W200, Summer 2020, Section 2

Team 5: Jay Venkata, Ziling Huang, Leyla Greengard, George Jiang
Link to Github repository: Project-2_Venkata_Greengard_Huang_Jiang

COVID-19 Analysis across US States

Introduction

In late December 2019 news came of the emergence of a new virus in Wuhan, China. At first people thought that it could not be passed from human to human (NY Times, Jan 10) but that turned out not to be correct. People started dying, and the disease spread quickly, first in East Asia, then to the Middle East and Europe and finally it arrived in the US. By January 20th, 2020 many countries including the United States had confirmed cases. At the end of January the World Health Organization declared a global emergency, and as the cases continued to rise across the world, it was declared a global pandemic. In this study we will analyze the evolution of the pandemic in the United States. We will address the following questions and provide a few takeaways for state governments to deal with the pandemics of the future.

1. What are the mortality trends across all the US states during this health crisis?
2. Is state-level policy action effective in slowing down the infections?
3. What is the relationship between mortality rate, infection rate and the percentage of seniors and children living in the same household?
4. How does health insurance coverage correlate with mortality rate and infection rate?
5. How are hospital provider capacities related to mortality rates?
6. Are socio-economic factors correlated with higher infection rate?

Data cleaning and sanity checks

The data leveraged for our analysis came from a github site maintained by the Center for Systems Science and Engineering at Johns Hopkins University[1]. The file also included the data for two cruise ships, Princess Diamond and Grand Diamond. These ships were not attached to any state. We ignored them in the analysis. We also ignored the data from American territories such as American Samoa, Mariana Islands, and others. Sanity checks were performed by graphing the data and making sure that the numbers seem reasonable. When in doubt we checked against other sources.

We performed transpositions on the data to ensure consistency for dataframe merges. We observed the shape of the data to ensure consistency after joining datasets. Additionally, we checked for the presence of null values which could affect the analysis.

1. Mortality trends across all the US states

[1] CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE

Methodology and approach

The file contained cumulative numbers of confirmed cases and deaths, on a daily basis from 1/22/2020 to 7/17/2020. The information was provided at the "Admin" level, which was similar to county level. However, in some cases, such as for example New York City, the "admin" data was not divided by county. We aggregated all the data by state.

Analysis

As of July 17, NY and NJ were the two states with the highest casualties in the United States, with about 404,000 and 176,000 to date, while the next higher number is MA with 112,000. Some of this can be attributed to the very large population density in those areas. We will explore factors that influence the numbers for NY and NJ in the sections below.

Mortality by state: We show in the graph below the number of deaths per 1,000,000 inhabitants in the state. Here again New Jersey is an outlier, followed by New York State. So the large number of deaths is not only a result of the large population. It may also be due to the pervasive use of public transportation. Many people in the tri-state area (NY-NJ-CT) commute by train or bus to NY city for work. Note that Connecticut has the third largest number of casualties behind NY and NJ. .

