Group 5: COVID-19

Kyle Cielencki, Artie Edwards, Brian Yu, Christion Lankford

Background/Timeline:

- In an attempt to discover how different parts of the US are being impacted by COVID-19, we built analyzed Johns Hopkins data to explore the statistics on the coronavirus pandemic. This includes visualizations, explanations of the presented metrics, and the details on the sources of the data.
  - **Hypotheses**: Population density plays a major role of infection rate. Infections rates of change are mutually exclusive to state.
  - **Outcome**: Although there is correlation between population and confirmed cases. There is no correlation between infection rate by state
- **Timeline:**
  - **December 31, 2019**: WHO says mysterious pneumonia sickening dozens in China
  - **January 11, 2020**: China reports first death
  - **January 20, 2020**: Other countries, including U.S., confirmed cases
  - **January 30, 2020**: The W.H.O declares a global health emergency
  - **February 6, 2020**: The U.S. report first death in California
  - **February 11, 2020**: The W.H.O proposed an official name for the disease, “COVID-19”
  - **March 26, 2020**: U.S. becomes the planet’s most infected nation
  - **April 11, 2020**: *Group 5* investigates the trends and factors to the COVID-19 outbreak in the U.S. Leveraging data from John Hopkins and US Census to explore areas, such as, but not limited to: population, density, and demographics (data from Jan 2020 – Apr 2020)

Description of Data:

- We utilized the Johns Hopkins COVID-19 dataset as our primary source *(please refer to the appendix)*. We joined this data with demographic data from the Census API to create a single dataset for more information on the regions analyzed. Additionally, we utilized Google Maps to create heatmaps of the outbreak (please refer to the “Mapping the Outbreak” section.
  - Joining the COVID and census data sources required retrieving the ZIP Codes for each confirmed case as the COVID data is stored in “latitude, longitude” format. From here, we were able to retrieve the median household income, income per capita and education levels for the areas analyzed. This provided important information. We notice a timing discrepancy between the census data and the COVID data as the populations differed. To combat this, we calculated a percent change between the census population and the John Hopkins population and applied the growth rate to the education figures as they were in total numbers (as opposed to percent or median).
  - We also looked into the evolution of confirmed cases in different parts of the country. We primarily focused on comparing the U.S. with Texas and the U.S. (without New York) with Texas. This is due to New York being an outlier to the data set. Our initial examination was done through excel to provide a starting off point to code in Pandas.
    - We broke down the analysis into two major categories: daily differences and the percent daily change (please refer to the “Evolution of the Outbreak” section).

Data Analysis Process:

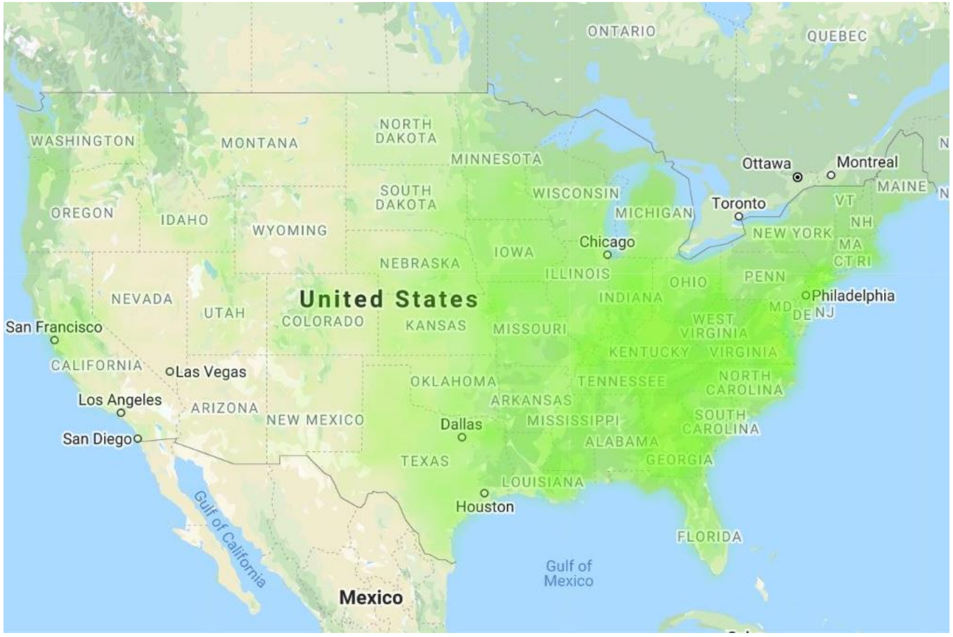
- We leveraged excel to break down the scope of the data set to determine:
    - How data was aggregated
    - The different levels of data (country, state, county)
    - Time frame (January 2020 – April 2020)
    - How the data and charts should look
- Data munging in excel as a starting off point to build in pandas
- We determined how to maximize the data with the given timeframe:
    - Decided to focus solely on the United States
    - Make comparisons of country, state, county (zip code), and latitude / longitude
    - Discuss how to create a cohesive data frame to perform individual analysis

