Behavioral Economic Factors and COVID-19

COVID-19 has been raging across the entire world since early this year. This pandemic shut down the world for months. No one has been able to escape it.

Several behavior factors had been said to affect the spread of COVID-19. Mask use and social distancing are mentioned the most. Less has been mentioned about economic factors such as income and healthcare costs. How strong is the relationship between COVID19 infections and mortalities? Do states with higher healthcare costs per capita have higher COVID-19 infections and/or mortality rates per capita? Do states with higher uninsured healthcare rates have higher COVID-19 infections and/or mortality rates? Do states with low mask usage have higher rates of covid19 infections? Does higher mask use correlate to income levels? This document will discuss mask use, income and healthcare cost and tries to determine the effects these factors have on COVID-19 infections and mortalities.

# **Datasets**

At the beginning our project, we reviewed several datasets. A couple of days later we narrowed that focus down to four main data sources:

* Census 2018 ACS-5
* New York Times Mask Usage
* Covid Tracking Project
* Center for Medicaid and Medicare

## **Census API**

The Census API contains several hundred tables. Our analysis includes three tables from the Census API: Total Population, Health Care Coverage by Sex by Age, and Median Household Income. Additional tables Race by Sex By Age and Poverty By Age were also reviewed but not included in the final analysis.

Jupyter notebooks was used to connect to the Census API and download table data. All columns from each table was downloaded and columns names were updated to more meaningful names. Some columns were also combined.

## **New York Times Mask Use**

The New York Times (NYT) has a COVID-19 repository on GitHub. Even though it contains several interesting datasets, the team focused on the mask use by county data. This csv file included counties listed by FIPS and five response categories: Never, Rarely, Sometimes, Frequently and Always.

The NYT mask use data is a based on a survey conducted in July of 2020. 250,000 people surveyed. No demographic information is listed for this sample.

## **Covid Tracking Project**

The COVID Tracking Project (CTP) bills itself as “a volunteer organization launched from The Atlantic and dedicated to collecting and publishing the data required to understand the COVID-19 outbreak in the United States.” The project is a collecting several different datasets. Our teams focused on the current covid positives and deaths data by state.

The CTP current covid positives and death data was gathered using the API provided by the CTP. Connections to the API was created in a jupyter notebook by saving the json response to lists, combining lists into a dictionary and converting the dictionary to a pandas dataframe.

## **Center for Medicaid and Medicare Services**

Healthcare cost data was difficult to discover. However, we were able to find information for the Center for Medicaid and Medicare Services. This dataset is called Total All Payers Per Capita State Estimates by State of Residence 2014 - Personal Health Care (Millions of Dollars).

As stated in the title, this data is from 2014. No later data has been found on their site. This data was taken from an excel spreadsheet.

# **COVID-19 Infections and Mortalities**

To determine how strong the relationship was between COVID-19 infections and mortalities, data from the COVID Tracking Project was used. Plotting the COVID-19 infections and mortalities by state shows a large variation between each state. Are some states outliers for COVID19 infections?

To answer this, the box plot of all the states data, shows a median of 74,325.5 COVID19 infections overall. Based on our calculations with an IQR of 121,811 COVID19 infections, we determined states with greater than 322,393 COVID19 infections as outliers. This represented the states of California, Florida, Texas, and New York. Are there also states that are outliers for COVID-19 deaths.

Similarly, we found a median of 1,253 COVID19 deaths from all of the state’s data. We determined that states with greater than 9407 deaths as outliers. This also represented California, Florida, Texas, and New York with the addition of New Jersey.

we wanted to find the linear regression model of COVID-19 infections and COVID-19 mortalities in the United States. With an extremely small p-value of 1.32e-14, not surprisingly a strong positive correlation was found. The r-squared value was 0.67 which shows that our linear regression model accounts for 67% of the variation. With a slope of 0.02, our model would predict that for every 100,000 increase in COVID19 positive cases, there would be an additional 2,000 COVID19 deaths. This gives us a case fatality rate of 2% which is in line with estimated CFR in the literature of 2-3%.

# **COVID-19 and Healthcare Costs**

We wanted to study if there is a correlation between total healthcare cost per capita and COVID-19 infection and mortality rates. Our null hypothesis says that there is no correlation between the total per capita healthcare cost and Infection rates and Mortality rates.

From our analysis, we found there is a negative correlation. A moderate inverse relationship between the healthcare cost per capita and infection rates. Therefore, as the healthcare cost per capita increases increases the infection rate decreases. Since the P value is very low, we can confidently reject the null hypothesis and thus say there is a relationship between total healthcare cost per capita and infection rates. With a R-squared value of 0.16, the data shows that the linear regression model itself accounts for 16% of the variation.

There is no relationship between total healthcare cost per capita and mortality rates. Hence, it proves our null hypothesis that there is no clear relationship.

# **COVID-19 and Uninsured Rates**

These are plots showing the relationship between the state uninsured as a percent of the total population vs state Infection and Mortality Rates as percentages of the population.

For the first graph: With a P-Value of .002 and correlation coefficient (r) of ? indicates that there is a statistically significant positive correlation between the uninsured population and infection rates in the United States.

In our linear regression model, the R squared value is .17 which indicates that around 17 percent of the variation in the infection rates can be explained in due in part to the states uninsured rates.

Why states with higher uninsured rates are linked with higher COVID19 positive rates is an interesting finding. We speculate that other factors related to being uninsured, such as social and economic factors like poverty and race may contribute to this trend. Moreover, are states with higher uninsured rates less able to practice social distancing behaviors?

For the second plot: the p-value is 0.39 suggesting that these variables are most likely not related, and we cannot infer any relationship between state uninsured rate and state mortality rates from our data.

# **COVID-19 and Mask Use**

The choropleth map shown below represent NY times survey conducted in July 2020. Participants were asked how likely they were to wear a mask when leaving their house. Participants were given the options of responding ‘ALWAYS’, ‘FREQUENTLY’, ‘SOMETIMES’, ‘RARELY, or ‘NEVER’. The map below represents states ranked by and scaled according to the response of ‘ALWAYS’.

The lighter end of the spectrum represents the states with the highest responses for ‘ALWAYS’ wear a mask when leaving the house, the darker end represents those with lowest tallies for the same response. This map visually depicts attitudes toward mask use around the country in the month of July. It is important to note that highest percentages for mask use are seen around parts of the country that were experiencing the largest number of COVID-19 cases at the time which implies that mask use may be related to COVID-19 risk.

A box-and-whisker plot was used for our descriptive statistical analysis of COVID-19 mask use data. The median is skewed in four of the five mask use categories, except for the ‘SOMETIMES’ mask use category. Outliers exist in the ‘NEVER’ (North Dakota) and ‘FREQUENTLY’ (Alaska) mask use categories. Overall, COVID-19 data is not normally distributed within the five mask use categories.

The figure below displays scatter plots of the median household income vs each mask use category along with the linear regression. A causal relationship between median household income and mask use can be concluded for only the ‘ALWAYS’ mask use category. This positive correlation exists for median household incomes of less than $60,000 per year. Essential workers fall into this income level and would ‘ALWAYS’ wear mask because of occupational requirements.

# **Conclusion**

COVID-19 infections and COVID-19 mortality to be strongly related and predictive. Economic factors of healthcare costs and uninsured rates were linked to COVID-19 infections only. There is no link to mortality.

Economic factor of median household income partly predictive of ‘ALWAYS’ mask use. High median household income and ‘ALWAYS’ mask use are clustered in the northeast states. Interestingly this region in the country with the lowest COVID-19 infections since survey data time.