COVID-19 Health Effects & Social Media

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

Abstract

Our group has decided to translate large COVID-19 data sets into visual illustrations to help depict the larger effects of COVID-19 on our society's health and its prevalence among social media platforms. Briefly, our primary dataset includes COVID-19 positive-case counts by age group, sex, race, and geography. Our secondary dataset includes Twitter user data with tweets that contain COVID-19 related keywords. Graphs and figures were generated using the R language for graphing and SQLite for large database manipulations.

Table of Contents

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.

B. Tweets and Twitter

Our second dataset includes COVID-19 Twitter data. Specifically, this data set was obtained from openicpsr.org from a research paper "COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes" by Gupta, et al. The authors highlight that this project was designed to collect and layout Twitter conversations around the COVID-19 subject matter spanning from January 2020 to September 2021. The authors generated the large dataset by string searching keywords such as "Corona", "covid", "nCov", and "wuhan". Furthermore, this dataset provided a unique aspect of the Tweets; The authors leverage artificial intelligence and machine learning to quantify the intrinsic emotion in the tweet and output an "Emotion score" that encapsulates the language and tone of the Twitter users spanning from happiness, sadness, anger, and fear. The data set was almost 5 GBs worth of COVID-19 related twitter data with over 158 million tweets worldwide and emotional quantification.

C. Hypertension and Life Expectancy

Our group ambitiously tackled three other datasets and attempted to intertwine all five datasets together and portray the effect of COVID-19 with other health data. Our ancillary datasets included 1) life expectancy, 2) hypertension, and 3) state population data. Briefly, our life expectancy data was obtained from the CDC and provided life expectancy data from the US by state. Our hypertension data came from the Division for Heart Disease and Stroke Prevention in the CDC. The state population data came from the census.gov.

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

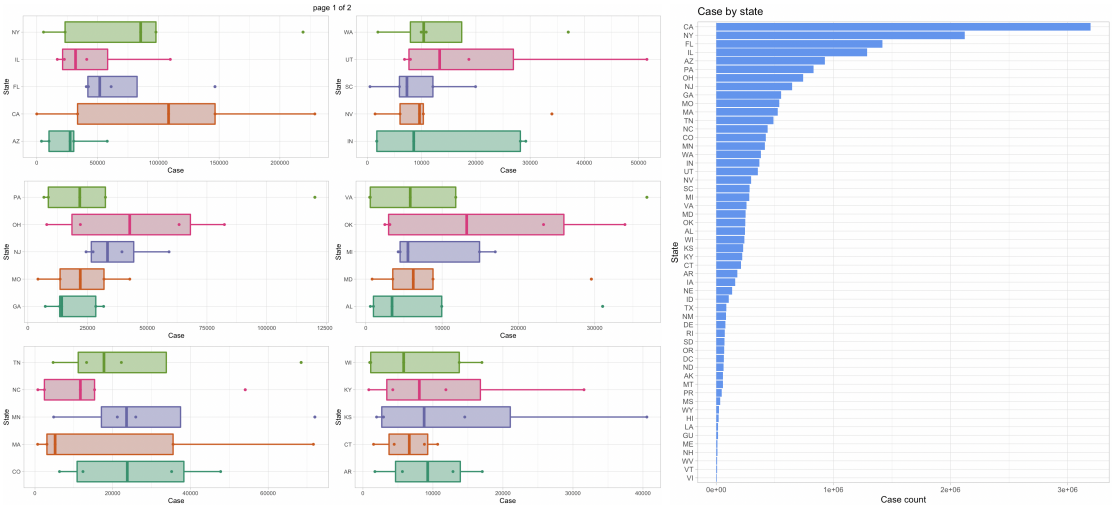
Based on the raw data provided, we use several procedures to clean the data. Firstly, the omit of "N/A" data entries and split the data frame by different columns of interest using functional programming. New data frames are created for each state using a *for* loop, and the graphs are placed into a 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.

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

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.



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

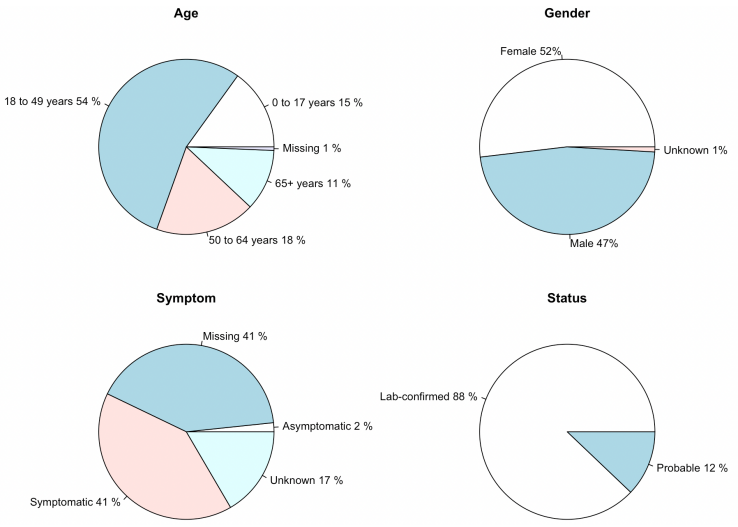


Figure 1: Pie chart

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.

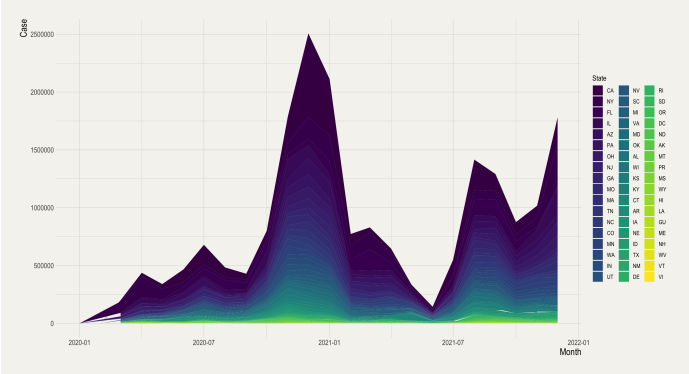


Figure 2: Stacked Plot

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.

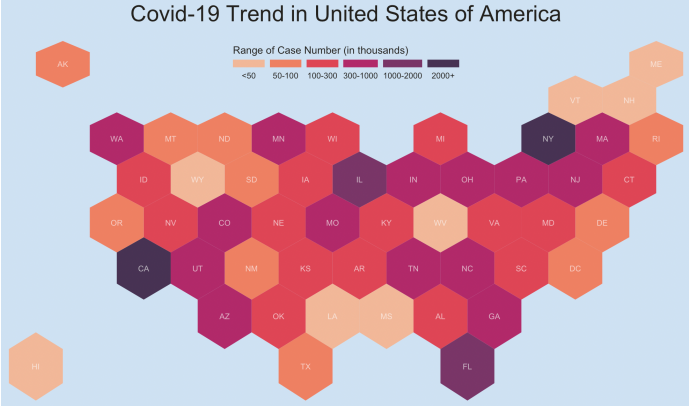


Figure 3: U.S. choropleth map

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.

3. Secondary Data: Tweets with COVID buzz words

Data Set Summary

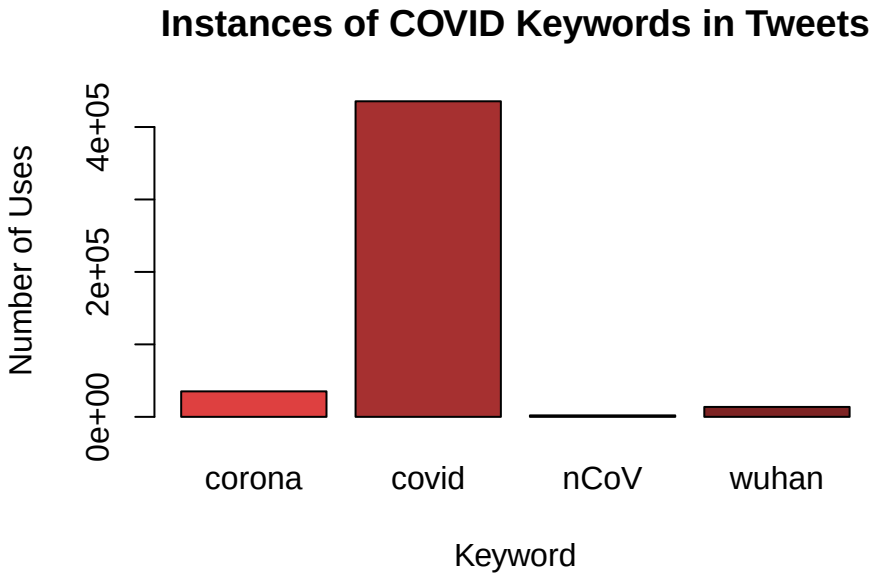
Our next data set contained data on tweets relating to COVID-19. Included is the tweet id, the timestamp, and what country the tweet was from. The data set has over 198 million entries for each tweet recorded relating to COVID-19 internationally.

The authors also made use of AI to generate an emotional rating of each tweet. With this being split up into happiness, sadness, anger, and fear.

Our hypothesis is that the number of tweets relating to COVID-19 could be used to predict the number of COVID cases at a given time.

Figures

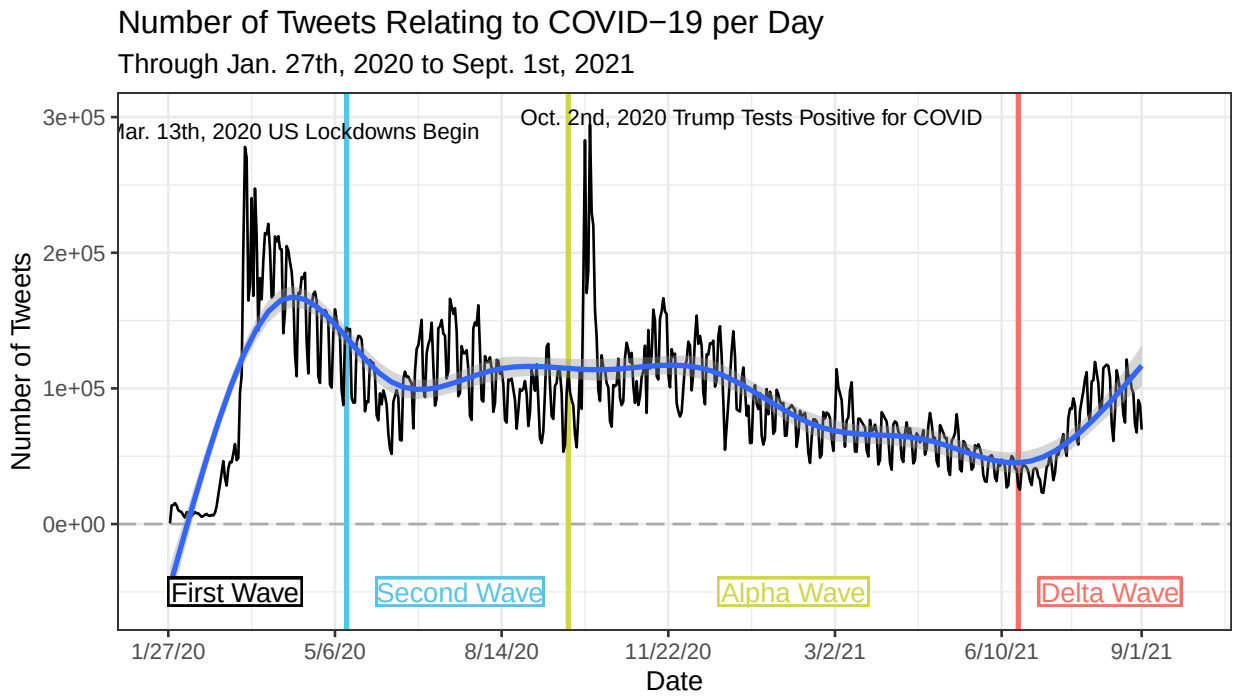
We can see how "covid" was by far the most common keyword captured by the data set. This is most likely because it is less formal than something like "nCoV" and would find more use by common people.



Next, we looked at time series of the amount of tweets relating to COVID-19 per day. We also annotated the graph with the dates of different COVID-19 case waves. This data was only for the United States and because of that we can see some interesting abnormalities. We can see the first large spike in tweets in this data set is around March 13th, 2020 which was around the time state lockdowns were beginning to go into effect. We see that interest slowly wanes in COVID after.

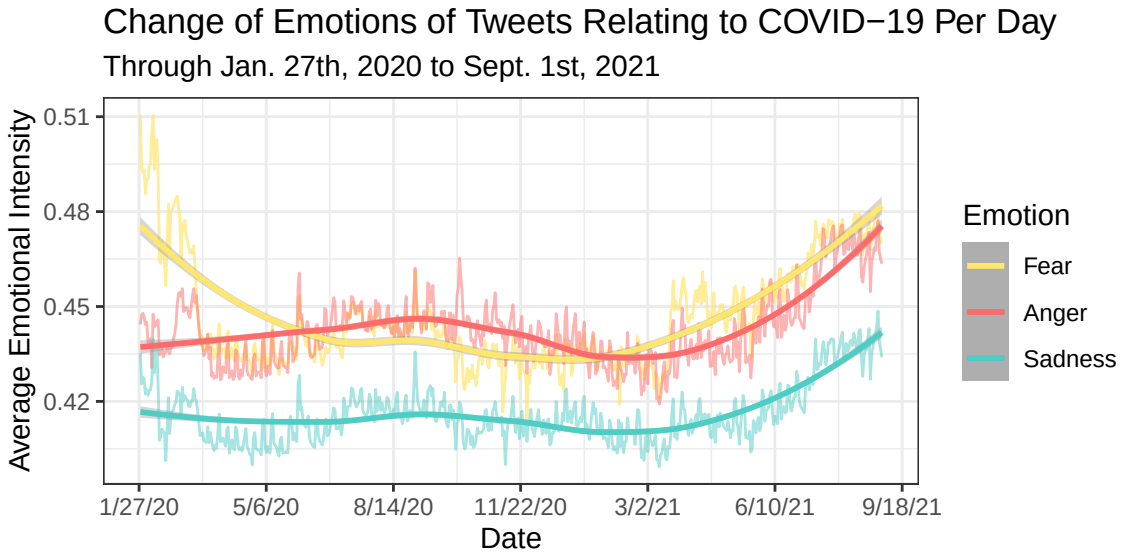
This is until a large spike around October 2nd, 2020, which corresponds to the day that former US President Donald Trump was confirmed to be positive for COVID-19. Given Trump's notorious use of Twitter at the time, it makes sense that this event contributed to a large spike in Twitter activity in the United States.

We lastly begin to see a rise in US Twitter activity as the delta wave sets in, again understandable given the higher lethality that came with that wave.



After this we looked at the emotional data provided by the data set. We graphed the average intensity of sadness, anger, and fear. The data set also included happiness, however, it was substantially lower on average, understandably so, and this made it difficult to include in the graph.

From this graph, we can see that there was a high amount of fear on average initially, and all emotions decreased over time before roughly the time of the delta wave, where the negative sentiment began to rise again on average.



4. Combining COVID Cases and Tweets related to COVID keywords.

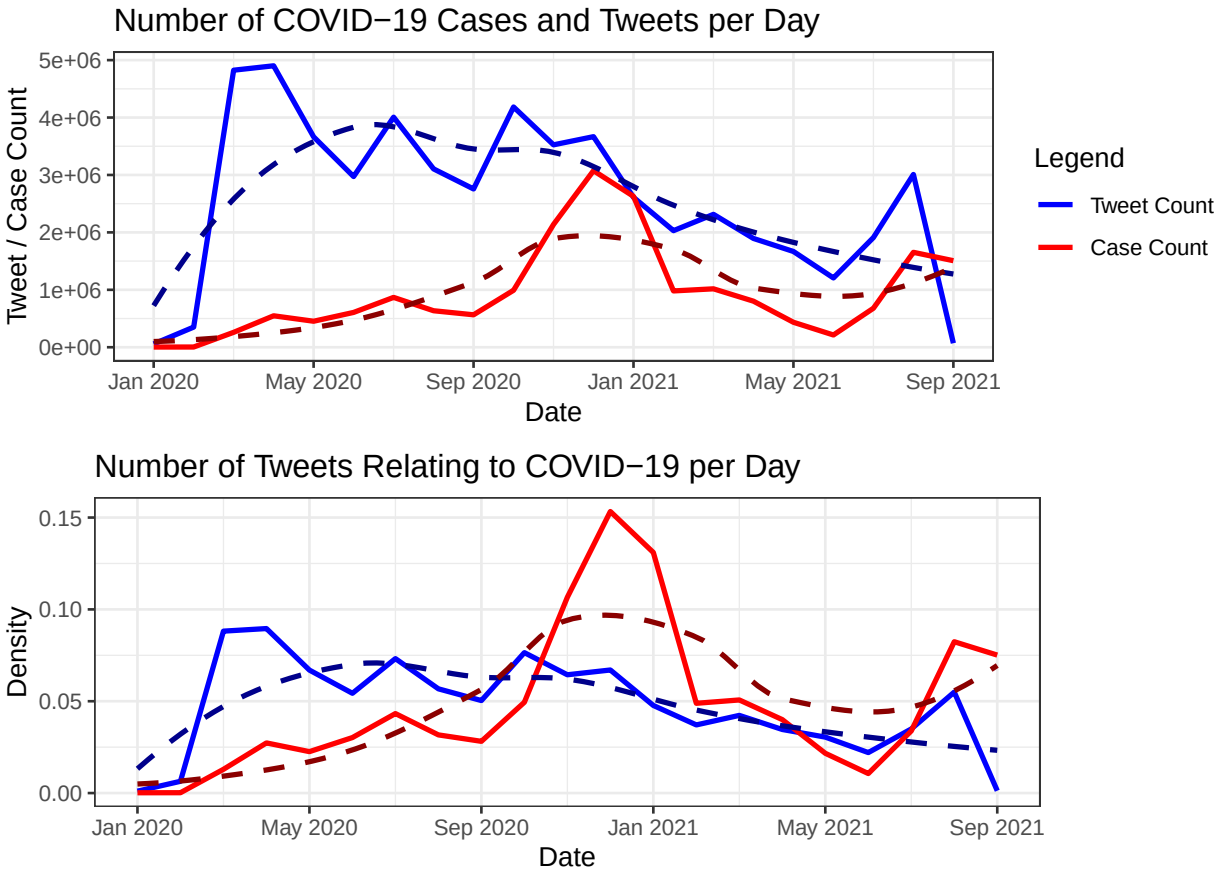
Summary and Rationale

We wanted to see whether the twitter data we've found could be used to predict the amount of COVID-19 cases in the United States. To do this, we summarized the data by finding counts of both tweets and cases per month. We then combined this data into one data set. We also found the amount of COVID-19 cases for males and females, and used this to calculate the proportion of cases per day that were male and those that were female.

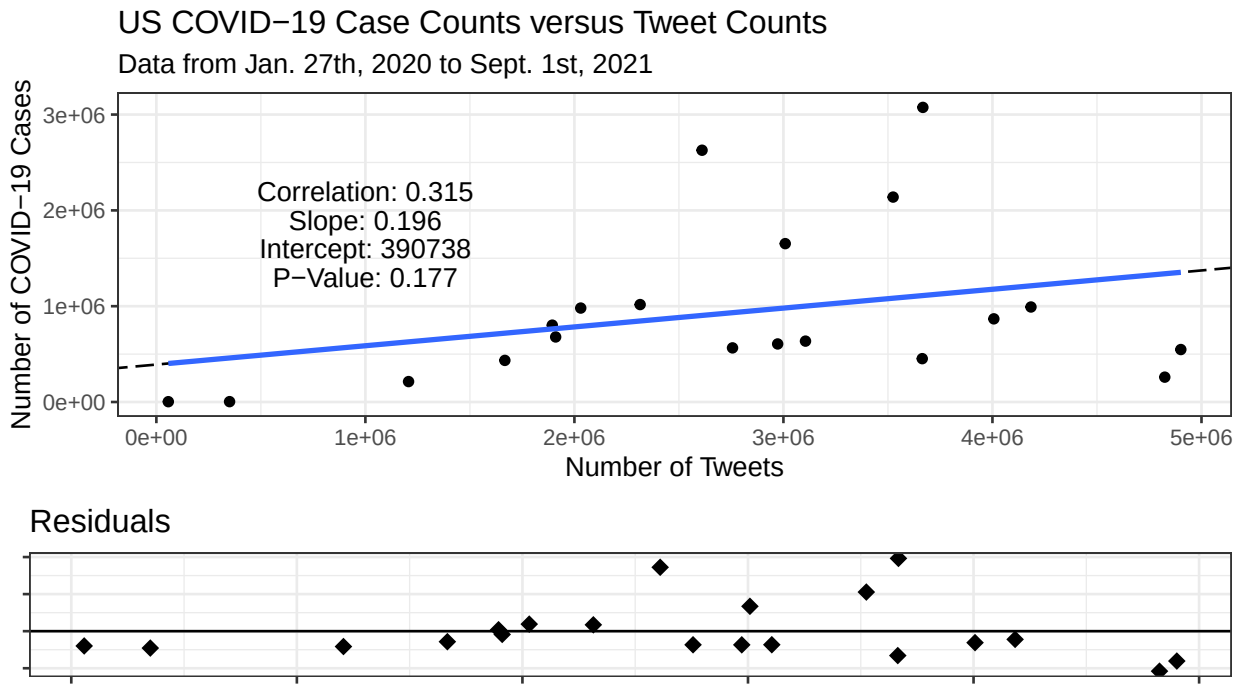
Figures

This first plot shows the monthly amount of tweets relating to COVID-19 and the monthly amount of COVID-19 cases over time. From it we can see that at first there were far more tweets than cases, but eventually both of them fell in number. It is interesting to note near the end that tweets and cases rose by similar amounts per month.

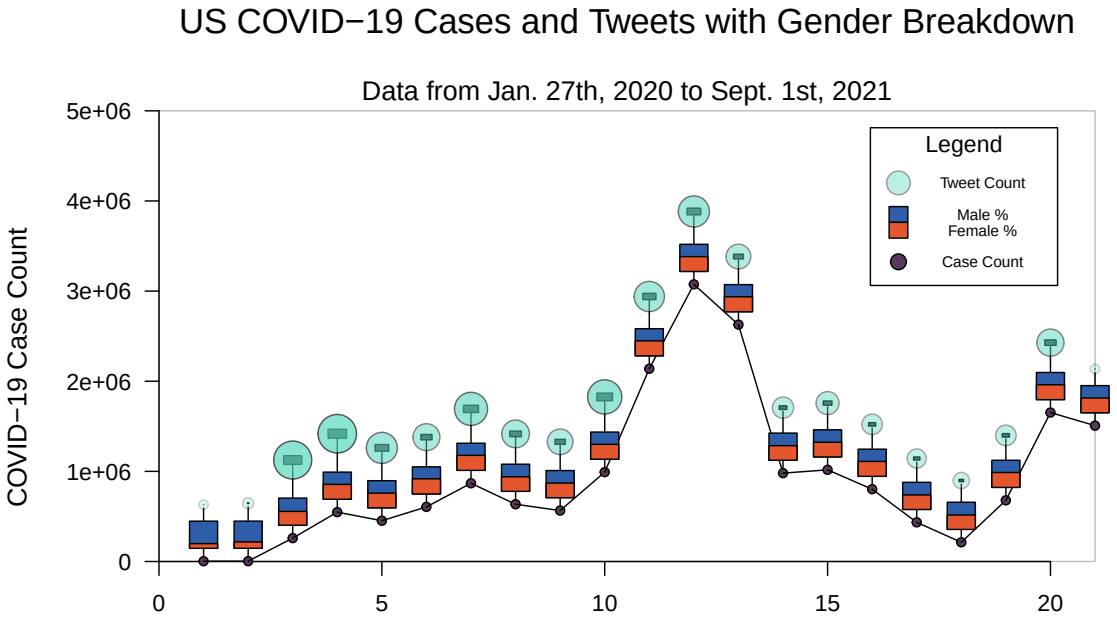
The second plot transforms the first into a density plot, and from this we can see how the case data was a lot more unevenly distributed, with the majority of cases occuring in the central spike.



Next, we checked for a linear relationship between the tweet and case counts, and found a correlation of 0.315 with a p-value of 0.177. This p-value is too high for us to conclude there is a linear relationship, however the correlation coefficent does make a good case for a loose relationship.



Finally, we wanted to analyse case count, tweet count, and the gender breakdown all at once. We created the following original visualization to do so. From this graph we can see a similar pattern that tweet counts were higher initially before the spike in cases. Interestingly the proportion between male and female was heavily skewed to include more males initially in the data before balancing out later. This is likely due to limitations of the data set because later on women actually accounted for slightly more cases consistently.



5. Combining COVID Cases with Hypertension and Life Expectancy Data

Summary and Rationale

We also analyzed secondary datasets that contained information about the Life Expectancy and Hypertension Mortality rates by state. These factors were chosen because Life Expectancy has been shown in other literature to correlate with overall quality of life and Hypertension mortality rates tend to indicate overall heart health levels in a state. We hypothesized that these factors may correlate with COVID case numbers on a state level.

Life Expectancy

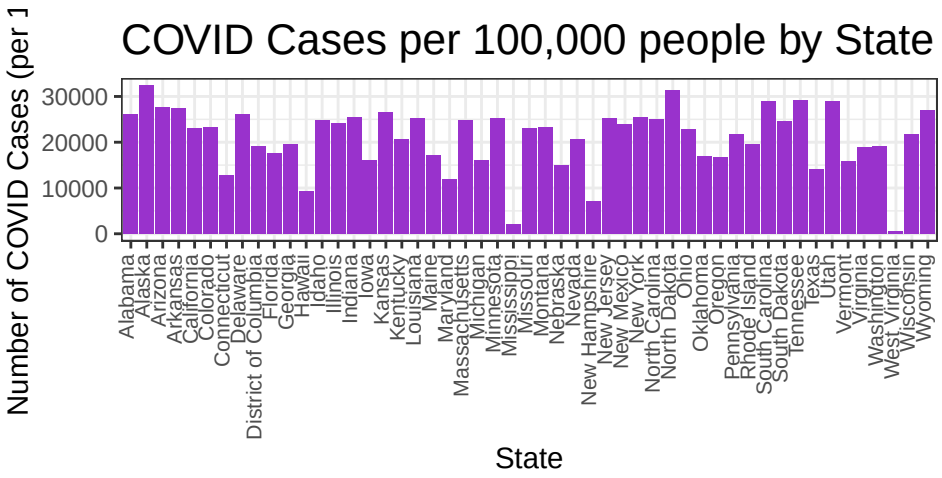
Life expectancy is a measure of the amount of time (in years) that a child born today in a certain state would be expected to live for on average. Many personal and biological factors influence life expectancy, including lifestyles, diet, genetics, and hygiene. Additionally, environmental influences can factor into life expectancy, such as health care quality, proximity to pollution, and safety of the neighborhoods in which one lives. Given how encompassing the factors that affect life expectancy are, it is a pretty solid measure of the general quality of life that one lives. As such, it would stand to reason that people who live higher quality lives that have would tend to have stronger immune systems and therefore might be less likely to contract COVID. Extending that logic, we hypothesize that we will see a negative relationship between State Life Expectancies and COVID case densities; States that have a higher life expectancy will generally have a lower COVID density.

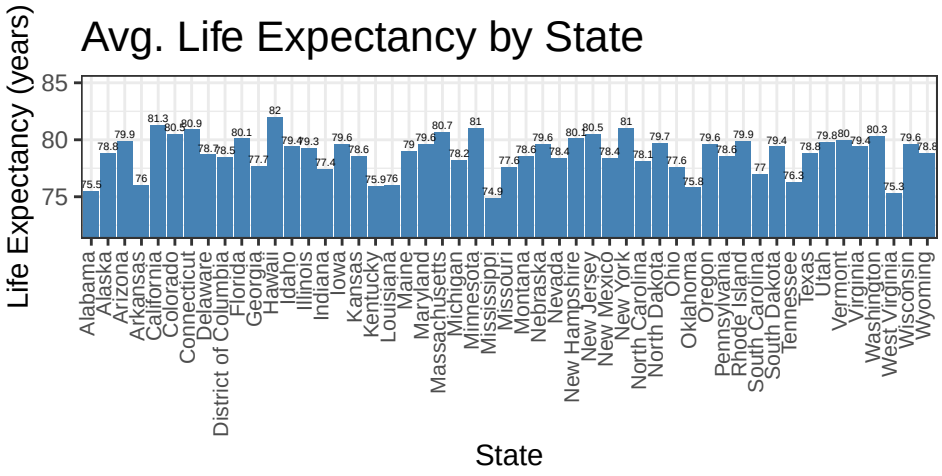
Hypertension

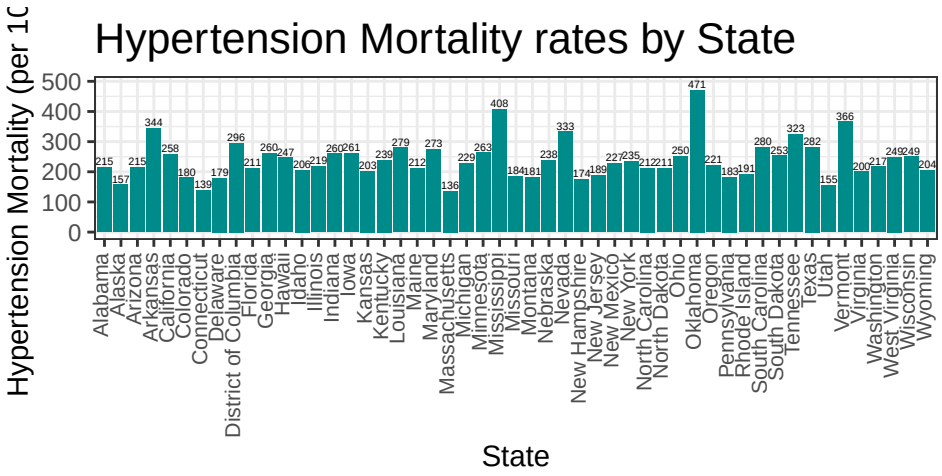
Hypertension, also known as high blood pressure, is known to be correlated with increased risk for cardiovascular diseases, and is generally a good indicator of a person's heart health. As such, a measure of a State's Hypertension based Mortality rates can be an indicative of the overall heart health of a state as well as the general efficacy of the state's health care system. COVID has been shown to affect multiple organs in the body, including the heart, and as such we suspect that it is possible that people that have weaker cardiovascular systems may be more prone to contracting COVID. To put that theory to the test, we analyzed the correlations between Hypertension Mortality rates and COVID Case densities at a state level.

Data Manipulation

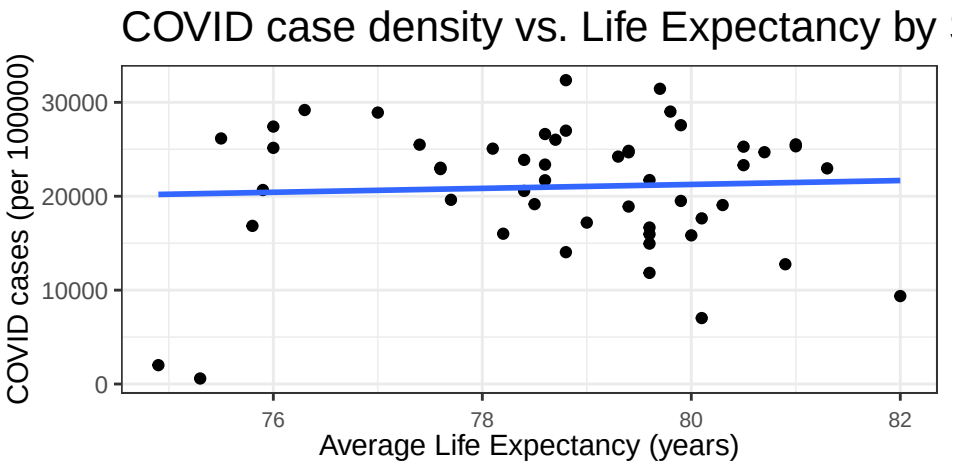
We wanted to test how COVID case numbers correlated with the lifestyles and the general effectiveness of the health care systems in each state. To that end, we chose the factors of **Life Expectancy** and **Hypertension Mortality** to consider and created our variable, State COVID Density, as a measure of COVID cases in each state divided by the population of each state and multiplied by 100,000. This was done so that states such as Texas and California that have large populations which likely result in higher COVID case numbers don't skew the data. Here's a view of the COVID density, life expectancy, and hypertension data:







Initially, we calculated a correlation of 0.051838 with a p-value of 0.717903. for the association between life expectancy and COVID case density. However, when graphically viewing this data, it seems there are a couple of states that could be outliers.

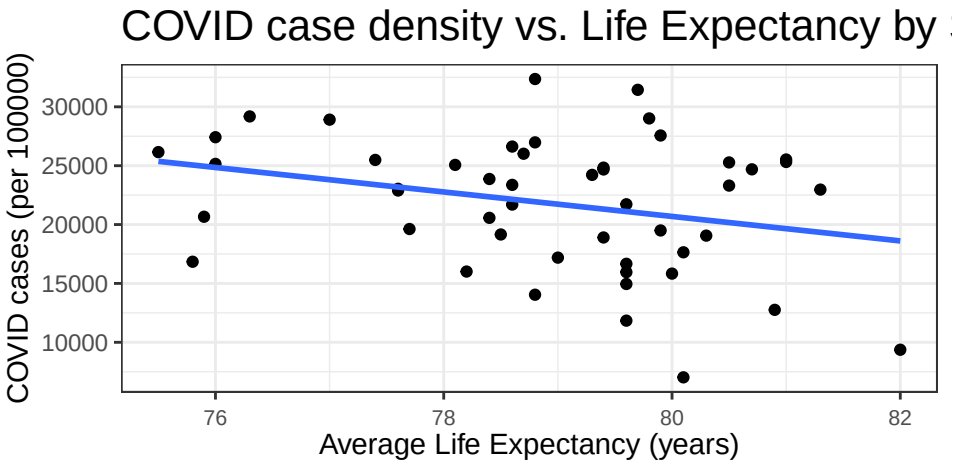


Those two states in the bottom left corner of the graph, having both very low COVID density and low life expectancy are West Virginia and Mississippi. However, it's possible that these two states are outliers to the general trend.

To test this, we used the basic IQR*1.5 method of outlier testing.

Given the distribution of the data, data values should fall within a range of approximately 4440 and 38000 COVID cases per 100000 people. West Virginia and Mississippi reported had about 589 and 2009 COVID cases per 1000000 people. These are below the expected minimum of the data, and therefore it is reasonable to conclude that these data values are potentially outliers. It is important to note that this could arise from the fact that in our dataset, there were many cases in which the state that the person is from was missing or excluded, and maybe West Virginia and Mississippi were states where COVID cases were under-reported.

Re-plotting the correlation after removing these outliers results in the following:

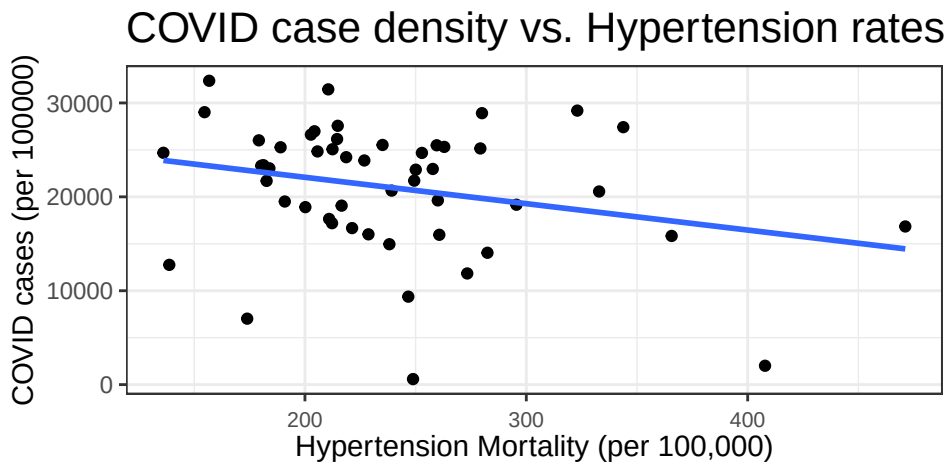


Calculated a correlation of -0.286958 with a p-value of 0.0456.

After removing those two outliers, we can see that we have a statistically significant, moderate, negative correlations between a State's Life Expectancy and its COVID density. This seems to support our hypothesis; as a state's life expectancy (and therefore it's quality of life) increases, it's COVID density tends to decrease. Citizens of states with higher qualities of living are also generally less like to be diagnosed with COVID.

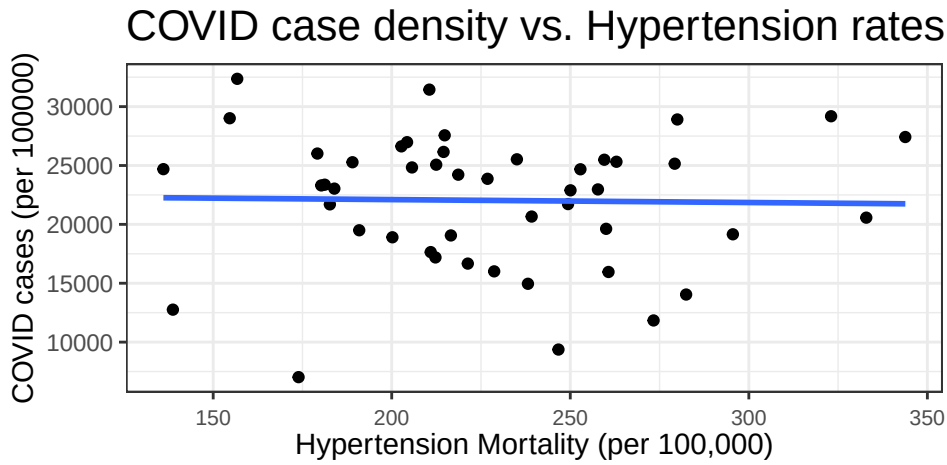
Hypertension

When checking the correlation between State COVID densities and State Hypertension rates, we once again get a p-value larger than 0.05.



Following a similar process for checking for outliers, we end up removing the same two states, West Virginia and Mississippi because they fall outside the range for State COVID density.

The range for normal data values of Hypertension Mortality per 100,000 people for each state is between approximately 113 and 349. There are three states that fall above this range - Oklahoma at 471, Mississippi at 408, and Vermont at 366. Removing all the outliers resulting in the following correlation and graph:



Interestingly, in this case removing the outliers resulting in an even higher p-value of 0.889, and the graph shows effectively no correlation.

6. 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.

B. Secondary Data Set

We analyzed and graphed our secondary data set about COVID-19-related tweets in the United States. We looked at the frequency of each covid keyword and the overall amount of covid-related tweets each day from January 27, 2020, to September 1, 2021. We also separately graphed the average emotion levels of tweets within this timeframe.

C. Primary + Secondary

Analyzing our primary and secondary datasets, we found that the number of tweets are loosely correlated with case count. While this discovery is not particularly useful for prediction alone, it may be useful as a feature in a more complex Machine Learning model. It should not be used in a linear model, given that we were unable to prove a statistically significant linear relationship.

D. Primary + Tertiary

There was a statistically significant ($\alpha = 0.05$) inverse linear relationship between average life expectancy per state and COVID-19 case density. This indicates that the factors that lead to a higher average life expectancy also decrease the number of COVID-19 cases in an area.

There's no demonstrable correlation between Hypertension Mortality rates (and heart health in general) and COVID density by State. That is to say, COVID cases were not moderated by poorer heart health and quality of health care in a state. However, it is possible that the states with poorer heart health and care quality saw higher COVID death rates, and that could be an aspect for future consideration. In this case, we were limited since the data set we used was often missing COVID death information, but it is possible in the near future there will be more reliable data about COVID deaths.

For future studies, we would recommend focusing in on smaller regions that are more homogeneous such as districts within a city, if such data is made available, rather than using a whole state to test correlations between COVID densities and the outcomes of other. Given the size and diversity of most states in the US, true correlations might be masked and false correlations may be seen.

Data Credits

COVID-19 case data: Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Public Use Data with Geography (version date: April 04, 2022)

COVID-19 tweet data: Gupta, Raj, Vishwanath, Ajay, and Yang, Yinping. COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-11-04. https://doi.org/10.3886/E120321V11

Life Expectancy: National Center for Health Statistics. U.S. Life Expectancy at Birth by State and Census Tract - 2010-2015. Date accessed [4/20/22]. Available from https://data.cdc.gov/d/5h56-n989.

Hypertension: Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention

State Population Totals: Population, Population Change, and Estimated Components of Population Change: April 1, 2010 to July 1, 2019 (NST-EST2019-alldata)