8-2023

Statistical Graph Quality Analysis of Utah State University Master of Science Thesis Reports

Ragan Astle
Utah State University

STATISTICAL GRAPH QUALITY ANALYSIS OF UTAH STATE UNIVERSITY MASTER OF SCIENCE THESIS REPORTS

by

Ragan Astle

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Jürgen Symanzik, Ph.D. Major Professor	Kady Schneiter Ph.D. Committee Member
Alan Wisler, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2023

ABSTRACT

Statistical Graph Quality Analysis of Utah State University Master of Science Thesis Reports

by

Ragan Astle, Master of Science

Utah State University, 2023

Major Professor: Jürgen Symanzik, Ph.D.
Department: Mathematics and Statistics

The usage of graphical software packages has become increasingly popular in our modern world. However, there is growing concern within the statistical visualization field that the default settings provided by these packages make it challenging to create good quality graphs that align with standard graph principles. This MS thesis aims to investigate whether the quality of graphs from Utah State University (USU) Plan A Master of Science (MS) thesis reports from the years 1930 to 2019 was affected by the rise of graphical software packages. We extracted all data stored on the USU Digital Commons website since November 2021 using regular expressions to determine the population of interest. The data was stratified based on plan and degree types and aggregated into five-year intervals. We established a sampling process to obtain the graphs and sampled a total of 90 graphs, five graphs within each aggregated five-year interval. To accurately judge graph quality, we compiled and condensed good graphic standards from the statistical literature to develop our own set of graph quality criteria. We grouped the criteria within four distinct categories, such as: *Labeling*, *Clear Understanding*, *Meaningful*, and *Scaling and Gridlines*. We constructed a scoring system to rate the quality of graphs against these criteria and collected the raw graph quality scores of the sampled graphs. Using R, we explored the raw graph quality scores through visualizations such as bar graphs, scatterplots, heat maps, and dendrograms, as well as through statistical methods, including locally estimated scatterplot smoothing, least squares regression lines, and hierarchical

clustering. We analyzed various relationships among variables such as overall accuracy, overall rating, criteria category accuracies, graph type, graph creation method, and time. We performed a hierarchical cluster analysis to explore the relationship between the developed graph quality criteria and sampled graphs. Finally, we evaluated whether the rise of graphical software packages impacted the quality of graphs within the USU Plan A MS thesis reports based on our assessment of the results.

(147 pages)

PUBLIC ABSTRACT

Statistical Graph Quality Analysis of Utah State University Master of Science Thesis Reports

Ragan Astle

Graphical software packages have become increasingly popular in our modern world, but there are concerns within the statistical visualization field about the default settings provided by these packages, which can make it challenging to create good quality graphs that align with standard graph principles. In this thesis, we investigate whether the quality of graphs from Utah State University (USU) Plan A Master of Science (MS) thesis reports from the years 1930 to 2019 was affected by the rise of graphical software packages. We collected all data stored on the USU Digital Commons website since November 2021 to determine the specific group of graphs we wanted to investigate and developed a sampling process to obtain a sample size of 90 graphs evenly distributed over the time range. To accurately judge graph quality, we compiled and condensed good graphic standards from the statistical literature and developed our own set of graph quality criteria, grouped within four distinct categories: *Labeling*, *Clear Understanding*, *Meaningful*, and *Scaling and Gridlines*. We constructed a scoring system to rate the quality of graphs against these criteria and explored the results by constructing several visualizations and performing various statistical analyses. Our analysis assessed whether the rise of graphical software packages impacted the quality of graphs within the USU Plan A MS thesis reports.

ACKNOWLEDGMENTS

I express my deepest and most sincere gratitude to my advisor, Dr. Jürgen Symanzik, for his incredible support and guidance throughout my thesis research. His attention to detail and ability to provide constructive feedback have taught me profound life lessons that have not only made me a better writer and statistician, but have also taught me to be more thorough and thoughtful in all different aspects of my life. His patience and dedication to my success have truly been invaluable to me. I am grateful for the countless hours Dr. Symanzik has spent advising me throughout my research and I feel incredibly fortunate to have had the opportunity to work with him.

I thank my committee members, Dr. Kady Schneiter and Dr. Alan Wisler, for their continued support and guidance as they have served time on my thesis committee and have provided me with very meaningful feedback and insights. I am grateful to the Utah State University Mathematics & Statistics Department for their assistance in helping me to achieve my goals and progress as a student. The staff, specifically the Graduate Program Coordinator, Gary Tanner, have been incredibly helpful with answering questions and providing me with a path to success.

I am grateful for my amazing husband, Zach, for his continued love, support, and strength. His encouragement to push through difficult situations has inspired and helped me achieve far more than I ever thought I would be able to. Finally, I thank my sweet daughter, Monroe, who has been with me throughout the hardest moments of this process and has been my biggest cheerleader from inside the womb.

RA

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Research Overview

In the statistical visualization and data science world today, there exist numerous, powerful software and tools for graphing data. These tools allow for people across countless fields and with varying levels of experience to easily create graphs.

In 1976, John Chambers and his colleagues at Bell Laboratories initiated S, which is a computational language and environment for data analysis and graphics (Venables and Ripley, 2000). This was one of the first statistical software packages that became widely available and changed the way people analyzed, visualized, and manipulated data. In 1988, S-PLUS was developed as an extension to S, with added functionality for statistical modeling and analysis (Venables and Ripley, 2000). While S was built as an open-source language, S-PLUS was created as a commercial product, which increased the popularity of the software among users. The R software environment for statistical computing and graphics (R Core Team, 2021), which was developed by Robert Gentleman and Ross Ihaka (Ihaka and Gentleman, 1996), was released in 1995 as a free, open-source alternative to S and S-PLUS (Peng, 2015). Since R does not require a license to use and contains highly advanced features for statistical analysis, visualization, and data science, it has become one of the most popular statistical software packages today.

In 1983, Mitchell Kapor and Jonathon Sachs developed a spreadsheet software application called Lotus 1-2-3 (Raković et al., 2014). Lotus 1-2-3 followed the success of the first spreadsheet software, VisiCalc, which was introduced in 1979 and only had a few basic data capabilities. According to Raković et al. (2014), Lotus 1-2-3 "enabled chart creation, as well as some database capabilities. It was the first program that supported operations on call ranges, cell naming and macros." In 1985, Microsoft released Excel, which was the first spreadsheet software to offer a graphical user interface (Raković et al., 2014). As the capabilities within Excel have become more

refined and advanced, its popularity among users has increased. Now Excel is one of the most widely used software packages for data storage, organization, and visualization.

Since the developments of these and other similar software packages, the popularity and usage of graphs has increased considerably as these graph creation tools are utilized in a wide range of settings, from professional businesses and organizations to elementary-level computer classes. The ability to compile and convey data in a visual way has never been so convenient.

Graphs are essentially diagrams that are used to represent data in an organized manner and communicate relationships between variables, both quantitatively and categorically. Through graphs, clear messages can be relayed to viewers in the form of visual representations rather than by numbers on a spreadsheet. Moreover, graphs are highly effective because they present information quickly and easily. In many cases, graphs invite further investigation and analysis of topics that may not otherwise be explored.

The purpose behind graph creation seems to be shifting from what it used to be. Instead of focusing on emphasizing the data as much as possible, the motivation seems to center around designing graphs to be eye-catching, attractive, and to bring data to life. While it is ideal for graphs to be engaging, there are general guidelines and practices that should be adhered to when creating graphs. Since graphical software packages have been made so widely available, it is a growing concern in the statistical visualization field that many of the standard graph principles are being forgotten, ignored, or have never been learned before (Gordon and Finch, 2015).

Most statisticians learn at some point in their education how to construct good quality graphs based on standards set by various experienced and educated data scientists, such as John Tukey (Tukey, 1972), Bill Cleveland (Cleveland, 1985), Howard Wainer (Wainer, 1984), Edward Tufte (Tufte, 1983), Naomi Robbins (Robbins, 2005), and more. However, many statisticians and scientists do not use graphs frequently enough (Gordon and Finch, 2015). This creates a serious problem since statisticians, as Cleveland (1984) emphasized, play a vital, leading role in improving graphical visualizations and data communication.

In 2015, Ian Gordon and Sue Finch (Gordon and Finch, 2015) conducted research to investigate whether or not statisticians were creating high quality graphs that aligned with the good graphic

principles taught in their education. For this research, Gordon and Finch (2015) looked at top-rated applied science and statistics journals and analyzed the graphs published in them. They determined five principles of graphical excellence based on insights from Cleveland and Tufte and evaluated a sample of the published graphs against them. In their research results, they found that no graphs were rated as exemplary according to their principles. In fact, 39% of the graphs sampled were rated as poor. While the overall quality in the applied science graphs was slightly lower than the statistics graphs, fewer statistics graphs stood alone compared to the applied science graphs. When a graph stands alone it means that it can be interpreted without the viewer needing to reference text or search for explanations. When a graph does not stand alone, it creates a communication gap for the viewer as they have to spend more time and energy deciphering the graph. Gordon and Finch (2015) reasoned that when creating graphs, statisticians tend to have too much familiarity with the data and assume details are obviously laid out for an outside viewer. Statisticians should not treat the creation process as straightforward, but rather rely heavily on the skills and expertise that they acquired and continue to learn. Furthermore, the results from the study conducted by Gordon and Finch (2015) called attention to the lack of connection between valued good graphic principles and actual practice.

Many people are not taught what statisticians and scientists are taught about graphics, yet can easily create graphs through available graphical software packages and present them to vast audiences. Thus, while it is important for statisticians to set an example of graphical excellence, it is also essential that graphical software packages provide adequate functions and settings so that everyday users can avoid making fatal mistakes when constructing graphs.

An extensive amount of common graphical mistakes made by users stem from the default settings that many graphical software packages implement (Gordon and Finch, 2015). Ideally, the default settings of graphical software packages should make it easy to produce graphs that meet the fundamental graph quality standards, especially since most users do not have the necessary knowledge to guide them in the graph creation process. Unfortunately, it seems that most available graphical software packages make it difficult to produce adequate graphs. Since graphical software packages have become so much more advanced over time, many of their default settings are now

centered around artistry and encouraging the user to feel like they can create complicated yet beautiful visuals through simple button clicks (Su, 2008). However, a dangerous line is crossed when added features start to obscure the information trying to be portrayed. Tufte (1997) said, "Data graphics should draw the viewer's attention to the sense and substance of the data, not to something else. The data graphical form should present the quantitative contents. Occasionally artfulness of design makes a graphic worthy of the Museum of Modern Art, but essentially statistical graphics are instruments to help people reason about quantitative information."

Symanzik et al. (2016) observed that the most frequent problems encountered among the winning posters of the American Statistical Association (ASA) Poster Competition from the years 2013 to 2016 were three-dimensional (3D) bar charts and pie charts. Interpreting information presented in a 3D format can be misleading since 3D elements distort data in a way that makes estimating angles and deciphering the involvement of shadows difficult. Additionally, Symanzik et al. (2016) indicated that "each dimension of a graph represents one variable of the data" and that "adding a third dimension suggests that we know a third thing about the data." Thus, incorporating 3D elements in a graph is misleading and deceptive to the viewer. As Tufte (1997) expressed, artistry should never distract the viewer from the importance of the data. However, it seems to be increasingly common to find graphs that contain unnecessary features and designs.

Microsoft Excel is an example of a software product that does not reflect appropriate statistical visualization standards and practices. In fact, many of the default graph types that Excel offers can easily distort essential information with meaningless symbols and purposeless graphical elements (Su, 2008). These included features, which are referred to as 'chartjunk' by Edward Tufte (Tufte, 1997), do not add any substance to the information presented and tell the viewer nothing new about the data. Instead, these unnecessary decorations distract the viewer from important takeaways. Chartjunk is becoming easier to access and include in visualizations due to the default settings in graphical software packages. As mentioned, Excel offers numerous default charts that violate the basic statistical graph principles, with 3D elements being one of the most common issues found in these graphs (Su, 2008). Excel guides users to make the 3D graph mistake by aligning the 3D graph options directly next to the 2D alternatives in the main graph menu. Settings such as this and

more require users to be educated in the main principles of good statistical graphs, yet so many are not. Common users of Excel typically rely on convenience and artistry when creating graphs, thus falling back on the default settings provided to them. Su (2008) indicated that "to make junk charts in Excel, users just need a few clicks of the buttons. To clean up chartjunk, they require more than a few." While unrealized, it takes a great effort on the users part to create adequate graphs using the Excel software.

Additionally, there are also several statistical numerical issues that can be found in the Excel software. As is shown through a variety of examples by Leo Knüsel (Knüsel, 2002), Excel has the capability to compute negative variances, the random number generation feature does not meet basic statistical requirements, and multiple errors can arise from the computation of fundamental statistical distributions. Thus, for statistical purposes, Excel should be used with caution since it is unreliable.

Highlighting the statistical flaws of Excel, one of the most used graphical software packages, demonstrates how easily other available graphical software packages and tools can lead users to create inadequate graphs. Gordon and Finch (2015) stated: "Many graphical faults would disappear if software packages came with defaults that were consistent with principles of good graphics." As a society, we rely heavily on available graphical software packages and their corresponding default settings to create graphs. However, graphical software packages have not always existed. What was the quality of graphs like in the years prior to the introduction of graphical software packages? Were people required to be more knowledgable in the principles of good graphics? Has the quality of graphs changed with the evolution of graphical software packages? Have we as a society fallen into the trending trap of chartjunk and convenient default graph options?

1.2 Research Hypotheses

For this research, we wanted to investigate graph quality trends over time. Based on the discussion in Section 1.1, our hypothesis is that since the rise of graphical software packages, graph quality has declined due to the inappropriate reflection of statistical visualization standards and practices in these widely available tools. Furthermore, we believe that prior to the invention of graphical soft-

ware packages, graph creators had to rely heavily on principles of good graphics, whether taught in their education or researched solely for the purpose of creating graphs. Based on this belief, we suspect that in earlier created graphs, principles of good graphics are more pronounced. Before graphical software packages, default templates and settings did not exist, which leads us to further suspect that the typical chartjunk we find in graphs today will not be as prevalent in earlier graphs. Thus, we ultimately believe that graphs created before the development of graphical software packages are of higher quality than graphs created since the rise of graphical software packages.

1.3 Exploratory Questions

Before beginning any initial research, we wanted to develop a few questions that would be the driving force and focus of our analysis. We established two exploratory questions to center our analysis around. For the first exploratory question, we asked: What kinds of statistical graphs have been used primarily in a certain set of publications? We then broke down this general question into more specific questions, such as: Which types of graphs are the most common overall? What do the use trends of these statistical graph types look like over a certain time period? How were the graphs created, i.e., by hand, typewriter, computer software? For the second exploratory question, we asked: How has the quality of the graphs changed over time? Similar to the first exploratory question, we broke this question down into more distinct components, such as: What kind of criteria and rating system will be the most effective in order to accurately judge the quality of statistical graphs? How has the overall quality of the graphs changed? How has the quality of different graph types changed? Can we analyze trends of the developed criteria over time as well? Using these two exploratory questions to guide our investigation, we explored how graph quality has changed over time and examined the effects that graphical software packages have on graph quality.

1.4 Thesis Outlook

In order to conduct this analysis, we needed to determine a population of interest to sample graphs from and develop a set of criteria to judge and rate the sampled graphs by, similar to the study conducted by Gordon and Finch (2015). In Chapter 2, we discuss how we determined graphs from Utah State University (USU) Plan A Master of Science (MS) theses as our population of interest.

We also highlight the data extraction and graph sampling processes. In Chapter 3, we explain the creation of the graph quality criteria and scoring system that were used to evaluate the quality of the sampled USU Plan A MS thesis graphs. In Chapter 4, we discuss the statistical methods and software that were used to analyze the raw graph quality scores that were collected from scoring the USU Plan A MS thesis sampled graphs according to the developed graph quality criteria and scoring system. In Chapter 5, we examine and assess the collected raw graph quality scores. In Chapter 6, we discuss, interpret, and compare the results presented in Chapter 5 to our initial hypotheses made in Section 1.2. Chapter 7 outlooks on other potential ways we could investigate graph quality trends over time and how the outcomes may change depending on different parameters or populations used.

Appendices A, B, C, D, and E provide additional information within this research. In Appendix A, we provide a lookup table for all the sampled USU Plan A MS thesis graphs. In Appendix B, we list all of the good graphic standards from the various graph quality literature resources discussed in Chapter 3. In Appendix C, we display the Excel files containing the raw graph quality scores of the USU Plan A MS thesis graphs that were not featured in Chapter 5. In Appendix D, we summarize the observed counts of the different notes used to identify unique traits within the graphs sampled. In Appendix E, we present scatterplots of the overall ratings of the graph creation methods used within our sample over time and assess the significance of their relationships.

CHAPTER 2

Available Data and Population of Interest

A necessary element to begin the research process of assessing statistical graph quality over time was the data collection procedure. In Section 2.1, we discuss the overall data contained within the Digital Commons website at Utah State University (USU). In Section 2.2, we explain how the data was extracted from the Digital Commons website at USU, how the population of interest was determined, and how the individual graphs, which are the object of study, were selected to be judged and analyzed against the developed graph quality criteria that are detailed in Chapter 3.

2.1 Utah State University Digital Commons Overview

All assessed graphs were collected and sampled from the USU Digital Commons website using methods that will be described in Section 2.2. The Digital Commons is an online library that allows numerous institutions, including USU, to manage, publish, and exhibit research completed by both faculty and students (DigitalCommons, 2023). It provides universities and colleges with unlimited storage and file sizes, as well as ensures all uploaded content is safe and protected.

The USU Digital Commons website contains past student theses, dissertations, projects, and reports. There are two separate webpages, a Plan A webpage and a Plan B webpage. The Plan A webpage features all theses and dissertations (DigitalCommons@USU, 2021b), whereas the Plan B webpage houses all projects, reports, and some theses that are considered of Plan B type (DigitalCommons@USU, 2021a). The documents stored on the USU Digital Commons website date from 1923 to the present day and include files from all degree types and departments that have existed throughout the history of USU. As of November 14, 2021, there were a total number of 9,602 files recorded within the USU Digital Commons website, with 8,074 files contained on the Plan A webpage, and 1,528 files contained on the Plan B webpage.

Documents are stored and accessed within the USU Digital Commons through clickable links that are organized chronologically by year. Each individual link is associated with either a Plan

A or Plan B document depending on whether it is stored on the Plan A or Plan B webpage. The first segment of the link is to directly access the portable document format (PDF) of the respective file, whereas the second segment, which provides the title of the file as well as the name of the author, takes users to a separate webpage that contains all the document information. This document webpage displays the title, the author, and provides the option to download the PDF of the file. This webpage also contains the following headings: *Date of Award*, *Document Type*, *Degree Name*, *Department*, *Committee Chair(s)*, *Committee*, *Abstract*, *Checksum*, *Recommended Citation*, and *Digital Object Identifier (DOI)*.

The *Date of Award*, formatted as MM-YYYY, refers to the month and year that the document was officially completed and approved. The *Document Type* signifies whether the file is a dissertation, thesis, creative project, or a report. The *Degree Name* classifies what field the document is associated with, such as: Master of Science (MS), Master of Arts (MA), Master of Computer Science (MCS), etc. The *Committee Chair(s)* and *Committee* headings name the individuals that worked with and guided the student through the research they conducted. The *Abstract* gives a brief summary on the subject of the paper. The *Checksum* is an alphanumeric value that specifically denotes the contents of the file and can act as a fingerprint in which comparisons can be made in order to detect errors and check the integrity of the file. The *Recommended Citation* displays the citation of the file that is to be used in order to properly give credit to the author and corresponding material. The *DOI* is a string of numbers, letters, and symbols assigned to the file in order to uniquely identify it and provides a link directly to the document and its information.

To understand the background of the overall data on the USU Digital Commons website, we reached out to the USU library for additional information. In September 2022, Becky Thoms, head of the Digital Initiatives department that manages the USU Digital Commons website, provided some insight on the history of the Digital Commons website at USU. As of September 2022, not all historic theses, dissertations, projects, and reports have been scanned. There was a serious effort made in the years 2013 to 2015 to digitize the earliest documents, i.e., those within the years 1920 to 1930. However, the effort was not as comprehensive as was initially understood. From 1920 to 1931, there were about 124 documents produced at USU. Upon an investigation conducted by

Becky Thoms in September 2022, the Digital Initiatives department had only digitized and made 33 of those files available on the USU Digital Commons website during that effort. Apparently, the decision about what to digitize or not digitize at the time was mainly based on the condition of the material. Unfortunately, many of the files that were not digitized and uploaded were too fragile for the robotic scanner that USU used. Since the effort made back in 2013 to 2015, USU has updated and improved their scanning equipment. After learning about the unscanned documents, Becky Thoms stressed the need for the Digital Initiatives department to restart the project so that the remaining documents would be added to the USU Digital Commons website. She believes that the new scanning equipment will allow for those additional files produced during the years 1920 to 1931 to be made available on the USU Digital Commons website. For this research, no files added after November 14, 2021 have been considered.

Furthermore, according to the information received from Becky Thoms, 2008 was the year that USU decided to fully transition from physical to electronic deposit of dissertations, theses, creative projects, and reports. This implies that almost all of these documents since 2008 have been made available on the USU Digital Commons website, possibly with the exception of copyrighted or otherwise restricted materials. From the years 1931 to 2008, before all submissions were received in electronic format, the Digital Initiatives department primarily digitized based on demand. In the past, the Digital Initiatives department received requests from patrons for individual documents as well as from other USU departments to digitize all of their respective documents. Not all requests from USU departments could be accommodated, but priority was typically given to the Geology, Mathematics, Nutrition, Dietetics and Food Sciences (NDFS), and Psychology departments.

As of September 2022, the Digital Initiatives department was under the impression that prior to 2008, no year had all files completely scanned in and made available on the USU Digital Commons website. To be able to provide more in-depth statistics on the files uploaded from the different years, the Digital Initiatives department would need to complete a comprehensive audit. This would compare the physical holdings in the USU Library's Special Collections and Archives with lists of graduate students as well as with the USU Digital Commons website. Unfortunately, this is an unplanned future effort but the information will be useful for the Digital Initiatives department once

obtained. Given the imbalance of documents according to quantity, type, and department over the years, we knew we could not work with the full data on the USU Digital Commons website. Rather, we needed to find a population of interest that was consistent over time.

2.2 Data Collection

For the data collection process, the data first needed to be extracted into a format that could be easily grouped and analyzed. Next, the construction of plots was required in order to comprehend what types of documents and how much of each were stored on the USU Digital Commons website and how the overall data may have changed over time. From the created plots, our population of interest was straightforwardly determined based on the different groupings of the overall data. Samples then needed to be obtained in an appropriate but random manner, thus a hierarchical sampling process was developed. This data collection process took place in November 2021, making use of all 9,602 documents that were accessible on the USU Digital Commons website at that time.

2.2.1 Data Collection for the Determination of the Population of Interest

To begin, it was essential to obtain all information that was housed within USU's Digital Commons website, as of November 2021. Every file on both the Plan A and Plan B webpages needed to be accessed and the corresponding document webpage described in Section 2.1 was to be scraped in order to retrieve all listed information. Web scraping is a technique that uses a computer program to import data from websites into files or spreadsheets for various purposes, some of which include data manipulation and formatting. Lastly, all of the extracted data needed to be laid out in a data frame to make later groupings and assessments easier to manage.

Based on past educational experience and for simplicity of the code required, the R software environment for statistical computing and graphics (R Core Team, 2021) was used to extract all USU Digital Commons data. The most thorough way to approach the scraping process was through regular expressions. Regular expressions are defined as a pattern that describes a set of strings. Using a loop, each webpage was read into R in HyperText Markup Language (HTML) format. Once in this format, regular expressions allowed for the simple creation of variables based on the headings contained on the document webpages. Regular expressions also provided an uncomplicated way to

clean and transform text and resolve any inconsistencies in the data.

Next, the data was organized into a data frame. Using the R software, the compiled data was then exported into an Excel sheet to be further analyzed in order to determine a reasonable population of interest. A segment of this Excel file is found in Figure 2.1. It is important to note that when scraping the data on the USU Digital Commons website through regular expressions, other variables were created from headings picked up in the HTML code. These headings are not always, if ever, displayed on the document's webpage. These headings include: *Department When Degree Awarded*, *Award Number*, an extra *DOI* variable, *FundRef*, *Comments*, *Related Content*, *Previous Versions*, and *Streaming Media*. Based on the constructed Excel file, these columns most always contain the value 'NULL'. While these columns were useless with regards to determining a population of interest, they were still extracted and included in the Excel file so as to not lose any documented information pertaining to a file. Using the R software, the variables *Plan* and *Year* were created during the data extraction process in order to keep track of what USU Digital Commons webpage, Plan A or Plan B, and what year each document belonged to.

Abstract	Author	Checksum	Committee	Committee Chair(s)	Date of Award	Degree Name	Department	Document Type	Included In	Recommended Citation	Share	Title	doi	Department When Degree Awarded	Award Number	DOI	FundRef	Comments	Related Content	Previous Versions	Plan	Streaming Media	Year
The NASA	Troy A. A	ca504fa8eb	Douglas F. Hu	Douglas F. Hu	5-2021	Master o	Mechanical an	Thesis		<!-- FILE: /srv/seq	NA	Sonic Boom Loudness	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
This study	Motaser	e05f5c1677	Randal Martir	Randal Marti	12-2021	Master o	Civil and Envir	Thesis		<!-- FILE: /srv/seq	NA	Ammonia Emission As	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
In a relatic	M. A. Ma	9b5ac69e93	Curtis Dyreso	Curtis Dyreso	12-2021	Master o	Computer Scie	Thesis		<!-- FILE: /srv/seq	NA	Achieving a Sequencec	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
This resear	Shubhar	284f9933c1	Curtis Dyreso	Curtis Dyreso	8-2021	Master o	Computer Scie	Thesis		<!-- FILE: /srv/seq	NA	MetaXMorph: Hierarcl	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Our resear	Anthony	c7bc1faa67	Fernanda Bati	Fernanda Bat	12-2021	Master o	Animal, Dairy,	Thesis		<!-- FILE: /srv/seq	NA	Impact of Fish Oil on Ir	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
In many di	Al-Wale	78331db36	Todd K. Moor	Todd K. Moo	5-2021	Doctor o	Electrical and (	Dissertation		c("<!-- FILE: /srv/s	NA	Development of a Two	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
In the wes	Sara Ma	966b44f1a2	Belize A. Lane	Belize A. Lane	5-2021	Master o	Civil and Envir	Thesis	NULL	<!-- FILE: /srv/seq	NA	Summer Stream Temp	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
This study	Britney (	3928e7cb7	Jennifer Mac	Jennifer Mac	5-2021	Master o	Plants, Soils, a	Dissertation		<!-- FILE: /srv/seq	NA	Fatty Acid Compositio	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
The equin	Sarah J. .	01cd6a7bb2	Michael L. Pa	Michael L. Pa	12-2021	Master o	Applied Scienc	Thesis		<!-- FILE: /srv/seq	NA	Development and Vali	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Earthquak	Emma M	3b4e93c6c6	Alexis K. Ault,	Alexis K. Ault	8-2021	Master o	Geosciences	Thesis		<!-- FILE: /srv/seq	NA	Multi-Proxy Approach	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Approxim	Elizabet	784d817c0	Kari E. Veblen	Kari E. Vebler	8-2021	Master o	Wildland Reso	Thesis		<!-- FILE: /srv/seq	NA	Transplanting Mature	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
For centur	Karen Ba	a4822f46e8	Joseph Wheat	Joseph Whea	12-2021	Master o	Watershed Sci	Thesis		<!-- FILE: /srv/seq	NA	Valley Bottom Inundat	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
To design	Simon B	2601c38b2	Zachary B. Sh	Zachary B. Sh	5-2021	Master o	Civil and Envir	Thesis		<!-- FILE: /srv/seq	NA	Numerical Simulation	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
As water s	Kade J. E	8e7639b1e	Michael C. Jol	Michael C. Jo	8-2021	Doctor o	Civil and Envir	Dissertation		<!-- FILE: /srv/seq	NA	An Analysis of Electron	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
There are	Arun J. E	21d6ee6c5	David K. Gelle	David K. Gell	8-2021	Doctor o	Mechanical an	Dissertation		<!-- FILE: /srv/seq	NA	Attitude and Reflectio	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Radar is a	Thomas	5963dc7e2	Todd Moon, J	Todd Moon	8-2021	Master o	Electrical and (	Thesis		<!-- FILE: /srv/seq	NA	Alternative Doppler Ex	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Morbidity	Tevan J.	4b260e55c	Kara J. Thornt	Kara J. Thorn	5-2021	Master o	Animal, Dairy,	Thesis		<!-- FILE: /srv/seq	NA	The Effects of Trace Mi	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Our relatic	Michael	019a62270	Jennifer E. Gi	Jennifer E. Gi	8-2021	Doctor o	Sociology, Soc	Dissertation		<!-- FILE: /srv/seq	NA	Alienation, Moderniza	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
For certair	Christop	9c2eba0b7(	R. Douglas Ra	R. Douglas Ra	8-2021	Master o	Wildland Reso	Thesis		<!-- FILE: /srv/seq	NA	Using Unmanned Aeria	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Adaptive (	Emalee E	a929102ee	Lacee Bosche	Lacee Bosche	8-2021	Master o	Applied Scienc	Thesis		<!-- FILE: /srv/seq	NA	The Relationship of Ad	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Monitorir	Michael	15176e79e	Mary M. Conr	Mary M. Con	5-2021	Master o	Wildland Reso	Thesis		<!-- FILE: /srv/seq	NA	Monitoring Populatio	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Following	Jared Fo	6f4ac29121	Sulaiman K. N	Sulaiman K. M	12-2021	Master o	Nutrition, Diet	Thesis		<!-- FILE: /srv/seq	NA	Injection of Iodoacetic	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Socioscier	Michelle	eaed1a3f5a	Tyson J. Sorer	Tyson J. Sore	5-2021	Doctor o	Applied Scienc	Dissertation		<!-- FILE: /srv/seq	NA	Socioscientific Issues i	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
This proje	Stuart C	706627d99	Tammy Proct	Tammy Proct	8-2021	Master o	History	Thesis		<!-- FILE: /srv/seq	NA	Cultural Memory and I	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
The purpo	Patrick L	ead0af9ff9d	Zachary B. Sh	Zachary B. Sh	5-2021	Master o	Civil and Envir	Thesis		<!-- FILE: /srv/seq	NA	Application of Compu	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Digital sac	Neil Car	1c13ec792(	Cindy D. Jone	Cindy D. Jone	5-2021	Doctor o	School of Teac	Dissertation		<!-- FILE: /srv/seq	NA	Digital Scripture: An Ir	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Obtaining	Yessica A	e3e217e817	Foster A. Agbl	Foster A. Agb	5-2021	Doctor o	Biological Engi	Dissertation		<!-- FILE: /srv/seq	NA	Biomethanation and A	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
This thesis	Marie Vi	ca778c9504	Philip L. Barlo	Philip L. Barl	5-2021	Master o	History	Thesis		<!-- FILE: /srv/seq	NA	The Challenge of Hybri	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Climate ch	Ryan T. (	0fc678adc5	Karen H. Bear	Karen H. Bear	8-2021	Doctor o	Wildland Reso	Dissertation		<!-- FILE: /srv/seq	NA	Climate-Driven Impact	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
"Nobody i	Darcie C	f91cd6a3fe	Idalis Villanue	Idalis Villanu	8-2021	Doctor o	Engineering Ed	Dissertation		<!-- FILE: /srv/seq	NA	A Mixed-Methods App	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
The relatic	Jhonata	af2f854301	Jürgen Syman	Jürgen Symar	8-2021	Master o	Mathematics a	Thesis		<!-- FILE: /srv/seq	NA	Housing Variables and	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Examinati	Shelby R	NULL	Mollie Murph	Mollie Murpl	8-2021	Master o	Languages, Phi	Thesis		<!-- FILE: /srv/seq	NA	The Rhetoric of the Do	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
The state (	Matthev	446054a09	Patrick Single	Patrick Singl	12-2021	Master o	Civil and Envir	Thesis		<!-- FILE: /srv/seq	NA	Active Transportation	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021
Sepsis is d	David Sc	366199e41	Robert E. Wa	Robert E. Wa	5-2021	Master o	Nutrition, Diet	Thesis		<!-- FILE: /srv/seq	NA	Effect of Dietary Short	https://doi.org/10.26076/	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	2021

Fig. 2.1: Segment of the Excel file containing all extracted data from the USU Digital Commons website.

2.2.2 Determining the Population of Interest

After the data were properly organized in a data frame format in R, we determined a subset of documents that would be reasonable to sample from for further graph quality research. As will be described in this section, the data were grouped and sorted using the R software and the developed data frame.

First, the data were separated into Plan A and Plan B groups. By grouping the documents and their respective information in this manner, the amount and types of documents within the two plan types could be easily assessed. Based on this grouping, there were a total of 8,074 Plan A documents and a total of 1,528 Plan B documents. Thus, the Plan A webpage contained over five times as many documents as the Plan B webpage as of November 2021.

After separating the data based on plan type, the two groups were further stratified into the separate document types. As highlighted in Section 2.1, the document types include: dissertations, theses, creative projects, and reports. From Table 2.1, we can see that as of November 2021, there were over two times as many theses as dissertations within the Plan A grouping. From Table 2.2, it is apparent that reports, as of November 2021, were the most common type of document within the Plan B grouping. When comparing Tables 2.1 and 2.2, we can see that theses were more commonly classified as Plan A rather than Plan B. Furthermore, the only common document types between the two plan type groups were theses, however, there were nowhere near the same amount of theses between the two groups. This suggested that the population of interest should source from one plan type instead of being sampling from both since there was little to no commonality between the two groups as far as document types.

Table 2.1: Plan A document type counts as of November 14, 2021.

Dissertations	2,459
Theses	5,615
Total	8,074

Table 2.2: Plan B document type counts as of November 14, 2021.

Creative Projects	281
Reports	1,057
Theses	190
Total	1,528

Next, it was important to look at how the grouped data appeared over time. Since the records date from 2021 back to 1923, it made sense to group years together in order to more easily visualize trend patterns. Thus, the data were aggregated into five-year intervals and were classified by the lower bound year within each interval, i.e., 1970, 1971, 1972, 1973, and 1974 were all organized and collected into the 1970 five-year interval, whereas 1975, 1976, 1977, 1978, and 1979 were allotted to the 1975 five-year interval.

After the data were grouped based on plan type, further stratified according to document type, and the years spanning the data were aggregated into five-year intervals, plots were created to assess the sampling strategy and determine which subset of the overall data would be the most ideal to evaluate. Figure 2.2 displays the document counts for the different document types over time for each plan type group. As previously mentioned, time is aggregated into five-year intervals from the years 1920 to 2020.