Obstacles/Limitations:

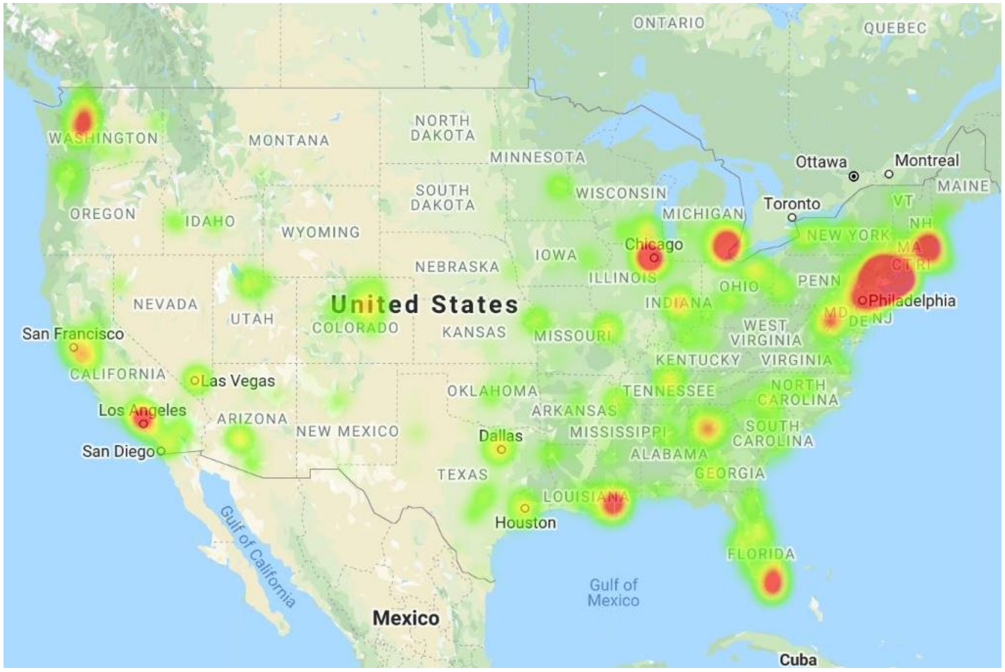
- The initial obstacle was consuming the information and figuring out how to present the findings.
    - The COVID data was aggregated daily rather than the assumed daily totals.
    - Additionally, the team had to define the columns of the data before any analysis could take place.
- Combining the John Hopkins data with the U.S. Census information presented a problem as the COVID-19 data was in latitude and longitude format and the census data was also broken out by zip code. This required extensive data cleanup and matching coordinates to each ZIP Code.
- An additional limitation was comfortability of working within pandas. Surprisingly, we did not have as many issues parsing and the transforming the data set. However, we struggled with formatting which consumed several hours to, at times, find no resolution. There were also other instances where we would apply the code to different data frames with similar structures but the solution would not properly format.
- A significant obstacle was time. We believe, if given additional time, there would be opportunities to explore different ratios within the dataframes. Ratios such as confirmed cases to county population or testing by state. Efficiency within the pandas program would have helped extrapolate more insights from the data source.

Mapping the outbreak over time:

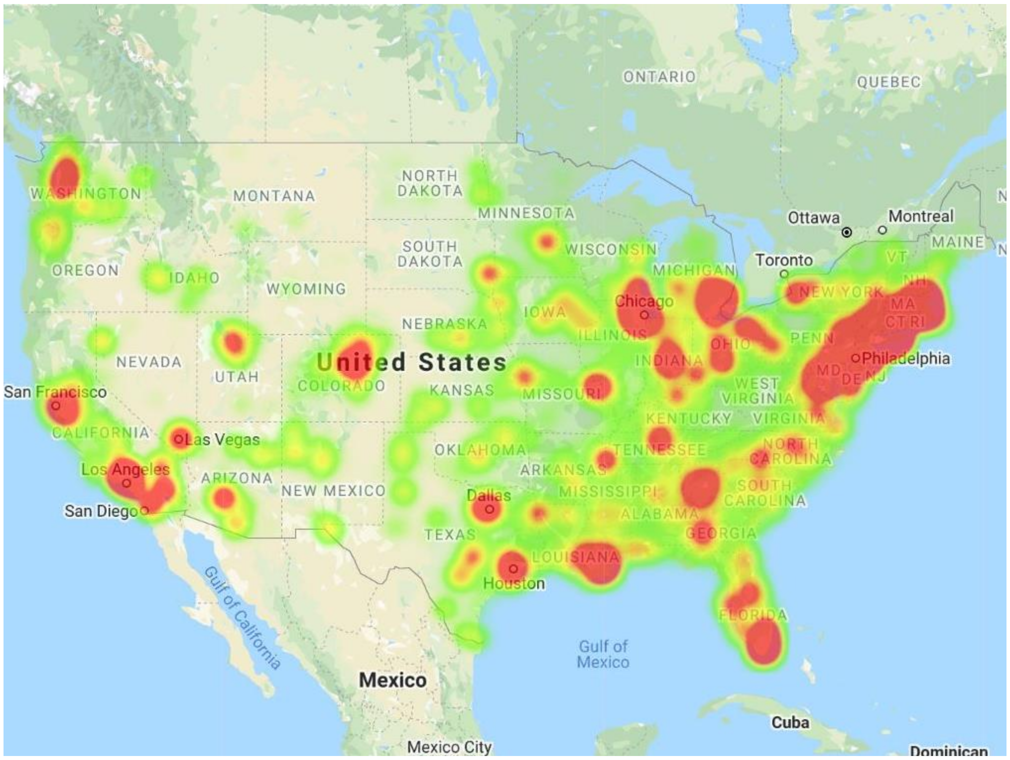
- February 28, 2020:



- **March 31, 2020**:



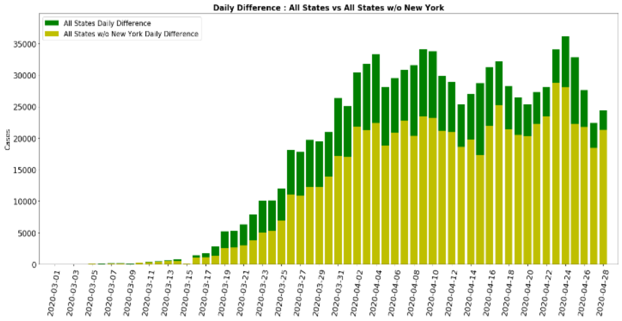
- **April 28, 2020**:

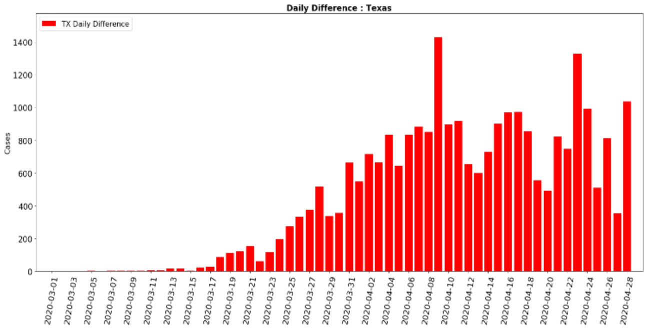


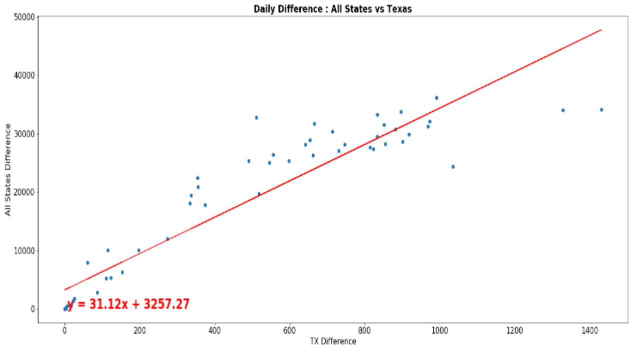
- The above heatmaps illustrate the outbreak over the months observed.
    - While the outbreak is nationwide, it is mostly concentrated in urban areas. Most notably in the Northeast, where the cities are most densely populated. This makes sense as human-to-human contact assists the spread of the virus.

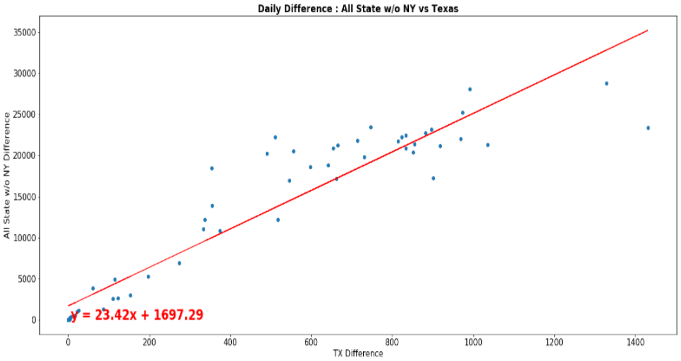
Evolution of the Outbreak:

- We looked into the evolution of confirmed cases in the United States and the relationship it had across different parts of the country. In lieu of the time, we focused on comparing the U.S. and Texas as well as U.S. without New York and Texas since New York's was an outlier to the data set. Our initial examination was done through excel to provide a starting off point to code in Pandas. We broke down the analysis into two major categories: daily differences and the percent daily change. The findings for daily difference, visually, showed a similarity when looking at the peaks and valleys of the chart. We also performed both a regression and correlation analysis to confirm the visual with high relationships with both tests (All states vs. TX: r-squared = 0.87 correlation = 0.93 ; All states w/o NY vs TX: r-squared = 0.89 correlation: 0.93) . This was interesting because we were under the assumption that each state's confirmed case trend would differ due to the unique characteristics of the state (i.e population, density, income level, etc.):

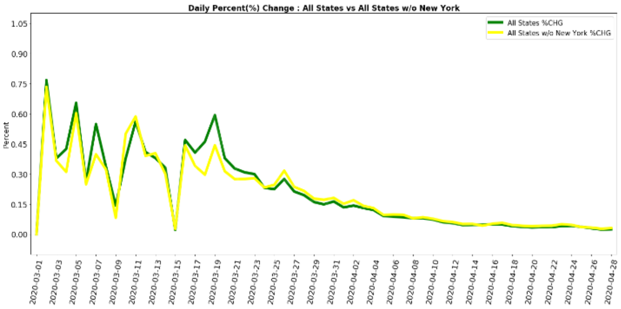


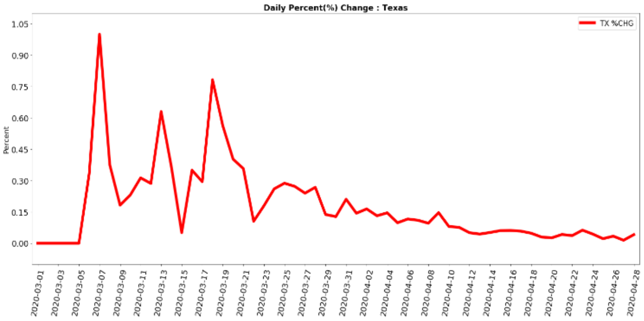






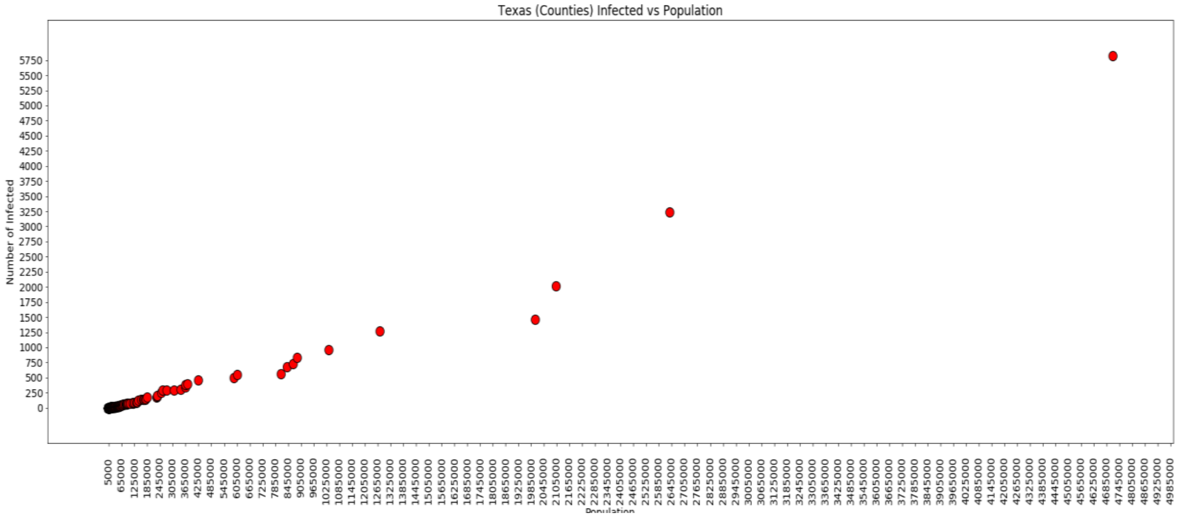
- To investigate further, we looked at the daily percent change to try and remove the noise of population size. When charting the data, we found, again, it looked very similar visually. However, when performing the both the regression and correlation analysis it proved that there was actually a low to weak relationship between the percent changes of the cumulative states and Texas (All states vs. TX r-squared = 0.32 correlation = 0.56 ; All states w/o NY vs TX r-squared = 0.23 correlation = 0.48). Thus providing evidence that our hypothesis was directionally accurate:



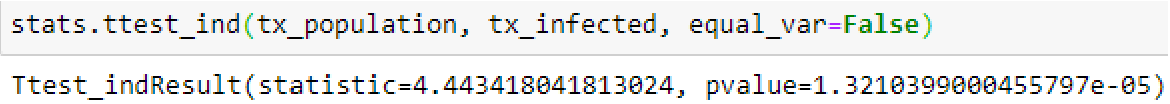


Cases across Texas & New York *(Counties)*:

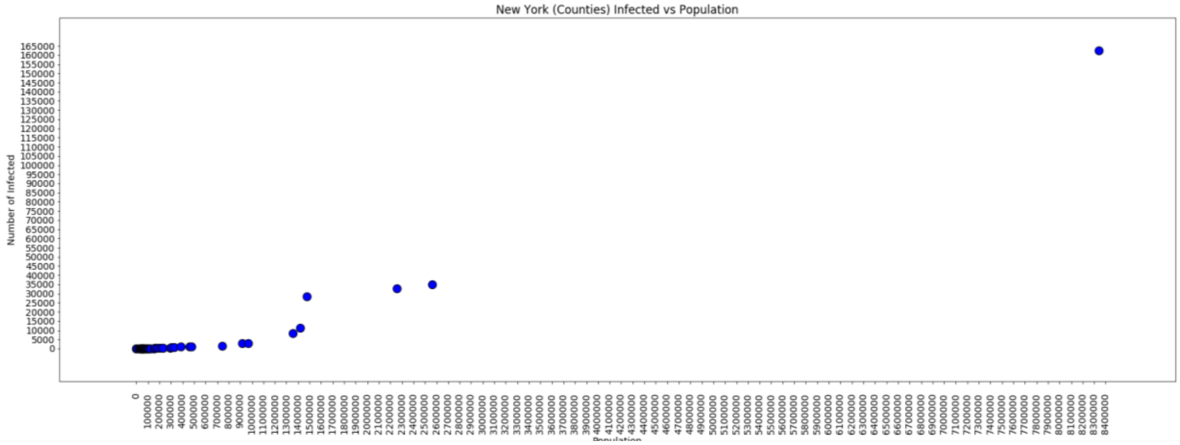
- The charts below show the relationship between the number of people infected in Texas and New York vs the population. The chart shows that as the population goes up the amount of infected people rises also for each county in the respective states.
- The leading county in Texas is Jefferson and has 5500-5800 cases.
- The leading county in New York is New York and has 164000-165000 cases.



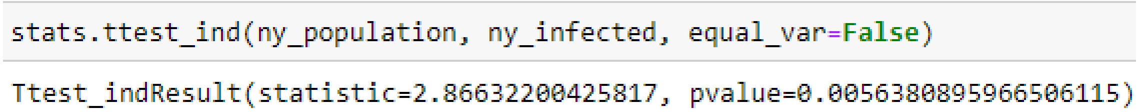
- At the 0.05 significance level, we collected enough evidence that the population and the number of infected people in Texas counties share a relationship.



```
stats.ttest_ind(tx_population, tx_infected, equal_var=False)

Ttest_indResult(statistic=4.443418041813024, pvalue=1.3210399000455797e-05)
```



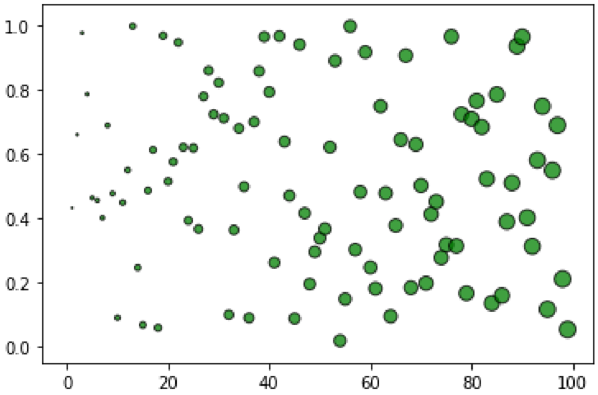
- At the 0.05 significance level, we collected enough evidence that also the population and the number of infected people in New York counties share a relationship.



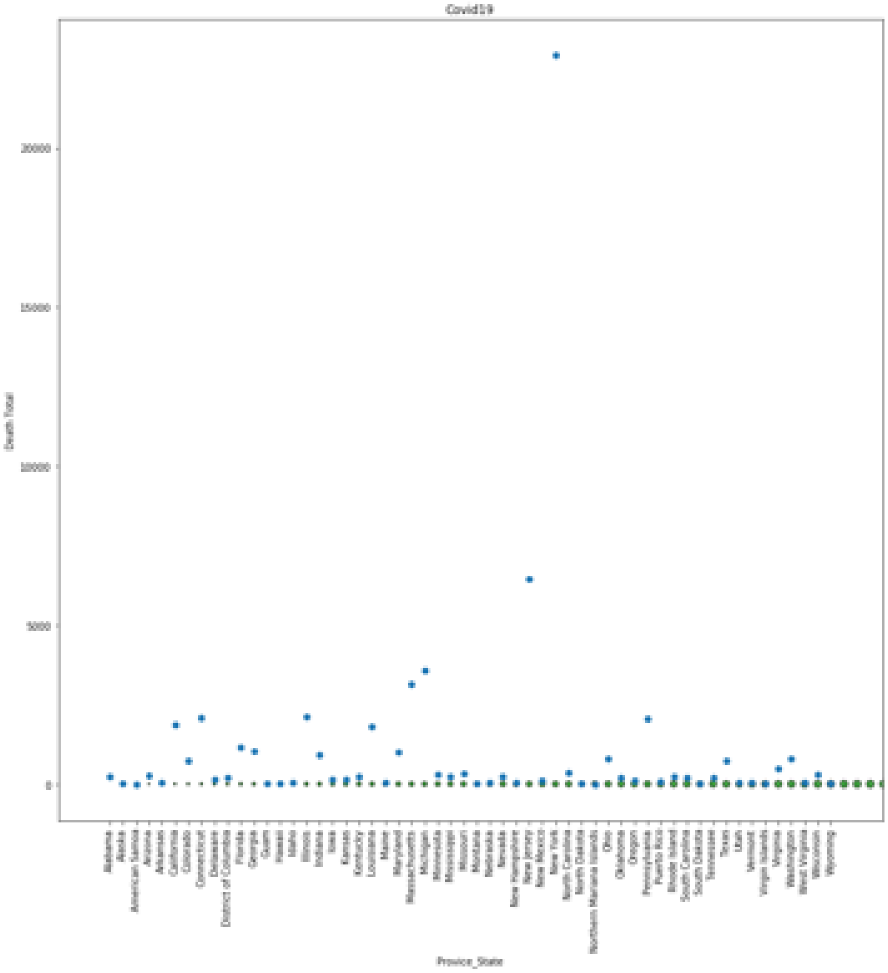
```
stats.ttest_ind(ny_population, ny_infected, equal_var=False)

Ttest_indResult(statistic=2.86632200425817, pvalue=0.0056380895966506115)
```

Deaths across the States:

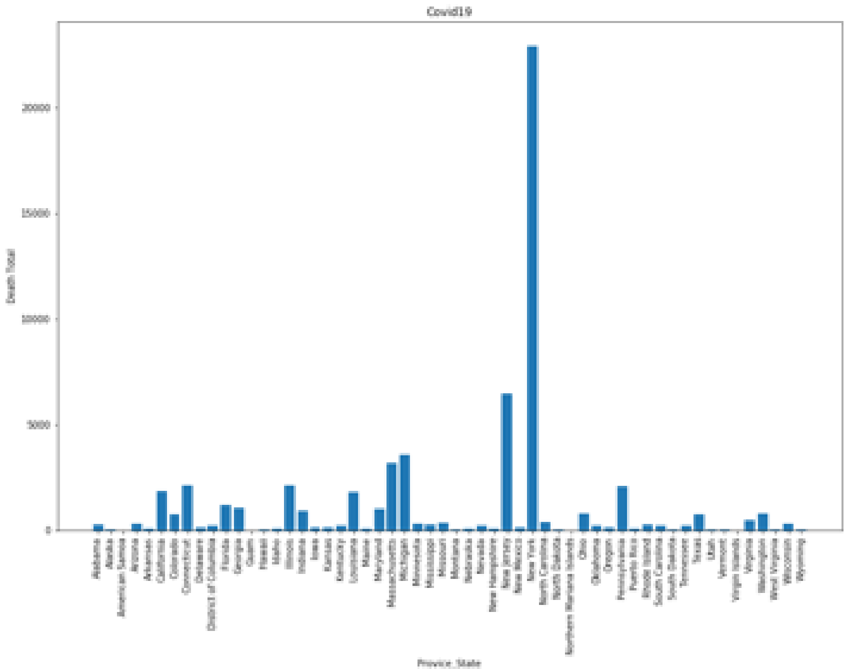
- The first scatter chart *(below)* plots the deaths for each state. The inaccurate numbers of deaths and how they are displaced on the chart doesn't give enough information that could determine number people in each state for comparison purposes:



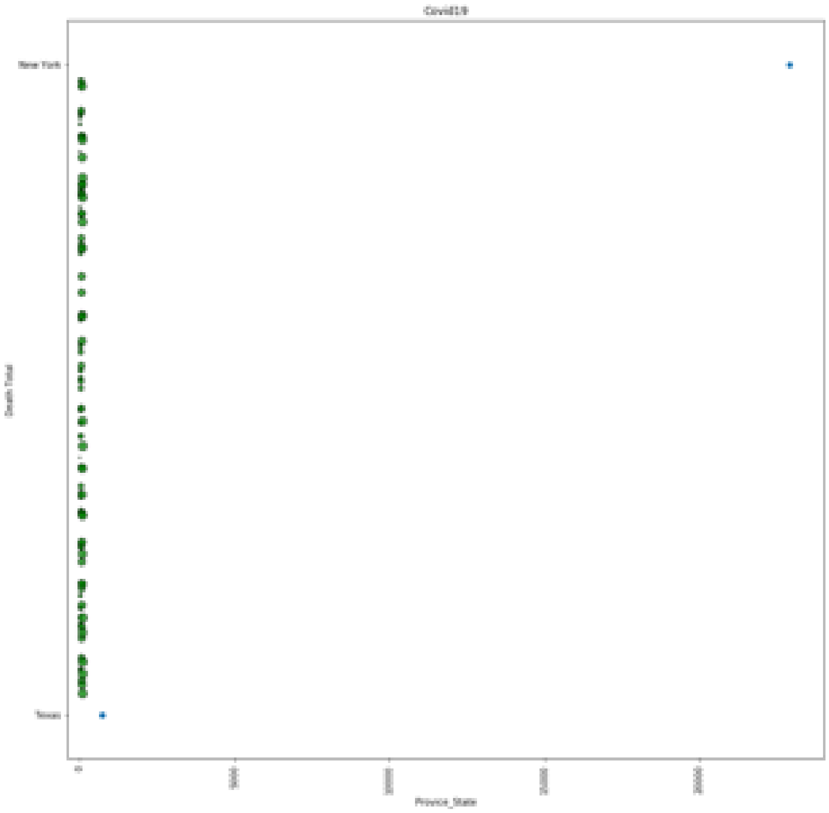
- The below scatter plot shows the total deaths across the U.S. As stated earlier, New York leads the country for total deaths. There are notable patterns throughout the chart as certain states lead in death count over the course of 4 months by a significant margin:



- The below bar chart displays the number of confirmed cases for each state compared to the total death count. New York & Texas have larger population; however, New York is much more densely populated. With the number of cases increasing, Texas does not report the same numbers as other states with large urban areas:

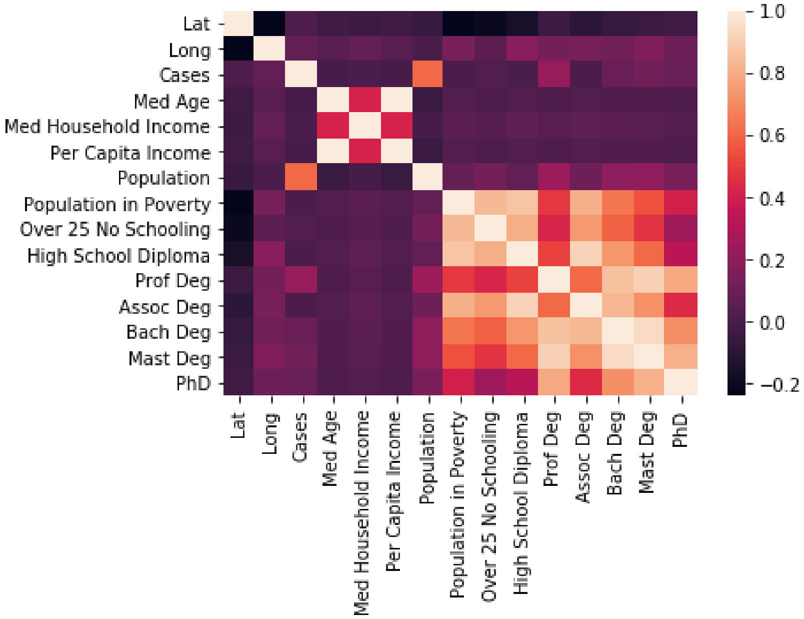


- The below compares New York verse Texas. New York has 22,912 COVID-related deaths, while Texas has 719:

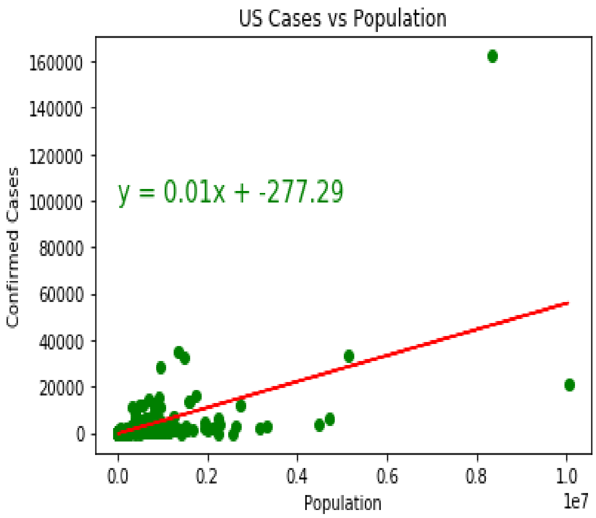


Outbreak as a Function of Population:

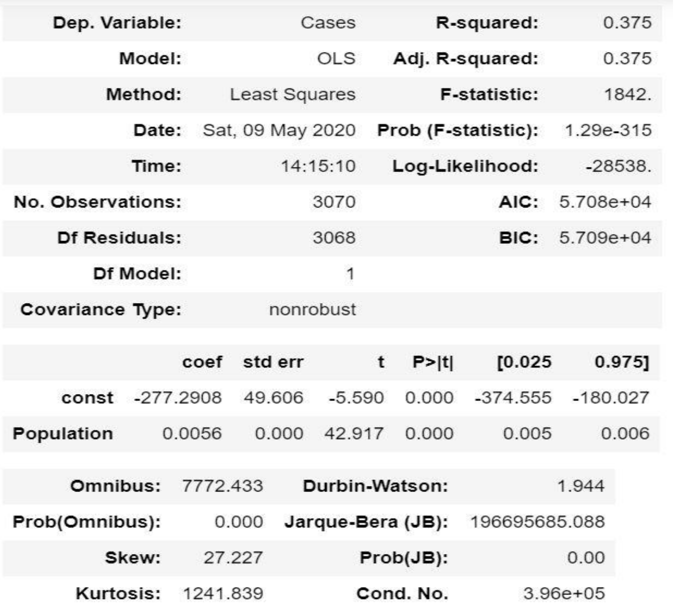
- The correlation heatmap below shows the relationships between zip code demographics and confirmed cases (based on 2010 US Census information):



  - The red line highlights the correlation between total number of confirmed cases vs the above characteristics. This illustrates that confirmed cases is not correlated to education level or income (as expected). It instead appears to mostly be a function of population.
    - This makes sense as the more people there are in a particular zip code, the more human-to-human contact there will be (consequentially increasing the spread).
    - This also shows how income or education level vs total cases might not be as correlated as previously thought.
      - We assumed education/income level would be a factor as higher income/more educated people would travel more and live in areas with international commercial activity.
    - It should be noted that there is a slight correlation between outbreak and "professional degrees." These are defined by the US census as, "degrees beyond a bachelor's degree, including law (e.g. JD, LLB), medical (e.g. MD), veterinary (e.g. DVM) and dental degrees (e.g. DDS)." The advanced degrees category includes all degrees beyond a bachelor's degree (including master's degrees, professional degrees, and doctorate degrees).
      - This was not included in our regression model as it did not improve the measure of fit for our regression line.
- The below is our regression model (noted by the red line) plotted against confirmed cases and population.



- The above accounts for the entire United State. There are outliers in the above data with New York in the top right-most corner of the graph.
- While total population has the strong correlation to the total number of confirmed cases, it piece of the data alone is not strong enough to provide a model to predict the total number of confirmed cases *(as reflected in a low adjusted R-squared)*:



Dep. Variable:	Cases	**R-squared:**	0.375
Model:	OLS	**Adj. R-squared:**	0.375
Method:	Least Squares	**F-statistic:**	1842.
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	1.29e-315
Time:	14:15:10	**Log-Likelihood:**	-28538.
No. Observations:	3070	**AIC:**	5.708e+04
Df Residuals:	3068	**BIC:**	5.709e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-277.2908	49.606	-5.590	0.000	-374.555	-180.027
Population	0.0056	0.000	42.917	0.000	0.005	0.006

Omnibus:	7772.433	**Durbin-Watson:**	1.944
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	196695685.088
Skew:	27.227	**Prob(JB):**	0.00
Kurtosis:	1241.839	**Cond. No.**	3.96e+05

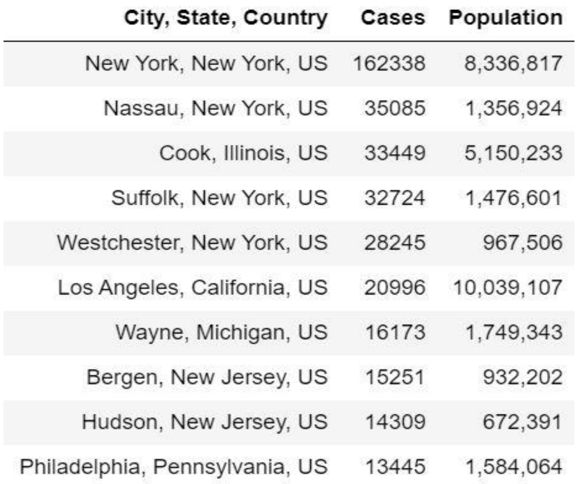
Discovery/Conclusion:

- The outbreak of COVID-19 is largely concentrated in dense urban areas. This is illustrated by the heatmaps earlier. The more people in each respective county (population) the more confirmed cases of COVID-19, with Manhattan, New York leading the country, with roughly 163 - 165k confirmed cases.
  - There is a strong correlation between population and confirmed cases.
  - Additionally, there is a strong correlation and effect when comparing daily differences.
    - Steady decline in daily percentage change as stay-at-home orders are put in place towards the end of March.
- There is a low correlation between confirmed cases and income level or education level. This points to the spread of the virus to be a function of person-to-person contact.
- **Conclusion**:
  - Applying "one size fit all" policy for lifting/easing the shutdowns is not an effective policy tool, since the characteristics of each state varies greatly. The spread coronavirus appears to be largely a function of population/density as person-to-person contact is more significant in these

areas. The lockdowns have proven to be effective tools to slow the spread of the virus. If there is to be any re-opening of the economy, it should be handled at the county level as the spread of outbreak varies greatly across the country.

<u>Appendix</u>:

- **Sources**:
  - COVID Data: https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=COVID-19+Cases.csv
  - Google Maps API: https://maps.googleapis.com/maps/api/geocode/json
  - US Census API
- **Additional Charts**:



City, State, Country	Cases	Population
New York, New York, US	162338	8,336,817
Nassau, New York, US	35085	1,356,924
Cook, Illinois, US	33449	5,150,233
Suffolk, New York, US	32724	1,476,601
Westchester, New York, US	28245	967,506
Los Angeles, California, US	20996	10,039,107
Wayne, Michigan, US	16173	1,749,343
Bergen, New Jersey, US	15251	932,202
Hudson, New Jersey, US	14309	672,391
Philadelphia, Pennsylvania, US	13445	1,584,064

(Top 10 Cities for COVID-19 outbreak)