Introduction

The Coronavirus disease (COVID-19) has tremendously impacted public health and economic and social well-being [1][2]. In addition to the dramatic loss of human life, the pandemic produced a mental health crisis, significant unemployment, and increased social barriers [2]. There was immense stress in keeping oneself and loved ones safe. Everyone missed out on major life milestones from graduations to weddings due to quarantining and social distancing mandates.

However, when we take a look from state to state, not every population seems to have been affected equally. My college friend in New York experienced the seriousness of COVID-19 earlier than I did in southern California since the spread of the virus was initially more rapid in the city. Friends who couldn't afford to take accurate, rapid tests had a more difficult time traveling to get to see family for the holidays. Some areas struggled more with having access to vaccines while others had difficulty with compliance to mask mandates. With all of these varying experiences, how could we group states that were similarly challenged by the pandemic?

Data Wrangling

I found eight raw datasets from a range of sources like Kaggle, CORGIS, United States Department of Agriculture and the Bureau of Transport Statistics. I imported, cleaned and reorganized them pertaining to total COVID-19 deaths by state, macroeconomic indicators, general health/wellness indicators and domestic air traffic data.

Exploratory Data Analysis

COVID-19 deaths and possible state classifying factors were then investigated through visualizing their relationships using Matplotlib and Seaborn graphs. An important question I wanted to explore is if the total number of deaths per state also means those states' populations were affected similarly.

One positively correlated relationship was between deaths, population and the percent of the population affected as seen in Figure 1. In this graph we can see as the states' population increases the number of deaths also increases steadily. The top states with the highest number of deaths (California, Texas, Florida, New York and Pennsylvania) don't appear to have the highest percent population affected like Tennessee, Alabama or Oklahoma. California has the highest number of deaths as well as population, yet the percent of the population affected was in the mid-lower range.

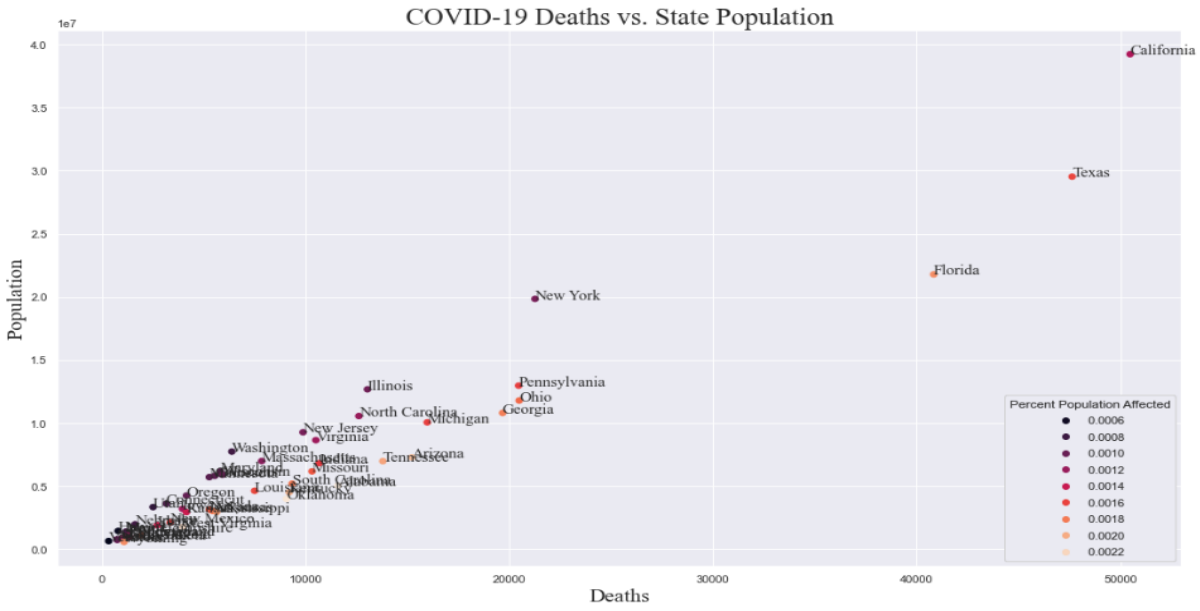

Figure 1. COVID-19 total deaths per state graphed against each states' population. Each data point is colored based on the percentage of the population belonging to the states' deaths.