Based on Figure 2.2, it is evident that the Plan A group had much higher document counts than the Plan B group overall, and this appears to be the consistent pattern throughout the years. It is important to note that the dip in the 2020 aggregated five-year interval for both plots is due to the fact that the interval only contains the years 2020 and most of 2021, whereas all other intervals, except for 1920, contain their full respective five years. The 1920 aggregated five-year interval only includes the years 1923 and 1924. It is clear that theses are the dominant document type for the Plan A group and reports are the dominant document type for the Plan B group. Plan A theses have the longest time span compared to all the other document types within the two plan type groups. Plan A dissertations only date back to the 1950 aggregated five-year interval whereas Plan A theses date back to the 1920 aggregated five-year interval. When looking at the Plan B plot, the earliest

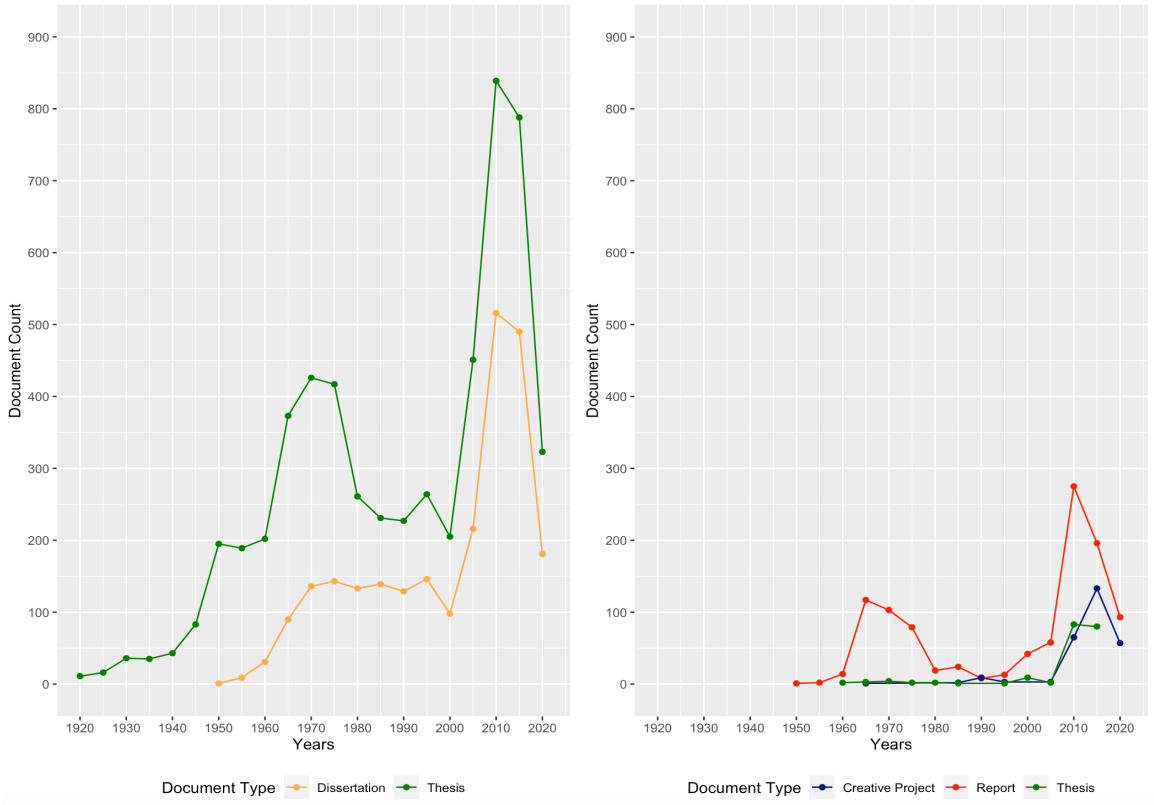


Fig. 2.2: Line Graphs of Plan A documents (left) and Plan B documents (right), stratified by document type and aggregated into five-year intervals from 1920-2020.

recorded document type is in the 1950 aggregated five-year interval.

From Figure 2.2, there appears to be a general increasing trend for both plan types. However, it is interesting to note a common dip in document counts between the aggregated years of 1980-2000 for both plan types. Based on the insight Becky Thoms provided on document requests, which is highlighted in Section 2.1, we might assume that the demand for digitized documents was lower for the years 1980 to 2000.

After analyzing the different document types among Plan A and Plan B documents, it was determined to focus research on Plan A theses as they date back to the 1920s and contain the largest counts overall for each aggregated five-year interval. Upon narrowing the scope for the population of interest to only consider Plan A theses, the data needed to be stratified further to define an even more specific population of interest. Thus, the Plan A theses population was additionally grouped by the unique degree names, see Figure 2.3.

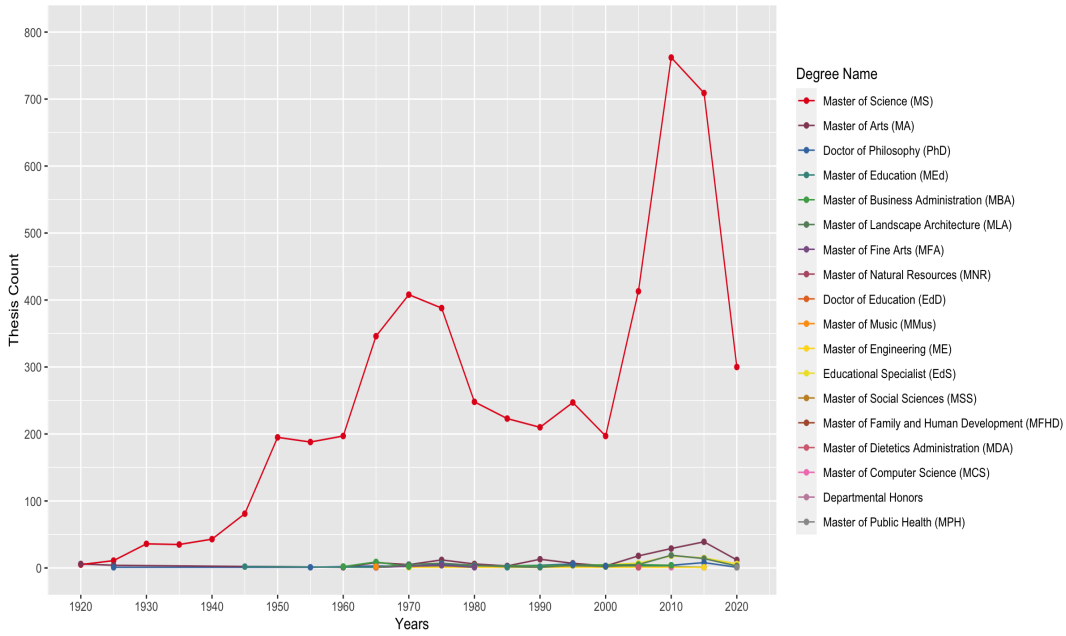


Fig. 2.3: Line Graph of Plan A theses, stratified by degree name and aggregated into five-year intervals from 1920-2020.

Figure 2.3 displays the breakdown of thesis counts by degree names within the Plan A group. The Master of Science (MS) degree has the highest thesis counts throughout the years. The rest of the degree names have counts so small it is almost impossible to correctly read off these counts from Figure 2.3. Thus, after analyzing the different degree types among Plan A theses, it was determined to focus research specifically on MS theses as they contain the largest counts for each aggregated five-year interval.

Next, it was essential to pick an appropriate year range that the Plan A MS theses would be sampled from. While it was important that the developed graph quality criteria were tested on external, unrelated graphs, it was also vital to test the criteria on a subset of the population of interest once determined (see Section 3.4). This was done to ensure that the criteria are clear, easily translate, and appropriately reflect the graph quality of the population of interest. It also aided in estimating how much time would be needed to assess each individual graph during the actual graph quality assessment process in order to determine a suitable quantity of graphs that should be sampled from each aggregated five-year interval. In order to provide a large enough time-span for testing, and after investigating that there were graphs contained in a reasonable number of theses from the

earliest aggregated five-year intervals, it was decided that the samples would be selected from each of the aggregated five-year intervals contained within the years 1930 to 2019, thus finalizing our population of interest. The year range 1930 to 2019 consists of 18 aggregated five-year intervals total, i.e., 1930, 1935, ..., 2015. Data within the same category, Plan A MS theses, but from the 1920, 1925, and 2020 aggregated five-year intervals, was used for testing purposes only as discussed in Section 2.2.3.

2.2.3 Testing Process

The next step was to test the developed criteria (see Section 3.2) on graphs in Plan A MS thesis reports outside of the designated sampling region. As discussed in Section 2.2.2, this provided a way to train for the real sampling and scoring process and to determine a reasonable sample size for each aggregated five-year interval within the sampling region. The Plan A MS thesis reports from 1923 to 1929 and 2020 to 2021 were used for the testing process. Several graphs contained within the theses in this testing time period were judged against the developed graph quality criteria (see Section 3.4). The amount of time it took to complete the scoring process was recorded and the amount of graphs contained in each of the reports was investigated. Two examples of graphs from the time period used for the testing process can be found in Figures 2.4 and 2.5. In Section 3.5, the developed graph quality criteria are discussed according to Figures 2.4 and 2.5, highlighting what problems exist within these graphs and which criteria are not met.

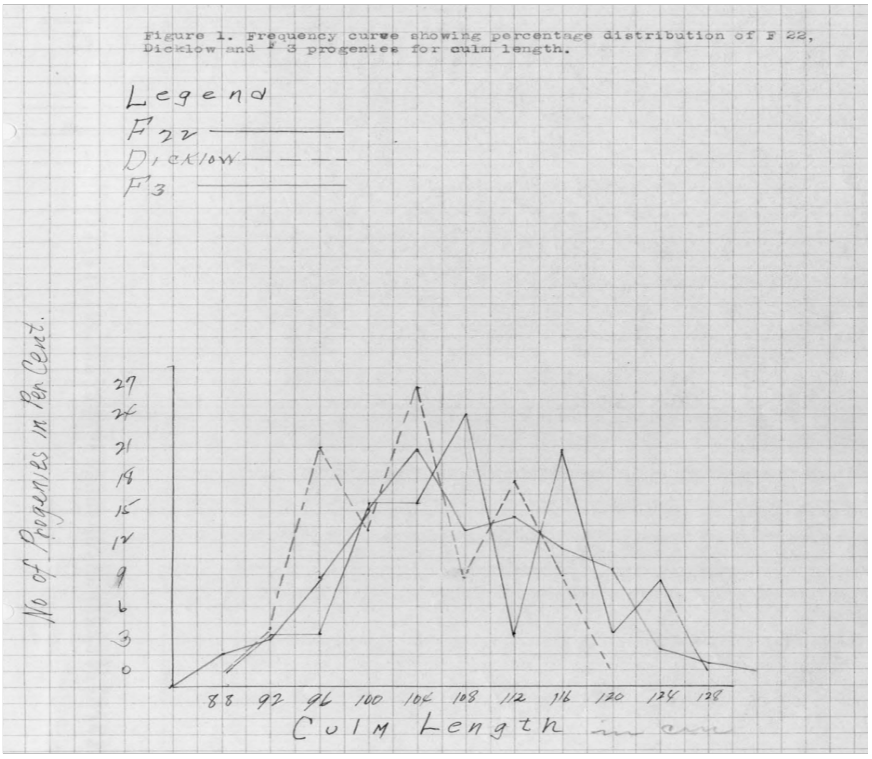


Fig. 2.4: Example of a graph within the 1925 aggregated five-year interval that was examined during the testing process.

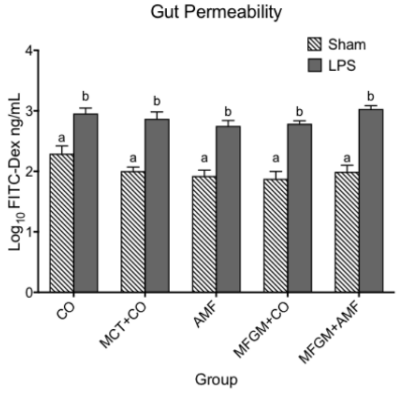


Figure 11. Gut permeability measurements. Effect of diets on fluorescence levels after Sham or LPS challenge. An n=6 is assigned per group. Data are log10 transformed mean nanogram per milliliter. No significant differences in fluorescence levels were seen between any diets for LPS challenged mice.

Fig. 2.5: Example of a graph within the 2020 aggregated five-year interval that was examined during the testing process.

2.2.4 Number of Graphs in Each Time Interval

After analyzing various graphs within the testing time period and judging them based on the developed graph quality criteria, an average of 10 to 15 minutes was typically spent scoring each graph. Based on this average time, a sample size of five graphs was determined to be taken from each aggregated five-year interval within the sampling region. Thus, with five graphs being randomly selected from each of the 18 aggregated five-year intervals, a total of 90 graphs would be sampled overall. If time had permitted, more than five graphs could have been sampled and assessed per five-year interval, however, with the given time constraints on this research, five graphs per interval was the resulting quantity. Due to the small sample size amount, the resulting five graphs from each five-year interval needed to come from separate thesis reports to avoid sampling bias.

2.2.5 Collecting Graph Samples and Determining the Order of Assessment

Next, the Plan A MS thesis graph samples from the years 1930 to 2019 needed to be obtained. Theses within each five-year interval were permutated with a seed set as the year value corresponding to the interval. Figure 2.6 displays a snippet of the permutation for the 1930 aggregated five-year interval. Using R, the rows, or rather, theses, within the 1930 five-year interval were reordered using a seed value of 1930. This restructured theses sequence was then exported to a sheet labeled '1930' in an Excel file containing all other interval permutations.

Once a random order of theses was established for each five-year interval, the next step was to select the graphs. This process consisted of collecting the first five theses in the permutation order for each aggregated five-year interval and then sampling one graph from each report. It is important to note that in some cases, a thesis report had multiple graphs contained within it. In this scenario, the graphs within that thesis would be assigned numbers from one up to the maximum number of graphs contained within the thesis, and one number, or in other words, graph, was chosen at random using a seed set as the report's corresponding document number. In other cases, a thesis report did not contain any graphs at all. In this circumstance, the next thesis in the permutation order would be used instead, and so forth, until a total of five graph samples were obtained for each aggregated five-year interval. A lookup table identifying the 90 sampled graphs can be found in Appendix A.

Abstract	Author	Checksum	Committee	Committee	DateofAwa	DegreeNar	Departmer	Document	Included Ir	Recommer	Share	Title	doi	Departmer	Award Nun	DOI	FundRef	Comments	Related Co	Previous V	Plan	Streaming	Year	Year1
The questic	Duncan Wa	dd22d108	Franklin Da	Franklin Da	5-1930	Master of S	Economics	Thesis		<!-- FILE: /s	NA	Some Econ	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1930	1930
The purpos	Roy A. Wes	361270afd	Joseph A. G	Joseph A. G	5-1931	Master of S	Sociology,	Thesis		<!-- FILE: /s	NA	The Young I	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
The benefic	Clarence Bu	a37793e3e	D. W. Pittm	D. W. Pittm	5-1931	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	A Study of I	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
Here-in is g	Ernest W. P	7c58aa4c2	J. S. Stanfor	J. S. Stanfor	5-1933	Master of S	Wildland R	Thesis		<!-- FILE: /s	NA	Bird Studie	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1933	1930
Antidating	Orval E. Wi	bd7efd7ae	D. W. Pittm	D. W. Pittm	5-1934	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Studies on	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1934	1930
The reclam	Lemoyne W	8dec7fcc0	D. W. Pittm	D. W. Pittm	5-1932	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Fertilizer St	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1932	1930
This study v	Carmen D.	0f108e1bc	Laura Watk	Laura Watk	5-1934	Master of S	Sociology,	Thesis		<!-- FILE: /s	NA	A Study of I	https://doi	Sociology	NULL	NULL	NULL	This work n	NULL	NULL	A	NULL	1934	1930
By reason o	George R. H	d87e168dc	Johnson	Johnson	5-1930	Master of S	Animal, Da	Thesis		<!-- FILE: /s	NA	A Research	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1930	1930
Present-da	C. Leland D	7b361773	George Ste	George Ste	5-1931	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Inheritance	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
Perennial v	Lionel Harr	b40ed7ef0	D. C. Tinge	D. C. Tinge	5-1930	Master of S	Agricultura	Thesis		<!-- FILE: /s	NA	The Use of (	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1930	1930
This paper i	Leslie W. N	525af2370	George Ste	George Ste	5-1931	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Inheritance	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
What pract	Elmer Jepp	749ccbeb1	E. A. Jacobs	E. A. Jacobs	5-1932	Master of S	School of T	Thesis		<!-- FILE: /s	NA	Fire Insurar	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1932	1930
The flavor,	George F. Ja	b50ba209c	A. J. Morris	A. J. Morris	8-1934	Master of S	Nutrition, I	Thesis		<!-- FILE: /s	NA	Some Facto	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1934	1930
"A coopera	Arthur J. M	e5763b015	George B. C	George B. C	5-1930	Master of S	Animal, Da	Thesis		<!-- FILE: /s	NA	A Study of I	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1930	1930
The purpos	Lowell E. St	f521709b1	E. A. Jacobs	E. A. Jacobs	8-1934	Master of S	Education	Thesis		<!-- FILE: /s	NA	An Evaluati	https://doi	Education	NULL	NULL	NULL	This work n	NULL	NULL	A	NULL	1934	1930
The potato	Wylie L. Th	7094e99bb	G. F. Knowl	G. F. Knowl	5-1934	Master of S	Biology	Thesis		<!-- FILE: /s	NA	A Study of t	https://doi	Entomolog	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1934	1930
"Pasture is	Newel Was	638e3352a	George B. C	George B. C	5-1931	Master of S	Animal, Da	Thesis		<!-- FILE: /s	NA	Influence o	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
The chief p	Alfred B. Ha	bdd86619l	Dean E. A. J	Dean E. A. J	5-1934	Master of S	School of T	Thesis		<!-- FILE: /s	NA	The Evoluti	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1934	1930
National Er	W. Wendel	4bde9036a	E. A. Jacobs	E. A. Jacobs	5-1934	Master of S	School of T	Thesis		<!-- FILE: /s	NA	An Evaluati	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1934	1930
Alfalfa has a	George Wh	059abd8f7	D. S. Jennir	D. S. Jennir	5-1932	Master of S	Agricultura	Thesis		<!-- FILE: /s	NA	Alfalfa Seed	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1932	1930
During rece	Bion Tolma	a21dc6f2a	D. C. Tinge	D. C. Tinge	5-1933	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Inheritance	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1933	1930
For many y	Lealand A. (	6ae8d3fe9	John W. Ca	John W. Ca	5-1931	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Temperatu	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
The purpos	Wilford Pri	e4dd9b3a2	E. A. Jacobs	E. A. Jacobs	5-1932	Master of S	School of T	Thesis		<!-- FILE: /s	NA	The Constru	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1932	1930
This study v	Eleonora Je	5f03b4f6e	Christine B	Christine B	5-1933	Master of S	Nutrition, I	Thesis		<!-- FILE: /s	NA	A Dietetic A	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1933	1930
The prime c	Z. Doyle Ro	0791a583c	Arthur J. N	Arthur J. N	5-1933	Master of S	Nutrition, I	Thesis		<!-- FILE: /s	NA	Pasteurizec	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1933	1930
Irrigation is	Prabh Dyal	0803b853a	O. W. Israel	O. W. Israel	5-1930	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Principles c	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1930	1930
Administra	Dean F. Pet	84896890	NA	NULL	5-1933	Master of S	School of T	Thesis		<!-- FILE: /s	NA	Certain Fac	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1933	1930
Due to the	George Tho	8694f46c4	W. P. Thom	W. P. Thom	5-1931	Master of S	Applied Ecc	Thesis		<!-- FILE: /s	NA	A Farm Org	https://doi	Agricultura	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
It was early	Thelma Hul	c4d175edd7e2a25eec	NULL		5-1931	Master of S	Family Con	Thesis		<!-- FILE: /s	NA	A Study of t	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
This paper i	LaGrande S	c4d6569b7	NA	NULL	5-1930	Master of S	Ecology	Thesis		<!-- FILE: /s	NA	The Spruce	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1930	1930
Due to the	Dwight Koc	d06826a55	D. W. Robe	D. W. Robe	5-1931	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Genetic Stu	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
Present-da	C. Leland D	46cdde734	George Ste	George Ste	5-1931	Master of S	Agricultura	Thesis		<!-- FILE: /s	NA	Inheritance	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1931	1930
This investi	Leah D. Me	81486b25	E. A. Jacobs	E. A. Jacobs	5-1933	Master of S	School of T	Thesis		<!-- FILE: /s	NA	A Study of t	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1933	1930
Many facto	Martha C. E	cbb625a6e	Christine B	Christine B	5-1932	Master of S	Family Con	Thesis		<!-- FILE: /s	NA	A Study of t	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1932	1930
In 1927 an	Delos Zobe	e89ec6adf8	D. W. Pittm	D. W. Pittm	5-1932	Master of S	Plants, Soil	Thesis		<!-- FILE: /s	NA	Fertilizer St	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1932	1930
The reclam	Orville L. El	8b24be10e	Willard Gai	Willard Gai	5-1932	Master of S	Physics	Thesis		<!-- FILE: /s	NA	The Design	https://doi	NULL	NULL	NULL	NULL	NULL	NULL	NULL	A	NULL	1932	1930

Fig. 2.6: Permutation example of the 1930 aggregated five-year interval. The Year column is highlighted in yellow to emphasize the randomized order of the theses contained within the interval.

The last step in the sampling process was to determine an appropriate order of assessment for the 90 sampled graphs. It is not proper technique to assess the graphs working through one aggregated five-year interval at a time, nor is it adequate to assess the five graphs for each five-year interval in the exact same chronological sequence, i.e., 1930, 1935, ..., 2010, 2015. Thus, to suitably handle the order of graph evaluation, the list of 18 aggregated five-year intervals was permutated five times, with a seed set using the values 1001, 1002, ..., 1005, respectively. These five distinct five-year interval sequences were created to determine the order by which the graphs should be evaluated. The five different permutation sequences can be found in Figure 2.7. The first sequence indicates the order of assessment of the first graphs selected in each of the five-year intervals, i.e., starting with the graph from 1940, then from 2000, then from 2005, and so on. The second sequence indicates the order of assessment of the second graphs selected in each of the five-year intervals (this time starting with the graph from 1985, then from 1970, then from 2000, ...), and so on. Upon establishing an order for the graph evaluation, the assessment process was ready to begin using the developed graph quality criteria and determined scoring system discussed in Chapter 3.

	First Permutation	Second Permutation	Third Permutation	Fourth Permutation	Fifth Permutation
1	1940	1985	1955	1980	1935
2	2000	1970	1950	1930	2005
3	2005	2000	1970	1990	1985
4	1960	1990	1985	1950	1970
5	1980	1960	1935	1945	2015
6	1955	1945	1980	1940	1990
7	1945	1955	1960	1975	1930
8	1985	1940	1995	1935	1965
9	1990	1965	2000	1955	1960
10	2015	1930	1945	1970	1980
11	1935	1950	1940	1960	1975
12	1965	2010	1930	2015	2010
13	1995	1935	2005	2005	1945
14	1970	1975	1990	1985	1940
15	1975	2005	2015	2000	1950
16	1930	1995	1965	2010	1995
17	2010	1980	1975	1965	1955
18	1950	2015	2010	1995	2000

Fig. 2.7: Order of graph assessment for the 90 sampled graphs, where the selected graphs across each of the aggregated five-year intervals are judged in the order of their respective numbered permutation sequence.

CHAPTER 3

Statistical Graph Quality Criteria

The primary focus of this research is to understand how the quality of statistical graphs has changed over time. To be able to properly judge the quality of the sampled graphs from our population of interest, Utah State University (USU) Plan A Master of Science (MS) thesis graphs (see Section 2.2.2), which were sampled as described in Section 2.2.5, we needed to develop a criteria set to assess statistical graph quality.

As highlighted in Section 1.1, there exist several standards on how to create high quality graphs from various experienced data scientists. In Section 3.1, we examine numerous articles from the statistical literature that contain good graphic principles to gain an understanding of the different fundamental attributes that make up a high quality graph. In Section 3.2, we compile the reviewed literature resources and the standards highlighted within them into an overarching list of statistical graph quality standards and further group, condense, and summarize the standards to develop a set of graph quality criteria to judge the quality of the sampled USU Plan A MS thesis graphs. It is important to note that for the purposes of this research, we do not design the graph quality criteria to be specific to individual graph types. Rather, we want the graph quality criteria to be generalizable to all types of graphs to allow for consistency, flexibility, and simplicity when conducting the statistical graph quality assessments. In Section 3.3, we create a system to score and rate the quality of graphs against the developed graph quality criteria. In Section 3.4, we test the developed criteria and scoring system on graphs unrelated to our population of interest as well as on graphs within our population of interest, i.e., from the time period used for the testing process (see Section 2.2.3). In Section 3.5, we illustrate the assessment process using the developed graph quality criteria and scoring system on two graph examples from the time period used for the testing process.

3.1 Graph Assessment Standards from the Literature

The first resource that was examined for good graph standards was the article titled, "Statistician Heal Thyself: Have We Lost the Plot?" (Gordon and Finch, 2015), which was summarized in Section 1.1. As a reminder, Gordon and Finch (2015) conducted research to assess whether statisticians were creating high quality graphs that aligned with good graphic principles. By sampling graphs from top-rated applied science and statistics journals, they evaluated whether the sampled graphs aligned with good graphic principles using their own developed methodology and criteria. Gordon and Finch (2015) assembled five principles of graphical excellence based on insights from the statisticians Cleveland (1985) and Tufte (1983). The five principles from Gordon and Finch (2015) include: 1. '*show the data clearly*', 2. '*use simplicity in design*', 3. '*use good alignment on a common scale for quantities to be compared*', 4. '*keep the visual encoding transparent*', and 5. '*use graphical forms consistent with Principles 1 to 4*'.

Principle 1, '*show the data clearly*', is based on the idea that the most important part of a graph is the data. Gordon and Finch (2015) highlighted detection as a component of this principle, where the viewer should be able to easily detect important elements of the data. Principle 2, '*use simplicity in design*', revolves around the data-to-ink ratio concept, which is that every added feature, or every bit of ink on a graph, should have a reason behind it and present new information. Therefore, graphs should have a high data-to-ink ratio. However, it is important to understand that the data itself does not have to be uncomplicated to achieve a high data-to-ink ratio. It is possible to create simple designed graphs that represent complex data sets. Principle 3, '*use good alignment on a common scale for quantities to be compared*', implies that good quality graphs should present information on constant measurement scales where possible so that graphical elements can be appropriately judged and compared. As an example, stacked bar charts violate this principle since only one category is aligned against the common scale. This is not visually straightforward and often results in miscommunication of information. Principle 4, '*keep the visual encoding transparent*', follows from the third principle. The graph creator should make it as easy as possible for the viewer to decode the information displayed in a graph. Gordon and Finch (2015) stated, "If possible, the decoding necessary should be transparent: the viewer should be barely aware of doing it. If it is

hard work to understand and explain a graph, the visual encoding is not transparent." The graph creator must decide what is necessary to most appropriately translate the information to the viewer, which involves choices such as: graph type, color, scaling, etc. Principle 5, '*use graphical forms consistent with Principles 1 to 4*', is primarily centered around the creator choosing graph types that will best suit the data and aid the viewer in understanding key takeaways. Gordon and Finch (2015) suggested the following as standard and effective graph types: histograms, dotplots, boxplots, line plots, bar or dot charts, and scatterplots.

Using these five principles of graphical excellence as their guide, Gordon and Finch (2015) then developed over 60 different features to be coded for each graph they analyzed, where the features included both good and bad criteria to judge the graphs by. The five principles of graphical excellence, along with the 60 coded features, can be found in Appendix B.1. In September 2021, we reached out to Ian Gordon and Sue Finch to understand more about these features and how they utilized them when assessing the sampled graphs from the applied science and statistics journals. Based on their response, we learned that the five principles of graphical excellence they developed helped identify the 60 features to code for each graph. While the features went into detail, they were not always applicable to every graph, i.e., p-values were coded only if represented, and if represented, the nature of the representation was coded with options such as: stars, relative p-values, or exact p-values. The authors provided us with their list of 60 features but included a message of caution that the coding process was extremely detailed and time consuming. After learning about the USU Plan A MS thesis graph quality research we were conducting, they advised that we select critical features to focus on and only code a subset of what they considered.

Lastly, using qualitative judgement, Gordon and Finch (2015) assigned an overall quality rating to each of the sampled graphs, ranging from poor to exemplary. The authors individually assigned each graph an overall rating and the assessments were compared. If there were any discrepancies between the two assigned overall quality ratings, they were solved through discussion between the authors. It is important to note that for each graph, the count of undesirable or desirable coded features were not used to rate the overall quality since different features have different weights in terms of their impact on the overall quality. Instead, the clarity, transparency, and context of the

information plotted were used to make the overall judgement. Each assessment was examined in relation to the corresponding number of undesirable coded features as a way to partially validate the authors' overall judgement.

The second resource explored for good graph standards was the book titled, "Creating More Effective Graphs" (Robbins, 2005). This book serves as a guide of general principles to creating good quality graphs. It offers basic techniques that can be generalized within a broad range of applications. Additionally, a variety of real-world examples are provided for many of the principles. Robbins (2005) emphasized two main categories of principles: '*visual clarity*' and '*clear understanding*'. The principles within the '*visual clarity*' category focus on the need for the viewer to be able to clearly see the data being graphed, as well as identify other key features, i.e., tick marks, axes, labels, etc. The principles within the '*clear understanding*' category focus on the need for the viewer to be able to clearly discern and interpret the data being graphed, as well as understand the included labels and captions. Since there are many principles within each category, not all will be discussed in this thesis. However, the full list of principles can be found in Appendix B.2.

Within the '*visual clarity*' category, the principle '*Make the data stand out. Avoid superfluity*' suggests that within a graph, the most striking feature should be the data. As Robbins (2005) recommended, the data should be the first thing that is noticed when looking at a graph. Alongside this idea, the principle 'Use visually prominent graphical elements to show the data.' further clarifies that all added features should only emphasize the data, not distract or take away from the information presented. Several of the principles within this category discourage the addition of any graph clutter, which can take form in a variety of ways. Examples of such principles include: '*Do not overdo the number of tick marks*' and '*Do not overdo the number of tick mark labels*'. Tick marks and tick mark labels have the potential to overcrowd a graph if an unnecessary amount is created. Often, not all the data needs to be labeled. The amount of tick marks and tick mark labels used should be the amount necessary for the viewer to still easily understand and interpret the data without needing any major clarifications.

Within the '*clear understanding*' category, the principle '*Draw the data to scale*' is vital for a viewer to be able to appropriately interpret the data being displayed in a graph. Since graphs

are visual representations of numerical data, it is essential to illustrate the information in a way that reflects the truth so that the viewer can easily understand the reality of the information. In the case of three-dimensional (3D) bar graphs or histograms, bars in the front can appear much larger than bars in the back. The principle '*Do not show changes in one dimension by area or volume*' highlights how changes in one dimension should be represented by proportional changes in one dimension alone. For example, when displaying the numbers one and two in a graph where variables are represented by area, this rule applies. The variable with a value of one would need to have a length and width of one to represent the area as one. For the variable with a value of two, which is double the value of one, it would be incorrect to double the length and width as the area for this variable would be reflected as four. Rather, the length and width would need to each be $\sqrt{2}$ for the area to be correctly represented as two. Another important principle is '*Use a common baseline wherever possible*'. When different baselines are used, comparing data becomes distorted and therefore requires more work for the viewer. In general, many of the principles within the '*clear understanding*' category emphasize consistency among all aspects of a graph.

In addition to the two main categories of general principles to creating effective graphs, Robbins (2005) included other general strategies for creating effective graphs, as well as a checklist of possible graph defects to aid graph creators through the process of constructing graphs. The additional strategies and checklist can be found in Appendix B.2. Using the checklist, graph creators can ask themselves the series of questions that Robbins (2005) developed, which will guide them to make appropriate decisions throughout the graph creation process.

The third resource researched for good graph standards was the article titled, "How to Display Data Badly" (Wainer, 1984). This article provided 12 bad graph rules to avoid when creating graphs, along with illustrated examples of each rule. The bad graph rules are centered around ways that data can be displayed poorly, which leave the viewer confused and uninformed. Wainer (1984) defined three components of the overall aim of good graphics, which are: '*showing data*', '*showing data accurately*', and '*showing data clearly*'. Thus, based off of these three components, there are consequently three avenues that data can be displayed inappropriately. Wainer (1984) categorized the 12 different bad graph rules within these three components. Not all 12 rules will be discussed in

this thesis, however, the full list of rules can be found in Appendix B.3.

Within the first component, '*showing data*', Rule 1, '*Show as Few Data as Possible (Minimize the Data Density)*', is typically followed when graph creators believe that a graph looks empty, uninformative, or boring. Often, the determined solution is to add unnecessary graphical features, or as Tufte (1997) named, 'chartjunk'. However, as previously discussed, the quality of a graph suffers as the data-to-ink ratio gets closer to zero. Rule 2, '*Hide What Data You Do Show (Minimize the Data-Ink Ratio)*', follows this concept. There are multiple ways to effectively hide data, i.e., in the grid, in the scale, in the excess of chartjunk, and more. Hiding data can be easily avoided if the graph creator approaches the graph creation process with awareness and understanding of the main purpose behind graphs, which is to convey data.

Within the second component, '*showing data accurately*', Rule 3, '*Ignore the Visual Metaphor Altogether*', highlights two different cases. The first case is when smaller numbers are represented by higher bars or larger areas than those for bigger numbers. The second case is when graphical components such as colors, line styles, and plot symbols are used in an inconsistent way in related graphs or sub-graphs. Rule 4, '*Only Order Matters*', describes when data is represented based only on the order of the numerical values. For example, when displaying the numbers 15 and 20 in a bar graph with heights one and two, respectively, this rule applies. While 20 is bigger than 15, the heights are not proportional to the actual numbers. Thus, the order of the bar heights is still correct, but the magnitudes are not adequately represented.

Within the third component, '*showing data clearly*', Rule 9, '*Austria First!*', highlights how ordering a graph alphabetically can conceal important structures in the data. By sorting the data in a more meaningful way, the viewer can quickly identify patterns, trends, and insights that are most relevant to the presented data. Rule 12, '*If It Has Been Done Well in the Past, Think of Another Way to Do It*', is followed when graph creators try to reimagine and reinvent visualizations. The primary graph types that have worked for years are sometimes abandoned so that more exciting and glamorous numerical illustrations can be used instead. When a viewer is unfamiliar with a graph type, this can lead to misinterpretation of data or even a complete lack of understanding. In most cases, choosing a graph type that has been successfully used and recommended in the past is the

most appropriate choice.

The fourth and last resource explored for good graph standards was the article titled, "Ten Guidelines for Effective Data Visualization in Scientific Publications" by Kelleher and Wagener (2011). They offered ten guidelines to adequately communicate and represent data in graphs. The ten guidelines are each based on certain references (Brewer, 1994; Chambers et al., 1983; Cleveland, 1994; Cleveland and Devlin, 1980; Cleveland and McGill, 1984; Few, 2004, 2009; Harrower and Brewer, 2003; Kosslyn and Chabris, 1992; Robbins, 2005; Strange, 2007; Tufte, 1983, 2006). The guidelines address common mistakes that can be made when creating graphs and how to avoid them. These guidelines are supportive of what is deemed as the primary objective of data visualization, to effectively convey information. Furthermore, they are generalized to cover a wide variety of applications and disciplines. Not all ten guidelines will be discussed in this thesis, however, the full list of guidelines can be found in Appendix B.4.

Guideline 1, '*create the simplest graph that conveys the information you want to convey*', indicates that the most important part of a graph is the data. Any additional features that do not add purpose can unnecessarily overcomplicate a graph. Guideline 4, '*select meaningful axis ranges*', emphasizes the need to choose an appropriate scale to properly convey data. A common mistake made by graph creators is to exclude the value zero when displaying absolute magnitudes along the vertical axis of a graph. In these cases, starting the vertical axis range anywhere other than zero can misrepresent the data and hide magnitude differences between quantities. Guideline 6, '*plot overlapping points in a way that density differences become apparent in scatter plots*', is vital to effectively communicate information to the viewer when data points overlap. If overlapping data points are plotted without any transparency or tactic to make them distinguishable, value differences can be difficult to decipher. In some cases, the viewer may not even be able to recognize that there are overlapping data points and assume they are seeing one data point. By using a technique to make data points distinguishable, it becomes easier to visualize density differences, thus strengthening the truthfulness of the information communicated to the viewer. Guideline 10, '*select an appropriate color scheme based on the type of data*', emphasizes that color can play an important role when displaying data. It has the power to strengthen the data portrayed in a graph if executed correctly, or

can severely misdirect the viewer if not. While color is not always needed, it can be an effective tool to support the objective of a graph. There are three main color scheme types, namely: sequential, diverging, and qualitative. Sequential schemes should primarily be used for quantitative data since the colors in this scheme graduate from light to dark, which can emphasize differences between low and high values, respectively. Diverging schemes should be used to show the contrast between low and high values in relation to an average value since the colors in this scheme typically use contrasting dark colors to distinguish low and high values and use a neutral color to represent an average value. Qualitative schemes should be used when representing categorical data since the colors in this scheme consist of contrasting colors, which can highlight differences between categories without alluding to magnitude.

In addition to some of the sources that formed the basis of the ten guidelines from Kelleher and Wagener (2011), it is important to note that additional sets of rules exist, such as graph color rules as suggested by Kosslyn (2006) and Zeileis et al. (2020), but have not been directly used as a basis for the graph quality criteria developed in this thesis.

3.2 Categorized and Condensed Graph Quality Criteria

After reviewing the four graph quality literature resources discussed in Section 3.1, we compiled the standards from each of the resources into an Excel sheet so that we could group, combine, and condense the various material in order to develop a criteria set to judge the quality of the sampled USU Plan A MS thesis graphs. This Excel sheet can be found in Figure 3.1. Overall, there are 108 standards compiled among the four resources. A color was assigned to the standards within each resource, as can be seen by the key provided in the bottom left of Figure 3.1. It is important to note that of the 60 coded features provided by Gordon and Finch (2015) (see Appendix B.1), only the first 25 features were included in Figure 3.1. The 35 features not included were partially categorized under Principle 5 of the five principles of graphical excellence created by Gordon and Finch (2015), '*use graphical forms consistent with Principles 1 to 4*', as well as within a list of other additional features. These 35 features mostly contained extremely specific and descriptive graph characteristics, some of which can be applied to certain graph types only, rather than generalized

graph principles. Additionally, one of these coded features was the overall quality rating Gordon and Finch (2015) developed.

Next, we looked for any similarities, repetition, or redundancy among the compiled standards. Based on this investigation, we were able to determine four main categories to group the compiled standards by, namely: *Labeling*, *Clear Understanding*, *Meaningful*, and *Scaling and Gridlines*. While the standards could have been separated into even more distinct groups, we felt that these four categories targeted a majority of the standards in a way that would be easy to condense and develop our own set of criteria from. The grouped standards can be seen in Figure 3.2. There are 19 standards in the *Labeling* category, 38 standards in the *Clear Understanding* category, 16 standards in the *Meaningful* category, and 33 standards in the *Scaling and Gridlines* category. The same colors assigned to the standards within each graph quality literature resource in Figure 3.1 are used to distinguish the standards in Figure 3.2, as can be seen by the key provided in the bottom left of Figure 3.2.

Note that two of the standards covered multiple ideas relevant to more than one category, thus were included in each applicable category. The principle '*Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward*' from Robbins (2005) was included in both the *Labeling* and *Scaling and Gridlines* categories, as can be seen in Figures 3.2a and 3.2d, respectively. While this principle largely refers to scaling, we felt that it also addresses labeling since the direction of tick marks can be applied to label clutter within a graph. Additionally, Rule 10, '*Label (a) Illegibly, (b) Incompletely, (c) Incorrectly, and (d) Ambiguously*', from Wainer (1984) was also included in both the *Labeling* and *Scaling and Gridlines* categories, as can be seen in Figures 3.2a and 3.2d, respectively. While this principle largely refers to labeling, we felt that the reference to incomplete labeling could be applicable to scales and axes, and their requirement for consistency and completeness. Additionally, four of the standards were too broad to fit into any of the four categories, therefore they were left uncategorized. The principles '*Proofread graphs*', '*Graph data two or more times when needed*', '*Graphing data should be an iterative experimental process*', and '*Many useful displays require careful, detailed study*' from Robbins (2005) are the four standards that were left out.

Guidelines Compilation

Gordon & Finch	
Feature number	
	Principle 1 - show the data clearly
1	Includes a caption
2	Caption is adequate
3	Suitable axis labels
4	Variable labelled rather than estimate (on axis)
5	Legend
6	Legend could be replaced by direct text
7	Boxed legend
8	Graph has detection problems
9	Contains undefined graphical elements
10	Contains undefined abbreviations
11	Unused white space that could be used for data
12	Incorrect scaling
	Principle 2 - use simplicity in design
13	Number of dimensions (2d or 3d)
14	Two scales on one axis
	Principle 4 - keep the visual encoding transparent
15	Uses colour
16	Colour use redundant
17	Uses cross hatching including stripes
18	Ordering would improve transparency
19	Does the graph stand alone?
	Principle 3 - use good alignment on a common scale for quantities to be compared
20	Elements to be compared aligned on common scale
21	Non horizontal tick mark labels
22	Gridlines, including reference line(s)
23	Gridlines too heavy
24	Additional gridlines could be used
25	Transposition would improve the graph

Wainer	
Rule	
	Showing data
1	Show as few data as possible (minimize the data density)
2	Hide what data you do show (minimize the data-ink ratio)
	Showing data accurately
3	Ignore the visual metaphor altogether
4	Only order matters
5	Graph data out of context
	Showing data clearly
6	Change scales in mid-axis
7	Emphasize the trivial (ignore the important)
8	Jiggle the baseline
9	Austria first
10	Label (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously
11	More is murkier: (a) more decimal places and (b) more dimensions
12	If it has been done well in the past, think of another way to do it

Kelleher and Wagener	
Guideline	
1	Create the simplest graph that conveys the information you want to convey
2	Consider the type of encoding object and attribute used to create a plot
3	Focus on visualizing patterns or on visualizing details, depending on the purpose of the plot
4	Select meaningful axis ranges
5	Data transformations and carefully chosen graph aspect ratios can be used to emphasize rates of change for time-series data
6	Plot overlapping points in a way that density differences become apparent in scatter plots
7	Use lines when connecting sequential data in time-series plots
8	Aggregate larger datasets in meaningful ways
9	Keep axis ranges as similar as possible to compare variables
10	Select an appropriate color scheme based on the type of data

Key:	
	Gordon & Finch (2015)
	Robbins (2005)
	Wainer (1984)
	Kelleher and Wagener (2011)

Robbins	
Chapter 6	
First set: reader clearly sees what is graphed	
6.2	Visual clarity
6.2.1	Clarity of data
	Make the data stand out. Avoid superfluity
	Use visually prominent graphical elements to show the data
	Overlapping plotting symbols must be visually distinguishable
	Superposed data sets must be readily visually assembled
	Do not clutter the interior of the scale-line rectangle
6.2.2	Clarity of other elements
	Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward
	Do not overdo the number of tick marks
	Do not overdo the number of tick mark labels
	Deemphasize grid lines and distinguish grid lines from data
	Avoid putting notes and keys inside the scale-line rectangle. Put a key outside and put notes in the caption or in the text
	Visual clarity must be preserved under reduction and reproduction
	Proofread graphs
	The chart or graph must be consistent with the text
Second set: reader clearly understands what is graphed	
6.3	Clear understanding
	Draw the data to scale
	Do not show changes in one dimension by area or volume
	Use a common baseline wherever possible
	Label data sets directly when it doesn't clutter the graph
	Don't require the reader to make calculations that a computer can make more easily
	Plot the variable of interest. If interested in improvement, plot improvement rather than *before* and *after*
	Strive for clarity
	Groups of charts need consistency in order, color, and other graphical elements
	Choose the principle least likely to mislead if more than one applies and they conflict with one another
6.4	General strategy
	A large amount of quantitative information can be packed into a small region
	Graphing data should be an iterative experimental process
	Graph data two or more times when needed
	Many useful displays require careful, detailed study
Appendix A	
Checklist of possible graph defects	
Can the reader clearly see the graphical elements?	
	Do the data stand out? Are there superfluous elements?
	Are all graphical elements visually prominent?
	Are overlapping plotting symbols visually distinguishable?
	Can superposed data sets be readily visually assembled?
	Is the interior of the scale-line rectangle cluttered?
	Do data labels interfere with the quantitative data or clutter the graph?
	Is the data rectangle within the scale-line rectangle?
	Do tick marks interfere with the data?
	Do tick mark labels interfere with the data?
	Are axis labels legible?
	Are there too many tick marks?
	Are there too many tick mark labels?
	Do the grid lines interfere with the data?
	Are there notes or keys inside the scale-line rectangle?
	Will visual clarity be preserved under reduction and reproduction?
Can the reader clearly understand the graph?	
	Are the data drawn to scale?
	Is there an informative title?
	Is area or volume used to show changes in one dimension?
	Are there too many dimensions in the graph (more than one in the data)?
	Are common baselines used wherever possible?
	Are all labels associated with the correct graphical elements?
	Is the reader required to make calculations?
	Are groups of charts drawn consistently?
Are the scales well chosen and labeled?	
	Is zero included for all bar graphs?
	Are there any unnecessary scale breaks?
	Is there a forceful indication of a scale break?
	Are there numerical values on two sides of a scale break that are connected?
	Does the aspect ratio allow the reader to see variations in the data?
	Are scales included for all axes?
	Are the scales labeled?
	Are tick marks at sensible values?
	Do the axes increase in the conventional direction?
	Does the data rectangle fill as much of the scale-line rectangle as possible?
	Are uneven time intervals handled correctly?
	Are the scales appropriate when different panels are compared?

Fig. 3.1: Compiled list of standards extracted from graph quality literature resources.

Graph has adequate labeling	
1	Includes a caption
2	Caption is adequate
3	Suitable axis labels
4	Variable labelled rather than estimate (on axis)
5	Legend
6	Legend could be replaced by direct text
7	Boxed legend
	Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward
	Avoid putting notes and keys inside the scale-line rectangle. Put a key outside and put notes in the caption or in the text
	Label data sets directly when it doesn't clutter the graph
	Do data labels interfere with the quantitative data or clutter the graph?
	Is the data rectangle within the scale-line rectangle?
	Do tick marks interfere with the data?
	Do tick mark labels interfere with the data?
	Are axis labels legible?
	Are there notes or keys inside the scale-line rectangle?
	Is there an informative title?
	Are all labels associated with the correct graphical elements?
10	Label (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously

(a) *Labeling* standards.

Graph has clear understanding	
8	Graph has detection problems
9	Contains undefined graphical elements
10	Contains undefined abbreviations
18	Ordering would improve transparency
19	Does the graph stand alone?
	Make the data stand out. Avoid superfluity
	Plot the variable of interest. If interested in improvement, plot improvement rather than *before* and *after*
	Groups of charts need consistency in order, color, and other graphical elements
	Use visually prominent graphical elements to show the data
	Overlapping plotting symbols must be visually distinguishable
	Superposed data sets must be readily visually assembled
	The chart or graph must be consistent with the text
	Visual clarity must be preserved under reduction and reproduction
	Do not show changes in one dimension by area or volume
	Don't require the reader to make calculations that a computer can make more easily
	Strive for clarity
	Choose the principle least likely to mislead if more than one applies and they conflict with one another
	Do the data stand out? Are there superfluous elements?
	Are all graphical elements visually prominent?
	Are groups of charts drawn consistently?
	Are overlapping plotting symbols visually distinguishable?
	Can superposed data sets be readily visually assembled?
	Will visual clarity be preserved under reduction and reproduction?
	Is area or volume used to show changes in one dimension?
	Is the reader required to make calculations?
	Use a common baseline wherever possible
	Are common baselines used wherever possible?
3	Ignore the visual metaphor altogether
4	Only order matters
5	Graph data out of context
7	Emphasize the trivial (ignore the important)
8	Jiggle the baseline
9	Austria first
2	Consider the type of encoding object and attribute used to create a plot
3	Focus on visualizing patterns or on visualizing details, depending on the purpose of the plot
6	Plot overlapping points in a way that density differences become apparent in scatter plots
7	Use lines when connecting sequential data in time-series plots
8	Aggregate larger datasets in meaningful ways

(b) *Clear Understanding* standards.

Graphical elements add meaning to information displayed	
11	Unused white space that could be used for data
13	Number of dimensions (2d or 3d)
15	Uses colour
16	Colour use redundant
17	Uses cross hatching including stripes
	Do not clutter the interior of the scale-line rectangle
	A large amount of quantitative information can be packed into a small region
	Is the interior of the scale-line rectangle cluttered?
	Are there too many dimensions in the graph (more than one in the data)?
	Does the data rectangle fill as much of the scale-line rectangle as possible?
1	Show as few data as possible (minimize the data density)
2	Hide what data you do show (minimize the data-ink ratio)
11	More is murkier: (a) more decimal places and (b) more dimensions
12	If it has been done well in the past, think of another way to do it
1	Create the simplest graph that conveys the information you want to convey
10	Select an appropriate color scheme based on the type of data

(c) *Meaningful* standards.

Graph has adequate scaling and gridlines	
12	Incorrect scaling
14	Two scales on one axis
20	Elements to be compared aligned on common scale
21	Non horizontal tick mark labels
22	Gridlines, including reference line(s)
23	Gridlines too heavy
24	Additional gridlines could be used
25	Transposition would improve the graph
	Deemphasize grid lines and distinguish grid lines from data
	Draw the data to scale
	Do not overdo the number of tick marks
	Do not overdo the number of tick mark labels
	Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward
	Do the grid lines interfere with the data?
	Are the data drawn to scale?
	Is zero included for all bar graphs?
	Are there any unnecessary scale breaks?
	Is there a forceful indication of a scale break?
	Are there numerical values on two sides of a scale break that are connected?
	Does the aspect ratio allow the reader to see variations in the data?
	Are scales included for all axes?
	Are the scales labeled?
	Are tick marks at sensible values?
	Do the axes increase in the conventional direction?
	Are uneven time intervals handled correctly?
	Are the scales appropriate when different panels are compared?
	Are there too many tick marks?
	Are there too many tick mark labels?
6	Change scales in mid-axis
10	Label (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously
4	Select meaningful axis ranges
5	Data transformations and carefully chosen graph aspect ratios can be used to emphasize rates of change for time-series data
9	Keep axis ranges as similar as possible to compare variables

(d) *Scaling and Gridlines* standards.

Key:	
	Gordon & Finch (2015)
	Robbins (2005)
	Wainer (1984)
	Kelleher and Wagener (2011)

Fig. 3.2: Compiled standards from the graph quality literature resources, categorized according to the determined four main categories.

The *Labeling* category, shown in Figure 3.2a contains all of the standards that relate to any kind of graph labeling, i.e., axes, variables, legends, captions, and titles. This category also includes standards about incorrect, illegible, or interfering labels that create unnecessary clutter in a graph.

The *Clear Understanding* category, shown in Figure 3.2b, contains all of the standards that relate to the interpretability of a graph. This category includes standards about data clarity, ordering, graphical elements, common baselines, overlapping data points, consistency, readability, and more.

The *Meaningful* category, shown in Figure 3.2c, contains all of the standards that relate to meaningful and purposeful graphical elements. This category includes standards about color selection, number of dimensions, clutter, data-to-ink ratio, choice of graph type, and more.

The *Scaling and Gridlines* category, shown in Figure 3.2d, contains all of the standards that relate to axes scaling and gridline use. This category includes standards about the number of axis scales, use of a common scale when comparing elements, gridlines, tick mark values, axis ranges, and more.

Once all the standards were grouped into the four distinct categories, the next task was to condense the standards in order to remove any overlapping ideas, combine similar topics, and summarize the main points. From condensing and summarizing the standards, the statistical graph quality criteria were developed using the four main categories previously highlighted: *Labeling*, *Clear Understanding*, *Meaningful*, and *Scaling and Gridlines*. This set of criteria can be found in Figure 3.3. In total, there are 24 distinct criteria among the four different criteria categories. As a reminder, we initially gathered 108 individual standards from the four graph quality literature resources. By drastically reducing and consolidating the standards, we have ensured that the resulting statistical graph quality criteria are highly refined, relevant, and efficient.

The criteria from the *Labeling* catgeory can be found in Figure 3.3a. Within this category, there are five different criteria. These criteria require the caption and legend of a graph to adequately explain the information being represented, the graph labels to be legible, sufficient, and unobtrusive, and for all the graphical elements to be clearly described.

The criteria from the *Clear Understanding* category can be found in Figure 3.3b. Within this category, there are eight different criteria. These criteria require the data in a graph to be visually

Graph has adequate labeling
1. Adequate use of caption and legend
2. Graph labels are legible
3. Graph labels are sufficient
4. All graphical elements are clearly and correctly defined
5. Labeling does not interfere with or clutter the graph

(a) *Labeling* criteria (L1-L5).

Graph has clear understanding
1. Data is visually clear
2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph
3. Data is accurately displayed in the most effective and appropriate way
4. Data is consistent with labeling/text
5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)
6. Clarity of data will be maintained through reproduction/reduction of graph
7. Overlapping data points or superposed data sets are distinguishable
8. A common baseline is used wherever possible

(b) *Clear Understanding* criteria (U1-U8).

Graphical elements add meaning to information displayed
1. Show as much data as possible (maximize data/ink ratio)
2. Simplest graph of the chosen type is used to convey the necessary information
3. Minimal dimensions used
4. Color is used appropriately and effectively

(c) *Meaningful* criteria (M1-M4).

Graph has adequate scaling and gridlines
1. One scale is included for every axis
2. No changes or breaks in scale
3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Adequate use of gridlines, do not hide or interfere with data
6. Data transformations are adequately used to display necessary data information
7. Size and aspect ratio display data adequately

(d) *Scaling and Gridlines* criteria (S1-S7).

Fig. 3.3: Statistical graph quality criteria according to the four main criteria categories.

clear, accurately displayed in an effective way, consistent with labeling, and to sustain clarity. Furthermore, the criteria require the graph to have clear purpose and interpretability, overlapping points to be distinguishable, and a common baseline to be used if possible. In the case of multiple graphs or multiple components in a single graph, the criteria within this category expect consistency among ordering, coloring, and all other similar characteristics.

The criteria from the *Meaningful* category can be found in Figure 3.3c. Within this category, there are four different criteria. These criteria require a graph to maximize the data-to-ink ratio by showing as much data as possible, be in its simplest form possible, contain minimal dimensionality, and for color to be used appropriately.

The criteria from the *Scaling and Gridlines* category can be found in Figure 3.3d. Within this category, there are seven different criteria. These criteria require a graph to only include one scale per axis, have no changes or breaks in the scales, have meaningful axis ranges and tick mark values, use gridlines appropriately, and use an effective aspect ratio and graph size to sufficiently display data. If data transformations are used, they must represent the data adequately. In the case that variables need to be compared within a single graph or across multiple graphs, the criteria within this category expect the axis ranges to be as similar as possible.

3.3 The New Graph Scoring System

Each graph sampled from the determined population of interest, USU Plan A MS thesis reports (see Section 2.2.2), underwent a scoring session against the developed graph quality criteria discussed in Section 3.2. This allowed for a standardized way to assess the quality of statistical graphs.

Within the scoring system, it was determined that the individual criterion within each category would receive a score based on whether the criterion was met or not. A score of 1 was given if the criterion was met, a score of 0 was given if the criterion was not met, and in the cases that the pre-condition for the criterion was not met, the criterion was not given a score.

After collecting the criteria scores for each sampled graph, an overall accuracy metric was calculated. First, the assigned criteria scores within each category were summed and divided by the

total number of criteria within the corresponding category to obtain criteria category scores. It is important to note that this calculation excluded any criteria that did not receive a score so as to not punish for a pre-condition that was not met. Next, the criteria category scores had to be equally weighted. Since there are four criteria categories, each criteria category score was multiplied by 0.25. Lastly, the weighted criteria category scores were summed and multiplied by 100 to obtain a percent correct score, or overall accuracy metric, for each graph.

To make the scoring process as effortless as possible, a scoring template in Excel was created to organize the individual criterion scores, criteria category scores, and the percent correct score for each sampled graph. This template also includes columns to insert all the descriptive and identifying information of each sampled graph, along with the assigned overall rating, which will be discussed later in this section. Figure 3.4 displays the template used.

From Figure 3.4, we can see the following columns are included in the template: *Year Interval*, *Graph Name*, *Type of Graph*, *How Created*, *Labeling 1 - Labeling 5*, *Clear Understanding 1 - Clear Understanding 8*, *Meaningful 1 - Meaningful 4*, *Scaling and Gridlines 1 - Scaling and Gridlines 7*, *Total 1 - Total 4*, *Percent Correct*, and *Overall Rating*.

The *Year Interval* column is used to indicate the aggregated five-year interval that each sampled graph was extracted from. As a reminder from Section 2.2.2, the sampled graphs from the USU Plan A MS thesis reports were extracted from the years 1930 to 2019 and aggregated into 18 five-year intervals from the years 1930 to 2015. The *Graph Name* column is used to identify the individual sampled graphs. The names given to each graph include the four-digit five-year interval the graph was sampled from, the three to four-digit report number of the thesis that contains the graph, and the thesis page number that the graph is located on, i.e., 1940-3925-20, where 1940 is the five-year interval that the graph was sampled from, 3925 is the report number of the thesis that contains the graph, and 20 is the page number the sampled graph can be found on.

The *Type of Graph* column classifies each sampled graph as a certain graph type, such as: line graph, bar graph, scatterplot, dot plot, and histogram. The 'other' graph type contains all other graph type options. Within this column, certain stand-out characteristics about the graphs are noted alongside the determined graph type to identify unique traits within these generic categories.

Year Interval	Graph Name	Type of Graph	How Created	Labeling					Clear Understanding								Meaningful				Scaling and Gridlines							Total 1	Total 2	Total 3	Total 4	Percent Correct	Overall Rating (1 = Worst, 5 = Best)
				1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	5	6	7						

Criteria:			
Graph has adequate labeling	**Graph has clear understanding**	**Graphical elements add meaning to information displayed**	**Graph has adequate scaling and gridlines**
1. Adequate use of caption and legend	1. Data is visually clear	1. Show as much data as possible (maximize data/ink ratio)	1. One scale is included for every axis
2. Graph labels are legible	2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph	2. Simplest graph of the chosen type is used to convey the necessary information	2. No changes or breaks in scale
3. Graph labels are sufficient	3. Data is accurately displayed in the most effective and appropriate way	3. Minimal dimensions used	3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. All graphical elements are clearly and correctly defined	4. Data is consistent with labeling/text	4. Color is used appropriately and effectively	4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Labeling does not interfere with or clutter the graph	5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)		5. Adequate use of gridlines, do not hide or interfere with data
	6. Clarity of data will be maintained through reproduction/reduction of graph		6. Data transformations are adequately used to display necessary data information
	7. Overlapping data points or superposed data sets are distinguishable		7. Size and aspect ratio display data adequately
	8. A common baseline is used wherever possible		

Fig. 3.4: Scoring template used for recording and organizing the quality data of the sampled graphs.

The ‘multiple graphs’ note describes a graph that contains more than one graph within the graph figure, whether side-by-side or on top of one other for comparison, or two separate distinct graphs representing completely different data sets. The ‘+ intervals’ note highlights when a graph has some form of a confidence interval included for a few or all of the data points. The ‘+ line’ note signifies if a line is incorporated with the data to highlight a trend or pattern, or to simply connect data points, whereas the ‘+ fitted line’ note indicates if a line is, by some mathematical method, fit to the data to demonstrate a relationship between data points. The ‘3D’ note describes a graph that contains 3D elements. Any sampled graph that is classified as a bar graph includes either the note ‘single’ or ‘grouped’ to denote how the bars are structured, and further uses notes like ‘side-by-side’, ‘stacked’, or ‘overlayed’ to describe the layout of the bars.

The *How Created* column categorizes how the graphs were produced, whether by hand, using a typewriter, or generated with some form of software. A combination of these processes could also have been used to construct the graphs, i.e., a generic layout could have been designed using a typewriter and the actual data along with other graph details could have been added in by hand afterwards. The ‘other’ graph creation method contains all other graph creation method options.

The *Labeling 1 - Labeling 5*, *Clear Understanding 1 - Clear Understanding 8*, *Meaningful 1 - Meaningful 4*, and *Scaling and Gridlines 1 - Scaling and Gridlines 7* columns refer to the set of graph quality criteria and corresponding categorizations that were developed in Section 3.2 and can be seen in Figure 3.3. From this section, 24 refined and distinct criteria were grouped into four categories, namely: *Labeling*, *Clear Understanding*, *Meaningful*, and *Scaling and Gridlines*. In the Excel template shown in Figure 3.4, each criterion within the four different categories has its own respective column, i.e., there are five distinct *Labeling* columns included in the template for the five different criteria within the *Labeling* category. This pattern follows for the other three criteria categories as well. The data entered in all of these columns follows the scoring system previously discussed, where an entry of 1 indicates that the criterion was met, an entry of 0 indicates that the criterion was not met, and the column left blank suggests that the pre-condition for the criterion was not met.

The *Total 1*, *Total 2*, *Total 3*, and *Total 4* columns contain the weighted criteria category scores

of each sampled graph, which were previously explained in this section, where *Total 1* contains the weighted *Labeling* criteria category score, *Total 2* contains the weighted *Clear Understanding* criteria category score, *Total 3* contains the weighted *Meaningful* criteria category score, and *Total 4* contains the weighted *Scaling and Gridlines* criteria category score. The *Percent Correct* column contains the calculated percent correct score of each sampled graph, which was also previously explained in this section. This percentage signifies the overall accuracy of a sampled graph according to the developed graph quality criteria. It appropriately accounts for each criterion within the four criteria categories, providing a statistic that can be easily understood on its own, as well as be used to be compared with the overall accuracies of other sampled graphs.

The *Overall Rating* column contains the assigned overall rating for each sampled graph. These overall ratings follow a similar concept to the overall quality ratings that Gordon and Finch (2015) incorporated into their research study, as discussed in Section 3.1. For our research, the overall ratings were determined using human judgement, before the actual assessment process using the developed graph quality criteria took place. This judgement was based on what visible errors and mistakes could be seen, and how badly we believed they affected the quality of a graph at an overview level. After subjectively weighing and considering the different graph elements performed correctly and incorrectly, a rating was given. These ratings are within the range one to five, where one is the worst possible score and five is the best. While the overall rating is a general first impression of the total graph quality, the graph quality criteria are very specific to individual graph features. Thus, while these processes were intended to be supportive of each other, they were conducted independently. A perfect score of five may be awarded to a graph even if some of the criteria were determined not to be met later on. Dr. Jürgen Symanzik was involved in the overall rating process so that more than one opinion and perspective could be utilized. We each individually rated the quality of the sampled graphs and then discussed and reasoned through any discrepancies before finalizing the overall ratings.

As discussed in Section 2.2.5, five distinct five-year interval permutation sequences were created to determine the assessment order of the 90 sampled graphs across the 18 aggregated five-year intervals from 1930 to 2015. As explained in Section 5.1, an Excel file was created for each of

the permutation sequences using the scoring template shown in Figure 3.4 to store and organize the raw graph quality scores retrieved from the respective graphs. The Excel file containing the scoring results for the first five-year interval permutation sequence can be found in Figure 5.1 while the Excel files containing the scoring results for the second, third, fourth, and fifth five-year interval permutation sequences, respectively, can be found in Appendix C.

3.4 Testing and Refining the Graph Quality Criteria and New Graph Scoring System

To ensure that the 24 developed graph quality criteria, introduced in Section 3.2, capture the different components of a good quality graph, we tested the criteria against a variety of graphs using the determined scoring system, which was highlighted in Section 3.3. We wanted to fine-tune the criteria if needed so that they were as clear, relevant, and concise as possible before assessing the quality of the sampled graphs from the USU Plan A MS thesis reports. Dr. Jürgen Symanzik was involved in the testing process so that more than one opinion and perspective could be utilized.

We first tested the developed graph quality criteria and scoring system on graphs unrelated to our population of interest. This testing was done against sampled graphs from the winning posters of the 2016 American Statistical Association Poster Competition and Project Competition (Symanzik et al., 2016). Using the scoring template shown in Figure 3.4, we individually assessed a total of ten graphs according to the developed graph quality criteria and established scoring system. We held weekly meetings to compare and discuss any discrepancies between our assigned criteria scores. This allowed for the fine-tuning of wording, functionality, and overall acceptance of the initially developed criteria, finally resulting in the criteria shown in Figure 3.3.

We then tested the criteria on graphs in USU Plan A MS thesis reports from the years reserved for the testing process, which was from the years 1923 to 1929 and 2020 to 2021, as highlighted in Section 2.2.3. Testing the criteria on graphs from actual thesis reports provided a way to appropriately train for the real sampling and scoring process and helped determine a reasonable sample size for each aggregated five-year interval within the sampling region, as mentioned in Section 2.2.2. There were a total of eight graphs sampled from the time period used for the testing process. Two examples of graphs assessed within this time period can be found in Figures 2.4 and 2.5. Similar

to the process executed with the unrelated graphs, we used the scoring template, shown in Figure 3.4, to assess the sampled graphs from the time period used for the testing process and held weekly meetings to discuss any discrepancies between our assigned criteria scores.

3.5 Scoring Process Examples

To illustrate how the quality of graphs are assessed using the developed graph quality criteria and scoring system, we will discuss the scoring process according to the graphs shown in Figures 2.4 and 2.5. For these two graphs, an Excel file was created using the scoring template shown in Figure 3.4, and has been filled in with the corresponding scoring information for the respective graphs, which will be discussed in this section. The completed Excel file can be seen in Figure 3.5.

First, we will discuss the scoring process for the graph shown in Figure 2.4. The graph was given the name '1925-3939-7' using the graph naming procedure described in Section 3.3. The first step in the scoring process for the graph shown in Figure 2.4 was to determine the graph type and graph creation method. Based on the plotted point values along two numerical axes, we determined the graph to be a scatterplot with the '+ line' added note since the data points are connected by lines. We also decided that the graph was created by hand. Next, we had to assign the graph an overall rating. Based on our initial inspection of the graph, we gave the graph shown in Figure 2.4 an overall rating of 4. The graph includes most of the necessary elements, such as: reasonable axis ranges, consistent scaling, clear and sufficient labels, good use of caption, etc. However, two of the variables are difficult to distinguish. Additionally, some of the data points are not very prominent. Overall, the graph contains most of the fundamental components of a good quality graph, however, minor improvements could be made for easier interpretation and understanding of the data.

The next step in the scoring process for the graph shown in Figure 2.4 was to assign each of the 24 graph quality criteria (see Section 3.2) a score using the scoring system outlined in Section 3.3. Under the *Labeling* criteria category, all five of the criteria are met since the graph shown in Figure 2.4 demonstrates adequate use of a caption and legend, has legible and sufficient labels, includes definitions for all added graphical elements, and does not contain any clutter within the graph. Under the *Clear Understanding* criteria category, the first, fourth, sixth, seventh, and eighth criteria

				Labeling					Clear Understanding								Meaningful				Scaling and Gridlines											
Year Interval	Graph Name	Type of Graph	How Created	1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	5	6	7	Total 1	Total 2	Total 3	Total 4	Percent Correct	Overall Rating (1 = Worst, 5 = Best)
1925	1925-3939-7	Scatterplot (+ line)	Hand-drawn	1	1	1	1	1	1	0	0	1		1	1	1	1	1	1	0	1	1	1	1	1		1	0.25	0.178571	0.1875	0.25	86.61	4
2020	2020-8052-38	Bar Graph (grouped, side-by-side)(+ intervals)	Software	0	1	0	0	0	1	0	0	0	1	1		1	0	1	1	1	1	1	1	1		1	1	0.05	0.142857	0.1875	0.25	63.04	3

Criteria:			
Graph has adequate labeling	**Graph has clear understanding**	**Graphical elements add meaning to information displayed**	**Graph has adequate scaling and gridlines**
1. Adequate use of caption and legend	1. Data is visually clear	1. Show as much data as possible (maximize data/ink ratio)	1. One scale is included for every axis
2. Graph labels are legible	2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph	2. Simplest graph of the chosen type is used to convey the necessary information	2. No changes or breaks in scale
3. Graph labels are sufficient	3. Data is accurately displayed in the most effective and appropriate way	3. Minimal dimensions used	3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. All graphical elements are clearly and correctly defined	4. Data is consistent with labeling/text	4. Color is used appropriately and effectively	4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Labeling does not interfere with or clutter the graph	5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)		5. Adequate use of gridlines, do not hide or interfere with data
	6. Clarity of data will be maintained through reproduction/reduction of graph		6. Data transformations are adequately used to display necessary data information
	7. Overlapping data points or superposed data sets are distinguishable		7. Size and aspect ratio display data adequately
	8. A common baseline is used wherever possible		

Fig. 3.5: Excel file containing the raw graph quality scores for the graphs shown in Figures 2.4 and 2.5, respectively.

are met since the data is fairly clear, is consistent with the labeling and caption, will maintain clarity through graph reproduction or reduction, contains overlapping yet distinguishable points, and uses a common baseline. The second criterion is not met since it is difficult to determine key details within the graph as some of the data points and line types are not very distinct. Following this issue, the third criterion is not met since the data is not displayed in the most effective way. The fifth criterion does not apply since there does not exist multiple components within the graph that necessitate consistency. Under the *Meaningful* criteria category, the first, second, and third criteria are met since the data-to-ink ratio is maximized, the scatterplot is simple and concise, and only two dimensions are included. The fourth criterion is not met since color was not used effectively as two of the variables, F_{22} and F_3, use the same line type and were assigned almost indistinguishable colors. Under the *Scaling and Gridlines* criteria category, all criteria, except for the sixth criterion, are met since one scale is included for each of the axes, there are no changes or breaks in the scales, the axis ranges and tick mark values are reasonable for the represented data and allow for the variables to be effectively compared, the gridlines do not interfere with the data, and the size of the figure displays the data adequately. The sixth criterion does not apply since there were no transformations performed on the data.

The last step in the scoring process for the graph shown in Figure 2.4 was to obtain the percent correct score, or overall accuracy of the graph. Based on the procedure outlined in Section 3.3, the percent correct score, or overall accuracy of the graph, therefore is calculated as 86.61%. This aligns fairly well with the assigned overall rating of 4.

Next, we will discuss the scoring process for the graph shown in Figure 2.5. The graph was given the name '2020-8052-38' using the graph naming procedure described in Section 3.3. As explained with the graph shown in Figure 2.4, the first step in the scoring process for the graph shown in Figure 2.5 was to determine the graph type and graph creation method. Based on the representation of the data, we determined the graph to be a bar graph with the 'grouped' and 'side-by-side' added notes to describe the position and layout of the bars, and the '+ intervals' added note since some form of a confidence interval was included with each of the bars. We also determined that the graph was created using software. For the overall rating, we gave the graph shown in Figure

2.5 a score of 3. While the graph contains some of the fundamental aspects of a good quality graph, such as consistency among groups, appropriate color choices, adequate scaling, and a reasonable axis range, the graph does not label all the included features. In this case, the confidence intervals and 'a' and 'b' labels included with the bars are not explained or defined anywhere on the graph figure. This can be detrimental for a viewer since there is no way to interpret the meaning behind these features, thus leading to confusion and incorrect analysis of the data.

The next step in the scoring process for the graph shown in Figure 2.5 was to assign the 24 graph quality criteria (see Section 3.2) a score using the scoring system outlined in Section 3.3. Under the *Labeling* criteria category, the second criterion is the only criterion met. The first, third, fourth, and fifth criteria are not met since both the caption and labels do not include any explanation of the included confidence intervals, which leaves an important graphical element undefined. Furthermore, the 'a' and 'b' labels on the two different bar types are undefined, redundant, and clutter the graph. The second criterion is met because the graph labels that are included are legible. Under the *Clear Understanding* criteria category, the first, fifth, sixth, and eighth criteria are met since the data is visually clear, the variables are consistent in both order and color, the data will maintain clarity through graph reproduction or reduction, and a common baseline is used. The second criterion is not met since it is difficult to interpret the key details of the graph due to the inadequate ordering of the groups and the insufficient labeling previously discussed. Following this issue, the third criterion is not met since the data is not displayed in the most effective way. The fourth criterion is not met because the confidence intervals and 'a' and 'b' labels shown in the graph are not explained in the provided caption, causing inconsistency between the data and caption. The seventh criterion does not apply since there are no overlapping data points or superposed data sets. Under the *Meaningful* criteria category, the second, third, and fourth criteria are met since the bar graph is simple and concise, only two dimensions are included, and color is used appropriately and effectively to be able to distinguish between the variables. The first criterion is not met since the data-to-ink ratio is not maximized due to the excess of 'a' and 'b' labels used. Under the *Scaling and Gridlines* criteria category, all the criteria, except for the fifth criterion, are met since one scale is included for the y-axis, there are no changes or breaks in the scale, the axis range and tick mark values are reason-

able for the represented data and allow for the variables to be effectively compared, the logarithmic transformation used portrays the data adequately, and the size of the figure sufficiently displays the data. The fifth criterion does not apply since there are no gridlines included on the graph.

The last step in the scoring process for the graph shown in Figure 2.5 was to obtain the percent correct score, or overall accuracy of the graph. Based on the procedure outlined in Section 3.3, the percent correct score, or overall accuracy of the graph, therefore is calculated as 63.04%. This aligns fairly well with the assigned overall rating of 3.

CHAPTER 4

Statistical Methods and Software

Various statistical methods were used in this research to analyze the raw graph quality scores obtained from the sampled Utah State University (USU) Plan A Master of Science (MS) thesis graphs using the developed graph quality criteria and scoring system discussed in Chapter 3. These methods include: locally estimated scatterplot smoothing (see Section 4.1), jittering (see Section 4.2), least squares regression lines and t-tests (see Section 4.3), Bonferroni Correction (see Section 4.4), confidence intervals and prediction intervals (see Section 4.5), Pearson's correlation coefficient (see Section 4.6), complete linkage clustering with Euclidean distance measure (see Section 4.7), and heat maps and dendrograms (see Section 4.8). Additionally, the various packages utilized from the R software environment for statistical computing and graphics (R Core Team, 2021) are discussed in Section 4.9.

4.1 Locally Estimated Scatterplot Smoothing

Locally estimated scatterplot smoothing (LOESS) is a nonparametric method used for smoothing a series of data in which no assumptions are made about the underlying data (Jacoby, 2000). In this procedure, the span parameter, α, dictates the width of the sliding window, which is essentially a window that moves over the data, sample by sample, and computes a sequence of local linear regressions. The span parameter indicates the proportion of data points that are to be used within each local regression. The value of α can range between 0 and 1 and is determined through a subjective process. The smaller the span value, the smaller the sliding window width is, resulting in a noisier and more jagged curve. Alternatively, the larger the span value, the larger the sliding window width is, thus resulting in a smoother curve.

Smoothing curves were created using the LOESS procedure to analyze graph type and graph creation method counts over time.

4.2 Jittering

Jittering adds random noise to data to prevent overplotting in graphs. Jittering allows for data observations to not be visually clumped together and for value differences to become more apparent. Essentially, jittering provides a neat and simple way to understand the relationships, or lack of relationships, that exist between independent and dependent variables.

The raw graph quality scores obtained from the sampled graphs were jittered since there are five data points, or accuracy observations, within each of the aggregated five-year intervals. To jitter the data, the five years corresponding to each aggregated five-year interval were permutated using a set seed and then distributed as a new year variable to the data observations contained within each of the respective aggregated five-year intervals. This permutation was applied to the entire graph data set so that each set or subset of data had consistent jittering among the data points.

4.3 Least Squares Regression Line and t-test

A least squares regression line is a line that provides the best fit to a set of data points by minimizing the sum of the squared differences between the actual data points and their corresponding predicted values on the line. Mathematically, this line can be represented as seen in Equation 4.1, where x is the independent variable, y is the dependent variable, $\hat{y}$ is the predicted value of y for a particular value of x, $\hat{\beta}_1$ is the slope of the regression line, and $\hat{\beta}_0$ is the y-intercept value:

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0. \tag{4.1}$$

The estimate of $\hat{\beta}_1$ can be obtained using the formula found in Equation 4.2, where x_i and y_i are the observed values of the independent and dependent variables, respectively, $\bar{x}$ and $\bar{y}$ are the means of the independent and dependent variables, respectively, $i = 1, ..., n$, and n is the total number of observations:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{4.2}$$

The estimate of $\hat{\beta}_0$ can be obtained using the formula found in Equation 4.3, where $\bar{x}$, $\bar{y}$, and $\hat{\beta}_1$ are as previously defined:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{4.3}$$

Once the estimates of the slope and intercept are obtained, they can be used to predict the value of the dependent variable for any given value of the independent variable.

In linear regression, a t-test is a statistical hypothesis test that is used to determine the significance of the estimated slope of a regression line. It evaluates whether the slope of the line is significantly different from zero. The null hypothesis, H_0, is that the slope is equal to zero, meaning that there is no linear relationship between the independent and dependent variables. The alternative hypothesis, H_a, is that the slope is not equal to zero, indicating that there is a significant relationship between the variables.

A one-sample t-test is used in linear regression to test H_0. The formula can be found in Equation 4.4, where t is the t-test statistic, $\hat{\beta}_1$ is the estimated slope coefficient as previously defined, β_1 is the hypothesized slope coefficient (which in this case is zero under H_0), and $SE_{\hat{\beta}_1}$ is the standard error of the slope coefficient:

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}}. \tag{4.4}$$

The resulting value of t follows a t-distribution with $n - 2$ degrees of freedom, where n is the total number of observations, as previously defined. After obtaining t, the p-value is calculated. The p-value is essentially the probability of obtaining a more extreme value than the one observed. If the p-value is less than the determined significance level, α, H_0 is rejected, thus concluding that there is statistical significant association between the independent and dependent variables. If the p-value is greater than α, we fail to reject H_0 and the opposite is assumed, i.e., that there is no statistical significant association between the independent and dependent variables. For more information on regression, the reader is invited to visit the work by Lewis-Beck and Lewis-Beck (2015).

Least squares regression lines were fit and t-tests were conducted to analyze the significance of various relationships within the collected graph quality data.

4.4 Bonferroni Correction

The Bonferroni Correction is a statistical method used to prevent data from incorrectly appearing as statistically significant when multiple statistical tests are performed simultaneously on a data set (Armstrong, 2014). Essentially, when multiple statistical tests are executed on a single data set, the chances of obtaining at least one significant result increases with each test performed. The Bonferroni Correction adjusts the significance level, or α value, by dividing the α value by the number of hypothesis tests performed. This adjustment reduces the chances of a type I error (false-positive result) occurring by lowering the α value so that the probability of identifying a significant result is less likely due to chance.

In this research, the Bonferroni Correction was used to adjust α in order to properly interpret the significance results of multiple significance tests that were performed simultaneously on a single data set.

4.5 Confidence and Prediction Intervals

A confidence interval and prediction interval are two different types of intervals used in statistical analysis. A confidence interval is an interval estimate of a population parameter, such as a mean or proportion. It is based on a sample statistic and is used to estimate the range of values in which the true population parameter is likely to fall (Bruce and Bruce, 2007). The confidence level for confidence intervals represents the level of confidence that a population parameter lies within an interval, i.e., a 95% confidence interval suggests that if the sampling process were repeated multiple times, on average, 95% of the intervals generated would contain the true population parameter. For a linear regression line, the confidence interval represents the range in which the true regression lies at the determined confidence level. In general, the confidence interval is wider near the ends of the fitted regression line and is more narrow towards the middle of the line, typically around the mean of the independent variable values. This suggests that the uncertainty in the predicted values increases as the distance from the central values of the independent variable increases.

A prediction interval is an interval estimate of an individual value. It takes into account both the uncertainty in the estimate of the population parameter and the variability of the individual observations, thus the prediction interval is generally wider than the confidence interval (Bruce and Bruce, 2007). The confidence level for prediction intervals represents the probability that a future observation will fall within an interval, i.e., a 95% prediction interval means that we are 95% confident that a new observation will fall within the interval. For a linear regression line, a prediction interval defines a range of values within which a response is likely to occur given a predictor value.

Confidence and prediction intervals were calculated to analyze the reliability of the estimates produced by the least squares regression lines fit within the collected graph quality data.

4.6 Pearson's Correlation Coefficient

Pearson's correlation coefficient, r, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables (Emerson, 2015). The correlation coefficient ranges between the values -1 to 1, where a correlation coefficient of 1 indicates a perfect positive linear relationship between two variables, i.e., as one variable increases, the other variable increases proportionally. A correlation coefficient of -1 indicates a perfect negative linear relationship between two variables, i.e., as one variable increases, the other variable decreases proportionally. This correlation coefficient is calculated using the formula shown in Equation 4.5, where x_i and y_i are the values of the independent and dependent variables, respectively, $\bar{x}$ and $\bar{y}$ are the means of the independent and dependent variables, respectively, $i = 1, ..., n$, and n is the total number of observations:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{4.5}$$

In this research, the Pearson's correlation coefficient was calculated to assess the strength of relationships within the collected graph quality data.

4.7 Complete Linkage Clustering with Euclidean Distance Measure

Cluster analysis is a statistical technique used to group similar data points together based on their attributes and features (Kettenring, 2006). The goal of clustering is to find natural groupings or clusters in the data that maximize the similarity within clusters and minimize the similarity between clusters. Hierarchical clustering, a type of clustering algorithm, creates a hierarchy of clusters by iteratively merging smaller clusters into larger ones based on their similarity (Bridges Jr., 1966). Hierarchical clustering is a useful technique in exploratory data analysis as it can reveal natural patterns and structures within data.

Complete linkage clustering is one method used in hierarchical clustering that calculates the maximum distance between clusters before they are merged (Großwendt and Röglin, 2017). Within each step, the clusters with the smallest maximum pairwise distance are merged. When using the Euclidean distance measure, the distance between two points is calculated using the formula found in Equation 4.6, where x' and y' are two arbitrary data points, x'_i and y'_i are a single coordinate pair of x' and y', respectively, $i = 1, ..., n'$, and n' is the total number of dimensions of the data points x' and y':

$$d(x', y') = \sqrt{\sum_{i=1}^{n'} (x'_i - y'_i)^2}. \tag{4.6}$$

A cluster analysis was conducted using the complete linkage clustering method with Euclidean distance measure to cluster the graph quality criteria and sampled graphs.

4.8 Heat Maps and Dendrograms

Heat maps and dendrograms are two common visualization techniques used in data analysis and clustering. Heat maps are graphical representations of data that come in various forms, but are essentially constructed to communicate relationships between data values in a unique yet digestible way through color-coded systems (Gehlenborg and Wong, 2012). In heat maps, the data values are encoded as colors in a grid-like structure. Heat maps can be useful to identify patterns and relationships between variables as well as highlight outliers and correlations.

Dendrograms are a type of tree diagram that shows hierarchical relationships between data

points (Jolliffe et al., 1989). They are commonly created as an output from hierarchical clustering, which as highlighted in Section 4.7, is a type of clustering algorithm that creates a hierarchy of clusters based on the attributes and features of data points. In a dendrogram, the individual data points are represented by leaves at the bottom of the diagram and the branches above the leaves represent clusters of similar data points. The length of the branches and position of the clusters on the dendrogram indicate the degree of similarity between the data points. The longer the branch, the greater the distance between the clusters. The closer the clusters are on the dendrogram, the more similar the data points within those clusters are to each other. Often, heat maps and dendrograms are used in combination to identify patterns and relationships between variables.

Heat maps and dendrograms were created to visualize the cluster analysis conducted on the graph quality criteria and sampled graphs using the complete linkage clustering method with Euclidean distance measure (see Section 4.7).

4.9 R Packages

In this section, the various R software environment for statistical computing and graphics (R Core Team, 2021) packages that were used for this research are discussed. These packages will only be briefly highlighted, but any documentation or sources utilized will be referenced for further study.

In this research, many of the functions used to read, parse, and extract information from the USU Digital Commons Website (see Section 2.2) came from the base R package, which is essentially the default package that is available with the installation of the R software (R Core Team, 2021). The readLines function was used to access and extract all information from the USU Digital Commons website in HyperText Markup Language (HTML) format. The regular expression functions, grep and gsub, were utilized throughout the extraction process to create variables and to clean and transform the data.

4.9.1 tidyverse

The tidyverse is a collection of data science R packages (Wickham, 2022). These packages allow users to utilize a variety of capabilities and tools for data manipulation, visualization, pro-

gramming, and more (Wickham et al., 2019). A particular set of packages from the tidyverse were used throughout this research, namely: dplyr, ggplot2, magrittr, and readxl R packages, which will be discussed in the next sections.

4.9.2 dplyr

The dplyr R package allows users to easily handle, format, and transform data through the numerous data capabilities it offers (Wickham et al., 2022b). There are several functions that dplyr provides to perform essential data manipulation tasks, such as: creating, selecting, filtering, summarizing, and arranging both variables and data values. The dplyr R package can be installed directly or loaded in combination with the other tidyverse R packages.

In this research, the functions within dplyr were utilized to arrange and transform the extracted data from the USU Digital Commons website (see Section 2.2) to be able to retrieve necessary variables and to format the data so that appropriate plots could be created.

4.9.3 ggplot2

The ggplot2 R package provides useful functions and tools to create intricate and sophisticated plots (Wickham et al., 2022a). The capabilities within ggplot2 allow users to construct good quality graphs and incorporate various graphical elements and aesthetics that can be easily altered and customized (Wickham, 2016). The ggplot2 R package can be installed directly or loaded in combination with the other tidyverse R packages.

Several functions within ggplot2 were used to create a variety of plots for this research. Based on the data being represented, line graphs, bar graphs, scatterplots, and more were constructed. The geom_smooth function was used to create smoothing curves using the locally estimated scatterplot smoothing (LOESS) method (see Section 4.1), which is recommended in the case of less than 1,000 observations (Wickham et al., 2022a).

4.9.4 magrittr

The magrittr R package is designed to make R code more readable and maintainable through the use of the forward-pipe operator (Bach et al., 2022). This tool is used to chain commands

together by forwarding values or expression results into the following expression or function. The magrittr R package can be installed directly or loaded in combination with the other tidyverse R packages.

The forward-pipe operator was used throughout this research to make code more straightforward and concise.

4.9.5 readxl

The readxl R package is used to read data from Excel files into R (Wickham and Bryan, 2022). The package makes it easy for users to import tabular data into a data frame format in R without any external dependencies. The readxl R package can be installed directly or loaded in combination with the other tidyverse R packages.

The read_excel function was used throughout this research to read Excel files into R data frames to easily store, manipulate, and transform data.

4.9.6 WriteXLS

The WriteXLS R package is used to create Excel files from R data frames (Schwartz, 2022). If multiple data frames are written to the same Excel file, each is placed in a separate worksheet.

The WriteXLS function was used throughout this research to write many of the created data frames in R into various Excel files to record, organize, and store data.

4.9.7 gridExtra

The gridExtra R package enables multiple grid objects to be positioned on a single plot (Auguie, 2017). It contains a variety of functions to arrange and customize the layout of both graphs and tables. Since the ggplot2 R package (see Section 4.9.3) was built on grid graphics, the gridExtra R package allows multiple ggplot2 objects to be plotted on a single page.

The grid.arrange function was used in this research to arrange and position plots created from the ggplot2 R package in one figure.

4.9.8 RColorBrewer

The RColorBrewer R package provides several different color palettes to use when creating plots (Neuwirth, 2022). These color palettes, designed by Cynthia Brewer (Brewer, 1994), can be viewed at http://colorbrewer.org/. The palettes are classified into three main color scheme types, namely: sequential, diverging, and qualitative. These color scheme types are explained in Section 3.1. Additionally, the RColorBrewer R package provides a variety of options to specify a desired palette, including a colorblind feature that selects the palettes that display colorblind-friendly colors.

In this research, a variety of functions from RColorBrewer were used to assign colors within the constructed plots.

4.9.9 gplots

The gplots R package contains a variety of tools to plot data (Warnes et al., 2022). It offers multiple functions to construct complex plots as well as enhanced versions of standard plots.

In this research, the heatmap.2 function was used to construct heat maps and dendrograms (see Section 4.8) using the complete linkage clustering method with the Euclidean distance measure (see Section 4.7).

4.9.10 NbClust

The NbClust R package offers 30 indices that each provide the user with a suggestion of the optimal number of clusters in a data set (Charrad et al., 2022). The user can select all or any subset of indices to be used in the process. Additionally, NbClust offers the user the best discovered clustering scheme as well as a function to perform k-means and hierarchical clustering (Charrad et al., 2014).

The NbClust R package was used to determine the optimal number of clusters for both the graph quality criteria and sampled graphs. In each of these cases, not all of the 30 indices were suitable for the data, thus only those that were suitable were used. For the graph quality criteria, 19 different indices were computed, which include the following: kl, ch, hartigan, cindex, db, silhouette, duda, pseudot2, ratkowsky, ball, ptbiserial, gap, mcclain, gamma, gplus, tau, dunn, sdindex,

and sdbw. For the sampled graphs, 20 different indices were computed, which include all the same indices computed for the graph quality criteria, with the addition of the beale index.

CHAPTER 5

Results

After the sampled Utah State University (USU) Plan A Master of Science (MS) theses graphs were obtained using the procedures and steps highlighted in Chapter 2, the statistical graph quality data of these individual graphs were collected and recorded using the developed graph quality criteria and scoring system that are explained in Chapter 3. In this chapter, Section 5.1 overviews the raw graph quality scores and Section 5.2 presents various plots and provides an assessment of the raw graph quality scores.

5.1 Scoring Results

As discussed in Section 2.2.5, five distinct five-year interval permutation sequences were created to determine the assessment order of the 90 sampled graphs across the 18 aggregated five-year intervals from 1930 to 2015. An Excel file was created for each of the permutation sequences to store and organize the raw graph quality scores retrieved from the respective graphs. Each Excel file was built from the same template, shown in Figure 3.4, so that the raw graph quality scores collected during the assessment process could be easily filed, accessed, and later grouped with all other raw graph quality scores gathered within the different permutation sequences.

As each sampled thesis graph was assessed, the scoring information discussed in Section 3.3 was input into the corresponding columns within the respective Excel files. The completed Excel file containing the scoring results for the graphs within the first five-year interval permutation sequence can be found in Figure 5.1. The four Excel files containing the scoring results for the graphs within the second, third, fourth, and fifth five-year interval permutation sequences, respectively, can be found in Appendix C. The five Excel files were read into the R software environment for statistical computing and graphics (R Core Team, 2021) and formatted as data frames. These data frames were then concatenated into one large data frame in order to easily create plots and properly assess the raw graph quality scores.

				Labeling					Clear Understanding								Meaningful				Scaling and Gridlines											Overall Rating
Year Interval	Graph Name	Type of Graph	How Created	1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	5	6	7	Total 1	Total 2	Total 3	Total 4	Percent Correct	(1 = Worst, 5 = Best)
1940	1940-3925-20	Line Graph (multiple graphs)	Hand-drawn/Typewriter	1	1	1	0	1	1	0	0	1	0	1	1	0	0	0	1	1	0	0	0	0	0		1	0.2	0.125	0.125	0.04167	49.17	2
2000	2000-7269-147	Histogram (multiple graphs)	Software	0	1	0	0	1	1	0	0	1	0	1		1	0	1	1	1	1	1	1	0			1	0.1	0.142857	0.1875	0.2	63.04	3
2005	2005-469-57	Line Graph	Software	0	1	1	1	1	0	0	0	1		0	0	1	1	1	1	1	1	1	0	1	1		1	0.2	0.071429	0.25	0.20833	72.98	4
1960	1960-2882-74	Bar Graph (grouped, stacked)	Typewriter	0	1	1	1	1	1	1	1	1	1	1		0	0	0	1	1	1	1	1	1			1	0.2	0.214286	0.125	0.25	78.93	4
1980	1980-7177-75	Line Graph	Typewriter	0	1	0	0	1	0	0	0	1	1	1		1	0	0	1	0	1	1	1	1			1	0.1	0.142857	0.0625	0.25	55.54	2
1955	1955-2689-22	Bar Graph (single, side-by-side)	Hand-drawn/Typewriter	1	1	0	0	1	1	0	0	0	1	1		1	0	1	1	1	0	1	0				1	0.15	0.142857	0.1875	0.125	60.54	2
1945	1945-4740-74	Line Graph	Hand-drawn/Typewriter	0	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	0	1	1	0	1	1		1	0.15	0.1875	0.1875	0.20833	73.33	3
1985	1985-6676-68	Scatterplot	Hand-drawn/Typewriter	0	1	0	0	0	1	0	0	1		1	1	1	0	0	1	0	1	1	1	1			1	0.05	0.178571	0.0625	0.25	54.11	3
1990	1990-4194-47	Line Graph (+ intervals)	Software	1	1	0	0	1	1	0	1	1	0	1			1	1	1	1	1	1	1				1	0.15	0.166667	0.25	0.25	81.67	4
2015	2015-7601-72	Bar Graph (grouped, side-by-side) (+ intervals)	Software	1	1	0	0	1	0	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1		1	0.15	0.21875	0.125	0.20833	70.21	4
1935	1935-1615-27	Line Graph	Hand-drawn/Typewriter	1	0	1	1	1	0	1	0	1		0	1	1	1	1	1	1	1	0	0	1			1	0.2	0.142857	0.25	0.15	74.29	4
1965	1965-2892-74	Line Graph	Hand-drawn/Typewriter	1	1	1	0	1	1	0	0	1		1	1	1	1	0	1	0	0	1	0	0	1		1	0.2	0.178571	0.125	0.125	62.86	3
1995	1995-6580-34	Line Graph	Software	1	1	1	1	0	1	1	1	1	1	1			1	1	1	1	1	1	1				1	0.2	0.25	0.25	0.25	95	5
1970	1970-5082-55	Bar Graph (grouped, side-by-side)	Typewriter	0	1	1	0	1	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	1			1	0.15	0.125	0.0625	0.1	43.75	2
1975	1975-6328-32	Line Graph (multiple graphs) (+ intervals)	Typewriter	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0	0	0			1	0.05	0.09375	0.125	0.1	36.88	2
1930	1930-1570-49	Line Graph	Hand-drawn/Typewriter	0	0	0	0	1	1	0	0	1		0	1	0	1	0	1	0	0	1	0	0	0	1	1	0.05	0.107143	0.125	0.10714	38.93	2
2010	2010-696-60	Bar Graph (grouped, side-by-side) (+ intervals)	Software	1	1	0	0	1	1	1	0	1	1	1	1	1	0	0	1	1	0	1	1	1	1		1	0.15	0.21875	0.125	0.20833	70.21	4
1950	1950-6249-52	Bar Graph (single, side-by-side)	Hand-drawn/Typewriter	1	1	1	0	1	1	1	1	1	1	1		1	0	0	1	1	1	1	1				1	0.2	0.25	0.125	0.25	82.5	5

Criteria:			
Graph has adequate labeling	**Graph has clear understanding**	**Graphical elements add meaning to information displayed**	**Graph has adequate scaling and gridlines**
1. Adequate use of caption and legend	1. Data is visually clear	1. Show as much data as possible (maximize data/ink ratio)	1. One scale is included for every axis
2. Graph labels are legible	2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph	2. Simplest graph of the chosen type is used to convey the necessary information	2. No changes or breaks in scale
3. Graph labels are sufficient	3. Data is accurately displayed in the most effective and appropriate way	3. Minimal dimensions used	3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. All graphical elements are clearly and correctly defined	4. Data is consistent with labeling/text	4. Color is used appropriately and effectively	4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Labeling does not interfere with or clutter the graph	5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)		5. Adequate use of gridlines, do not hide or interfere with data
	6. Clarity of data will be maintained through reproduction/reduction of graph		6. Data transformations are adequately used to display necessary data information
	7. Overlapping data points or superposed data sets are distinguishable		7. Size and aspect ratio display data adequately
	8. A common baseline is used wherever possible		

Fig. 5.1: Excel file containing the raw graph quality scores for the first five-year interval permutation sequence.

5.2 Assessment of the Raw Graph Quality Scores

In Section 1.3, two main exploratory questions were posed with regard to this research. The first exploratory question was to investigate what kinds of statistical graphs have been used primarily in a certain set of publications. More specifically, the types of graphs that were the most common overall, the use trends that exist for these graphs over time, and how these graphs were created. The second exploratory question was to explore how the quality of the graphs changed over time. More specifically, how the overall quality changed over time, how the quality of the different graph types changed over time, and if there exist any trends in the developed graph quality criteria over time.

To answer these questions, several series of plots were created and various assessments were conducted to analyze the quality of the sampled Plan A MS theses graphs. For the first exploratory question, Section 5.2.1 provides a variety of plots analyzing count metrics of the sampled graph types and the graph creation methods used to construct the sampled graphs. Smoothing curves have been fit to the data to analyze any patterns or trends that occur over time. While not included in this chapter, Appendix D contains a summary of the counts of the different notes used to identify unique traits within the sampled graphs (see Section 3.3). For the second exploratory question, Section 5.2.2 displays several scatterplots analyzing different accuracies of the sampled graphs, i.e., combined overall accuracies, individual graph type overall accuracies, and individual criteria category accuracies. Regression lines have been fit to the data and significance tests have been conducted to analyze the relationships between time and the different accuracies of the graphs. Section 5.2.3 provides an analysis on both the average scores of the individual graph quality criteria as well as the criteria category accuracy distributions. Section 5.2.4 presents several scatterplots analyzing the overall ratings of the sampled graphs, i.e., combined overall ratings, individual graph type overall ratings, and individual graph creation method overall ratings. Regression lines have been fit to the data and significance tests have been conducted to analyze the relationships between time and the overall ratings of the graphs. Section 5.2.5 provides a scatterplot analyzing the overall accuracies and the overall ratings of the sampled graphs. A regression line has been fit to the data, a confidence and prediction interval have been calculated, a significance test has been conducted, and a correlation metric has been obtained to analyze the relationship between the overall accuracies and

the overall ratings of the graphs. In addition, a cluster analysis has been conducted and the results have been visualized via a heat map, see Section 5.2.6.

5.2.1 Temporal Analysis of the Count Metrics

As mentioned in Section 5.2, plots displaying count metrics were created and smoothing curves were fit to the data to analyze the first exploratory question, which was to investigate what kinds of statistical graphs have been used primarily in a certain set of publications. Plots were developed to analyze the counts of both the different types of sampled graphs and the different methods of graph creation. Each of these categories were presented as overall totals as well as displayed individually over time within the sampling window, which was determined to be between the years 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015, as described in Section 2.2.2.

In each of the plots that present count totals over time, a smoothing curve, displayed as a light grey line, has been fit to the data to help illustrate any general patterns that occur over time. Smoothing curves were created as described in Section 4.9.3 using the locally estimated scatterplot smoothing (LOESS) procedure, see Section 4.1. For each of the plots that include a smoothing curve, the span parameter was set to 0.75. Therefore, with regards to the data contained in each plot, 75% of the data points were included within each local regression. Different options of span values were examined, but in the end, 0.75 was deemed as the most appropriate to highlight the relationships that are represented throughout these figures.

5.2.1.1 Temporal Analysis of the Graph Type Counts

The different graph types that were sampled include the following: line graphs, bar graphs, scatterplots, dot plots, histograms, and other graphs. The other graph type includes two non-standard graphs. It is also important to note that there was not a single pie chart contained in the sample. Figure 5.2 displays the overall counts of each of the different graph types sampled.

We can see from Figure 5.2 that line graphs and bar graphs had the two highest counts, with line graphs having a slightly higher count than bar graphs. About one-third of all sampled graphs were line graphs and about one-third of all sampled graphs were bar graphs. Scatterplots, which had the next highest count, accounted for about one-sixth of all sampled graphs. The next highest

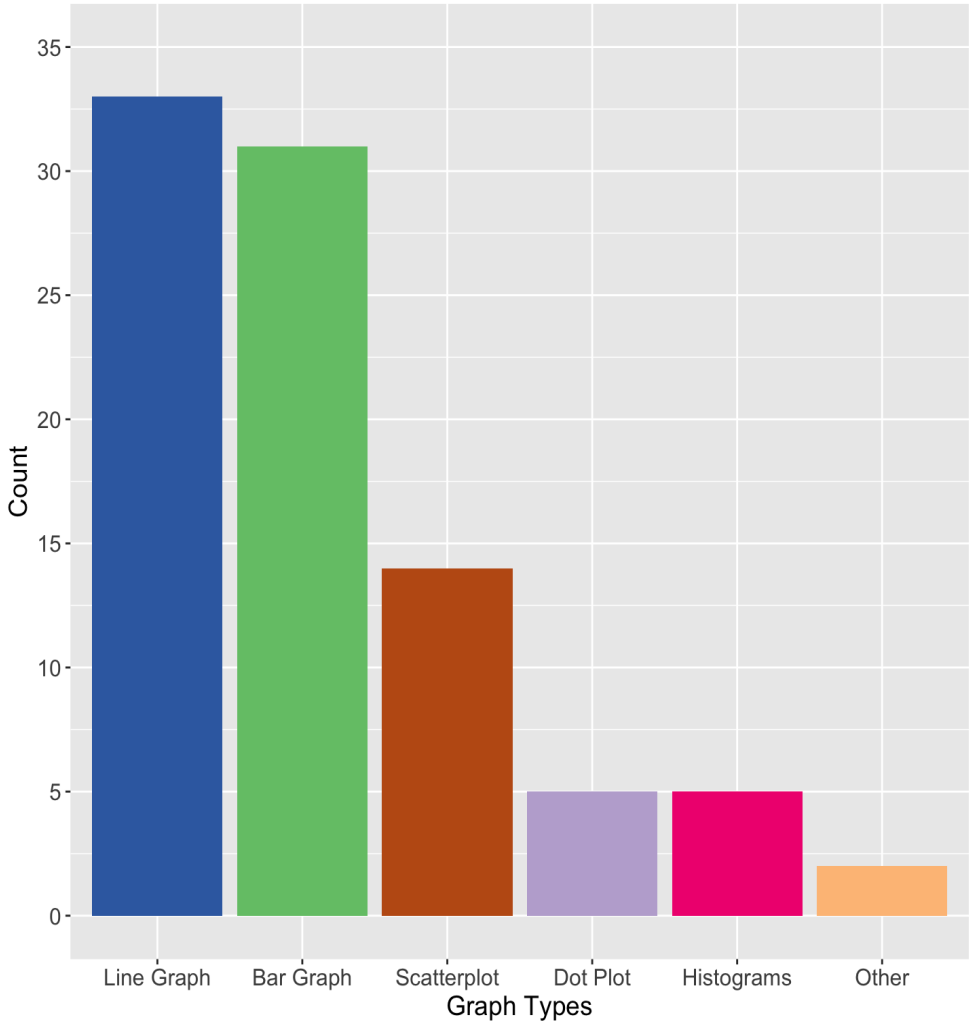
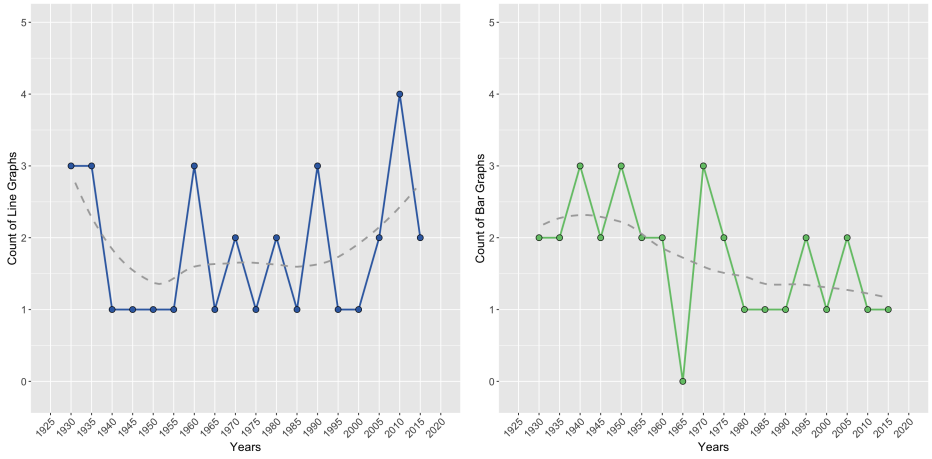


Fig. 5.2: Bar graph of the counts of the sampled graph types.

counts were both dot plots and histograms, with five graphs of each. There were only two other graphs.

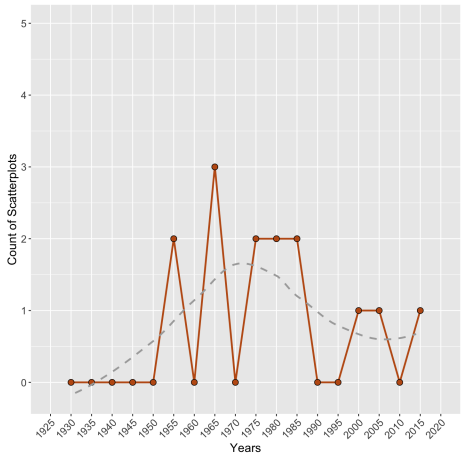
Figure 5.3 depicts the counts of each of the individual graph types from 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015.

We can see from Figures 5.3a and 5.3b that the sample counts over the aggregated five-year intervals for both line graphs and bar graphs seem to typically fluctuate between values one and three. In Figure 5.3a, the highest count for line graphs occurred in the 2010 five-year interval with a value of four. Furthermore, none of the aggregated five-year intervals had a line graph sample count of zero. For bar graph samples, seen in Figure 5.3b, we can identify one instance of a count of zero

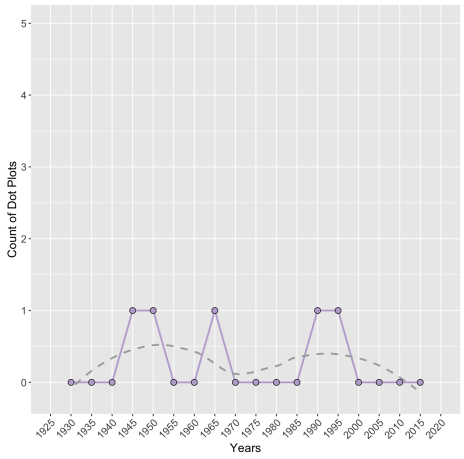


(a) Line graph.

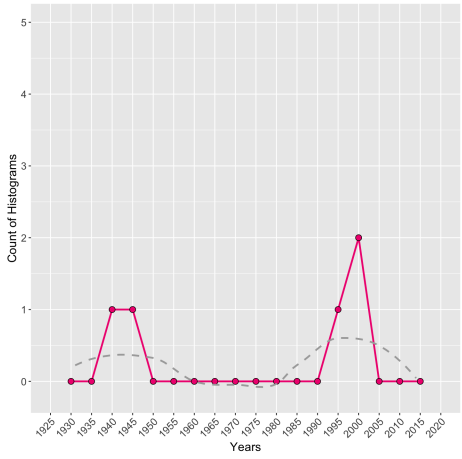
(b) Bar graph.



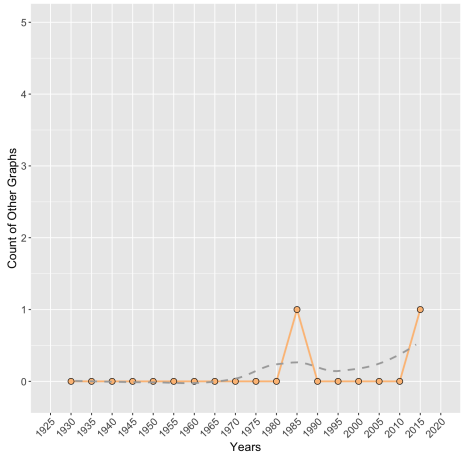
(c) Scatterplot.



(d) Dot plot.



(e) Histogram.



(f) Other graph.

Fig. 5.3: Counts of individual graph types over aggregated five-year intervals from 1930-2015.

in the 1965 five-year interval. In general, it appears that the counts for the line graph and bar graph samples stayed fairly consistent, however, the bar graph sample counts seem to slightly decrease over the aggregated time range.

From Figure 5.3c, it appears that for the first few five-year intervals, the scatterplot sample counts were zero, however, after the 1950 aggregated five-year interval, the scatterplot sample counts fluctuated between values zero and three. There appears to be a slight decline in counts towards the later five-year intervals.

Based on Figure 5.3d, we can see that the sample counts for dot plots were minimal over the aggregated five-year intervals, with values slightly fluctuating between zero and one.

The histogram count samples, shown in Figure 5.3e, appear to spike at both ends of the aggregated time range, with a small spike during the 1940 to 1945 five-year intervals and a slightly larger one during the 1995 to 2000 five-year intervals. Otherwise, the counts were zero for all other five-year intervals.

From Figure 5.3f, we can see that the other graph type only had one count in the 1985 five-year interval and one count in the 2015 five-year interval. Otherwise, the counts were zero for all other five-year intervals.

5.2.1.2 Temporal Analysis of the Graph Creation Method Counts

Next, we want to look at counts of the various graph creation methods that were used to construct the sampled graphs. The methods that were used include the following: software, typewriter, hand-drawn, a combination of hand-drawn and typewriter (hand-drawn/typewriter), and other. The other graph creation method includes one non-standard graph creation method. Figure 5.4 displays the overall counts of the different graph creation methods.

We can see from Figure 5.4 that the hand-drawn/typewriter method had the highest count overall, accounting for nearly 48% of all sampled graphs. The software method had the next highest count, accounting for about 41% of all sampled graphs. The typewriter, hand-drawn, and other methods had the three lowest counts, respectively. The counts for these methods were considerably lower than the counts for the hand-drawn/typewriter and software methods.

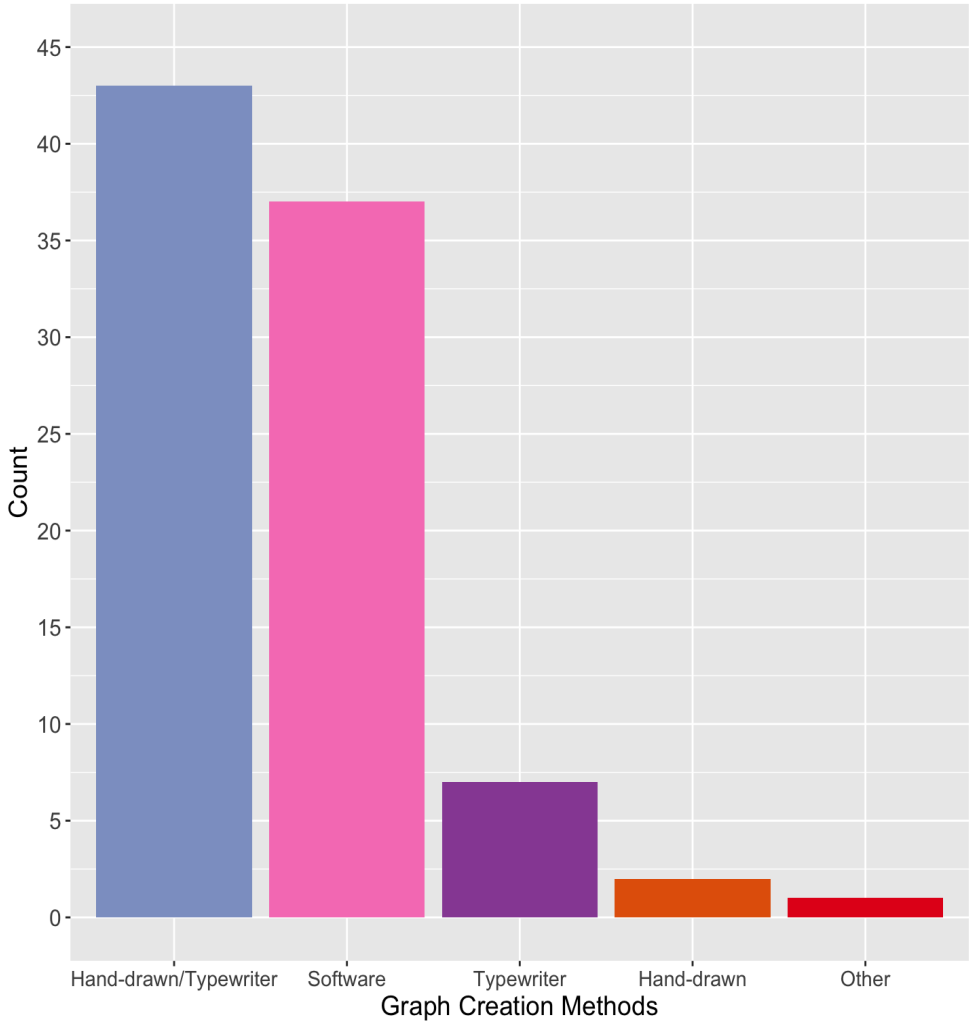
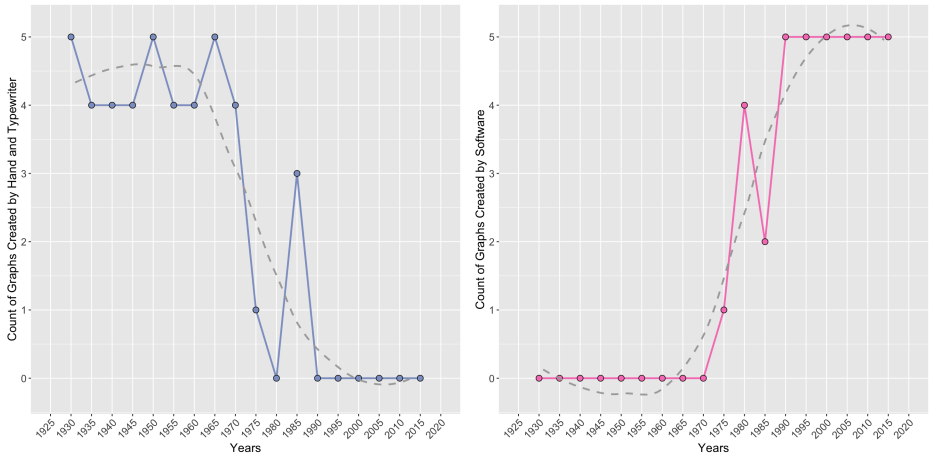


Fig. 5.4: Bar graph of the counts of the graph creation methods.

Figure 5.5 depicts the counts of each of the individual graph creation methods from 1930 to 2019, where time is aggregated into the five-year intervals from 1930 to 2015.

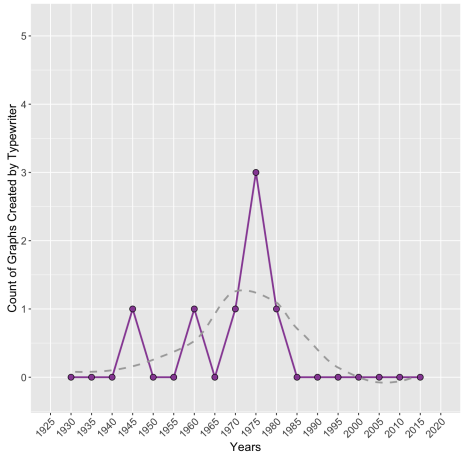
From Figure 5.5a, we can see that the counts for the hand-drawn/typewriter method were very high in the earlier aggregated five-year intervals, however, drastically dropped in the 1975 five-year interval. Other than the spike in the 1985 five-year interval, the counts were zero for the remaining aggregated time range.

The counts for the software method, displayed in Figure 5.5b, show an almost opposite pattern over time compared to the hand-drawn/typewriter counts seen in Figure 5.5a. Here, the counts were zero during the earlier aggregated five-year intervals and then began to increase considerably after

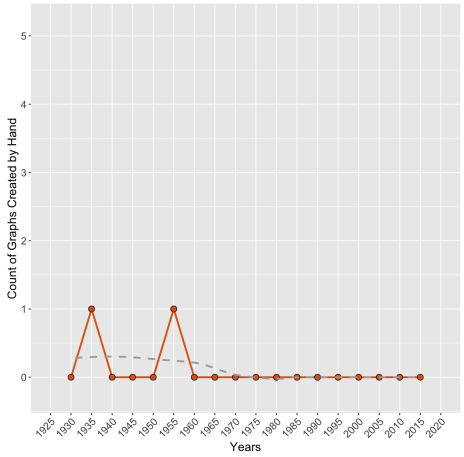


(a) Hand-drawn/Typewriter method.

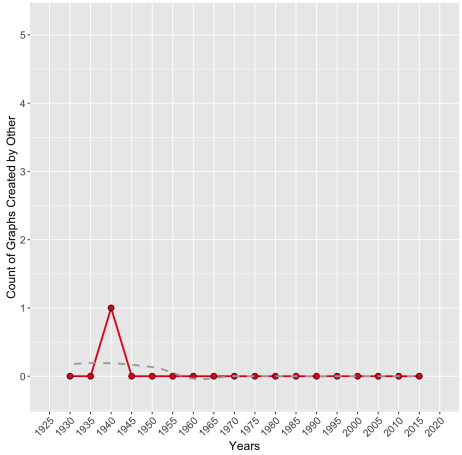
(b) Software method.



(c) Typewriter method.



(d) Hand-drawn method.



(e) Other method.

Fig. 5.5: Counts of individual graph creation methods over aggregated five-year intervals from 1930-2015.

the 1970 five-year interval. The count values from the 1990 to the 2015 five-year interval were five, which is the maximum possible number of counts for a five-year interval.

Based on Figure 5.5c, we can see that the typewriter method counts were zero from the 1930 to the 1940 five-year interval, and then slightly fluctuated between values zero and three from the 1945 to the 1980 five-year interval. It appears that from the 1985 five-year interval and on, the typewriter method counts were zero.

From Figure 5.5d, we can see that the hand-drawn method only had one count in the 1935 five-year interval and one count in the 1955 five-year interval. From the 1960 five-year interval and on, the counts were zero.

The other graph creation method, shown in Figure 5.5e, only had one instance in the 1940 five-year interval. Otherwise, the counts for this method were zero for all other five-year intervals.

5.2.2 Temporal Analysis of the Accuracy Metrics

In this section, scatterplots displaying accuracies were created, regression lines were fit to the data, and significance tests were conducted to investigate the second exploratory question, which was to explore how graph quality has changed over time (see Section 5.2). The various scatterplots presented in this section contain different groupings of data that the second exploratory question considers: combined overall accuracies, individual graph type overall accuracies, and individual criteria category accuracies. These data groupings are displayed over time from the years 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015.

Within each of these plots, a linear regression line, $\hat{y}$, has been fit to the data and a p-value has been provided to indicate whether the regression line is statistically significant or not, as described in Section 4.3. The significance result for each regression line that was part of a series of tests are interpreted using both a significance level of $\alpha = 0.05$ and an adjusted significance level, α', which is calculated using the Bonferroni Correction as explained in Section 4.4. The fitted regression is depicted as a grey line. Additionally, every set or subset of the graph data used to create the scatterplots were jittered, as explained in Section 4.2.

5.2.2.1 Temporal Analysis of the Combined and Individual Graph Type Overall Accuracies

Figure 5.6 displays all the sampled Plan A MS thesis graph accuracies by graph type over time within the sampling window of 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015. The graph accuracies depicted in Figure 5.6 source from the calculated values in the *Percent Correct* column, which is explained in Section 3.3 and can be seen in Figure 5.1, and are referred to as the overall accuracies of the sampled graphs. Based on this plot, we can see that the accuracy observations are very spread out across the time intervals, thus there is no clear accuracy trend for the sampled graphs over the aggregated time range. Additionally, the different graph types show no clear or recognizable relationships over time. The regression line depicted in Figure 5.6 was fit on all data points, independent of graph type. Since the significance test performed on the regression line was a single test, the significance level was not adjusted using the Bonferroni Correction. While there appears to be a slight increasing pattern shown by the fitted regression line, the p-value of 0.159 confirms that there is no significant relationship between time and the overall accuracies of the graphs according to the significance level of $\alpha = 0.05$.

To get a better understanding, Figure 5.7 displays the overall accuracies of the sampled graphs according to the individual graph types over time so that the different graph type accuracy trends can be easily identified. As a note, since the Plan A MS thesis graph sample only contained two observations of the other graph type, the regression line and p-value were not calculated for Figure 5.7f. Since there were five different graph type significance tests performed, the significance level of $\alpha = 0.05$ was adjusted using the Bonferroni Correction (see Section 4.4), thus resulting in an adjusted significance level of $\alpha' = 0.01$.

In Figure 5.7a, we see a slight increasing trend over time for the line graph accuracies. However, since the p-value is 0.081, we fail to reject the null hypothesis according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$ and conclude that there is no association between the independent and dependent variables. From this plot, we can see that the lowest overall accuracy took place in the 1975 five-year interval with a value of about 37%. The highest overall accuracy was within the 1995 five-year interval with a value of about 95%.

From Figure 5.7b, we see that the bar graph accuracy observations are fairly spread out between

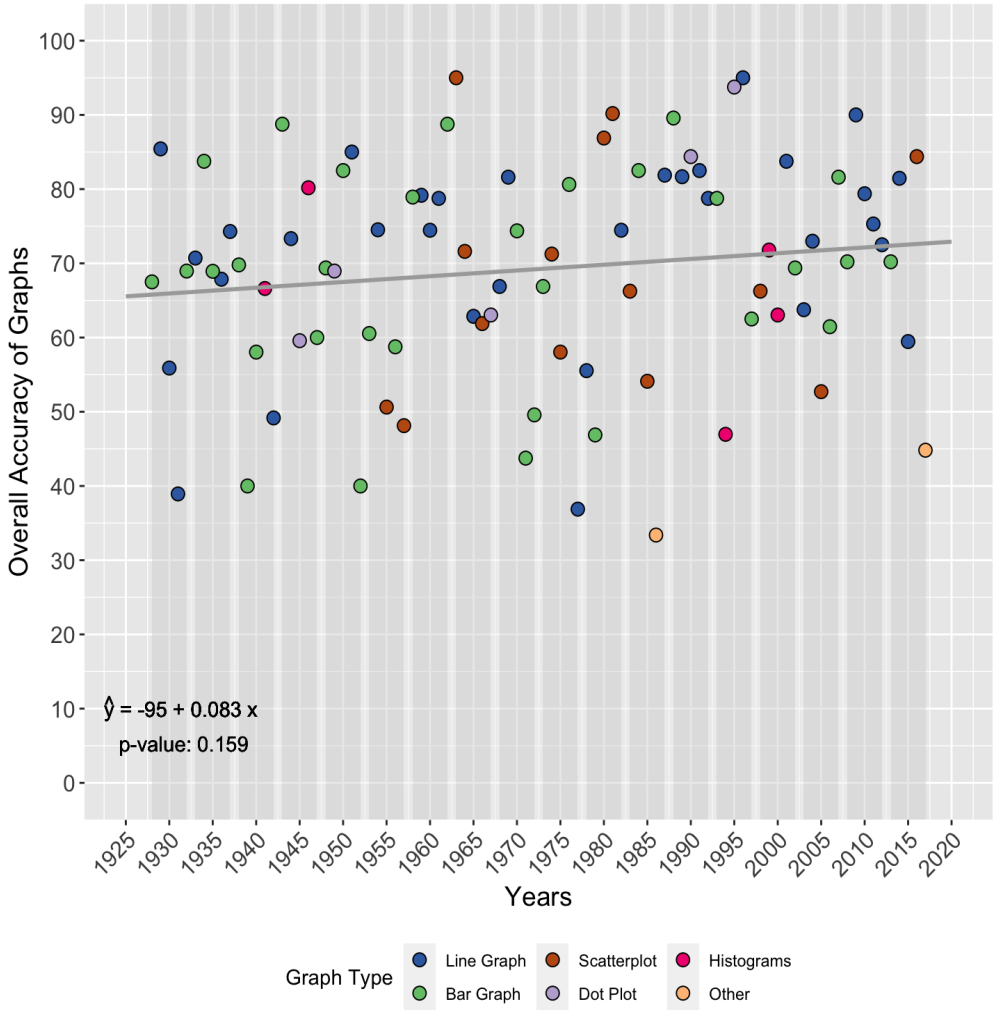


Fig. 5.6: Scatterplot of overall accuracies by graph type over aggregated five-year intervals from 1930-2015.

the values of 40% and 90%, with no clear trend. The p-value of 0.473 confirms that there is no association between time and the overall accuracies of the bar graphs according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$.

The scatterplot accuracies shown in Figure 5.7c demonstrate no clear pattern between the independent and dependent variables. While the fitted regression line indicates an increasing trend, the p-value of 0.524 establishes that there is no association between the variables according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$.

Based on Figure 5.7d, we can see that there is a clear increasing trend between time and the

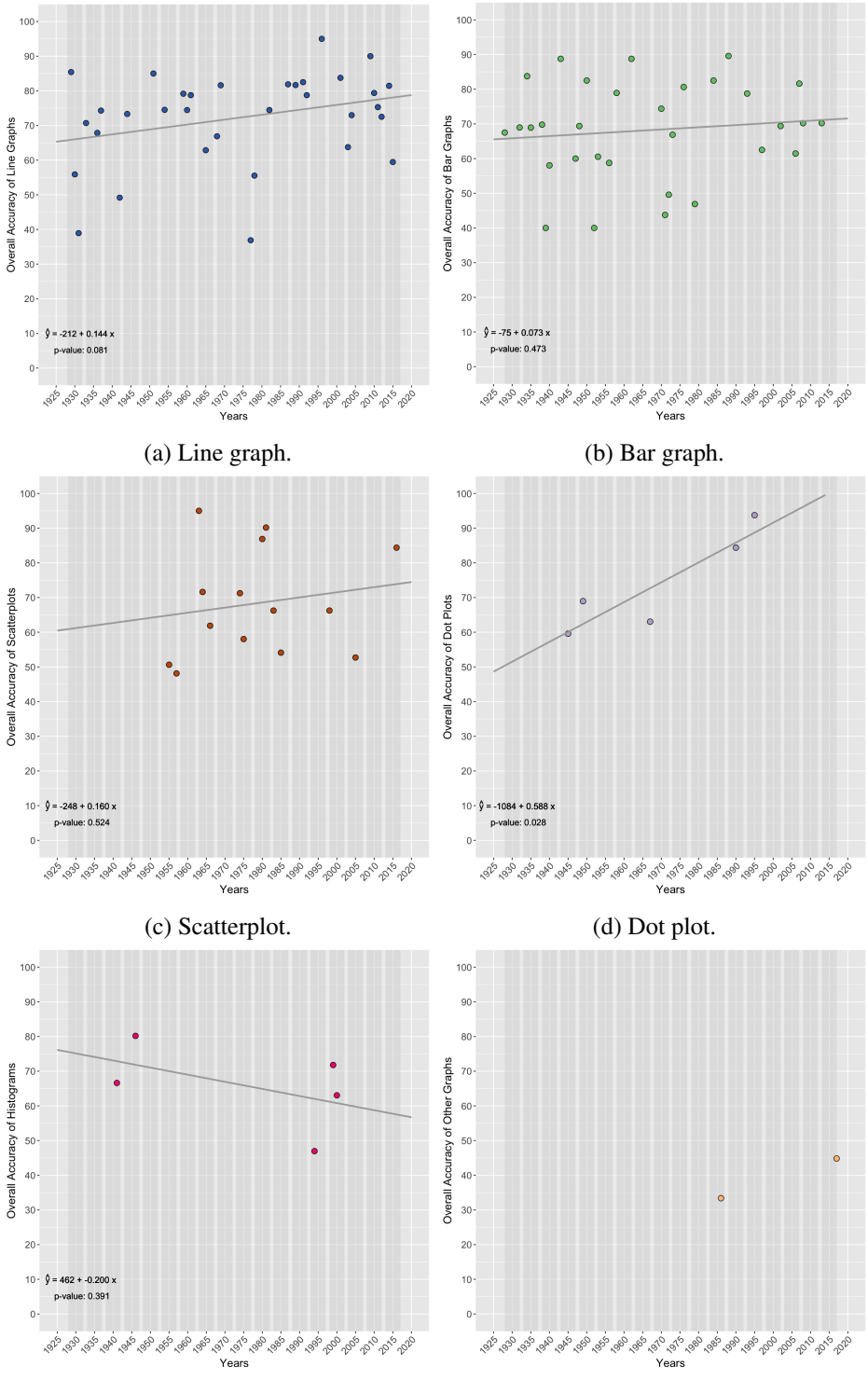


Fig. 5.7: Scatterplots of individual graph type overall accuracies over aggregated five-year intervals from 1930-2015.

overall accuracies of the dot plots. With a p-value of 0.028, according to the original significance level of $\alpha = 0.05$, we would reject the null hypothesis and assume that there is an association between the variables. However, according to the adjusted significance level of $\alpha' = 0.01$, we fail to reject the null hypothesis and conclude that there is no association between the variables. It appears that the lowest overall accuracy took place in the 1945 five-year interval with a value of about 60%, whereas the highest overall accuracy was within the 1995 five-year interval with a value of about 94%.

In Figure 5.7e, we see that there is a distinct decreasing relationship between time and the overall accuracies of the histograms. However, the p-value of 0.391 indicates that there is no significant relationship between these variables according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$. We can see that the lowest overall accuracy took place in the 1995 five-year interval with a value of about 47%, whereas the highest overall accuracy was within the 1945 five-year interval with a value of about 80%.

Figure 5.7f shows the two overall accuracies of the other graphs. As noted previously in this section, due to the small sample size of the other graph type, the regression line and p-value were not included. The first observation took place in the 1985 five-year interval with an overall accuracy value of about 34%. The second observation occurred in the 2015 five-year interval, with an overall accuracy value of about 45%.

5.2.2.2 Temporal Analysis of the Criteria Category Accuracies

In this section, the graph data were grouped by the four different developed criteria categories, which include: *Labeling*, *Clear Understanding*, *Meaningful*, and *Scaling and Gridlines*. Figure 5.8 displays the graph accuracies according to the individual criteria categories from the years 1930 to 2019, where time is is aggregated into five-year intervals from 1930 to 2015. The graph accuracies represented in the scatterplots for each of the four criteria categories originate from the calculated values in the *Total 1* - *Total 4* columns, respectively, which are explained in Section 3.3 and can be seen in Figure 5.1. However, the *Total 1* - *Total 4* columns were weighted for the calculation of the *Percent Correct* column, thus the values in the *Total 1* - *Total 4* columns were reverted back to

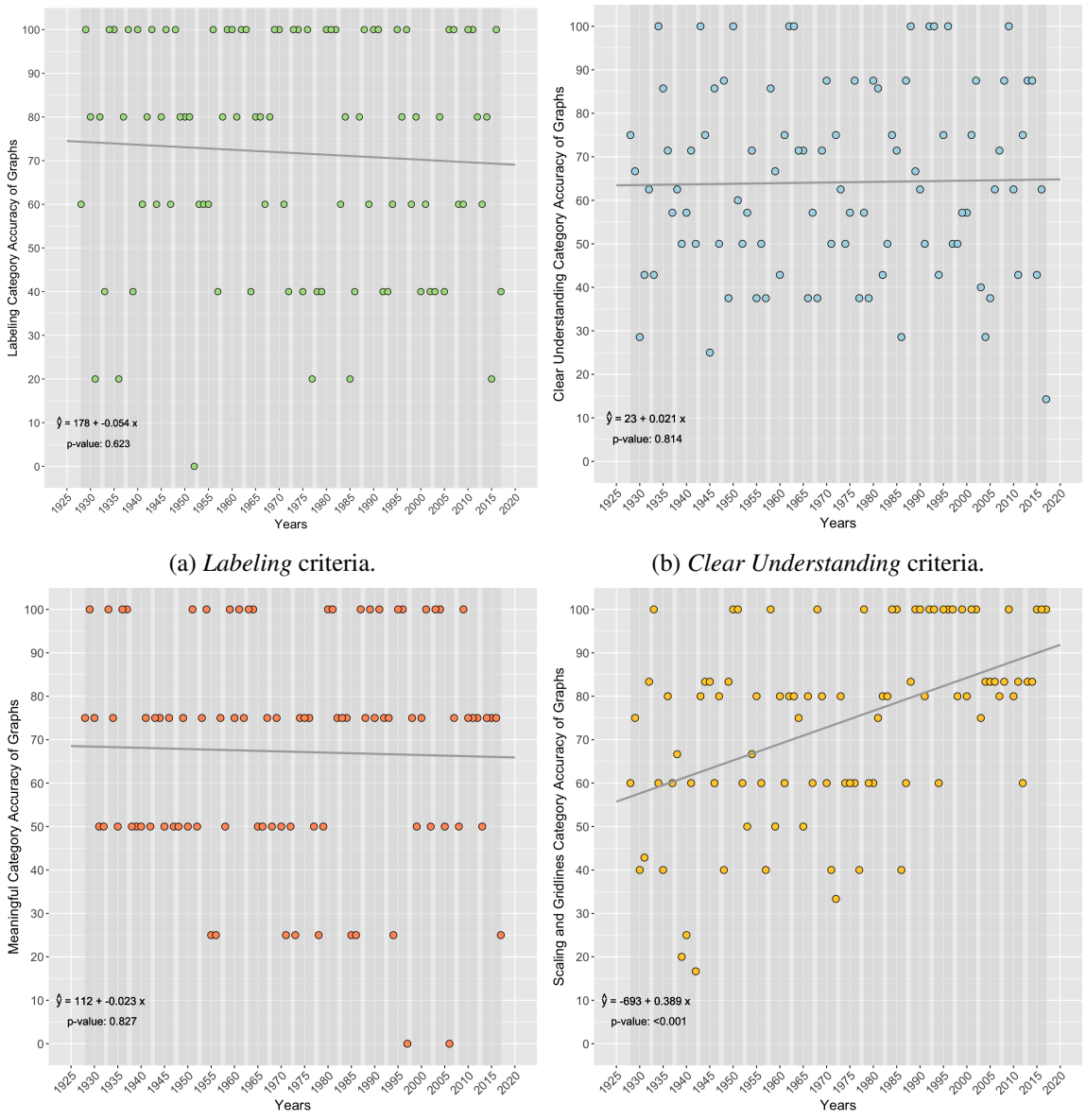
their unweighted form by dividing the values by 0.25. These unweighted criteria category scores, or accuracy values, were plotted for their respective criteria categories in Figure 5.8. Since there were four different criteria category significance tests performed, the significance level of $\alpha = 0.05$ was adjusted using the Bonferroni Correction (see Section 4.4), thus resulting in an adjusted significance level of $\alpha' = 0.0125$.

From Figure 5.8a, we can see that the accuracy values for the *Labeling* criteria category are very spread out over the aggregated time range. The lowest accuracy appears to be within the 1950 five-year interval with a value of 0%. As far as the highest accuracy, there are several observations with values of 100% across multiple aggregated five-year intervals. The regression line shows a slight decreasing trend, however, the p-value of 0.623 indicates that there is no association between the years and accuracy values of the *Labeling* criteria category according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0125$.

Similar to the *Labeling* criteria category accuracies shown in Figure 5.8a, the *Clear Understanding* criteria category accuracies are very spread out, as can be seen in Figure 5.8b. The fitted linear regression line is nearly flat, with a p-value of 0.814, thus concluding that there is no existing relationship between the years and accuracy values of the *Clear Understanding* category according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0125$. From Figure 5.8b, we can see that the observation with the lowest accuracy took place in the 2015 five-year interval with a value of about 14%.

Looking at Figure 5.8c, the accuracy values of the *Meaningful* criteria category appear to be very spread out. It seems that towards the later aggregated five-year intervals, there are several observations that have lower accuracy values, with the lowest accuracies taking place in the 1995 and 2005 five-year intervals, both with a value of 0%. The p-value of 0.827 for the fitted linear regression line indicates no association between the years and accuracy values of the *Meaningful* category according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0125$.

In Figure 5.8d, we can see that the *Scaling and Gridlines* criteria category accuracies demonstrate a general increasing trend over the aggregated time range. Based on the p-value of < 0.001,



(a) *Labeling* criteria.

(b) *Clear Understanding* criteria.

(c) *Meaningful* criteria.

(d) *Scaling and Gridlines* criteria.

Fig. 5.8: Scatterplots of individual criteria category graph accuracies over aggregated five-year intervals from 1930-2015.

we reject the null hypothesis according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0125$ and conclude that there exists a relationship between the years and accuracy values of the *Scaling and Gridlines* category. The lowest accuracy appears to be within the 1940 five-year interval with a value of about 17%. For the highest accuracy, there are several observations with values of 100% across multiple aggregated five-year intervals, however, observations with this value appear more frequently in the later five-year intervals.

5.2.3 Analysis of the Criteria Scores and Criteria Category Accuracies

In this section, we want to analyze the average scores of the individual graph quality criteria and the accuracy distributions of the four criteria categories.

For each of the 24 graph quality criteria, the scores of the 90 sampled graphs were averaged and are displayed as percents in Table 5.1. The criteria scores for the sampled graphs source from the criteria columns discussed in Section 3.3 and can be seen in Figure 5.1. The criteria, along with their corresponding labels, are defined in Figure 3.3. In Section 3.3, the scoring system for the criteria is explained. As a reminder, the sampled graphs did not always meet the precondition for certain criteria and were therefore not assigned a score for those particular criteria. Thus, the number of graphs that met the precondition for each criterion are included in Table 5.1 as well.

From Table 5.1, we can see that the U3 and S3 criteria had the two lowest average scores, respectively, where the U3 criterion had an average score of 21.1% and the S3 criterion had an average score of 32.2%. We can also see that all 90 sampled graphs met the precondition for these two criteria. Alternatively, the S6, S7, and M3 criteria had the three highest average scores, respectively, where the S6 criterion had a perfect score of 100% and the S7 and M3 criteria had the exact same average score of 96.7%. However, it is important to note that for the S6 criterion, only four out of the 90 sampled graphs actually met the precondition. All 90 sampled graphs met the preconditions for the S7 and M3 criteria.

Next, we wanted to look at the distributions of the criteria category accuracies for the 90 sampled graphs. Figure 5.9 displays each of the accuracy distributions for the four different criteria categories, as well as their corresponding average accuracy. The accuracies represented for the

Table 5.1: Average scores of the individual criteria along with the number of graphs that met the pre-condition.

Criterion	Average Score (%)	Number of Graphs With Precondition Met
U3	21.1	90
S3	32.2	90
L4	47.8	90
M1	47.8	90
U2	48.9	90
U1	53.3	90
U7	57.8	64
M2	62.2	90
M4	62.2	90
L3	68.9	90
L1	70.0	90
S5	71.4	28
S2	73.0	89
U6	76.7	90
U5	80.3	71
L5	81.1	90
S4	85.3	68
S1	85.6	90
U4	88.9	90
U8	89.4	85
L2	91.1	90
M3	96.7	90
S7	96.7	90
S6	100.0	4

criteria categories are the unweighted criteria category scores, as explained in Section 5.2.2.2. The average accuracy for each criteria category is shown by a red dot in Figure 5.9.

From Figure 5.9, we can see that there are no substantial differences between the criteria category accuracy distributions and averages. The medians of the criteria categories range between about 63% to 80%, with the highest medians in the *Labeling* and *Scaling and Gridlines* categories and the lowest median in the *Clear Understanding* category. The interquartile range of the *Labeling* criteria category is larger than for the other categories, indicating more accuracy variability. The spreads of the distributions are fairly similar, however, the *Labeling* and *Meaningful* categories exhibit lower accuracies compared to the other two categories. We can see that the *Meaningful* cat-

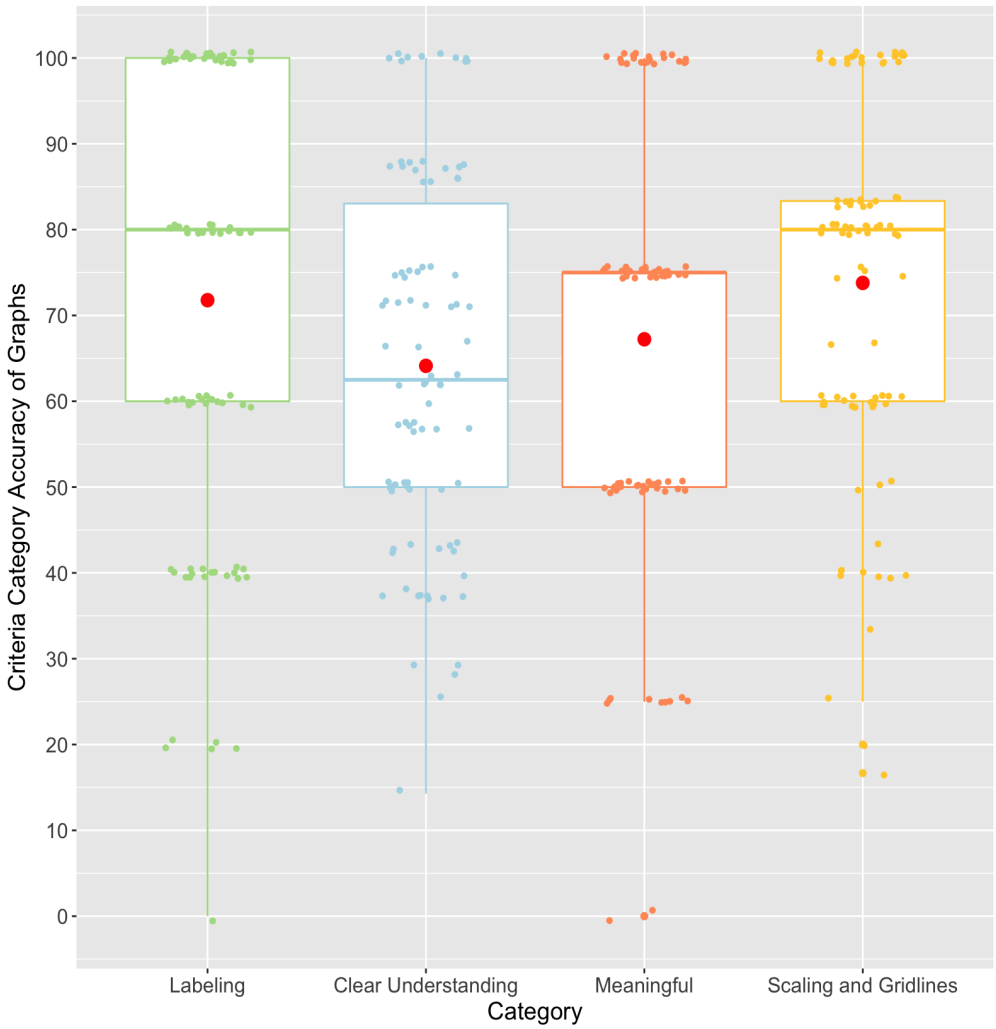


Fig. 5.9: Dot-boxplot of the criteria category accuracies for the 90 sampled graphs. The average accuracy for each criteria category is shown by a red dot.

egory has a few noticeable outliers around 0%. Lastly, the differences between the criteria category accuracy averages are not very drastic. The averages range from about 64% to 74%. We can see that the *Scaling and Gridlines* category had the highest average accuracy and the *Clear Understanding* category had the lowest average accuracy.

5.2.4 Temporal Analysis of the Overall Ratings

Next, scatterplots displaying the overall ratings were created, regression lines were fit to the data, and significance tests were conducted to further investigate the second exploratory question,

which was to explore how graph quality has changed over time (see Section 5.2). In this section, the data were grouped by combined overall ratings, individual graph type overall ratings, and individual graph creation method overall ratings. These data groupings are displayed over time from the years 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015.

The overall ratings depicted in this section source from the *Overall Rating* column, which is explained in Section 3.3 and can be seen in Figure 5.1. Similar to the scatterplots presented in Section 5.2.2, within each of these graphs, a linear regression line, $\hat{y}$, has been fit to the data and a p-value has been provided to indicate whether the regression line is statistically significant or not, as described in Section 4.3. The significance result for each regression line that was part of a series of tests are interpreted using both a significance level of $\alpha = 0.05$ and an adjusted significance level, α', which is calculated using the Bonferroni Correction as explained in Section 4.4. The fitted regression is depicted as either a blue or grey line, depending on the colors used in the scatterplot. Additionally, every set or subset of the graph data used to create the scatterplots were jittered, as explained in Section 4.2.

5.2.4.1 Temporal Analysis of the Overall Ratings of All Sampled Graphs

Figure 5.10 displays all the sampled Plan A MS thesis graph overall ratings over time within the sampling window of 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015. We can see that the overall ratings are typically between the values two and four. It appears that the lowest overall rating took place in the 1970 five-year interval with a value of one. For the highest overall rating, there are two observations with values of five that took place in the 1950 and 1995 five-year intervals. The regression line depicted in Figure 5.10 was fit on all data points independent of any grouping, thus the significance level was not adjusted using the Bonferroni Correction. The fitted regression line indicates an increasing trend and the p-value of 0.021 establishes that there is a significant relationship between time and the overall ratings of the graphs according to the significance level of $\alpha = 0.05$.

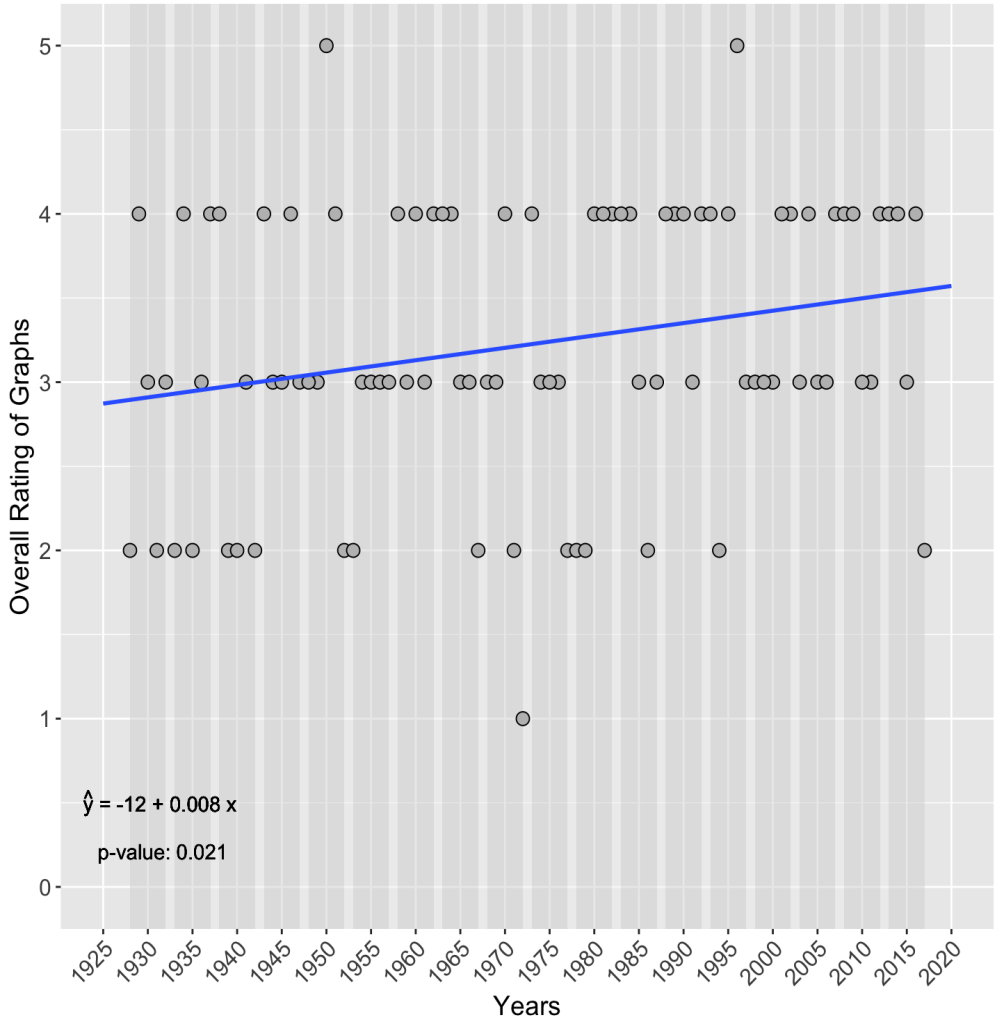


Fig. 5.10: Scatterplot of overall ratings of the sampled USU Plan A MS thesis graphs over aggregated five-year intervals from 1930-2015.

5.2.4.2 Temporal Analysis of the Graph Type Overall Ratings

Next, we want to look at the overall ratings of the different graph types. Figure 5.11 displays the overall ratings of the sampled graphs according to the individual graph types over time from the years 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015. Similar to Figure 5.7f, since the Plan A MS thesis graph sample only contained two observations of the other graph type, the regression line and p-value were not calculated for Figure 5.11f. Since there were five different graph type significance tests performed, the significance level of $\alpha = 0.05$ was adjusted using the Bonferroni Correction (see Section 4.4), thus resulting in an adjusted significance

level of $\alpha' = 0.01$.

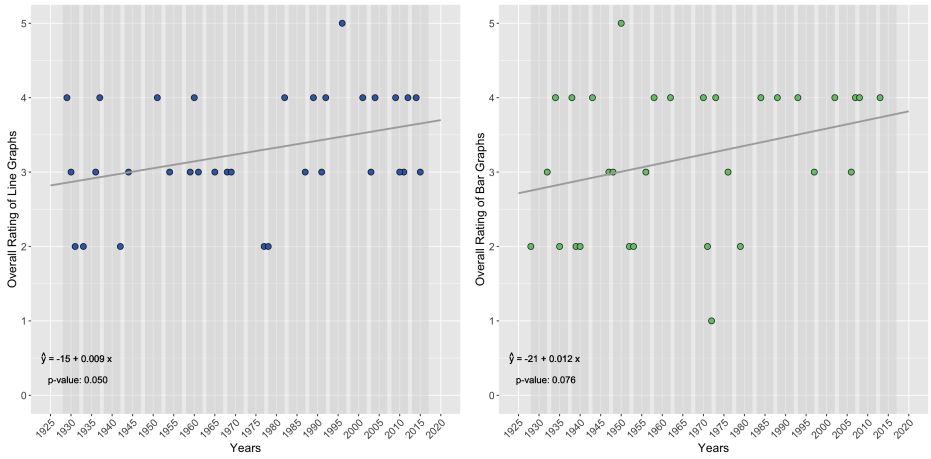
In Figure 5.11a, the line graph overall ratings are somewhat spread out over the aggregated time range between the values two and five. After the 1980 five-year interval, we can see that no observations had a value of two. We can also see that there was only one observation that had a value of five, which took place in the 1995 five-year interval. The fitted regression line demonstrates an increasing trend. However, since the p-value is 0.050459, or 0.05 rounded, we fail to reject the null hypothesis according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$ and conclude that there is no association between the independent and dependent variables.

Figure 5.11b shows the overall ratings of the bar graphs. Similar to Figure 5.11a, after the 1980 five-year interval, we can see that no observations had a value of two. It appears that the lowest overall rating took place in the 1970 five-year interval with a value of one, whereas the highest overall rating was within the 1950 five-year interval with a value of five. While the fitted regression line indicates an increasing trend over time, the p-value of 0.076 establishes that there is no association between the variables according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$.

In Figure 5.11c, we can see that the scatterplot overall ratings range between the values three and four. The linear regression demonstrates a slightly increasing trend, however, the p-value of 0.643 indicates that there is no significant relationship between the variables according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$.

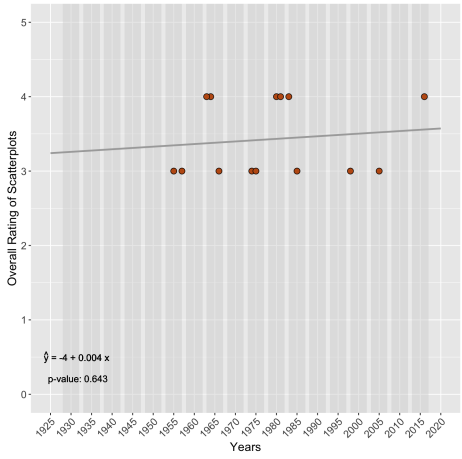
Based on Figure 5.11d, there appears to be an increasing trend between time and the overall ratings of the dot plots. However, the p-value of 0.216 indicates that there is no significant relationship between these variables according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$. We can see that the lowest overall rating took place in the 1965 year-interval with a value of two. The highest overall ratings took place in the 1990 and 1995 five-year intervals, both with a value of four.

In Figure 5.11e, we can see that there is a distinct decreasing relationship between time and the overall ratings of the histograms. However, the p-value of 0.310 indicates that there is no significant

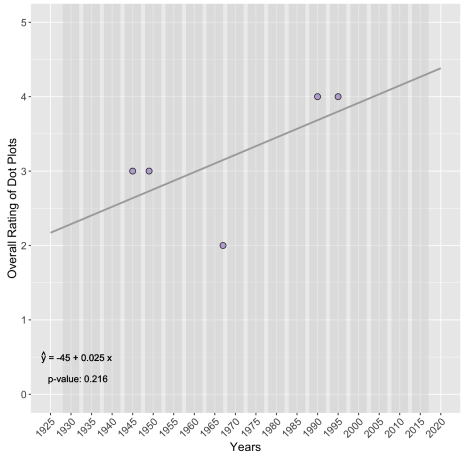


(a) Line graph.

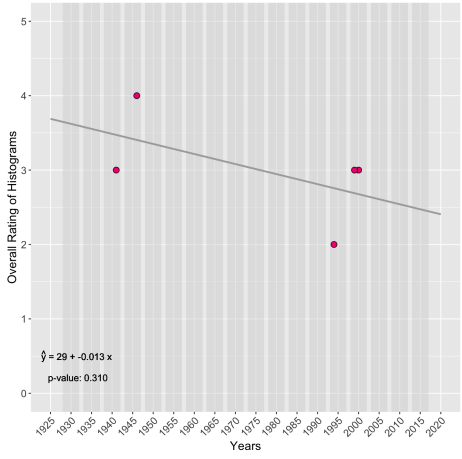
(b) Bar graph.



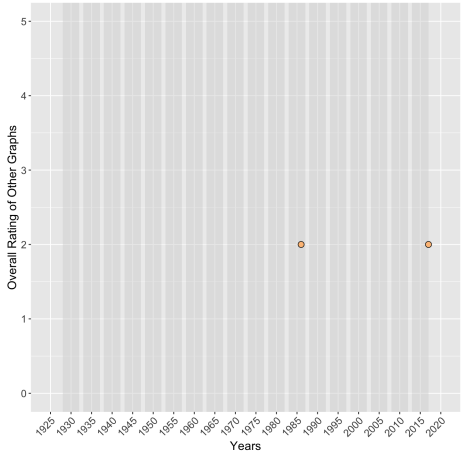
(c) Scatterplot.



(d) Dot plot.



(e) Histogram.



(f) Other graph.

Fig. 5.11: Scatterplots of individual graph type overall ratings over aggregated five-year intervals from 1930-2015.

relationship between these variables according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.01$. We can see that the lowest overall rating took place in the 1995 five-year interval with a value of two, whereas the highest overall rating was within the 1945 five-year interval with a value of four.

Figure 5.11f shows the two overall ratings of the other graphs. As noted previously in this section, due to the small sample size of the other graph type, the regression line and p-value were not included. The observations took place in the 1985 and 2015 five-year intervals, both with an overall rating value of two.

5.2.4.3 Temporal Analysis of the Graph Creation Method Overall Ratings

Next, the overall ratings of the different graph creation methods were plotted over time from the years 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015. Since the Plan A MS thesis graph sample only contained two observations of the hand-drawn graph creation method and one observation of the other graph creation method, the regression lines and p-values were not calculated for these two graph creation methods. Since there were three different graph creation method significance tests performed, the significance level of $\alpha = 0.05$ was adjusted using the Bonferroni Correction (see Section 4.4), thus resulting in an adjusted significance level of $\alpha' = 0.0167$.

No significant relationships between time and the overall ratings of the individual graph creation methods were detected according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0167$. While not included in this section, Figure E.1, which contains the scatterplots of the overall ratings of the individual graph creation methods, can be found in Appendix E.

5.2.5 Analysis of the Overall Accuracies and the Overall Ratings

In this section, a scatterplot was created to analyze and assess the relationship between the overall accuracies and the overall ratings of the 90 sampled graphs, as can be seen in Figure 5.12. A linear regression line, $\hat{y}$, has been fit to the data and a p-value has been provided to indicate whether the regression line is statistically significant or not, as described in Section 4.3. The significance

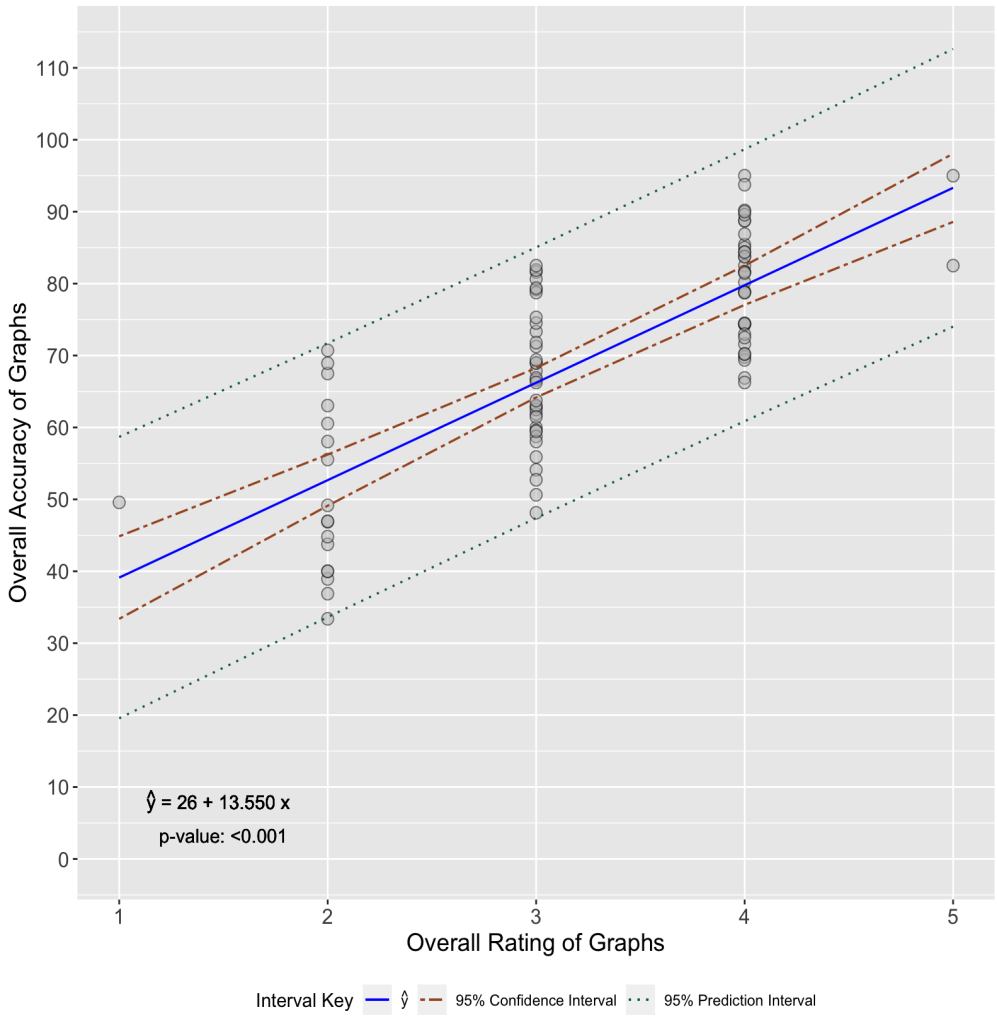


Fig. 5.12: Scatterplot of the overall accuracies vs the overall ratings of the sampled USU Plan A MS thesis graphs.

level is interpreted using a significance level of $\alpha = 0.05$. The fitted regression is depicted as a blue line. Additionally, a 95% confidence interval and 95% prediction interval (see Section 4.5) were calculated and included in Figure 5.12 as well. These intervals are outlined by the spaces between the dashed brown lines and dotted green lines, respectively. The overall accuracies and the overall ratings of the sampled graphs depicted in Figure 5.12 source from the calculated values in the *Percent Correct* column and the values in the *Overall Rating* column, respectively, which were explained in Section 3.3 and can be seen in Figure 5.1.

From Figure 5.12, we can see that there appears to be a clear increasing trend between the

overall accuracies and the overall ratings of the sampled graphs. Based on the p-value of < 0.001, we reject the null hypothesis according to the significance level of $\alpha = 0.05$ and conclude that there exists a relationship between the overall accuracies and the overall ratings of the sampled graphs. To measure the strength of the relationship between the two variables, we calculated the Pearson's correlation coefficient (see Section 4.6) as 0.764. This value indicates a moderately strong positive linear relationship. While it does not suggest a perfectly correlated relationship, it does indicate that there is a strong tendency for the two variables to move in the same, positive direction.

The 95% prediction interval appears to be fairly wide. All observed values fall into this interval, thus suggesting that there is a high degree of uncertainty in the predictions made by the model. As discussed in Section 4.5, the 95% confidence interval is wider near the ends of the fitted linear regression line and is more narrow towards the middle of the line, suggesting that the uncertainty in the estimated overall accuracies increases as we move away from the central values of the overall ratings. The confidence interval appears to be the most narrow, or most certain, around the overall rating value of three. This aligns with the average overall rating, which is calculated to be 3.2.

Figure 5.12 illustrates that the majority of sampled graphs have overall ratings between two and four. There is only one observation with an overall rating of one and two observations with an overall rating of five. For the overall ratings between two and four, the corresponding lowest and highest observed overall accuracies increase as the overall ratings increase. We can see that for the observation with an overall rating of one, the overall accuracy is about 50%. While this is not the lowest observed overall accuracy among the 90 sampled graphs, it is important to note that the sampled graphs were independently assigned an overall rating and assessed against the developed graph quality criteria, as mentioned in Section 3.3. Although intended to be supportive of each other, these processes were not required to be entirely reflective. This same reasoning applies to the two observations that received an overall rating of five, as the overall accuracies obtained are not the highest observed overall accuracies between the 90 sampled graphs. As mentioned in Section 3.3, receiving an overall rating of five does not require a graph to meet all criteria. The overall accuracies for these two graphs appear to be about 82% and 95%.

For the sampled graph that received an overall accuracy of about 82%, the fourth criterion in

the *Labeling* criteria category was not met because the titles for the horizontal and vertical axes were not included. Additionally, the first criterion in the *Meaningful* criteria category was not met because a bar graph was used to represent the data when a dot plot could have been used, which would have further minimized the data-to-ink ratio. Lastly, the second criterion in the *Meaningful* criteria category was not met because the graph itself contained the horizontal axis labels, which made the graph less clear and simple than it could have been. However, the overall rating of five was assigned since both axes were reasonably defined in the overall title of the graph as well as in the caption, and the graph still effectively represented the data in a straightforward and understandable way.

For the sampled graph that received an overall accuracy of about 95%, the fifth criterion in the *Labeling* criteria category was not met since there was an excessive amount of labels included on the horizontal axis. However, the overall rating of five was assigned since the data was still easily interpretable despite of the unnecessary amount of horizontal axis values.

5.2.6 Cluster Analysis

In this section, a cluster analysis was performed and the results have been visualized by means of a heat map (see Section 4.8), as can be seen in Figure 5.13. The heat map created for this research was constructed as described in Section 4.9.9. The 24 graph quality criteria and 90 sampled graphs were clustered using the complete linkage clustering method with Euclidean distance measure, which is explained in Section 4.7.

The heat map, exhibited in Figure 5.13, presents the scored values of the 24 criteria for all 90 sampled graphs. These scored values source from the criteria columns discussed in Section 3.3 and can be seen in Figure 5.1. The criteria, along with their corresponding labels, are defined in Figure 3.3. The main key in the top left corner of Figure 5.13 reflects the values used to score the criteria (see Section 3.3). The 0.5 value in the main key refers to blank score entries, when an individual criterion does not apply to the graph. The 0 and 1 values in the main key directly translate to the 0 and 1 values used within the scoring system.

Two dendrograms (see Section 4.8) were created in Figure 5.13 that represent the results of the

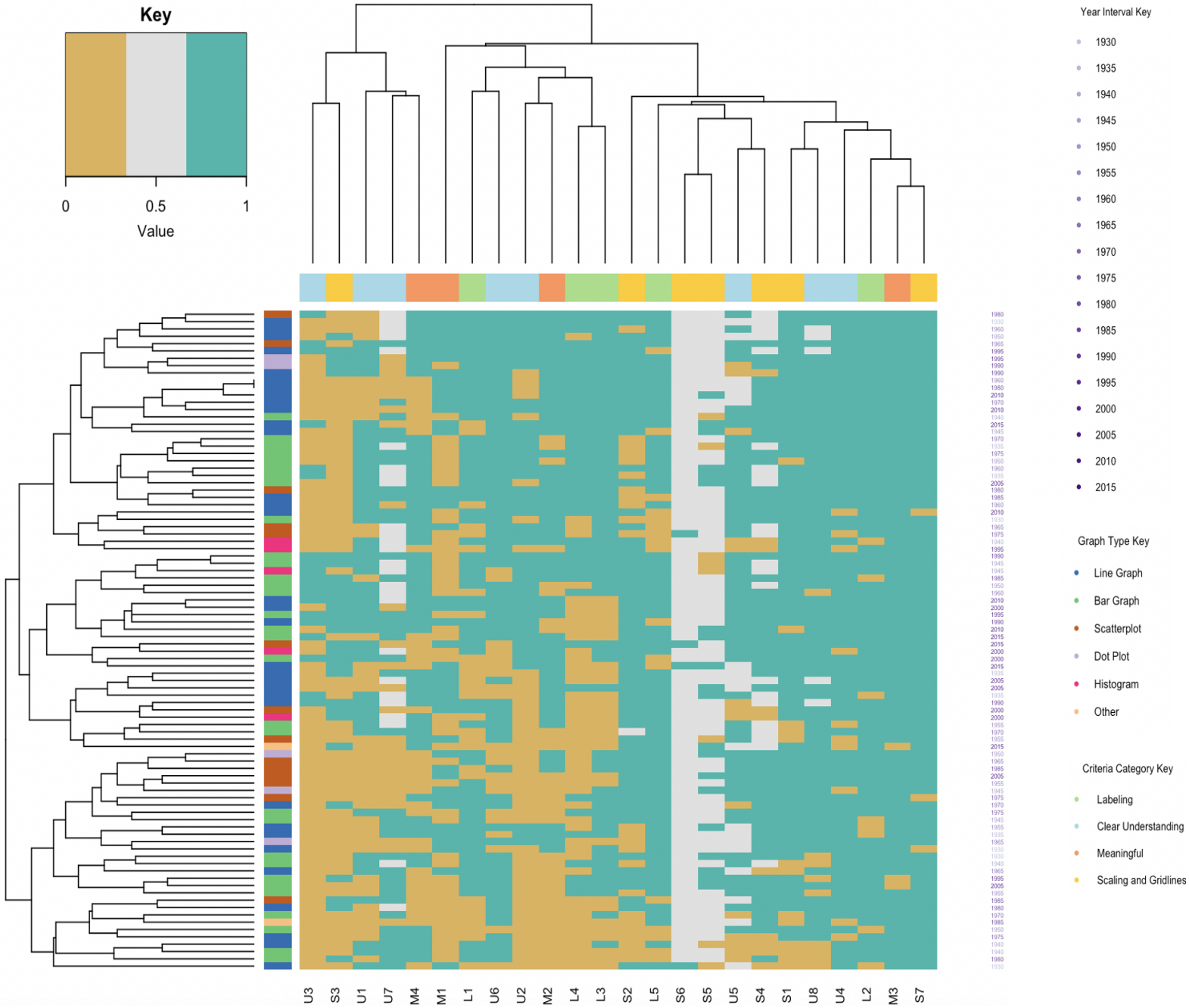


Fig. 5.13: Heat map summary of the criteria scorings for all 90 sampled graphs using the complete linkage clustering method with the Euclidean distance measure. Both a graph type dendrogram and criteria dendrogram were created. Labels were included for the corresponding aggregated five-year intervals of the 90 sampled graphs using a sequential color scheme, with lighter shades indicating earlier five-year intervals and darker shades representing later five-year intervals.

clustering algorithm for the selected parameters. The dendrograms were constructed as described in Section 4.9.9. A dendrogram for graph type can be found vertically on the left side of the plot and a dendrogram for the criteria can be found horizontally at the top of the plot. The corresponding aggregated five-year intervals of the 90 graphs are labeled along the right side of the heatmap, where the five-year intervals are displayed using a sequential purple color scheme, where a darker purple represents a later aggregated five-year interval. The respective color keys for these dendrograms and the aggregated five-year interval labels are displayed on the right side of Figure 5.13.

When looking at the main body of Figure 5.13, we can see that based on the complete linkage hierarchical clustering, the M3 and S7 criteria appear to be the most similar as both of these criteria mostly contain criteria scores of 1. Additionally, the clustering method indicates that the S5 and S6 criteria are the second most similar as both of these criteria mostly contain missing data.

For the criteria dendrogram depicted in Figure 5.13, there visually appear to be two main clusters, however, with a dendrogram of this complexity, a visual interpretation alone is not very meaningful. Using the NbClust R package (Charrad et al., 2014, 2022), 19 different indices, which are highlighted in Section 4.9.10, were each computed to determine the optimal number of clusters for the criteria, see Table 5.2. Based on the results seen in Table 5.2, six of the indices agreed with our visual assessment of two main clusters, however, many of the indices interpreted the optimal number of clusters differently. Again, these inconsistent results indicate that there is no clear optimal number of clusters for the 24 criteria. Furthermore, we can see from the criteria dendrogram that no noteworthy criteria category groupings were formed from the overall clustering.

Table 5.2: Optimal number of clusters for the 24 criteria based on 19 indices.

Number of Clusters	**Frequency**
2	6
3	2
5	2
21	3
23	6

For the graph type dendrogram depicted in Figure 5.13, it is difficult to determine any prominent clustering patterns. While there visually appear to be two main clusters, as previously noted, a visual interpretation alone is not very meaningful with a dendrogram of this complexity. Using the NbClust R package, 20 different indices, which are highlighted in Section 4.9.10, were each computed to determine the optimal number of clusters for the sampled graphs, see Table 5.3. Based on the results seen in Table 5.3, four of the indices agreed with our visual assessment of two main clusters, however, many of the indices interpreted the optimal number of clusters differently. These inconsistent results confirm that there is no clear optimal number of clusters for the 90 sampled graphs. Furthermore, we can see in Figure 5.13 that there are a few noteworthy graph type groups that are formed from the overall clustering. Towards the top of the graph type dendrogram, there is a cluster that contains five line graphs (shown in blue). Within this cluster, there are two line graphs that have identical assessments of all criteria. Slightly below this grouping is a cluster that contains seven bar graphs (shown in green). In the middle section of the graph type dendrogram, there is a cluster that contains four line graphs (shown in blue). There does not appear to be any other noticeable graph type groupings based on the overall clustering.

Table 5.3: Optimal number of clusters for the 90 sampled graphs based on 20 indices.

Number of Clusters	**Frequency**
2	4
3	1
6	2
10	1
73	1
86	2
88	1
89	8

Lastly, when looking at the aggregated five-year interval labels on the right side of Figure 5.13, there do not appear to be any noticeable five-year interval groupings that align with the graph type dendrogram clusters.

CHAPTER 6

Discussion

The main objective of this research is to explore how statistical graph quality has changed over time. In Section 6.1, we interpret the results depicted throughout Chapter 5 according to the exploratory questions defined in Section 1.3. In Section 6.2, we compare our findings according to the literature discussed in Section 1.1 and research hypotheses developed in Section 1.2.

6.1 Explanation and Interpretation of Result Findings

Two main exploratory questions pertaining to this research were presented in Section 1.3. The first exploratory question was to investigate what kinds of statistical graphs have been used primarily in a certain set of publications. More specifically, the types of graphs that were the most common overall, the use trends that exist for these graphs over time, and how these graphs were created. The second exploratory question was to explore how the quality of the graphs changed over time. More specifically, how the overall quality changed over time, how the quality of the different graph types changed over time, and if there exist any trends in the developed graph quality criteria over time.

In Section 2.2.2, we established our population of interest as Utah State University (USU) Plan A Master of Science (MS) thesis reports. We determined that graph samples from these reports would be selected from the years 1930 to 2019, where time is aggregated into 18 five-year intervals from 1930 to 2015. Based on the discussion in Section 2.2.4, we sampled five graphs from each of the 18 aggregated five-year intervals, thus sampling a total of 90 graphs overall. We assessed the quality of the sampled graphs from the USU Plan A MS thesis reports using the developed graph quality criteria (see Section 3.2) and scoring system (see Section 3.3) and collected the respective raw graph quality scores. These scores were presented and assessed throughout Chapter 5.

6.1.1 First Exploratory Question Discussion

For the first exploratory question, Section 5.2.1 analyzed and assessed count metrics of the different graph types that were sampled as well as the different methods of graph creation that were used to create the graphs within the sample. Smoothing curves were fit to the data to illustrate any patterns over time using the locally estimated scatterplot smoothing (LOESS) procedure (see Section 4.1). It is important to note that only five graphs were sampled from each aggregated five-year interval, so the trends shown for the graph types and graph creation methods could be easily influenced by a larger sample size. Thus, while we can discuss the various trends observed in our sample, it is difficult to generalize these trends to be applicable outside of this research (see Section 7.2).

The first component we will discuss from the first exploratory question is: *Which types of graphs are the most common overall?* From Figure 5.2, we can see the overall counts of the different graph types sampled. Based on our sample, line graphs were the most common type of graph overall. As mentioned in Section 5.2.1.1, a little over one-third, or about 36%, of all sampled graphs were line graphs. Bar graphs were also very common, accounting for about 34% of all sampled graphs. Scatterplots, dot plots, histograms, and a few non-standard (other) graphs were also sampled, but were not as common as line graphs and bar graphs. All of these graph types, other than the non-standard graphs, align with the standard and effective graphs suggested by Gordon and Finch (2015), which are highlighted under Principle 5 of their five principles of graphical excellence (see Section 3.1).

The second component we will discuss from the first exploratory question is: *What do the use trends of these statistical graph types look like over a certain time period?* From Figure 5.3, we can see the different graph type counts over time. While the use trend for line graphs, shown in Figure 5.3a, appears to fluctuate somewhat over time, line graphs were still the most consistently used graph type over time compared to the other different graph types. The highest line graph count, as mentioned in Section 5.2.1.1, occurred in the 2010 five-year interval with a value of four. This was the highest graph type count to occur across the aggregated five-year intervals out of the six different graph types observed in the sample. The use trend for bar graphs, as can be seen in Figure

5.3b, appears to slightly decrease over time, but not drastically. It appears that the scatterplot graph type, shown in Figure 5.3c, began to increase in use after its first appearance in the 1950 five-year interval, before slightly declining during the later five-year intervals. While the use trends for dot plots and histograms, shown in Figures 5.3d and 5.3e, respectively, are not very noteworthy, they do show that these graph types were fairly uncommon within this sample over the aggregated time range. The two spikes in histogram counts at the beginning and end of the aggregated time range could be due to the small sample size and not suggest much as far as trend patterns. In Figure 5.3f, the two counts for the other graph type do not occur until the later end of the aggregated time range, with the first count occurring in the 1985 five-year interval. This seems reasonable since we would typically expect non-standard graphs to appear in later years, during or after the rise of graphical software packages, which as highlighted in Section 1.1, began to rise around the year 1976 with the development of the S software. Based on the discussion in Section 1.1, graphical software packages provide default options that offer more creative and obscure graph types, thus providing reasoning as to why we observe the two non-standard graphs during the later end of the aggregated time range.

The third component we will discuss from the first exploratory question is: *How were the graphs created?* From Figure 5.4, we can see the overall counts of the different graph creation methods. Based on our sample, the hand-drawn/typewriter graph creation method was the most common graph creation method overall. As mentioned in Section 5.2.1.2, this method accounts for nearly 48% of all sampled graphs. The software graph creation method was the next most common graph type, accounting for about 41% of all sampled graphs. The typewriter and hand-drawn graph creation methods, as well as a single non-standard (other) graph creation method were also used within the sample, but were not as commonly used as the hand-drawn/typewriter and software methods. From Figure 5.5, we can see the different graph creation method counts over time. Based on the sample conducted, we can see from Figure 5.5a that the hand-drawn/typewriter graph creation method was very common throughout the earlier aggregated five-year intervals and then dropped drastically in the 1975 five-year interval. We can compare the use trend of this graph creation method to the use trend of the software graph creation method in Figure 5.5b, which shows an almost opposite pattern. We can see that the software graph creation method was not used

during the earlier aggregated five-year intervals and then increased in use during and after the 1975 five-year interval. As highlighted in Section 1.1, graphical software packages began to rise around the year 1976 with the development of the S software. These two use trends highlight the rise of graphical software packages, where the hand-drawn/typewriter graph creation method was the primary graph creation method used up until the rise of graphical software packages. Once graphical software packages existed, the software graph creation method quickly became the dominant graph creation method used, and has since been the primary graph creation method used according to our sample. The typewriter and hand-drawn graph creation methods, shown in Figures 5.5c and 5.5d, respectively, also highlight the rise of graphical software packages as the counts for these two methods largely occur before the rise of graphical software packages. Since there is only one count of the other graph creation method, shown in Figure 5.5e, not much can be interpreted from its use trend.

6.1.2 Second Exploratory Question Discussion

For the second exploratory question, Sections 5.2.2 and 5.2.4 assessed the accuracies and the overall ratings of the sampled graphs, respectively, Section 5.2.3 examined the average scores of the individual graph quality criteria and the accuracy distributions of the four criteria categories, Section 5.2.5 analyzed the relationship between the overall accuracies and overall ratings of the sampled graphs, and Section 5.2.6 presented a cluster analysis of the graph quality criteria and sampled graphs. In Sections 5.2.2, 5.2.4, and 5.2.5, a linear regression line, $\hat{y}$, was fit to the data and a p-value was provided to indicate the statistical significance of the regression lines, as described in Section 4.3. The significance result for each regression line that was part of a series of tests were interpreted using both a significance level of $\alpha = 0.05$ and an adjusted significance level, α'. For each series of significance tests, α' was calculated using the Bonferroni Correction as explained in Section 4.4. In Section 5.2.6, the 24 graph quality criteria and 90 sampled graphs were clustered using the complete linkage clustering method with Euclidean distance measure, which is explained in Section 4.7.

The first component we will discuss from the second exploratory question is: *How has the*

overall quality of the graphs changed over time? We will first discuss this question according to the overall accuracies of the sampled graphs and then discuss it according to the overall ratings of the sampled graphs. From Figure 5.6, we can see all of the sampled Plan A MS thesis graph accuracies by graph type over time. Based on our sample, the overall accuracies of the graphs do not appear to have changed over time. As explained in Section 5.2.2.1, the accuracy observations are very spread out over the time intervals and the calculated p-value indicates that there is no significant relationship between time and the overall accuracies of the graphs. From Figure 5.10, we can see all of the sampled Plan A MS thesis overall ratings over time. Based on our sample, the overall ratings of the graphs appear to have changed over time. As explained in Section 5.2.4.1, the fitted regression line indicates an increasing trend and the calculated p-value confirms that there is a significant relationship between time and the overall ratings of the graphs. Thus, we can conclude that according to our sample, the overall quality of graphs, in relation to overall accuracy, has not changed over time. However, the overall quality of graphs, in relation to overall rating, has increased over time.

The second component we will discuss from the second exploratory question is: *How has the quality of different graph types changed over time?* We will discuss this question for each of the graph types according to their overall accuracies and overall ratings. In Figures 5.7 and 5.11, we can see the overall accuracies and overall ratings of the different graph types over time, respectively. Figures 5.7a and 5.11a, Figures 5.7b and 5.11b, and Figures 5.7c and 5.11c show the overall accuracies and overall ratings for line graphs, bar graphs, and scatterplots, respectively. While we can see a somewhat increasing trend over time for both the overall accuracies and overall ratings of line graphs, bar graphs, and scatterplots, the significance test results indicate that there is no association between time and the overall accuracies or overall ratings for any of these graph types. Figures 5.7d and 5.11d and Figures 5.7e and 5.11e, show the overall accuracies and overall ratings for dot plots and histograms, respectively. It is important to note that the sample counts for these two graph types are very small. When a regression line is fit to a small sample size, the estimates of the slope and intercept are less precise, making it difficult to detect a statistically significant relationship between variables. If the sample size were larger, we would be able to detect

more clear and confident trends over time for both the overall accuracies and overall ratings. While each of the regression lines within these plots depict strong trends over time, the significance test results indicate that there is no association between time and the overall accuracies or overall ratings for dot plots and histograms. Figures 5.7f and 5.11f display the overall accuracies and overall ratings for the other graph type, respectively, but due to the small sample count of the other graph type, there is not much to interpret as far as the quality trends over time. Thus, we can conclude that according to our sample, the quality of line graphs, bar graphs, scatterplots, dot plots, histograms, and other graph types have not changed over time.

The third component we will discuss from the second exploratory question is: *Can we analyze trends of the developed criteria over time?* From Figure 5.8, we can see the accuracies of the different criteria categories over time. Figures 5.8a, 5.8b, and 5.8c show the accuracy values for the *Labeling*, *Clear Understanding*, and *Meaningful* criteria categories, respectively. The significance test results indicate that there is no association between time and the accuracies of these criteria categories. Figure 5.8d displays the accuracy values for the *Scaling and Gridlines* criteria category. As explained in Section 5.2.2.2, the fitted regression line indicates an increasing trend and the p-value confirms that there is a significant relationship between time and the *Scaling and Gridlines* criteria category accuracies. Thus, we can conclude that according to our sample, the accuracy of the *Labeling*, *Clear Understanding*, and *Meaningful* criteria categories did not change over time. However, the accuracy of the *Scaling and Gridlines* criteria category increased over time.

Next, we will consider the relationship between the overall accuracies and the overall ratings of the sampled graphs, which can be seen in the scatterplot depicted in Figure 5.12. It is clear that there is an increasing, moderately strong linear relationship between the overall accuracies and overall ratings of the sampled graphs, as indicated by the conducted significance test and calculated Pearson's correlation coefficient (see Section 4.6). While the overall rating is a valuable measure of graph quality, it does not capture all of the information contained in the overall accuracy measure. The overall rating may be sufficient enough in some instances, however, the overall accuracy more adequately represents graph quality and is particularly useful when one is interested in analyzing the quality of specific graphical features in depth.