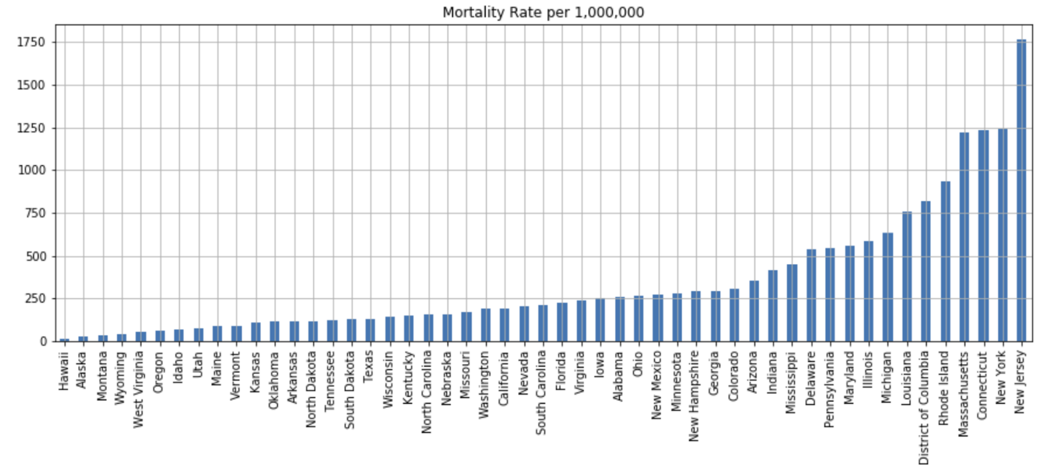

Figure 1

Recent Evolution of Mortality rate

Methodology and approach

For this graph, the number is calculated as the percentage increase of the average daily mortality over the past 14 days as of 7/17/20 over the 14-day average daily mortality as of 7/3/20. For example, for Montana, the graph shows that the mortality rate has increased by 60% over the past 2 weeks. This graph displays the *rate* at which casualties are increasing as of 7/17/20. In order to smooth out the daily fluctuations, we calculated the average increase in mortality rate over the past 14 days, and compared it to the same number from the previous 2 weeks.

Analysis

Montana, Texas, Arizona, Arkansas and South Carolina are the states with the highest rates of growth. New York has one of the slowest rates and New Jersey is in the middle of the pack.

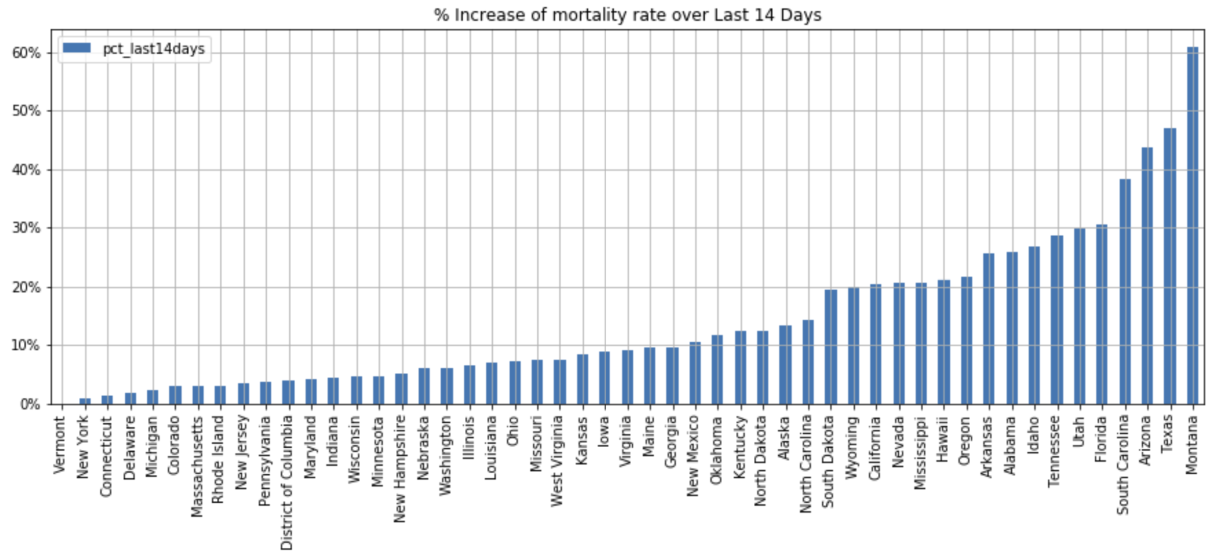

Figure 2

2. Is state-level policy action effective in slowing down the infections and mortality?

Reopening Dates and Mask Mandates

Data and context

In addition to the data above, we used states' open/close/mask mandate dates which were obtained from the National Governors Association[2]

Methodology

We first look at the effect of locking down and reopening the states on the number of cases. In the graphs below we show the evolution of the infection (total number of cases) in conjunction with the closing and reopening dates (red lines), as well as the mask mandate (green line). We chose Arizona and New York, at the two extremes of the change in mortality rate (Montana had only a

[2] https://www.nga.org/coronavirus-state-actions-all/#NE

small number of cases, hence rates may be skewed, and Texas did not have a clear stay-at-home mandate).

Analysis

As can be seen in the graphs, both states opened while the numbers of infections were increasing (upward sloping blue lines), but Arizona reopened (on May 15) while the rate was still increasing (the curve was becoming steeper) while New York did it on June 8, when the rate of growth was going down. The enactment of the mask mandate in NY does not show a notable decrease, and it is too early to determine if it has had any effect in AZ.

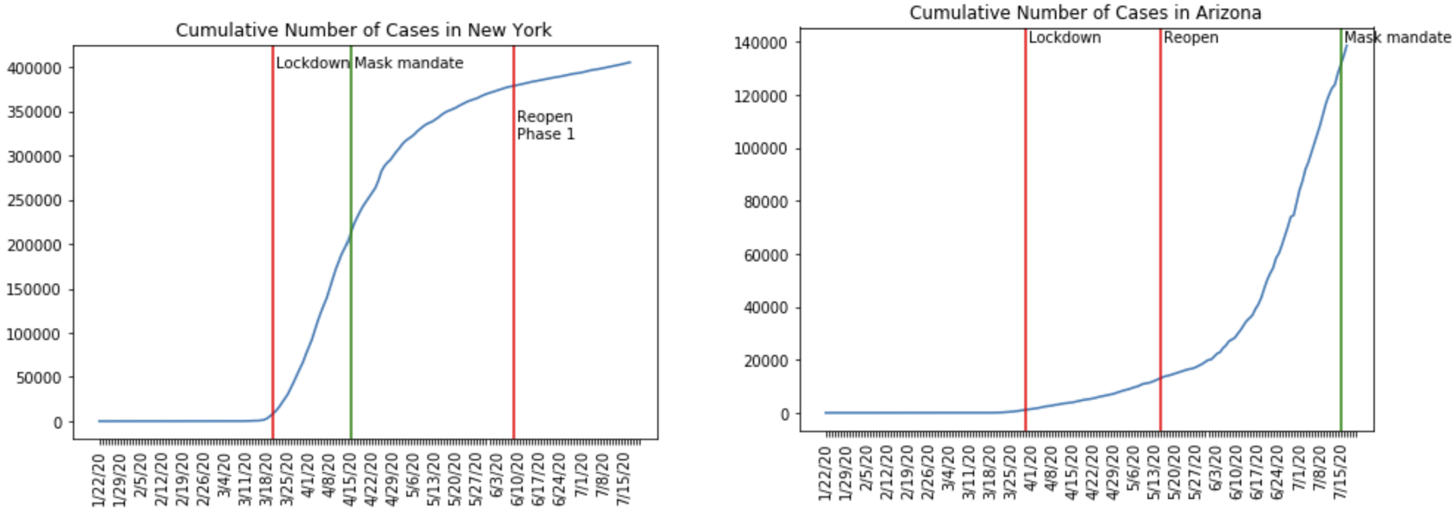

Figure 3

Figure 4 displays the evolution of mortality rate over time. It may be surprising that the mortality rate in Arizona was always higher than in New York. The reason is that the number of casualties in Arizona was always (and still is) lower than in New York, hence any increase is likely to represent a higher percentage of the existing number. This also explains the fact that the AZ line is much less smooth than the NY line. The graph shows that Arizona opened when its death rate was stable while New York opened it while its death rate was decreasing.

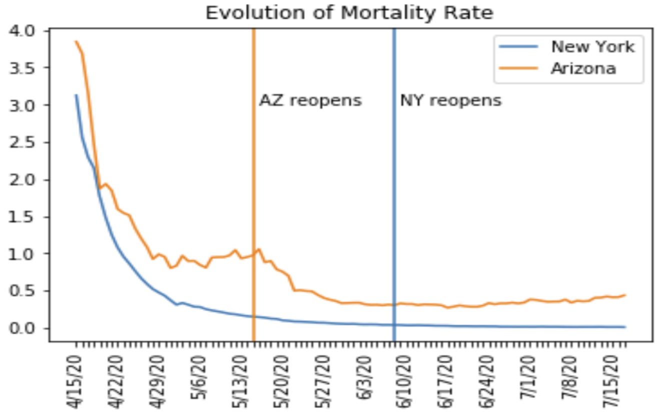

Figure 4

3. The Relationship between Mortality Rate or Transmission Rate and Percentage of Seniors Living with School-age Children

Data and Context

The data for 50 states as of July was taken from JHU[3], KFF[4], RT [5]and joined. This has been a central topic on many people's minds given the push to reopen schools.

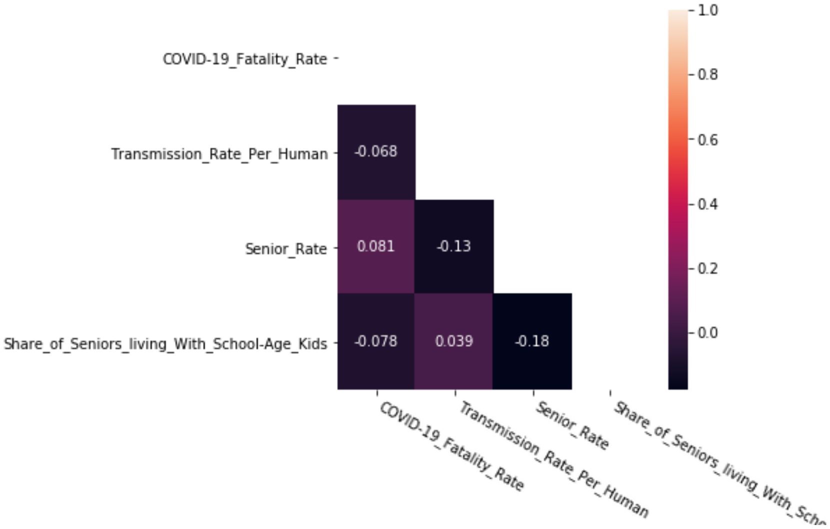

Figure 5: Correlation heatmap of Proportion of Seniors with Transmission and Mortality Rates

Methodology and Approach

[3] CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE
[4] https://www.kff.org/statedata/collection/covid-19-and-related-state-data
[5] https://rt.live/

We calculated the Senior Rate by taking Total Seniors in each State and dividing by Total Population in each State. Share of seniors living with school age children was derived by taking the number of seniors living with school children divided by total seniors in each state. Fatality rate was available per state from John Hopkins University data. Transmission rate per human was taken from RT and tells us the average number of people who will contract the disease from an infected person.

Analysis

A correlation heatmap was in Figure 5 plotted using data points from these 50 states. The results showed that there was low correlation between the Transmission Rate and Mortality Rate with Percentage of Elderly Living with School Children Per State. The results seem logical given that most schools are not slated to open till fall. The effects of schools reopening will not be felt until that goes into effect.

4. Is there a relationship between insurance coverage and mortality rate or infection rate?

Data and Context

Data on insurance coverage in 50 states was taken from KFF[6] and joined with mortality rate taken from JHU[7] and transmission rate from RT[8]. The question stemmed from general concern around people losing insurance coverage tied to employment thereby exacerbating mortality rate from inability to afford medical treatment. For those who did not have insurance prior to the crisis, lack of regular health checkups and poor health may have increased their susceptibility to infection.

[6] https://www.kff.org/statedata/collection/covid-19-and-related-state-data

[7] CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE

[8] https://rt.live/

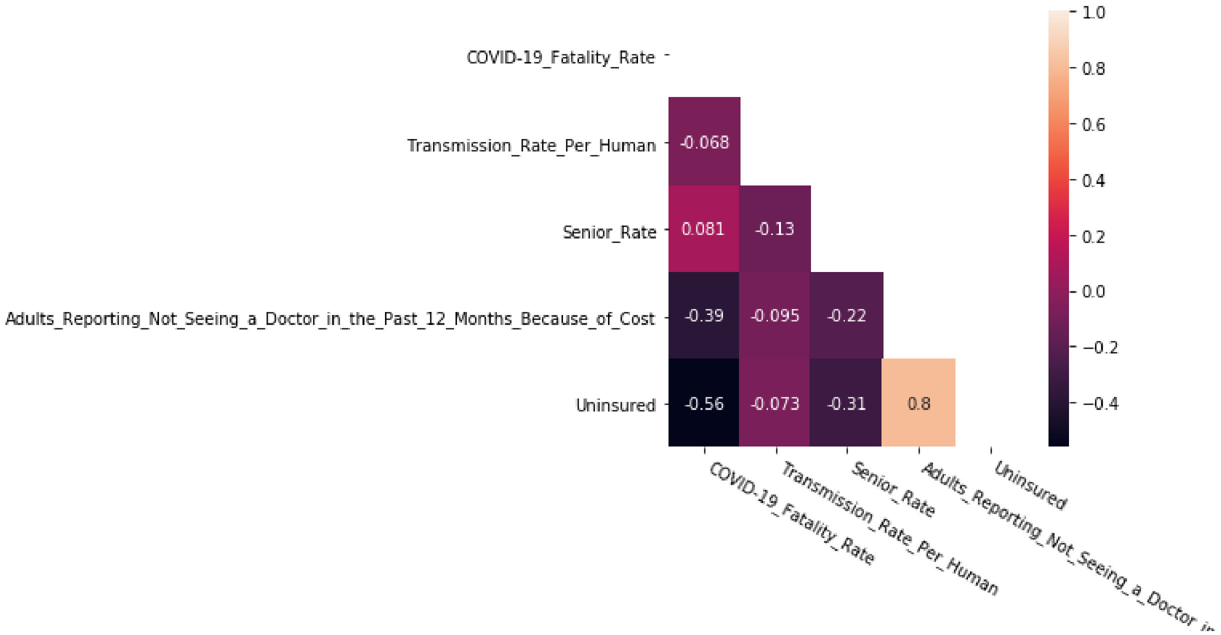

Figure 6 : Correlation heatmap of Proportion of Uninsured with Transmission and Mortality Rates

Methodology and Approach

Information did not exist around the number of elderly who were uninsured. We calculated the Senior Rate by taking Total Seniors in each State and dividing by Total Population in each State and included that in our plot. Next, data points such as Percentage of Adults Reporting not Seeing a Doctor in the Past 12 months, Uninsured Percentage by State and Transmission Rate were included.

Analysis

A correlation heatmap was in Figure 6 plotted using data points from these 50 states. The results showed that there was some negative correlation between the Uninsured Percentage by State and Mortality Rate by State which runs contrary to our initial expectations. On further consideration and research, it was found that most uninsured individuals are younger than 35. It makes sense that the mortality rate would be lower for younger individuals. A better test would involve comparing young individuals with or without health insurance. Unfortunately, that data was not available. Uninsured Rate per State and Senior Rate by State had low correlation. On the other hand, Adults Not Going to the Doctor due to Costs and Uninsured Percentage showed highly positive correlation. There was low correlation between Transmission rate and Uninsured Percentage as confirmed by the heatmap in Figure 6.

5. **How are hospital provider capacities related to mortality rates?**

Data and context

Hospitalization rate is one of the most important metrics that is being tracked through this health crisis. We wouldn’t have heard about COVID if there were a lot of COVID cases and none of them led to hospitalizations or deaths. Unlike positivity rate and number of cases, hospitalizations are a lagging indicator since it takes some time for new cases to turn into hospitalizations, ICU admissions and deaths, especially if testing is widespread and timely.

Methodology

For the hospitalization rate, we leverage these 2 metrics - Hospital admissions per 1,000 population by ownership type, 2019, and Hospital Beds per 1,000 population by ownership type, 2018 from our KFF dataset. We plot these 2 metrics against the COVID-19 deaths per 1,000,000 population. This would ultimately show a relation between hospital capacity and deaths due to COVID.

Analysis

From Figure 7A below, we can see that Hospital Admissions is linearly correlated with the COVID death rate. This can be explained based on the fact that only the most serious patients get admitted to hospitals.

From Figure 7B below, we can see that there is minimal correlation between hospital beds against COVID-19 deaths. This proves that hospital provider capacities by State doesn't play a role in COVID mortality rates in the United States. This metric could be attributed to the high quality of health infrastructure in the US in comparison to other countries.

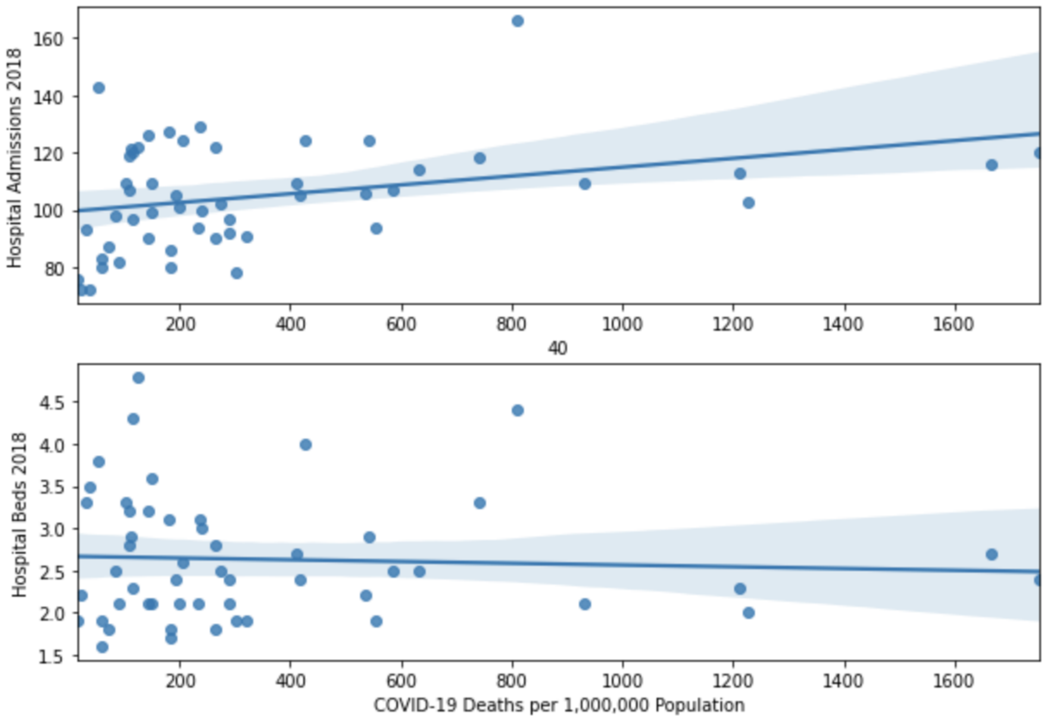

Figure 7A and 7B

6. Are socio-economic factors correlated with higher Covid-19 infection and death rate?

Data and Context

Socio-economic factors are an important metric to track in this health crisis, because intuitively people from low income households are most impacted by the Covid-19. This study explores whether that intuition is true. Its finding can help governments direct better policies to help control the crisis. The data used in the analysis is from the KFF database[9].

Methodology

Death rate is derived by dividing confirmed death by total confirmed cases. Infection rate is derived by dividing confirmed cases by total state population. Social economic factor is approximated by the unemployment rate during March, because we want to look at the social economic factor before Covid-19 infection took off to determine whether socio-economic factors are a meaningful predictor of infection and death rate.

Analysis

In figure 8 we see that at the state level there is no strong relationship between the socio-economic factor (approximated by unemployment rate) of the State and the infection rate of the virus. This

[9] https://www.kff.org/statedata/collection/covid-19-and-related-state-data/

seems counter intuitive, as we would expect that States with higher unemployment to have higher infection rates as lower income people are more likely to work in jobs that require human contacts. However, this is consistent with the findings that most infection cases are found in States with traditionally high income states such as NY and CA, because these States have higher population density and more tourism, which are the primary drivers of the spread of the diseases.

Interestingly there is no relationship between States' unemployment rate and death rate (figure 8). This suggests that wealth is not a factor on whether a person survives the disease. One possible reason is that currently we don't have effective treatment for the virus, so even if a person has better access to healthcare due to economic factors, she is not more likely to get better from the diseases.

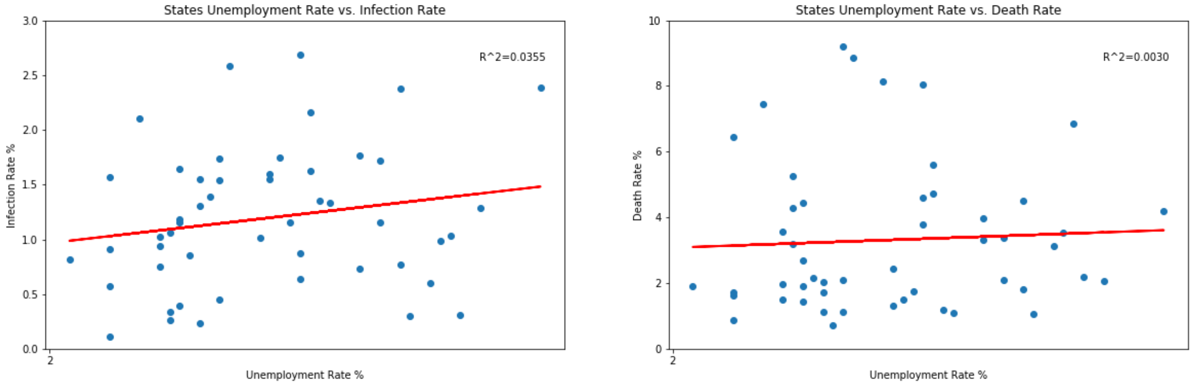

Figure 8: Relationship Between State Unemployment Rate and Covid-19 Infection, Death Rate

However, just because we don't see correlation between the State's socio-economic condition and infection/death rate, it doesn't mean there is no relationship between the two factors. There are so many confounding factors between States such as population density and tourism popularity, it is difficult to compare the States directly without controlling for these factors. Therefore, we should instead compare the States with similar population density and tourism popularity. Looking through the online data[10], we found several States with similar population density and tourism popularity. New York and Florida, Tennessee and Georgia are of special interest because not only do they have similar population density, they are also similar in terms of tourism attraction.

After controlling for population density and tourism attraction, we see in figure 9 that States with higher unemployment before Covid-19 took off also have higher infection rate and death rate. This is consistent with our expectation that States with higher unemployment have both higher infection and death rate as lower income people are more likely to work in jobs that require human contacts, and these people are also less likely to get quality healthcare.

[10] https://simple.wikipedia.org/wiki/List_of_U.S._states_by_population_density;
https://www.worldatlas.com/articles/the-most-visited-states-in-the-us.html;
https://www.businessinsider.com/the-most-popular-us-states-for-tourism-2014-10

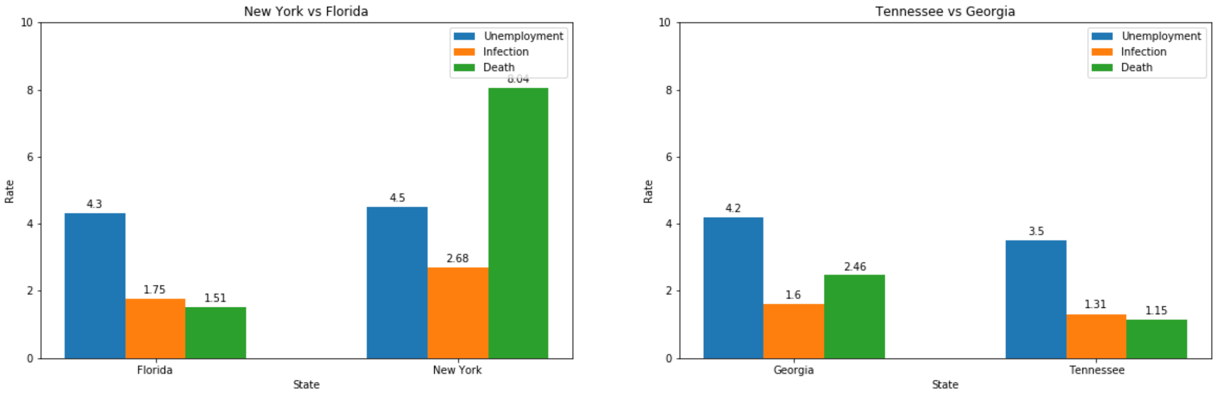

Figure 9: Bar Chart of State Unemployment, Infection Rate, and Death Rate

In conclusion, at first glance there seems to be no relationship between socio-economic factor and infection/death rate at State level. However, after digging deeper into the data and controlling for confounding factors, we see socio-economic factors do play a role in causing higher Covid-19 infection and death.

Takeaways for US States based on COVID-19 analysis conducted

1. Our analysis of the state level data shows that while some states, particularly in the Northeast have had a large number of deaths both on an absolute level and relative to their populations, the increase in their casualties is slowing down, while that of other states, such as Montana, Texas and Arizona are increasing.
2. Our analysis shows that the actions in NY state were effective, and while Arizona had the same actions, they were not effective. This may indicate that the timing of policy is as important as the policy itself.
3. Hospital capacities in the form of the number of available beds doesn't play a role in the mortality rate due to COVID between US states. This can be attributed to the higher standard of health infrastructure across all US states in comparison to other countries.
4. Because people from low income households are more likely to get infected and die from Covid-19, it is important for the government to provide extra support to them.
5. We are unable to prove that there is increased infection rate or mortality amongst the elderly living with school children. While the results are inconclusive, this is mainly because schools remain closed. The government should rely on the advice of public health experts when weighing in on the topic.
6. We are unable to prove a positive correlation between lack of health insurance and increased infection or mortality rate through this study. Age is a confounding factor. If the data were available, a better study could be conducted to compare health outcomes of people in the same age group with and without health insurance.