The Hispanic and Latino demographic had the highest positive correlation with deaths for the United States demographic data as seen in Figure 2. The states with a larger White not Hispanic or Latino population had the highest negative correlation with deaths.

Deaths	1.000000
Percent African American	0.284419
Percent American Indian And Alaska Native	-0.233846
Percent Asian	0.152693
Percent Native Hawaiian And Other Pacific Islander	-0.141103
Percent Two Or More Races	-0.138449
Percent Hispanic Or Latino	0.509375
Percent White Not Hispanic Or Latino	-0.456957
Median Household Income	0.017543
Percent Persons Below Poverty Level	0.160476

Figure 2. Table of correlation between COVID-19 deaths and various demographic or macroeconomic features.

Interestingly, median household income didn't seem to have as significant of a correlation as expected as seen in Figure 2. But when the percent population affected was graphed against the deaths and hue set by median household income, an interesting relationship was evident. As seen in Figure 3 the states with a higher median household income didn't have less deaths but had a lower percent of their population affected.

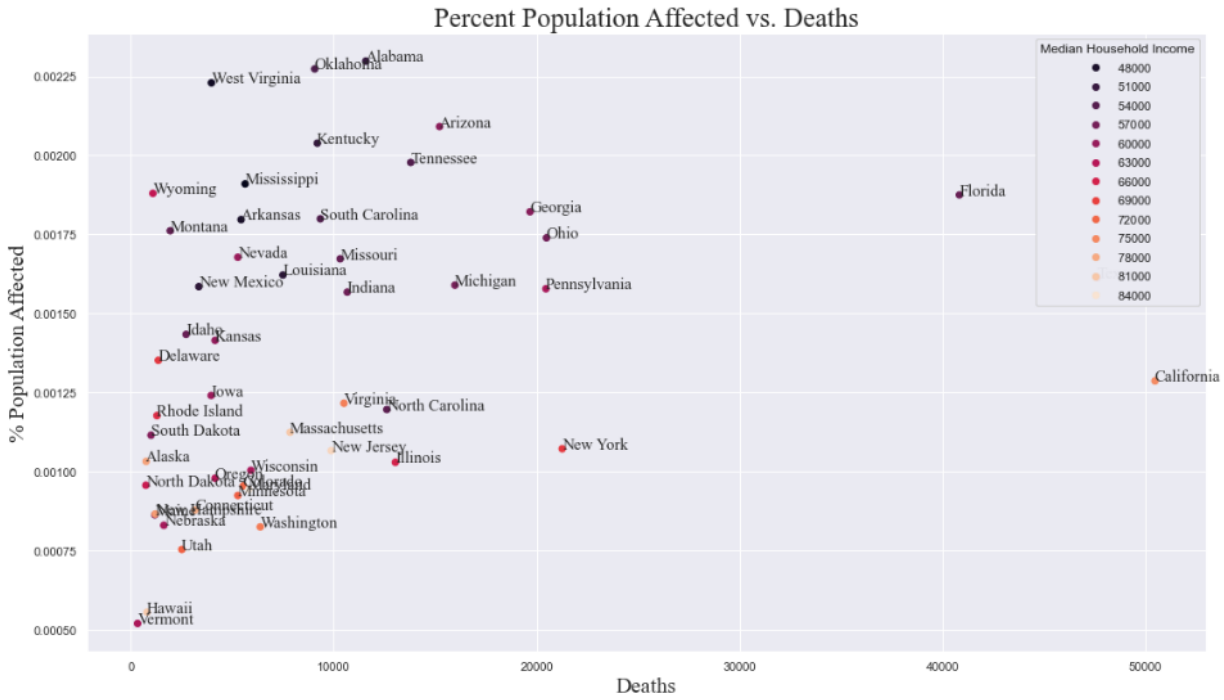

Figure 3. COVID-19 total deaths per state graphed against the percentage of the population affected. Each data point is colored based on the states' median household income.

The months and year where COVID-19 deaths peaked were also considered for the United States overall. In Figure 4 below we can see that COVID-19 peaked in the United States as a whole from March-April 2020, November-January 2020/2021, July-August 2021 and November-February 2022. The first time frame aligns with when COVID-19 was initially recognized and the world entered into the pandemic.

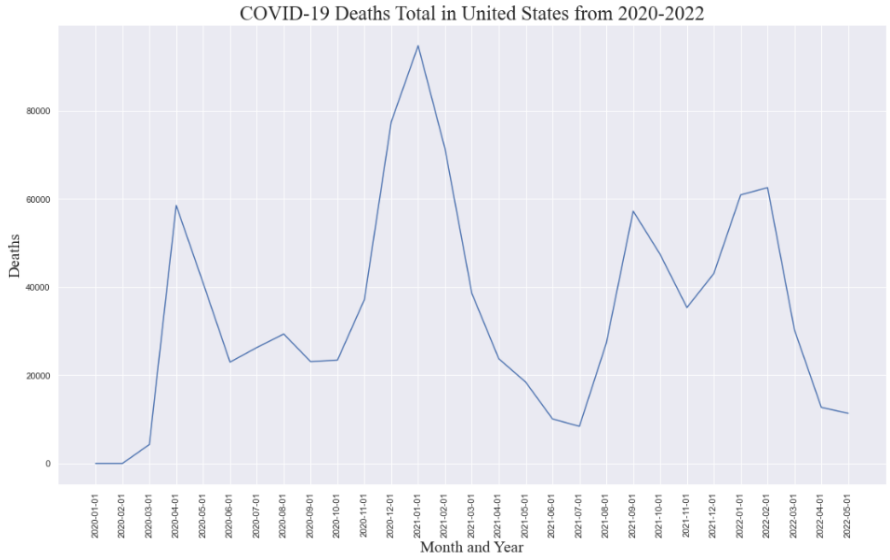

Figure 4. COVID-19 total deaths per month in the United States from 2020-2022.

I also studied the peaks for the top five states with the highest number of total deaths. These visualizations can be seen below in Figures 5-9.

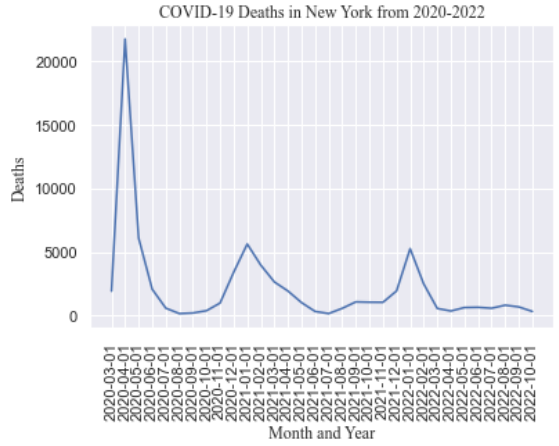

Figure 5. COVID-19 deaths in New York by month from 2020-2022.

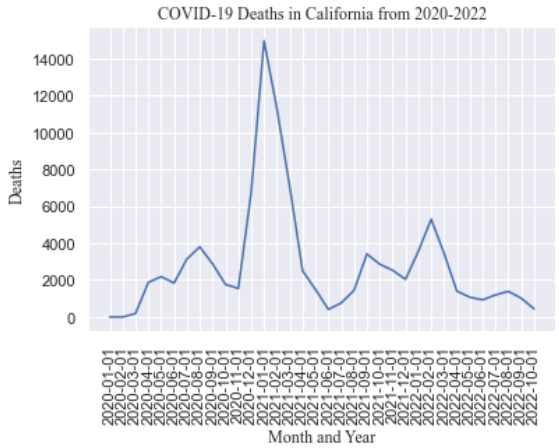

Figure 6. COVID-19 deaths in California by month from 2020-2022.

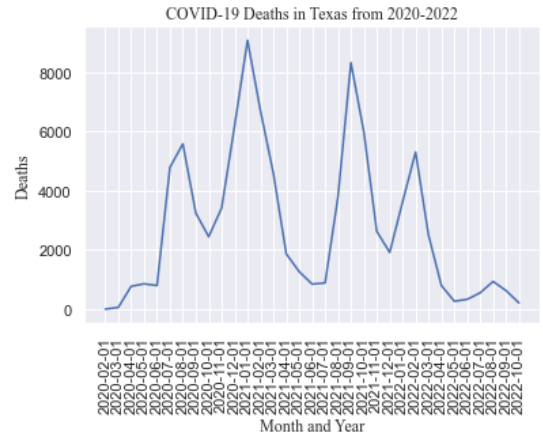

Figure 7. COVID-19 deaths in Texas by month from 2020-2022.

