# STAT 405/605: R for Data Science Project Report

Ben Allen, Anthony Cai, Rohit Jha, Daniel Prasca-Chamorro, Ivan Paredes

April 24, 2022

## 1. Introduction

### A. Pandemic Outbreak

In more than the past two years, coronavirus (CoV) has been the center of attention in our daily life. The surveys behind are increasing popular projects, which develops a statistical system to predict the trend of COVID-19 throughout the world. Monitoring the virus cases are essential both for authorities, administration, and citizens to develop the awareness of how to pandemic envolve, taking necessary measures to protect the public health system as well as stability of societal operation. At the beginning of outbreak after January 20, 2020 when CDC confirmed the first ever COVID-19 case in United States sampled from Washington state, there is a shortage of laboratory tests and available data analysis to represent the progress of a pandemic, but at this point plenty of surveys have been fully conducted and evne our group would be able to make it as a final project of this class. The opening section of this paper along with our project is the collective dense visualization of surveillance data with geographical information across the United States of America. We will present main components of this primary data set that is publicly available in CDC, and RStudio provides the computational development system as we mainly designed to split data by state and use functional programming for statistical computing and graphics. This empowers the portability of our analysis that can swiftly adapt to future works, possible extensions, and tracing method.

## 2. Primary Data Set: Covid-19 Case Surveillance Data

### A. Summary of the data set

The first and foremost object that lays the solid foundation of our project is the *COVID-19 Case Surveillance Public Use Data with Geography* made by CDC's cleaning algorithm. It includes 15 columns representing elements of demographics, residency, test result, exposure history, disease severity, coupled with underlying conditions in medical risk behaviors. Each row of the data is one diagnosed patient. The geographic information includes District of Columbia and U.S. territories such as Porto Rico. These case reports are shared by health departments and are using standardized formats. They are considered provisional but have quality assurance procedures and data suppression.

### B. Data Processing & Visualization

Based on the raw data provided, we use several procedures to clean the data. Firstly, the omit of "N/A" data entries and split the data frame by different columns of interest using functional programming. New data frames are created for each state using a *for* loop, and the graphs are placed into a list to create massive visualization. *ggplot2* is one of the most prevalent library we use to create plots, and we tried to use the SQL feature in RStudio to manipulate data. The subsets of the original data frame can be gathered as tables in a database, and by using queries we are able to obtain ideal outputs. The commends in SQL can also be pasted as a function with mutable key variables, which can ensure the generalization of database. Another feature we use is the *String*; it deals with the character variables in an numerical way. Lastly, we introduce some advanced works in visualization such as violin chart, stacked area chart, and hexbin choropleth map. Combining plots in grid view to save space and make the result more compact. Simple integrative plots are included at the end using *plotly*, *lubridate*, and *dygraphs*.

### C. Figures

This section will glance some visualization we did based on the primary data set. The histogram of case count of each state is shown in orange bars, with the 3-month and 5-month moving average lines. Cumulative mean is indicated by the blue line showing how the trend changes over time.

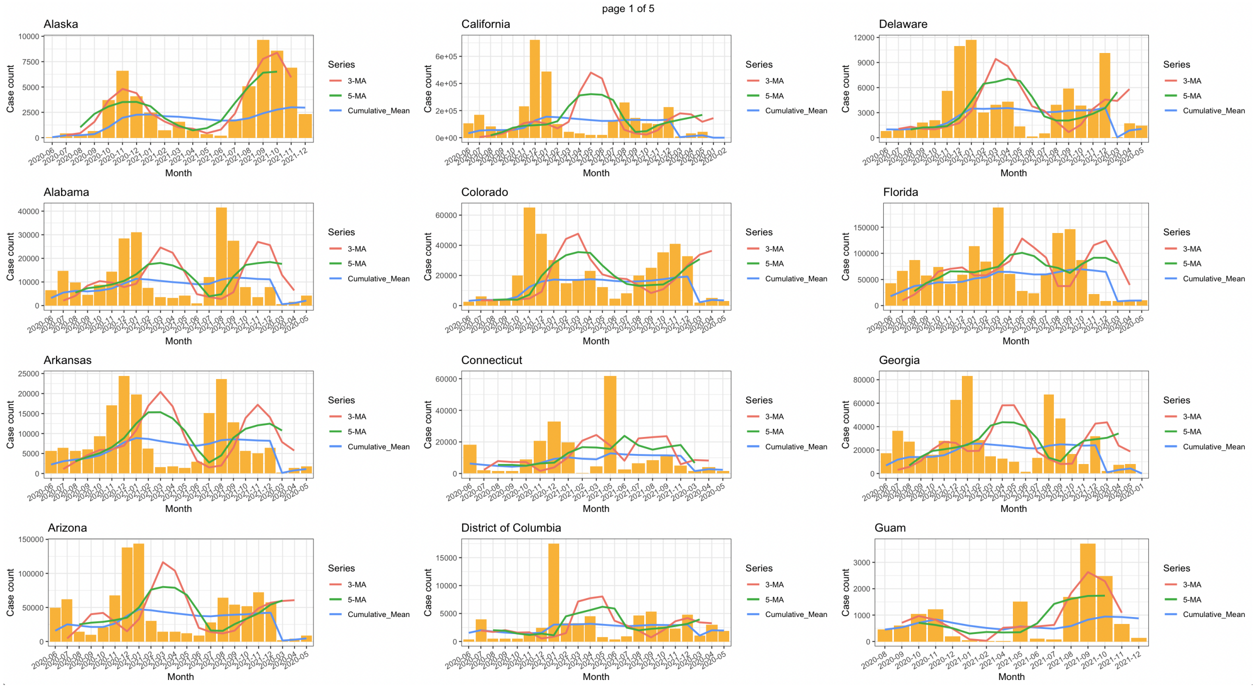

Figure 1: State case distribution with trend

The total case count of each state/territories is spread and sorted with flipped coordinate to better visualize the difference in between.

Pie plots indicate the percentage of the answers towards each reported case. We chose four most remarkable columns to generate the outputs.

Meanwhile, we use stacked area chart to demonstrate the evolution of a numeric variable for several state, and offers a dedicated reordering with ggplot2 by case count.

Box plots are useful to provide a summary of the data to quickly identify critical segments of the distribution, the dispersion of the data set, and signs of skewness. We group the 5 states for each plots since they may have closer case counts and make the plots on correct scales.

The last figure describes a hexbin map. It is built on a *geojson* file providing boundaries of states as hexagons. *geojason* gallery dedicates a whole section of geospatial object that one can plot using the plot() function.

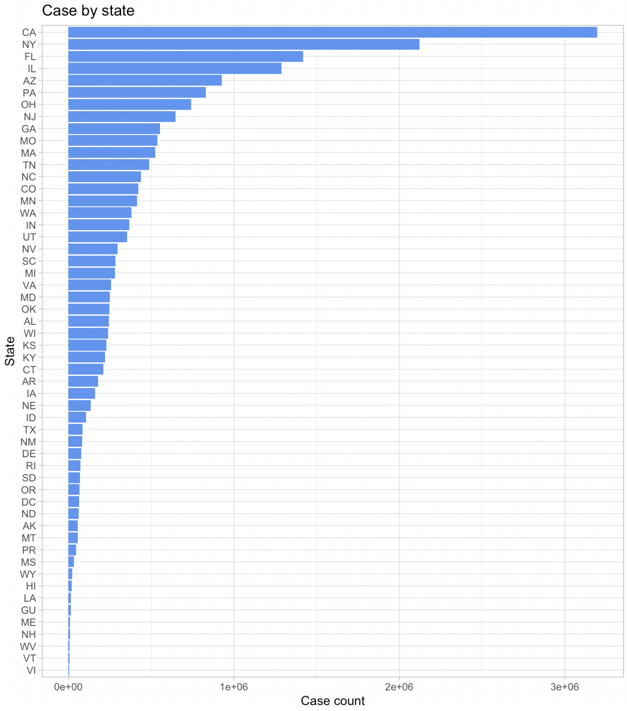

Figure 2: State case distribution

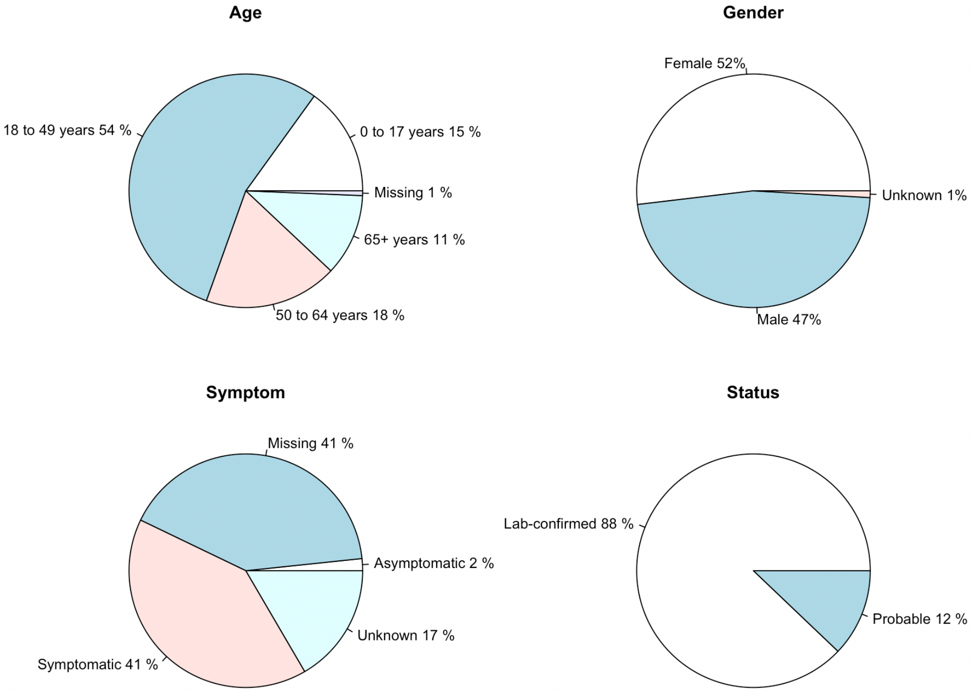

Figure 3: Pie chart

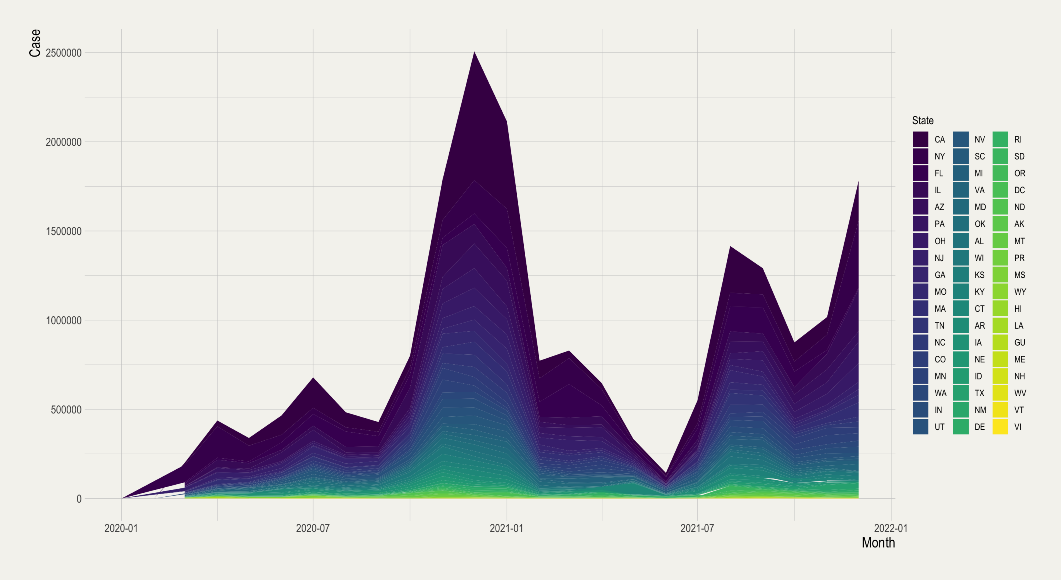

Figure 4: Stacked Plot

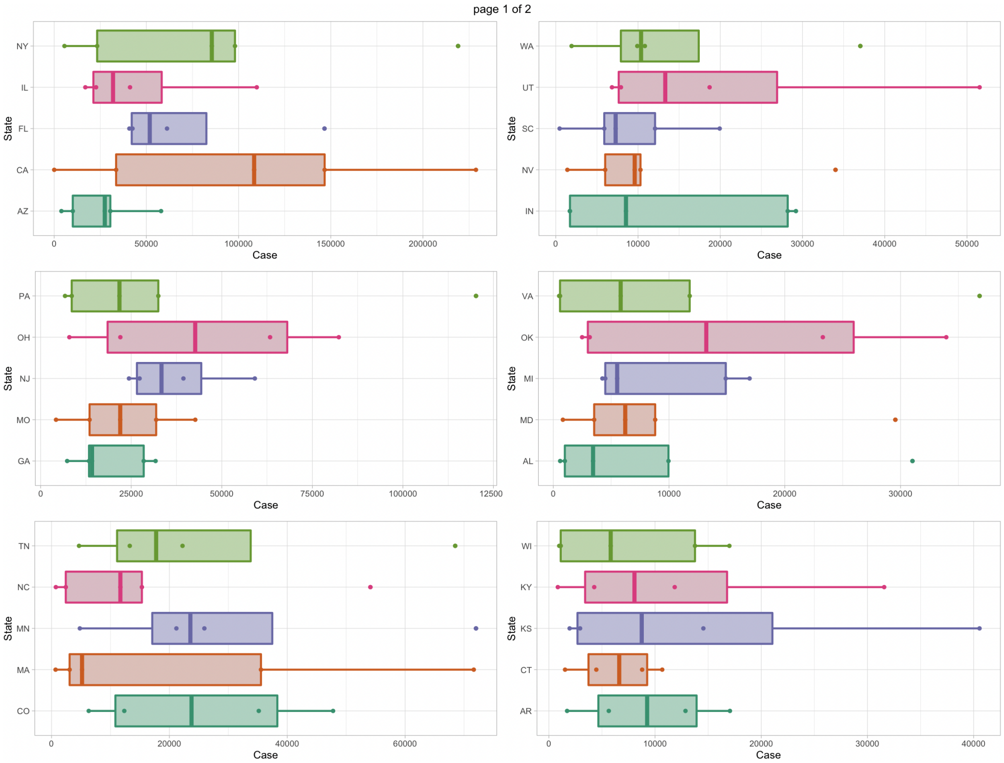

Figure 5: Boxplot of some top case count states

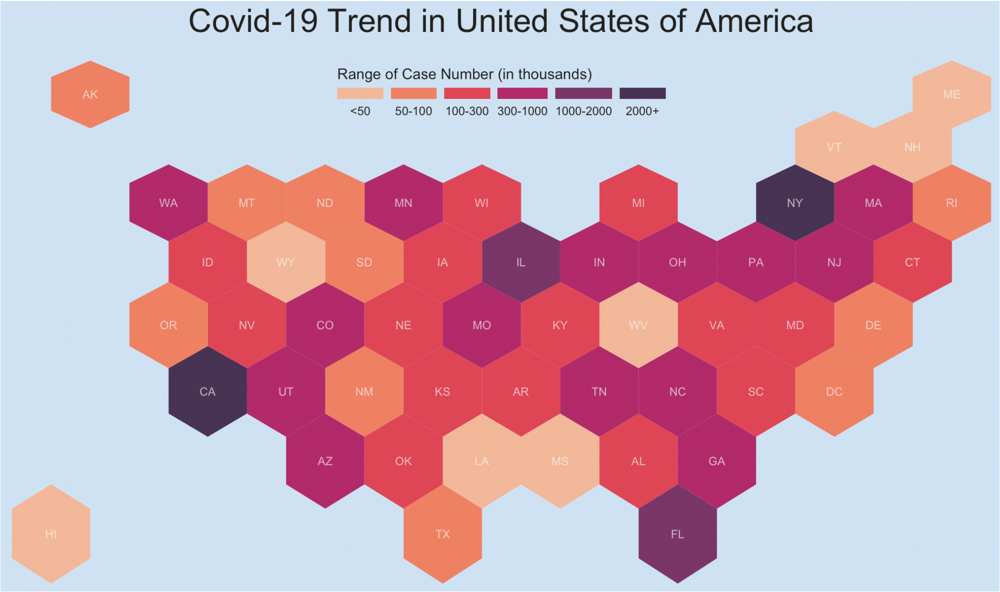

Figure 6: U.S. choropleth map

### D. Insights

There has already been many interesting works on pandemic visualization, and it has been shown that for many fundamental and agile managements of data sets, RStudio could provide ingenious approaches to generate the ideal outputs. The more advanced work that our group has not done on the COVID-19 cases is to find the causality of these known variables behind the obvious case counts and reported entries. Unlike correlation, causation explore the framework of statistical science to examine the potential outcome, counter factual reasoning, and address potential connection to many other factors from case surveillance.

## 7. Conclusion

### A. Pandemic Statistics & Inference

Our work on the primary data sets reinforces cleaning, scraping, rearranging, revision, functional programming, and visualization rather then analysis of in-depth concepts. But it is important to establish such a starting point to serve for more diverse resources about the pandemic.