

**EXPANDING ACCESS AND REMOVING
BARRIERS: DATA SCIENCE
EDUCATION WITH THE OPEN CASE
STUDIES DIGITAL PLATFORM**

by

Michael Robert Breshock

**A thesis submitted to Johns Hopkins University in conformity with the
requirements for the degree of Master of Science in Engineering**

Baltimore, Maryland

December 2021

© 2021 Michael Breshock

All rights reserved

Abstract

Open Case Studies (OCS) is a data science education platform created to provide examples of best practices for the next generation of data scientists [1]. The aim of this thesis project is to enhance the case studies and expand the services provided by the platform to further meet the needs of data science education today. The case studies were made accessible to non-English speakers with the addition of a feature that can toggle between translations with more than 100 languages. An RStudio "shiny" app was created to allow nontechnical users to create their own custom case studies [2]. Interactive versions of the case studies were created that incorporate live tutorials using R packages "learnr" and "gradethis" [3, 4]. This introduced the ability for case study readers to check their understanding along the way with feedback from interactive quizzes in the case study itself. Finally, an R package called "OCSdata" was developed to provide easy access to case study data and enable modular use of the case studies [5]. In tandem with these developments, the platform itself was analyzed with the help of user feedback and Google Analytics traffic data [6]. The results of this assessment indicate that OCS is already providing essential material for data science education globally. The

new developments detailed in this report will expand the reach and quality of the platform even further.

Primary Reader and Advisor: Carrie Wright (Biostatistics)

Faculty Readers: Stephanie C. Hicks (Biostatistics) and Casey Overby Taylor (Biomedical Engineering)

Acknowledgments

I am eternally grateful to my thesis advisor Carrie Wright for her abundance of guidance, encouragement and support in this process. Under Carrie's supervision this past year I have grown as an academic, a professional, and as a person. Her positive attitude and cheery outlook made this thesis project not only productive but also pleasant, even during a global pandemic. Again, thank you for everything, Carrie!

I would like to thank Stephanie C. Hicks and Casey Overby Taylor for volunteering to join my thesis committee as readers. Their feedback and expertise has been invaluable in making this thesis the best it can be.

I would also to thank the Biomedical Engineering Master's Program Manager Sam Bourne. Sam has answered countless questions from me over the past two years and always does so with a smile. Sam was very helpful with managing course requirements as well as helping Carrie and I navigate an interdisciplinary thesis project.

Thank you to the Johns Hopkins University Office of the Provost Digital Education & Learning Technology Acceleration (DELTA) Grant and the Johns Hopkins Data Science Lab for funding this project.

Dedication

This thesis is dedicated to my family whose emphasis and commitment to education inspired me to pursue this degree.

To my ancestors who first came to America in search of a better life through Fells Point in Baltimore. Earning my master's degree less than five miles away from Fells Point feels full circle.

To my parents, Bob and Julie, who worked so hard to provide the education and life I have today. Your loving support has given me the strength to push through any adversity. Thank you both so much.

Table of Contents

Abstract	ii
Acknowledgements	iv
Dedication	v
Table of Contents	vi
List of Tables	ix
List of Figures	x
1 Introduction	1
1.1 Challenges in Data Science	1
1.2 Data Science Education	3
1.3 Open Case Studies: A Data Science Education Platform	4
1.4 Case Study Structure	9
1.5 Project Goals	13

2	Methods	18
2.1	Translate Case Studies	18
2.2	MakeCaseStudies App	20
2.3	Interactive Case Studies	24
2.4	OCSdata Package	25
2.5	Assessment of Case Study Use and Interest	35
3	Results	45
3.1	Translate Case Studies	45
3.2	MakeCaseStudies App	47
3.3	Interactive Case Studies	50
3.4	OCSdata Package	50
3.5	User Assessment of Case Studies: Survey Responses	66
3.5.1	Summary of Study Population	66
3.5.2	Educators	70
3.5.3	Students & Self-learners	77
3.6	Assessment of Popularity and Reach: Google Analytics Traffic Data	86
4	Discussion and Conclusion	95
4.1	Translate Case Studies	95
4.2	MakeCaseStudies App	96
4.3	Interactive Case Studies	98

4.4	OCSdata Package	99
4.5	User Assessment of Case Studies: Survey Responses	102
4.6	Assessment of Popularity and Reach: Google Analytics Traffic Data	105
4.7	Conclusion	108
5	Appendices	111
5.1	Appendix A: Open Case Studies User Feedback Survey	111
5.2	Appendix B: MakeCaseStudies Guide	143
	Bibliography	150
	Curriculum Vitae	161

List of Tables

1.1	Sections of an Open Case Study	11
1.2	Open Case Studies Contributors	12
2.1	Example Uses of Data Types	32
2.2	OCSdata Package Dependencies	34
2.3	Adhering to CRAN Policy	34
2.4	Google Analytics Metrics & Dimensions Analyzed	43
3.1	Adhering to CRAN Policy Solution	57
3.2	OCSdata Functions	62
3.3	Numerical Results of Case Studies Used by Participants	68
3.4	Numerical Results of Likelihood to Recommend to Others . .	70
3.5	Numerical Results of Student & Self-learner Familiarity with Skills	80

List of Figures

1.1	Original Case Study	5
1.2	Data Science Over Time	8
1.3	Open Case Studies GitHub Repositories	10
1.4	DELTA Grant Project Proposal	17
2.1	International Google Analytics Data	21
2.2	MakeCaseStudies Prototype	23
2.3	Data Retrieval Workflow: Before Package	27
2.4	Case Study Data Folder Structure	31
2.5	Tweet Advertising OCS Survey	35
2.6	Survey Embedded in Case Studies	36
2.7	Survey Distribution Logic	38
2.8	Google Analytics Dashboard	42
3.1	Google Translate Button in Case Studies	46
3.2	MakeCaseStudies: Create Tab	48

3.3	MakeCaseStudies: Preview Tab and Export	49
3.4	Interaction: Multiple Choice Quiz	51
3.5	Interaction: Code Chunk Free Response	52
3.6	Interaction: Fill in the Blank	53
3.7	Counts of Interactive Exercises	54
3.8	OCSdata Algorithm Flowchart	59
3.9	OCSdata on GitHub	60
3.10	Open Case Studies Hexstickers	61
3.11	How to Use OCSdata	63
3.12	Data Retrieval Workflow: With OCSdata	64
3.13	OCSdata Download Metrics	65
3.14	Survey Participant Categories	67
3.15	Case Studies Used by Participants	69
3.16	Likelihood to Recommend Open Case Studies to Others	71
3.17	Students and Courses Taught by Educator Participants	72
3.18	Subjects of Interest for Educators	74
3.19	Case Study Materials & Sections of Interest	75
3.20	Types of Student Participants & Reason for Use	78
3.21	Student & Self-learner Familiarity with Case Study Topics	79
3.22	Student & Self-learner Familiarity with Case Study Skills	81
3.23	Student & Self-learner Learning Interests	82
3.24	Student & Self-learner Learning Outcomes	84

3.25 Student & Self-learner Satisfaction Review	85
3.26 Open Case Studies Daily Users	88
3.27 Open Case Studies Weekly Engagement	90
3.28 Total Case Study Sessions	91
3.29 Open Case Studies Global Users	93
3.30 Top Ten Cities Using Open Case Studies	94

Chapter 1

Introduction

1.1 Challenges in Data Science

Many aspects of life are becoming significantly supported by digital technology. This digital infrastructure has the ability to continuously collect data like never before and has caused an unprecedented growth in data in all facets of life [7, 8]. This seemingly never-ending wave of data has been coined by Adams 2020 as the "data deluge, in which commercial companies, the natural sciences, the social sciences, professional sports teams, government agencies, and other institutions are generating ever-increasing quantities of data" [9]. Contained within this data is information that may provide important insights. However, due to its massive nature, this information is almost never immediately obvious to the human eye with basic descriptive statistics. It would be both impractical and inefficient to comb through these massive datasets data

point by data point.

Motivated by this need for methods to extract information from big data, the field of data science was born. Data science relies on the foundations of mathematics and computer science to automate the analysis of large datasets with algorithms and programming [7, 8, 10, 11]. This requires many different skills and disciplines depending on the specific application [8]. For example, the duties of a data scientist today can include data extraction, preparation, exploration, transformation, storage and retrieval, computing infrastructures and communication of analysis results [7].

Another major focus of data science today is artificial intelligence (AI) and machine learning (ML) [7]. These are techniques where algorithms use data to learn how to achieve a desired outcome in a specific task. This has made for an unprecedented ability to automate tasks and caused an explosion of interest in AI/ML and their statistical foundations.

Data scientists today need to be well versed in a wide range of theoretical foundations and practical applications to succeed in such a dynamic role. The rapid pace of growth in this field makes keeping up with best standards and practices even more difficult. Similar to other professional fields such as public health or epidemiology, data scientists will need to continue educating themselves to stay up to date.

1.2 Data Science Education

In part due to the rapid growth of this field, academia has struggled to provide the educational resources needed to train the next generation of data scientists [12]. Consequently, much of data science education today happens on the job, usually taught by an expert practitioner [13]. While learning from an expert might be ideal for specific applications, this approach reaches a very small audience, sets little to no standards for data science education and fails to provide material to teach what standards do exist.

In an effort to address this education deficit, Hicks et al. started the Open Case Studies (OCS) project with the goal to provide self-contained, multimodal, peer-reviewed, and open-source guides of vetted real-world examples for active educational experiences [1]. The idea to create such experiential guides came from the case study model developed by Nolan and Speed 1999 [14] where students are given the opportunity to apply statistical thinking to real data [15]. Based on this model, Hicks et al. 2018 listed one of the five principles of teaching data science is to "organize the course around a set of diverse case studies" [16]. However, curating a diverse set of case studies independently can be very difficult and usually impractical for instructors. OCS was created to address this challenge and provide data science educators with a centralized source of case studies. The case studies could be used in the classroom (both onsite and online) and also outside of the classroom by acting as an archive of stand-alone examples of best practices.

1.3 Open Case Studies: A Data Science Education

Platform

When this thesis project began, the OCS team had already written and produced eleven case studies [1]. Each case study provides an online lesson in applying data science and statistics fundamentals to current public health issues with real world data. The case studies are delivered and distributed as HTML files, making them easy to read with any web browser. They look like a typical web page with the case study content displayed in the middle panel and a navigable table of contents on the left margin that scrolls along the page (see Figure 1.1). All case studies can be found on the organization's website opencasestudies.org.

The case studies are written to provide the reader with sufficient background and context on the public health topic at hand to make sense of the data being analyzed. After explaining the background, motivation, source of the data, and learning objectives, the case study transitions into sections explaining step by step how to analyze the data. Each case study has a data import, exploration, wrangling, visualization, and analysis section which takes the reader through a standard data science workflow where inferences about the data can be made along the way. See Table 1.1 for a summary of the case study sections.

All of this is done with R programming [17]. Python [18] is another programming language with similar capabilities to R in terms of statistical computing. However, the OCS developers work in the field of biostatistics where

Open Case Studies: Exploring CO₂ emissions across time

The screenshot shows a web browser window with the URL `opencasestudies.org/ocs-bp-co2-emissions/`. The page title is "Open Case Studies: Exploring CO₂ emissions across time". On the left is a sidebar with a "OPEN CASE STUDIES" logo and a table of contents including: Motivation, Main Questions, Learning Objectives, Context, Limitations, What are the data?, Data Import, Data Wrangling, Data Visualization, Data Analysis, Summary, Suggested Homework, and Additional information. The main content area contains four data visualizations:

- World CO₂ Emissions per Year (1751-2014)**: A line graph showing emissions in Metric Tonnes from 1800 to 2014. The y-axis ranges from 0e+00 to 3e+07. The x-axis is labeled "Year" with "Limited to reporting countries".
- Top 10 CO₂ Emission-producing Countries**: A heatmap titled "Ordered by Emissions Produced in 2014". The countries listed are China, United States, India, Russia, Japan, Germany, Saudi Arabia, Iran, South Korea, and Canada. The x-axis shows years from 1970 to 2010. A color scale for $\ln(\text{CO}_2 \text{ Emissions (Metric Tonnes)})$ ranges from 4 to 15.
- CO₂ Emissions (Metric Tons)**: A scatter plot with a blue trend line showing emissions from 1980 to 2014. The y-axis ranges from 4,500,000 to 5,500,000.
- Temperature (Fahrenheit)**: A scatter plot with a blue trend line showing temperature from 1980 to 2014. The y-axis ranges from 52 to 55.
- US CO₂ Emissions and Temperature (1980-2014)**: A scatter plot with a blue trend line showing the relationship between scaled temperature (Fahrenheit) and scaled emissions (Metric Tonnes). The y-axis ranges from -2 to 2, and the x-axis ranges from -2 to 1.

 Below the visualizations is a disclaimer: "Disclaimer: The purpose of the Open Case Studies project is to demonstrate the use of various data science methods, tools, and software in the context of messy, real-world data. A given case study does not cover all aspects of the research process, is not claiming to be the most appropriate way to analyze a given dataset, and should not be used in the context of making policy decisions without external consultation from scientific experts." At the bottom is a license notice: "This work is licensed under the Creative Commons Attribution-NonCommercial 3.0 (CC BY-NC 3.0) United States License."

Figure 1.1: Original Case Study | Image of the beginning of the CO₂ Emissions case study viewed on a web browser. This is one of eleven case studies written by the OCS team before this project began. Left: Navigable table of contents. Top: Title. Middle: Preview of data visualizations made at the end of the case study. Bottom: Disclaimer about the purpose of OCS.

R is the most common programming language used. For this reason, the first case studies were written in R, but the OCS developers would like to offer Python case studies in the future.

R is a programming language and environment particularly useful for statistical computing and graphics [17]. It is an excellent tool for data scientists and practically any data science related task can be executed in R. In order to execute these tasks, R programmers use what are called functions.

A function is a command that can receive an input (or several) and produces the requested output. Functions are the key to programming in R efficiently. R comes with a base set of functions. These functions cover the basic needs of any R programmer. For more complex and specific tasks, R programmers can write their own functions. Writing a function saves you time, energy, and lines of code by allowing you to use a single command to achieve the desired output instead of what can be up to thousands of lines of code. Functions can be written such that their behavior changes depending on certain input parameters, allowing them to be flexible to different contexts.

R programmers are able to save a set of functions in what is called a package. Packages are how R programmers can share their functions, data, and more with each other in an official format and structure. Packages can be downloaded, installed, and loaded into R with at most a few commands. Once downloaded and installed, the package can be loaded at the user's convenience. When a package is loaded, users have access to all of its contents. This includes the functions of the package. This allows programmers to use functions written by other developers, potentially saving them a lot of

time and energy.

The sharing of packages is a huge driving force in the growth of data science by allowing the community to benefit from each other's work [11]. Using a function available in a package that was developed by another programmer (or yourself in some cases) instead of writing a new function saves an enormous amount of time (see Figure 1.2). Using packages also enables introductory level programmers to use functions that they wouldn't have the necessary programming knowledge to write themselves. Packages are at the foundation of statistical programming in R. The nature of this free access to packages and their source material can be considered what is called "open source."

Open source data science is a pillar of the OCS education platform. As is seen in the sharing of R packages and the rapid development of data science, the availability of open source material is a powerful driver of innovation and growth by providing access to knowledge for the most amount of people [11]. For the case studies to be free and accessible to all, the programming language used within them must also be open source. The open source nature of R and its statistical packages makes it the ideal language to use in OCS.

R is continuously updated and any version, including the most recent version, can be downloaded for free from cran.r-project.org for Linux, macOS, and Windows. CRAN also hosts an official repository of R packages that have been vetted and approved to be safe for public use. The open source nature of R makes it a perfect fit for the case studies.

Another programming language that is similarly suited for OCS is Python

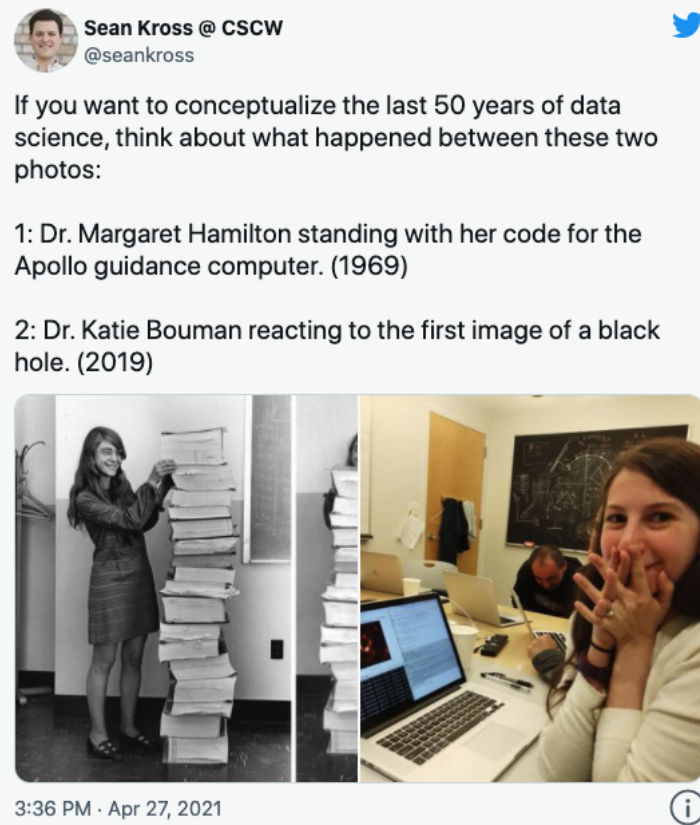


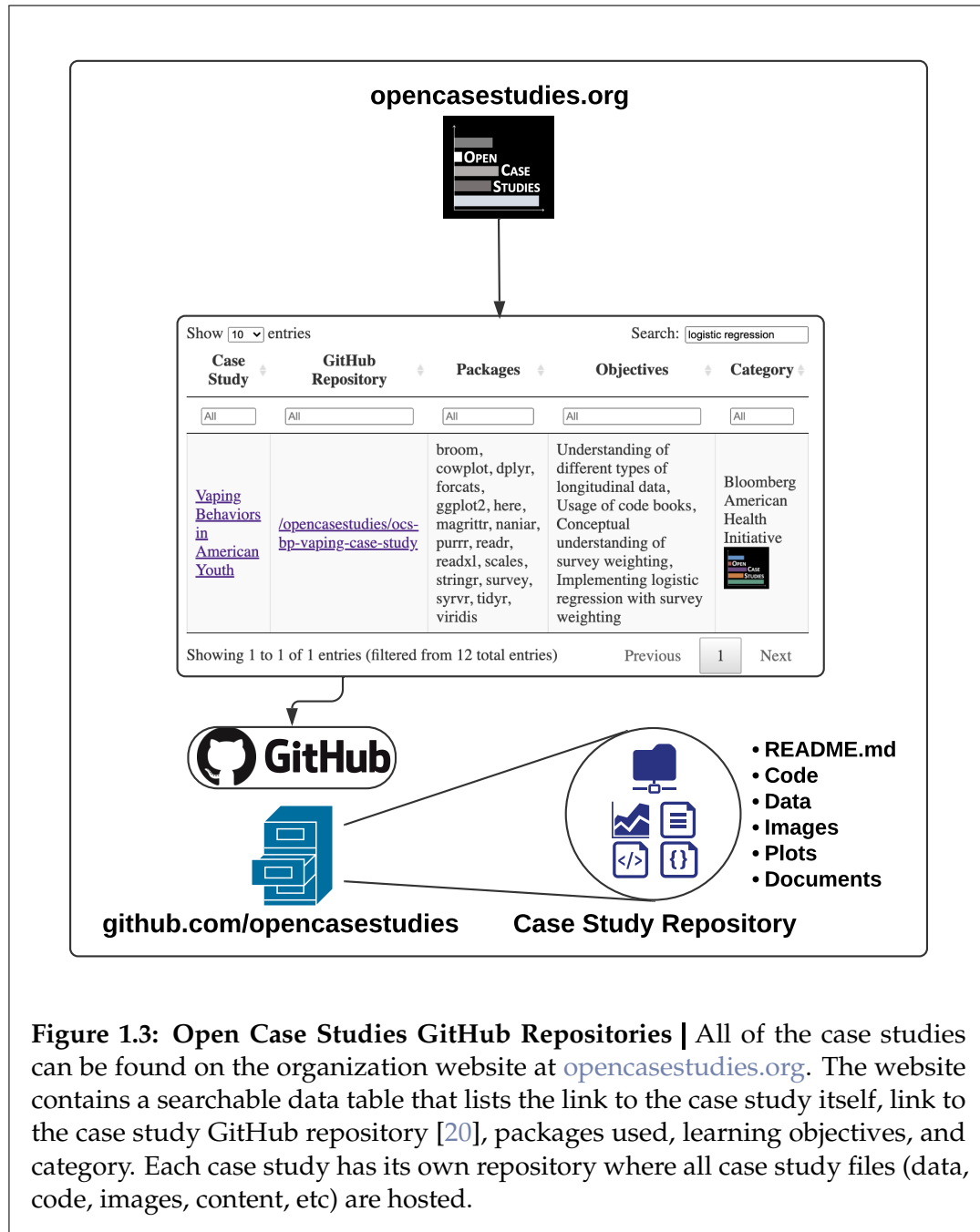
Figure 1.2: Data Science Over Time | Tweet comparing the application of data science in the NASA Apollo space mission in 1969 to the first captured image of a black hole in 2019. As can be seen in the left photograph, the number of lines of code used in 1969 filled so many books that made a stack taller than the author herself. Compare this to the photograph from 2019, where all the code is contained within a commercial personal computer. This visible change in volume is an example of how much data science has grown in efficiency and accessibility. Tweet by Sean Kross, an expert in Human-Computer Interaction [19].

[18]. Python is also an open source programming language and environment that has freely distributed statistical packages. Different fields will use different programming languages for a variety of reasons. OCS focuses on public health topics and was developed by biostatistics faculty. These fields today mostly use R, so the case studies were initially developed in R. However, OCS developers plan to offer Python based case studies in the future.

1.4 Case Study Structure

In the spirit of open source material, the case study web pages are hosted through GitHub [20] and each case study comes with its own GitHub repository. Users can find the case study data in these repositories, as well as all the other files that make up a case study (see Figure 1.3).

The case studies are divided into a set of fundamental sections that walk the reader through a standard data science workflow [14]. All case studies begin with a standardized set of section headers that provide the reader with some background on the subject of the data to be analyzed. These sections cover the motivation for analysis, main questions to be investigated, case study learning objectives, context, limitations, and a description of the data itself. Once the background is sufficiently covered the case studies jump into sections that go over how to analyse the data. Case studies will have most if not all of the following sections: data import, data wrangling, data exploration, data visualization, and data analysis. The case studies end with a summary conclusion, suggestions about homework for readers to do on their own, and



Case Study Section	Description
Motivation	Motivating figure and text at the start of the case study
Main questions	Scientific question(s)
Learning objectives	Both data science and statistics learning objectives
Context	Context of question(s) or data
Limitations	Any limitations in case study or with data used
What are the data?	Summary of where the data came from and what the data contain
Data import	Analyses for importing data
Data wrangling	Analyses for wrangling data
Data exploration or data visualization	Analyses for data exploration and visualization
Data analysis	Analyses containing statistical concepts and methods to answer question(s)
Summary	Summary of results and conclusion
Suggested homework	Question(s) to explore further
Additional information	Helpful links and packages used

Table 1.1: Sections of an Open Case Study. Each case study will include most if not all of the listed sections. Left column corresponds to section title and right column contains a description of the corresponding section. Sections listed in the order they generally appear in a case study (top to bottom).

resources to find additional information. See Table 1.1 for a list of each section with descriptions.

The first case studies developed were a tremendous achievement in providing open source data science lessons with practical applications. Based on initial feedback, OCS is already filling a gap in data science education and providing much needed material to budding data scientists. This initial success motivated the OCS team to expand and enhance the platform. See Table 1.2 for a detailed list of all OCS authors and contributors.

Type of Contributor	Names of Contributors
Authors	Stephanie C. Hicks, Carrie Wright, Leah Jager, Margaret Taub, Michael Ontiveros, Kexin (Sheena) Wang, John Muschelli, Qier Meng, Michael Breshock
Public Health Consultants	Jessica Fanzo, Brendan Saloner, Megan Latshaw, Renee M. Johnson, Daniel Webster, Elizabeth Stuart, Aboozar Hadavand, Roger Peng, Kirsten Koehler, Alex McCourt, Ashkan Afshin, Erin Mullany
Data Science Reviewers	Leslie Myint (Macalester College), Shannon E. Ellis (University of California – San Diego), Christina Knudson and Students: Jensen Stanton, Tina Trinh, Ruby Ho, Lukas Buhler and Anonymous (University of St. Thomas), Michael Love (University of North Carolina), Nicholas Horton (Amherst College), Mine Çetinkaya-Rundel (University of Edinburgh, Duke University, RStudio)
Funding	Bloomberg Philanthropies: Bloomberg American Health Initiative (Director: Joshua M. Sharfstein, Associate Director: Michelle Spencer, Special Projects Officer: Paulani Mui), Digital Education & Learning Technology Acceleration (DELTA) Grants (Executive Vice Provost: Stephen Gange, Provost’s Fellow: Ira Gooding)

Table 1.2: Open Case Studies Contributors. Names of the all the individuals who contributed to the creation and review of the case studies. OCS was funded as a [high impact project](#) for the [Bloomberg American Health Initiative](#) to Stephanie C. Hicks as principal investigator [21]. The expansion of the project and the work done in this thesis was funded by the [Digital Education & Learning Technology Acceleration \(DELTA\) Grants](#) awarded to Carrie Wright as principal investigator.

1.5 Project Goals

With the core content of the current case studies well established, the OCS team identified a few actionable areas for improvement that would broaden the case studies' reach and impact. These areas were identified by the OCS developers themselves with the help of early feedback from educators asked to review the case studies (see Table 1.2 for the names of OCS contributors). The educators provided helpful comments on what might help improve the case study teaching experience. Based on this feedback, this project set out to improve the case studies by enabling translation, increasing interactivity, enabling the creation of custom case studies, and expanding access to case studies.

One of the first glaring barriers to the previous case studies was that they were only offered in English. However, the whole world is experiencing a data revolution [9]. To make OCS accessible to a broader audience, the language barrier to our case studies should be minimized as much as possible. Based on previous experience, the team felt that offering the case studies in at least a handful of the world's most predominant languages would significantly grow our user base and provide data science education more globally.

Another identified aspect with room for improvement was the case studies' interactivity. While the case studies are built to be an effective classroom exercise, they are also intended to be a useful resource for students and self-learners when studying on their own. In a classroom setting, instructors can engage a variety of tactics to engage their students with the material. There's also plenty of opportunity for the instructor to pause and check students'

understanding. However, the case studies alone may not provide enough opportunities for independent learners to interact with the material and check their understanding on their own. There was worry that this lack of built-in engagement and feedback may negatively impact learning outcomes during independent use.

To make OCS more engaging for readers both in and out of the classroom, interactive elements would be incorporated into the case study structure. Interactive elements such as multiple choice questions, fill in the blank and fill in the chunk coding exercises would engage OCS users by providing opportunities to check their comprehension. Allowing users to test their skills and receive feedback improves learning outcomes. Such interactive elements could be embedded into the case studies with minimal changes to the original version with the help of the "learnr" and "gradethis" packages by RStudio [3, 4]. Inspiration for this idea was taken from other online coding resources such as [w3schools.com](https://www.w3schools.com) where users are able to practice coding in their web-browser [22].

This aspect of the project was mostly completed by my lab partner Qier Meng, a Master's student in Biostatistics. As such, the interactive case studies will not be a focus of this thesis, but will still be touched on to provide a complete report of the project's accomplishments.

OCS is a platform dedicated to providing its users with a wide range of options for educational material and resources. To expand user's options with OCS, a web application would be developed to provide a simple, user-friendly tool for constructing case studies. The app would be designed so that users

can easily input their own text, figures, and videos into the app. Once the user is finished and satisfied, they would just need to press a button, and the app would knit the case study together into the standard OCS format and download the final case study to the user's local computer. Users would then be free to use and distribute their case study as they wish.

The main improvement that will be discussed in this thesis is the enhanced case study accessibility. All of OCS data is available in the respective case study's GitHub repository [20]. However, case study users new to GitHub can find it a confusing process to download files from repositories. Additionally, users must move the downloaded data to the appropriate project directory. Overall, this process leaves room for error and acts as a barrier to introductory students. Troubleshooting these errors can be a headache for both students and instructors and eats away at valuable learning time. Inspired by the base R "datasets" package [17], the OCS team set out to develop an R package that would simplify the data retrieval process and provide easy access to case study repositories.

Another idea to make the case studies more accessible was to make them modular. Modular means that users would be able to jump to any section of interest for their particular needs and be able to learn from that section without needing to work through any of the previous sections. The structure of the original case studies prevented much modular usage. Each section involving data exercises will reformat the data in such a way that prepares it for the next step in analysis. As such, each section (except for the first) relies on the work done in the previous section. The case studies were designed

this way to provide a cohesive story that simulates data science in the real world. Unfortunately, this also makes it very difficult to use the case study modularly. In tandem with package development, the data files themselves would be reorganized to both support the package infrastructure as well as enable modular use of case studies. The package would also support this modular use by allowing users to download only the specific files they need for the sections they work on.

In sum, the main contribution of this project is the expansion of case study accessibility through the new developments detailed in this report. See Figure 1.4 for a diagram of the project improvements proposed in the application to the Johns Hopkins University Office of the Provost Digital Education & Learning Technology Acceleration (DELTA) Grants program. The final improvements made in this project were founded on the ideas outlined in this proposal. See Table 1.2 for more information on project funding.

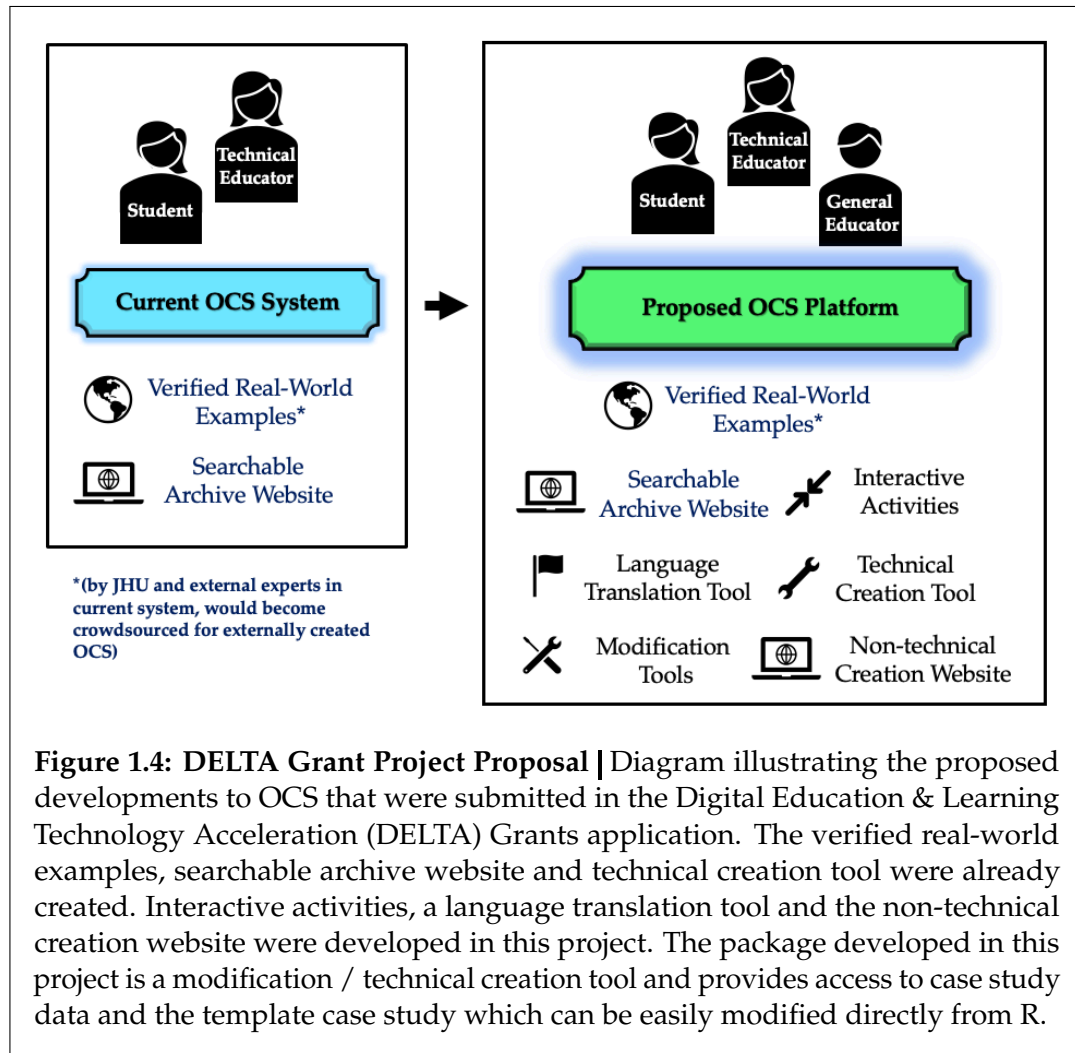


Figure 1.4: DELTA Grant Project Proposal | Diagram illustrating the proposed developments to OCS that were submitted in the Digital Education & Learning Technology Acceleration (DELTA) Grants application. The verified real-world examples, searchable archive website and technical creation tool were already created. Interactive activities, a language translation tool and the non-technical creation website were developed in this project. The package developed in this project is a modification / technical creation tool and provides access to case study data and the template case study which can be easily modified directly from R.

Chapter 2

Methods

2.1 Translate Case Studies

In order to translate the case studies, we initially planned to develop a package that would automatically translate the case studies. We aimed to produce case studies in at least a few common global languages, such as Mandarin, Spanish, and French. Based on the previous work of OCS team members, we believed this to be reasonably achievable. The plan was to repurpose a package previously made by OCS developer Dr. John Muschelli which translated text into audio [23]. It would still take significant work, but using the previous package as a starting point made translating the case studies with a package seem possible.

While conducting an exploratory analysis of global traffic to the OCS website, we noticed that the page titles in some events were showing up in

other languages. We could see some of the titles listed in Simplified Chinese, Spanish, Turkish, and Arabic (see Figure 2.1). This amazed us as somehow our users were already viewing the case studies in their local language, even languages with different alphabets! After some further investigation and brainstorming, we came to the conclusion that the users were taking advantage of Google Translate's ability to translate web pages. This reminded me of how websites often have a Google Translate button somewhere on their pages that will translate to any language the user picks. Inspired by our users' ingenuity, we investigated how to incorporate a Google Translate button into the case studies. With a quick search, we discovered that we could do this with a few lines of HTML code [24]. See the code below:

```
<div id="google_translate_element"></div>
```

```
<script type="text/javascript">
```

```
function googleTranslateElementInit() {  
    new google.translate.TranslateElement(  
        {pageLanguage: 'en'},  
        'google_translate_element');  
}
```

```
</script>
```

```
<script type="text/javascript"
```

```
src="//translate.google.com/translate_a/element.js?
```

```
cb=googleTranslateElementInit"></script>
```

The first line initiates a new web page element that will contain the translate button. The script in the next line defines a JavaScript [25] function that creates the translate button. The script in the last line adds a reference to the Google Translate API [26] which allows the function defined in the previous line to access Google Translate.

To give ourselves time to focus on other aspects of the project that needed more attention, the Google Translate button was used to translate the case studies, rather than developing a translating package. Google Translate will provide our users with significantly more language options than is currently available. Additionally, the ability to change the language with the click of a button is attractive. Users could use this functionality to toggle back and forth to learn the words for technical terms in a different language. The quality of translation provided by Google Translate was assessed with a literature review and a face validity check done by a colleague fluent in Chinese.

2.2 MakeCaseStudies App

Case studies in general are an excellent education tool for many fields and disciplines outside of data science as well. For example, case studies are fundamental for education in medicine, law, psychology, and counseling to name a few [27, 28, 29, 30]. This ubiquitous teaching with case studies is likely due to their ability to explore complex issues in depth with real-life settings [31]. This would be attractive to any educator needing to teach practical and

Page title and screen class ▾	+	↓ Views
9 Open Case Studies: Opioids in United States		146
10 Open Case Studies: School Shootings in the United States		111
11 Open Case Studies: Influence of Multicollinearity on Measured Impact of Right-to-Carry Gun Laws Part 1		72
12 Exploring CO2 emissions across time		10
13 Exploring CO2 emissions across time		6
14 Vaping Behaviors in American Youth		6
15 (not set)		1
16 Exploring CO2 emissions across time \		1
17 Açyk mysallar: Wagtyň geçmegi bilen CO2 zyňyndylaryny öwrenmek		0
18 Estudios de caso abiertos: disparidades en la desconexión de los jóvenes		0
19 Estudios de caso abiertos: exploración de las emisiones de CO2 a lo largo del tiempo		0
20 Estudios de casos abiertos: comportamientos de vapeo en la juventud estadounidense		0
21 Estudios de casos abiertos: exploración de patrones globales de obesidad en regiones rurales y urbanas		0
22 دراسات الحالة المفتوحة: استكشاف انبعاثات ثاني أكسيد الكربون عبر الزمن		0
23 开放案例研究: 多重共线性对运载工具权法的可衡量影响		0
24 开放案例研究: 探索跨时间的二氧化碳排放		0

Figure 2.1: International Google Analytics Data | Image of the results identified in Google Analytics where the case study web-page titles appear in languages other than English. This discovery suggested that users were able to translate the case studies with Google Translate and indicated that OCS could take advantage of this tool. Note: The views for these pages are listed as zero. This is due to the fact that Google Analytics only counts a view when the session is first initiated. The OCS website defaults to English, so the session would already be initiated when users change to the desired language. Google Analytics records this change in page title, but does not count it as a new view.

applied skills.

The use of case studies in education is in no way new, but the method in which they are delivered has changed drastically. Education resources today are almost exclusively stored online. This has only become more true in the wake of the COVID-19 pandemic. Case studies now need to be in digital format in order to be properly distributed to the appropriate channels.

However, constructing a digital case study can pose too much of a barrier to educators. Many teachers find learning the skills needed to create their own digital case study would take too much time on top of all of their other responsibilities. Even for teachers who already have the skills may find this too time consuming when needing to create course content in a crunch.

To address the needs of educators today, OCS set out to develop a web application for users to create their own case studies. This would enable educators to easily convert their own course content into a digital case study format. Before this project began, a prototype app was developed as a proof-of-concept. It was built using RStudio's "shiny" package [2]. The prototype consisted of a main panel and sidebar user interface where users can input headings, text, image and video links to be included in the case study (see Figure 2.2). The users would then click a download button to have their inputs knitted into an HTML file that could be shared with students or hosted online. The app would be called MakeCaseStudies.

This project was able to build off this prototype. With the basic function of knitting a case study together from user inputs already achieved, improving the app's user interface and experience was the main focus of this project. To

Create an Online Case Study

This tool is provided to help users create online lessons *quickly and easily* like our [open case studies](#), which are online step-by-step lessons that guide users through a *real-world problem solving challenge*.

Logo Image URL

<https://opencasestudies.github.io/img/logo.jpg>

Title

Gram Negative Vs Gram Positive Bacteria

Document format

HTML

Start by clicking the **Make Case Study** button to download an example lesson.

[Make Case Study](#)

Delete and replace the existing content for your own content and press the **Make Case Study** button again to download your own lesson!

Powered by:



Hosted by:

[Open Case Studies](#)



Photo by [Samantha Borges on unsplash](#)

Main Image URL

<https://thisonevsthatone.com/wp-content/uploads/Grar>

Header1

Gram-positive vs Gram-negative, what's 1

Narrative Text Section 1

Gram-positive is a type of bacteria that h:

Image 1 URL

<https://cdn1.byjus.com/wp-content/uplo:>

Header2

Gram Stain Method

Narrative Text Section 2

During Gram staining both bacteria are si

Image 2 URL

<https://cdn.technologynetworks.com/tn/>

Figure 2.2: MakeCaseStudies Prototype | Image of the prototype app developed as a proof-of-concept for MakeCaseStudies. The version shown has had minimal updates to the app design, but the core structure is the same as the original. This app was able to make a simple case study, however needed improvements in its user interface and flexibility to be useful for educators.

bring the prototype up to OCS standards, the layout was reoriented, color-coordinated with the OCS website, and a preview tab function was added. In the preview tab, users would be able to view the case study being constructed before downloading it in HTML format. The app was also improved to be more flexible in the number of sections users are allowed to make. This was all done again using RStudio's "shiny" package. This package makes it very simple to build interactive web apps in R [2, 32].

No method to officially evaluate the MakeCaseStudies app has yet to be implemented. Unfortunately, no data was collected on the prototype app, so we are unable to analytically compare the new and old versions. Future work on this project should make it a priority to collect data on the app either through Google Analytics, user surveys, or other methods.

2.3 Interactive Case Studies

The following aspect of this project was mostly developed by Qier Meng, biostatistics master's student and my lab partner on this project. All of the case studies were made interactive by Meng except for one. I worked on the interactive version of the "School Shootings in the United States" case study [33]. Meng will not be writing a thesis on her work with the case studies, so her success in developing the interactive case studies will be touched on to provide a comprehensive review of the project results.

The case studies were converted into a new format in order to incorporate interactive exercises. These exercises include multiple choice questions, fill

in the blank, and writing chunks of code. The exercises provide feedback and allow the user to check their progress along the way. This would all be achieved with the help of the "learnr" and "gradethis" R packages. All of the case studies are written using what is called an R Markdown document. R Markdown documents make it very easy to write reports that include R code chunks and code outputs [34, 35, 36]. This was ideal for writing the case studies which include a lot of coding examples. The "learnr" package makes it possible to write interactive tutorials inside R Markdown [3]. Furthering this functionality, the "gradethis" package allows feedback to be incorporated into "learnr" tutorials [4]. This feedback could be hints and tips or comments on why an answer is correct or incorrect. The interactive case studies would be hosted on [RStudio Connect](#) to allow for the new live tutorials [32, 37]. RStudio Connect is a publishing platform that can host live reports, dashboards, and more [38].

An official assessment of the interactive case studies has yet to be conducted. Future work on the case studies shall include research on the effectiveness of interactive case studies along with a survey of user experience.

2.4 OCSdata Package

While all of the case study data are available in their respective GitHub repositories [20], accessing this data has proved troublesome for some users. Previously, using case study data required users to download the data files from GitHub and move them into the correct folder on their personal computer

(see Figure 2.3). Based on the teaching experiences of the OCS team, including myself, this process leaves a lot of room for error for students, particularly when the file needs to be moved. Inspired by packages like the "datasets" package included in base R [17], this project was motivated to develop its own package that would provide easier access to OCS data. This method would be faster, allowing for students and educators to focus on the material at hand.

Package development initiated with an investigation into how the "datasets" package and others like it are able to package data files. It was discovered that R packages can contain what is called an RDA file. RDA stands for R Data and is one of the official file formats used to export R objects [39]. Object is the technical term used to refer to a variety of data structures in R. They can be vectors, matrices, data frames (think spreadsheet), functions, plots and more. As such, any object created in R, such as a data frame, can be saved to an RDA file. These RDA files can be included in a package so that when the package is installed, the R objects saved in the files can be loaded directly into the user's environment. This is how "datasets" and similar packages work. Unfortunately for OCS, this method does not work for non-RDA files.

The data used in OCS comes from a variety of different sources. Each source often has its own data standards and formats used. As such, the case study data comes in a variety of different file formats. Some of the data is in RDA format, but many files are not. Case study data files may also be in CSV, XLS/XLSX, PDF formats and more. As more case studies are added, this list will likely grow.

It is necessary for the case studies to include data files in their original

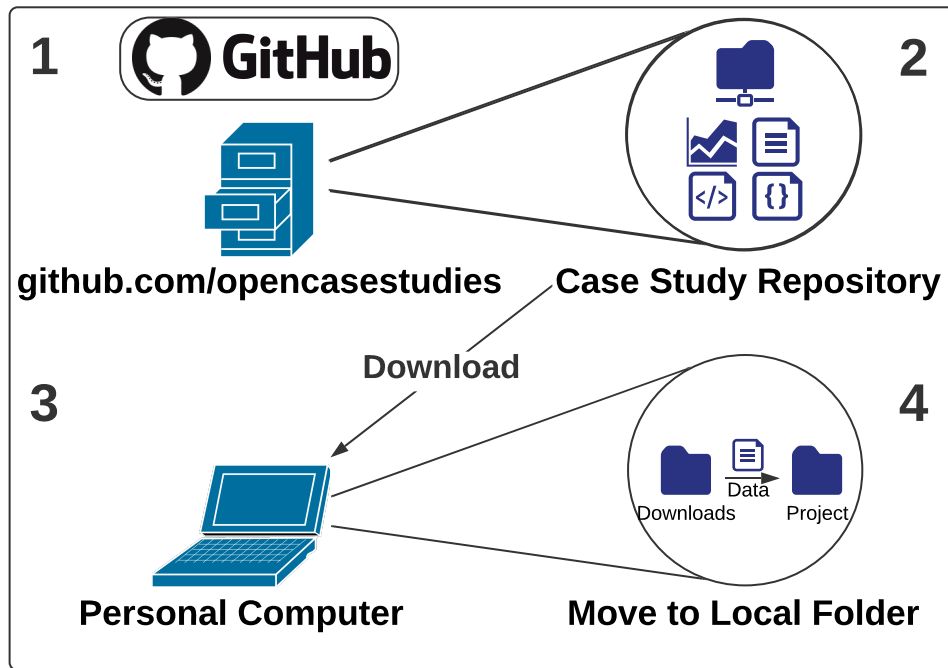


Figure 2.3: Data Retrieval Workflow: Before Package | Diagram of the data retrieval process before package development. Users must navigate to the case study’s GitHub repository [20] and download the data files needed from there. Then the user must move the downloaded files to the appropriate local project folder. This process leaves room for error, especially for first time users. For example, some users may not find GitHub easy to navigate and fail to download the necessary data file.

format. The case studies are intended to provide a hands-on tutorial of a typical data science workflow that emulates as closely as possible what a professional data scientist would do in the real world. Data in the real world is messy and will vary from source to source. A data scientist will be expected to analyze data in whatever format it is provided. As such, using the unaltered source data in the case studies is the best way to prepare students for the real world. In order to provide access to all case study data, the package would need to be able to handle more than one file type.

This initial investigation concluded that the OCS package would have to operate differently than the "datasets" package. After some further research, a possible solution was found. Rather than including the data files in the package itself, the package could provide functions that would automatically download case study data files to a local directory that could be easily accessed by users. The method would still be very simple by requiring minimal input from the user. In case of any confusion, each section of the case studies would be updated to include instructions on how to download the required data files using the package. After some initial testing, a basic algorithm was established for downloading case study files:

1. Create a new folder named OCSdata to save files in. This is done to avoid overwriting users local files, which would happen if a file is downloaded that coincidentally has the same name as one of the user's files. The folder is named "OCSdata" instead of "data" because "data" is a common word used to name folders, and any pre-existing folder named "data" would be overwritten.

2. Scrape data from the case study repository selected by user. In this step the `GET()` function from the "httr" package [40] is used to extract data from the case study repository websites hosted on <https://github.com/opencasestudies/>.
3. Once the data is extracted, save it to the user's hard drive in the folder that was specified by user with `write_disk()`, also from "httr" [40].

This algorithm would be flexible enough to download any data type.

In parallel with package development, the data itself would be restructured to make modular use of the case studies possible. To do this, new data files were created at the end of each section of the respective case study. These data files were posted to the case study repositories and organized into various sub-folders based on which section the file comes from. This infrastructure was already in place for the wrangled data files, but was expanded to all versions of the data in this project.

These folders were named "raw," "imported" and "wrangled." The raw folder contains the data files as they came from the source. These files would be used in the data import section. The imported folder contains data files created in the data import section. They have been imported and saved as R objects and would be used in data exploration and data wrangling sections. The wrangled folder contains data files created in the data wrangling section. They have been cleaned and prepared for analysis. These files would be used in data visualization and data analysis sections.

Two more sub-folders were created and named "simpler_import" and "extra." The former was created at the request of an OCS reviewer. The

reviewer commented that in some case studies the data import step is quite complicated due to the format of the raw data. They went on to suggest that providing the raw data in more user friendly file formats (such as CSV) in these cases would be beneficial to educators who want their students to practice data import, but don't want it to be overly complicated. As such, CSV versions of the raw data were created for case studies where the raw data was not already in CSV format. These files are placed in the "simpler_import" folder for users looking for a simple import process. The "extra" folder was created to contain any of the source data files that were not used in the case study analysis. Users are free to use the extra data files for further analyses as they wish. The "simpler_import" and "extra" folders are offered in only select case studies. See Figure 2.4 and Table 2.1 for more information on the data folders.

To keep consistent with this modular structure, the package would be developed so that each data sub-directory (category) would have its own corresponding function. For example, if I wanted to download raw data from a case study, I would use a raw data specific function. For downloading data from the imported folder, I would use a imported data specific function, and so on. Additionally, the package would provide functions to automatically download whole case study repositories. This would allow users to obtain all case study files directly from R to use, edit, and distribute as they please.

There were many R packages that supported the development of the OCS-data package and its functions. The "devtools" package is a fundamental tool

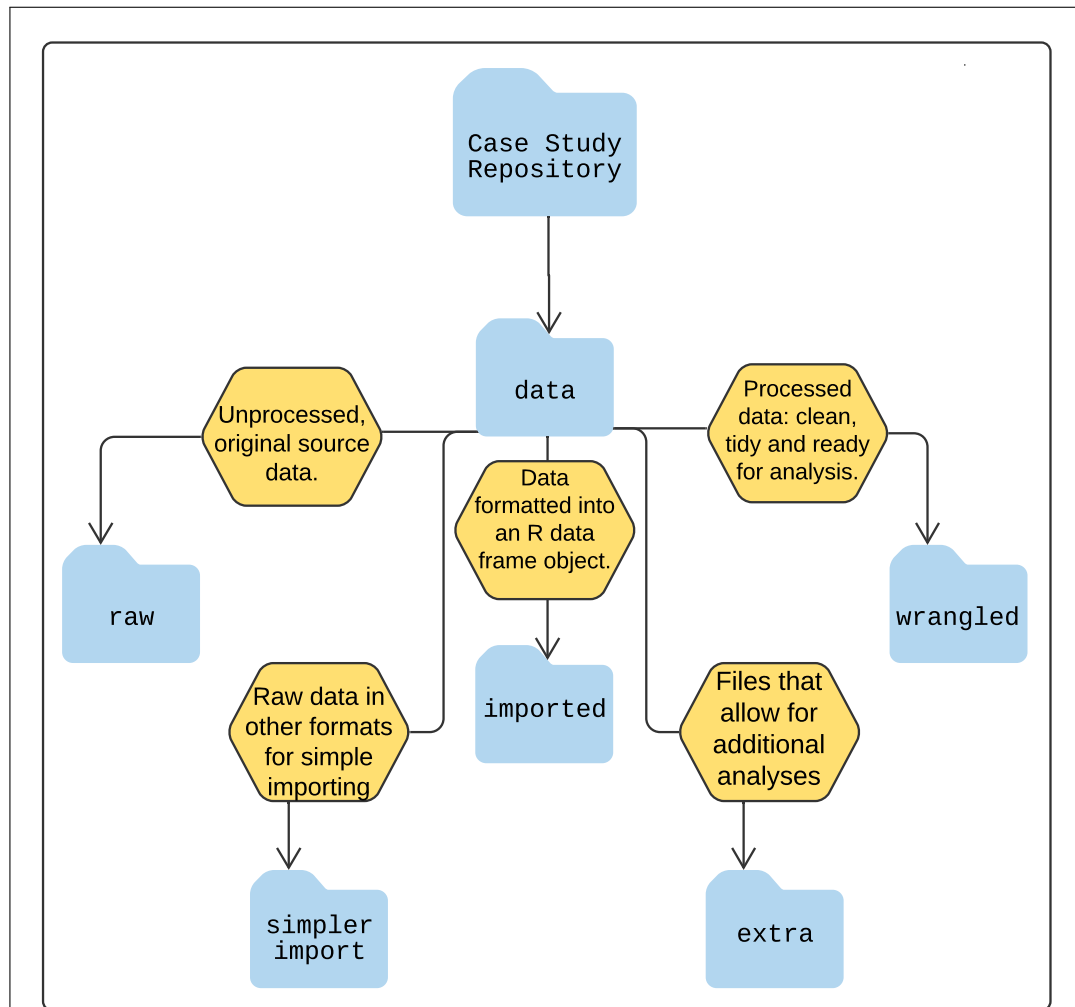


Figure 2.4: Case Study Data Folder Structure | Tree diagram illustrating the organization and structure of Open Case Study data files. The *data/raw* folder contains data files as they came from their source. Raw data that has been converted into an R object and saved as an R data file can be found in the *data/imported* folder. The *data/wrangled* folder contains clean data that's ready for analysis or visualizations. All case studies offer at least raw, imported, and wrangled data folders. Select case studies also offer *data/simpler_import* and/or *data/extra* folders. The former contains raw data files that have been converted to file formats that are friendlier for data import in R. The latter contains extra raw data files that were not used in the case study but are available for users to conduct their own analyses.

Data Folder	Case Study Section	User	Example Use
raw	Data Import	Student	Data science students looking for open source data for a group project.
imported	Data Exploration, Data Wrangling	Student	Public health student practicing data wrangling and visualization for a course assignment.
wrangled	Data Visualization, Data Analysis	Self-Learner	Statistics student looking for real world examples outside of class.
simpler_import	Data Import	Educator	Advanced data analysis instructor who wants students to practice data import without over-complication.
extra	Not Used in Case Study	Educator	Course instructor assigns homework using related but new data that expands beyond the case study.

Table 2.1: Example Uses of Data Types. OCS saves the data at the end of each section and makes these files available to users. Users can start off at any section in the case study and use these data files to skip the work completed in previous sections. The table lists which data to use in the data folder column based on the start point in the corresponding case study section column. Example uses and their users are also listed for each data folder category.

for any package developer. It offers extremely helpful functions that automate package documentation, testing, building and more [41]. "roxygen2" is another package that supports documentation by automatically creating manuals with instructions for how to use each function [42]. The package "rmarkdown" was used to write package documentation. Writing in R Markdown was advantageous because media such as images, videos, and GIFs can easily be included along with text in R Markdown documents [34, 35, 36]. Once R Markdown documents were finalized, they could be exported into readable PDF and HTML formats with the "knitr" package [43, 44, 45]. Inside the package functions themselves, functions from "usethis," "httr" and "purrr" were used as well as base R. The `create_from_github()` and `use_zip()` functions from "usethis" give OCSdata users the option to download a whole case study repository either by cloning or as a zip file [46]. The `GET()` function from "httr" was used to both scrape repository data and download the data files themselves [40]. The `map()` function from "purrr" was used to reshape the repository website data in a usable format [47]. See Table 2.2 for a summary of the packages used to develop OCSdata.

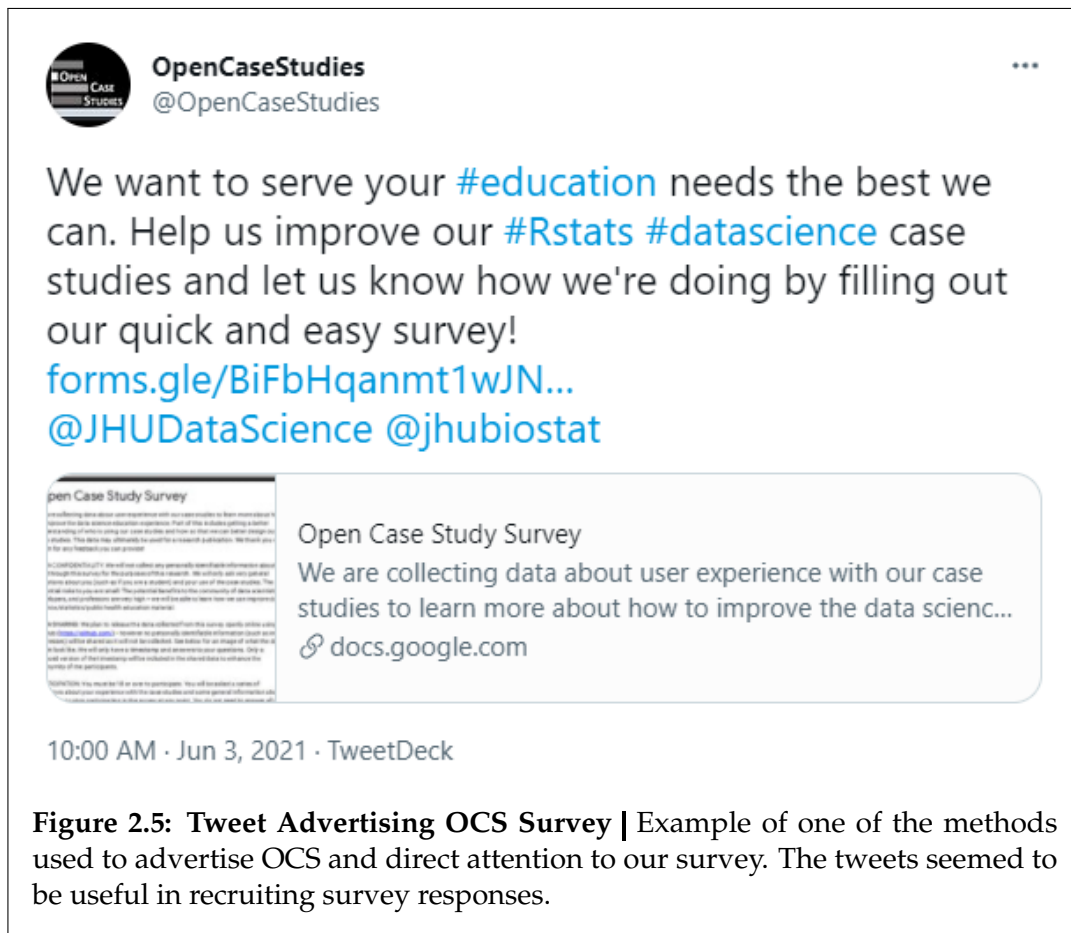
Once the package was ready for use in OCS, it was submitted to the Comprehensive R Archive Network (CRAN). CRAN is an official collection of R packages. Making the package available on CRAN would allow the number of downloads to be tracked and provide an official, safe source for our users. The first package version submitted to CRAN was denied mainly because the package broke two of CRAN's policies. The policies in question are discussed below in Table 2.3 as well as methods used to find solutions.

Package	Use	Documentation
devtools [41]	Package development	devtools.r-lib.org
roxygen2 [42]	Package documentation	roxygen2.r-lib.org
rmarkdown [34, 35, 36]	Documentation writing	rmarkdown.rstudio.com
knitr [43, 44, 45]	Documentation formatting	yihui.org/knitr
usethis [46]	Download repositories	usethis.r-lib.org
purrr [47]	Wrangling repository data	purrr.tidyverse.org
httr [40]	Scrape API data	httr.r-lib.org

Table 2.2: OCSdata Package Dependencies. Table of the packages that were used to develop OCSdata.

CRAN Policy	Method in Breach	Investigation
Packages may not alter the user's global environment.	In the first version of the package, RDA files were directly loaded into the user's environment rather than downloaded. This would allow users to skip a step in loading the data file.	The pros and cons of skipping the import step versus releasing the package on CRAN were weighed. It was decided that adding one more step to the process of loading an RDA file was a small sacrifice to adhere to CRAN's policies.
Packages may not default to saving files in the user's current working directory.	When user's do not provide any input on where to save the downloaded files to on their computer, the functions would default to saving the files in their current working directory.	Researched how other packages adhered to this policy by consulting with Julia Silge and investigating the methods used in the "textdata" package [48].

Table 2.3: Adhering to CRAN Policy. This table provides an outline of the process the package went through to adhere to CRAN policy. The CRAN policies that the first package submission was in breach of are listed in the first column. The second column discusses the methods used in the first package version that broke these policies. The third column explains what methods were used to investigate the issue at hand and research possible solutions.



2.5 Assessment of Case Study Use and Interest

The quality of material provided by the OCS platform is assessed in this project with a user feedback survey. The survey was made with Google Forms and can be viewed online at forms.gle/Wu1Q1mNZhsfyncs66 [49]. It is advertised on our website, within the case studies themselves, and on our new Twitter page. See Figure 2.5 and Figure 2.6 for images of the survey in a tweet and in a case study, respectively. Approval was obtained from the institutional review board to advertise the survey in this way.

To cite this case study please use:

Wright, Carrie and Ontiveros, Michael and Jager, Leah and Taub, Margaret and Hicks, Stephanie. (2020). <https://github.com/opencasestudies/ocs-bp-co2-emissions>. Exploring CO2 emissions across time (Version v1.0.0).

To access the GitHub repository for this case study see here: <https://github.com/opencasestudies/ocs-bp-co2-emissions>. This case study is part of a series of public health case studies for the Bloomberg American Health Initiative.

Please help us by filling out our survey.

Open Case Study Survey

We are collecting data about user experience with our case studies to learn more about how to improve the data science education experience. Part of this includes getting a better understanding of who is using our case studies and how so that we can better design our case studies. This data may ultimately be used for a research publication. We thank you very much for any feedback you can provide!

DATA CONFIDENTIALITY: We will not collect any personally identifiable information about you through this survey for the purposes of this research. We will only ask very general questions about you (such as if you are a student) and your use of the case studies. The potential risks to you are small. The potential benefits to the community of data scientists, developers, and professors are very high – we will be able to learn how we can improve data science/statistics/public health education material.

DATA SHARING: We plan to release the data collected from this survey openly online using GitHub (<https://github.com/>) - however no personally identifiable information (such as email addresses) will be shared as it will not be collected. See below for an image of what the data might look like. We will only have a timestamp and answers to your questions. Only a reduced version of the timestamp will be included in the shared data to enhance the anonymity of the participants.

Figure 2.6: Survey Embedded in Case Studies | The survey was also included near the beginning of the case studies themselves. This ensured that case study users would see the survey to increase responses.

The survey asks users questions about how they used the case studies and their general opinion of them. In some cases, users are asked specifically to compare learning data science with versus without case studies. The survey begins by gauging the participant's general impression of OCS. Participants are then asked to identify themselves as either an educator, student or self-learner. The survey will pose different types of questions to participants depending on how they identify.

Participants must be 18 years of age or older to be included in this study. They also must consent to their responses being used for research. If a participant is under 18 years old or does not consent, they are excluded from the study. See Figure 2.7 for a diagram of the survey distribution logic.

All participants were asked to select the case studies they had already looked at when taking the survey. Results were broken into the individual categories as well as the absolute total (all participants) to analyze which case studies were popular for each type of user as well as which case studies were most popular overall. They were also asked how likely they were to recommend the case studies to others. Responses were ranked out of 5, with one being not likely and five being very likely.

Educators were asked to indicate the type of education material they were searching for that brought them to OCS and which case study materials and sections were most interesting to them from a teaching perspective. Educators who used a case study were asked questions about how they taught with the case study. They were also asked to give feedback on the case studies and their experience using the material in an open response format.

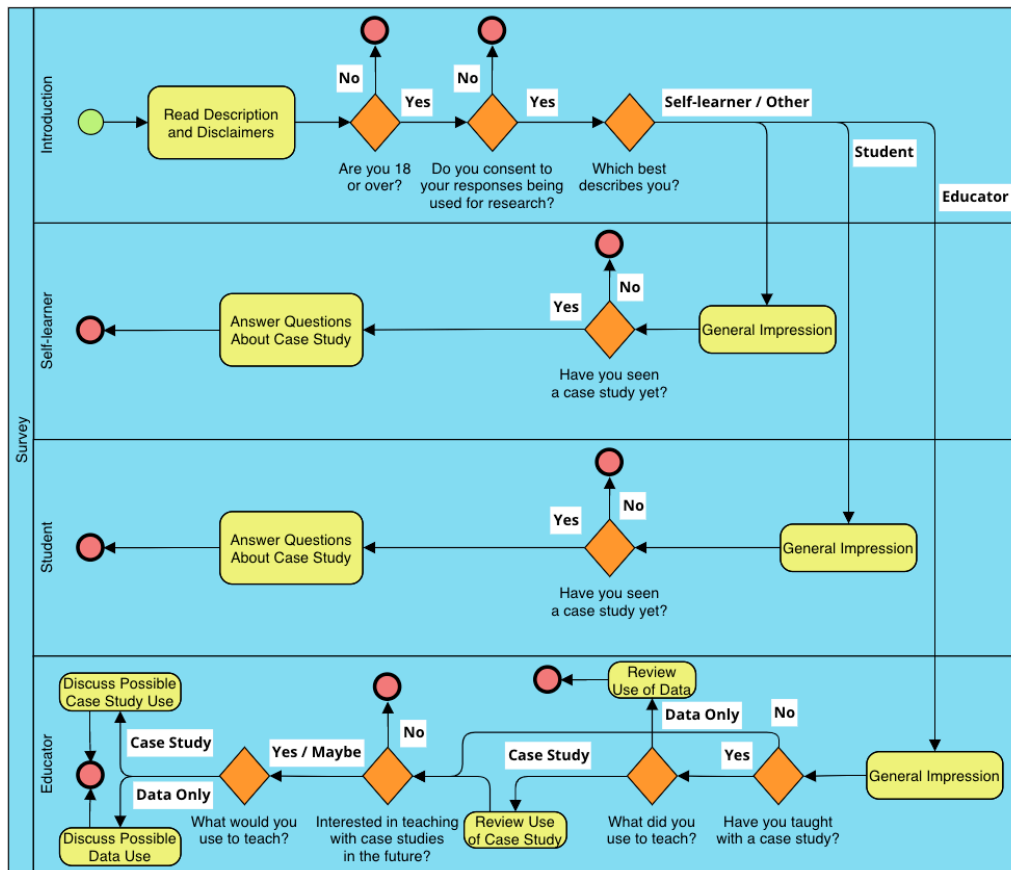


Figure 2.7: Survey Distribution Logic | The survey is structured to split into different sections depending on the user’s response to a several key questions. The logical flow of the survey is visualized here and indicates how different types of responses are distributed. The green circle indicates the beginning of the survey. Yellow rectangles indicate a task to be done by the survey participant. Orange diamonds represent exclusion gateways, where the participant’s response to the question at hand dictates the path they follow after. Red circles with a black border indicate a terminal section where participants are asked their final questions and the survey ends. The bold text in white boxes represent the possible responses to the question at hand. There are many more questions posed in the survey than shown, but only questions that split up the survey based on responses are shown here. You can view this image online on our [GitHub!](#)

To assess the types of students using OCS and how they came to it, student participants were asked to identify their academic level and their reason for using OCS. Other than the previous question, student and self-learner participants were posed with nearly identical survey sections. Since the two participant categories answered nearly the same questions and are both learning from case studies (rather than teaching with them), their responses have been grouped together in this report.

Students and self-learners were asked to rank their familiarity with different case study topics to assess what level of knowledge OCS users already have. Familiarity with a topic would also suggest experience with other education material and thus the ability to compare OCS to similar resources. They were also asked to rank their familiarity with other programming languages, R programming, and the "tidyverse" package out of five. To assess the learning needs of our student and self-learner users, these participants were asked to identify the case study topics they were interested in learning about.

Certain questions in the survey were only posed to students and self-learners who had already seen a case study. Four out of eight student participants and two out of ten self-learners indicated they had already seen a case study, making for six total participants who could answer questions about using the case studies. These participants were asked to identify which case study topics they learned something new about from reading a case study.

Students and self-learners were also asked to give feedback on their experience using OCS. These participants were asked to rank out of five the usefulness of case studies, their likeliness to refer back to the case studies, and

their enjoyment of case studies compared to other resources. To gain some insight on the backgrounds of our participants, self-learners had the option to report their field of work. See Appendix A for the survey in its entirety.

The survey data was analysed in R with the help of a few key packages. The package "googlesheets4" was used to import the survey data into R [50]. Google Forms automatically creates a spread sheet of all responses in Google Sheets, where each column is a survey question and each row is a survey participant [51]. Their responses to each question fill the cells. Using this package to extract the survey data is beneficial as it ensures that the most recent version of data is being used every time the analysis is run. Without this package, the spread sheet would need to be re-downloaded and imported every time a new response is recorded.

Once the survey data is imported into R with "googlesheets4," the responses need to be wrangled. The response data in Google Sheets comes in a format that needs to be reorganized to be easier to read and analyze. The package "tidyverse" was particularly useful for wrangling the survey data [52]. Package "magrittr" was also used throughout the analysis for the 'pipe' functionality it provides [53]. A 'pipe' in R is the symbol `%>%` and can be used in R code to make the output of one command be the input to another command. This can make code writing much more efficient and reduce the number of lines needed.

The packages "ggplot2" and "ggpubr" were used to visualize the survey data. "ggplot2" is one of the packages included in the "tidyverse" and is used for almost all graphics made in R [54]. This package was used to create nearly

all the plots presented in the Results chapter. The package "ggpubr" was used to combine multiple plots into one figure in a publication friendly format [55].

In addition to the survey, we are also measuring website and case study traffic with Google Analytics [6]. This traffic data provided insights on the breadth and depth of OCS' reach and impact. The traffic data includes number of users (new and returning), geographic location of users, pages visited, time spent on page, events (clicks, scrolls, etc), and more. This is helpful in quantifying OCS' impact on a global scale.

Google Analytics also automatically provides visualizations of the traffic data in a dashboard (see Figure 2.8). As mentioned earlier, an initial exploratory analysis of this data was helpful in identifying the need for translation of the case studies as well as indicating that some users were already translating case studies on their own. Ultimately, we will be creating our own visualizations of the traffic data, but Google Analytics provides a helpful interface for initial explorations of the data.

Google Analytics has a large list of metrics and dimensions that are included in the traffic data recorded. A list of all the metrics and dimensions available in Google Analytics can be viewed on the [Google Analytics API page](#) [56]. This page was used to identify metrics and dimensions of interest. In R, the names of the metrics and dimensions were used to extract the data of interest. See Table 2.4 for more information on the specific metrics and dimensions used for the analysis conducted in this project.

Similar to "googlesheets4," the package "googleAnalyticsR" was used to import the traffic data from Google Analytics directly into R [57]. This package

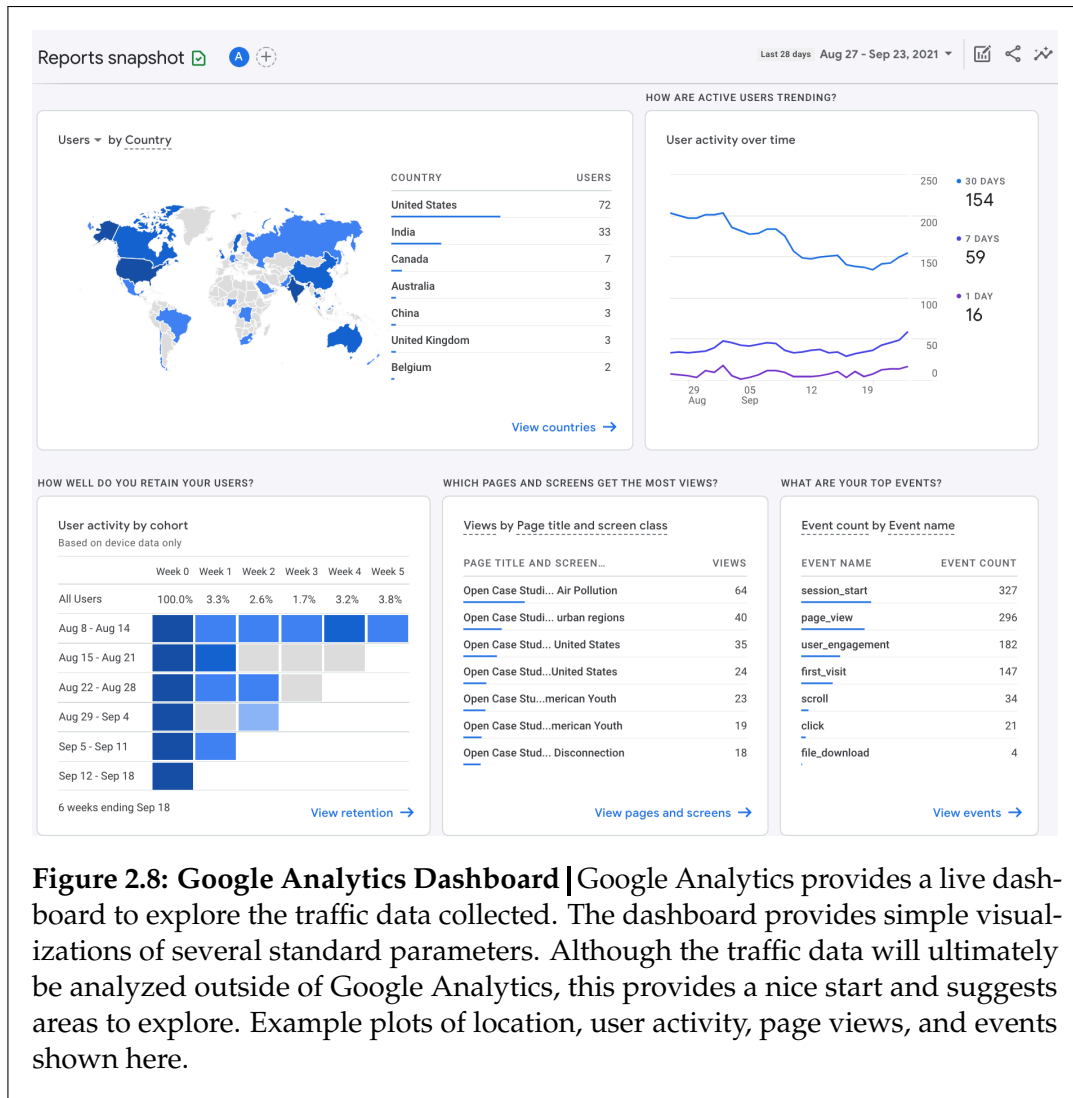


Figure 2.8: Google Analytics Dashboard | Google Analytics provides a live dashboard to explore the traffic data collected. The dashboard provides simple visualizations of several standard parameters. Although the traffic data will ultimately be analyzed outside of Google Analytics, this provides a nice start and suggests areas to explore. Example plots of location, user activity, page views, and events shown here.

Metric / Dimension	Description	Relevance
newUsers	The number of users who interacted with your site or launched your app for the first time.	Counts how many OCS visitors are visiting for the first time. Useful in tracking the organizations growth, reach and exposure.
active1DayUsers	The number of distinct active users on your site or app within a 1 day period. The 1 day period includes the last day in the report's date range.	This count provides the most accurate representation of daily total user count. Total count means both new and returning visitors. Useful in measuring case study popularity and interest.
sessions	The number of sessions that began on your site or app.	Used to compare the popularity and usage of the different case studies.
engagedSessions	The number of sessions that lasted longer than 10 seconds, or had a conversion event, or had 2 or more screen views.	A more impactful session count. Engaged sessions are of interest since they indicate the visitor interacted with the content rather than immediately leaving.
engagementRate	The percentage of engaged sessions (Engaged sessions divided by Sessions). This metric is returned as a fraction; for example, 0.7239 means 72.39% of sessions were engaged sessions.	This metric is helpful in measuring the percent of visitors that stay interested in the content when arriving to the organization webpage. It can be used to track engagement over time and assess the impact of platform updates.
fullPageUrl	The hostname, page path, and query string for web pages visited.	Used to keep track of data for different platform pages independently, including the case studies themselves. Also helpful in distinguishing static and interactive versions of case studies.
city	The city from which the user activity originated.	Provides information on where OCS is being looked at and used. Helpful in understanding the organization's global reach.

Table 2.4: Google Analytics Metrics & Dimensions Analyzed. This table provides an outline of the Google Analytics metrics and dimensions used for the analysis reported in this thesis. The first column lists the variable's name as it is listed in Google Analytics. The second column provides the variable's description from Google Analytics [56]. The third column describes why this variable is relevant to the project and how it was used in analysis.

will scrape data based on a specified date range and the input metrics and dimensions. This was very helpful in not only importing the traffic data but also filtering out data not relevant to this project.

The data from Google Analytics was also analyzed with the help of the "tidyverse", "magrittr", "ggplot2" and "ggpubr" packages. These were used for the same reasons they were used for the survey data: wrangling, pipes, and visualization, respectively. The analysis of the traffic data also enlisted the help of the "ggmap" and "leaflet" packages [58, 59]. These packages were used together to visualize OCS users on a geographical map.

Chapter 3

Results

3.1 Translate Case Studies

The case studies were effectively translated by adding the Google Translate button to the top of the case studies. This has made all of the case studies translatable to 100+ global languages [26]. The translation quality has not been officially assessed, however translations from Google Translate, although not perfect, are generally understandable. This was confirmed by fellow research assistant Qier Meng who is a fluent Chinese speaker.

The thoroughness of Google Translate was impressive. Nearly every element of the case study HTML file was translated, including the interactive exercise elements in the interactive versions (see Figure 3.1).

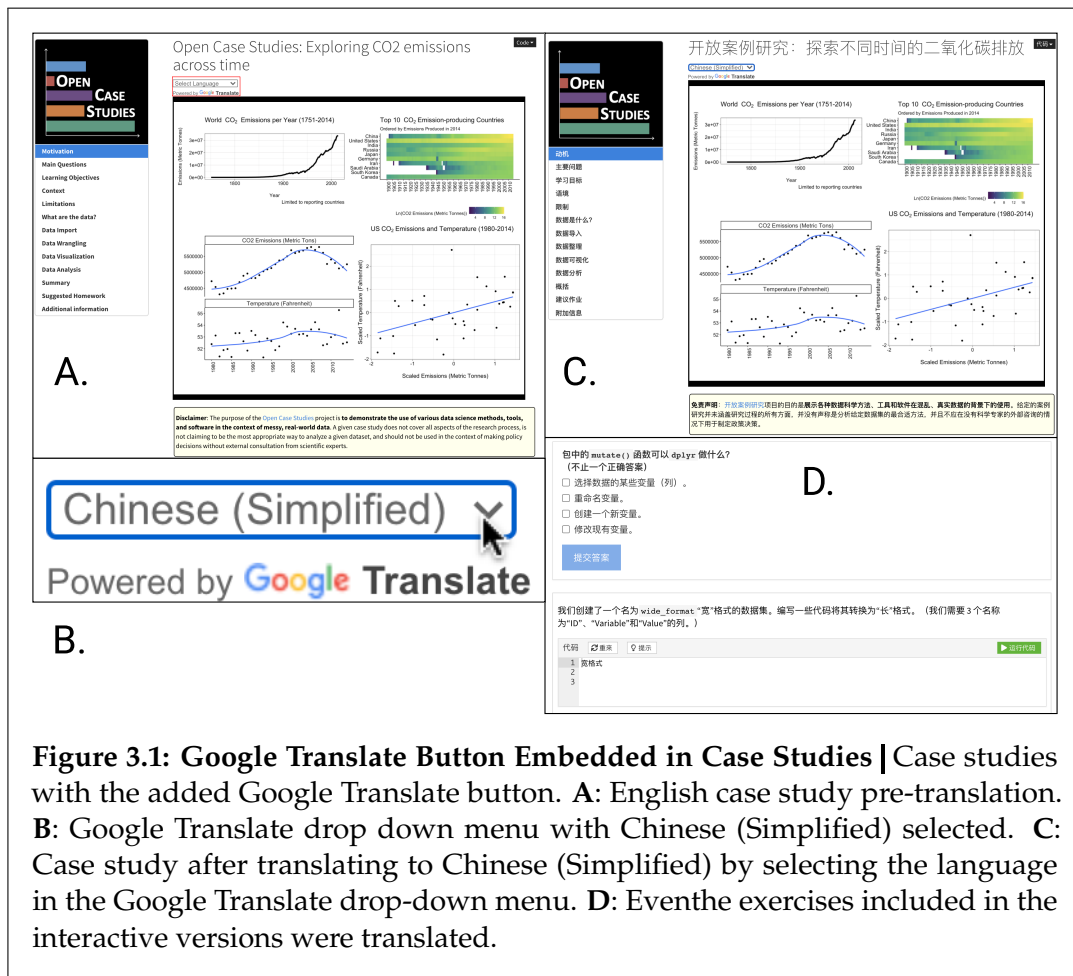


Figure 3.1: Google Translate Button Embedded in Case Studies | Case studies with the added Google Translate button. A: English case study pre-translation. **B:** Google Translate drop down menu with Chinese (Simplified) selected. **C:** Case study after translating to Chinese (Simplified) by selecting the language in the Google Translate drop-down menu. **D:** Eventhe exercises included in the interactive versions were translated.

3.2 MakeCaseStudies App

Using the prototype as a start point, a web app for individuals to develop their own digital case studies in the OCS format was developed. The final app is called MakeCaseStudies and can be found at <https://rsconnect.biostat.jhsph.edu/MakeCaseStudies/>.

The latest version of MakeCaseStudies comes with a few upgrades. The first is an update to the user interface that makes for a more pleasant appearance and experience. The second update is the addition of buttons that allow users to add and remove case study sections as needed (see Figure 3.2). The third update is the introduction of the preview tab. Now users can cycle between the create and preview tabs to view the case study being created in real time. This should allow for a much more user friendly editing process (see Figure 3.3). See Appendix B for a guide on using the app.

The MakeCaseStudies app had a measurable increase in its complexity. The prototype app only supported two case study sections. The updated app now allows for more than two and allows users to specify the number of sections they want via the new "insert" and "remove" buttons (see Figure 3.2). For further analysis of the app, traffic data for MakeCaseStudies should be tracked with Google Analytics [6]. This will allow for the success of the app to be measured quantitatively. User input should also be collected to assess the user experience. Unfortunately, there is no data to compare the new app to the prototype as the prototype was never in a state that could be officially released. Beginning to collect data on the app now will also allow for comparative analyses with future developments.

Create a Case Study

This tool is provided to help users create online lessons *quickly and easily* like our [open case studies](#), which are online step-by-step lessons that guide users through a **real-world problem solving challenge**.

Powered by:



Photo by Kari Shea on Unsplash

Start by clicking the **Make Case Study** button to download an example lesson.

↓ Make Case Study

Document Format: HTML

Delete and replace the existing content within the Create tab for your own content, check the preview in the Preview tab, and finally press the **Make Case Study** button again to download your own lesson!

Create

Preview

Logo Image URL

Title

Main Image URL

Image 2 URL

youtube video code

Insert Header

Remove Header

Insert Narrative Section

Remove Narrative Section

Figure 3.2: MakeCaseStudies: Create Tab | MakeCaseStudies is an online web application that allows users to create their own digital lessons in the OCS format. Users can simply replace the defaults in the text boxes to input their own educational content. The newest iteration of this app contains buttons at the bottom of the create tab that allow users to add or remove sections as needed. Find app at rsconnect.biostat.jhsph.edu/MakeCaseStudies.

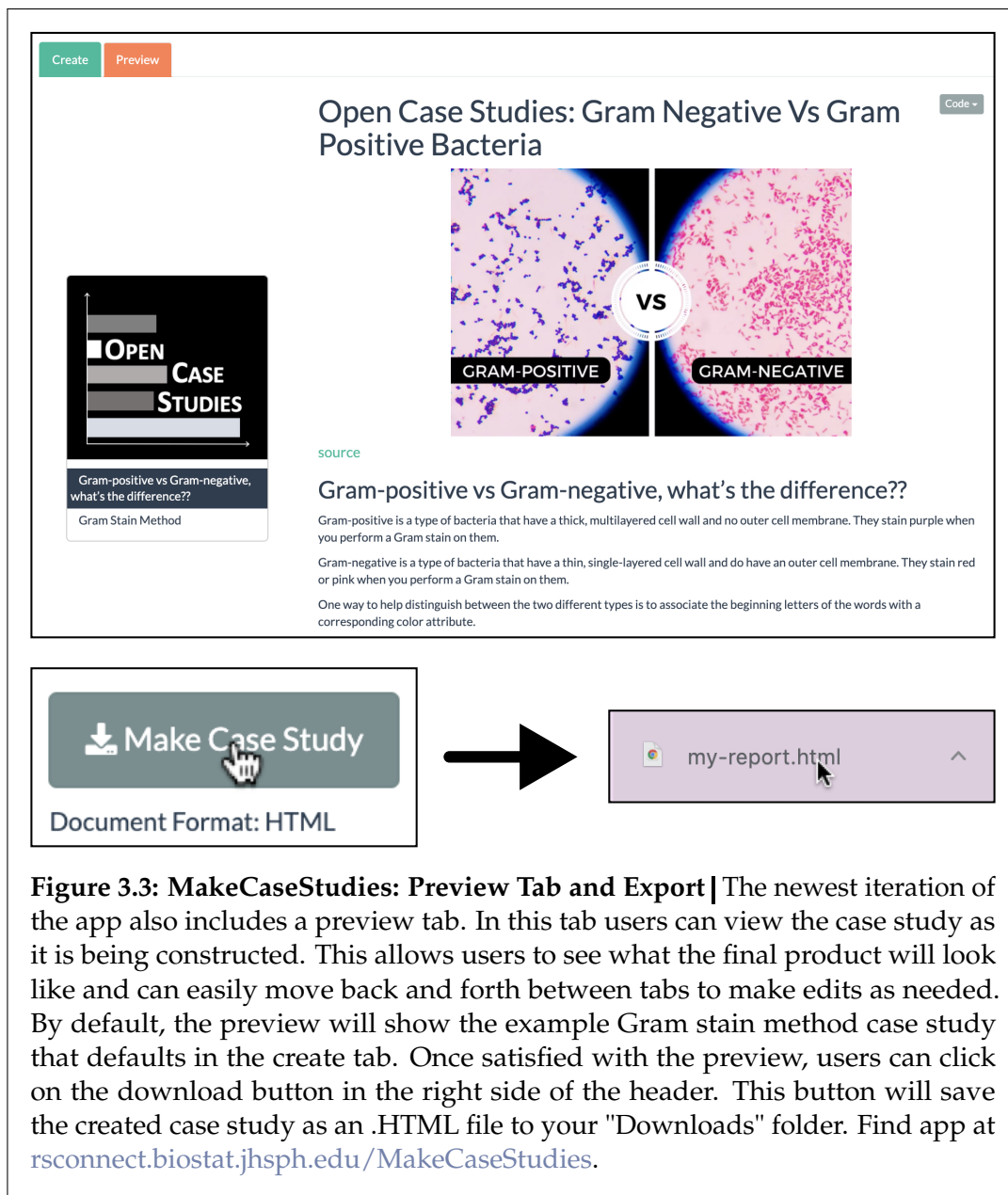


Figure 3.3: MakeCaseStudies: Preview Tab and Export | The newest iteration of the app also includes a preview tab. In this tab users can view the case study as it is being constructed. This allows users to see what the final product will look like and can easily move back and forth between tabs to make edits as needed. By default, the preview will show the example Gram stain method case study that defaults in the create tab. Once satisfied with the preview, users can click on the download button in the right side of the header. This button will save the created case study as an .HTML file to your "Downloads" folder. Find app at rsconnect.biostat.jhsph.edu/MakeCaseStudies.

3.3 Interactive Case Studies

The interactive versions of the case studies have been successfully created for 10 out of 11 total case studies. The interactive versions of the Opioids, Youth Mental Health, Obesity, Right-to-Carry Gun Laws Analysis, CO2 Emissions, and Diet case studies are now live and available for use at www.opencasestudies.org. This has been achieved thanks to the hard work of OCS research assistant Qier Meng with the supervision of OCS developer Dr. Carrie Wright.

The interactive versions include live exercises for users to check their understanding and get feedback while reading through a case study. These exercises include quizzes with multiple choice and true or false questions (see Figure 3.4) as well as free response and fill in the blank code chunks. In the code chunk exercises, users get to practice writing code using the functions and methods described in the previous section (see Figure 3.5 and Figure 3.6). The exercises provided in the interactive versions of the case study should improve user learning outcomes. See Figure 3.7 for the total number of exercises created.

3.4 OCSdata Package

An R package to improve accessibility and automate the data retrieval process was successfully developed. The package is called "OCSdata" and can be found on the OCS GitHub [20] at github.com/opencasestudies/OCSdata (see

What does the `filter()` function do?

- Selects specified columns. X
- Reorders rows of data by a specified variable. X
- Groups the data by a variable. X
- Extracts all rows that satisfy a condition. ✓

Correct!

Inside the `facet_grid()` function, `A ~ B` tells you to facet by A (columns) and B (rows).

- True
- False

Incorrect

Variable on the left side of `~` tells you to facet by rows. Variable on the right side tells you to facet by columns.

Try Again

Figure 3.4: Interaction: Multiple Choice Quiz | The interactive version of the case studies include multiple choice questions that provide feedback. If the submitted answer is incorrect, the feedback tells the user why that answer is incorrect and allows them to try again. The example questions shown are from the case study on CO2 Emissions [60].

Suppose that we have a dataset called `weight`. Write some code to plot weight change by individual. We want the plot to:

- have title as "Weight Change"
- have *x* axis as "Time_day" and *y* axis be "Weight_lb"
- be a line plot
- have different colors representing IDs 1, 2, and 3
- have legend on the right of the plot

Note: first we need to use `as.character()` to convert "ID" from number to character.

R Code Start Over Hints Run Code

```
1 weight
2
3
```

Suppose that we have a dataset called `weight`. Write some code to plot weight change by individual. We want the plot to:

- have title as "Weight Change"
- have *x* axis as "Time_day" and *y* axis be "Weight_lb"

Hints Next Hint >> Copy to Clipboard

```
1 weight$ID <- as.character(weight$ID)
```

R Code Start Over Hints Run Code

```
1 weight
2
3
```

Figure 3.5: Interaction: Code Chunk Free Response | The interactive version of the case studies allows users to apply their newfound knowledge in coding exercises. In these exercises, users are asked to practice writing their own code in a free response format. If the user is stuck, they can use the hint button to provide them the next step. The example exercise shown is from the "CO2 Emissions" case study [60].

The image shows a screenshot of an interactive coding exercise interface. It is divided into two main sections: a 'Hints' panel at the top and an 'R Code' panel at the bottom.

Hints Panel: This panel contains a code editor with the following R code:

```

1 mtcars %>%
2 # step 1
3 tibble::rownames_to_column("car")

```

At the top right of the hints panel, there is a 'Next Hint >>' button and a 'Copy to Clipboard' button.

R Code Panel: This panel contains a code editor with the following R code:

```

1 # Note there might be other ways to do this that give us the same results
2 # For practice purposes try to fill in the blanks
3 mtcars %>%
4 # step 1
5 tibble::rownames_to_column("___") %>%
6 # step 2
7 dplyr::____(____, ____, ____ ) %>%
8 # step 3
9 dplyr::mutate(mpg_group = _____(____ 16.7 ~ "___",
10 _____ 16.7 ~ "___", 21.4 ~ "___",
11 _____ 21.4 ~ "___"))

```

At the top of the R Code panel, there are buttons for 'Start Over', 'Hints', 'Run Code', and 'Submit Answer'.

Figure 3.6: Interaction: Fill in the Blank | The interactive version of the case studies allows users to apply their newfound knowledge in coding exercises. In these exercises, users are asked to practice writing their own code in a fill in the blank format. Similar to the free response exercises, users can click the hint button if they get stuck. The example exercise shown is from the "Opioids in the United States" case study [61].

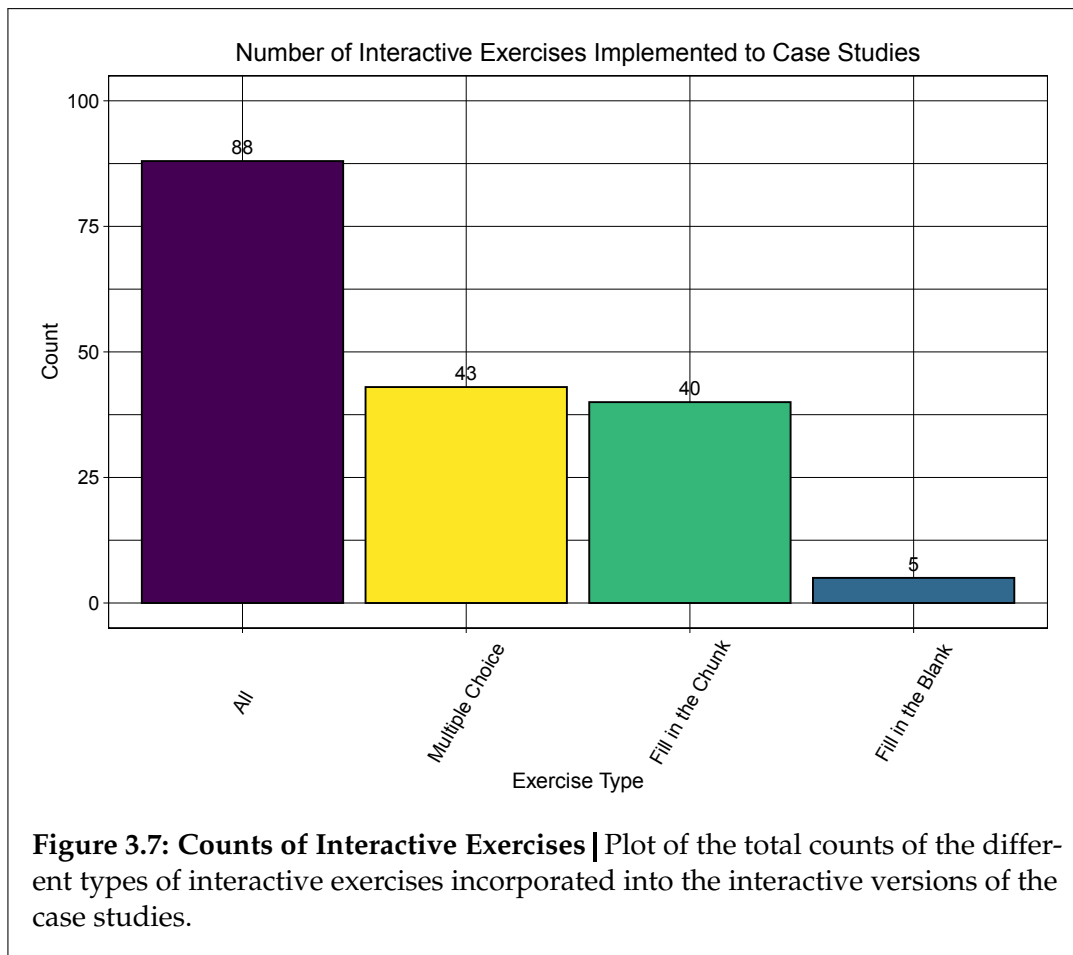


Figure 3.9). It can now also be found on the Comprehensive R Archive Network (CRAN) at CRAN.R-project.org/package=OCSdata. Providing "OCSdata" through CRAN allows users to more easily find and install the package from an official, centralized source.

Getting the package on CRAN was a very arduous process and exciting achievement for a first time package developer. The process of getting a package ready to submit to CRAN has many steps, which was no small undertaking. The most important step is to ensure the package to be submitted conforms to all CRAN policies [62]. There are many guides to CRAN submissions available for package developers, but Karl Broman's guide was particularly helpful for getting OCSdata on CRAN [63].

After many rounds of documentation updates and checking the package on the release and development versions of R as well as operating systems Windows, OS X, and Linux, OCSdata was ready for its first submission to CRAN. It's common for new packages to go through multiple rounds of submissions before receiving approval to be added to CRAN. The CRAN reviewer for OCSdata replied to the first submission with feedback on what needed to be changed.

Luckily there were only a few changes to be made, most of which were simple documentation fixes. However, there were two comments about how the package functions were not in accordance with CRAN policies. The first issue had to do with case study RDA files. The original functions loaded .RDA files directly into the users global environment. This is not allowed in CRAN policies, probably to prevent users from losing or corrupting data. This

was fixed simply by downloading the RDA files in the same way all other file formats were downloaded. Users would just need to load the RDA files into their environment themselves, which is a simple and quick process. The second issue was not solved as easily.

The other issue with the package identified by the CRAN reviewer was that the functions defaulted to saving the downloaded data files in the users current working directory. This is also disallowed by CRAN policies, again to protect users' local files. This posed a problem for OCS, as the intention for the package is to allow users to download data directly into their active project directory. This feedback was also confusing, as it was apparent that other packages were able to default to saving files in users' current working directories. With the help of RStudio's [Julia Silge](#) and using R package "textdata" [48] as a reference a solution for OCSdata was identified. Instead of defaulting to users' current working directories, users would be required to input a path to the desired download directory. If no input is provided, an interactive session would be triggered asking the user to confirm a download directory, suggesting their current working directory as an option. See Table 3.1 for a summary of how the package was updated to adhere to CRAN policies.

This process led to a cascade of fail-safes and error checks implemented in the package that were missing from the version first submitted. Overall, this process made the package code more elegant and safer for users. The final basic algorithm for any OCSdata function is as follows:

1. Check for valid case study code input. If no valid input, return appropriate error message.

CRAN Policy	Method in Breach	Investigation	Solution
Packages may not alter the user's global environment.	In the first version of the package, RDA files were directly loaded into the user's environment rather than downloaded. This would allow users to skip a step in loading the data file.	The pros and cons of skipping the import step versus releasing the package on CRAN were weighed. It was decided that adding one more step to the process of loading an RDA file was a small sacrifice to adhere to CRAN's policies.	The RDA files are downloaded using the same method as all other file types. The files are downloaded into a specified directory, ready to be imported into R by the user. Users can easily load the RDA file by either double clicking the file in RStudio or using the 'load()' function.
Packages may not default to saving files in the user's current working directory.	When user's do not provide any input on where to save the downloaded files to on their computer, the functions would default to saving the files in their current working directory.	Researched how other packages adhered to this policy by consulting with Julia Silge and investigating the methods used in the "textdata" package [48].	The package now requires users to specify a download location in the function inputs. If no location is provided, the function interactively requests input from the user, suggesting their current working directory as an option. The package functions will not download any files without input from the user.

Table 3.1: Adhering to CRAN Policy Solution. This table provides an outline of the process the package went through to adhere to CRAN policy. The CRAN policies that the first package submission was in breach of are listed in the first column. The second column discusses the methods used in the first package version that broke these policies. The third column explains what methods were used to investigate the issue at hand and research possible solutions. The final solution to the problem and the method used that was accepted by CRAN is presented in the fourth column. With these updates the package was accepted to CRAN and released accordingly.

2. Check for a download directory input. If no input:
 - (a) If the R session is interactive, ask for user input interactively, suggesting their current working directory as an option.
 - (b) If the R session is not interactive, return appropriate error message.
3. Check that the download directory input exists. If it does not exist, return appropriate error message.
4. Check that the requested data exists for the case study input. If it does not exist, return appropriate error message.
5. If all checks pass, proceed to download steps.
6. Create "OCSdata" sub-directory in the directory specified by user input.
7. Curl list of data files from input case study repository. Filter for type of data requested.
8. Use list of repository file names to download data from GitHub API [20]. Save in the previously created "OCSdata" sub-directory.
9. Once downloads are completed, return success message with the file path to the downloaded data folder.

See Figure 3.8 for a flowchart of this algorithm.

Once these changes were implemented, package OCSdata was approved and posted to CRAN on August 6th, 2021 after its second submission. To celebrate this achievement and make the package even more official, I designed an OCSdata logo using the common hex sticker format (see Figure 3.10). This

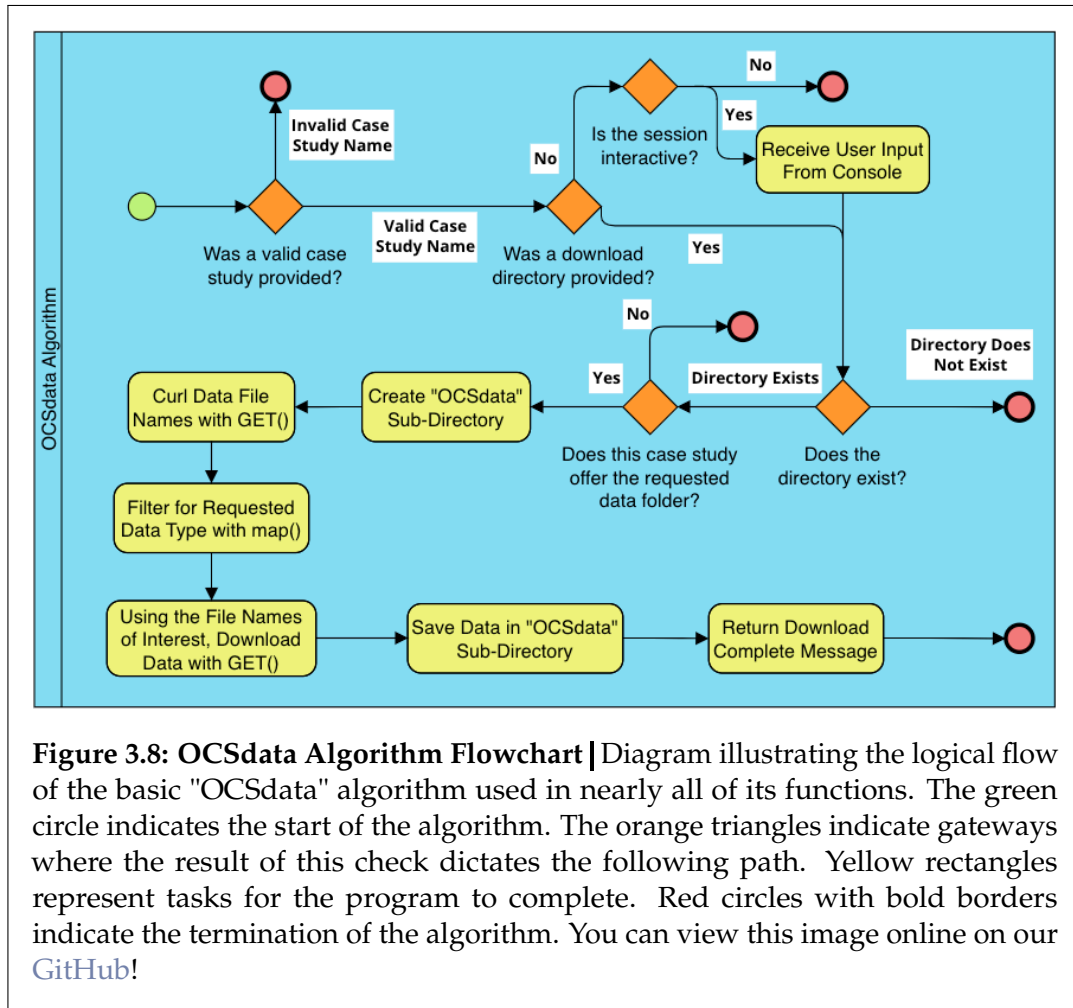


Figure 3.8: OCSdata Algorithm Flowchart | Diagram illustrating the logical flow of the basic "OCSdata" algorithm used in nearly all of its functions. The green circle indicates the start of the algorithm. The orange triangles indicate gateways where the result of this check dictates the following path. Yellow rectangles represent tasks for the program to complete. Red circles with bold borders indicate the termination of the algorithm. You can view this image online on our [GitHub!](#)

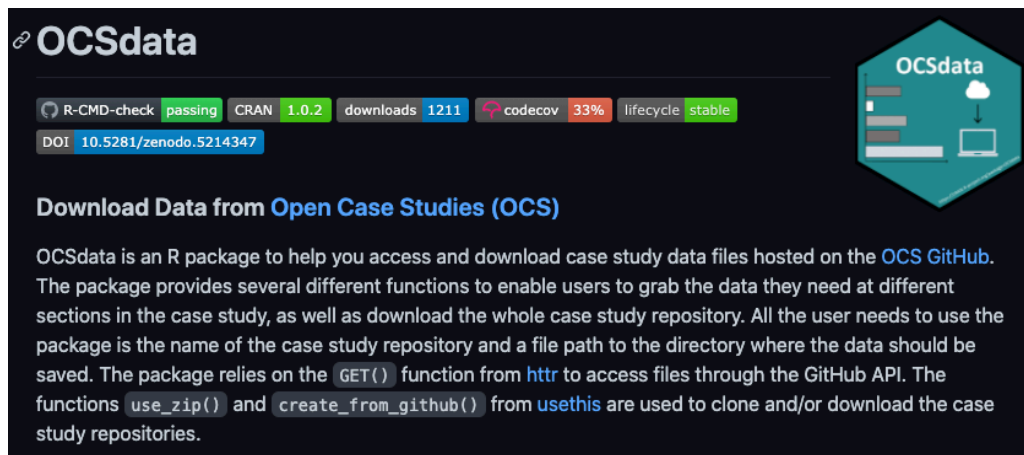


Figure 3.9: OCSdata README on GitHub | The most recent version of OCSdata can be found in the OCS GitHub repositories [20]. Here is a screenshot of the README file displayed on the repository page. The buttons indicate the results of integrated testing, current CRAN version, number of CRAN downloads, Zenodo DOI, code coverage, and package status.

was created with the help of the "hexSticker" package [64]. The sticker can also be seen in the package's README file (Figure 3.9).

The package can be installed simply using the following commands.

1. Install the Release Version from CRAN:

```
install.packages("OCSdata")
```

2. Install the Development Version from GitHub:

```
# If not already installed
install.packages("devtools")
devtools::install_github("opencasestudies/OCSdata")
```

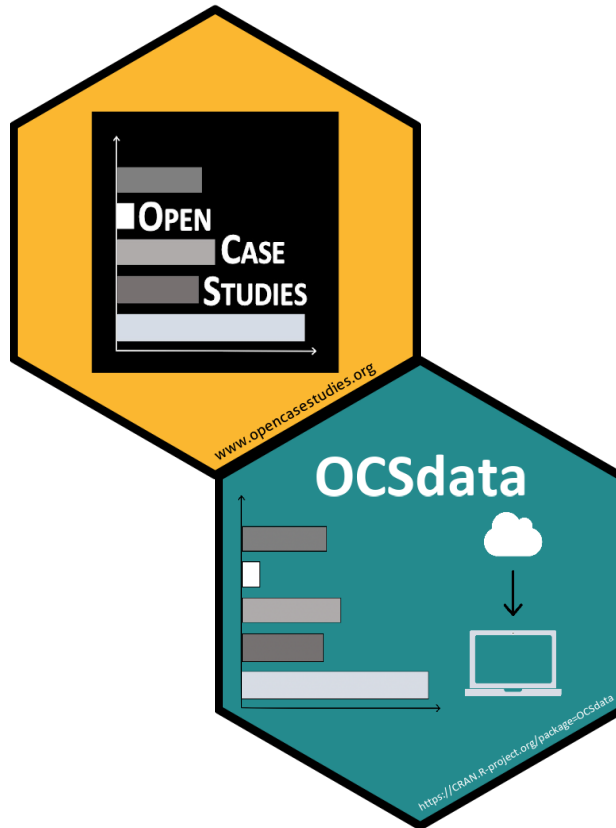



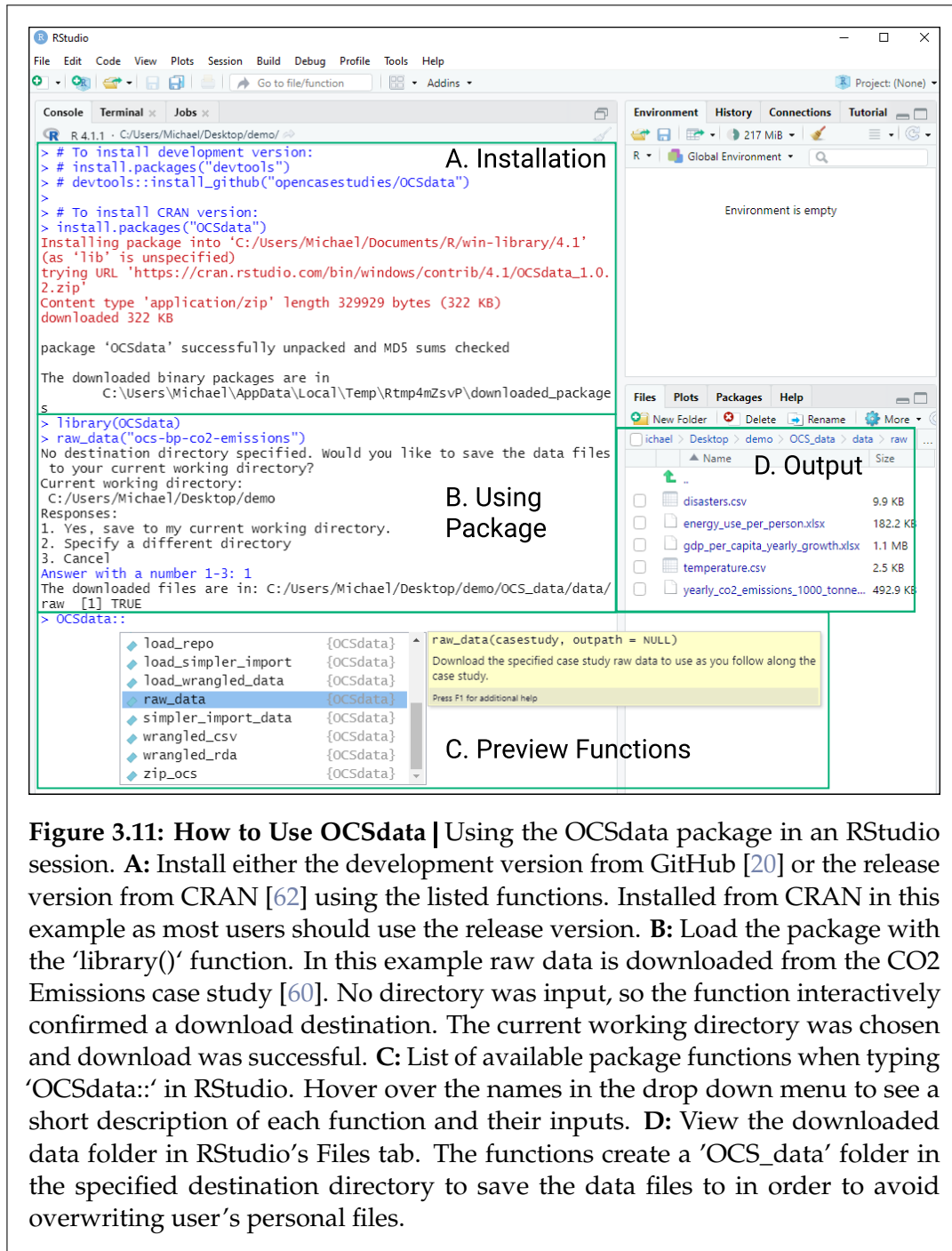
Figure 3.10: Open Case Studies Hexstickers | Hexstickers are a common format used for branding R-based projects, especially R packages. As a part of the package development process, hexstickers were made for both the OCS organization and the OCSdata package. This logo was created after the package was approved for CRAN as a fun exercise to celebrate its official release. This image can be found on [GitHub!](#) [20]

Function	Description Section	Corresponding Section
'raw_data()'	Download case study raw data	Data Import
'imported_data()'	Download raw data as premade R objects in .RDA files	Data Exploration, Data Wrangling
'wrangled_csv()'	Download pre-wrangled data in .CSV format	Data Visualization, Data Analysis
'wrangled_rda()'	Download pre-wrangled data in .RDA format	Data Visualization, Data Analysis
'simpler_import_data()'	Download raw data in formats that are simple to import	Data Import
'extra_data()'	Download extra data for further analysis	Not used in case study
'zip_ocs()'	Download case study repository as a .ZIP file	All
'clone_ocs()'	Clone and download case study repository with Git	All

Table 3.2: OCSdata Functions. OCSdata offers eight main functions that users might find useful. There is a different function for each of the different data sub-types. There are also functions to allow users to download or clone a whole case study repository.

The package comes with eight main functions of interest for OCS users. These functions enable users to work through a whole case study, use just part of a case study, and/or access case study files to repurpose for personal needs. See Table 3.2 below for their descriptions and corresponding case study sections. The basic instructions to using OCSdata functions and finding their output can be seen in Figure 3.11.

OCSdata has successfully removed barriers to case study data by simplifying the data retrieval process. Now, case study users are able to download the data they need to follow along a case study directly into their active project without leaving RStudio or their preferred R development environment. See



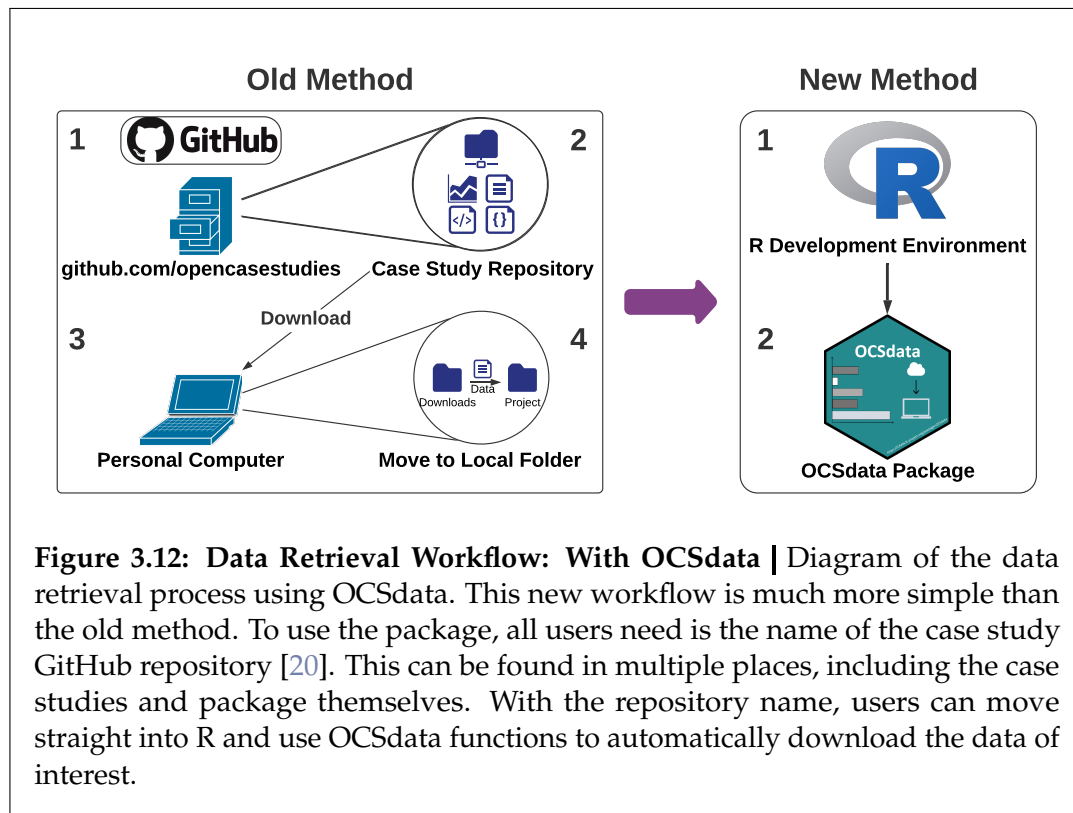
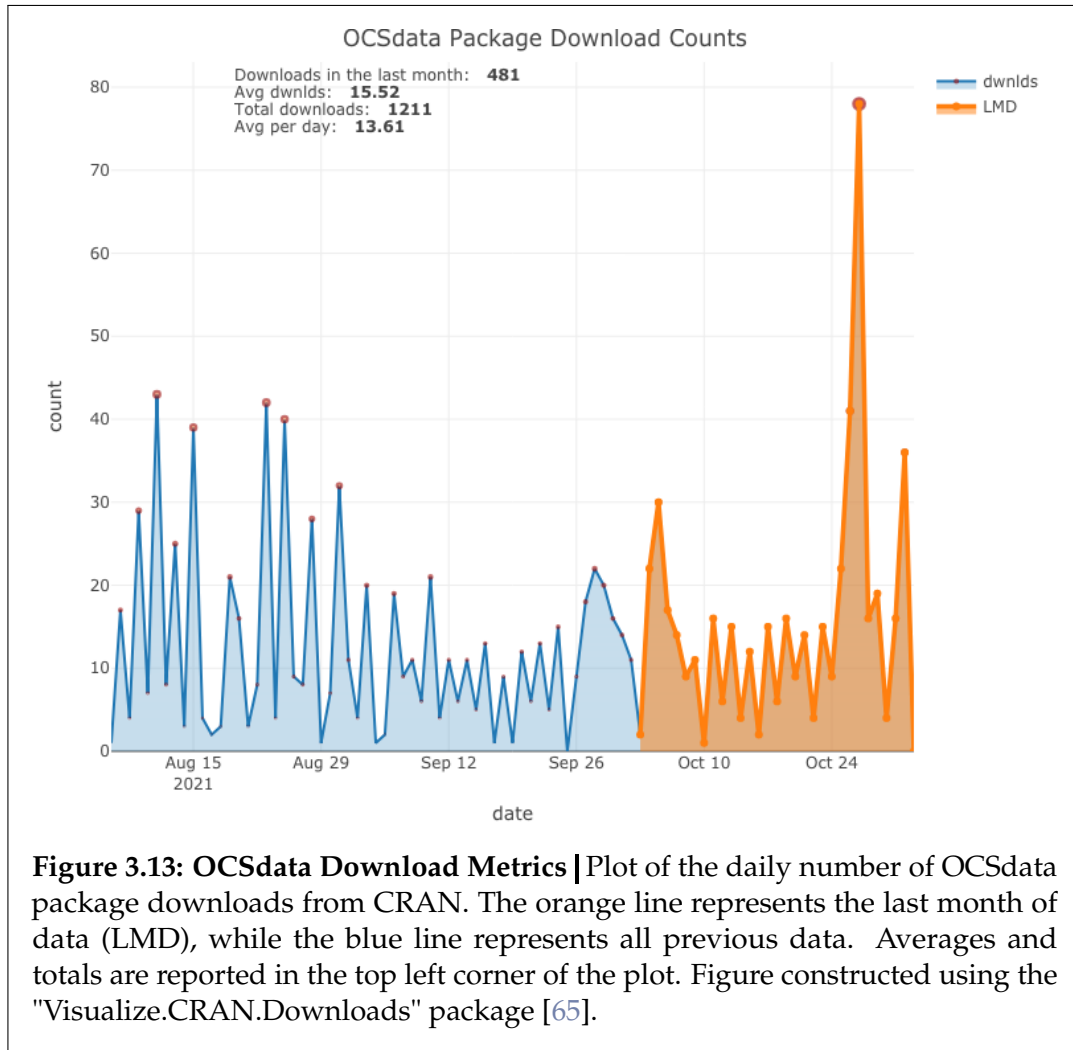


Figure 3.12 for a visualization of this new retrieval process. This should significantly reduce troubleshooting time for students, educators, and self-learners and free up more time to practice learning objectives.

The package also enables users to easily download a whole case study repository. With the package, users can get access to any case study file of interest, including the data, images, plots, even the files used to write the case studies themselves. This allows users to view our source code and edit the case study for their own personal needs without leaving R.

The number of package downloads is automatically tracked and recorded through CRAN. As of November 2nd, 2021, the package has a total of 1,211



downloads since it was first released on CRAN on August 6th, 2021. This number of downloads has exceeded our expectations and indicates that OCS users are finding the package useful. See Figure 3.13 for a visualization of daily downloads and information on total and average counts.

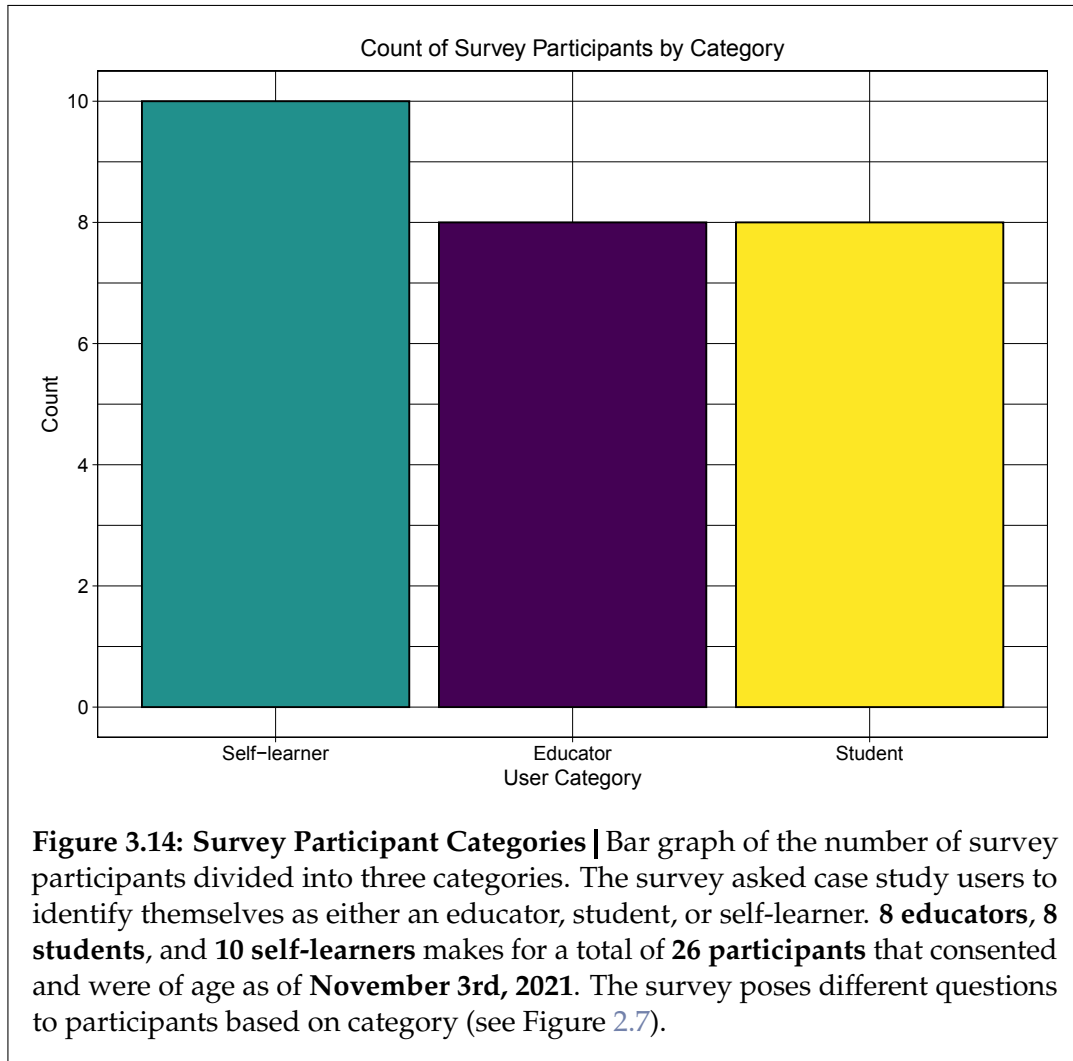
3.5 User Assessment of Case Studies: Survey Responses

The OCS survey recruited case study users to review our education content, share their experience using it, and provide feedback. The results of this survey provide a glimpse into who OCS users are, what their needs are, and how well the platform is addressing those needs.

3.5.1 Summary of Study Population

As of November 3rd, 2021 the survey has recruited a total of 29 responses. Survey participants were excluded from this study if they did not consent to their data being used for research or they were under 18 years of age. Two participants (6.9%) did not consent to data usage and one participant (3.4%) was underage. No other responses were excluded. This makes for a total of 26 (89.7%) consenting, of age participants.

Once the exclusion criteria questions were passed, the first question posed to participants was to identify themselves as an educator (someone looking for data and education material for instruction purposes), student (someone using case studies for a course or to help with a course), or self-learner (someone interested in learning more about the skills, topics, and or concepts covered in the case studies but not for an academic course). Of the 26 included survey participants, eight (30.8%) identified as educators and eight (30.8%) identified as students. One participant identified themselves using the "other" option



as a "data analyst - using data science in the corporate world." For analysis purposes, this participant was grouped into the self-learner category, making for a total of ten (38.5%) self-learners (see Figure 3.14).

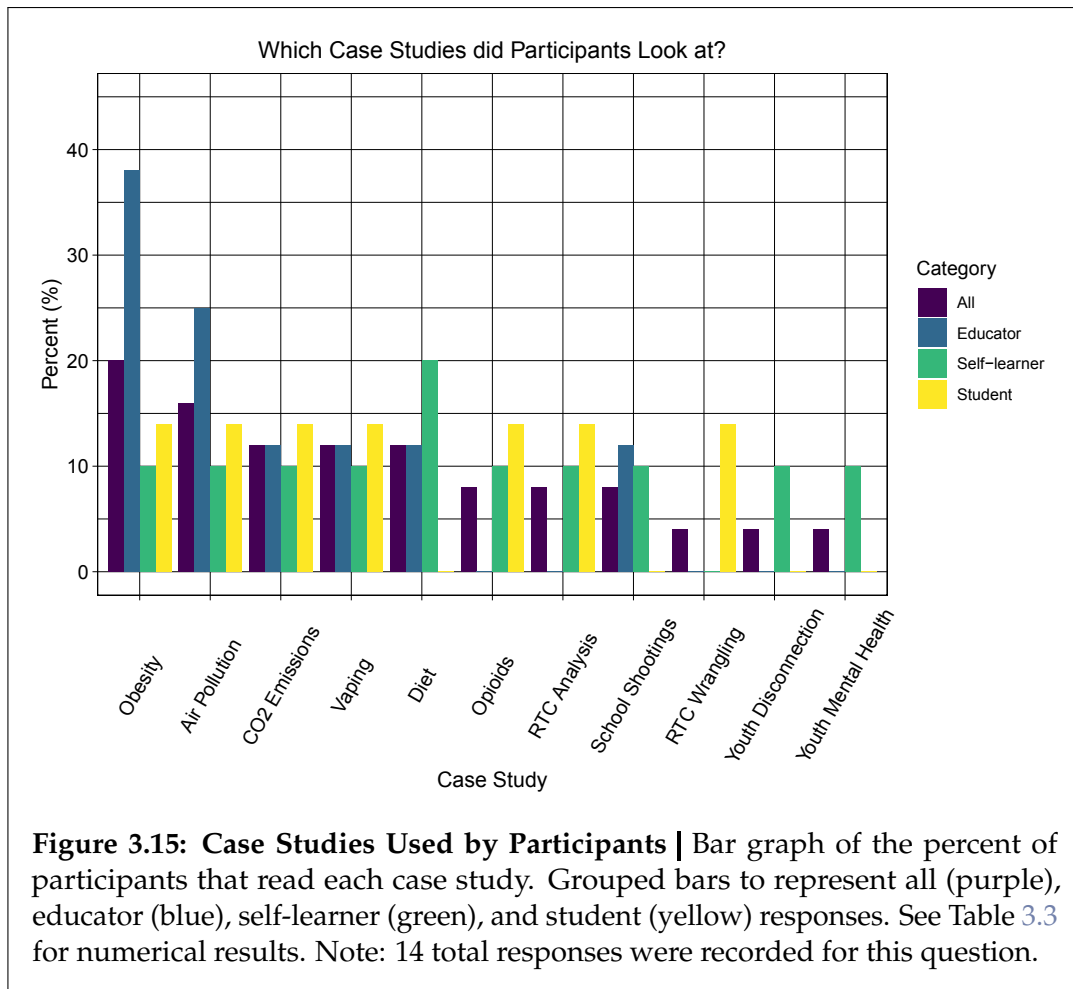
The survey gave participants different sections of questions depending on their category. There are some common questions throughout, but generally different questions were posed to assess teaching with case studies from educator participants while students and self-learners were asked questions to

Case Study	All	Educators	Self-learners	Students
Obesity	20%	38%	10%	14%
Air Pollution	16%	25%	10%	14%
CO2 Emissions	12%	12%	10%	14%
Vaping	12%	12%	10%	14%
Diet	12%	12%	20%	0%
Opioids	8%	0%	10%	14%
RTC Analysis	8%	0%	10%	14%
School Shootings	8%	12%	10%	0%
RTC Wrangling	4%	0%	0%	14%
Youth Disconnection	4%	0%	10%	0%
Youth Mental Health	4%	0%	10%	0%

Table 3.3: Numerical Results of Case Studies Used by Participants. Survey participants were asked to identify which case studies they had already looked at. This table presents the percent of participants that looked at each case study for all participants as well as each category of user.

assess learning with case studies. See Figure 2.7 for a diagram illustrating the different break points in the survey where participants were sent to different sections based on their answers to particular questions.

The survey indicated that the Obesity case study was most popular overall and for educators. The Obesity case study would be appealing to educators as it involves many standard statistical tests such as the Wilcoxon signed-rank test [1]. These tests are commonly taught in statistics courses. On the other hand, the most popular case study for self-learners was the Diet case study. Student participants were much more evenly spread across the case studies, with no clear favorite. RTC Wrangling, Youth Disconnection, and Youth Mental Health were the least popular case studies based on all survey participants. See Table 3.3 and Figure 3.15 for the numerical and graphical representations of the results, respectively.



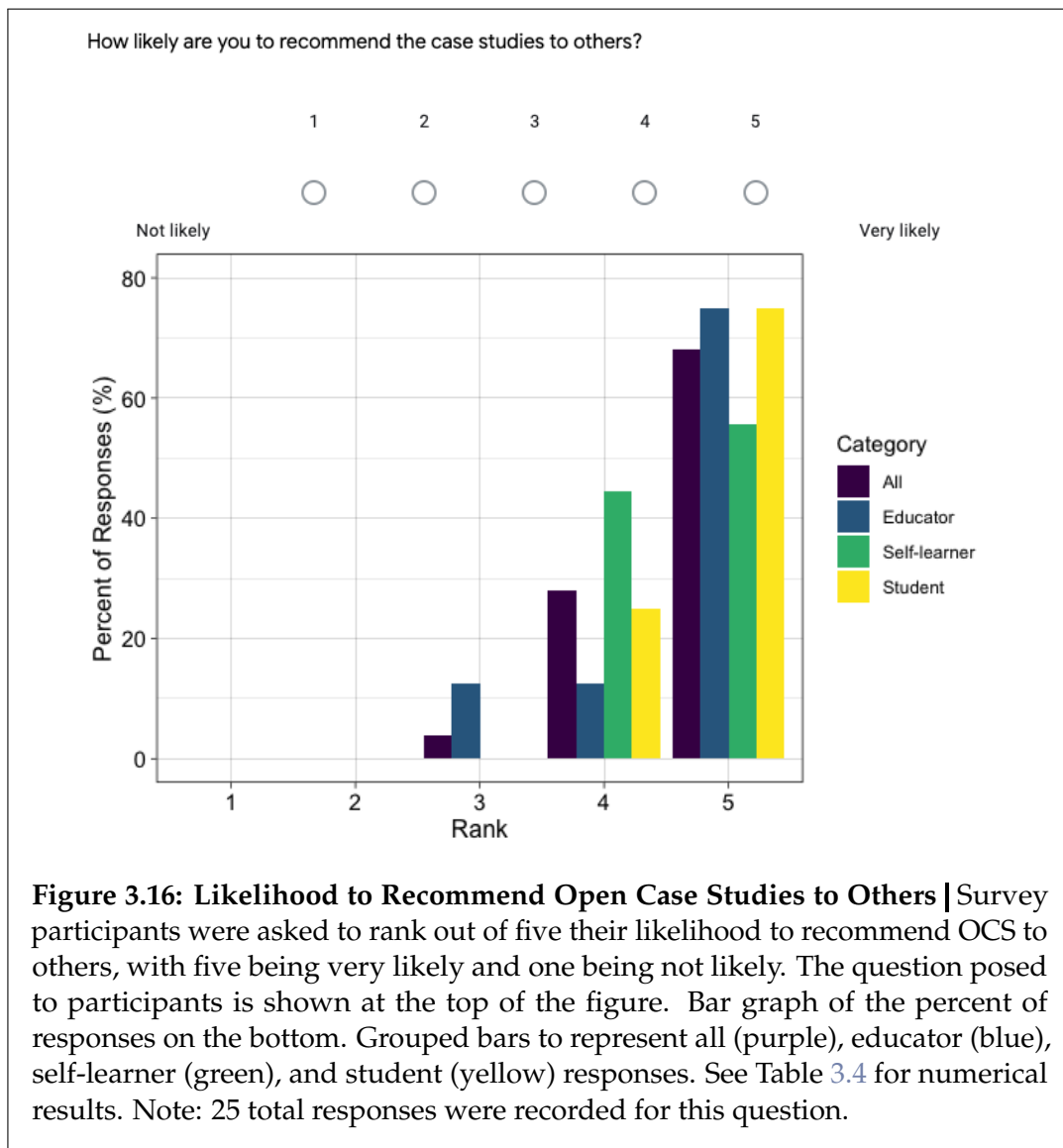
Rank	All	Educators	Self-learners	Students
1 (Not likely)	0%	0%	0%	0%
2	0%	0%	0%	0%
3	4.0%	12.5%	0%	0%
4	28.0%	12.5%	44.4%	25.0%
5 (Very likely)	68.0%	75.0%	55.6%	75.0%

Table 3.4: Numerical Results of Likelihood to Recommend to Others. Survey participants were asked to rank out of five their likelihood to recommend OCS to others, with five being very likely and one being not likely. This table presents the percent of participants that gave an answer of each option 1-5.

Overall, the most popular response from participants about their likelihood to recommend the Case Studies was five out of five, followed by a likelihood of four out of five. The lowest rank given was a likelihood of three out of five. This came from only one participant who was an educator. Rank five was the most popular answer for educators, self-learners, and students. See Table 3.4 and Figure 3.16 for the numerical and graphical representations of results, respectively.

3.5.2 Educators

To assess the types of educators using OCS, educator participants were asked to identify the types of courses and students they teach. The majority of educators taught data science and/or statistics courses, with a small portion teaching mathematics and/or public health courses. Half of educators taught undergraduate and/or graduate level students. Interestingly, three of the educators indicated teaching high school students. No educators taught middle school students. See Figure 3.17.



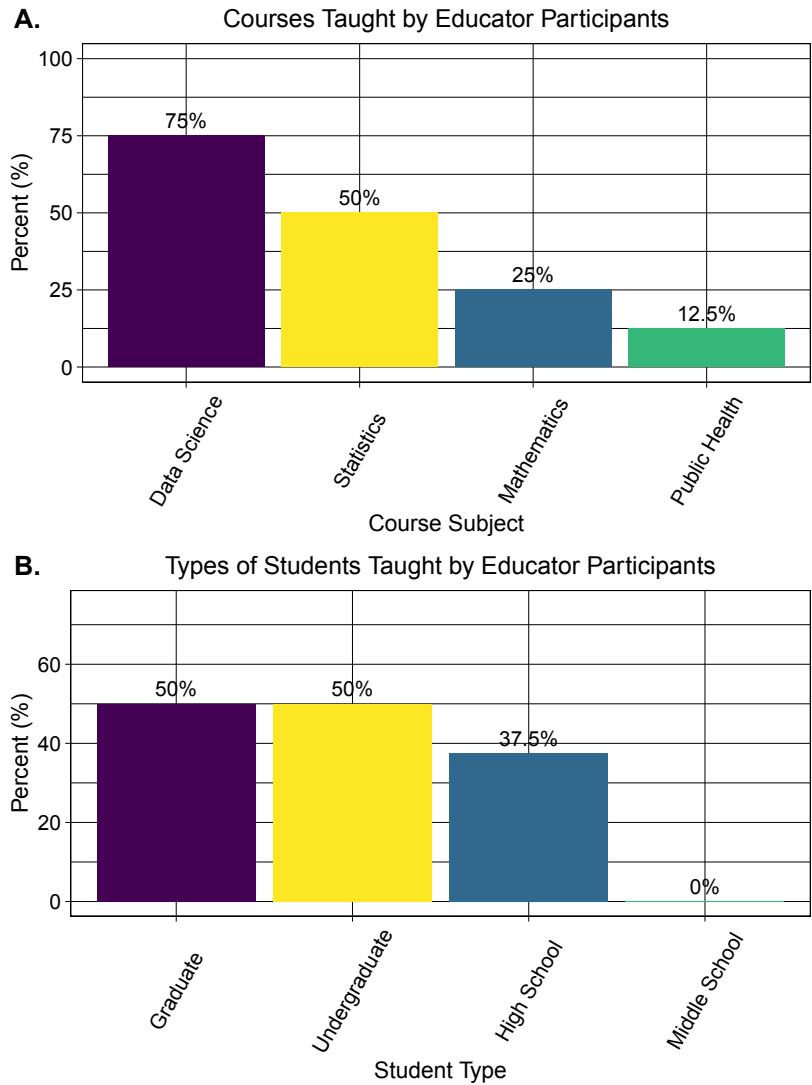


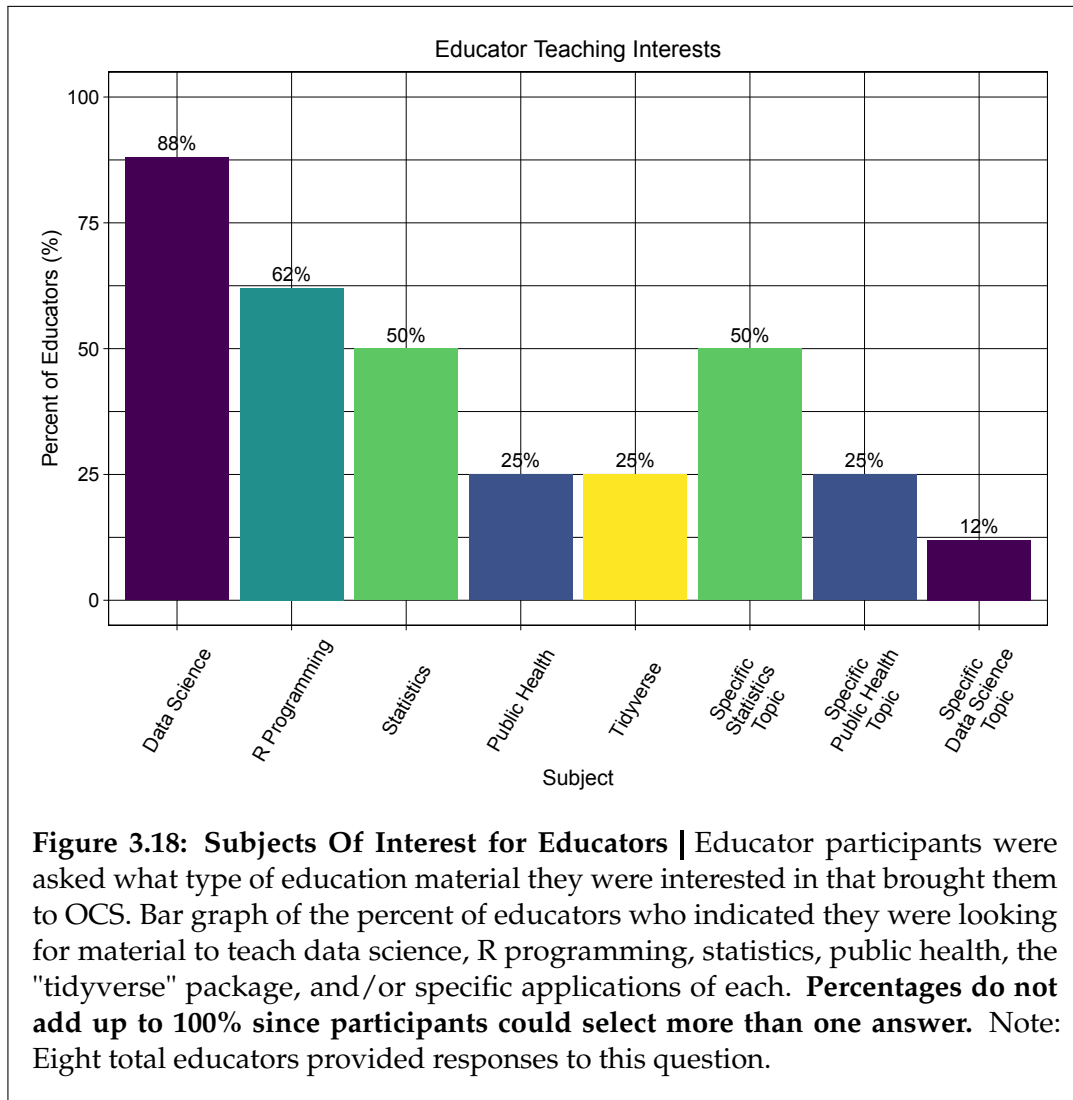
Figure 3.17: Students and Courses Taught by Educator Participants | Educator participants were asked to identify the types of courses and academic levels taught. **A.** Bar graph of the percent of educators teaching data science (purple), statistics (yellow), mathematics (blue) and public health (green) courses. Note: Seven total educators provided responses to this question. **B.** Bar graph of the percent of educators teaching graduate (purple), undergraduate (yellow), high school (blue) and middle school (green) level students. Note: Eight total educators provided responses to this question. **Percentages do not add up to 100% since participants could select more than one answer.**

Nearly all educators were interested in data science education material. Over half of educators were also interested in teaching R programming with OCS. Half of participants would teach statistics in general and/or specific statistical methods with case studies. The least amount of participants were interested in teaching just a specific data science topic. A quarter of educator participants were interested in education material on public health in general, specific public health topics, and the "tidyverse" package. See Figure 3.18.

The majority of educators indicated that they would most likely use a combination of either full case studies, parts of case studies, or just the data to teach depending on the situation. One educator was interested in using only a full case study and another educator was interested in only part of a case study. No educators were only interested in the data.

Educators had a wide range of case study sections they were interested in. A quarter of participants indicated interest in all of the case study sections. Half of participants were interested in the motivation/context and data analysis sections as well as the data itself. Data visualization was the next most popular, followed by a tie between the data exploration, import, and wrangling sections. See Figure 3.19

Out of seven responses, six educators said they were interested in using case study material in the future, while one responded maybe. Generally, educators were interested in using case studies to teach data science. Educators with more specific responses said they would use OCS to restructure their own content, introduce students to R, develop projects for high school students, and provide data for independent student work. Two educators



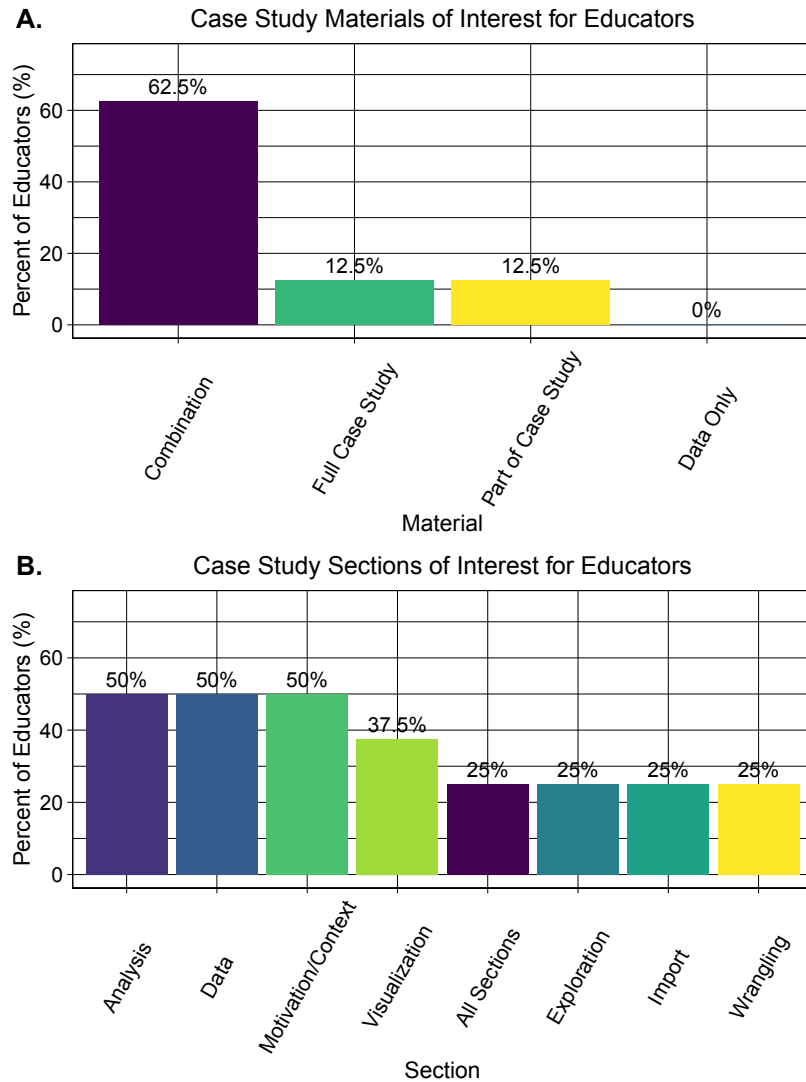


Figure 3.19: Case Study Materials & Sections Of Interest for Educators | Educator participants were asked which parts of the OCS platform they were interested in using. **A.** Bar graph of the percent of educators who were interested in data only (blue), part of a case study (yellow), a full case study (green), or a combination (purple) of all three depending on the context. **B.** Bar graph of the percent of educators who were interested in each section of the case study. **Percentages do not add up to 100% since participants could select more than one answer.** Note: Seven total educators provided responses to these questions.

indicated they were interested in using the Obesity case study to teach, while the rest weren't sure yet.

Three of the eight educator participants had already used case studies to teach when taking the survey. Two of these educators used the Air Pollution case study, while the other used Obesity. These case studies were used to teach statistics, data science, and independent studies. One educator had no issue accessing case study data, while two educators managed to find the data but with some trouble.

The following responses came from the three educators who used case studies to teach. In most cases, only one or two of participants responded to the questions in this section, making this data incomplete but still of interest.

One educator used case studies to teach linear modeling, while another assigned students to analyse data and write a report about it. Two educators used the full case study to teach and one used part of a case study. Helpful sections for these educators included the motivation, context, "What are the Data?" data import, data visualization, and data analysis sections.

Based on their feedback, educators were generally very appreciative of OCS. One educator commented "excellent resource for teaching and for students." Another reported that using our data saved them a lot of time and effort. Two educators ranked their enjoyment of teaching with the case studies. Both found it more enjoyable than using previous material. One educator found that students enjoyed learning with the case studies about the same as before, while another reported students enjoying it less. All three educators found it very clear how to use the case studies modularly. Two educators

indicated that using the case studies saved them time and effort. The same educators learned something new using the case studies and were able to incorporate something new into their teaching. Two educators indicated that student comprehension of the material was somewhat better than before and both indicated that teaching with the case studies was somewhat similar to their previous teaching method. Educators commented that the biggest benefit to the case studies was the readily available data and students' interaction with the material.

3.5.3 Students & Self-learners

Student demographics were distributed as follows: five out of eight students identified as graduate students and three identified as undergraduate students. Half of the student participants used the case studies to supplement their learning in a course. The other half used the case studies for an assignment from their course instructor. See Figure 3.20.

Student and self-learner participants were asked to choose a statement out of a list that best described their familiarity with data science, public health, and statistics. On the topic of data science, 16.7% of participants had heard briefly about the topic, 66.7% had an understanding about the topic, and 16.7% already knew all or most of the material. On the topic of public health, a third (33.3%) had heard briefly about the topic, a third had an understanding about the topic, and the remaining third already knew most of the material. Participant familiarity with statistics had the same distribution as data science. See Figure 3.21.

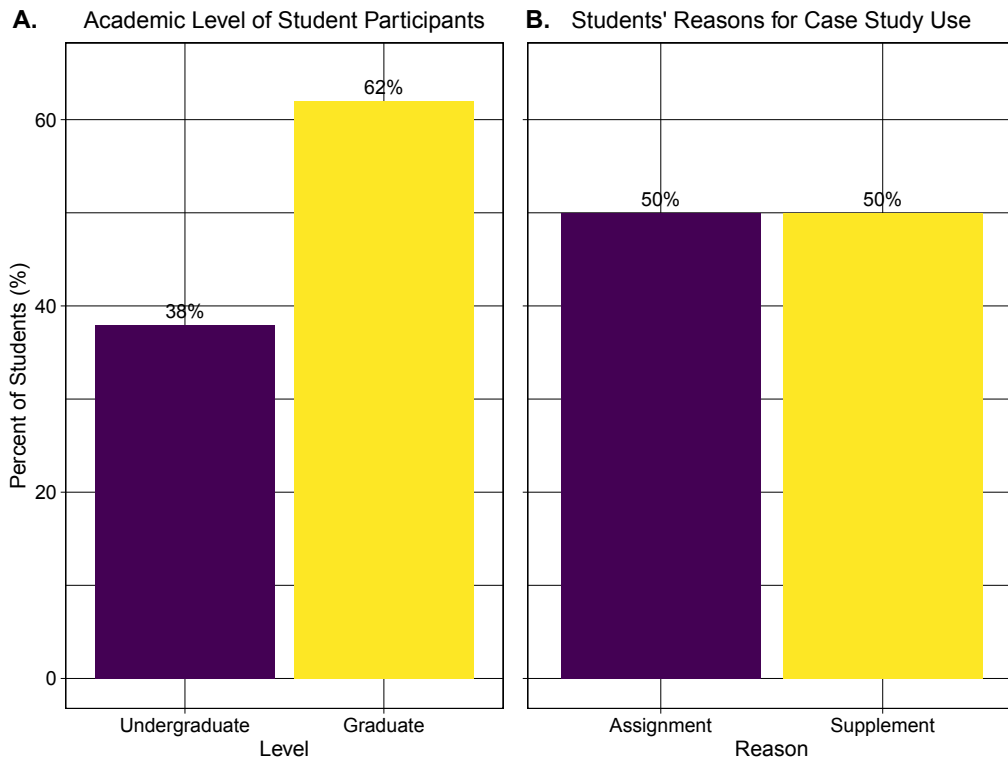
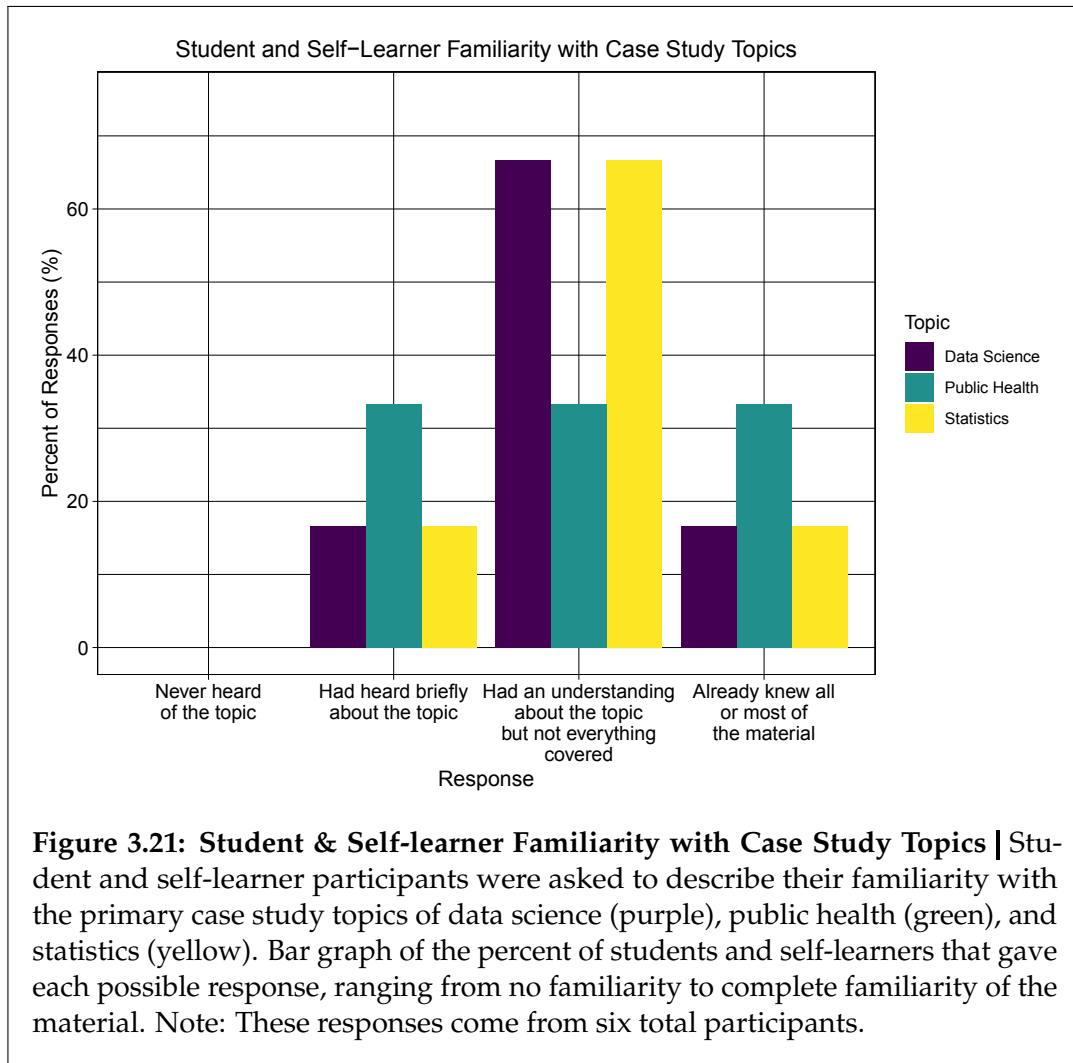


Figure 3.20: Types of Student Participants & Reason for Use | Student participants were asked to identify their current academic level as well as how they came to use OCS. **A.** Bar graph of the percent of students who identified as undergraduate (purple) or graduate (yellow) students. No students identified themselves as middle or high school students. **B.** Bar graph of the percent of student responses identifying the reason they used OCS. Assignment (purple) refers to students who were assigned to read a case study by a course instructor. Supplement (yellow) refers to students who found OCS on their own when searching for resources to supplement course instruction. Note: All student participants responded to these questions, making eight total responses.



Topic / Skill	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
Other Programming	33.3	16.7	0	0	33.3
R Programming	0	16.7	16.7	16.7	50
Tidyverse	16.7	16.7	16.7	16.7	33.3

Table 3.5: Numerical Results of Student & Self-learner Familiarity with Skills. Student and self-learner participants were asked to rank their familiarity with various case study skills. This table presents the percent of participants that gave an answer of each option on a scale 1-5. See Figure 3.22 for more information.

Most students and self-learners were very familiar (five out of five) with R programming. The participants were split on other programming languages, with 33.3% indicating no familiarity with other programming (one out of five) and 33.3% being very familiarly. A third (33.3%) of participants were very familiar with the "tidyverse" package and 16.7% had no familiarity. The rest of participants were spread in between. See Table 3.5 and Figure 3.22 for numerical and graphical representations of results.

Nearly all students and self-learners were interested in learning more about data science in general. The next most popular topics were statistics and R programming, however students and self-learners were flipped on which was most popular between the two. Specific data science topics were the fourth most popular for both students and self-learners. Students were more interested in specific statistical topics while self-learners were more interested in public health. Specific public health topics had the least amount of interest from either group. See Figure 3.23.

The majority of students and self-learners that had already read through a case study reported learning something new about statistics, data wrangling, and data communication. The smallest number learned something new about

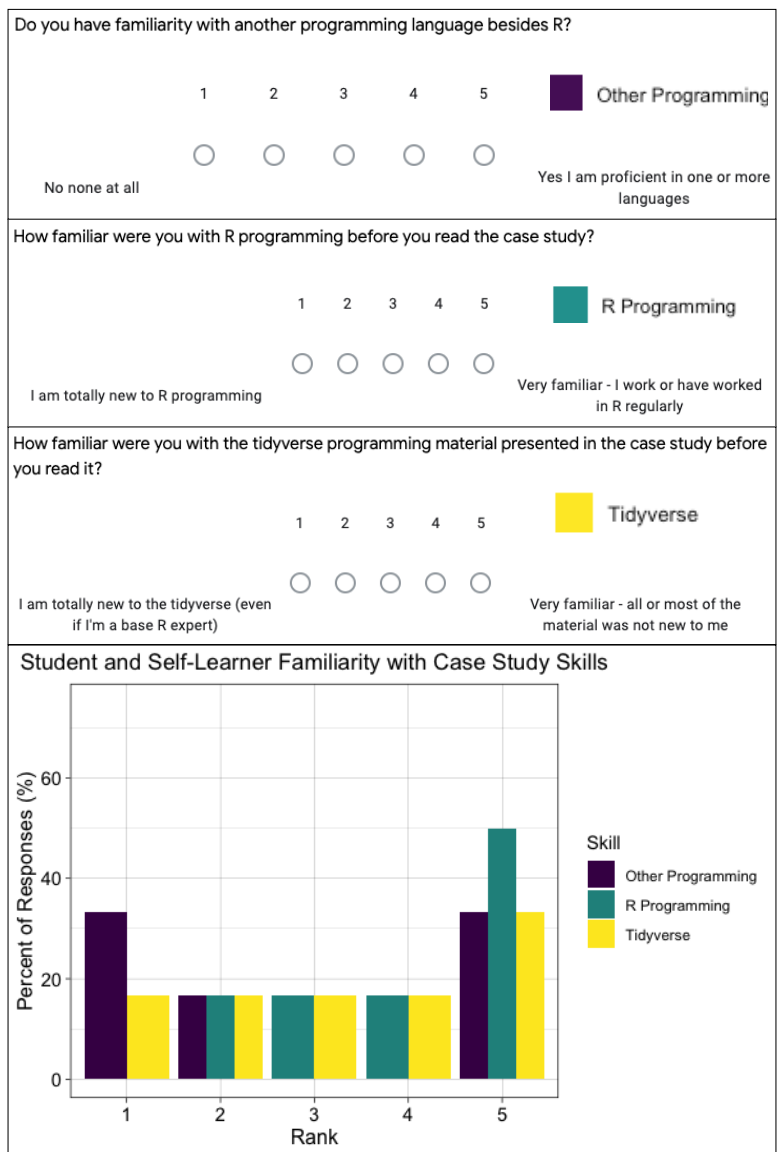


Figure 3.22: Student & Self-learner Familiarity with Case Study Skills | Student and self-learner participants were asked to describe their familiarity with some skills related to the case studies. The skills surveyed were programming languages other than R, R programming, and the "tidyverse" package. The tree images in the top half of the figure show the three questions posed to the participants. On the bottom is a bar graph of the percent of students and self-learners ranking out of five their familiarity with other programming languages (purple), R programming (green), and the "tidyverse" package (yellow). The color scheme and order of the questions matches the plot legend. See Table 3.5 for numerical results. Note: These responses come from six total participants.

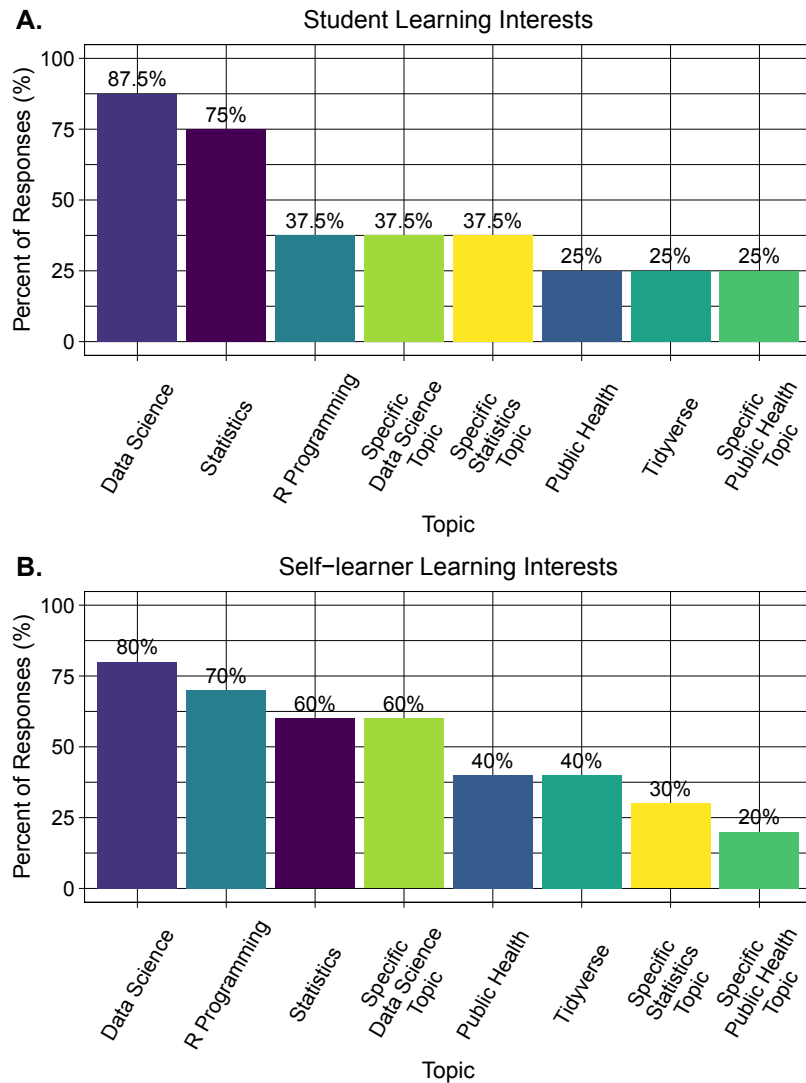


Figure 3.23: Student & Self-learner Learning Interests | Student and self-learner participants were asked to identify topics they were interested in learning about from resources like OCS. Topics include statistics, data science, public health, R programming, the "tidyverse" package, and/or specific applications of each. **A.** Bar graph of the percent of student participants who were interested in learning about each topic. Note: Eight total students responded to this question. **B.** Bar graph of the percent of self-learner participants who were interested in learning about each topic. Note: Ten total self-learners responded to this question. **Percentages do not add up to 100% since participants could select more than one answer.**

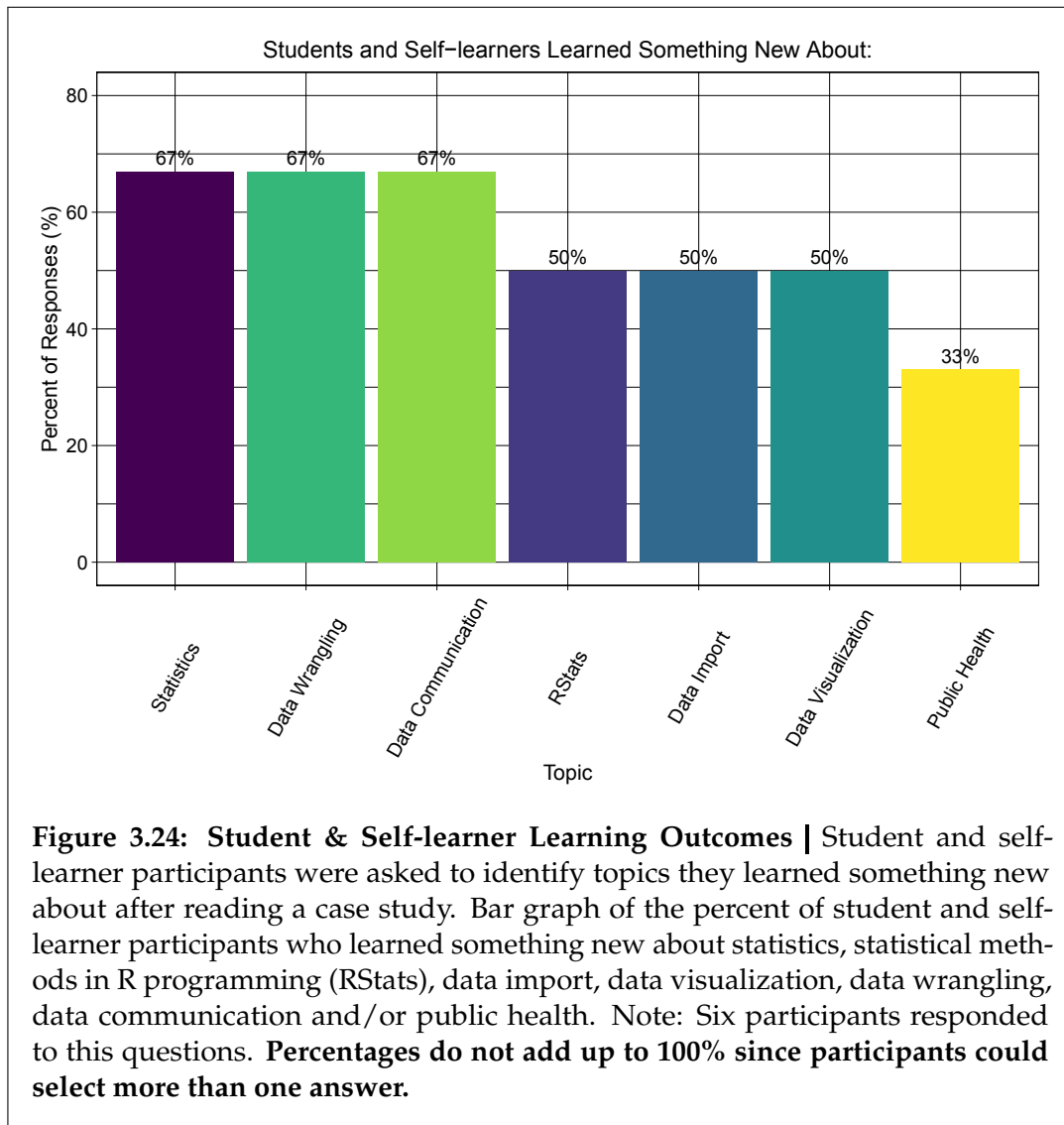
public health with 33% of students and self-learners. See Figure 3.24.

For the case studies' usefulness, all participants gave at least a rank of four out of five, with five being very useful. In total, 67% of participants gave a five out of five. Most participants gave at least a four out of five for their likeliness to refer back to case studies in the future, with five being very likely. One participant gave a response of three out of five in likeliness to refer back. Half of participants ranked their enjoyment of the case studies five out of five, with five being very enjoyable. A third of participants ranked their enjoyment 3 out of five and the remaining participants ranked it four out of five. See Figure 3.25.

The student and self-learner free response feedback sections of the survey were also very informative. All students and self-learners had no trouble accessing case study data. These participants were also very appreciative of the case studies, with one student specifically grateful for the R source code available in the case study GitHub repository [20]. One self-learner had a particularly kind message for the project:

“Open Case Studies is something I wish I had back in college! The closest I got to a resource similar to this was Kaggle, but that took a lot of digging through various formats, and individuals' projects were unreliable. Open Case Studies is intuitive, informative, and easy to access. I am excited to see where the project goes and will definitely use this for my personal research and education.”

The following list is a summary of the fields of work that self-learners came from:



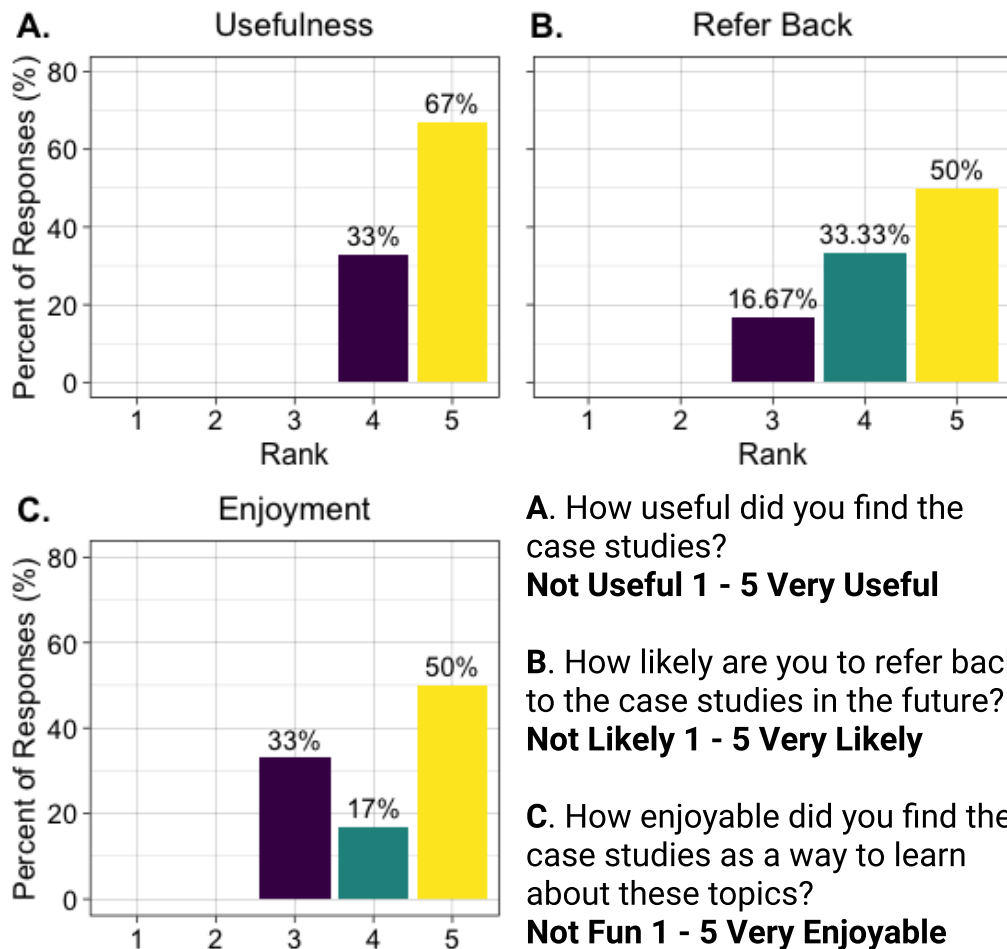


Figure 3.25: Student & Self-learner Satisfaction Review | Student and self-learner participants were asked a few questions to review the overall quality of OCS. These questions and the scale provided in the survey are shown in the bottom right. **A.** Participants were asked to rank out of 5 the usefulness of case studies. Bar graph of the percent of responses for each rank given, with one being not useful and five being very useful. **B.** Participants were asked to rank out of 5 the likelihood that they would refer back to a case study in the future. Bar graph of the percent of responses for each rank given, with one being not likely and five being very likely. **C.** Participants were asked to rank out of 5 their enjoyment of case studies as an educational resource. Bar graph of the percent of responses for each rank given, with one being not enjoyable and five being very enjoyable. Note: Six participants responded to these questions.

- Neuroscience
- Clinical Trials
- Population Genetics
- Human Resources
- Research & Development
- Psychology
- Education
- Data Science
- Analytics

3.6 Assessment of Popularity and Reach: Google Analytics Traffic Data

Tracking the OCS website traffic (including the case studies themselves) provides lots of insight into our users and their habits. Google Analytics provides several dimensions and metrics to track and analyze different aspects of website traffic. This data provides helpful insights that can indicate measurable improvements. Many of these metrics and dimensions are irrelevant (ads, sales, etc) to OCS or beyond the scope of this project, however, but the metrics relevant to the project are reported here.

Google Analytics tracks website users with several different metrics. Two of these user metrics that seemed most relevant to our study were "newUsers" and "active1DayUsers." Users who are visiting the website for the first time are considered "newUsers," while all total users visiting the website in a day are counted in "active1DayUsers." In Figure 3.26 these two metrics are plotted over time with daily counts. It can be seen that the "newUsers" count (blue) is usually just below the "active1DayUsers" count (orange), suggesting that in general, the majority of OCS daily users are new. There are some gaps of time where this is less true, and there are even some days that indicate there were more new users than total active users. But how could this be true, shouldn't all new users be counted in total active users? This comes from errors/inconsistencies in the Google Analytics tracking system. When a user visits the site from a different device or browser with the same IP address they may be incorrectly counted as a new user more than once [66]. Keeping this in consideration, Google Analytics reports 1,653 total users and 2,205 new users from January 1st, 2021 to November 3rd, 2021. This makes for a daily average of 5 total users and 7 new users.

In addition to users, the "sessions," "engagedSessions," and "engagementRate" metrics were analyzed to view the total activity from all users on our websites. A "session" begins when a user visits one of our websites (home page, case study, etc.). It becomes an "engaged" session if it has a duration time of ten seconds or more. The rate of engagement is measured with "engagementRate" which is the number of total "sessions" divided the number of total "engagedSessions." Figure 3.27 consists of two plots: one of the weekly

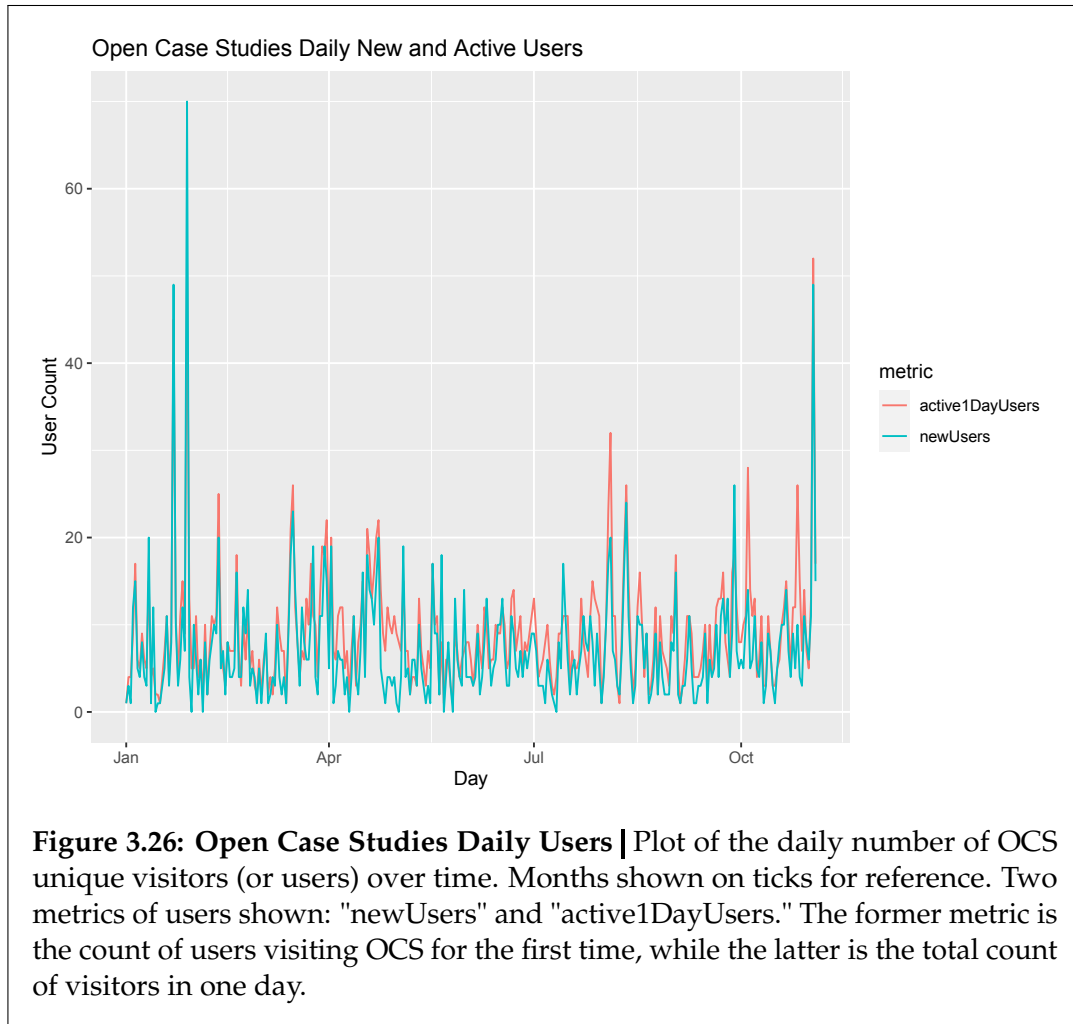


Figure 3.26: Open Case Studies Daily Users | Plot of the daily number of OCS unique visitors (or users) over time. Months shown on ticks for reference. Two metrics of users shown: "newUsers" and "active1DayUsers." The former metric is the count of users visiting OCS for the first time, while the latter is the total count of visitors in one day.

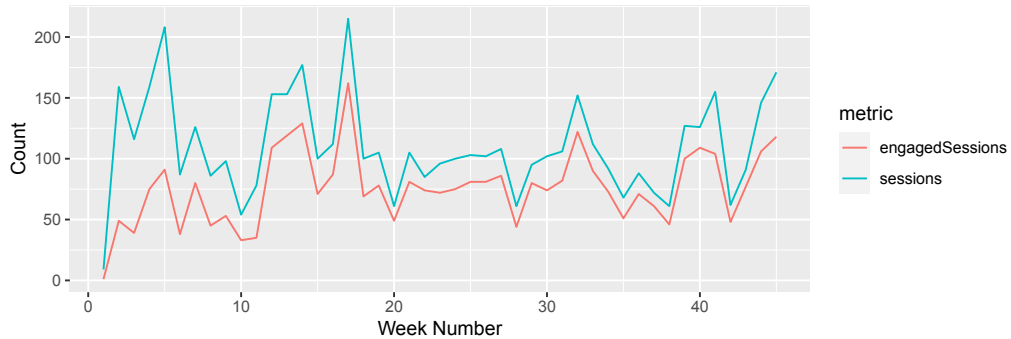
number of sessions (blue) and engaged sessions (orange), the other of the weekly average engagement rate. As the "engagedSessions" count approaches the total "sessions" count in the first plot, the "engagementRate" increases in the second plot. Google Analytics reports 4,945 total sessions, 3,420 total engaged sessions, and a 0.692 average engagement rate in total from January 1st, 2021 to November 3rd, 2021. This makes for a daily average of 16 sessions and 11 engaged sessions.

Note how the data in plot A of Figure 3.27 drops off in the first and last week. Google Analytics defines a week as beginning on Sunday [6]. The data presented is recorded from January 1st, 2021, a Friday, to November 3rd, 2021, a Wednesday. This makes it so the first and last week start and end before Sunday and results in an artificial decrease of the session counts in the tails of the data.

Google Analytics also allows us to track traffic data for each specific case study. To view which case studies were the most popular, total session counts for each case study were compared. Although the interactive versions have been online for much less time than the static versions, traffic data was collected for the interactive case studies that were made live during this project. Figure 3.28 display visualizations of both total session counts for all case studies and the interactive versions.

One of the more intriguing dimensions that Google Analytics tracks is user location. This data gives an important view into the global reach of OCS. To visualize this data, a bubble map was created where each bubble corresponds to a city where OCS recorded visits from. The size and color of

A. Open Case Studies Weekly Sessions and Engaged Sessions



B. Open Case Studies Weekly Engagement Rate

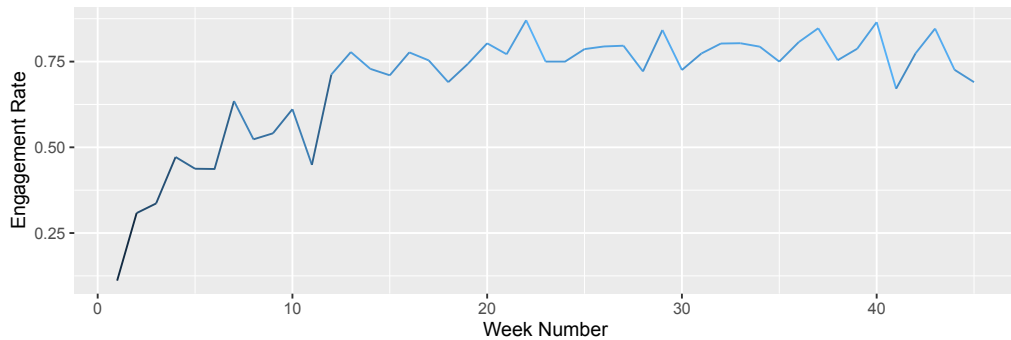
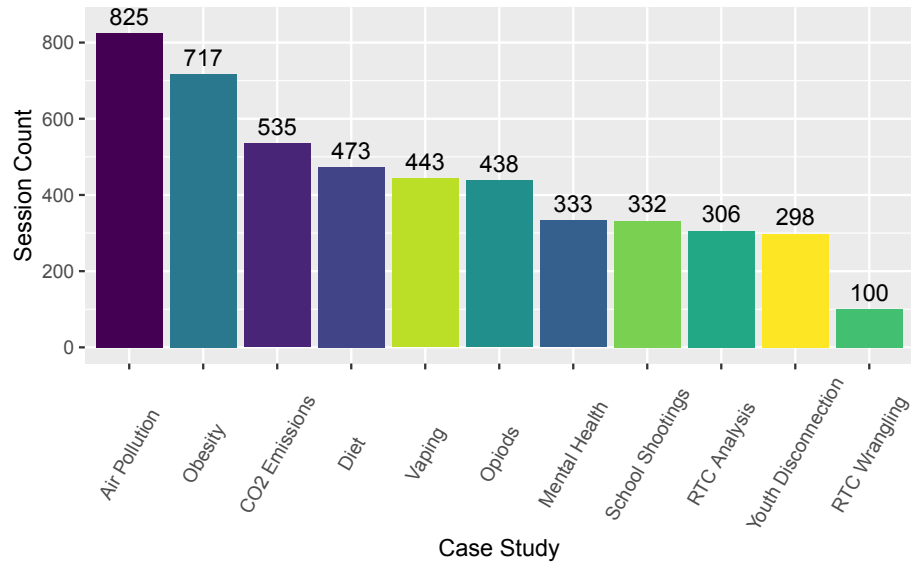


Figure 3.27: Open Case Studies Weekly Engagement | Visualizations of OCS engagement over time. Week zero on the horizontal axis starts on January 1st, 2021 and the last week of data shown ends on November 3rd, 2021. **A.** Plot of the weekly number of sessions over time. Two metrics shown: "sessions" and "engagedSessions." The former metric is the total number of sessions per week, while the latter is the total number of engaged sessions per week. A session begins when a user opens an OCS website. A session becomes engaged when it lasts longer than 10 seconds or has two or more screen views. Counts drop off in the first and last week on the plot because these "weeks" account for less than seven days of data **B.** Plot of the weekly average engagement rate. Engagement rate is the percentage of total sessions that were engaged.

A. Number of Total Sessions by Case Study



B. Number of Total Sessions by Interactive Case Study

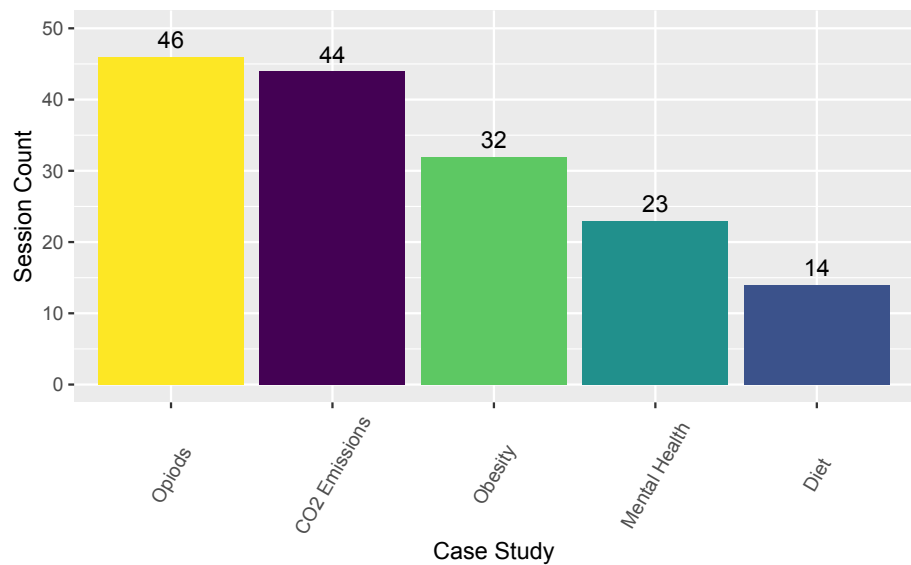


Figure 3.28: Total Case Study Sessions | Total sessions recorded on case study websites with Google Analytics. **A.** Bar graph of the number of sessions recorded for each case study including data for both static and interactive versions. **B.** Bar graph of the total number of sessions recorded for each interactive version of the case studies. Not all interactive case studies have been posted yet, so there are less case studies listed here than in **A.** Numerical results for both plots are shown in the figure above the corresponding bars.

the bubbles correspond to the total number of sessions coming from the city in question (see Figure 3.29). This map shows that the case studies have been used across the globe on a total of six continents (all except Antarctica). The largest bubble on the map corresponds to Baltimore, Maryland, USA. This is likely because Johns Hopkins University, the institution where OCS was developed, is located in Baltimore.

In Figure 3.30, the top ten cities with the most OCS sessions were plotted in a bar graph to make clear which cities were using OCS the most. Again, Baltimore is clearly the most popular for the same reason as explained previously. Seattle is the second most popular which may also be in part due to this author working remotely from Seattle. The other cities on the list, however, have no such explanations, with three of the top five cities coming from outside of the United States.

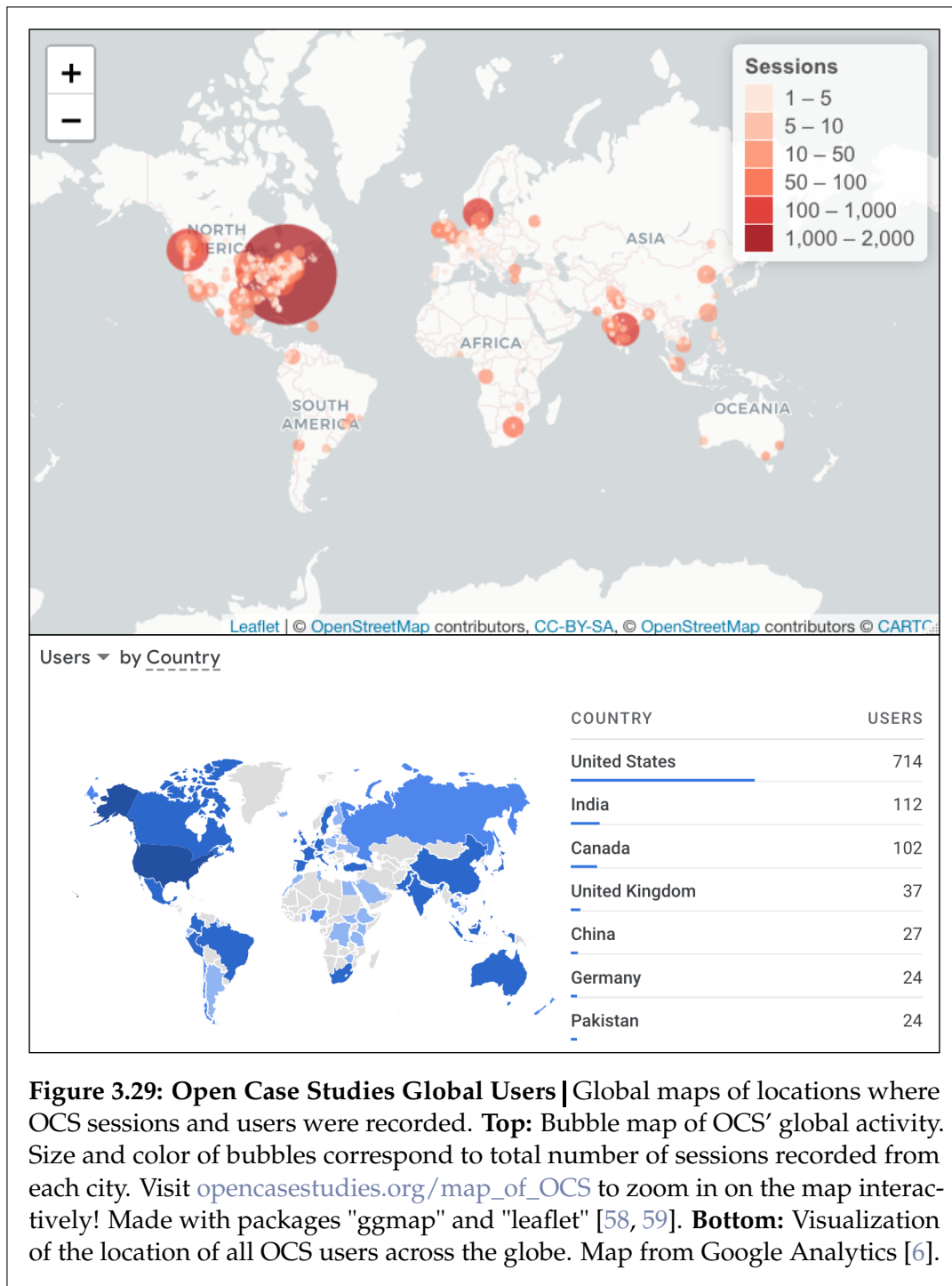


Figure 3.29: Open Case Studies Global Users | Global maps of locations where OCS sessions and users were recorded. **Top:** Bubble map of OCS' global activity. Size and color of bubbles correspond to total number of sessions recorded from each city. Visit opencasestudies.org/map_of_OCS to zoom in on the map interactively! Made with packages "ggmap" and "leaflet" [58, 59]. **Bottom:** Visualization of the location of all OCS users across the globe. Map from Google Analytics [6].

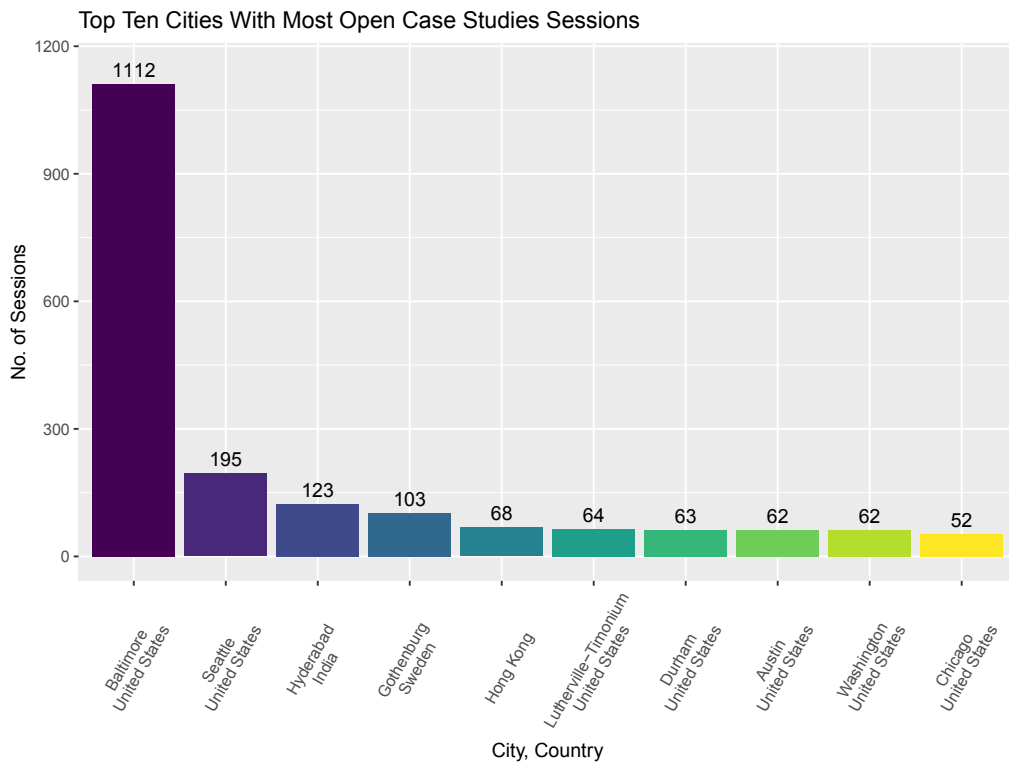


Figure 3.30: Top Ten Cities Using Open Case Studies | Bar graph of the 10 cities with the most total sessions recorded. Baltimore is significantly more popular than others, likely due to the fact that the OCS developers' home institution is located in the city.

Chapter 4

Discussion and Conclusion

Improvements to OCS are ongoing and this report provides a snapshot of the work today. This is leading to future work in translation capabilities, interactive options, the MakeCaseStudies app, the OCSdata package, accessibility and more. Such future directions are discussed in more detail in the following sections.

4.1 Translate Case Studies

The case studies translated by Google Translate are not perfect, but an important step in the right direction. Based on a literature review, the translation accuracy varies from language to language. The most accurate translations are for Latin based languages such as Spanish, and the least accurate translations are for languages with alphabets significantly different than English such as

Japanese [67]. While many OCS users will be served well by Google Translate, the translation quality for some will be less than ideal. According to Google Analytics, India, China and Pakistan are all countries with some of the most OCS users globally (see Figure 3.29). These users deserve quality translations of the case studies that Google Translate currently cannot offer. One benefit of Google Translate, however, is that it constantly continues to improve its translations as it gathers more data and algorithms improve.

In the mean time, OCS should conduct an official investigation into the quality of translations for the case studies. This should be done at least for the most popular languages as indicated by Google Analytics. Based on this investigation, methods should be implemented to improve the translations as needed. The initial idea to develop a software package specifically to translate case studies might be considered again, however this was out of scope for this project. Nevertheless, the translate function, as is, provides significantly more accessibility to non-English speakers than before. The case studies can also be translated into over 100 languages, drastically more than any number imagined in this project's proposal. This addition to the case studies is an important development.

4.2 MakeCaseStudies App

The MakeCaseStudies web-application is a basic tool that can be used to create custom case studies in the OCS format. This project improved upon a previous prototype by redesigning the user interface, adding a preview

tab, and implementing a function to add and remove case study sections. One limitation, however, is its lack of flexibility. The prototype app had no flexibility, limiting users to a format consisting of two main sections, one image, and one video.

This was addressed somewhat in the most recent version of the app with the "insert/remove" buttons. However, there are still limitations to this function. There is a maximum number of sections the users can add. Right now, this is set to 20, which should be enough for most users. The maximum number would have been set higher to cover any possibility, but this causes problems with the case studies' table of contents. Currently, the table of contents includes all possible sections available to the user, even if the section is left blank or wasn't used. This is another problem that needs to be addressed, but for now the maximum number of sections has to be a reasonable number to preserve the table of contents. Additionally, users are still only able to add or remove text sections and headings as needed. A function to add extra videos and images has yet to be developed.

The web-app is imperfect as is, but requires only a few tweaks to be optimal for OCS users. Until then, the app is still helpful for users looking to make short to medium length digital case studies in the OCS format. In the survey, one educator responded that they were looking to restructure their own content into the OCS format. MakeCaseStudies will be a very useful tool for this educator and others like them. Future work should assess the quality and popularity of the app with user feedback surveys and website traffic data.

4.3 Interactive Case Studies

The interactive versions of the case studies are a major development. These versions will make the case study experience more engaging and should improve user learning outcomes. These will also be useful for educators in need of practical exercises for students to do either in class or at home. The importance of this development is evident from the survey results, as one educator responded that the "interaction of students with the material" was the biggest benefit of OCS. Publishing the rest of the interactive case studies will be an impactful development for the OCS platform.

Future work on the case studies themselves could take this interactivity even further. One idea is to create interactive versions of the preview plots in the introduction of a case study. Current case studies offer a preview of the case study at the very beginning with a figure of a few of the plots made with the case study's data. Making these plots interactive would allow users to explore a case study even further before reading it. Ideally, these plots could come with a control panel that allows users to select different kinds of data, metrics, and plots to preview the full range of what can be achieved in the case study. This addition would make for a more engaging introduction to the case study as well as help users decide which case study is best for their needs.

Another idea for future development of the case studies' interactivity is to implement a text-to-speech option. Giving users the option to listen to case studies as well as read them would improve accessibility for visually impaired users. In theory, this would also allow any user to immerse themselves in the

material as if it were a recorded lecture. Expanding the ways users can interact with the case studies will improve user experience, reduce barriers, and widen OCS' reach. This would be an impactful feature by providing access to those who may have been previously excluded from similar education material. R package "ari" [23] would likely be helpful in implementing a text-to-speech option.

A related, but separate idea for future improvements to OCS in general is to create an RStudio Addin (extension) [68] that would contain all case studies. This format would likely make it easier for developers to incorporate more interactive elements in the case studies, including the elements previously described. OCS as an RStudio Addin will be discussed in more detail in the following section.

The interactive case studies created during the course of this project are already improving user experience and engagement. They are a significant development and will be a foundation to build upon for further interactivity.

4.4 OCSdata Package

The development of the R package "OCSdata" is also an excellent addition to the OCS education platform. This package will give educators a quick fix option for troubleshooting errors related to data acquisition on students' computers. It will do the same for students, self-learners, or any type of user struggling to access data on their own.

Issues in the early steps of the data science process such as data acquisition

may not seem like the most pertinent. However, many first time instructors (including myself) are surprised by the amount of classroom time that is consumed by these issues. If the issue isn't solved quickly it can pose a major dilemma for instructors.

On one hand, the students facing the issues are likely to panic as they are no longer able to follow along with the lecture. If this happens early in the student's programming education, when it is most likely to occur, the experience may be uncomfortable enough to drive them away from the discipline entirely.

On the other hand, students who are not having issues have to sit patiently waiting for the instructor to return to lecturing. This may be only a minor inconvenience if a solution is found quickly. However, sometimes (possibly often) a solution is not immediately found. As time ticks on, instructors are faced with either leaving the students with issues to fend for themselves, or have the students without issues continue to wait patiently. With the help of "OCSdata," OCS users can avoid this problem completely.

Getting the package on CRAN was also a very exciting achievement. This makes it easier for our users to access the package. It also earns the package some credit and official status which lets new users know the package is safe to use.

When the package functions had to be restructured to comply with CRAN policies, it was worrisome that the package wouldn't be able to automatically save the data files to the user's current project. However, in the end the changes made to the package achieve this goal sufficiently and in an even

more elegant and safer fashion.

The best advice I have for R developers submitting their package to CRAN for the first time is to thoroughly read and understand CRAN policy before submitting. This cannot be reiterated enough. If there are any policies that seem confusing, utilize your resources to help understand them. Before submitting, double check that your package is in compliance with all CRAN policies as CRAN will allow for almost no exceptions. An important note is that policies are not applied retroactively. This means that packages already released on CRAN may not have to comply to all of the same policies as your package.

Future work on the package will need to be done to maintain proper function. Additionally, there are a couple of improvements to the package that could be made in the future. The first development to the package should make it so the functions are able to work with new case studies in the future. The functions could be made flexible to future case studies by updating a list of current case studies contained within the package that gets updated every time the package is loaded.

As discussed briefly above, a future development in OCS could be to implement the case studies as an Rstudio Addin. Rstudio Addins are graphical user interfaces (GUI) that allow RStudio users to interactively run R functions by selecting parameters and pressing buttons rather than writing code [68]. An RStudio Addin could be developed that not only contains all of the case studies, but also the OCSdata package. The package functions could be repurposed into this theoretical addin so that users could access both the case

study content and data, all from within RStudio. This would make the case studies easier to use, improve user experience, and likely improve learning outcomes.

Overall, the package has been a very exciting development for OCS. The number of downloads the package already has from CRAN is outstanding and suggests that the package is worthwhile for our users. Another survey in the future should be conducted to obtain user feedback about the package to assess its quality and any need for improvement or bug fixes. Currently, users are able to report bugs or provide suggestions on the GitHub repository issues page at github.com/opencasestudies/OCSdata/issues [20]. I am excited to see how the package will continue to develop in the future.

4.5 User Assessment of Case Studies: Survey Responses

The results of the survey were intriguing and gave important feedback on how effective the case studies are as an education platform. There was a relatively even spread between the number of responses from educators, students and self-learners. It was a bit surprising that self-learners accounted for most of the responses, although there were only two more self-learners than educators or students.

The case studies looked at by the survey participants is also very informative. Obesity was the most popular case study overall and particularly

popular with educators. This case study may be popular amongst educators due to its application of several standard statistical tests, methods that many educators would teach. The second most popular case study overall was the air pollution case study. Coincidentally, this case study is also second most popular for educators. The popularity of this case study is likely due to its machine learning content, a hot topic in the world right now and something many educators are beginning to teach.

The "Right-to-Carry" (RTC) Wrangling, Youth Disconnection, and Youth Mental Health case studies were the least popular among survey participants. Wrangling, the process of cleaning and preparing data for analysis, is often an underappreciated step in data science education. OCS users are no exception to this trend, as seen in the difference between the results of the RTC Wrangling and RTC Analysis case studies. However, wrangling can often be the most time consuming step, hence why the "Right-to-Carry" case study had to be split into two parts. The lack of attention towards the Youth Disconnection and Youth Mental Health case studies is more surprising. These are two particularly important topics in the current state of the world that would very much relate to the student experience. This author suspects, however, that these topics may hit too close to home for students at the moment. Keeping their students in mind, educators may have avoided these case studies as a precaution.

The survey was also very helpful in providing insight to who our users are and what they are looking for. Most of our educators taught undergraduate and/or graduate students, while a few taught high school. None taught

middle school. All of the student participants were undergraduate or graduate students, none in middle or high school. However, to use this data for research, users were required to be 18. There may be high school students who responded to the survey, but were not old enough to fully participate. Most participants were interested in learning/teaching data science but few participants indicated interest in a specific data science topic. This is representative of both the popularity and novelty of the field. People want to learn data science to be able to fill the growing demand for data scientists, but many of them don't know what they need or want to learn specifically.

Another interesting result is that one of the topics that the most amount of students and self-learners learned something new about was data communication. This suggests that there may be a lack of material and focus on data communication methods in current data science resources. This would be on par with many other STEM fields where educating how to communicate effectively is often not prioritized. However, in data science especially, being able to communicate results and present information is a very important and fundamental skill. Other education resources should follow OCS' example and begin highlighting data communication in their material as well.

In general, survey participants rated OCS very highly on a handful of important aspects. A significant majority of responses said they were likely or very likely to refer other people to OCS. The questions asking participants how they felt about the case studies' usefulness, their likeliness to refer back, and their enjoyment of the case studies had very similar results. This is highly positive feedback and is motivating to see that people are appreciative of this

material and regard it as high quality. However, that doesn't mean there isn't room for improvement. There were still one or two neutral responses (3 out of 5) to these questions. The future directions discussed here will likely improve the user experience of even the harshest critics.

It wasn't an easy process to recruit survey responses. Advertising OCS and the survey on Twitter was found to be the most effective and reasonable method for our team. Team members also advertised the survey in relevant courses they participated in. This method wasn't always available like Twitter was, but was very reliable when available. New recruitment methods should continue to be explored and documented.

One major limitation of this survey is that it does not reflect the majority of developments made to OCS with this project. The survey was being conducted while the project was developed, and thus the survey responses cannot be used to assess the changes made in this thesis. An important next step for future work on the project is to conduct further surveys to measure user satisfaction with these developments.

4.6 Assessment of Popularity and Reach: Google

Analytics Traffic Data

The website traffic data tracked with Google Analytics also provided insights into who uses OCS and how they use it. This data was very helpful in measuring overall usage, individual case study popularity, and also provides

information on where users are located. The knowledge gained from this analysis and those in the future will be very helpful for improving the OCS platform and measuring growth overtime.

The daily user count for OCS surpassed expectations. Seven new users a day on average implies that OCS is consistently growing its user base. This is exciting for the project developers as it is evidence that OCS is becoming a popular resource. Even more exciting is that the number seems to be trending upward in the last couple months, indicating that the case studies' reach is continuing to expand beyond the Johns Hopkins University network. OCS is not only growing its user base, but also increasing engagement with these users. The weekly engagement rate plot clearly shows a significant increase in the average engagement rate in the more recent weeks versus the first weeks. According to Figure 3.27, the last couple months have stabilized around a 75% engagement rate, meaning that well over half our visitors are engaging with the website content. This would suggest that most visitors find the content interesting enough to continue reading for at least some duration, rather than exiting the page immediately. The growth in engagement rate from the first week also suggests that the advertising methods used since the start of the year have been effective in recruiting case study users.

One of the most helpful metrics provided from the traffic data is the number of total sessions accrued for each case study. The number of sessions is particularly helpful as it indicates which case studies are the most used. Knowing which case studies are the most popular can help improve case studies in the future by suggesting which elements in a case study our users

are looking for.

Interestingly, the top two case studies with the most sessions matches the survey results for the top two most popular case studies. Both data sets indicate Air Pollution and Obesity were the most used case studies. However, Obesity was the most popular for survey participants, while Air Pollution had the most sessions according to the traffic data. The traffic data shows that case study use is generally more evenly distributed between case studies than what is suggested by the survey results. One exception to this is the "Right-To-Carry" (RTC) Wrangling case study which has significantly less sessions than any other case study. This is likely due to the lack of appreciation for data wrangling methods, as discussed previously.

It is to be expected that survey and traffic data would not match perfectly. The survey participants make a very small sample of total case study users and likely are not a perfect reflection of the greater population. Regardless, the feedback from this sample, even if small, is invaluable.

One exciting analysis conducted on the traffic data was visualizing the locations of all OCS users. Baltimore was the most popular city by a large margin. This is not surprising since the home institution of OCS is located in Baltimore and many of the first educators to use OCS are colleagues also located in Baltimore.

The next city with the most number of sessions was Seattle. This author is located in Seattle and is likely the source of many of these sessions. However, it's highly doubtful that this is the source of all the sessions recorded. Seattle is also a city with a large academic community, particularly in medicine, public

health, and biostatistics. It's not unlikely that OCS has been used in this community, especially considering the relationship between the biostatistics departments at Johns Hopkins University in Baltimore and the University of Washington in Seattle.

The rest of the cities reported by Google Analytics have no such explanation, and thus all sessions would be counted from visitors with no association to the project or developers. It's very exciting to see that three of the top five cities with the most OCS sessions are located outside of the US. Additionally, sessions have been recorded from every continent in the world except for Antarctica. This is evidence that OCS' reach expands across the globe and is providing educational material for the world. It also further indicates the necessity of a platform like OCS, and is motivating to develop the platform even further.

Overall, Google Analytics has proved to be an excellent tool for measuring OCS' usage and progress. As the project continues to grow, this data should continue to be tracked to provide further insights and to measure the growth of the platform over time. In the future, this data could be used as evidence of the project's success and impact when applying for more funding to grow OCS.

4.7 Conclusion

While the survey and traffic data were a great start in measuring and assessing the quality of the case studies, this should be taken even further in

the future. One initial idea that was ultimately out of scope for this project is to conduct a controlled, comparative study where one teacher teaches one section of a course without case studies and another section with case studies. Student outcomes would be analyzed and both the instructor and students would be interviewed and surveyed to collect a detailed assessment of case studies' impact on data science education.

Following in that suit, another comparative study could be conducted within the same paradigm but instead comparing the static and interactive case studies. This would make it possible to measure any significant differences in learning outcomes with the interactive versions. This could validate the impact the interactive elements have on the learning experience and suggest possible methods for improving said interactivity.

OCS would also benefit from getting feedback from sources outside of academia. Industry experts could be asked to look at the platform web pages and the case studies themselves to provide expert feedback on website design. Having an outside perspective such as this would be very helpful in improving user experience. Companies such as [usertesting.com](https://www.usertesting.com) [69] could be utilized to acquire such feedback.

This project brought a lot of exciting developments to OCS. The case studies were translated and made to be interactive. A package was developed that will remove barriers to case study data and increase accessibility for new users. Educators now can use the MakeCaseStudies app to create their own digital content. The insights extracted from the survey and website traffic data has been very informative and will help grow OCS further. These developments

will improve the education platform for current and future users alike.

Chapter 5

Appendices

5.1 Appendix A: Open Case Studies User Feedback Survey

The user feedback survey was created and distributed using Google Forms [49]. The survey in its entirety is attached below. Note that the attached version looks a bit different than the online version. This is because surveys are converted to a handwriting friendly format when printed from Google Forms. The OCS survey can also be viewed online at <https://forms.gle/GtvyS5JgypF6Los1A>.

Open Case Study Survey

We are collecting data about user experience with our case studies to learn more about how to improve the data science education experience. Part of this includes getting a better understanding of who is using our case studies and how so that we can better design our case studies. This data may ultimately be used for a research publication. We thank you very much for any feedback you can provide!

DATA CONFIDENTIALITY: We will not collect any personally identifiable information about you through this survey for the purposes of this research. We will only ask very general questions about you (such as if you are a student) and your use of the case studies. The potential risks to you are small. The potential benefits to the community of data scientists, developers, and professors are very high – we will be able to learn how we can improve data science/statistics/public health education material.

DATA SHARING: We plan to release the data collected from this survey openly online using GitHub (<https://github.com/>) - however no personally identifiable information (such as email addresses) will be shared as it will not be collected. See below for an image of what the data might look like. We will only have a timestamp and answers to your questions. Only a reduced version of the timestamp will be included in the shared data to enhance the anonymity of the participants.

PARTICIPATION: You must be 18 or over to participate. You will be asked a series of questions about your experience with the case studies and some general information about you. You can stop participating in the survey at any point. You do not need to answer all of the questions. Only answers that are submitted by clicking the submit button at the end of the survey will be recorded. Depending on how you answer the survey, this will take 1-10 minutes to complete. You will have an opportunity at the end of the survey to edit your responses, including removing all of them. Any answers that are ultimately submitted (and not removed by the participant) will continue to be used for research.

General information about traffic to the case study websites and our main website (www.opencasestudies.org) is being tracked by Google Analytics (<https://analytics.google.com>) - which provides us with summary information such as the number of visitors from different countries. A running count of the number of people who have visited our main website (www.opencasestudies.org) is also being tracked with ClusterMaps (<https://clustrmaps.com/>). We will not use this to try to identify who has submitted a survey response.

RISK: Risk of participation is minimal. If our google account was hacked, it is possible that someone could attempt to leak this information. However, the questions we ask pose minimal security risk to participants as they are largely about if participants found the case studies useful. Recall that you do not need to answer most questions. Only a few questions are required to determine what additional questions to ask.

CONSENT: The first questions will ask about your age and if you consent (or agree) that the data from your responses be used for research purposes.

If you have different thoughts for different case studies, we kindly ask that you submit to our survey more than once.

QUESTIONS OR PROBLEMS: If you have questions or problems, email Carrie Wright at cwri60@jh.edu.

You may also contact the Johns Hopkins Bloomberg School of Public Health IRB Office if you have questions about your rights as a participant/parent of a study

participant. Contact the IRB if you feel you have not been treated fairly or if you have other concerns.

This study has been deemed exempt by the Johns Hopkins Bloomberg School of Public Health IRB. No personal or private information will be collected. Thus, the proposed activity does not qualify as human subjects research as defined by DHHS regulations 45 CFR 46.102, and does not require IRB oversight.

IRB No.: 14965
Date Approved: 2020-02-26

The IRB contact information is:
Address: Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street, Suite E1100, Baltimore, MD 21205
Telephone: 410-955-3193; Toll Free: 1-888-262-3242
E-mail: jhsph.irboffice@jhu.edu

* Required

Here is an example of what the data from this survey will eventually look like to us (each row contains the responses from a participant):

A	B	C
Timestamp	How would you describe yourself (choose best option)?	Do you intend to use a case study (or part of a case study) in one of your classes?
2/17/2021 16:11:28	Educator of Data Science/Stats/Public Health/other - looking for data and education material to use in courses	Maybe
2/17/2021 16:12:30	Student - using case studies for a course or to help with a course	

1. Are you 18 or over? You cannot participate in this survey if you are under 18. *

Mark only one oval.

- Yes Skip to question 2
 No

Consent

2. Do you consent to your responses being used for research purposes? Please see the description at the beginning of the survey for more information (use the back button within this survey if you wish to see it again). *

Mark only one oval.

- Yes Skip to question 3
 No

3. How would you describe yourself (choose best option)? Note that you can say "wish not to say" for the "other" category. *

Mark only one oval.

- Educator of Data Science/Stats/Public Health/other - looking for data and education material to use in courses *Skip to question 4*
- Student - using case studies for a course or to help with a course *Skip to question 52*
- Self-learner - interested in learning more about the skills, topics, and or concepts covered in the case studies *Skip to question 71*
- Other: _____

For
Educators

We would like to know your thoughts about how we can create content to better serve your needs. Please get in touch with us (email: opencasestudies@gmail.com) if you would like to work with us to conduct research about the use of our case studies in the classroom, if you are interested in creating a case study, if you have ideas, or you would like to know more about our project. Also check out our website at www.opencasestudies.org if you have not already.

4. What specifically brought you to the case studies? (select all that apply)

Check all that apply.

- General interest in teaching material about statistics
- Interested in teaching about a specific statistical topic (paired t-test, regression, etc.)
- General interest in teaching materials about data science/data analysis
- Interested in teaching about a specific data science skill (such as creating a dashboard)
- General interest in teaching materials about R programming
- General interest in teaching materials about the tidyverse
- General interest in teaching materials about public health
- Interested in teaching about a specific public health topic (such as mental health, obesity etc.)

Other: _____

5. What type of students do you teach? (select all that apply)

Check all that apply.

- Graduate students
- Undergraduate students
- High school students
- Middle school students

Other: _____

6. How likely are you to recommend the case studies to others?

Mark only one oval.

	1	2	3	4	5	
Not likely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very likely

7. What case study(ies) have you looked at? (select all that apply)

Check all that apply.

- Case study about obesity: <https://www.opencasestudies.org/ocs-bp-rural-and-urban-obesity/>
 - Case study about diet: <https://www.opencasestudies.org/ocs-bp-diet/>
 - Case study about school shootings: <https://www.opencasestudies.org/ocs-bp-school-shootings-dashboard/>
 - Case study about multicollinearity and RTC laws (wrangling): <https://www.opencasestudies.org/ocs-bp-RTC-wrangling/>
 - Case study about multicollinearity and RTC laws (analysis): <https://www.opencasestudies.org/ocs-bp-RTC-analysis/>
 - Case study about predicting air pollution: <https://www.opencasestudies.org/ocs-bp-air-pollution/>
 - Case study about CO2 emissions: <https://www.opencasestudies.org/ocs-bp-co2-emissions/>
 - Case study about youth disconnection: <https://www.opencasestudies.org/ocs-bp-youth-disconnection/>
 - Case study about youth mental health: <https://www.opencasestudies.org/ocs-bp-youth-mental-health/>
 - Case study about opioid shipments in the US: <https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/>
 - Case study about vaping behaviours among US youths: <https://www.opencasestudies.org/ocs-bp-vaping-case-study/>
 - None
- Other: _____

8. Have you already used our case studies or the data from our case studies to teach?

Mark only one oval.

- Yes *Skip to question 9*
- No *Skip to question 37*
- Other: _____

Skip to question 37

Already used a case study or data

9. Which case studies (or data from a case study) did you use? (select all that apply)

Check all that apply.

- Not sure yet
 - Case study about obesity: <https://www.opencasestudies.org/ocs-bp-rural-and-urban-obesity/>
 - Case study about diet: <https://www.opencasestudies.org/ocs-bp-diet/>
 - Case study about school shootings: <https://www.opencasestudies.org/ocs-bp-school-shootings-dashboard/>
 - Case study about multicollinearity and RTC laws (wrangling): <https://www.opencasestudies.org/ocs-bp-RTC-wrangling/>
 - Case study about multicollinearity and RTC laws (analysis): <https://www.opencasestudies.org/ocs-bp-RTC-analysis/>
 - Case study about predicting air pollution: <https://www.opencasestudies.org/ocs-bp-air-pollution/>
 - Case study about CO2 emissions: <https://www.opencasestudies.org/ocs-bp-co2-emissions/>
 - Case study about youth disconnection: <https://www.opencasestudies.org/ocs-bp-youth-disconnection/>
 - Case study about youth mental health: <https://www.opencasestudies.org/ocs-bp-youth-mental-health/>
 - Case study about opioid shipments in the US: <https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/>
 - Case study about vaping behaviours among US youths: <https://www.opencasestudies.org/ocs-bp-vaping-case-study/>
- Other: _____

10. What type of course(s) did you use our case studies or data for? (check all that apply)

Check all that apply.

- Public Health
 - Statistics
 - Math
 - Data Science
- Other: _____

11. Was it clear how to locate, access, and download files from a case study repository?

Mark only one oval.

- Yes, very clear
- I found the files with a bit of difficulty
- No, very unclear
- Other: _____

12. Was it clear how to use just part of the case study in a modular way?

Mark only one oval.

- Yes, very clear
- I found the info in the readme file with a bit of difficulty
- No, very unclear
- Other: _____

13. Please tell us how you taught with our data or case studies - Example: for a lecture about data wrangling or to have students write a report about a data analysis.

14. What did you use to teach?

Mark only one oval.

- Just the data *Skip to question 15*
- Parts of a case study *Skip to question 22*
- Full case study *Skip to question 22*
- Combination (sometimes just the data or part of the case study... etc.)
Skip to question 22

Just used data

15. What type of data did you use? (select all that apply)

Check all that apply.

Raw data files or original source (ex. a website)

Wrangled data files (as rda files)

Wrangled data files (as csv files)

Other: _____

16. Did using our data save you time?

Mark only one oval.

No

Somewhat

Yes, a lot

17. Did using our data save you effort?

Mark only one oval.

No

Somewhat

Yes, a lot

18. Do you plan to use our data again in the future?

Mark only one oval.

1 2 3 4 5

No For sure!

19. Do you plan to use our case studies (not just data) in the future?

Mark only one oval.

1 2 3 4 5

No For sure!

20. If not, what might make you more inclined?

21. Please provide any additional feedback you might have.

Used part or all of a case study

22. Which sections were most helpful for you? (select all that apply)

Check all that apply.

- Motivation/Context/What are the data?
- Data Import
- Data Wrangling
- Data Visualization
- Data Analysis

Other: _____

23. How different was incorporating the case study or parts of the case study compared to how you typically teach similar material?

Mark only one oval.

- Very different
- Somewhat similar
- About the same

24. If it was different from how you typically teach, how?

25. How do you typically teach similar material?

26. Did incorporating our case study materials into your class, save you time?

Mark only one oval.

1 2 3 4 5

No, not at all. Yes, very much!

27. Did incorporating our case study materials into your class, save you effort?

Mark only one oval.

1 2 3 4 5

No, not at all. Yes, very much!

28. Compared to how you taught similar material previously, did you enjoy using the case studies to teach with?

Mark only one oval.

1 2 3 4 5

Not at all Absolutely, will do again!

29. Compared to how you taught similar material previously, how well did your students seem to enjoy the case studies?

Mark only one oval.

- My students did not enjoy learning with case study as well
- They seemed to enjoy it about the same
- My students seemed to enjoy it better
- I have not taught similar material previously
- Other: _____

30. Compared to how you taught similar material previously, how well did students seem to learn the content that you taught using a case study?

Mark only one oval.

- Less well
- About the same
- Better
- I have not taught similar material before
- Other: _____

31. Did you learn anything new using the case studies?

Mark only one oval.

- Yes
- No
- Other: _____

32. Did using our case studies allow you to incorporate something new into your teaching?

Mark only one oval.

- No, not at all
- Yes, somewhat
- Yes, a lot

33. What was the biggest benefit of using the case study?

34. What was the most difficult aspect of using the case study?

35. Is there anything you would do differently next time?

36. Please provide any additional feedback you might have.

Interested in using our materials?

37. Are you interested in using our case study materials or the data from one of our case studies to teach in the future?

Mark only one oval.

- Yes *Skip to question 39*
- Maybe *Skip to question 39*
- No - I am not interested *Skip to question 38*
- Other: _____

Skip to question 39

Do not intend to use a case study

38. Please provide any additional feedback you might have.

Use of case studies or data

39. Are you interested in using just the data, part of case study, or a full case study?

Mark only one oval.

- Just the data *Skip to question 40*
- Only part of a case study (or case studies) *Skip to question 46*
- Full case study (or case studies) *Skip to question 46*
- Depends (sometimes just data, sometimes part or all of a case study)
Skip to question 46
- Other: _____

Skip to question 46

Interested in data only

40. Which case study or case studies data are you interested in using?
(select all that apply)

Check all that apply.

- Not sure yet
 - Case study about obesity: <https://www.opencasestudies.org/ocs-bp-rural-and-urban-obesity/>
 - Case study about diet: <https://www.opencasestudies.org/ocs-bp-diet/>
 - Case study about school shootings: <https://www.opencasestudies.org/ocs-bp-school-shootings-dashboard/>
 - Case study about multicollinearity and RTC laws (wrangling):
<https://www.opencasestudies.org/ocs-bp-RTC-wrangling/>
 - Case study about multicollinearity and RTC laws (analysis):
<https://www.opencasestudies.org/ocs-bp-RTC-analysis/>
 - Case study about predicting air pollution:
<https://www.opencasestudies.org/ocs-bp-air-pollution/>
 - Case study about CO2 emissions: <https://www.opencasestudies.org/ocs-bp-co2-emissions/>
 - Case study about youth disconnection: <https://www.opencasestudies.org/ocs-bp-youth-disconnection/>
 - Case study about youth mental health: <https://www.opencasestudies.org/ocs-bp-youth-mental-health/>
 - Case study about opioid shipments in the US:
<https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/>
 - Case study about vaping behaviours among US youths:
<https://www.opencasestudies.org/ocs-bp-vaping-case-study/>
- Other: _____

41. What type of course(s) would you (or did you) use our case study data for? (check all that apply)

Check all that apply.

- Public Health
 - Statistics
 - Math
 - Data Science
- Other: _____

42. What type of data do you intend to use?

Mark only one oval.

- Raw data files / the original source (ex. a website)
- Wrangled data (as rda files)
- Wrangled data (as csv files)
- Other: _____

43. Is it clear how to locate, access, and download files from a case study repository?

Mark only one oval.

- Yes, very clear
- Not super clear
- No, I don't know where to obtain the data
- Other: _____

44. Please tell us how you might teach with our data (or did already) - Example: for a lecture about data wrangling.

45. Please provide any additional feedback you might have.

Interested in data or more

46. Please tell us which materials you are most interested in using (or you already used to teach). (select all that apply)

Check all that apply.

Motivation/Context

Data Import

Data Wrangling

Data Exploration

Data Visualization

Data Analysis

The data itself

All sections

Other: _____

47. Which case study or case studies might you be interested in using.
(select all that apply)

Check all that apply.

- Not sure yet
 - Case study about obesity: <https://www.opencasestudies.org/ocs-bp-rural-and-urban-obesity/>
 - Case study about diet: <https://www.opencasestudies.org/ocs-bp-diet/>
 - Case study about school shootings: <https://www.opencasestudies.org/ocs-bp-school-shootings-dashboard/>
 - Case study about multicollinearity and RTC laws (wrangling):
<https://www.opencasestudies.org/ocs-bp-RTC-wrangling/>
 - Case study about multicollinearity and RTC laws (analysis):
<https://www.opencasestudies.org/ocs-bp-RTC-analysis/>
 - Case study about predicting air pollution:
<https://www.opencasestudies.org/ocs-bp-air-pollution/>
 - Case study about CO2 emissions: <https://www.opencasestudies.org/ocs-bp-co2-emissions/>
 - Case study about youth disconnection: <https://www.opencasestudies.org/ocs-bp-youth-disconnection/>
 - Case study about youth mental health: <https://www.opencasestudies.org/ocs-bp-youth-mental-health/>
 - Case study about opioid shipments in the US:
<https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/>
 - Case study about vaping behaviours among US youths:
<https://www.opencasestudies.org/ocs-bp-vaping-case-study/>
 - None
- Other: _____

48. Is it clear how to locate, access, and download files from a case study repository?

Mark only one oval.

- Yes, very clear
- Not super clear
- No, I don't know where to obtain the data
- Other: _____

49. What type of course(s) would you (or have you) used our case studies for? (check all that apply)

Check all that apply.

- Public Health
- Statistics
- Math
- Data Science

Other: _____

50. Please tell us how you might teach (or already taught) with one or more of our case studies. (Example: for materials for students to practice writing about the results of a data analysis)

51. Please provide any additional feedback you might have.

For Students

52. How did you find our case studies?

Mark only one oval.

- Using case studies as part of a course (asked to do so by an instructor)
- Using case studies to supplement other instruction (not asked to do so by an instructor)
- Other: _____

53. What type of student would you describe yourself as?

Mark only one oval.

- High school
- College
- Graduate
- Do not wish to say
- Other: _____

54. What specifically brought you to the case studies? (check all that apply)

Check all that apply.

- interested in learning more about statistics in general
- interested in learning more about data science/data analysis in general
- interested in learning more about public health in general
- interested in learning more about R programming in general
- interested in learning more about the tidyverse in general
- interested in a specific case study public health topic (such as environment, mental health, etc.)
- interested in a specific data science skill (such as how to create a dashboard, machine learning, etc.)
- interested in a specific statistical topic (such as how to perform a paired t-test)
- Other: _____

55. How likely are you to recommend the case studies to others?

Mark only one oval.

1 2 3 4 5

Not likely Very likely

56. If you have not seen a case study yet, select "Not yet" , otherwise select "Yes" to answer questions about a case study you have seen.

Mark only one oval.

Not yet! *Skip to question 57*

Yes *Skip to question 58*

Student (no case studies)

57. Please provide any feedback you might have.

Student continued...

58. What case study(ies) have you looked at? (select all that apply)

Check all that apply.

- Case study about obesity: <https://www.opencasestudies.org/ocs-bp-rural-and-urban-obesity/>
- Case study about diet: <https://www.opencasestudies.org/ocs-bp-diet/>
- Case study about school shootings: <https://www.opencasestudies.org/ocs-bp-school-shootings-dashboard/>
- Case study about multicollinearity and RTC laws (wrangling): <https://www.opencasestudies.org/ocs-bp-RTC-wrangling/>
- Case study about multicollinearity and RTC laws (analysis): <https://www.opencasestudies.org/ocs-bp-RTC-analysis/>
- Case study about predicting air pollution: <https://www.opencasestudies.org/ocs-bp-air-pollution/>
- Case study about CO2 emissions: <https://www.opencasestudies.org/ocs-bp-co2-emissions/>
- Case study about youth disconnection: <https://www.opencasestudies.org/ocs-bp-youth-disconnection/>
- Case study about youth mental health: <https://www.opencasestudies.org/ocs-bp-youth-mental-health/>
- Case study about opioid shipments in the US: <https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/>
- Case study about vaping behaviours among US youths: <https://www.opencasestudies.org/ocs-bp-vaping-case-study/>
- Other: _____

59. Were you able to locate, access, and download the data for the case study with relative ease?

Mark only one oval.

- Yes, I found the files I needed with no issue
- I found the files but had some difficulty
- No, I could not find the files I needed
- Other: _____

60. How familiar were you with the statistical methods presented in the case study before you read it?

Mark only one oval.

- Never heard about the statistical topics covered
- Had heard briefly about the statistical topics covered
- Had an understanding about the topics but not as much as the case study covered
- Regularly perform the statistical methods covered/already knew all or most of the material
- Other: _____

61. How familiar were you with the public health topics presented in the case study before you read it?

Mark only one oval.

- Never heard about the public health topics
- Had heard briefly about the topics
- Had an understanding about the topics but not as much as the case study covered
- Already knew all or most of the material

62. How familiar were you with the data science topics presented in the case study before you read it?

Mark only one oval.

- Never heard about the data science topics
- Had heard briefly about the topics
- Had an understanding about the topics but not as much as the case study covered
- Already knew all or most of the material

63. I learned something new about: (select all that apply)

Check all that apply.

- A statistical method or concept
 - How to implement a statistical method using R
 - Data import methods
 - Data visualization methods
 - Data wrangling methods
 - Data communication methods
 - The public health topics presented
 - All of the above
 - None
- Other: _____

64. How useful did you find the case study/studies?

Mark only one oval.

1 2 3 4 5

Not useful Very useful

65. How enjoyable did you find the case study/studies as a way to learn about these topics?

Mark only one oval.

1 2 3 4 5

Not interesting/fun Very enjoyable!

66. How likely are you to refer back to the case study/studies in the future?

Mark only one oval.

1 2 3 4 5

Not likely Very likely

67. How familiar were you with R programming before you read the case study?

Mark only one oval.

1 2 3 4 5

I am totally new to R programming Very familiar - I work or have worked in R regularly

68. How familiar were you with the tidyverse programming material presented in the case study before you read it?

Mark only one oval.

1 2 3 4 5

I am totally new to the tidyverse (even if I'm a base R expert) Very familiar - all or most of the material

69. Do you have familiarity with another programming language besides R?

Mark only one oval.

1 2 3 4 5

No none at all Yes I am proficient in one or more languages

70. Please provide any additional feedback you might have.

For Self-Learners

71. What specifically brought you to the case studies? (select all that apply)

Check all that apply.

- interested in learning more about statistics in general
 - interested in learning more about data science/data analysis in general
 - interested in learning more about public health in general
 - interested in learning more about R programming in general
 - interested in learning more about the tidyverse in general
 - interested in a specific public health topic (such as environment, mental health, etc.)
 - interested in a specific data science skill (such as how to create a dashboard, machine learning, etc.)
 - interested in a specific statistical topic (such as how to perform a paired t-test)
- Other: _____

72. How likely are you to recommend the case studies to others?

Mark only one oval.

	1	2	3	4	5	
Not likely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very likely

73. What field do you work in? (we would like to know more about who is interested in our case studies - you are not required to answer)

74. If you have not seen a case study yet, select "Not yet" , otherwise select "Yes" to answer questions about a case study you have seen.

Mark only one oval.

Not yet! *Skip to question 75*

Yes *Skip to question 76*

Self-learner (no case studies yet)

75. Please provide any feedback you might have.

Self-learner continued...

76. What case study(ies) have you looked at? (select all that apply)

Check all that apply.

- Case study about obesity: <https://www.opencasestudies.org/ocs-bp-rural-and-urban-obesity/>
- Case study about diet: <https://www.opencasestudies.org/ocs-bp-diet/>
- Case study about school shootings: <https://www.opencasestudies.org/ocs-bp-school-shootings-dashboard/>
- Case study about multicollinearity and RTC laws (wrangling): <https://www.opencasestudies.org/ocs-bp-RTC-wrangling/>
- Case study about multicollinearity and RTC laws (analysis): <https://www.opencasestudies.org/ocs-bp-RTC-analysis/>
- Case study about predicting air pollution: <https://www.opencasestudies.org/ocs-bp-air-pollution/>
- Case study about CO2 emissions: <https://www.opencasestudies.org/ocs-bp-co2-emissions/>
- Case study about youth disconnection: <https://www.opencasestudies.org/ocs-bp-youth-disconnection/>
- Case study about youth mental health: <https://www.opencasestudies.org/ocs-bp-youth-mental-health/>
- Case study about opioid shipments in the US: <https://www.opencasestudies.org/ocs-bp-opioid-rural-urban/>
- Case study about vaping behaviours among US youths: <https://www.opencasestudies.org/ocs-bp-vaping-case-study/>
- Other: _____

77. Were you able to locate, access, and download files from the case study repository with relative ease?

Mark only one oval.

- Yes, I found the files I needed with no issue
- I found the files but had some difficulty
- No, I could not find the files I needed
- Other: _____

78. How familiar were you with the statistical methods presented in the case study before you read it?

Mark only one oval.

- Never heard about the statistical topics covered
- Had heard briefly about the statistical topics covered
- Had a basic understanding about the topics but not as much as the case study covered
- Regularly perform the statistical methods covered/ already had a great deal of knowledge about the topics covered

79. How familiar were you with the public health topics presented in the case study before you read it?

Mark only one oval.

- Never heard about the public health topics covered
- Had heard briefly about the topics covered
- Had a basic understanding about the topics but not as much as the case study covered
- Had in depth knowledge about the topic - already knew all or most of the material covered

80. How familiar were you with the data science topics presented in the case study before you read it?

Mark only one oval.

- Never heard about the data science topics covered
- Had heard briefly about the topics covered
- Had a basic understanding about the topics but not as much as the case study covered
- Had in depth knowledge about the topic - already knew all or most of the material covered

81. I learned something new about: (select all that apply)

Check all that apply.

- A statistical method or concept
- How to implement a statistical method using R
- Data import methods
- Data visualization methods
- Data wrangling methods
- Data communication methods
- The public health topics presented
- All of the above
- None

Other: _____

82. How useful did you find the case study/studies?

Mark only one oval.

1 2 3 4 5

Not useful Very useful

83. How enjoyable did you find the case study/studies as a way to learn about these topics?

Mark only one oval.

1 2 3 4 5

Not interesting/fun Very enjoyable!

84. How likely are you to refer back to the case study/studies in the future?

Mark only one oval.

1 2 3 4 5

Not likely Very likely

85. How familiar were you with R programming before you read the case study?

Mark only one oval.

1 2 3 4 5

I am totally new to R programming Very familiar - I work or have worked in R regularly

86. How familiar were you with the tidyverse programming material presented in the case study before you read it?

Mark only one oval.

1 2 3 4 5

I am totally new to the tidyverse (even if I'm a base R expert) Very familiar - all or most of the material

87. Do you have familiarity with another programming language besides R?

Mark only one oval.

1 2 3 4 5

No none at all Yes I am proficient in one or more languages

88. Please provide any feedback you might have.

5.2 Appendix B: MakeCaseStudies Guide

A step by step guide was created to walk new users through how to use the MakeCaseStudies web-app. The guide comes with screen shots and red boxes to highlight important parts at each step. This guide shall continue to be updated as the app is updated.

MakeCaseStudies Walkthrough

Welcome to MakeCaseStudies! This web-app is here to help you create a case study. To create a case study, start by filling in the text boxes highlighted in red with the information you want to be included in the case study. In this format you can include a case study title, body text, headers, images, and a video. By default, these text boxes are filled in with the content from a case study on the Gram Stain method to act as an illustrative example. Provide links to images using the image URL and attach a video by inputting its YouTube video code.

Create a Case Study



This tool is provided to help users create online lessons quickly and easily like our open case studies, which are online step-by-step lessons that guide users through a real-world problem solving challenge.

Powered by:



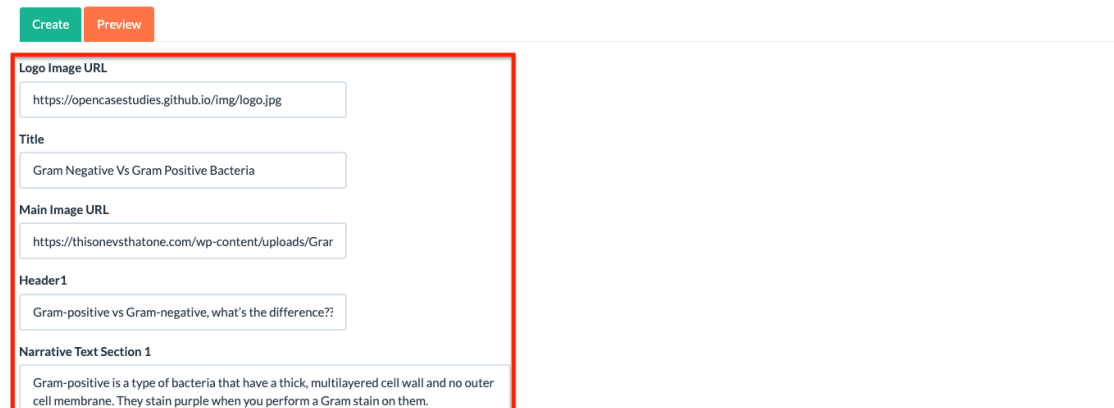
Start by clicking the **Make Case Study** button to download an example lesson.

[Make Case Study](#)

Document Format: HTML

Delete and replace the existing content within the Create tab for your own content, check the preview in the Preview tab, and finally press the **Make Case Study** button again to download your own lesson!

Photo by Kari Shea on Unsplash



Create Preview

Logo Image URL

Title

Main Image URL

Header1

Narrative Text Section 1

Figure 1: Opening page of the MakeCaseStudies app. Use this app to create your own case study.

Click the orange “Preview” tab (highlighted in red) in the top left to view what the case study in-progress looks like. Check this tab before downloading to ensure the final product looks as expected.

If you need more than the two sections provided by default, return to the “Create” tab and scroll to the bottom to find “insert” and “remove” buttons. Use the “Insert Header” and “Insert Narrative Section” buttons (highlighted in red) to add a new case study section with a header and text. The “remove” buttons are provided to remove any unused sections. Buttons to add more images and videos are not available at this time but they are under development.

Once satisfied with your case study, click on the download button at the top of the page

Image 1 URL

<https://cdn1.byjus.com/wp-content/uploads/2018/11/bi>

Header2

Gram Stain Method

Narrative Text Section 2

During Gram staining both bacteria are stained with a purple dye, but the gram-negative does not retain it. So, you could also associate that concept to their names as well.

So gram-positive (plus) bacteria have the purple stain, and gram-negative bacteria (minus) do not have the purple stain.

Of course, there is much more to these two bacterial types than that, so let's delve a bit deeper into both types.

Gram staining was invented by Hans Christian Gram, and it's sometimes referred to as Gram's method. This process is where gram-positive and gram-negative bacteria derived their names. Since bacteria are so tiny, Gram staining is used to determine the bacterial type (positive or negative).

This is incredibly important for treating ailments caused by bacteria. Different bacterial types react in different ways to different treatments, so we need to know

Image 2 URL

<https://cdn.technologynetworks.com/tn/images/body/g->

youtube video code

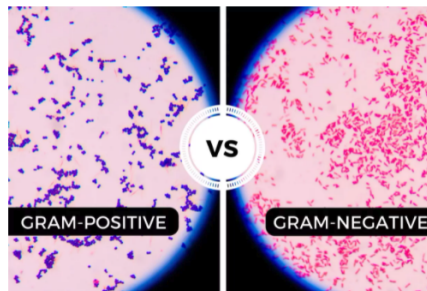
AZS2wb7pMo4

Figure 2: Create a case study by filling in the case study material into the provided text boxes.

Create Preview

Open Case Studies: Gram Negative Vs Gram Positive Bacteria

Code ▾



OPEN CASE STUDIES

Gram-positive vs Gram-negative, what's the difference??

Gram Stain Method

source

Gram-positive vs Gram-negative, what's the difference??

Gram-positive is a type of bacteria that have a thick, multilayered cell wall and no outer cell membrane. They stain purple when you perform a Gram stain on them.

Gram-negative is a type of bacteria that have a thin, single-layered cell wall and do have an outer cell membrane. They stain red or pink when you perform a Gram stain on them.

One way to help distinguish between the two different types is to associate the beginning letters of the words with a corresponding color attribute.

So gram-positive bacteria stain purple, and gram-negative bacteria do not.

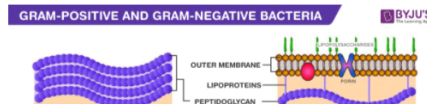


Figure 3: Preview tab of the MakeCaseStudies app.

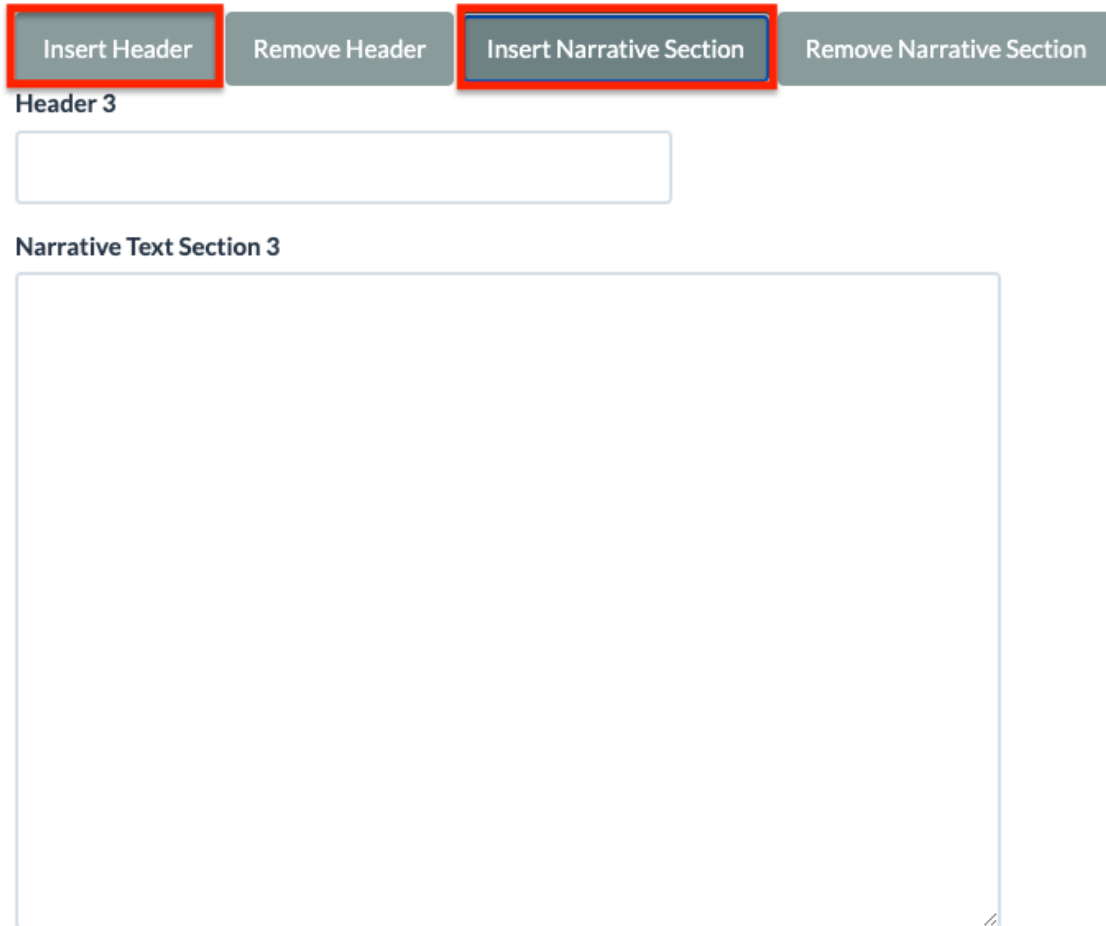


Figure 4: Insert and Remove Header and Narrative Text Section buttons.

labeled “Make Case Study.” The button is highlighted in red. This will download an HTML file containing the case study.

Create a Case Study

This tool is provided to help users create online lessons *quickly and easily* like our [open case studies](#), which are online step-by-step lessons that guide users through a real-world problem solving challenge.

Powered by:



Photo by Kari Shea on Unsplash

Start by clicking the **Make Case Study** button to download an example lesson.

Make Case Study

Document Format: HTML

Delete and replace the existing content within the Create tab for your own content, check the preview in the Preview tab, and finally press the **Make Case Study** button again to download your own lesson!

Figure 5: Make Case Study download button.

Create a Case Study

This tool is provided to help users create online lessons *quickly and easily* like our [open case studies](#), which are online step-by-step lessons that guide users through a real-world problem solving challenge.

Powered by:





Photo by Kari Shea on Unsplash

Start by clicking the **Make Case Study** button to download an example lesson.

Make Case Study

Document Format: HTML

Delete and replace the existing content within the Create tab for your own content, check the preview in the Preview tab, and finally press the **Make Case Study** button again to download your own lesson!

Create **Preview**

Logo Image URL

Title

Main Image URL

Header1

Narrative Text Section 1

Show All ×

Figure 6: Downloaded case study file.

Open the downloaded file (highlighted in red) named `my-report.html` to view the case study created. This file can be shared with students or hosted online.

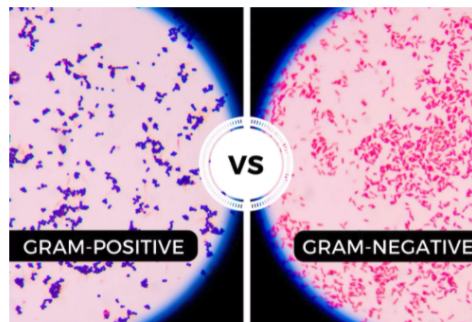
OPEN CASE STUDIES

Gram-positive vs Gram-negative, what's the difference??

Gram Stain Method

Open Case Studies: Gram Negative Vs Gram Positive Bacteria

Code ▾



[source](#)

Gram-positive vs Gram-negative, what's the difference??

Gram-positive is a type of bacteria that have a thick, multilayered cell wall and no outer cell membrane. They stain purple when you perform a Gram stain on them.

Gram-negative is a type of bacteria that have a thin, single-layered cell wall and do have an outer cell membrane. They stain red or pink when you perform a Gram stain on them.

One way to help distinguish between the two different types is to associate the beginning letters of the words with a corresponding color attribute.

So gram-positive bacteria stain purple, and gram-negative bacteria do not.

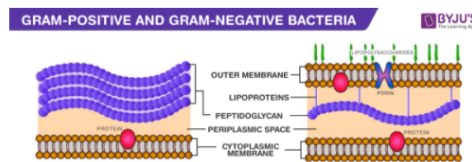


Figure 7: Open the case study HTML file.

Bibliography

- [1] Stephanie Hicks et al. *Open Case Studies*. en-us. URL: <https://www.opencasestudies.org/> (visited on 11/02/2021).
- [2] Winston Chang et al. *shiny: Web Application Framework for R*. 2021. URL: <https://CRAN.R-project.org/package=shiny>.
- [3] Barret Schloerke et al. *learnr: Interactive Tutorials for R*. 2021. URL: <https://CRAN.R-project.org/package=learnr>.
- [4] Garrick Aden-Buie et al. *gradethis: Automated Feedback for Student Exercises in 'learnr' Tutorials*. 2021. URL: <https://pkgs.rstudio.com/gradethis/>.
- [5] Michael Breshock, Carrie Wright, and Stephanie Hicks. *OCSdata: Download Data from the 'Open Case Studies' Repository*. 2021. URL: <https://CRAN.R-project.org/package=OCSdata>.

- [6] *Analytics Tools & Solutions for Your Business - Google Analytics*. Google Marketing Platform. URL: <https://marketingplatform.google.com/about/analytics/> (visited on 10/20/2021).
- [7] Wil M. P. van der Aalst. "The Data Science Revolution: How Learning Machines Changed the Way We Work and Do Business". en. In: 555 (2020). Ed. by Leon Strous et al., pp. 5–19. DOI: [10.1007/978-3-030-64246-4_2](https://doi.org/10.1007/978-3-030-64246-4_2). URL: http://link.springer.com/10.1007/978-3-030-64246-4_2.
- [8] Wil M. P. van der Aalst. "Data Scientist: The Engineer of the Future". en. In: (2014). Ed. by Kai Mertins et al., pp. 13–26. DOI: [10.1007/978-3-319-04948-9_2](https://doi.org/10.1007/978-3-319-04948-9_2). URL: http://link.springer.com/10.1007/978-3-319-04948-9_2.
- [9] Joel C. Adams. "Creating a Balanced Data Science Program". In: SIGCSE '20 (2020), pp. 185–191. DOI: [10.1145/3328778.3366800](https://doi.org/10.1145/3328778.3366800). URL: <https://doi.org/10.1145/3328778.3366800>.
- [10] Wil van der Aalst. "Data Science in Action". en. In: *Process Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 3–23. ISBN: 978-3-662-49850-7 978-3-662-49851-4. DOI: [10.1007/978-3-662-49851-4_1](https://doi.org/10.1007/978-3-662-49851-4_1). URL: http://link.springer.com/10.1007/978-3-662-49851-4_1.

- [11] Katie Malone. “Doing Data Science on the Shoulders of Giants: The Value of Open Source Software for the Data Science Community”. en. In: *Harvard Data Science Review* (2020). DOI: 10.1162/99608f92.268cc8e4. URL: <https://hdsr.mitpress.mit.edu/pub/xsrt4zs2>.
- [12] Thomas Donoghue, Bradley Voytek, and Shannon E. Ellis. “Teaching Creative and Practical Data Science at Scale”. In: *Journal of Statistics and Data Science Education* 29.sup1 (2021), S27–S39. ISSN: null. DOI: 10.1080/10691898.2020.1860725. URL: <https://doi.org/10.1080/10691898.2020.1860725>.
- [13] Sean Kross and Philip J. Guo. “Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–14. ISBN: 978-1-4503-5970-2. URL: <https://doi.org/10.1145/3290605.3300493>.
- [14] D. Nolan and T. P. Speed. “Teaching Statistics Theory through Applications”. In: *The American Statistician* 53.4 (1999), pp. 370–375. ISSN: 0003-1305. DOI: 10.1080/00031305.1999.10474492. URL: <https://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474492>.

- [15] C. J. Wild and M. Pfannkuch. “Statistical Thinking in Empirical Enquiry”. en. In: *International Statistical Review* 67.3 (1999), pp. 223–248. ISSN: 1751-5823. DOI: [10.1111/j.1751-5823.1999.tb00442.x](https://doi.org/10.1111/j.1751-5823.1999.tb00442.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.1999.tb00442.x>.
- [16] Stephanie C. Hicks and Rafael A. Irizarry. “A Guide to Teaching Data Science”. en. In: *The American Statistician* 72.4 (2018), pp. 382–391. ISSN: 0003-1305, 1537-2731. DOI: [10.1080/00031305.2017.1356747](https://doi.org/10.1080/00031305.2017.1356747). URL: <https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1356747>.
- [17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: <https://www.R-project.org/>.
- [18] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [19] Sean Kross @ CSCW on Twitter. en. URL: <https://twitter.com/seankross/status/1387173741759582211>.
- [20] *GitHub: Build software better, together*. en. URL: <https://github.com/about> (visited on 12/07/2021).

- [21] *Open Case Studies* | *Bloomberg American Health Initiative*. en. URL: <https://americanhealth.jhu.edu/open-case-studies> (visited on 12/05/2021).
- [22] *W3Schools Free Online Web Tutorials*. en-US. URL: <https://www.w3schools.com/default.asp>.
- [23] Sean Kross. *ari: Automated R Instructor*. 2020. URL: <https://CRAN.R-project.org/package=ari>.
- [24] *How To Google Translate*. en-US. URL: https://www.w3schools.com/howto/howto_google_translate.asp.
- [25] *What is JavaScript? - Learn web development* | MDN. en-US. URL: https://developer.mozilla.org/en-US/docs/Learn/JavaScript/First_steps/What_is_JavaScript (visited on 12/07/2021).
- [26] *Google Translate - A Personal Interpreter on Your Phone or Computer*. URL: <https://translate.google.com/intl/en/about/>.
- [27] Chris Smith. *Case Studies in Geriatric Medicine and Patient Care*. en. URL: <https://www.hopkinsmedicine.org/gec/studies/>.
- [28] *US Department of Justice Case Studies*. en. 2016. URL: <https://www.justice.gov/olp/case-studies>.
- [29] Juliet Kaarbo and Ryan K. Beasley. "A Practical Guide to the Comparative Case Study Method in Political Psychology". en. In: *Political*

- Psychology* 20.2 (1999), pp. 369–391. ISSN: 0162-895X, 1467-9221. DOI: 10.1111/0162-895X.00149. URL: <https://onlinelibrary.wiley.com/doi/10.1111/0162-895X.00149>.
- [30] John J. Schmidt. *Counseling in schools: Essential services and comprehensive programs, 4th ed.* Needham Heights, MA, US: Allyn & Bacon, 2003. ISBN: 978-0-205-34056-9.
- [31] Sarah Crowe et al. “The case study approach”. en. In: *BMC Medical Research Methodology* 11.1 (2011), p. 100. ISSN: 1471-2288. DOI: 10.1186/1471-2288-11-100. URL: <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-11-100>.
- [32] RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, PBC. Boston, MA, 2020. URL: <http://www.rstudio.com/>.
- [33] Carrie Wright et al. *Open Case Studies: School Shootings in the United States (Version v1.0.0).* 2020. URL: <https://github.com//opencasestudies/ocs-bp-school-shootings-dashboard>.
- [34] JJ Allaire et al. *rmarkdown: Dynamic Documents for R.* 2021. URL: <https://github.com/rstudio/rmarkdown>.
- [35] Yihui Xie, J.J. Allaire, and Garrett Grolemond. *R Markdown: The Definitive Guide.* Boca Raton, Florida: Chapman and Hall/CRC, 2018. URL: <https://bookdown.org/yihui/rmarkdown>.

- [36] Yihui Xie, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook*. Boca Raton, Florida: Chapman and Hall/CRC, 2020. URL: <https://bookdown.org/yihui/rmarkdown-cookbook>.
- [37] Jonathan McPherson and JJ Allaire. *rsconnect: Deployment Interface for R Markdown Documents and Shiny Applications*. 2021. URL: <https://CRAN.R-project.org/package=rsconnect>.
- [38] *RStudio Connect*. en. URL: <https://rstudio.comhttps://www.rstudio.com/products/connect/> (visited on 12/07/2021).
- [39] *Load, Save, and .rda files | R-bloggers*. en-US. 2017. URL: <https://www.r-bloggers.com/2017/04/load-save-and-rda-files/> (visited on 10/28/2021).
- [40] Hadley Wickham. *httr: Tools for Working with URLs and HTTP*. 2020. URL: <https://CRAN.R-project.org/package=httr>.
- [41] Hadley Wickham, Jim Hester, and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*. 2021. URL: <https://CRAN.R-project.org/package=devtools>.
- [42] Hadley Wickham et al. *roxygen2: In-Line Documentation for R*. 2020. URL: <https://CRAN.R-project.org/package=roxygen2>.

- [43] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. 2021. URL: <https://yihui.org/knitr/>.
- [44] Yihui Xie. *Dynamic Documents with R and knitr*. 2nd. Boca Raton, Florida: Chapman and Hall/CRC, 2015. URL: <https://yihui.org/knitr/>.
- [45] Yihui Xie. “knitr: A Comprehensive Tool for Reproducible Research in R”. In: *Implementing Reproducible Computational Research*. Ed. by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman and Hall/CRC, 2014. URL: <http://www.crcpress.com/product/isbn/9781466561595>.
- [46] Hadley Wickham and Jennifer Bryan. *usethis: Automate Package and Project Setup*. 2021. URL: <https://CRAN.R-project.org/package=usethis>.
- [47] Lionel Henry and Hadley Wickham. *purrr: Functional Programming Tools*. 2020. URL: <https://CRAN.R-project.org/package=purrr>.
- [48] Emil Hvitfeldt. *textdata: Download and Load Various Text Datasets*. 2020. URL: <https://CRAN.R-project.org/package=textdata>.
- [49] *Google Forms: Free Online Surveys for Personal Use*. URL: <https://www.google.com/forms/about/>.
- [50] Jennifer Bryan. *googlesheets4: Access Google Sheets using the Sheets API V4*. 2021. URL: <https://CRAN.R-project.org/package=googlesheets4>.

- [51] *Google Sheets: Free Online Spreadsheets for Personal Use*. URL: <https://www.google.com/sheets/about/>.
- [52] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: [10.21105/joss.01686](https://doi.org/10.21105/joss.01686).
- [53] Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*. 2020. URL: <https://CRAN.R-project.org/package=magrittr>.
- [54] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: <https://ggplot2.tidyverse.org>.
- [55] Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. 2020. URL: <https://CRAN.R-project.org/package=ggpubr>.
- [56] *API Dimensions & Metrics | Google Analytics Data API*. en. URL: <https://developers.google.com/analytics/devguides/reporting/data/v1/api-schema>.
- [57] Mark Edmondson. *googleAnalyticsR: Google Analytics API into R*. 2021. URL: <https://CRAN.R-project.org/package=googleAnalyticsR>.

- [58] David Kahle and Hadley Wickham. “ggmap: Spatial Visualization with ggplot2”. In: *The R Journal* 5.1 (2013), pp. 144–161. URL: <https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>.
- [59] Joe Cheng, Bhaskar Karambelkar, and Yihui Xie. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. 2021. URL: <https://CRAN.R-project.org/package=leaflet>.
- [60] Carrie Wright et al. *Open Case Studies: Exploring CO2 emissions across time (Version v1.0.0)*. 2020. URL: <https://github.com/opencasestudies/ocs-bp-co2-emissions>.
- [61] Carrie Wright et al. *Open Case Studies: Opioids in the United States (Version v1.0.0)*. 2020. URL: <https://github.com/opencasestudies/ocs-bp-opioid-rural-urban>.
- [62] *CRAN Repository Policy*. URL: <https://cran.r-project.org/web/packages/policies.html>.
- [63] *Getting your R package on CRAN*. URL: https://kbroman.org/pkg_primer/pages/cran.html.
- [64] Guangchuang Yu. *hexSticker: Create Hexagon Sticker in R*. 2020. URL: <https://CRAN.R-project.org/package=hexSticker>.

- [65] Marcelo Ponce. *Visualize.CRAN.Downloads: Visualize Downloads from 'CRAN' Packages*. 2021. URL: <https://CRAN.R-project.org/package=Visualize.CRAN.Downloads>.
- [66] *Misunderstood Metrics: Google Analytics Users | Analytics Edge Help*. URL: <https://help.analyticsedge.com/article/misunderstood-metrics-users/>.
- [67] Milam Aiken. "An Updated Evaluation of Google Translate Accuracy". en. In: *Studies in Linguistics and Literature* 3.3 (2019), p253. ISSN: 2573-6426, 2573-6434. DOI: 10.22158/sll.v3n3p253. URL: <http://www.scholink.org/ojs/index.php/sll/article/view/2180>.
- [68] *RStudio Addins*. URL: <http://rstudio.github.io/rstudioaddins/>.
- [69] *UserTesting: The Human Insight Platform*. URL: <https://www.usertesting.com/>.

Michael Robert Breshock

Email: mbreshock@gmail.com • Website: mbreshock.github.io

Github: [mbreshock](https://github.com/mbreshock) • LinkedIn: [michaelbreshock](https://www.linkedin.com/in/michaelbreshock) • Twitter: [@michaelbreshock](https://twitter.com/michaelbreshock)

Research Interests

Computational neuroscience, machine learning/artificial intelligence, data science, biostatistics, cognitive science, brain computer interfaces

Education

2019 – Present	Johns Hopkins University – Baltimore, MD
(Expected	MSE in Biomedical Engineering — GPA: 3.92
Graduation	Focus in Neuroengineering and Biomedical Data Science
December	<i>Advisor: Carrie Wright, PhD (Department of Biostatistics, Johns Hopkins</i>
2021)	<i>Bloomberg School of Public Health)</i>

Selected Coursework

- *Cognitive Science*: Information coding in neural activity, Computational cognitive science
- *Computational Neuroscience*: Computational medicine in imaging, Neuro-image processing, Neural implants and interfaces, Auditory and vestibular systems
- *Data Science*: Statistical learning, Advanced data science for biomedical engineering, Intro to data science

2015 – 2019	Santa Clara University – Santa Clara, CA
	BS in Bioengineering, Medical Device Track — GPA: 3.73
	<i>magna cum laude</i>

Research Experience

- November 2020 – Present **Open Case Studies – Johns Hopkins Data Science Lab**
 Advisor: Carrie Wright, PhD (*Johns Hopkins University*).
Open Case Studies (OCS) is a repository of open source data science case studies on current public health problems using real data. My project is updating the current case studies by removing barriers, implementing interactive elements, and expanding reach. My master's thesis will analyze the impact these updates have on data science education inside and outside the classroom. See the Software section below for more about my work with OCS.
- June 2020 – November 2020 **Psychiatric Neuroimaging – Bakker Memory Lab**
 Mentors: Farah Naaz, PhD and Arnold Bakker, PhD (*Johns Hopkins University*).
 Developed a custom script in Matlab using Psychtoolbox to display random-dot-motion as a visual stimulus for subjects under MRI brain scan in psychiatric studies. Produced scripts and functions to programmatically convert data from MRI machines to the brain imaging data structure (BIDS) standard format.
- May 2018 – June 2019 **NebuFlask: A User-Friendly Nebulizer – BioInnovation and Design Lab**
 Advisor: Prashanth Asuri, PhD (*Santa Clara University*).
 Lead an interdisciplinary team of engineers to design and prototype an upgraded nebulizer for my senior design project. Identified user needs and areas for improvement, created a design to address these needs, and constructed a prototype. Final design made for a lightweight, quiet, rechargeable and discrete nebulizer that would allow for asthma patients to inhale their medications with ease.
- June 2015 – September 2017 **Center for Industrial and Medical Ultrasound – UW APL**
 Mentors: Tatiana (Tanya) Khokhlova, PhD and Julianna Simon, PhD (*University of Washington*).
 Studied the effects of pressure on the twinkling artifact phenomenon that occurs when imaging a kidney stone under ultrasound (see publication below). Gained experience programming in Matlab, testing medical devices, developing ex vivo experiment models, and managing laboratory hazards. Assisted in live animal studies and observed the experimental process with a variety of projects and faculty.

Industry Experience

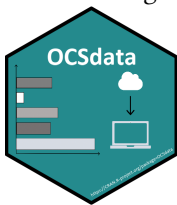
- June 2018 – June 2019 **InCube Labs (Engineering Intern)** – San Jose, CA
 Worked on a contract from Boston Scientific to develop a spinal cord stimulator (SCS) for the treatment of intractable, chronic pain. Supported the hardware and software teams' verification and validation process by identifying and removing bugs as well as drafting software specification requirements and test protocols.

Honors and Scholarships

- 2015 – 2019 University Honors Program (Santa Clara University)
 Member of the SCU University Honors Program (UHP) for the duration of my undergraduate career. UHP members must go above and beyond their normal degree requirements by taking special honors courses, participating in talks and seminars, completing an honors capstone and branching outside one's own field and culture.
- 2015 – 2019 Undergraduate scholarship, SCU Incentive Grant (Santa Clara University)
- 2018 Tau Beta Pi National Engineering Honor Society (Santa Clara University School of Engineering)
- 2018 Nominee, Barry Goldwater Scholarship (Santa Clara University)
- 2017 Finalist, Richard Osberg Undergraduate Research & Travel Fellowship (Santa Clara University)

Software

2021 Package



OCSdata

OCSdata is an R package to help you access and download case study data files hosted on the [Open Case Studies \(OCS\) GitHub](#). The package bridges the gap from web-browser to IDE, allowing users to automatically download the data they need with simple functions all within R. Now available on [CRAN](#), the Comprehensive R Archive Network.

2021 WebApp

MakeCaseStudies

MakeCaseStudies is a web-based application made with [Shiny](#). This dashboard enables [Open Case Studies](#) users to create their own digital lessons in a simple, user-friendly environment. DIY case studies.

See [Open Case Studies](#) above to learn more about the case studies.

Publications

- 2018 Simon JC, Sapozhnikov OA, Kreider W, **Breshock M**, Williams JC, Bailey MR. The role of trapped bubbles in kidney stone detection with the color Doppler ultrasound twinkling artifact. *Phys Med Biol.* 2018 Jan 9; 63(2): 025011. doi: [10.1088/1361-6560/aa9a2f](https://doi.org/10.1088/1361-6560/aa9a2f). PMID: 29131810; PMCID: [PMC5791757](https://pubmed.ncbi.nlm.nih.gov/PMC5791757/).

Teaching Experience

- June 2021 **Teaching Assistant, Introduction to R for Public Health Researchers (Johns Hopkins Bloomberg School of Public Health)**
Providing general assistance and troubleshooting errors as students practice coding in R using skills learned in lecture. Topics include basic syntax, data import, data wrangling, functions, packages, data visualization and statistical analysis.

Presentations and Posters

- May 2021 **Open Case Studies: an experiential lesson guide delivery platform**
Presentation *Provost's Teaching with Technology DELTA Symposium, Johns Hopkins University*
- June 2019 **NebuFlask: advancing usability of nebulizers to increase patient compliance**
Poster *University Honors Program Senior Poster Session, Santa Clara University*
- May 2019 **Nebufask: advancing usability of nebulizers to increase patient compliance**
Presentation *Senior Design Conference, Santa Clara University School of Engineering*
- February 2019 **Brain computer interface development for virtual reality applications**
Poster *School of Engineering Research Showcase, Santa Clara University*

Technical Skills

Programming languages

Proficient in: R, Python, MATLAB

Familiar with: Unix, SQL

Software

LaTeX, Git/GitHub, Google Analytics, Microsoft Office, Figma, Lucidchart

Projects

Data wrangling, statistical analysis, visualization, machine learning/artificial intelligence, package development, graphic design, web-based applications, continuous integration and deployment, experimental design

Personal Statement

I find the brain fascinating and want to further our understanding of it. I am excited by the discoveries made possible by computational models and analyses in cognition/neuroscience and want to contribute to these pursuits. Throughout my program I've focused my studies on neuroengineering and biomedical data science. The courses I've taken and my research experiences have allowed me to wrangle, analyze, and visualize real biological data including brain images, neural recordings, covid-19 cases and more. Post graduation, I hope to gain experience as a data scientist working with neurological data. Eventually, I plan to pursue a Ph.D in computational neuroscience.