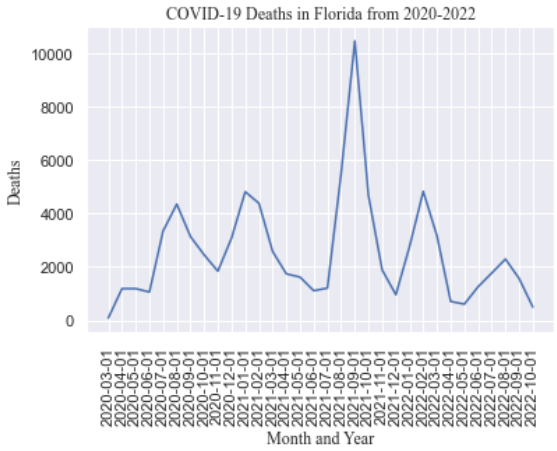

Figure 8. COVID-19 deaths in Florida by month from 2020-2022.

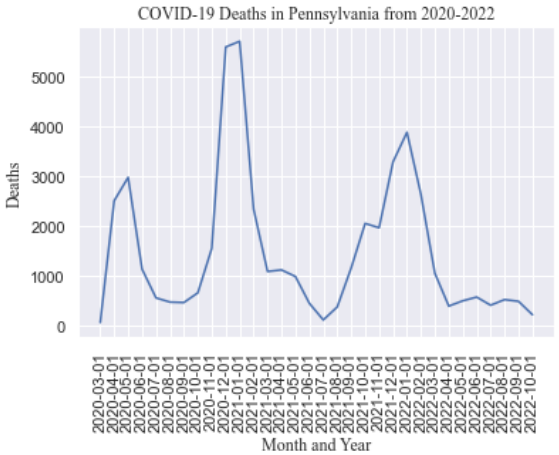

Figure 9. COVID-19 deaths in Pennsylvania by month from 2020-2022.

California's peaks most closely resemble the peaks of the United States as a whole. Did the state of California respond to and abide by various mandates most reflective of the entire population? New York's large peak during the beginning of the pandemic matches closely with the media attention they gained as one of the COVID-19 hotspots in April of 2020. The close housing proximity of cities there as well as heavy reliance on public transportation could've also affected the initial rapid spread of a virus we didn't know much about at the time. Texas seems to have had a steady pattern of peaks and troughs throughput 2020-2022.

November to January of 2020/2021 matches with the November-February peak in 2022. These are both holiday seasons so we looked into the possibility of deaths increasing from before Christmas to after by running ANOVA on those time periods. The boxplot of this relationship can be seen below in Figure 10.

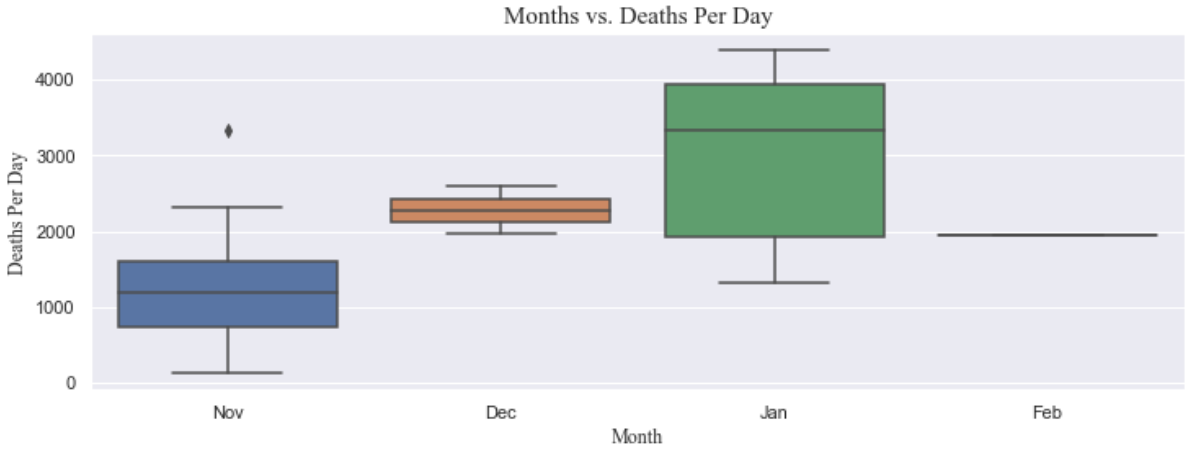

Figure 10. Boxplot of the range of deaths in the United States per day in each month from 2020-2022.

