**Behavioral Economic Factors and COVID-19**

The novel virus COVID-19 has been spreading across the world since early this year. The current pandemic has virtually shut down the world. Life, for many, is drastically different from our pre-pandemic time. The rate of COVID19 infections correlates with COVID19 deaths. We hypothesized that a linear regression model would be predictive of fatalities based on the total number of infections.

Next, we set out to analyze several behavior factors that affect the spread of COVID-19. Many scientists in the CDC and WHO urge mask use and social distancing as this virus primarily spreads from respiratory droplets and aerosolization. Although economic factors such as household income, healthcare costs, and health insurance rates are known determinants of health, less is known about these factors impacting COVID19. We hypothesized that higher healthcare costs per capita have higher COVID-19 infections and mortality rates per capita. We hypothesized that states with higher uninsured healthcare rates have higher COVID-19 infections and mortality rates.

We hypothesized that states with high mask usage would have lower rates of covid19 infections. Moreover, higher mask use would correlate with higher household income levels.

**Datasets**

At the beginning of our project, we reviewed various datasets on the internet. Our selection was narrowed to four primary data sources:

1) Census 2018 ACS-5 API data for uninsured rates and median household income

2) New York Times Mask Usage Survey Data

3) Covid Tracking Project API Data for COVID19 infections and COVID19 deaths

4)Center for Medicaid and Medicare for personal healthcare costs

*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 notebook was used to document and execute our code and plots. Jupyter was used to connect to the Census API and download table data. All fields from each table was downloaded, and column 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 impressive datasets, the team focused on county data's mask usage data. The NYT mask use data was a survey conducted in July of 2020. The sample size is 250,000 people surveyed from around the United States. This csv file included counties listed by FIPS and five response categories: Never, Rarely, Sometimes, Frequently, and Always for mask use. A limitation of this data set is that no demographic information is available.

*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 collecting several different datasets concerning COVID-19. Our team focused on the cumulative COVID positives and deaths data by state.

The data was gathered using the API provided by the CTP website. Connections to the API were created in a Jupyter notebook by saving the json response to lists, combining lists into a dictionary, and converting it to a Pandas dataframe.

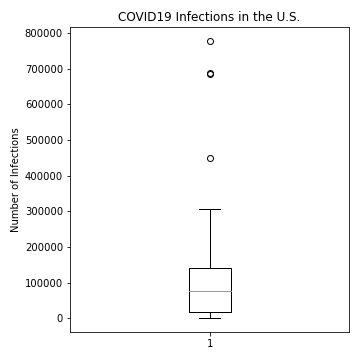
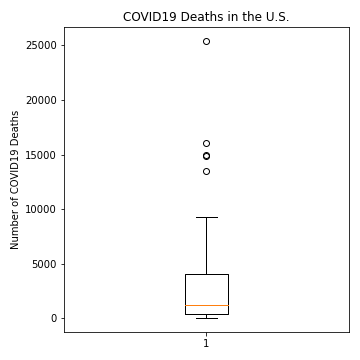
*Center for Medicaid and Medicare Services*

The dataset "Total All Payers Per Capita State Estimates by State of Residence 2014 - Personal Health Care (Millions of Dollars)."

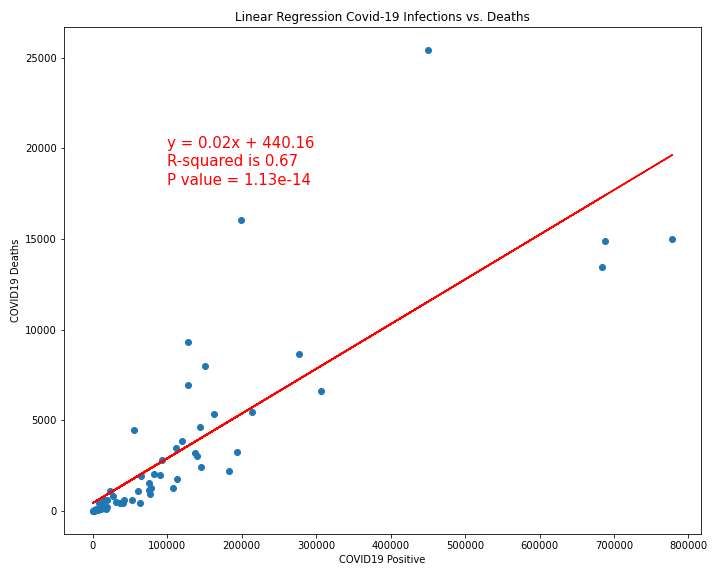
Of note, 2014 was the latest dataset were able to locate that had personal healthcare costs broken down by the state level. This data was taken from an excel spreadsheet.

**COVID-19 Infections and Mortalities Analysis**

We analyzed which states were outliers for the number of COVID infections and deaths? A box plot created visually to show the outliers. We used the Pandas quantile function for the quantitative analysis to determine the outliers. The median for COVID19 infections is 74,325 cases in the U.S. We determined states with greater than 322,393 COVID19 infections as outliers that represented California, Florida, Texas, and New York. For COVID19 deaths, we found a median of 1,253 deaths. We determined that states with greater than 9407 deaths as outliers. These outlier states were California, Florida, Texas, New York, and New Jersey.

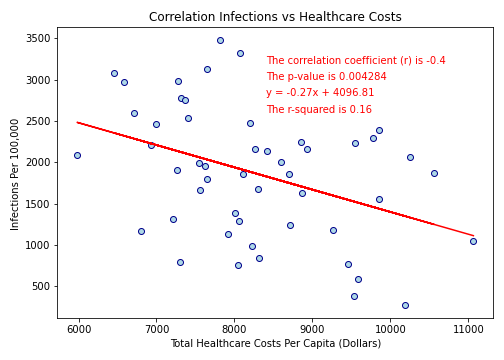
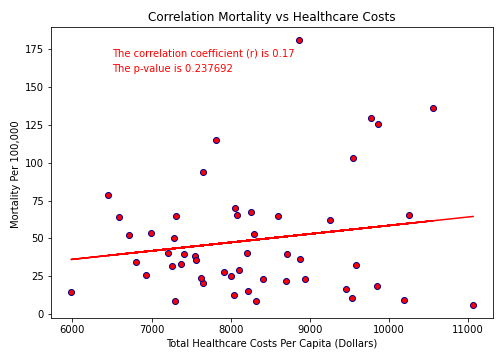
To determine a linear regression model of COVID19 positivity and COVID19 deaths in the United States. With an extremely small p-value of 1.13 X 10^-14, a strong positive correlation was found. The r-squared value was 0.67, which shows that our linear regression model accounts for 67% of the variation. Our linear regression model equation was y= 0.02x + 445.05. With a slope of 0.02, our model would predict that for every 100,000 increase in COVID19 cases, there would be an additional 2,000 COVID19 deaths. Thus, this gives us a CFR (case fatality rate) of 2%, which is in line with the estimated CFR in the scientific literature of 2-3% for COVID19.



**COVID-19 and Healthcare Costs Analysis**

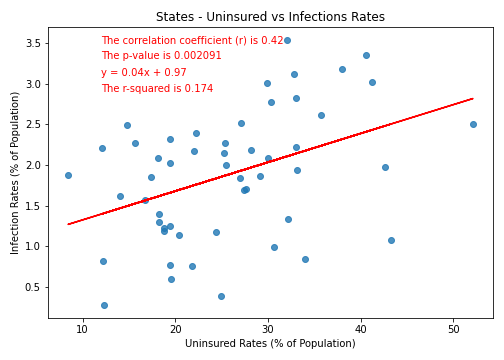
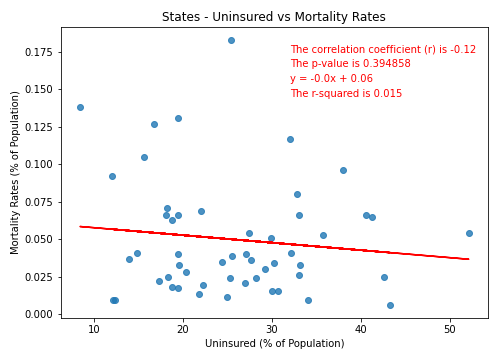
We wanted to study if there is a correlation between total healthcare costs per capita and COVID-19 infection and mortality rates. From our analysis, we found that there is a negative correlation. A moderate inverse relationship between healthcare costs per capita and infection rates (r = -0.4, p-value = 0.004). Therefore, as the healthcare cost per capita increases, the infection rate decreases. Since the P-value is very low, we can confidently reject the null hypothesis and say there is a relationship between total healthcare costs per capita and infection rates. With an R-squared value of 0.16, the data shows that the linear regression model accounts for 16% of the variation.

There is no relationship between total healthcare costs per capita and mortality rates (r=0.17, p-value 0.23). Hence, the null hypothesis could not be rejected, and a relationship was not found.

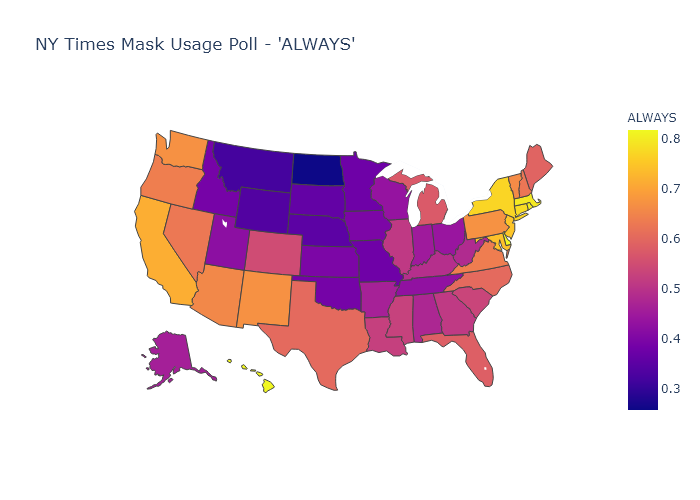
**COVID-19 and Uninsured Rates Analysis**

The plots show the relationship between the state uninsured as a percent of the total population vs. state infection and mortality Rates as percentages of the population. A moderate positive correlation was found between uninsured rates and infection rates (r=0.42, p-value = 0.002). The linear regression model showed that the R squared value is .17, which indicates that around 17 percent of the infection rates variation can be explained by the state uninsured rates. There was no correlation found between uninsured rates and mortality rates (r=-0.12, p-value = 0.397).

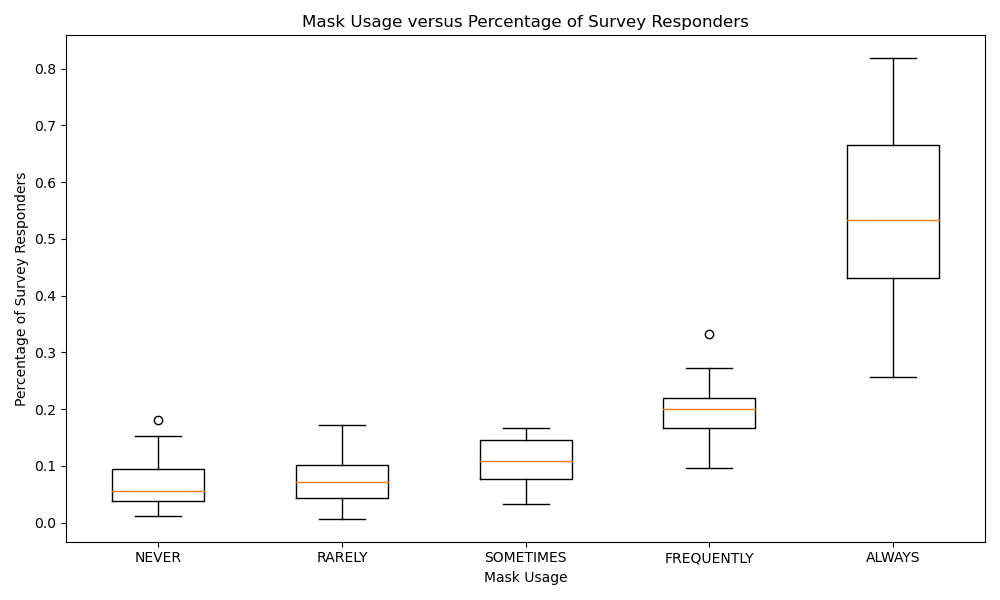
 

**COVID-19 and Mask Use Analysis**

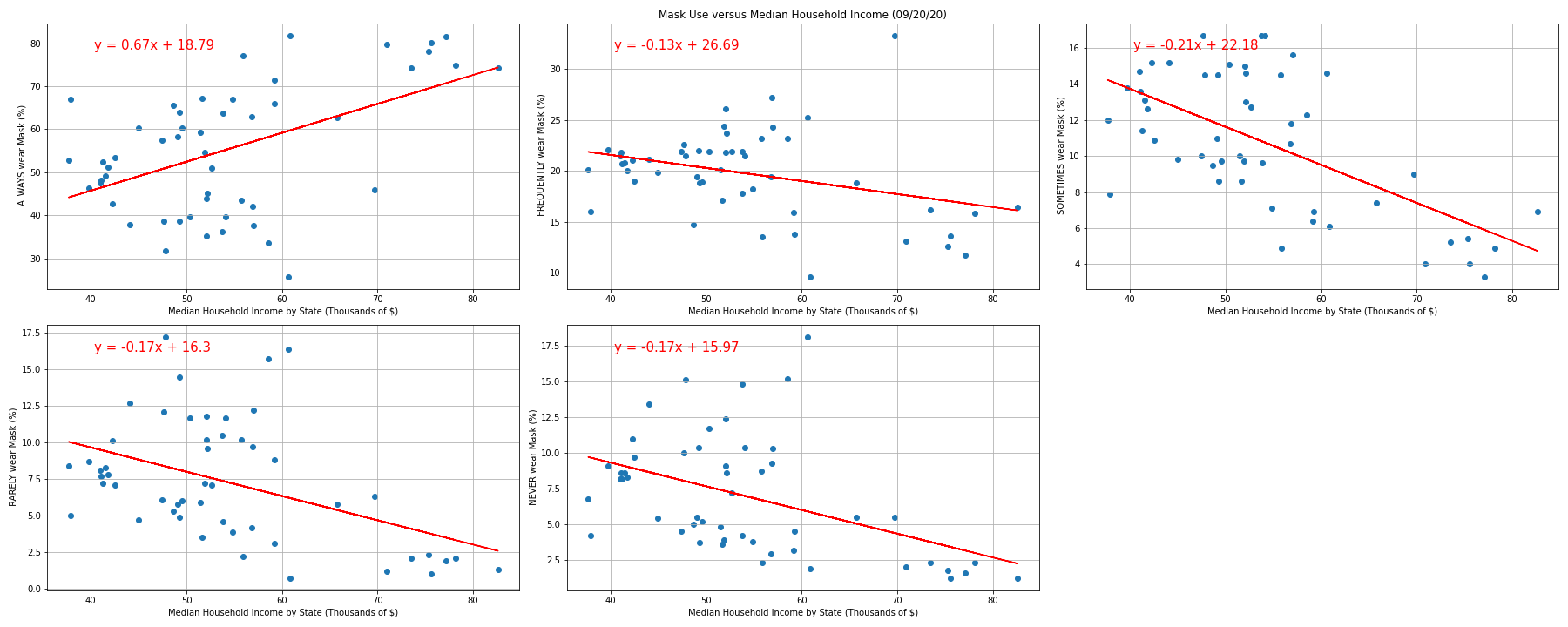
A choropleth map was created from the NY times mask usage survey conducted in July 2020. The map represented states ranked by and scaled according to 'ALWAYS' for mask use response. It is important to note that the highest percentages for mask use are seen around parts of the country experiencing the largest number of COVID-19 cases early in the pandemic, which may imply 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, mask usage data is not normally distributed within the five mask use categories.



Next, we created scatter plots of the median household income vs. each mask use category along with the linear regression. A moderate positive correlation between median household income and 'Always' mask use was found (r=0.5, p-value <0.001). A linear regression model with a r-squared value of 0.25 was found. Therefore 25% of the variation seen between 'Always" mask usage and median household income can be explained by our model.



**Conclusion**

Although we conclude that COVID19 infections data from COVID Tracking Project to be strongly correlated and predictive for COVID19 mortality, the economic factors that we analyzed—healthcare costs and uninsured rates were associated with only COVID19 infections, but not mortality. This implies that states with high uninsured citizens may be at higher risk of more widespread COVID19 infections, but it should not affect mortality rates. We are uncertain why a trend in states with lower healthcare costs per capita and higher COVID19 infections was seen. This was the opposite trend, as predicted by our hypothesis. Further investigation is needed to determine the implications of this relationship.

Finally, the economic factor—median household income—is partly predictive for "always" mask use wearers. Our geo-mapping showed that by July of 2020, high mask usage responders from the survey tended to cluster in the northeast region. Interestingly, this is also the region of the country with some of the highest household income rates. Since early July of 2020, when the mask usage survey was conducted, the northeast region continues to have the country's lowest COVID-19 infection rates. We believe that high mask use in the northeast region is a significant contributor to this low COVID-19 infection rate.