Additionally, we will discuss the cluster analysis of the 24 graph quality criteria and 90 sampled graphs, which can be seen by the heat map depicted in Figure 5.13. We will also consider the average scores of the graph quality criteria, which are shown in Table 5.1. As discussed in Section 5.2.6, the M3 criterion ('*Minimal dimensions used*') and S7 criterion ('*Size and aspect ratio display data adequately*') were determined to be the most similar according to the complete linkage hierarchical clustering with Euclidean distance measure (see Section 4.7). Using the provided key for the main body of Figure 5.13, we can see that these two criteria mostly contain criteria scores of one. As discussed in Section 5.2.3, these two criteria had the same average score of 96.7%, which was the second highest average score of all the criteria. Additionally, all 90 sampled graphs met the preconditions for these two criteria. Thus, we can conclude that these two criteria performed extremely well according to the 90 sampled graphs. This essentially means that the sampled graphs typically used minimal dimensionality and were appropriately sized to display the data adequately.

The S5 criterion ('*Adequate use of gridlines, do not hide or interfere with data*') and S6 criterion ('*Data transformations are adequately used to display necessary data information*') were determined to be the second most similar according to the complete linkage hierarchical clustering with Euclidean distance measure. In Figure 5.13, we can see that these two criteria mostly contain missing data. Thus, we can conclude that for the 90 sampled graphs, the preconditions for these two criteria were usually not met. In other words, the 90 sampled graphs typically did not contain gridlines or data transformations. However, as discussed in Section 5.2.3, the S6 criterion ('*Data transformations are adequately used to display necessary data information*') had an average score of 100%, which was the highest average score of all the criteria. While we might infer that this criterion is the easiest to fulfill of all the criteria, only four of the 90 sampled graphs met the precondition. Thus, we must take into account the very small sample size of graphs that actually used data transformations and treat potential generalizations with caution.

From Figure 5.13, we can see that the U3 criterion ('*Data is accurately displayed in the most effective and appropriate way*') and S3 criterion ('*Axis ranges and tick mark values are appropriate and meaningful for the type of graph used*') mostly contain criteria scores of zero. As highlighted in Section 5.2.3, these two criteria had the lowest average scores of all the criteria. Additionally,

all 90 sampled graphs met the preconditions for these two criteria. Thus, it is evident that these two criteria had the poorest performance among all the criteria according to the 90 sampled graphs. Based on the assessment of the sampled graphs, frequent problems occurred according to these two criteria. For the U3 criterion ('*Data is accurately displayed in the most effective and appropriate way*'), common issues included the addition of unnecessary or misleading graphical elements such as undefined confidence intervals, extreme use of color or labels, and the presentation of data using a more complicated graph type than necessary. This highlights a tendency among graph creators to overcomplicate their graphs by incorporating excessive or unnecessary elements. Graph creators should focus their attention more on the interpretability of their visualization and less on the aesthetic of their visualization. They should choose the most appropriate type of graph for the data being presented and simplify the graph as much as possible to avoid needless complexity. Additionally, graph creators should consider the audience when creating a graph, assess the level of understanding required, and determine what message needs to be conveyed. For the S3 criterion ('*Axis ranges and tick mark values are appropriate and meaningful for the type of graph used*'), common issues involved unreasonable axis scaling, overuse or incorrect spacing of tick mark values, and numerical axes of bar graphs not beginning at zero. These issues suggest the need for graph creators to exercise caution when selecting axis ranges and tick mark values that can best represent their data. Essentially, graph creators should choose axis ranges that are meaningful for the type of data being displayed and are not too narrow or wide. Tick mark values should be spaced evenly and should be aligned with the tick marks to help viewers accurately interpret the data. Additionally, the numerical axis of bar graphs should always begin at zero to adequately represent differences between magnitudes.

It was determined that the criteria dendrogram, which is shown at the top of Figure 5.13, had no clear optimal number of clusters (see Table 5.2) and showed no noteworthy criteria category groupings according to the overall clustering. Thus, we can conclude that the criteria categories contribute little to the groups of criteria formed in the overall clustering. While the graph type dendrogram, which is shown on the left side of Figure 5.13, also had no clear optimal number of clusters (see Table 5.3), there were a few noteworthy graph type groups formed from the overall

clustering, as were highlighted in Section 5.2.6. Overall, it seems that graph type contributes little to the groups of individual graphs formed in the overall clustering.

6.2 Comparison of Literature and Results

Many of the articles from the literature discussed in Section 1.1 emphasized that graphical software packages do not provide default settings that follow the fundamental statistical visualization standards and practices, making it difficult for the everyday user to create adequate, good quality graphs. As Gordon and Finch (2015) highlighted, a vast amount of the common graphical errors made by users are due to the default settings that many graphical software packages implement. Additionally, Su (2008) explained that, in general, graphical software packages focus their default settings around artistry and creativity, making it more difficult than ever to create good quality graphs. Based on this discussion, our overarching research hypothesis (see Section 1.2) was that graphs created before the rise of graphical software packages are of higher quality than graphs created since the rise of graphical software packages. Based on the discussion of the results in Section 6.1.2, the quality analysis performed on our sample of USU Plan A MS thesis graphs does not reinforce our overarching hypothesis. When looking at overall graph quality, the accuracies of the sampled graphs did not change over time and the overall ratings of the sampled graphs actually increased, or improved, over time. This contradicts our prior belief that graph quality would decrease over time largely due to the development and use of graphical software packages. When looking at overall graph quality by graph type, the quality of the sampled line graphs, bar graphs, scatterplots, dot plots, histograms, and other graph types did not change over time. Thus, based on our sample, it appears that the rise of graphical software packages did not negatively influence the quality of the sampled graphs.

Additionally, as discussed in Section 1.1, it is a growing concern that basic standard graph principles are being forgotten or ignored. The study conducted by Gordon and Finch (2015) demonstrated the gap between valued good graphic principles and actual practice as they observed that a large portion of the graphs they sampled from applied science and statistics journals were rated as poor. Our research also found that the sampled graphs did not completely adhere to good graphic

principles. According to Figure 5.9, based on the assessment of the 90 sampled graphs, the average accuracy for each of the four criteria categories ranged from about 64% to 74%. Additionally, the average overall rating was 3.2, as highlighted in Section 5.2.5. These averages indicate moderate levels of performance, thus suggesting that the sampled graphs did not fully exemplify good graphic principles. According to Su (2008), Excel, which is one of the most widely used graphical software packages, provides easy access to meaningless and unnecessary graphical elements, or 'chartjunk' as Tufte (1997) denoted. Users would need to know how and what to avoid within the default settings offered by many graphical software packages in order to create good quality graphs. In addition to our overarching hypothesis, we also hypothesized that prior to the invention of graphical software packages, graph creators had to rely heavily on principles of good graphics since they were required to construct graphs using their own knowledge and skillset. Thus, as highlighted in Section 1.2, we suspected that principles of good graphics would be more pronounced in earlier created graphs as opposed to graphs created since the rise of graphical software packages. Based on the discussion of the criteria category accuracy results from our sample in Section 6.1.2, these beliefs were invalidated. The accuracy of the *Labeling*, *Clear Understanding*, and *Meaningful* criteria categories did not change over time and the accuracy of the *Scaling and Gridlines* criteria category actually increased, or improved, over time. Thus, it is reasonable to assume that due to the rise of graphical software packages, graphical features such as gridlines, scaling, and axis ranges were more appropriately executed than before.

Symanzik et al. (2016) observed that the most frequent problems encountered among the winning posters of the American Statistical Association (ASA) Poster Competition from the years 2013 to 2016 were three-dimensional (3D) bar charts and pie charts, as discussed in Section 1.1. Prior to beginning the sampling process (see Section 2.2.5), we expected pie charts to be a very common graph type observed among our sample and believed we would see an increase in their use over time with the rise of graphical software packages. However, based on our graph sample from USU Plan A MS thesis reports, not one pie chart was examined. We do not think that the absence of pie charts in our sample necessarily means that pie charts are rarer than we anticipated. Rather, we believe the lack of pie charts within our sample is due to the small sample size from each aggregated five-year

interval. While obtaining the randomly selected graphs for our sample, many pie charts were observed in various thesis reports, however, none were selected according to our randomized sampling process. Thus, we speculate that a larger sample size would have resulted in a greater selection of pie charts. As discussed in Section 6.1.1, line graphs were the most commonly observed graph type within our sample, with bar graphs following closely as the second most common graph type. We also observed that these two graph types were the most consistently used graph types over time. These results are not too surprising since line graphs and bar graphs are very simple, yet useful graph types that can represent a wide range of data.

Furthermore, according to Su (2008), 3D elements are one of the most common issues found in the default charts offered by Excel. As previously mentioned, Symanzik et al. (2016) observed that the most frequent problems encountered among the winning posters of the ASA Poster Competition from the years 2013 to 2016 were three-dimensional (3D) bar charts and pie charts. From our discussion in Section 1.1, we know that 3D graphical elements can be misleading, distorting, and deceptive to the viewer, therefore are prominent examples of chartjunk. Thus, we suspected that 3D graphical elements would be much more prevalent in graphs created since the rise of graphical software packages as opposed to earlier created graphs. When compiling and developing the graph quality criteria (see Section 3.2), we made sure to acknowledge the appropriate dimensionality in graphs. The third criterion in the *Meaningful* category is '*Minimal dimensions used*'. Surprisingly, based on our discussion of the cluster analysis in Section 6.1.2, this criterion, along with the seventh criterion in the *Scaling and Gridlines* criteria category, performed the best among the sampled graphs compared to the other graph quality criteria. Thus, the sampled graphs generally did not contain 3D elements. Our belief that we would see an increase in 3D graphical elements with the rise of graphical software packages was not apparent in the results of our quality analysis on the Plan A MS thesis graph sample.

In Section 1.1, various graphical software packages were highlighted along with the years they were made available to the public. One of the first graphical software packages to become widely available was the S software in 1976 (Venables and Ripley, 2000). After the release of S, various updates were made and different versions were released, such as S-PLUS in 1988 (Venables

and Ripley, 2000). In 1995, the R software environment for statistical computing and graphics (R Core Team, 2021) was made available and has since become one of the most popular graphical software packages today (Peng, 2015). Excel, which is also now one of the most widely used software products, has evolved from various spreadsheet software applications such as Lotus 1-2-3, which was introduced in 1983 and was the first spreadsheet software product to enable chart creation (Raković et al., 2014). Based on our graph sample, the use trend of the software graph creation method reinforces the rise of graphical software packages, which apparently began to rise in 1976 with the development of the S software. As discussed in Section 6.1.1, the software graph creation method began to increase in use during and after the 1975 five-year interval, becoming the only graph creation method used within the graph sample throughout the later aggregated five-year intervals. Prior to the 1975 five-year interval, the software graph creation method was not used within the graph sample. This use trend directly aligns with the years corresponding to the rise of graphical software packages and demonstrates the popularity of graphical software packages among graph creators.

Based on the literature discussed in Section 1.1, we have highlighted several different groups of graph creators throughout this thesis. For instance, Gordon and Finch (2015) sampled from applied science and statistics journals, thus they investigated the work of graph creators primarily from academia or those pursuing doctoral degrees. On the other hand, Symanzik et al. (2016) analyzed graphs from students in elementary through secondary school. In this thesis, we investigated graphs from MS theses, which largely targeted academically trained students who have most likely entered the work force following the completion of their degree, with the exception of a few students who may have continued their studies to pursue a doctorate. By comparing these groups, it is evident that the group of graph creators studied in this thesis are most representative and best embody the everyday graph creators that exist in the real world.

CHAPTER 7

Conclusion and Future Work

In this thesis, we have analyzed the exploratory questions related to how statistical graph quality has changed over time, specifically with the rise of graphical software packages. In Section 7.1, we briefly overview and summarize this research. In Section 7.2, we highlight some research limitations. In Section 7.3, we outline potential options for future research.

7.1 Conclusion

Based on the literature discussed in Section 1.1, we learned that many of the default settings that graphical software packages use do not always align with standard graph principles. Users have to understand how to navigate around these inadequate default settings in order to create good quality graphs. From this discussion, we established our research hypotheses and developed a few exploratory questions to guide the direction of our analysis. As highlighted in Section 1.2, our hypotheses were primarily centered around the belief that graphs created before the rise of graphical software packages are of higher quality than graphs created since the rise of graphical software packages. For our exploratory questions (see Section 1.3), we wanted to investigate what kinds of statistical graphs have been used in a certain set of publications and how the quality of these graphs changed over time.

To begin our research, we first developed a set of graph quality criteria (see Section 3.2) and created a scoring system (see Section 3.3) to evaluate the quality of the sampled graphs. The set of graph quality criteria were constructed by condensing and summarizing the standards from the graph quality literature resources discussed in Section 3.1. The list of these standards can be found in Appendix B. Instead of being tailored to evaluate specific types of graphs, the criteria were generalized to evaluate the quality of all types of graphs. We then identified a certain set of publications we would explore for this research, referred to as our population of interest, which consists of Utah State University (USU) Plan A Master of Science (MS) thesis reports from the years 1930 to 2019

(see Section 2.2.2). In Section 2.2.5, we established a sampling process that was used to dictate how we sampled graphs from these reports.

Next, we used the developed graph quality criteria and scoring system to evaluate the quality of the sampled graphs and collect the raw graph quality scores, which were analyzed and assessed throughout Chapter 5. In Chapter 6, we discussed and interpreted the results from Chapter 5 according to the exploratory questions, literature, and hypotheses highlighted in Chapter 1. We found that there is no evidence that the quality of the sampled graphs from the USU Plan A MS thesis reports declined since the rise of graphical software packages. We also found that many of our research results did not directly align with or support the literature discussion in Section 1.1. However, as discussed in Section 6.2, we did observe that the timeline of the rise of graphical software packages (see Section 1.1) is supported by the observed use trend of the software graph creation method within our sample. This validated our discussion in Section 1.1 on how graphical software packages have become widely available to many different fields and settings, and have quickly become the primary method used to create graphs.

7.2 Research Limitations

The major limiting factor of this research was the sample size of our study. While sampling five graphs from each of the 18 aggregated five-year intervals was the most feasible option due to time constraints, it greatly impacted the power of our study. With a small sample size, it is difficult to detect significant effects and draw meaningful conclusions from the data. While we have interpreted our results and analyzed them according to the literature discussed in Section 1.1, it is difficult to generalize our findings and fully analyze our initial research hypotheses since our population of interest may not be adequately represented. Having a larger sample size would provide more insight into how the quality of USU Plan A MS thesis graphs has changed over time, which in turn would allow us to make stronger conclusions on how graph quality in general has changed over time due to the rise of graphical software packages.

While time did not severely impact our research, it did constrain the amount of graphs that could be sampled from each five-year interval, as previously discussed. More time would have

allowed for a larger sample size to be taken from each aggregated five-year interval, which would have enhanced the validity and generalizability of our research results.

Additionally, our research focused on analyzing graphs from USU Plan A MS thesis reports as our population of interest (see Section 2.2.2). However, as discussed in Section 2.1, the available theses on the USU Digital Commons website do not fully represent all USU departments over time. As of September 2022, almost all dissertations, theses, creative projects, and reports since 2008 have been uploaded to the USU Digital Commons website. For the years before 2008, not all files have been completely scanned and made available on the USU Digital Commons website. Before 2008, patrons and USU departments requested digitization of documents, but not all requests could be fulfilled, and certain departments such as Geology, Mathematics, Nutrition, Dietetics, and Food Sciences (NDFS), and Psychology were given priority. Furthermore, all documents had to be physically scanned and uploaded to the USU Digital Commons website before 2008, and many were too fragile for the scanners used at the time. Thus, the available theses in digital formats do not fully represent all USU departments over time. Analysis performed on a population of interest that is not fully represented can introduce bias, potentially resulting in findings that are inaccurate or unreliable. To be able to make stronger and more generalizable conclusions on statistical graph quality over time based on our population of interest, a more representative sample would need to obtained from the USU Digital Commons website. This would require digitizing and uploading more documents from the various USU departments over time.

Another limitation within this research pertains to the determined values used to score the criteria. Assigning scores of only 0 or 1 to the criteria does not provide a comprehensive evaluation or fully capture the quality of a graph within the individual criteria categories. A graph that partially satisfies a criterion but fails to meet all of its requirements cannot be accurately assessed using a score of 0 or 1 for that particular criterion.

7.3 Future Work

There are several options for potential future research that could be used to investigate statistical graph quality trends over time. These options exist among various parameters and populations

of interest.

Future work could include sampling thesis reports from other degrees within USU to see how the analysis conclusions compare. Since our population of interest for this research only focused on graphs from MS thesis reports, it would be interesting to analyze graph quality trends over time from degrees that do not encompass the scientist domain. As discussed throughout this thesis, graphical software packages have become widely available and accessible to people within all types of fields and with varying levels of graph creation experience. Typically, scientists are required to create graphs throughout their education, so examining the quality of graphs within other degrees that may be less exposed to graph creation could potentially bring insightful results to the investigation of graph quality over time. Such a study could further stratify by department and not only by five-year intervals, however, this would require that the same department existed for the entire 90-year period. One could also look at dissertations, but as seen in Figure 2.2, those have only been made available digitally since the 1950 five-year interval. Additionally, one could stratify based on the six main graph types and select a predetermined number of each graph type from every five-year interval instead of relying on chance to determine the selection of graph types in the sample.

Another avenue for potential future work would be to broaden the scope of this research to sample and analyze the quality of statistical graphs in a subset of published journals or textbooks, or even student research conducted at other universities. This could help to identify new research questions, increase the generalizability of the results, and enhance the significance of this research by demonstrating the relevance to a wider audience.

In this research, we developed a general set of graph quality criteria to assess the quality of all possible graph types. Instead of a general framework to be used on all graph types, future work could include developing graph quality criteria for specific graph types by building on and fine-tuning existing ideas from the statistical literature, such as the coded features categorized under Principle 5 of the five principles of graphical excellence created by Gordon and Finch (2015), '*use graphical forms consistent with Principles 1 to 4*', which can be found in Table B.2. This process could be done similar to the way the graph quality criteria were developed within this research (see Section 3.2). This would allow for a more targeted and accurate evaluation of graph quality among

the different graph types and would allow for the quality of graphs within the same graph type to be more comparable. Furthermore, it would better capture all the unique features and qualities that can be missed or overlooked with a generalized set of graph quality criteria.

As discussed in Section 7.2, the scoring values of 0 or 1 used to assess the individual graph quality criteria limited the ability to fully and adequately capture graph quality within this research. Thus, future work could include a more detailed and refined scoring scale to score the graph quality criteria, such as a Likert scale. Likert scales typically use a five-point or seven-point scale, which can measure responses with a greater degree of nuance. Using a Likert scale to score the graph quality criteria would allow for graph quality to be evaluated at a more comprehensive and detailed level.

Finally, future work could also include increasing the amount of graphs sampled over time. This would allow for more confident conclusions to be made about statistical graph quality trends. Increasing the sample size can increase the statistical power of relationships and the ability to detect significant results. It can also increase precision and reduce random sampling error by providing a more accurate representation of statistical graph quality trends over time. As previously outlined, such a study could further stratify by department as long as the departments existed for the entire 90-year period. Additionally, a multivariate regression analysis could be conducted to analyze the relationship between multiple independent variables and graph quality. This analysis could incorporate independent variables such as graph type, graph creation method, and department.

REFERENCES

Armstrong, R. A. (2014). When to Use the Bonferroni Correction. *Ophthalmic and Physiological Optics*, 34(5):502–508. https://doi.org/10.1111/opo.12131.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3 (https://CRAN.R-project.org/package=gridExtra).

Bach, S. M., Wickham, H., and Henry, L. (2022). *magrittr: A Forward-Pipe Operator for R*. R package version 2.0.3 (https://CRAN.R-project.org/package=magrittr).

Brewer, C. A. (1994). Chapter 7 - Color Use Guidelines for Mapping and Visualization. In MacEachren, A. M. and Taylor, D. R. F., editors, *Visualization in Modern Cartography*, volume 2 of *Modern Cartography Series*, pages 123–147. Elsevier Science, Tarrytown, New York. https://doi.org/10.1016/B978-0-08-042415-6.50014-4.

Bridges Jr., C. C. (1966). Hierarchical Cluster Analysis. *Psychological Reports*, 18(3):851–854. https://doi.org/10.2466/pr0.1966.18.3.851.

Bruce, P. and Bruce, A. (2007). *Practical Statistics for Data Scientists*. O'Reilly Media, Sebastopol, California.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Duxbury Press, Boston, Massachusetts. https://doi.org/10.1201/9781351072304.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6):1–36. https://doi.org/10.18637/jss.v061.i06.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2022). *NbClust: Determining the Best Number of Clusters in a Data Set*. R package version 3.0.1 (https://CRAN.R-project.org/package=NbClust).

Cleveland, W. S. (1984). Graphs in Scientific Publications. *The American Statistician*, 38(4):261–269. https://doi.org/10.1080/00031305.1984.10483223.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Wadsworth Advanced Books and Software, Monterey, California.

Cleveland, W. S. (1994). *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey, 2 edition.

Cleveland, W. S. and Devlin, S. J. (1980). Calendar Effects in Monthly Time Series: Detection by Spectrum Analysis and Graphical Methods. *Journal of the American Statistical Association*, 75(371):487–496. https://doi.org/10.1080/01621459.1980.10477500.

Cleveland, W. S. and McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554. https://doi.org/10.1080/01621459.1984.10478080.

DigitalCommons (2023). Digital Commons Network. https://network.bepress.com/.

DigitalCommons@USU (2021a). All Graduate Plan B and Other Reports. https://digitalcommons.usu.edu/gradreports/.

DigitalCommons@USU (2021b). All Graduate Theses and Dissertations. https://digitalcommons.usu.edu/etd/index.html.

Emerson, R. W. (2015). Causation and Pearson's Correlation Coefficient. *Journal of Visual Impairment & Blindness*, 109(3):242–244. https://doi.org/10.1177/0145482X1510900311.

Few, S. (2004). Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message. *Intelligent Enterprise*. https://www.perceptualedge.com/articles/ie/the_right_graph.pdf.

Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, Oakland, California.

Gehlenborg, N. and Wong, B. (2012). Heat Maps. *Nature Methods*, 9(3):213. https://doi.org/10.1038/nmeth.1902.

Gordon, I. and Finch, S. (2015). Statistician Heal Thyself: Have We Lost the Plot? *Journal of Computational and Graphical Statistics*, 24(4):1210–1229. https://doi.org/10.1080/10618600.2014.989324.

Großwendt, A. and Röglin, H. (2017). Improved Analysis of Complete-Linkage Clustering. *Algorithmica*, 78(4):1131–1150. https://doi.org/10.1007/s00453-017-0284-6.

Harrower, M. and Brewer, C. A. (2003). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37. https://doi.org/10.1179/000870403235002042.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314. https://doi.org/10.1080/10618600.1996.10474713.

Jacoby, W. G. (2000). Loess:: A Nonparametric, Graphical Tool for Depicting Relationships Between Variables. *Electoral Studies*, 19(4):577–613. https://doi.org/10.1016/S0261-3794(99)00028-1.

Jolliffe, I. T., Allen, O. B., and Christie, B. R. (1989). Comparison of Variety Means Using Cluster Analysis and Dendrograms. *Experimental Agriculture*, 25(2):259–269. https://doi.org/10.1017/S0014479700016768.

Kelleher, C. and Wagener, T. (2011). Ten Guidelines for Effective Data Visualization in Scientific Publications. *Environmental Modelling & Software*, 26(6):822–827. https://doi.org/10.1016/j.envsoft.2010.12.006.

Kettenring, J. R. (2006). The Practice of Cluster Analysis. *Journal of Classification*, 23(1):3–30. https://doi.org/10.1007/s00357-006-0002-6.

Knüsel, L. (2002). On the Reliability of Microsoft Excel XP for Statistical Purposes. *Computational Statistics & Data Analysis*, 39(1):109–110. https://doi.org/10.1016/S0167-9473(02)00035-X.

Kosslyn, S. M. (2006). Chapter 7 - Creating Color, Filling, and Optional Components. In *Graph Design for the Eye and Mind*, pages 157–200. Oxford University Press, New York, New York. https://doi.org/10.1093/acprof:oso/9780195311846.003.0007.

Kosslyn, S. M. and Chabris, C. F. (1992). Minding Information Graphics. *Folio*, 21(2):69–71. http://www.chabris.com/Kosslyn1992b.pdf.

Lewis-Beck, M. S. and Lewis-Beck, C. (2015). *Applied Regression: An Introduction*. Sage Publications Inc, Newbury Park, California, 2 edition.

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-3 (https://CRAN.R-project.org/package=RColorBrewer).

Peng, R. D. (2015). *R Programming for Data Science*. Leanpub, Victoria, British Columbia, Canada. https://www.cs.upc.edu/~robert/teaching/estadistica/rprogramming.pdf.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raković, L., Sakal, M., and Pavlicevic, V. (2014). Spreadsheets - How It Started. *International Scientific Journal of Management Information Systems*, 9(4):9–14. https://www.ef.uns.ac.rs/mis/archive-pdf/2014%20-%20No4/42_Rakovic%20v5.pdf.

Robbins, N. B. (2005). *Creating More Effective Graphs*. Wiley, Hoboken, New Jersey.

Schwartz, M. (2022). *WriteXLS: Cross-Platform Perl Based R Function to Create Excel 2003 (XLS) and Excel 2007 (XLSX) Files*. R package version 6.4.0 (https://CRAN.R-project.org/package=WriteXLS).

Strange, N. (2007). *Smoke & Mirrors: How to Bend Facts and Figures to Your Advantage*. A & C Black Publishers, London, United Kingdom.

Su, Y. (2008). It's Easy to Produce Chartjunk Using Microsoft Excel 2007 but Hard to Make Good Graphs. *Computational Statistics & Data Analysis*, 52(10):4594–4601. https://doi.org/10.1016/j.csda.2008.03.007.

Symanzik, J., Robbins, N. B., and Heiberger, R. M. (2016). Observations on the Type and Quality of Graphs Used in the ASA/NCTM Annual Poster Competition During the Years 2013 to 2016. In *2016 JSM Proceedings*, pages 2517–2531, Alexandria, Virginia. http://www.statlit.org/pdf/2016-Symanzik-Robbins-Heiberger-ASA.pdf.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, Connecticut.

Tufte, E. R. (2006). *Beautiful Evidence*. Graphics Press, Cheshire, Connecticut.

Tukey, J. W. (1972). Some Graphic and Semigraphic Displays. In Bancroft, T. A., editor, *Statistical Papers in Honor of George W. Snedecor*, pages 293–316, Ames, Iowa. Iowa State University Press. https://www.edwardtufte.com/tufte/tukey.

Venables, W. N. and Ripley, B. D. (2000). *S Programming*. Springer, New York, New York. https://doi.org/10.1007/978-0-387-21856-4.

Wainer, H. (1984). How to Display Data Badly. *The American Statistician*, 38(2):137–147. https://doi.org/10.1080/00031305.1984.10483186.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2022). *gplots: Various R Programming Tools for Plotting Data*. R package version 3.1.3 (https://CRAN.R-project.org/package=gplots).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York. https://ggplot2.tidyverse.org.

Wickham, H. (2022). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.3.2 (https://CRAN.R-project.org/package=tidyverse).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686. https://doi.org/10.21105/joss.01686.

Wickham, H. and Bryan, J. (2022). *readxl: Read Excel Files*. R package version 1.4.1 (https://CRAN.R-project.org/package=readxl).

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., and Dunnington, D. (2022a). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.6 (https://CRAN.R-project.org/package=ggplot2).

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2022b). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10 (https://CRAN.R-project.org/package=dplyr).

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., and Wilke, C. O. (2020). colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software*, 96(1):1–49. https://doi.org/10.18637/jss.v096.i01.

APPENDICES

APPENDIX A

List of Sampled Graphs From Utah State University Plan A Master of Science Thesis Reports

In this appendix, a lookup table identifying the 90 sampled graphs from Utah State University (USU) Plan A Master of Science (MS) thesis reports can be found in Table A.1. The graph names used to identify the graphs follow the naming procedure described in Section 3.3. The table is organized by the *Graph Name* column, where it is first sorted by the four-digit five-year interval and then by the three to four-digit report number, both in ascending order.

Table A.1: List of all 90 sampled graphs from Utah State University Plan A Master of Science thesis reports.

Graph Name	**Year**	**Report Number**	**Page Number**	**Figure Number**	**Permutation**
1930-1357-78	1934	1357	78	9	5
1930-1570-49	1930	1570	49	N/A	1
1930-3946-30	1934	3946	30	4	3
1930-3994-20	1932	3994	20	5	4
1930-3997-39	1931	3997	39	5	2
1935-323-20	1938	323	20	4	2
1935-1615-27	1935	1615	27	3	1
1935-1928-13	1938	1928	13	3	5
1935-3671-61	1935	3671	61	1	4
1935-3948-07	1937	3948	07	2	3
1940-1795-48	1940	1795	48	5	3
1940-1883-28	1941	1883	28	3	5
1940-3925-20	1942	3925	20	1	1
1940-4708-75	1940	4708	75	1	2
1940-4709-17	1940	4709	17	29	4

Table A.1: List of all 90 sampled graphs from Utah State University Plan A Master of Science thesis reports. (Continued)

Graph Name	Year	Report Number	Page Number	Figure Number	Permutation
1945-1778-45	1949	1778	45	4	2
1945-1798-43	1947	1798	43	1	5
1945-1922-34	1947	1922	34	7	4
1945-4740-74	1948	4740	74	32	1
1945-5741-67	1947	5741	67	4	3
1950-1836-70	1952	1836	70	8	5
1950-3610-54	1954	3610	54	3	4
1950-5545-35	1953	5545	35	5	3
1950-6249-52	1952	6249	52	11	1
1950-7868-95	1953	7868	95	18	2
1955-2689-22	1956	2689	22	1	1
1955-2697-28	1956	2697	28	1	2
1955-2755-27	1956	2755	27	2	4
1955-3641-56	1957	3641	56	9	5
1955-3731-59	1959	3731	59	4	3
1960-2761-26	1961	2761	26	1	4
1960-2882-74	1964	2882	74	1	1
1960-3653-29	1961	3653	29	3	5
1960-4380-32	1960	4380	32	3	2
1960-4947-27	1963	4947	27	2	3
1965-2892-74	1967	2892	74	1	1
1965-2914-28	1966	2914	28	3	2
1965-2951-49	1968	2951	49	6	4
1965-2974-59	1967	2974	59	1	3
1965-5583-36	1967	5583	36	9	5

Table A.1: List of all 90 sampled graphs from Utah State University Plan A Master of Science thesis reports. (Continued)

Graph Name	Year	Report Number	Page Number	Figure Number	Permutation
1970-1599-52	1970	1599	52	7	4
1970-3509-50	1971	3509	50	7	5
1970-5082-55	1971	5082	55	22	1
1970-5721-31	1974	5721	31	1	3
1970-6852-47	1971	6852	47	3	2
1975-3215-39	1976	3215	39	4	2
1975-3235-68	1976	3235	68	8	4
1975-3243-36	1976	3243	36	7	5
1975-3381-37	1978	3381	37	4a-b	3
1975-6328-32	1978	6328	32	5	1
1980-3343-68	1982	3343	68	14	4
1980-3411-38	1984	3411	38	3	3
1980-4200-30	1980	4200	30	3	2
1980-5293-37	1982	5293	37	4	5
1980-7177-75	1984	7177	75	17	1
1985-4083-37	1988	4083	37	1	3
1985-4090-47	1987	4090	47	4	4
1985-4093-39	1986	4093	39	1	2
1985-6676-68	1985	6676	68	14	1
1985-7263-22	1986	7263	22	2	5
1990-4194-47	1993	4194	47	2	1
1990-4655-40	1994	4655	40	2	3
1990-5400-65	1993	5400	65	11	2
1990-6434-54	1992	6434	54	6	5
1990-6517-23	1994	6517	23	1	4

Table A.1: List of all 90 sampled graphs from Utah State University Plan A Master of Science thesis reports. (Continued)

Graph Name	Year	Report Number	Page Number	Figure Number	Permutation
1995-4602-39	1996	4602	39	1	4
1995-6580-34	1998	6580	34	1-1	1
1995-6702-32	1995	6702	32	7	2
1995-6747-38	1996	6747	38	11	5
1995-7424-32	1998	7424	32	1	3
2000-5520-78	2004	5520	78	7	4
2000-6598-86	2002	6598	86	39	5
2000-6599-39	2003	6599	39	6	2
2000-6708-42	2001	6708	42	2-5	3
2000-7269-147	2003	7269	147	61	1
2005-120-48	2008	120	48	17	3
2005-450-37	2009	450	37	37	2
2005-469-57	2009	469	57	29	1
2005-483-129	2009	483	129	3.85	4
2005-2592-58	2007	2592	58	2	5
2010-696-60	2010	696	60	3-5	1
2010-771-58	2010	771	58	12	2
2010-947-63	2010	947	63	41	4
2010-964-72	2011	964	72	41	3
2010-1942-76	2013	1942	76	3.2	5
2015-4249-24	2015	4249	24	4	4
2015-4455-118	2015	4455	118	A.1	3
2015-7469-67	2019	7469	67	A2	5
2015-7586-65	2019	7586	65	3.3	2
2015-7601-72	2019	7601	72	4.5	1

APPENDIX B

Standards from Graph Quality Literature Resources

In this appendix, the good graph standards from the four graph quality literature resources discussed in Section 3.1 are presented. Appendix B.1 displays the principles and features from Gordon and Finch (2015), Appendix B.2 displays the principles from Robbins (2005), Appendix B.3 displays the rules from Wainer (1984), and Appendix B.4 displays the guidelines from Kelleher and Wagener (2011).

B.1 Principles and Features from Gordon and Finch (2015)

In this appendix section, the five principles of graphical excellence from Gordon and Finch (2015) are presented in Table B.1. Additionally, the 60 coded features they developed are shown in Table B.2. The features are grouped according to the five principles of graphical excellence and additional features that Gordon and Finch (2015) used to assess the sampled graphs in their study.

Table B.1: List of the five principles of graphical excellence determined by Gordon and Finch (2015).

Number	**Principle**
1	Show the data clearly
2	Use simplicity in design
3	Use good alignment on a common scale for quantities to be compared
4	Keep the visual encoding transparent
5	Use graphical forms consistent with Principles 1 to 4

Table B.2: List of the 60 coded features developed by Gordon and Finch (2015), categorized by the five principles of graphical excellence (see Table B.1) and other additional features defined by Gordon and Finch (2015).

	Number	**Feature**
Principle 1		
	1	Includes a caption
	2	Caption is adequate
	3	Suitable axis labels
	4	Variable labelled rather than estimate (on axis)
	5	Legend
	6	Legend could be replaced by direct text
	7	Boxed legend
	8	Graph has detection problems
	9	Contains undefined graphical elements
	10	Contains undefined abbreviations
	11	Unused white space that could be used for data
	12	Incorrect scaling
Principle 2		
	13	Number of dimensions (2d or 3d)
	14	Two scales on one axis
Principle 4		
	15	Uses colour
	16	Colour use redundant
	17	Uses cross hatching including stripes
	18	Ordering would improve transparency
	19	Does the graph stand alone?
Principle 3		
	20	Elements to be compared aligned on common scale
	21	Non horizontal tick mark labels
	22	Gridlines, including reference line(s)
	23	Gridlines too heavy
	24	Additional gridlines could be used
	25	Transposition would improve the graph

Table B.2: List of the 60 coded features developed by Gordon and Finch (2015), categorized by the five principles of graphical excellence (see Table B.1) and other additional features defined by Gordon and Finch (2015). (Continued)

	Number	**Feature**
Principle 5		
	26	Dotplot
	27	Barchart excluding stacked
	28	Horizontal bars on barchart
	29	Dotchart
	30	Dotchart would be better
	31	Boxplot
	32	Scatterplot
	33	Plot of means or proportions
	34	Time series
	35	Histogram
	36	Specify any other standard form
	37	Non standard form
	38	Estimates represented as bars
	39	Estimates represented as bars with error bars
	40	Number of panels
	41	Panels could be used
Other features		
	42	Raw data
	43	Summary statistics or estimates (include percentages)
	44	Estimates and uncertainty
	45	Fitted line/model
	46	A theoretical model
	47	Number of variables
	48	Type of “error bars” or bounds (choose one: SD, SE, CI, Not specified)
	49	P-values
	50	Stars on graphs
	51	Relative p-values
	52	Exact p-values
	53	Width in cm
	54	Height in cm

Table B.2: List of the 60 coded features developed by Gordon and Finch (2015), categorized by the five principles of graphical excellence (see Table B.1) and other additional features defined by Gordon and Finch (2015). (Continued)

	Number	**Feature**
	55	Number of distinct numerical values represented
	56	Number of points labelled with values
	57	Font easy to read
	58	Number of typos found
	59	Obvious statistical problem
	60	Overall rating

B.2 Principles from Robbins (2005)

In this appendix section, the general principles to creating good quality graphs from Robbins (2005) are presented in Table B.3. The principles are grouped according to the two main principle categories emphasized by Robbins (2005), '*visual clarity*' and '*clear understanding*', as well as the general strategies and checklist of possible graph defects that Robbins (2005) developed.

Table B.3: List of the general principles to creating good quality graphs developed by Robbins (2005), categorized according to the two main principle categories, general strategies, and checklist of possible graph defects defined by Robbins (2005).

	Principle
Visual clarity category	• Make the data stand out. Avoid superfluity • Use visually prominent graphical elements to show the data • Overlapping plotting symbols must be visually distinguishable • Superposed data sets must be readily visually assembled • Do not clutter the interior of the scale-line rectangle • Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward • Do not overdo the number of tick marks • Do not overdo the number of tick mark labels • Deemphasize grid lines and distinguish grid lines from data • Avoid putting notes and keys inside the scale-line rectangle. Put a key outside and put notes in the caption or in the text • Visual clarity must be preserved under reduction and reproduction • Proofread graphs • The chart or graph must be consistent with the text
Clear understanding category	• Draw the data to scale • Do not show changes in one dimension by area or volume • Use a common baseline wherever possible

Table B.3: List of the general principles to creating good quality graphs developed by Robbins (2005), categorized according to the two main principle categories, general strategies, and checklist of possible graph defects defined by Robbins (2005). (Continued)

	Principle
	• Label data sets directly when it doesn't clutter the graph • Don't require the reader to make calculations that a computer can make more easily • Plot the variable of interest. If interested in improvement, plot improvement rather than before and after • Strive for clarity • Groups of charts need consistency in order, color, and other graphical elements • Choose the principle least likely to mislead if more than one applies and they conflict with one another
General strategies	• A large amount of quantitative information can be packed into a small region • Graphing data should be an iterative experimental process • Graph data two or more times when needed • Many useful displays require careful, detailed study
Checklist of possible graph defects	• Do the data stand out? Are there superfluous elements? • Are all graphical elements visually prominent? • Are overlapping plotting symbols visually distinguishable? • Can superposed data sets be readily visually assembled? • Is the interior of the scale-line rectangle cluttered? • Do data labels interfere with the quantitative data or clutter the graph? • Is the data rectangle within the scale-line rectangle? • Do tick marks interfere with the data? • Do tick mark labels interfere with the data?

Table B.3: List of the general principles to creating good quality graphs developed by Robbins (2005), categorized according to the two main principle categories, general strategies, and checklist of possible graph defects defined by Robbins (2005). (Continued)

	Principle
	• Are axis labels legible? • Are there too many tick marks? • Are there too many tick mark labels? • Do the grid lines interfere with the data? • Are there notes or keys inside the scale-line rectangle? • Will visual clarity be preserved under reduction and reproduction? • Are the data drawn to scale? • Is there an informative title? • Is area or volume used to show changes in one dimension? • Are there too many dimensions in the graph (more than one in the data)? • Are common baselines used wherever possible? • Are all labels associated with the correct graphical elements? • Is the reader required to make calculations? • Are groups of charts drawn consistently? • Is zero included for all bar graphs? • Are there any unnecessary scale breaks? • Is there a forceful indication of a scale break? • Are there numerical values on two sides of a scale break that are connected? • Does the aspect ratio allow the reader to see variations in the data? • Are scales included for all axes? • Are the scales labeled? • Are tick marks at sensible values? • Do the axes increase in the conventional direction? • Does the data rectangle fill as much of the scale-line rectangle as possible? • Are uneven time intervals handled correctly? • Are the scales appropriate when different panels are compared?

B.3 Rules from Wainer (1984)

In this appendix section, the 12 bad graph rules to avoid when creating graphs from Wainer (1984) are presented in Table B.4. The principles are categorized according to the three components of the overall aim of good graphics defined by Wainer (1984), '*showing data*', '*showing data accurately*', and '*showing data clearly*'.

Table B.4: List of the 12 bad graph rules to avoid when creating graphs developed by Wainer (1984), categorized according to the three components of the overall aim of good graphics defined by Wainer (1984).

	Number	**Rule**
Showing data	1	Show as few data as possible (minimize the data density)
	2	Hide what data you do show (minimize the data-ink ratio)
Showing data accurately	3	Ignore the visual metaphor altogether
	4	Only order matters
	5	Graph data out of context
Showing data clearly	6	Change scales in mid-axis
	7	Emphasize the trivial (ignore the important)
	8	Jiggle the baseline
	9	Austria first
	10	Label (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously
	11	More is murkier: (a) more decimal places and (b) more dimensions
	12	If it has been done well in the past, think of another way to do it

B.4 Guidelines from Kelleher and Wagener (2011)

In this appendix section, the ten guidelines to adequately communicate and represent data in graphs from Kelleher and Wagener (2011) are presented in Table B.5.

Table B.5: List of the ten guidelines to adequately communicate and represent data in graphs defined by Kelleher and Wagener (2011).

Number	Guideline
1	Create the simplest graph that conveys the information you want to convey
2	Consider the type of encoding object and attribute used to create a plot
3	Focus on visualizing patterns or on visualizing details, depending on the purpose of the plot
4	Select meaningful axis ranges
5	Data transformations and carefully chosen graph aspect ratios can be used to emphasize rates of change for time-series data
6	Plot overlapping points in a way that density differences become apparent in scatter plots
7	Use lines when connecting sequential data in time-series plots
8	Aggregate larger datasets in meaningful ways
9	Keep axis ranges as similar as possible to compare variables
10	Select an appropriate color scheme based on the type of data

APPENDIX C

Scoring Results for Graphs Within the Second, Third, Fourth, and Fifth Five-Year Interval Permutation Sequences

In this appendix, Figures C.1 - C.4 display the four Excel files containing the scoring results for the sampled Utah State University (USU) Plan A Master of Science (MS) theses graphs within the second, third, fourth and fifth five-year interval permutation sequences, respectively. The scoring results for the sampled graphs within the first five-year interval permutation sequence can be found in Figure 5.1. The five-year interval permutation sequences were created as discussed in Section 2.2.5.

				Labeling					Clear Understanding								Meaningful				Scaling and Gridlines											Overall Rating
Year Interval	Graph Name	Type of Graph	How Created	1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	5	6	7	Total 1	Total 2	Total 3	Total 4	Percent Correct	(1 = Worst, 5 = Best)
1985	1985-4093-39	Other	Hand-drawn/Typewriter	0	1	0	0	1	0	0	0	0		1	0	1	0	0	1	0	0	0	0	1			1	0.1	0.071429	0.0625	0.1	33.39	2
1970	1970-6852-47	Bar Graph (grouped, side-by-side)	Hand-drawn/Typewriter	0	1	0	0	1	1	0	0	1	1	1	1	1	0	0	1	1	0		0				1	0.1	0.1875	0.125	0.08333	49.58	1
2000	2000-6599-39	Scatterplot (multiple graphs) (+ fitted line)	Software	1	1	0	0	1	1	0	0	1	0	1	0	1	1	1	1	0	1	1	1	0			1	0.15	0.125	0.1875	0.2	66.25	3
1990	1990-5400-65	Line Graph (+ intervals)	Software	1	1	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1			1	0.1	0.25	0.1875	0.25	78.75	4
1960	1960-4380-32	Line Graph	Hand-drawn/Typewriter	1	1	1	1	1	0	1	0	1	1	1			1	1	1	1	1	0	0				1	0.25	0.166667	0.25	0.125	79.17	3
1945	1945-1778-45	Bar Graph (grouped, overlayed)	Typewriter	1	0	1	0	1	0	0	0	1	1	0	1	1	0	0	1	1	1	1	0	1			1	0.15	0.125	0.125	0.2	60	3
1955	1955-2697-28	Line Graph	Hand-drawn	1	0	1	0	1	0	1	0	1		1	1	1	1	1	1	1	1	0	0	1	1		1	0.15	0.178571	0.25	0.16667	74.52	3
1940	1940-4708-75	Bar Graph (grouped, stacked)	Hand-drawn/Typewriter	1	1	0	0	0	1	0	0	1	0	1	1	0	0	0	1	1	0	0	0	0			1	0.1	0.125	0.125	0.05	40	2
1965	1965-2914-28	Dot Plot (+ line)	Hand-drawn/Typewriter	1	1	0	0	1	0	1	0	1		1	0	1	1	1	1	0	1	0	0	1			1	0.15	0.142857	0.1875	0.15	63.04	2
1930	1930-3997-39	Line Graph	Hand-drawn/Typewriter	1	1	1	0	1	0	0	0	1		0	0	1	1	1	1	0	1	0	0	1			0	0.2	0.071429	0.1875	0.1	55.89	3
1950	1950-7868-95	Line Graph	Hand-drawn/Typewriter	1	1	1	0	1	0	1	0	1		1			1	1	1	1	1	1	1				1	0.2	0.15	0.25	0.25	85	4
2010	2010-771-58	Line Graph (+ fitted line)	Software	1	1	1	1	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	1			0	0.2	0.1875	0.1875	0.15	72.5	4
1935	1935-323-20	Bar Graph	Hand-drawn/Typewriter	1	1	1	1	1	1	1	0	1	1	1		1	0	0	1	1	1	0	0		0		1	0.25	0.214286	0.125	0.1	68.93	2
1975	1975-3215-39	Scatterplot	Typewriter	1	1	1	1	1	0	0	0	1	1	1	0	1	1	0	1	1	1	1	0	1			0	0.25	0.125	0.1875	0.15	71.25	3
2005	2005-450-37	Line Graph	Software	0	1	0	0	1	0	0	0	1		1			1	1	1	1	1	1	0				1	0.1	0.1	0.25	0.1875	63.75	3
1995	1995-6702-32	Histogram (multiple graphs)	Software	0	1	1	1	0	1	0	0	0	0	1		1	0	0	1	0	1	1	0	0			1	0.15	0.107143	0.0625	0.15	46.96	2
1980	1980-4200-30	Scatterplot (+ fitted line)	Software	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1			1	0.25	0.21875	0.25	0.15	86.88	4
2015	2015-7586-65	Line Graph (+ intervals)	Software	0	1	0	0	0	0	1	0	1		0	0	1	1	1	1	0	1	1	1	1	1	1	1	0.05	0.107143	0.1875	0.25	59.46	3

Criteria:			
Graph has adequate labeling	Graph has clear understanding	Graphical elements add meaning to information displayed	Graph has adequate scaling and gridlines
1. Adequate use of caption and legend	1. Data is visually clear	1. Show as much data as possible (maximize data/ink ratio)	1. One scale is included for every axis
2. Graph labels are legible	2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph	2. Simplest graph of the chosen type is used to convey the necessary information	2. No changes or breaks in scale
3. Graph labels are sufficient	3. Data is accurately displayed in the most effective and appropriate way	3. Minimal dimensions used	3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. All graphical elements are clearly and correctly defined	4. Data is consistent with labeling/text	4. Color is used appropriately and effectively	4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Labeling does not interfere with or clutter the graph	5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)		5. Adequate use of gridlines, do not hide or interfere with data
	6. Clarity of data will be maintained through reproduction/reduction of graph		6. Data transformations are adequately used to display necessary data information
	7. Overlapping data points or superposed data sets are distinguishable		7. Size and aspect ratio display data adequately
	8. A common baseline is used wherever possible		

Fig. C.1: Excel file containing the raw graph quality scores for the second five-year interval permutation sequence.

				Labeling					Clear Understanding								Meaningful				Scaling and Gridlines											Overall Rating
Year Interval	Graph Name	Type of Graph	How Created	1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	5	6	7	Total 1	Total 2	Total 3	Total 4	Percent Correct	(1 = Worst, 5 = Best)
1955	1955-3731-59	Scatterplot (+ fitted line)	Hand-drawn/Typewriter	1	1	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0	1	1	0	1			1	0.15	0.09375	0.0625	0.2	50.63	3
1950	1950-5545-35	Dot Plot (+ line)	Hand-drawn/Typewriter	1	1	1	0	1	0	0	0	1	1	0	0	1	1	1	1	0	1	1	0	1	1		1	0.2	0.09375	0.1875	0.20833	68.96	3
1970	1970-5721-31	Bar Graph (grouped, side-by-side)	Hand-drawn/Typewriter	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	0	0	1			1	0.25	0.21875	0.125	0.15	74.38	4
1985	1985-4083-37	Line Graph	Software	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1			1	0.2	0.21875	0.25	0.15	81.88	3
1935	1935-3948-07	Line Graph	Hand-drawn/Typewriter	0	1	0	0	1	0	0	0	1		0	1	1	1	1	1	1	1	1	1	1			1	0.1	0.107143	0.25	0.25	70.71	2
1980	1980-3411-38	Line Graph	Software	1	1	1	1	1	0	0	0	1		1	0	1	1	1	1	0	1	1	0	1			1	0.25	0.107143	0.1875	0.2	74.46	4
1960	1960-4947-27	Line Graph	Hand-drawn/Typewriter	1	1	1	1	1	0	0	0	1		1	0	1	1	1	1	0	1	1	0	1			1	0.25	0.107143	0.1875	0.2	74.46	4
1995	1995-7424-32	Bar Graph (grouped, side-by-side)	Software	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1			1	0.1	0.25	0.1875	0.25	78.75	4
2000	2000-6708-42	Bar Graph (grouped, side-by-side)	Software	0	1	1	0	0	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1			1	0.1	0.21875	0.125	0.25	69.38	4
1945	1945-5741-67	Dot Plot (multiple graphs) (+ line)	Hand-drawn/Typewriter	1	1	1	0	1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	0	1	1		1	0.2	0.0625	0.125	0.20833	59.58	3
1940	1940-1795-48	Bar Graph (grouped, side-by-side)	Hand-drawn/Typewriter	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	0	1	1	0	1	0		1	0.25	0.15625	0.125	0.16667	69.79	4
1930	1930-3946-30	Line Graph	Hand-drawn/Typewriter	1	1	1	1	1	0	1	0	1		1		1	1	1	1	1	1	1	0				1	0.25	0.166667	0.25	0.1875	85.42	4
2005	2005-120-48	Scatterplot (+ fitted line)	Software	0	1	0	0	1	0	0	0	1	1	0	0	1	1	0	1	0	1	1	0	1	1		1	0.1	0.09375	0.125	0.20833	52.71	3
1990	1990-4655-40	Dot Plot	Software	1	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1			1	0.25	0.15625	0.1875	0.25	84.38	4
2015	2015-4455-118	Surface Plot	Software	0	1	0	0	1	0	0	0	0		0	0	1	0	0	0	1	1	1	1		1		1	0.1	0.035714	0.0625	0.25	44.82	2
1965	1965-2974-59	Scatterplot (+ fitted line)	Hand-drawn/Typewriter	0	1	1	0	0	0	1	0	1	1	1		1	1	1	1	1	1	1	0				1	0.1	0.178571	0.25	0.1875	71.61	4
1975	1975-3381-37	Bar Graph (single, side-by-side) (multiple graphs)	Typewriter	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1			1	0.25	0.21875	0.1875	0.15	80.63	3
2010	2010-964-72	Line Graph (multiple graphs) (+ intervals)	Software	1	1	0	0	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1			1	0.15	0.25	0.25	0.25	90	4

Criteria:			
Graph has adequate labeling	**Graph has clear understanding**	**Graphical elements add meaning to information displayed**	**Graph has adequate scaling and gridlines**
1. Adequate use of caption and legend	1. Data is visually clear	1. Show as much data as possible (maximize data/ink ratio)	1. One scale is included for every axis
2. Graph labels are legible	2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph	2. Simplest graph of the chosen type is used to convey the necessary information	2. No changes or breaks in scale
3. Graph labels are sufficient	3. Data is accurately displayed in the most effective and appropriate way	3. Minimal dimensions used	3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. All graphical elements are clearly and correctly defined	4. Data is consistent with labeling/text	4. Color is used appropriately and effectively	4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Labeling does not interfere with or clutter the graph	5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)		5. Adequate use of gridlines, do not hide or interfere with data
	6. Clarity of data will be maintained through reproduction/reduction of graph		6. Data transformations are adequately used to display necessary data information
	7. Overlapping data points or superposed data sets are distinguishable		7. Size and aspect ratio display data adequately
	8. A common baseline is used wherever possible		

Fig. C.2: Excel file containing the raw graph quality scores for the third five-year interval permutation sequence.

				Labeling					Clear Understanding								Meaningful				Scaling and Gridlines											Overall Rating
Year Interval	Graph Name	Type of Graph	How Created	1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	5	6	7	Total 1	Total 2	Total 3	Total 4	Percent Correct	(1 = Worst, 5 = Best)
1980	1980-3343-68	Bar Graph (grouped, stacked) (multiple graphs)	Software	1	1	0	0	0	0	0	0	1	0	1	1	0	0	0	1	1	0	0	1	1			1	0.1	0.09375	0.125	0.15	46.88	2
1930	1930-3994-20	Bar Graph (grouped, side-by-side)	Hand-drawn/Typewriter	1	1	1	0	0	1	0	0	1	1	1	1	1	0	1	1	1	1	0	0	1			1	0.15	0.1875	0.1875	0.15	67.5	2
1990	1990-6517-23	Line Graph (multiple graphs) (+ intervals)	Software	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	0			1	0.25	0.125	0.25	0.2	82.5	3
1950	1950-3610-54	Bar Graph (single, side-by-side) (multiple graphs)	Hand-drawn/Typewriter	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	1			1	0.25	0.21875	0.125	0.1	69.38	3
1945	1945-1922-34	Histogram	Hand-drawn/Typewriter	1	1	1	1	1	1	1	1	1	1	0		1	0	1	1	1	1	1	0		0		1	0.25	0.214286	0.1875	0.15	80.18	4
1940	1940-4709-17	Bar Graph (single, side-by-side)	Other	1	1	1	1	1	1	0	0	1	1	1		0	0	0	1	1	0	0	0				1	0.25	0.142857	0.125	0.0625	58.04	2
1975	1975-3235-68	Scatterplot (+ fitted line)	Hand-drawn/Typewriter	0	1	1	0	0	0	1	0	0	1	1		1	0	1	1	1	1	0	0			1	1	0.1	0.142857	0.1875	0.15	58.04	3
1935	1935-3671-61	Bar Graph (single, side-by-side)	Hand-drawn/Typewriter	1	1	1	1	1	1	1	1	1	1	1		1	0	1	1	1	1	0	0		1		1	0.25	0.25	0.1875	0.15	83.75	4
1955	1955-2755-27	Bar Graph (grouped, stacked)	Hand-drawn/Typewriter	1	1	1	1	1	0	0	0	1	1	1	1	0	0	0	1	0	1	0	0	1			1	0.25	0.125	0.0625	0.15	58.75	3
1970	1970-1599-52	Line Graph	Hand-drawn/Typewriter	1	1	1	0	1	0	0	0	1	0	1	0	1	1	0	1	0	1	1	1	1			1	0.2	0.09375	0.125	0.25	66.88	3
1960	1960-2761-26	Line Graph	Hand-drawn/Typewriter	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	0	1			1	0.2	0.1875	0.25	0.15	78.75	3
2015	2015-4249-24	Line Graph	Software	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1		1	0.2	0.21875	0.1875	0.20833	81.46	4
2005	2005-483-129	Bar Graph (grouped, side-by-side, 3D)	Software	1	1	1	1	1	0	0	0	1	1	1	1	1	0	0	0	0	1	1	0	1	1		1	0.25	0.15625	0	0.20833	61.46	3
1985	1985-4090-47	Bar Graph (grouped, side-by-side)	Hand-drawn/Typewriter	1	0	1	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1			1	0.2	0.1875	0.1875	0.25	82.5	4
2000	2000-5520-78	Line Graph (+ intervals)	Software	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1			1	0.15	0.1875	0.25	0.25	83.75	4
2010	2010-947-63	Line Graph	Software	1	1	1	1	1	0	0	0	1		1	0	1	1	1	1	0	1	1	0	1	1		1	0.25	0.107143	0.1875	0.20833	75.3	3
1965	1965-2951-49	Scatterplot (multiple graphs)	Hand-drawn/Typewriter	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1			1	0.25	0.25	0.25	0.2	95	4
1995	1995-4602-39	Bar Graph (single, side-by-side, 3D)	Software	1	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	1	1	1	1		1	0.25	0.125	0	0.25	62.5	3

Criteria:			
Graph has adequate labeling	**Graph has clear understanding**	**Graphical elements add meaning to information displayed**	**Graph has adequate scaling and gridlines**
1. Adequate use of caption and legend	1. Data is visually clear	1. Show as much data as possible (maximize data/ink ratio)	1. One scale is included for every axis
2. Graph labels are legible	2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph	2. Simplest graph of the chosen type is used to convey the necessary information	2. No changes or breaks in scale
3. Graph labels are sufficient	3. Data is accurately displayed in the most effective and appropriate way	3. Minimal dimensions used	3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. All graphical elements are clearly and correctly defined	4. Data is consistent with labeling/text	4. Color is used appropriately and effectively	4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Labeling does not interfere with or clutter the graph	5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)		5. Adequate use of gridlines, do not hide or interfere with data
	6. Clarity of data will be maintained through reproduction/reduction of graph		6. Data transformations are adequately used to display necessary data information
	7. Overlapping data points or superposed data sets are distinguishable		7. Size and aspect ratio display data adequately
	8. A common baseline is used wherever possible		

Fig. C.3: Excel file containing the raw graph quality scores for the fourth five-year interval permutation sequence.

				Labeling					Clear Understanding								Meaningful				Scaling and Gridlines											Overall Rating
Year Interval	Graph Name	Type of Graph	How Created	1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	5	6	7	Total 1	Total 2	Total 3	Total 4	Percent Correct	(1 = Worst, 5 = Best)
1935	1935-1928-13	Line Graph	Hand-drawn	0	0	0	0	1	1	0	1	1	1	0		1	1	1	1	1	1	1	0		1		1	0.05	0.178571	0.25	0.2	67.86	3
2005	2005-2592-58	Bar Graph (single, side-by-side)	Software	1	1	1	1	1	1	0	0	1	1	1		1	0	1	1	1	1	1	0		1		1	0.25	0.178571	0.1875	0.2	81.61	4
1985	1985-7263-22	Scatterplot (+ fitted line)	Software	0	1	1	0	1	0	0	0	1	1	1	0	1	1	1	1	0	1	1	0	1			1	0.15	0.125	0.1875	0.2	66.25	4
1970	1970-3509-50	Line Graph	Hand-drawn/Typewriter	1	1	1	1	1	0	1	0	1		1	1	1	1	1	1	0	1	1	0	1			1	0.25	0.178571	0.1875	0.2	81.61	3
2015	2015-7469-67	Scatterplot (multiple graphs)	Software	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	0	1	1	1	1		1	1	0.25	0.15625	0.1875	0.25	84.38	4
1990	1990-6434-54	Bar Graph (grouped, side-by-side)	Software	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0		1	0.25	0.25	0.1875	0.20833	89.58	4
1930	1930-1357-78	Bar Graph (grouped, stacked)	Hand-drawn/Typewriter	1	1	1	0	1	1	0	0	1	1	1	1	0	0	0	1	1	1	1	0	1	1		1	0.2	0.15625	0.125	0.20833	68.96	3
1965	1965-5583-36	Scatterplot (+ fitted line) (multiple graphs)	Hand-drawn/Typewriter	1	1	1	0	1	0	0	0	1	1	0	0	1	0	1	1	0	1	1	0	1			1	0.2	0.09375	0.125	0.2	61.88	3
1960	1960-3653-29	Bar Graph (single, side-by-side)	Hand-drawn/Typewriter	1	1	1	1	1	1	1	1	1	1	1		1	0	1	1	1	1	1	0		1		1	0.25	0.25	0.1875	0.2	88.75	4
1980	1980-5293-37	Scatterplot (+ fitted line)	Software	1	1	1	1	1	0	1	1	1	1	1		1	1	1	1	1	1	1	0				1	0.25	0.214286	0.25	0.1875	90.18	4
1975	1975-3243-36	Bar Graph (grouped, side-by-side)	Software	1	1	1	1	1	1	0	0	1	1	0	1	1	0	0	1	0	1	1	0	1			1	0.25	0.15625	0.0625	0.2	66.88	4
2010	2010-1942-76	Line Graph (multiple graphs)	Software	1	1	1	1	1	0	1	0	1	1	1	0	1	1	1	1	0	1	1	0	1			1	0.25	0.15625	0.1875	0.2	79.38	3
1945	1945-1798-43	Bar Graph (single, side-by-side)	Hand-drawn/Typewriter	1	1	1	1	1	1	1	1	1	1	1		1	0	1	1	1	1	1	1		0		1	0.25	0.25	0.1875	0.2	88.75	4
1940	1940-1883-28	Histogram (multiple graphs)	Hand-drawn/Typewriter	1	0	1	1	0	1	1	0	1	0	1		1	0	1	1	1	1	1	0	0			1	0.15	0.178571	0.1875	0.15	66.61	3
1950	1950-1836-70	Bar Graph (grouped, side-by-side)	Hand-drawn/Typewriter	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	1	1	0	0	1			1	0	0.125	0.125	0.15	40	2
1995	1995-6747-38	Dot Plot (multiple graphs)	Software	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1			1	0.25	0.1875	0.25	0.25	93.75	4
1955	1955-3641-56	Scatterplot (+ fitted line)	Hand-drawn/Typewriter	0	1	0	0	1	0	0	0	0	1	1	0	1	1	0	1	1	0	1	0		0		1	0.1	0.09375	0.1875	0.1	48.13	3
2000	2000-6598-86	Histogram (multiple graphs)	Software	1	1	1	0	1	1	1	0	0	1	0		1	0	1	1	0	1	1	1	1			1	0.2	0.142857	0.125	0.25	71.79	3

Criteria:			
Graph has adequate labeling	**Graph has clear understanding**	**Graphical elements add meaning to information displayed**	**Graph has adequate scaling and gridlines**
1. Adequate use of caption and legend	1. Data is visually clear	1. Show as much data as possible (maximize data/ink ratio)	1. One scale is included for every axis
2. Graph labels are legible	2. Key details/patterns of the data can be interpreted from the graph; clear purpose of the graph	2. Simplest graph of the chosen type is used to convey the necessary information	2. No changes or breaks in scale
3. Graph labels are sufficient	3. Data is accurately displayed in the most effective and appropriate way	3. Minimal dimensions used	3. Axis ranges and tick mark values are appropriate and meaningful for the type of graph used
4. All graphical elements are clearly and correctly defined	4. Data is consistent with labeling/text	4. Color is used appropriately and effectively	4. Axis ranges are as similar as possible when comparing variables across graphs or within a single graph
5. Labeling does not interfere with or clutter the graph	5. Groups of graphs or multiple components in a single graph are consistent with one another (ordering, coloring, etc)		5. Adequate use of gridlines, do not hide or interfere with data
	6. Clarity of data will be maintained through reproduction/reduction of graph		6. Data transformations are adequately used to display necessary data information
	7. Overlapping data points or superposed data sets are distinguishable		7. Size and aspect ratio display data adequately
	8. A common baseline is used wherever possible		

Fig. C.4: Excel file containing the raw graph quality scores for the fifth five-year interval permutation sequence.

APPENDIX D

Graph Note Counts

This appendix contains a count summary of the different notes used to identify unique traits within the sampled graphs, as can be seen in Table D.1. These graph notes are introduced and explained in Section 3.3. In Section 5.2.1.1, count metrics of the graph types observed within the sampled graphs from Utah State University (USU) Plan A Master of Science (MS) thesis reports were analyzed and assessed, however, the specific notes used to describe the graphs within these graph types were not discussed. Thus, Table D.1 displays the counts of each of the notes recorded for the sampled graphs, grouped by the six observed graph types. The overall counts of the graph types shown in Table D.1 match the graph type counts shown in Figure 5.2. As a reminder from Section 3.3, if relevant, the sampled graphs could be assigned multiple graph notes to describe their characteristics. Bar graphs included either the note ‘single’ or ‘grouped’ to describe the structure of the bars, and further used notes such as ‘side-by-side’, ‘stacked’, or ‘overlayed’ to describe the layout of the bars. As can be seen in Table D.1, the two graphs within the other graph type were not assigned any graph notes.

Table D.1: Graph note counts according to the different graph types.

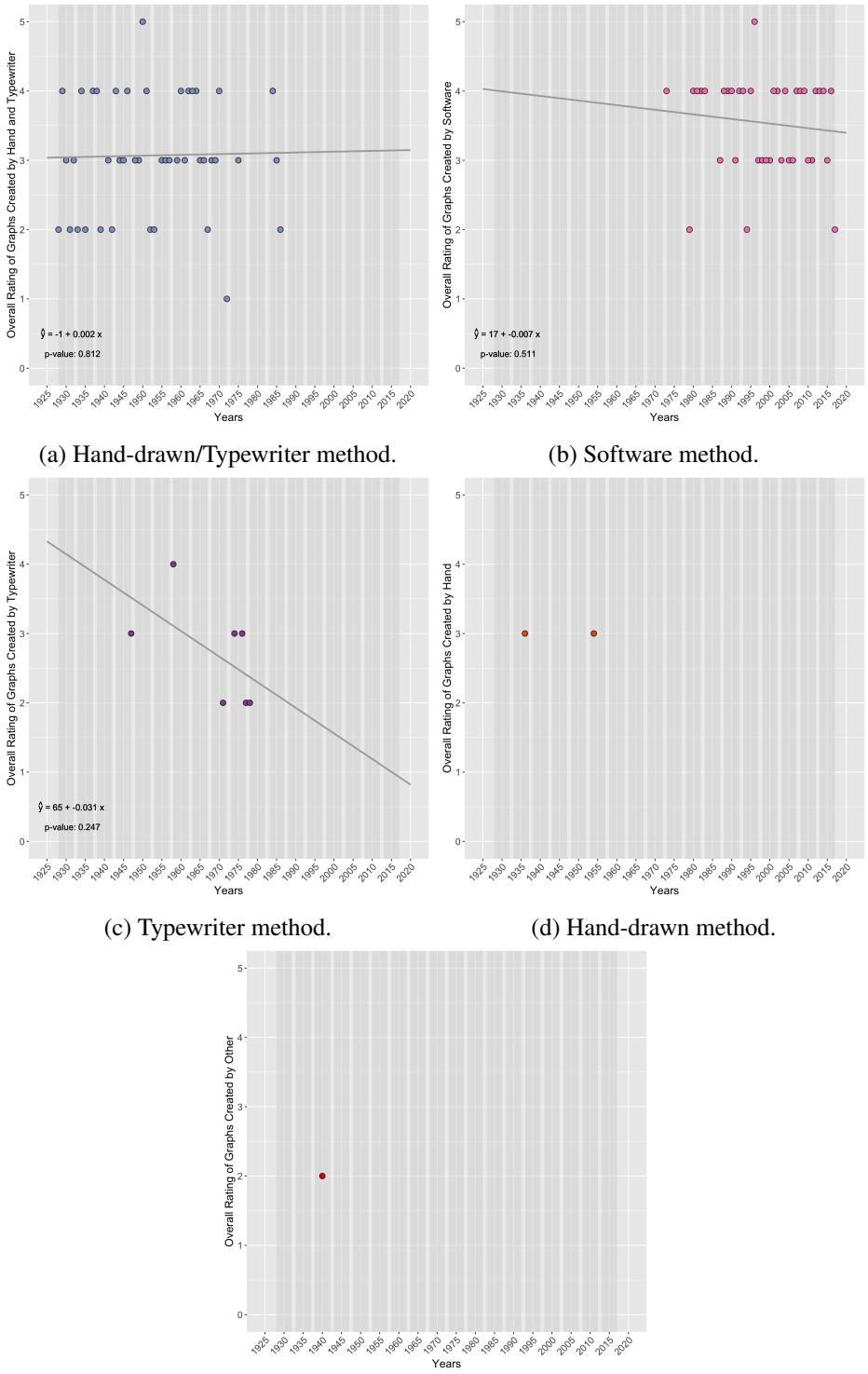
Graph Type	**Graph Note**	**Count**
Line Graph		33
	multiple graphs	5
	+ intervals	7
	+ fitted line	1
Bar Graph		31
	multiple graphs	3
	+ intervals	2
	3D	2
	single	10
	grouped	20
	side-by-side	24
	stacked	5
	overlayed	1
Scatterplot		14
	multiple graphs	4
	+ fitted line	10
Dot Plot		5
	multiple graphs	2
	+ line	3
Histogram		5
	multiple graphs	4
Other		2

APPENDIX E

Graph Creation Method Overall Ratings

In this appendix, scatterplots of the overall ratings of the different graph creation methods used to create the sampled graphs from Utah State University (USU) Plan A Master of Science (MS) thesis reports can be seen in Figure E.1. The overall ratings are plotted over time from the years 1930 to 2019, where time is aggregated into five-year intervals from 1930 to 2015. As discussed in Section 5.2.4, the overall ratings depicted in these scatterplots source from the *Overall Rating* column, which was explained in Section 3.3 and can be seen in Figure 5.1. In Section 5.2.4.3, it was highlighted that the USU Plan A MS thesis graph sample only contained two observations of the hand-drawn graph creation method and one observation of the other graph creation method. Thus, for each graph creation method, other than the hand-drawn graph creation method shown in Figure E.1d and the other graph creation method shown in Figure 5.5e, a linear regression line, $\hat{y}$, has been fit to the data and a p-value has been provided to indicate whether the regression line is statistically significant or not, as described in Section 4.3. The significance result for each regression line that was part of a series of tests are interpreted using both a significance level of $\alpha = 0.05$ and an adjusted significance level, α', which is calculated using the Bonferroni Correction as explained in Section 4.4. Since there were three different graph creation method significance tests performed, the significance level of $\alpha = 0.05$ was adjusted using the Bonferroni Correction, thus resulting in an adjusted significance level of $\alpha' = 0.0167$. Additionally, the data shown in these scatterplots were jittered, as explained in Section 4.2.

In Figure E.1a, we can see that the overall ratings for the hand-drawn/typewriter method are most commonly between the values two and four. The lowest overall rating took place in the 1970 five-year interval with a value of one, whereas the highest overall rating took place in the 1950 five-year interval with a value of five. The fitted regression line depicts no obvious trend between time and the overall ratings of the hand-drawn/typewriter graph creation method. The p-value of 0.812 establishes that there is no association between the variables according to both the original



(a) Hand-drawn/Typewriter method.

(b) Software method.

(c) Typewriter method.

(d) Hand-drawn method.

(e) Other method.

Fig. E.1: Scatterplots of individual graph creation method overall ratings over aggregated five-year intervals from 1930-2015.

significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0167$.

Figure E.1b shows the overall ratings for the software method. We can see that the values range between two and five, but the values three and four appear to be the most common. While the linear regression line demonstrates a decreasing trend over time, the p-value of 0.511 indicates that there is no significant relationship between the variables according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0167$.

In Figure E.1c, the overall ratings for the typewriter method range between the values two and four. The fitted linear regression shows a decreasing trend over time, however, according to the p-value of 0.247, we fail to reject the null hypothesis according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0167$. Thus, we conclude that there is no association between the independent and dependent variables.

Figures E.1d and E.1e show the overall ratings of the hand-drawn and other graph creation methods, respectively. As noted previously in this section, due to the small sample size of these two graph creation methods, the regression lines and p-values were not included. We can see that for the hand-drawn graph creation method, the observations took place in the 1935 and 1955 five-year intervals, both with an overall rating value of three. For the other graph creation method, the one observation took place in the 1940 five-year interval with an overall rating value of two.

As discussed in Section 5.2.4.3, no significant relationships between time and the overall ratings of the individual graph creation methods were detected according to both the original significance level of $\alpha = 0.05$ and the adjusted significance level of $\alpha' = 0.0167$.