The null hypothesis was that the holiday season did not affect the number of COVID-19 deaths in the United States. The f and p values for ANOVA test of these results were [13.05428214] [0.06879249] for the holidays period from 2020-2021 and [2.38099311] [0.26278734] for the holiday period from 2021-2022. According to the boxplot, the deaths per day seem to increase post-Christmas in January. One possible hypothesis for the increase in deaths can be due to holiday travel and visiting friends, family and loved ones during that time.

Pre-processing

In order to be able to group states based on their shared macroeconomic, demographic and health features, a dataframe with all this information needed to be aggregated on the state level. I altered the data frames to be uniform in column name capitalization, data type and replacing missing values with the median of the corresponding column. The values for these features also needed to be scaled using the standard scaler to fit and transform the final dataframe.

Modeling

The states were grouped using a Kmeans clustering model. Initially, the number of clusters was set at three as a way to explore the model. This cluster analysis can be seen in Figure 11.

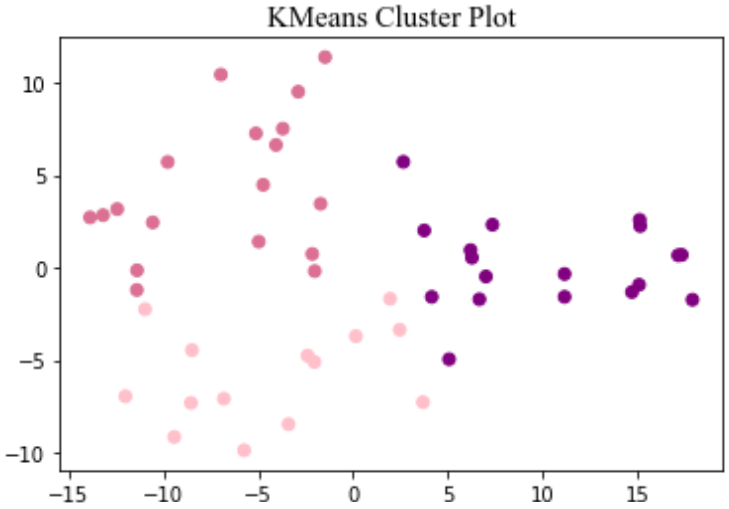

Figure 11. Kmeans cluster plot of all demographic, macroeconomic and health features for each state.

As seen in Figure 12, a principal component analysis (PCA) was then run to show that the elbow indicated the optimal number of clusters being three. This would be an ideal number from the business perspective as states could be easily clustered and recognized by high, medium and low level impact by COVID-19. The plot of inertia versus number of clusters also agreed with a lower inertia, ideal for states within the same group, at a number of clusters being three. The Kmeans cluster plot shows how the states could be generally grouped into three levels of response to the COVID-19 pandemic. Overall, the features used in the final dataframe covered a comprehensive list of demographic, macroeconomic and health indicators. Future analyses would benefit from additional data on each states' compliance to various mandates as well as the effectiveness of different responses to COVID-19 such as vaccine/booster availability.

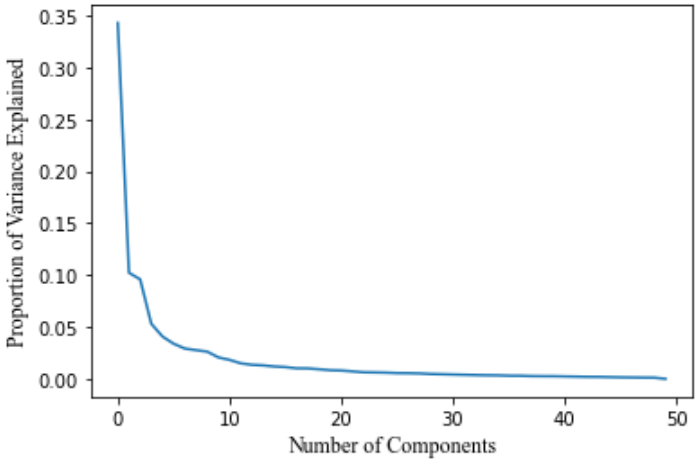

Figure 12. Principal component analysis to determine ideal number of clusters

Sources:

[1] “Coronavirus.” *World Health Organization*, World Health Organization, https://www.who.int/health-topics/coronavirus#tab=tab_1.

[2] *Impact of COVID-19 on People’s Livelihoods, Their Health and Our Food Systems*. 13 Oct. 2020, www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems.