Major Findings

**Is there a statistically significant difference in Covid-19 infection rates between U.S. States?**

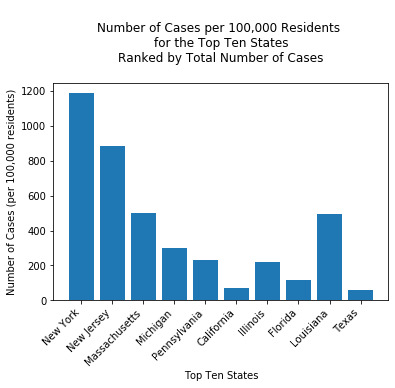
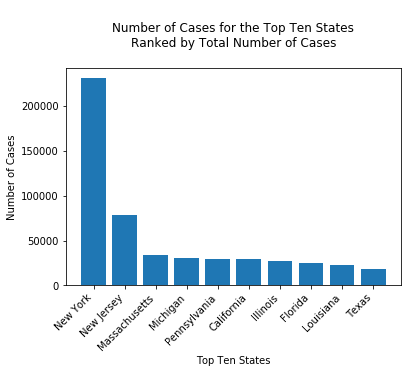
We conducted a chi-squared analysis to show that the difference between infection rates for individual states versus the national average was statistically significant.

Our chi-squared test-statistic (1309982.674) was larger than the critical value (66.339). We can reject the null hypothesis that there is no difference in infection rates between states. While this test tells us that the difference between the number of cases per state and their expected values is unlikely to be caused by chance alone, it does not offer additional insight as to why this difference occurred.

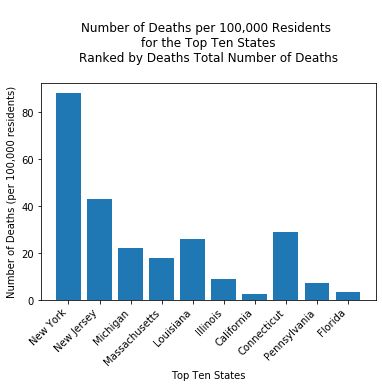
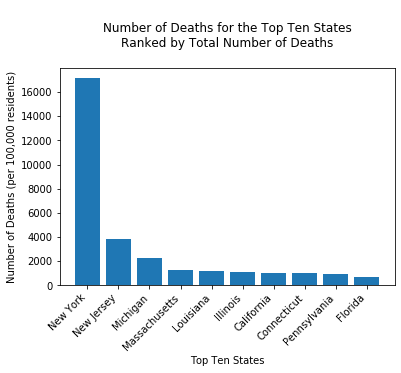
Knowing that the difference between state infection rates is significant allows us to continue our analysis. Had we been unable to reject the null hypothesis then the rest of our analysis would be irrelevant due to the lack of difference between states.

After determining that there was a difference between states, we first looked at the top ten states by total number of cases and plotted them as a bar chart to visualize the difference.

For each of those ten states we also plotted the number of cases per 100,000 residents to visualize how each state differed in terms of infection rate. See figures below.



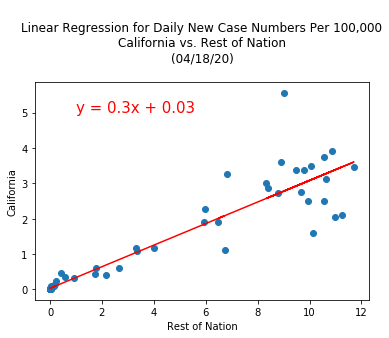
Next we took the top ten states by total number of deaths and then plotted the number of deaths, and the number of deaths per 100,000 residents for each of these states to see how each state differed in terms of death rates per capita. See figures below.



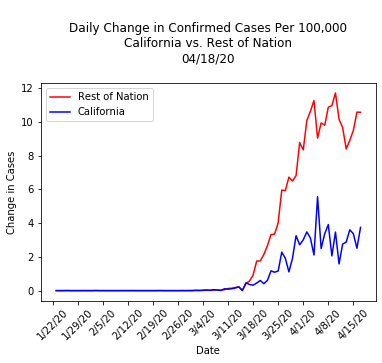
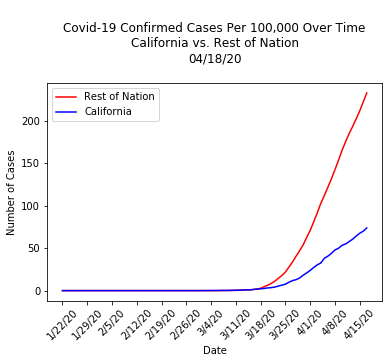
From these charts and our chi-squared analysis we can see that states are experiencing significantly different rates of infection. This analysis also helped to show the difference between total numbers and numbers per capita. One motivation for this analysis was to standardize numbers based on population to contrast the abundance of media reports of statistics without a valid basis for comparison (ie. headline news about USA passing Italy for number of cases when the two countries have very different total population counts).

**How does California compare to the rest of the nation?**

In this portion of our analysis we compare the state of California to the rest of the nation in order to determine whether we as a state are doing better, worse than, or equal to the rest of the nation.

We started with a linear regression model (below) to check for correlation between California and the rest of the nation. We can see that there is a strong positive correlation (r-squared = 0.0934) between the daily number of new cases for California and the rest of the nation. Using this model we predict that California's cases will follow the trend of the rest of the nation. The model also predicts California will experience roughly 1/3 of the cases experienced by the rest of the nation (line equation = 0.3x + 0.03).

Knowing that the two variables are correlated, we now want to investigate whether the apparent difference in cases numbers per day is statistically significant. The charts below illustrate the apparent difference in infection rates between California and the nation.



To test for differences between California and the rest of the nation we conducted 2 t-tests to test for difference in means.

The first t-test compared California’s daily change in cases to the rest of the nation. After conducting the t-test, we were able to reject the null hypothesis that there is no significant difference between the daily number of new cases for California vs. the rest of the nation (p-value=0.0001).

The second t-test compared the total number of cases for California vs. the rest of the nation. After conducting the second t-test above, we are able to reject the null hypothesis that there is no significant difference between the number of confirmed cases for California vs. the rest of the nation (p-value = 0.002).

Using these test results combined with the linear regression model created above, we determined that California's number of new cases per day is positively correlated with the rest of the nation but the number of new cases per day is significantly less than the rest of the nation.

One conclusion that could be drawn from these results is that California is experiencing the same general trend for daily new Covid-19 cases as the rest of the nation, but the severity of our infection rate is lower than the rest of the nation. This shows that California is doing better than the rest of the nation in terms of infection rate but does not tell us about the reason for the difference.

**Is there a correlation between temperature and number of infected cases?**

We wanted to look at how higher or lower temperatures increase the spread of infection. We used Meteostat data API to find historical weather data. We narrowed the scope to monthly average temperatures and top infected states.

Two main problems occurred while trying to analyze these data sets. One was that the Meteostat API stopped pulling data after February, so March and April statistics could not be run. Second, because of how I chose to narrow the scope, there was not enough data points to use for a better analysis.

I used two data sources to create a data frame. One source was for the monthly average temperatures and the other showed the total number of cases in a given state per month. I realized that there was not enough varying data to create an accurate graph of what I wanted to analyze. March and April is when more data for Covid 19 cases were released so it would have been useful if I had more information to run this analysis.

A simple scatter plot was used to graph the data and calculate the correlation coefficient and plotted a line of regression. The coefficient for January was 0.16 and 0.31 for February. The low correlation coefficient tells there is a weak correlation between temperature and infected cases. If we were to accept these results, we can assume that temperature and number of infected cases are not related.

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated