# SPREAD AND DYNAMICS OF COVID-19: A DATA VISUALIZATION STUDY

Research Thesis

Presented in partial fulfillment of the requirements for graduation
with research distinction in Data Analytics
in the undergraduate colleges of The Ohio State University

by

Fan Fei

Undergraduate Program in Data Analytics
The Ohio State University

2021

Thesis Committee:
Asuman S. Turkmen, Advisor
Christopher M. Hans

# Abstract

We live in a data-driven era, in which data are continuously acquired for a variety of purposes, and the ability to make timely decisions based on available data is crucial. Therefore, visualization techniques have proven to be effective for not only presenting essential information in vast amounts of data but also driving complex analyses.

A pandemic of severe acute respiratory syndrome, coronavirus 2 (SARS-CoV-19) by the World Health Organization (WHO), emerged in Wuhan, China, and it is currently underway resulting in worldwide severe morbidity and mortality. Since the beginning of the pandemic, visualizations have assisted us to understand the unpredictable nature of this virus by providing detailed information based on data and variables on the ground. The objective of this study is to overview some of the state-of-art visualization techniques to illustrate the dynamics of the COVID-19 spread and to create a dashboard using Tableau on publicly available COVID-19 nationwide and Ohio data. Secondly, the pandemic has highlighted and exposed underlying structural inequalities in various areas of social, economic, civil, and political life. The impacts of COVID-19 are known to disproportionately affect certain racial, national, or ethnic communities and population groups. Therefore, another goal of this study is to provide descriptive data analysis for

COVID-19 data to understand how affection/death rate differ by features such as race and ethnicity.

## Table of Contents

## List of Tables

## List of Figures

# Chapter 1. Introduction

The ability to convey a large amount of data or text in a single image is increasingly useful in today’s online culture where colossal amounts of data are generated every single day. Data (information) visualization is an effective way of exploring complex patterns or large quantities of data that cannot be easily perceived by looking at a table with numbers or reading a text. Visual elements such as charts and graphs provide powerful representations of huge amounts of data points that the human brain can process and see analytics presented visually. Data visualization can be used either as an exploratory tool before any statistical analysis to identify what to focus on during analysis or as a communicative tool to present our analysis findings to an audience. Nonetheless, data visualizations help users to grasp difficult concepts or identify new patterns, and therefore they can more effectively answer questions, tell stories, and put forth arguments than words alone [Card et al., 1999; Few, 2009; Tufte, 1983; Tufte, 1990; Tufte, 1997].

Over the past decade, the field of visualization has matured; a variety of techniques for a numerous of data types have been developed to solve problems in several domains. This chapter provides an overview of the intellectual history of data visualization followed with a summary of existing techniques categorized according to some criteria.

## 1.1. History of Data Visualization

Although data visualization is a relatively modern development, it is certainly not a new concept in human history. The earliest examples of visualization were geometric diagrams, tables with the positions of stars, and maps to help in navigation and exploration. Egyptian surveyors were first to use idea of coordinates to lay out towns by something like latitude and longitude [Friendly, 2006]. Funkhouser [1936] described an anonymous 10th century multiple time-series graph of the changing position of the seven most prominent heavenly bodies over space and time depictions of quantitative information (Figure 1).

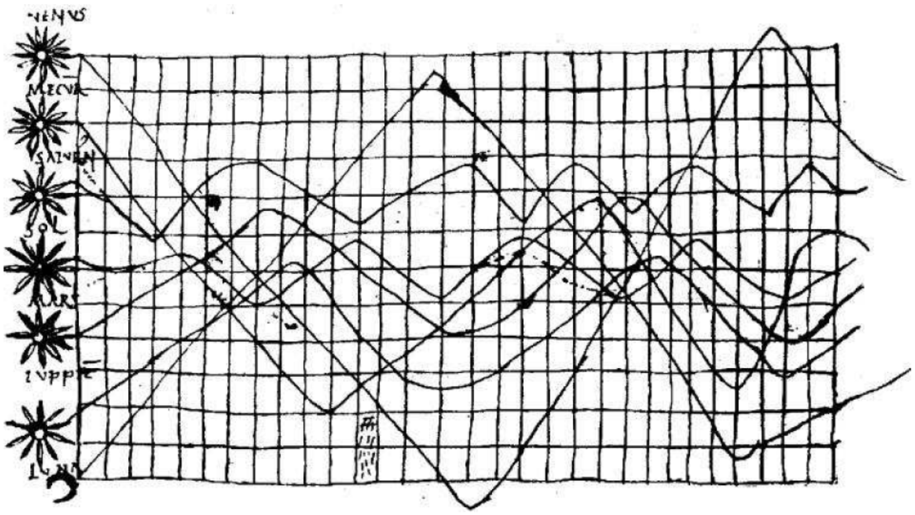


**Figure 1.** A 10th century plot showing planetary movements as cyclic inclinations over time (Source: Funkhouser, 1936, page 261).

The development of triangulation to determine mapping locations accurately, use of captured images directly, and the first modern cartographic atlas were important developments of the16th century [Friendly, 2006]. The 17th century was giving rise to the beginnings of visual thinking, as illustrated by the examples of Scheiner (who introduced the principle of "small multiples" [Tufte, 1983] to show the changing configurations of sunspots over time) and van Langren (a Flemish astronomer to the court of Spain believed to be the creator of the first visual representation of statistical data) [Tufte, 1997]. William Playfair, who is considered as the inventor of most of the graphical forms widely used today, introduced graph and bar chart [Playfair, 1786], later the pie chart and circle graph [Playfair, 1801]. During cholera epidemics in 1848–1849 and 1853–1854 resulting in thousands of deaths, Dr. John Snow produced his famous dot map [Snow, 1855] showing deaths due to cholera clustered around the Broad Street pump in London leading the discovery of water-born cause of the disease.

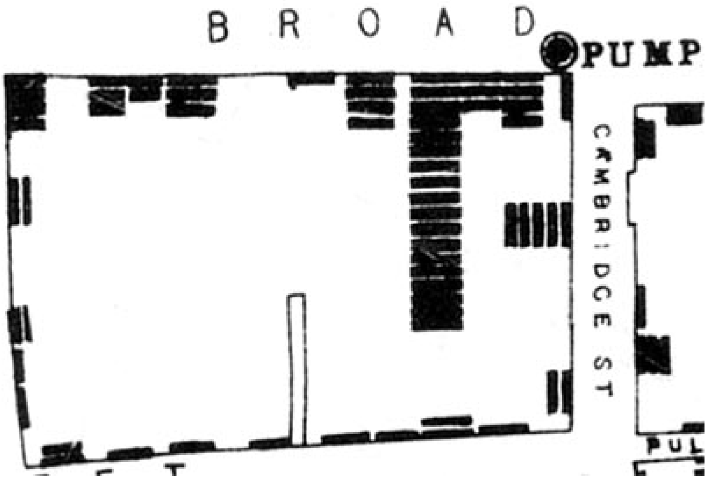


**Figure 2.** Dr. Snow’s dot map showing deaths due to cholera clustered around the Broad Street pump in London. Source: https://www.ph.ucla.edu/epi/snow/mapsbroadstreet.html

Between 1850 and 1900 is the golden age of statistical graphics. Some significant developments of the era include but not limited to 3D surface plots [Perozzo, 1880], circle diagrams [Minard, 1861], polar area charts [Nightingale, 1857], semi-logarithmic graphs [Jevons, 1879, 1958]. Francis Galton plays very important role in the development of the ideas of correlation and regression. He discovered the “anti-cyclonic” (counterclockwise) pattern of winds around low-pressure regions, combined with clockwise rotations around high-pressure zones [Galton, 1863]. He also used “isodic curves” to portray the joint effects of wind and current on the distance ships at sea could travel in any direction [Galton, 1881].

While the late 1800s were the “golden age” of statistical graphics and thematic cartography, the early 1900s can be called the “modern dark ages” of visualization with only a few graphical innovations [Friendly and Denis, 2000]. The visualization began to rise again in the mid-1960s with significant intersections and collaborations of computer science research at Bell Laboratories and with the developments in exploratory data analysis (EDA) introduced by Tukey [Tukey and Tukey, 1985]. During the end of the 20th century, data visualization has blossomed into a multi-disciplinary research area. All these developments have led to new paradigms, languages, and software packages for expressing statistical ideas and implementing data visuals and, eventually, result in an explosive growth in new visualization methods and techniques.

### 1.2.Visualization Methods

In the past decade, a variety of approaches have been introduced to visually convey data. Keim and Kriegel [1996, 1997] categorized visual data exploration techniques for data into six classes, namely geometric, icon-based, pixel-oriented, hierarchical, graph-based and hybrid techniques. Another study by Chan [2006] used similar ideas to classify multivariate data visualization techniques into four broad categories. We will adopt Keim and Kriegel [1996, 1997] and explain these six categories in detail.

Geometric projection techniques aim at finding informative projections and transformations of multidimensional datasets [Keim and Kriegel, 1996]. It generally maps the attributes to a typical Cartesian plane like scatterplot, or more innovatively to an arbitrary space such as parallel coordinates. The techniques in this category are ideal for detecting outliers and correlation amongst different dimensions and handling large datasets. This class includes exploratory statistics techniques typically used for data processing, such as principal component analysis, factor analysis, and multidimensional scaling. Scatter plots, prosection matrix [Furnas and Buja, 1994], hyperslice [van Wijk and R. van Liere, 1993], parallel coordinates [Inselberg, 1997], and Andrews Curve [Andrews, 1972] are some examples falling into this category. An example of a scatter plot and a parallel coordinate plot can be seen in Figure 3 (a) and (b), respectively. The main limitation of the methods in this category is the difficulty of visualization beyond 3-dimensional data.

The icon-based or iconographic display techniques map each multidimensional data item to an icon (or glyph) whose visual features vary depending on the data values by which users can study the overall features and relationships in the data. One of the first approaches is the Chernoff faces technique [Chernoff, 1973]. Two data attributes are mapped to the 2D position of a face icon in the display and the remaining attributes are mapped to the properties of the face icon.

In pixel-oriented techniques, a pixel is used to represent each data item, and different attributes are exhibited in different sub-windows where the range of possible data values are mapped to pixels according to a fixed color map [Keim, 2000]. The idea is to represent an attribute value by a pixel based on some color scale. For an *n*-dimensional dataset, *n*-colored pixels are used to represent one data item, with each attribute values being placed in separate sub-windows. The techniques are further categorized as “query independent” (the arrangement of the pixels in the sub-windows is fixed) or “query dependent” (a query item is provided and distances from the data values to the given query value are computed using some metrics).

Hierarchical techniques subdivide the data space and present subspaces in a hierarchical fashion. Well-known representatives of hierarchical visualization techniques are Dimension Stacking [LeBlanc et al., 1990], Treemaps [Shneiderman, 1992], and Cone Trees [Robertson et al., 1991]. The techniques in this category concern mainly hierarchical data, or data in which several attributes are more important to users or of more interest. While color has been used extensively to encode an addition dimension, it is also very common to replace its role by textures that obviously provide more graphical attributes for higher dimensional data.

Graph-based techniques visualize large graphs using specific layout algorithms, query languages, and abstraction techniques to convey their meaning clearly and quickly [Battista et al., 1994]. There are several approaches and systems targeted at this specific domain, which appear in Card et al. [1999]. Finally, hybrid techniques integrate multiple visualization techniques, either in one or multiple windows, to enhance the expressiveness of the visualizations. Linking between visualization windows is a useful resource and most techniques rely heavily on dynamics and interaction.

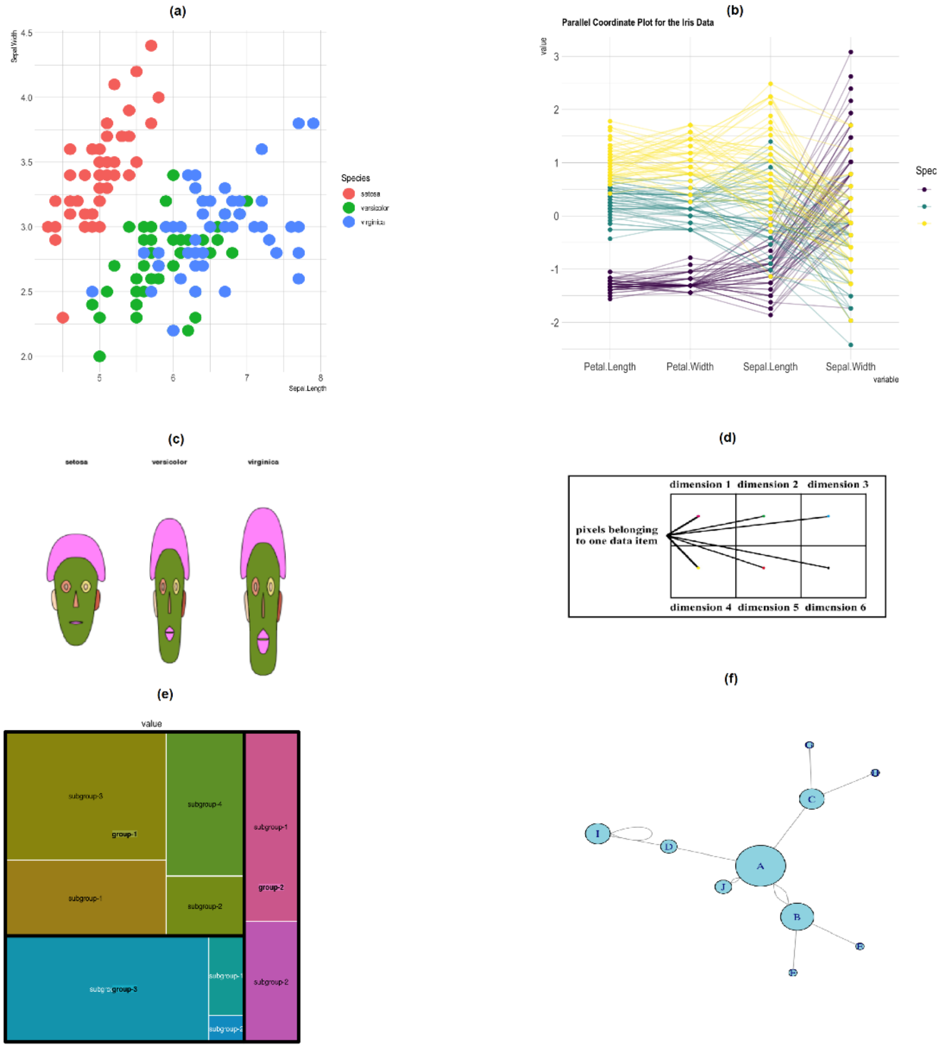


**Figure 3.** An illustration of graphical techniques: (a) scatter plot, (b) parallel coordinates, (c) Chernoff faces, (d) pixel plot, (e) treemap, (f) a network graph. Source: https://www.r-graph-gallery.com/ (for a, b, c, e, and f) and Keim, 2000 for (d).

Although there are plenty number of options described here, how well you represent data and select an appropriate tool based on your needs are very important. Abundance of available methods does not necessarily alleviate common pitfalls in visualization, which can inhibit the ability of readers to effectively understand the information presented. Nonetheless, data visualization faces some challenges including but not limited to finding a suitable mapping of high-dimensional data into a two-dimensional visual form, presenting dense structure of data in a single visual display that enables users to explore the data space interactively while discriminating individual dimensions and assessing effectiveness of a visualization technique to determine what knowledge is present in the data and what insight is gained by visualizing it. Therefore, developing effective and innovative data visuals is still a challenging task posing many interesting research questions which need to be solved.

Kelleher and Wagener [2011] provide some guidelines so that visualizations can effectively convey information and reduce pitfalls mentioned earlier. According to Kelleher and Wagener [2011], a good visual should

- create the simplest graph that conveys the information needs to be conveyed,
- consider the type of encoding object and attribute used to create a plot,
- focus on visualizing patterns or on visualizing details, depending on the purpose of the plot,
- select meaningful axis ranges,
- plot overlapping points in a way that density differences become apparent in scatter plots,
- use lines when connecting sequential data in time-series plots,

- aggregate larger datasets in meaningful ways,
- keep axis ranges as similar as possible to compare variables,
- select an appropriate color scheme based on the type of data.

Another important factor for creating effective and impactful data visualizations is the use of appropriate programming languages / software tools. To date, there are dozens of applications, tools, and scripts available to create successful visualizations of large datasets, and consequently, programming inclines to become a popular method to realize data visualization. Many are very basic and have a lot of overlapping features. However, there are standouts that either have more capability for the types of visualizations they can create or are significantly easier to use than the other available options. Tableau Public [https://public.tableau.com/en-us/s/] is a very popular software tool that could easily create dashboards and panels without advanced coding skills. The public version of Tableau is free to use for anyone looking for a powerful way to create data visualizations that can be used in a variety of settings. They have an extensive gallery of infographics and visualizations. For those with programming experience, R with the open-source R package Shiny, JavaScript, and Python are among popular programming languages offering special capacities to data visualization. Several libraries based on different programming languages have been constructed and shared in the world, which turns data presentation into simple coding.

Undoubtedly, significant development has been made in the field of data visualization in the past decades. However, there still exists many interesting directions deserving to explore, and data visualization will continue to be a prominent research area in future.

## 1.3.Organization of Thesis

The rest of the thesis is structured in three chapters. Chapter 2 describes how data visualizations have been front-and-center in the efforts to communicate the science around Coronavirus disease 2019 (COVID-19). This chapter briefly explains three well-known dashboards along with their comparisons. In Chapter 3, dashboards for the national and Ohio data are created using Tableau and a descriptive statistical analysis is carried out. Finally, Chapter 4 gives conclusions.

## Chapter 2. COVID-19 and Data Visualization

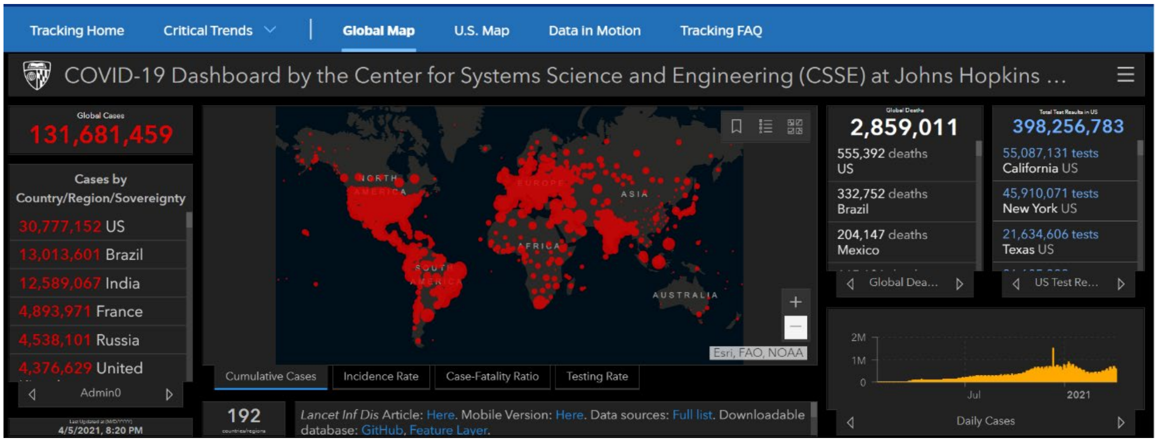
COVID-19, a previously unknown respiratory illness caused by the coronavirus SARS-CoV-2 was declared a pandemic by the World Health Organization (WHO) on March 11, 2020 [WHO, 2020]. Since then, visualization techniques have been front-and-center in the efforts to communicate the science around COVID-19 to the very broad audience of policy makers, scientists, healthcare providers, and public.

The COVID-19 pandemic has put the use of dashboards under the global spotlight. Many dashboards and reporting tools show data simply as a set of one or more basic charts, such as the bar plot, line, or pie chart. In response to this ongoing health emergency, several online interactive dashboards are developed to visualize and track reported cases of COVID-19 in real time. A few good examples of these dashboards that offer insight into what is happening in the United States and other countries include but not limited to John Hopkins University (JHU) Coronavirus Research Center [https://coronavirus.jhu.edu/], WHO [https://covid19.who.int/], and Centers for Disease Control and Prevention (CDC) [https://covid.cdc.gov/covid-data-tracker/#datatracker-home] dashboards. In addition to these national and global resources, many colleges have created dashboards showing their COVID-19 infection and testing data such as Amherst College, Wagner College, and the Ohio State University.

In this chapter, an overview of these three popular dashboards is provided along with a comparison chart to explore similarities and differences among them.

## 2.1 JHU Coronavirus Resource Center Dashboard

The dashboard developed by JHU was the first to track and display information on cases and death totals for different countries and states in the United States [https://coronavirus.jhu.edu/]. The dashboard is cited frequently by public health and government officials. It is run by more than two dozen people and funded by Johns Hopkins University, Bloomberg Philanthropies, and the Stavros Niarchos Foundation. It uses real-time mapping software provided by the company Esri.



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins ...

Global Cases
131,681,459

Cases by Country/Region/Sovereignty
30,777,152 US
13,013,601 Brazil
12,589,067 India
4,893,971 France
4,538,101 Russia
4,376,629 United

Last Updated at (M/D/YYYY)
4/5/2021, 8:20 PM

192
countries/regions

*Lancet Inf Dis* Article: Here. Mobile Version: Here. Data sources: Full list. Downloadable database: GitHub, Feature Layer.

Global Deaths
2,859,011
555,392 deaths US
332,752 deaths Brazil
204,147 deaths Mexico

Total Test Results in US
398,256,783
55,087,131 tests California US
45,910,071 tests New York US
21,634,606 tests Texas US

**Figure 4.** The Johns Hopkins' COVID-19 Dashboard that was captured on April 5, 2021.

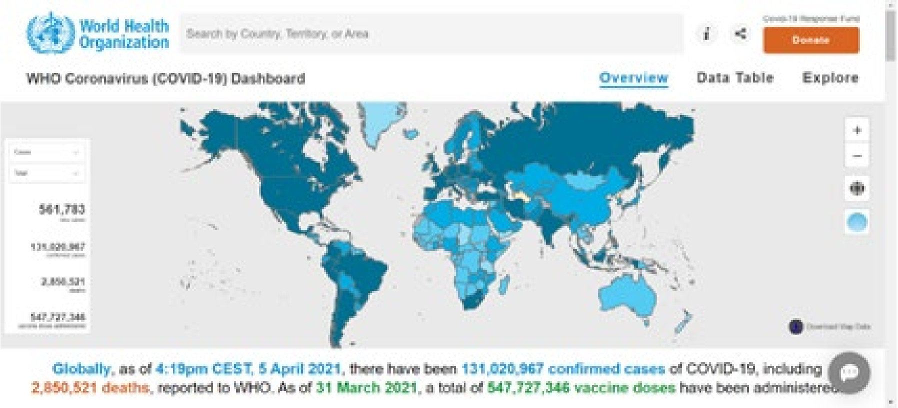
Data are provided from a variety of sources including CSSE GitHub and CCI GitHub. Using these data, JHU Coronavirus Resource Center created the world map dashboard shown in Figure 4 that highlights confirmed cases of COVID-19 by country and state, as well as other key counts such as total death (by country) and people tested (by state). The other two tabs within the dashboard offer a deeper dive into the states and their counties to allow users to find answers to questions such as whether the curve flattened with supporting evidence and visualizations. In addition, with the user-interactive feature, people can switch from viewing the cumulative cases throughout the world into incidence rate, case-fatality ratio, and testing rate by simply clicking on the buttons below the map.

### 2.2. WHO Dashboard

The World Health Organization dashboard [https://covid19.who.int/] is among the most accessed and relied upon visualizations in the COVID-19 pandemic. Everyone can navigate the WHO dashboard easily by exploring numbers of infected, deaths, and recoveries around the world. In addition, the WHO dashboard assists users to analyze data about COVID-19 in real-time as graphics change depending on the current situation. Figure 5 shows a snapshot of the dashboard. The overview section of COVID-19 on the WHO dashboard offers information about tracking guidelines to enable understanding of the pandemic while the explorer section guides the public on navigating through the dashboard and helping them make sense of the numbers on COVID-19. The data displayed in this dashboard are available for download as comma-separated values (CSV) files [https://covid19.who.int/info].

### 2.3. CDC COVID Data Tracker

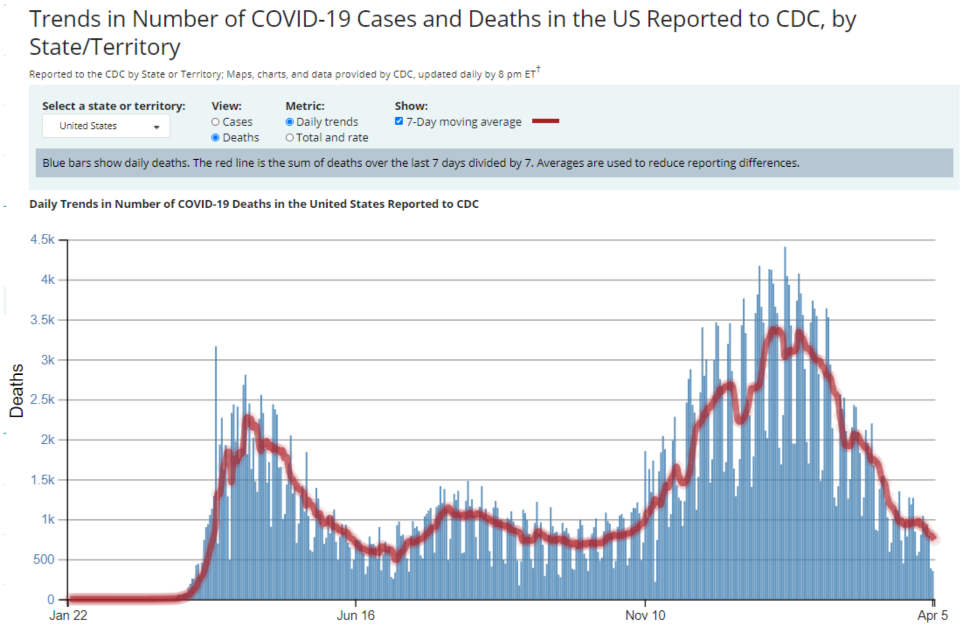
The Center for Disease Control and Prevention (CDC) is the United States national health protection agency, and a reliable resource for information about COVID-19. During the week of May 9, 2020, the Centers for Disease Control and Prevention launched a new COVID-19 data dashboard [https://covid.cdc.gov/covid-data-tracker/#datatracker-home] that illustrates the number of cases, deaths and people tested, school closures and other social impacts of the pandemic. It also includes state-by-state status of business openings/closings, healthcare facilities, state of emergency status, lock-down/shelter in place status. Figure 6 shows trends in number of COVID-19 cases and deaths in the United States. While referring to an individual states' website for updates is always the best course of action in keeping up to date with how each state is handling COVID-19, this is a good place to compare how all states are handling issues.


World Health Organization
Search by Country, Territory, or Area
Covid-19 Response Fund
Donate
WHO Coronavirus (COVID-19) Dashboard
Overview
Data Table
Explore
561,783
131,020,967
2,850,521
547,727,346
+
−

Globally, as of 4:19pm CEST, 5 April 2021, there have been 131,020,967 confirmed cases of COVID-19, including 2,850,521 deaths, reported to WHO. As of 31 March 2021, a total of 547,727,346 vaccine doses have been administered.

**Figure 5.** The WHO COVID-19 Dashboard that was captured on April 5, 2021.

To obtain timely and detailed data on COVID-19 cases in the United States, CDC uses two data sources. The first data source is an aggregate count based on a robust, multistep process to collect data and confirm the case and death numbers with jurisdictions daily. The second data source involves line-level data for each case, which provide additional information about whether the patient died and other details such as age and race and ethnicity. CDC receives the line-level data primarily from state health departments without personal identifiers such as names or home addresses.



Trends in Number of COVID-19 Cases and Deaths in the US Reported to CDC, by State/Territory

Reported to the CDC by State or Territory; Maps, charts, and data provided by CDC, updated daily by 8 pm ET†

Select a state or territory: United States

View: ○ Cases ● Deaths

Metric: ● Daily trends ○ Total and rate

Show: ☑ 7-Day moving average

Blue bars show daily deaths. The red line is the sum of deaths over the last 7 days divided by 7. Averages are used to reduce reporting differences.

Daily Trends in Number of COVID-19 Deaths in the United States Reported to CDC

**Figure 6.** The CDC COVID Data Tracker that was captured on April 5, 2021.

## 2.3. Comparing Existing Dashboards

A summary of the COVID-19 data sources is shown in Table 1. Data are updated at the end of each day in all cases except for the WHO, since they update some data weekly. The WHO and Johns Hopkins provide easy-to-access download portals, while the CDC only provides a dashboard without an option to download the data.

All of them has the user-interactive feature that allows the user to choose what they want to view an indicator of the fact that the data visualization technique has been largely improved in the last decade. WHO and JHU choose to use table as one of the visualization expression ways because viewers can directly see the comparison in clear numbers, and both the websites includes the worldwide results while CDC only has nationwide data. For JHU and CDC, people who live in the United States can narrow down and focus on the information about the county that they live. Lastly, only CDC dashboard includes the racial difference as a factor.

**Table 1.** A comparison of the popular COVID-19 dashboards.

| | WHO | John Hopkins | CDC |
|---|---|---|---|
| Interactive feature | Yes | Yes | Yes |
| Update status | Daily/Weekly | Daily | Daily |
| Open Access | Yes | Yes | No |
| Race factor | No | No | Yes |
| County Summary | No | Yes | Yes |
| Table | Yes | Yes | No |
| Vaccination | Yes | No | Yes |
| Region | Worldwide | Worldwide | Nationwide |

In summary, a vast number of dashboards, spanning across multiple fields of interest, with varying data sources, different contexts and levels of focus, purposes, have been developed recently so that rapidly growing COVID-19 data can be presented in a visually appealing manner. While there are several standout dashboards, there is no one-size-fits-all template or model to accomplish a perfect visualization.

## Chapter 3.  COVID-19 Data Analysis: An Exploratory Study

This chapter focuses on creating visualizations for real COVID-19 datasets for the United States and Ohio using Tableau Public [https://public.tableau.com/en-us/s/]. Descriptive data analyzes are provided to understand impact of COVID-19 on different ethnicity and race groups.

### 3.1. Data Description

This study analyzed the data obtained from the COVID Racial Data Tracker [https://covidtracking.com/race] that advocates for, collects, publishes, and analyzes racial data on the pandemic across the United States. This is a collaboration between the COVID Tracking Project [https://covidtracking.com/] and the Boston University Center for Antiracist Research [https://www.bu.edu/antiracism-center/the-center/]. The race and ethnicity data from 51 states have been launched on April 15, 2020 and have been updated every week. The dataset consists of daily numbers of cases, deaths, hospitalizations for different races in the United States.

The Ohio COVID-19 data are downloaded from the Ohio Department of Health (ODH) Dashboard that displays the most recent preliminary data reported to the ODH about cases, hospitalizations, and deaths in Ohio by selected demographics (age and sex) and county of residence.

The population estimates for 2020 are from the 2017 population projections for the United States based on the 2010 Census [https://www2.census.gov/programs-surveys/popproj/datasets/2017/2017] popproj/np2017_d1_mid.csv].

### 3.2. Descriptive Data Analysis

Using data extracted from the COVID Racial Data Tracker, the infection and mortality rates for COVID-19 (out of 100,000 people) are calculated for each race/ethnicity group separately and together between April and December 2020. The rates are calculated by dividing the number of cases/deaths by the population estimate of the United States in 2020 only considering 51 states (including District of Columbia). The grouped bar plots in Figure 7 clearly illustrates rising trend for both rates.

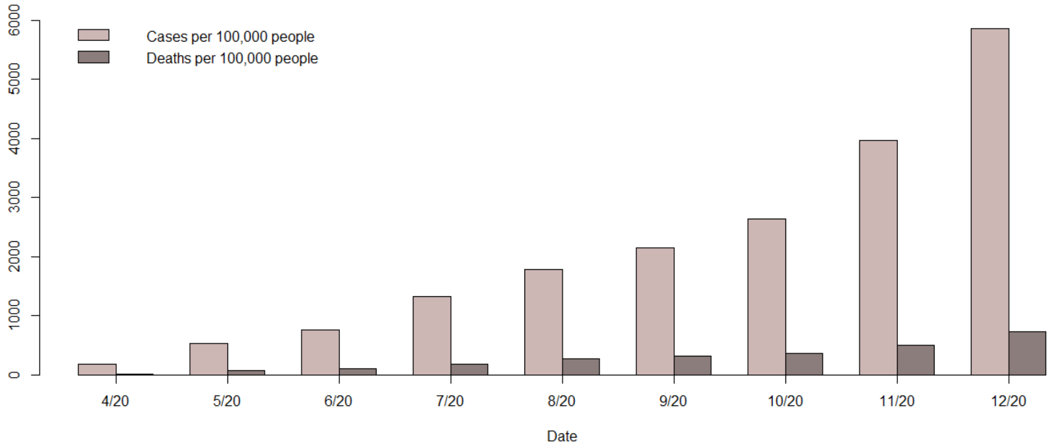


**Figure 7.** Grouped bar plots for the number of cases/deaths out of 100,000 between April and December, 2020.

Given the reported health disparities in COVID-19 infection and mortality by race/ethnicity [DiMaggio et al., 2020], these rates are also calculated for each race/ethnicity specifically

White, Black, Latin/Hispanic, American Indian and Alaska Native (AIAN), and Asian groups.

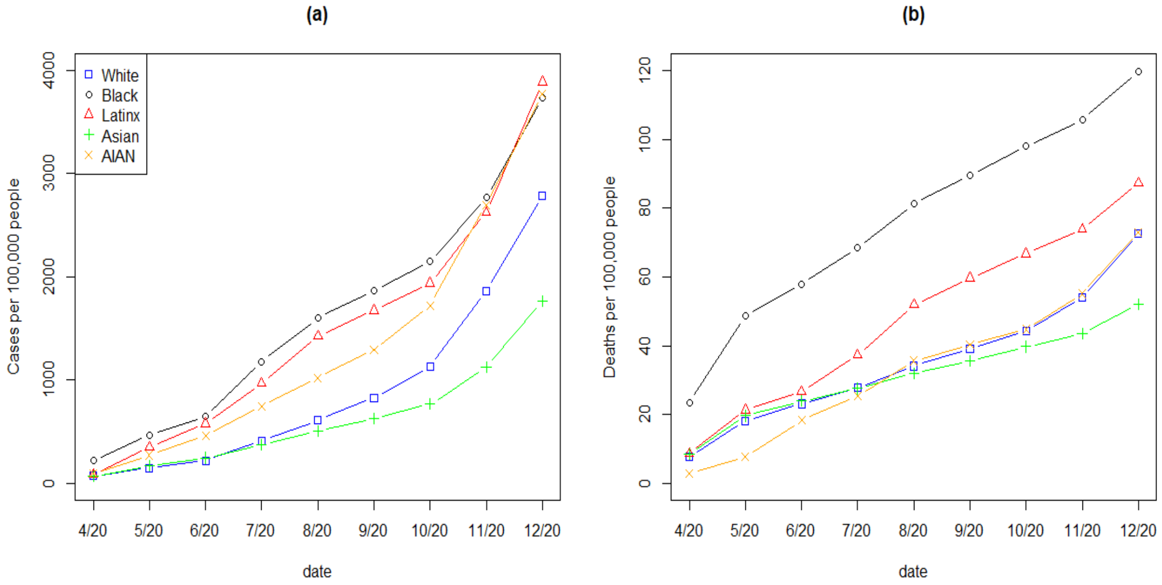


**Figure 8**. COVID-19 (a) infection, (b) mortality rates out of 100,000 for different race groups between April and December, 2020.

Figure 8 illustrates infection and mortality rates showing the enormous disparity in between Black and White individuals. In general, all groups other than Asians had higher rates of infection. The mortality rates for Black individuals were clearly distinguishable from those of others. In both figures, Black individuals are followed by Latin/Hispanic individuals. Figure 9 shows testing rates for each race group. White individuals have higher rate of testing and given that this group has the lowest infection rate after Asians, this can be considered as an indicator of easy health access. Although testing rate is also high for Black individuals, it is reasonable given that the infection rate is the highest in this group. Another important point is that AIAN individuals have the lowest testing rate while they are third race with the highest infection rate. These potential testing access barriers also reveal some ethnicity- and income-based healthcare disparities.

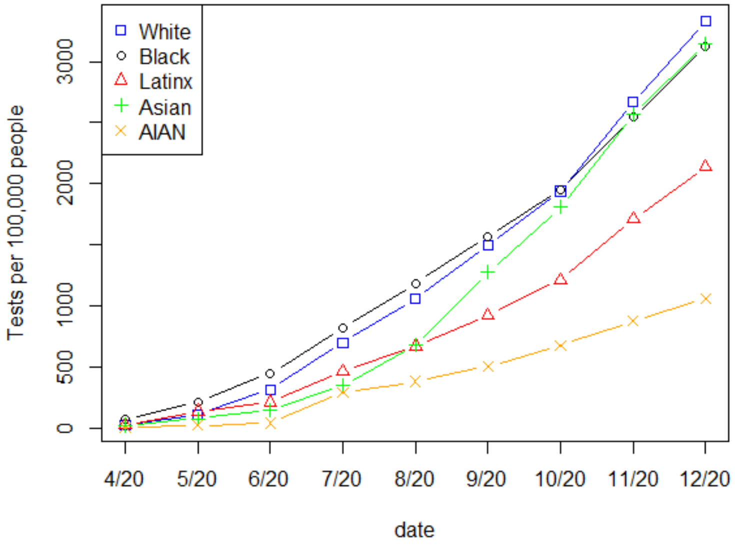


**Figure 9.** COVID-19 testing rates out of 100,000 for different race groups between April and December, 2020.

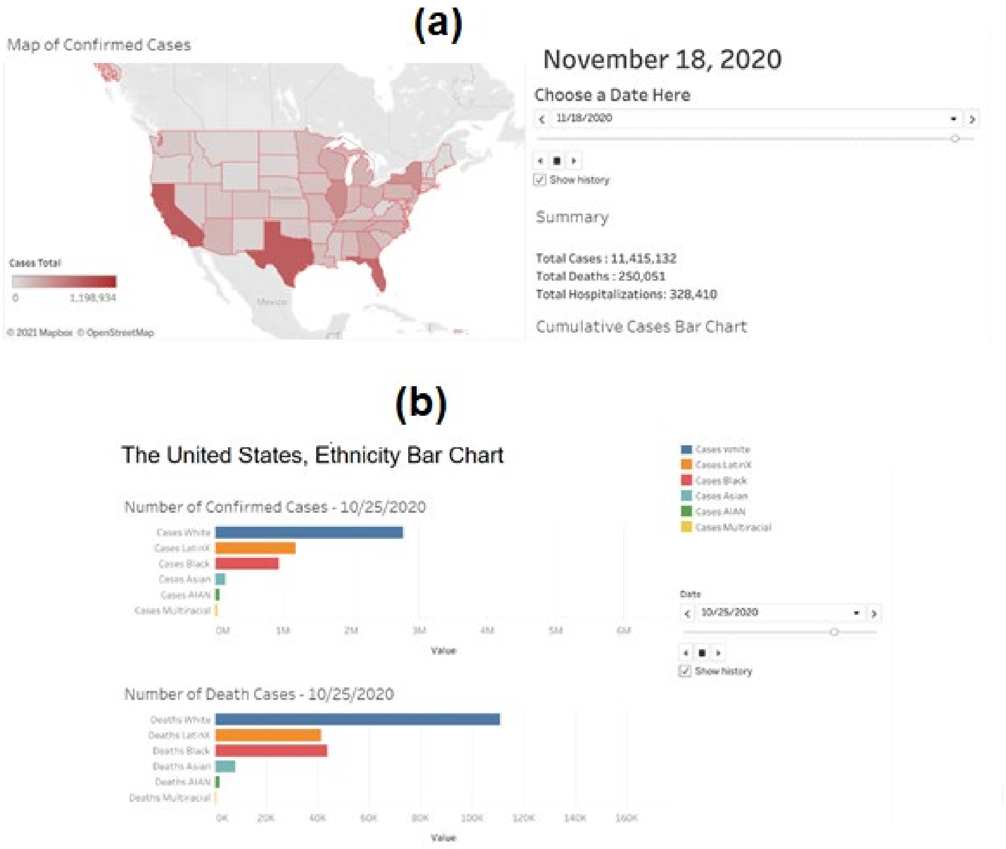
This analysis has indicated considerable disparities in the prevalence of COVID-19 across racial/ethnic subgroups of the population in the United States. Specifically, infection and death rate are disproportionately high for the Black population. In addition, it is observed high infection/death rates for Hispanic and AIAN populations. These racial/ethnic health disparities in the risk of infection (and mortality) are closely associated with less access to health care and adverse economic conditions. To achieve health equity, barriers must be removed so that everyone has a fair opportunity to be as healthy as possible.

### 3.3. Tableau Dashboards

In this subsection, 4 statewide and Ohio COVID-19 data visualizations are presented. These visuals are created using Tableau Public [https://public.tableau.com/en-us/s/] and can be accessible via https://public.tableau.com/profile/fan5584#!/

"COVID-19 Dashboard for US" and "The United States, Ethnicity Bar Chart" visuals use the nationwide data, and snapshots of these interactive plots can be seen in Figure 10 (a) and (b), respectively. They are both interactive plots. "COVID-19 Dashboard for US" allows one to see total number of cases, deaths, and hospitalizations along with a color-scaled and a bubble state map. "The United States, Ethnicity Bar Chart" illustrates a bar chart for numbers of cases and deaths among White, Black, Latin/Hispanic, Asian, AIAN, and other ethnic/race groups for available range of dates between 4/12/2020 and 12/20/2020.

Second set of visuals, named "COVID Dashboard for Ohio State" and "Ohio County Bar Chart", only represents the data from the ODH. The state dashboard gives total number of cases and deaths for 88 counties in Ohio between 3/9/2020 and 12/21/2020. Users can interactively click on each county to reveal number of cases and deaths. A bar plot representation of cumulative data can also be found on the same dashboard. "Ohio County Bar Chart" shows top 15 counties with the highest infection rate where users can change the date interactively.



11/18/2020

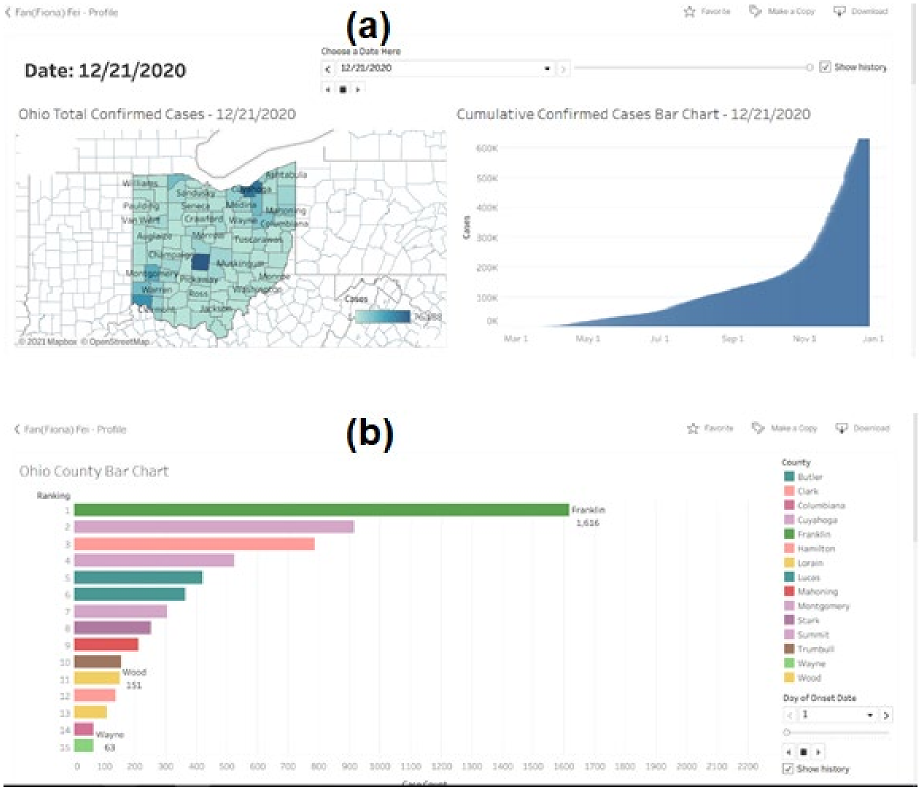
☑ Show history

Total Cases : 11,415,132
Total Deaths : 250,051
Total Hospitalizations: 328,410

Date: 10/25/2020

☑ Show history

**Figure 10.** Tableau visuals for (a) the confirmed cases, deaths, and hospitalizations, (b) the confirmed cases and deaths for different race/ethnicity groups in the United States between April and December, 2020.



Date: 12/21/2020

Choose a Date Here: 12/21/2020

☑ Show history

Day of Onset Date: 1

☑ Show history

**Figure 11.** Tableau visuals for (a) the confirmed cases and deaths on a color-scaled map between March and December, 2020, (b) a bar plot for the confirmed cases for different counties in Ohio in December, 2020.

## Chapter 4. Summary and Conclusions

The amount of available data continues to grow in an exponential rate and innovative tools to visualize such data is a key step towards understanding data. Visualization research over the past decades has discovered a wide range of effective visualization techniques that go far beyond the basic pie, bar and line charts used so pervasively in spreadsheets and dashboards. COVID-19 pandemic has shown the importance of dashboards and visualization for everyone. Various countries and even states/regions have created their own COVID-19 dashboards to educate people on the pandemic.

In this study, an overview of the intellectual history of data visualization and a summary of existing visualization techniques classified into six categories (geometric, icon-based, pixel-oriented, hierarchical, graph-based and hybrid techniques) are given. The fundamental role of dashboards in the efforts to communicate the science around COVID-19 is emphasized, and three popular dashboards are briefly described along with their comparisons. In this thesis, dashboards for the national and Ohio data are created using Tableau Public and a descriptive statistical analysis pointing out disparities among race/ ethnicity groups is carried out.

Despite many significant developments in the field of data visualization, there still exists numerous interesting directions deserving to explore. In this thesis, the focus was only exploratory data analysis and visualization. In future, the purpose is to utilize a data with more variables such as vaccine, age, sex that can be gather from existing COVID-19 data hubs, and to create online dashboards that update data automatically so that users of visualization in all disciplines can explore, communicate, and understand their results in real-time.

## Bibliography

Andrews, D. F.  (1972) “Plots of high-dimensional data,” *Biometrics*, 28(1), 125-136.

Battista, G.D., Eades, P., Tamassia, R., Tollis, I.G. (1994). “Annotated bibliography on graph drawing,” *Computational Geometry: Theory and Applications*, 4(5), 235-282.

Boston University Center for Antiracist Research, https://www.bu.edu/antiracism-center/the-center/ [Online].

Chernoff, H. (1973) “The use of faces to represent points in k-dimensional space graphically,” *Journal of the American Statistical Association*, 68, 361-368.

Card, S.K., Mackinlay, J.D., Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. Academic Press, San Diego, CA.

CDC COVID Data Tracker. Retrieved April 5, 2021 from https://covid.cdc.gov/covid-data-tracker/#datatracker-home [Online]

Chan W. W. Y. (2006). “A survey on multivariate data visualization,” Department of Computer Science and Engineering Hong Kong University of Science and Technology Clear Water Bay, Kowloon, Hong Kong.

DiMaggio, C., Klein, M., Berry, C., Frangos, S. (2020). “Blacks/African Americans are 5 times more likely to develop COVID-19: spatial modeling of New York City ZIP code-level testing results,” *Epidemiology*, 51, 7-13

Few, S. (2009). *Now You See It*. Analytics Press, Oakland, USA.

Friendly M. (2006). *A Brief History of Data Visualization*. Springer.

Friendly, M. and Denis, D. (2000). “The roots and branches of statistical graphics,” *Journal de la Soci´et´e Franc¸aise de Statistique*, 141(4), 51–60. (published in 2001).

Funkhouser, H. G. (1936). “A note on a tenth century graph,” *Osiris*, 1, 260–262.

Furnas, G. W.  and Buja, A. (1994). “Prosection views: dimensional inference through sections and projections,” *Journal of Computational and Graphic Statistics*, 3(4), 323-353.

Galton, F. (1863). *Meteorographica, or, Methods of Mapping the Weather*. London: Macmillan.

Galton, F. (1881). “On the construction of isochronic passage charts,” *Proceedings of the Geographical Section of the British Association*, n.v.(XI), 657, 704. published: *Roy. Geog. Soc. Proc.*, 1881, 657-658.

Inselberg, A. (1997). “Multidimensional detective,” *Proceedings of the IEEE Symposium on Information Visualization*, 100-107.

Jevons, W. S. ([1879] 1958). Graphical method. In *Principles of Science: A Treatise on Logic and Scientific Method*, (pp. 492–496). New York: Dover, 3rd edn. First ed.: 1874; page numbers from 3rd Ed. Dover reprint (1958).

Keim, D.A. (1997). "Visual techniques for exploring databases," *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining Tutorial Program.*

Keim, D.A. (2000). "Designing pixel-oriented visualization techniques: theory and applications," *IEEE Trans. Visualization and Computer Graphics*, 6(1), 59-78.

Keim, D.A., Kriegel, H.P. (1996). "Visualization techniques for mining large databases: a comparison," *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 923-938.

Kelleher, C., Wagener, T. (2011). "Ten guidelines for effective data visualization in scientific publications," *Environmental Modelling & Software,* 26(6), 822-827.

LeBlanc, J., Ward, M.O., Wittels, N. (1990). "Exploring n-dimensional databases," *Proceedings of* IEEE Visualization, 230-237.

Minard, C. J. (1861). "*Des Tableaux Graphiques et des Cartes Figuratives*." Paris: E. Thunot et Cie. ENPC: 3386/C161; BNF: V-16168.

Nightingale, F. (1857). *Mortality of the British Army*. London: Harrison and Sons.

Ohio Department of Health Dashboard, https://coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards/overview/ [Online].

Perozzo, L. (1880). "Della rappresentazione graphica di una collettivit`a di individui nella successione del tempo," *Annali di Statistica*, 12, 1–16.

Playfair, W. (1786). *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century*. London: Corry. Re-published in Wainer, H. and Spence, I. *The Commercial and Political Atlas and Statistical Breviary*, 2005, Cambridge University Press, ISBN 0-521-85554-3.

Playfair, W. (1801). *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. London: Wallis. Re-published in Wainer, H. and Spence, I. (eds.), *The Commercial and Political Atlas and Statistical Breviary*, 2005, Cambridge University Press, ISBN 0-521-85554-3.

Robertson, G., Card, S., Mackinlay, J. (1991). "Cone trees: animated 3D visualizations of hierarchical information," *Proc. ACM Int'l Conf. Human Factors in Computing*, 189-194.

Shneiderman, B., (1992). "Tree visualization with treemaps: a 2D spacefilling approach," *ACM Trans. Graphics*, 11(1), 92-99.

Snow, J. (1855). *On the Mode of Communication of Cholera*. London. 2nd edn.

Tableau Public Software, https://public.tableau.com/en-us/s/ [Online].

The COVID Data Tracker by The Atlantic Monthly Group, https://covidtracking.com/ [Online].

The COVID Racial Data Tracker. Retrieved March 20, 2021 from https://covidtracking.com/race/dashboard [Online].

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

Tufte, E.R. (1990). *Envisioning Information*. Cheshire, Conn. Graphics Press.

Tufte, E. R. (1997). *Visual Explanations*. Cheshire, CT: Graphics Press.

Tukey, J.W., Tukey, P.A. (1985). "Computer graphics and exploratory data analysis: an introduction," In *The Collected Works of John W. Tukey: Graphics 1965–1985*, Vol. 5, ed.WSCleveland, pp. 419–38. New York: Chapman & Hall.

United States Census Bureau, Retrieved from "Projected Population by Single Year of Age, Sex, Race, and Hispanic Origin for the United States: 2016 to 2060" https://www2.census.gov/programs-surveys/popproj/datasets/2017/2017-popproj/np2017_d1_mid.csv [Online].

van Wijk, J. J. and van Liere, R. (1993). "HyperSlice: visualization of scalar functions of many variables", *Proceedings of the 4th IEEE Conference on Visualization,* 119-125.

World Health Organization 2020. WHO Director-General's opening remarks at the media briefing on COVID-19. World Health Organization. https://www.who.int/dg/speeches/detail/who-director-general-s-openingremarks-at-the-media-briefing-on-covid-19---11-march-2020 [Online].

WHO Coronavirus (COVID-19) Dashboard. Retrieved April 5, 2021 from https://covid19.who.int/ [Online].

# Appendix

## A. R-code with data links for the analysis of Section 3.2

```
##########################################################################
#######
### COVID-19 Descriptive Data Analysis
##########################################################################
#######

####The population estimation information is obtained from
###https://www2.census.gov/programs-surveys/popproj/datasets/2017/2017-
popproj/np2017_d1_mid.csv

nUS=332639102
nB=44734497
nW=253280207
nAIAN=4232241
nH=62312770
nAs=20009001

###Set a home directory and put the following files in it
###https://docs.google.com/spreadsheets/d/e/2PACX-
1vS8SzaERcKJOD_EzrtCDK1dX1zkoMochlA9iHoHg_RSw3V8bkpfk1mpw4pfL5RdtSOyx_o
ScsUtyXyk/pub?gid=43720681&single=true&output=csv
###
C = read.csv("CRDT Data - CRDT.csv", header = TRUE)
head(C)

###There are 56 states, you can omit
#"VI", "AS","GU","MP",and "PR"
# and focus on only 51 states including "DC"

table(C[,1])
names(table(C[,2]))
C=C[-which(C[,2]%in%c("VI","AS","GU","MP","PR")),]


###Creating monthly data from 4/20 to 3/21
library(stringr)
dates=as.character(C[,1])
months=names(table(substring(dates, 5, 6)))

for (i in 1:length(months)){
Ai=C[which(substring(dates, 5, 6)==months[i]),]
save(Ai,file=paste(paste("covid",months[i],sep="_"),"RData",sep="."))
}

###All monthly data are saved in the directory

load("covid_04.RData")
apr=Ai
```

```
remove(Ai)

load("covid_05.RData")
may=Ai
remove(Ai)

load("covid_06.RData")
jun=Ai
remove(Ai)

load("covid_07.RData")
jul=Ai
remove(Ai)

load("covid_08.RData")
aug=Ai
remove(Ai)

load("covid_09.RData")
sep=Ai
remove(Ai)

load("covid_10.RData")
oct=Ai
remove(Ai)

load("covid_11.RData")
nov=Ai
remove(Ai)

load("covid_12.RData")
dec=Ai
remove(Ai)

###Since data are cumulative, we only look at the last day data for the
whole month
###2021 data are not used

sum_apr = apr[apr$Date == "20200429",]
sum_may = may[may$Date == "20200531",]
sum_jun = jun[jun$Date == "20200628",]
sum_jul = jul[jul$Date == "20200729",]
sum_aug = aug[aug$Date == "20200830",]
sum_sep = sep[sep$Date == "20200930",]
sum_oct = oct[oct$Date == "20201028",]
sum_nov = nov[nov$Date == "20201129",]
sum_dec = dec[dec$Date == "20201230",]

###Combine all states:

country_apr <- subset(sum_apr, select = -c(Date, State) )
# Compute column sums
all_apr = t(as.matrix(colSums(country_apr,na.rm="TRUE")))
```

```
country_may <- subset(sum_may, select = -c(Date,State) )
all_may = t(as.matrix(colSums(country_may,na.rm="TRUE")))

country_jun <- subset(sum_jun, select = -c(Date,State) )
all_jun = t(as.matrix(colSums(country_jun,na.rm="TRUE")))

country_jul <- subset(sum_jul, select = -c(Date,State) )
all_jul = t(as.matrix(colSums(country_jul,na.rm="TRUE")))

country_aug <- subset(sum_aug, select = -c(Date,State) )
all_aug = t(as.matrix(colSums(country_aug,na.rm="TRUE")))

country_sep <- subset(sum_sep, select = -c(Date,State) )
all_sep = t(as.matrix(colSums(country_sep,na.rm="TRUE")))

country_oct <- subset(sum_oct, select = -c(Date,State) )
all_oct = t(as.matrix(colSums(country_oct,na.rm="TRUE")))

country_nov <- subset(sum_nov, select = -c(Date,State) )
all_nov = t(as.matrix(colSums(country_nov,na.rm="TRUE")))

country_dec <- subset(sum_dec, select = -c(Date,State) )
all_dec = t(as.matrix(colSums(country_dec,na.rm="TRUE")))

country_monthly <-
rbind(all_apr,all_may,all_jun,all_jul,all_aug,all_sep,all_oct,all_nov,a
ll_dec)


###US summary for 2020
all=rbind((country_monthly[,1]*100000)/nUS,(country_monthly[,4]*100000)
/nUS)
lgd=c("Cases per 100,000 people","Deaths per 100,000 people")

###Figure 7

barplot(all, xlab
="Date",main="",col=c("mistyrose3","mistyrose4"),beside=TRUE,ylim=c(0,6
500),
legend.text = lgd,args.legend = list(title = "", x =
"topleft",box.lty=0))
axis(side=1, at=c(2,5,8,11,14,17,20,23,26), labels=month)


#Calculate cases out of 100,000 for five races.

total_cases=colSums(country_monthly[,2:6])
case_rate_White=(country_monthly[,2]*100000)/nW
case_rate_Black=(country_monthly[,3]*100000)/nB
case_rate_Latinx=(country_monthly[,4]*100000)/nH
case_rate_Asian=(country_monthly[,5]*100000)/nAs
case_rate_AIAN=(country_monthly[,6]*100000)/nAIAN
```

```
#Calculate deaths of 100,000 for five races.

total_death=colSums(country_monthly[,15:19])
death_rate_White=(country_monthly[,15]*100000)/nW
death_rate_Black=(country_monthly[,16]*100000)/nB
death_rate_Latinx=(country_monthly[,17]*100000)/nH
death_rate_Asian=(country_monthly[,18]*100000)/nAs
death_rate_AIAN=(country_monthly[,19]*100000)/nAIAN

#Calculate testing proportion for four races.

test_rate_White=country_monthly[,41]/nW*100000
test_rate_Black=country_monthly[,42]/nB*100000
test_rate_Latinx=country_monthly[,43]/nH*100000
test_rate_Asian=country_monthly[,44]/nAs*100000
test_rate_AIAN=country_monthly[,45]/nAIAN*100000

#Calculate proportion of cases who died for five races.
prop_death=total_death/total_cases

#Calculate proportion of hospitalizations among cases for four races.

total_hosp=colSums(country_monthly[,28:32])
prop_hosp=total_hosp/total_cases

#create month names
month <- c('4/20', '5/20', '6/20', '7/20', '8/20',
'9/20','10/20','11/20','12/20')

###Figure 8

par(mfrow=c(1,2))
plot(c(1:9),case_rate_White,col ="blue",pch=0,main="(a)",type =
"b",ylim=c(0,4000),xlab="date",ylab="Cases per 100,000
people",xaxt="n")
lines(c(1:9),case_rate_Black,col ="black",pch=1,type = "b")
lines(c(1:9),case_rate_Latinx,col ="red",pch=2,type = "b")
lines(c(1:9),case_rate_Asian,col ="green",pch=3,type = "b")
lines(c(1:9),case_rate_AIAN,col ="orange",pch=4,type = "b")
axis(side=1, at=c(1:9), labels=month)
legend("topleft", legend=c("White", "Black","Latinx","Asian","AIAN"),
       col=c("blue","black","red","green","orange"), pch=c(0,1,2,3,4))

plot(c(1:9),death_rate_White,col ="blue",pch=0,main="(b)",type =
"b",xlab="date",ylim=c(0,120),ylab="Deaths per 100,000
people",xaxt="n")
lines(c(1:9),death_rate_Black,col ="black",pch=1,type = "b")
lines(c(1:9),death_rate_Latinx,col ="red",pch=2,type = "b")
lines(c(1:9),death_rate_Asian,col ="green",pch=3,type = "b")
lines(c(1:9),death_rate_AIAN,col ="orange",pch=4,type = "b")
axis(side=1, at=c(1:9), labels=month)
```

###**Figure 9**

```
plot(c(1:9),test_rate_White,col ="blue",pch=0,type =
"b",xlab="date",ylab="Tests per 100,000 people",xaxt="n")
lines(c(1:9),test_rate_Black,col ="black",pch=1,type = "b")
lines(c(1:9),test_rate_Latinx,col ="red",pch=2,type = "b")
lines(c(1:9),test_rate_Asian,col ="green",pch=3,type = "b")
lines(c(1:9),test_rate_AIAN,col ="orange",pch=4,type = "b")
lines(c(1:9),test_rate_NHPI,col ="yellow",pch=5,type = "b")
axis(side=1, at=c(1:9), labels=month)
legend("topleft", legend=c("White", "Black","Latinx","Asian","AIAN"),
       col=c("blue","black","red","green","orange"), pch=c(0,1,2,3,4))
```