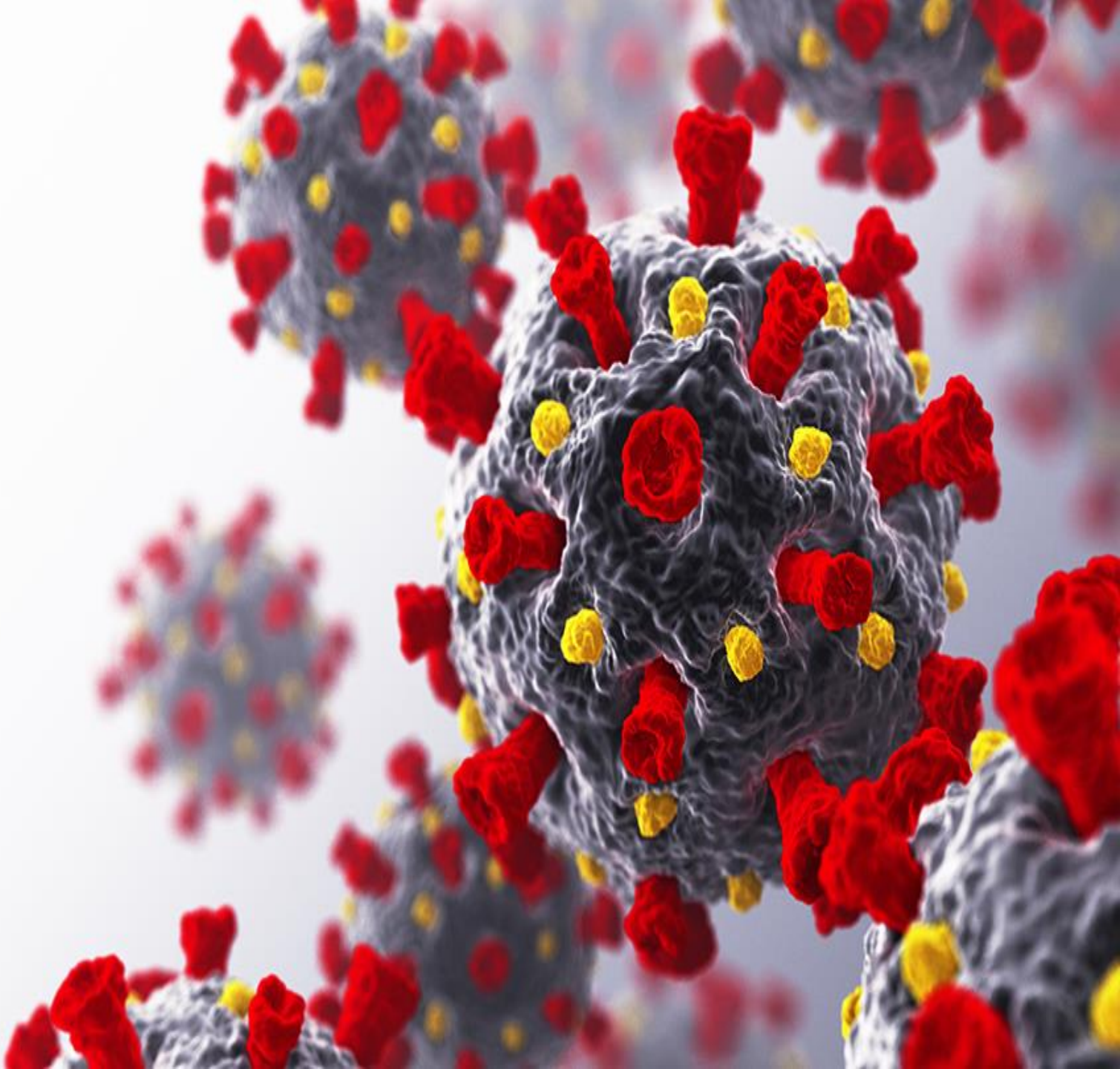
COVID – 19

CORONAVIRUS

Arthur Edwards
Brian Yu
Christion Lankford
Kyle Cielencki



History of Coronavirus (COVID-19)

DEC 31, 2019
WHO says mysterious pneumonia sickening dozens in China

JAN 11, 2020
China reports first death

JAN 20, 2020
Other countries, including U.S., confirmed cases

JAN 30, 2020
The W.H.O declares a global heath emergency

FEB 6, 2020
The U.S. report first death in California

FEB 11, 2020
The W.H.O proposed an official name for the disease : Covid-19

MAR 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)***

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

Data Analysis Process, Strategy, and, Exploration

Data Analysis Process

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

Obstacles

- The primary data source for this project was John Hopkins data (CSV). This dataset was combined with U.S. Census information to break out population demographics by zip code. This presented a problem as the COVID-19 data was in latitude and longitude format. Census data was also broken out by zip code which created an additional layer of complexity. Ultimately, requiring extensive data cleanup and matching coordinates to each zip code
- Consuming the information and how to present the findings.
- Data was aggregated daily rather than the assumed daily totals
- Combining datasets that use different geographic identifiers
- Timing discrepancies between datasets
- Defining data

Discovery and Conclusion

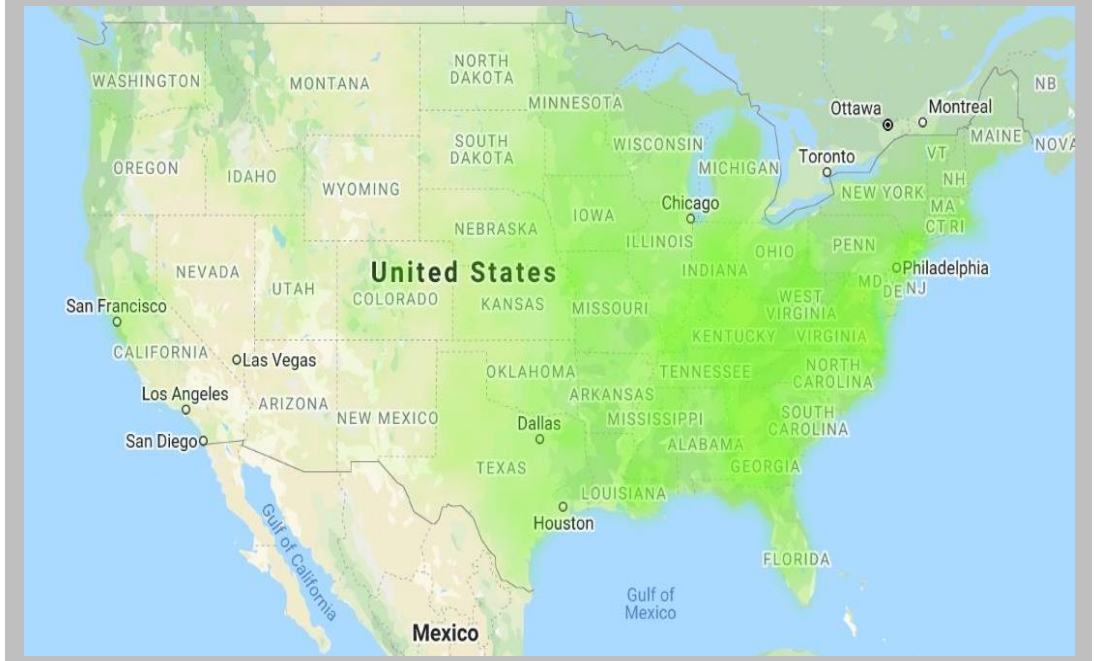
Discovery

- Outbreaks are concentrated in dense urban areas
- Low correlation between confirmed cases and income level
- Strong correlation between population and confirmed cases
- Strong correlation and strong 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
- The more people in each respective county (population) the more confirmed cases of COVID-19
- The leading county in New York is Manhattan, roughly 163 - 165k confirmed cases

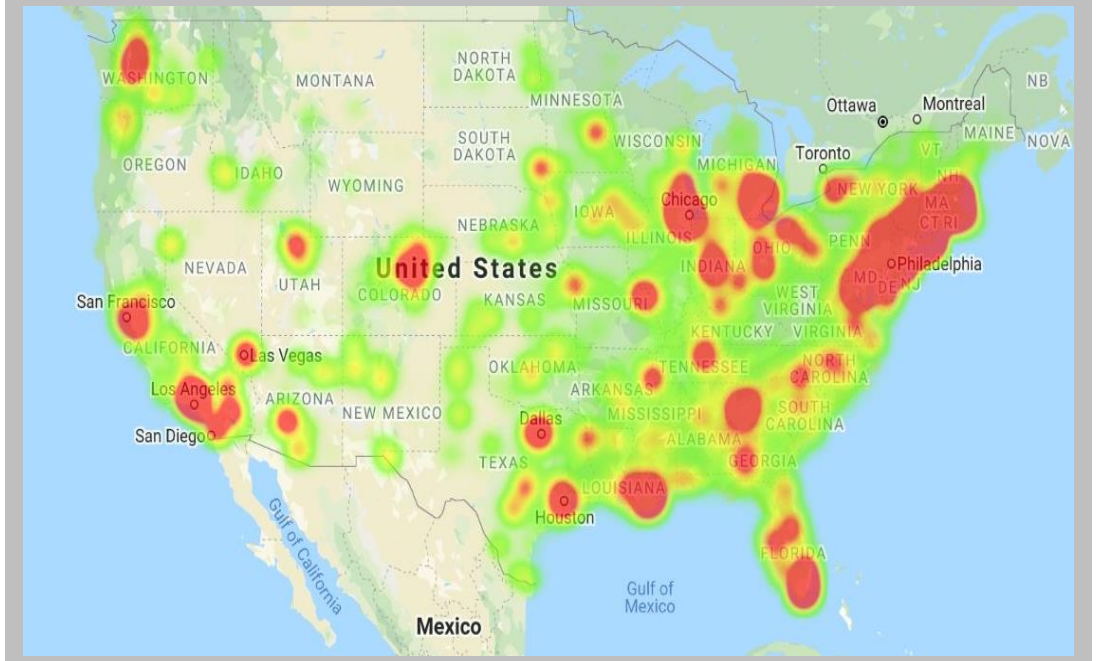
Conclusion: Applying "one size fit all" policy for lifting / easing shutdowns do not work, since the characteristics of each state varies greatly.

COVID – 19 outbreaks are mostly concentrated in urban area, specifically in the Northeast

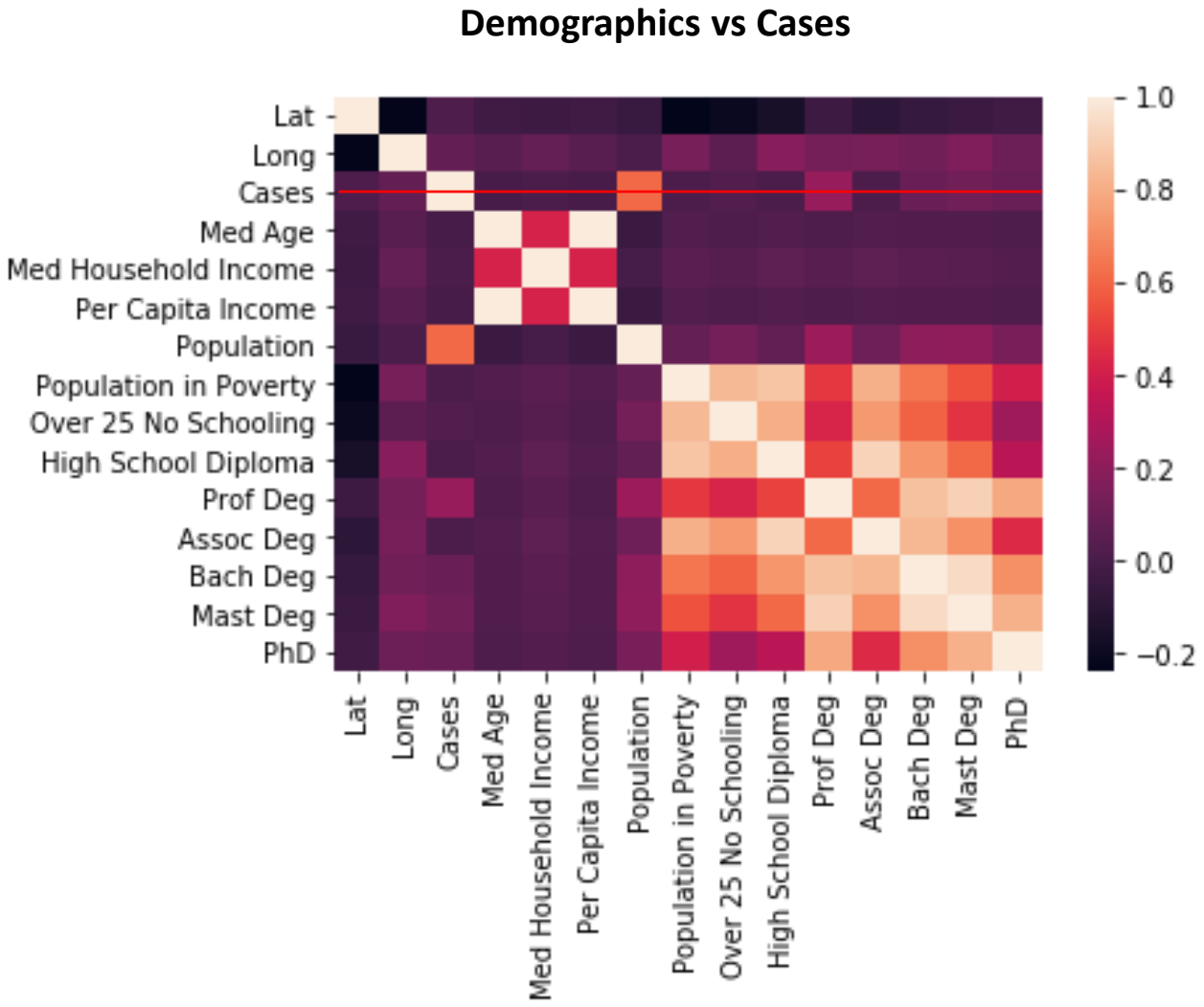
February 28, 2020



April 28, 2020



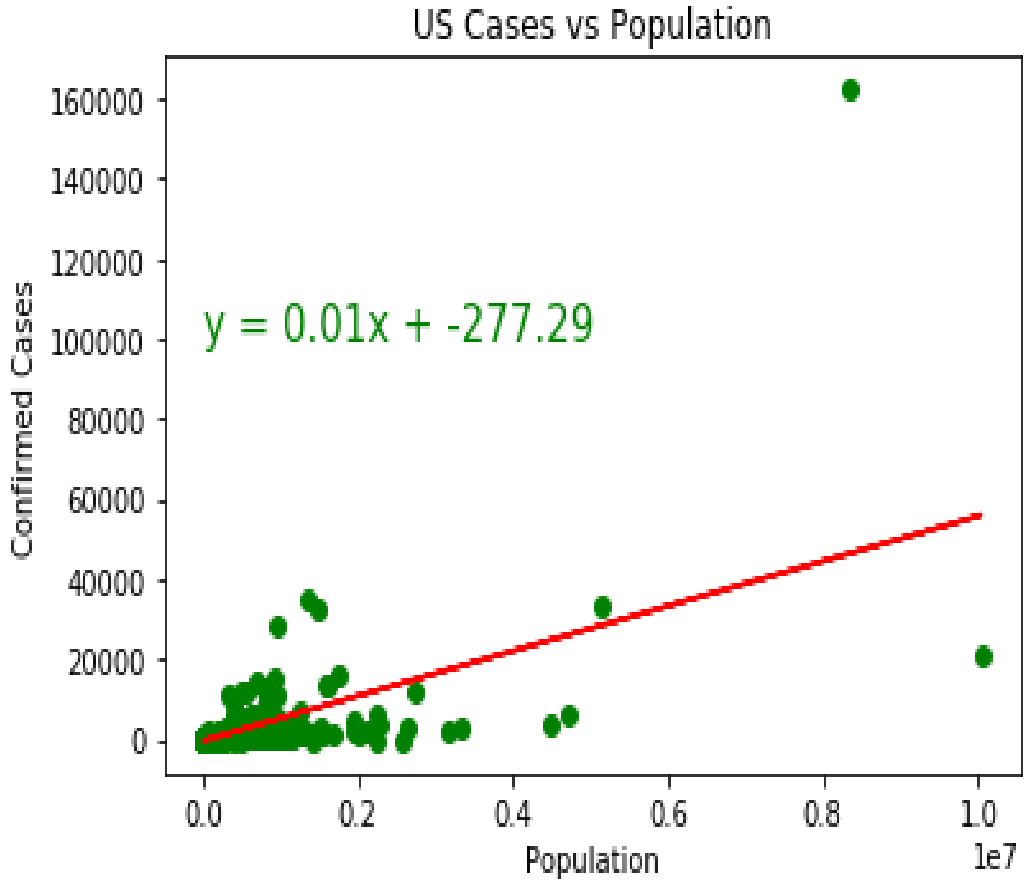
Heatmap confirms strong correlation between population and confirmed cases



The **correlation heatmap** to the left shows the relationships between different characteristics *(based on 2010 US Census information and sorted by zip code)*

- 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, 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 goes also shows how income or education level vs total cases might not be as correlated as previously thought.

Strong population correlation to cases not reflected in Adj. 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

Takeaways and Limitations

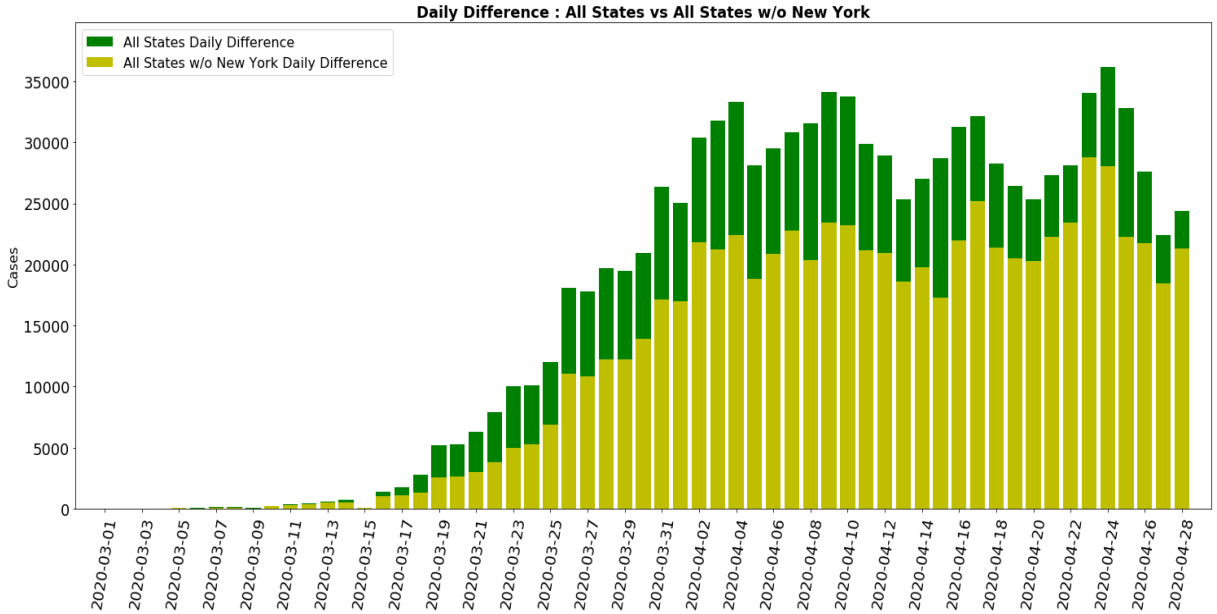
City, State, Country	Cases	Population
New York, New York, US	162338	8,336,817
Nassau, New York, US	35085	1,356,924
Cook, Illinois, US	33449	5,150,233
Suffolk, New York, US	32724	1,476,601
Westchester, New York, US	28245	967,506
Los Angeles, California, US	20996	10,039,107
Wayne, Michigan, US	16173	1,749,343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

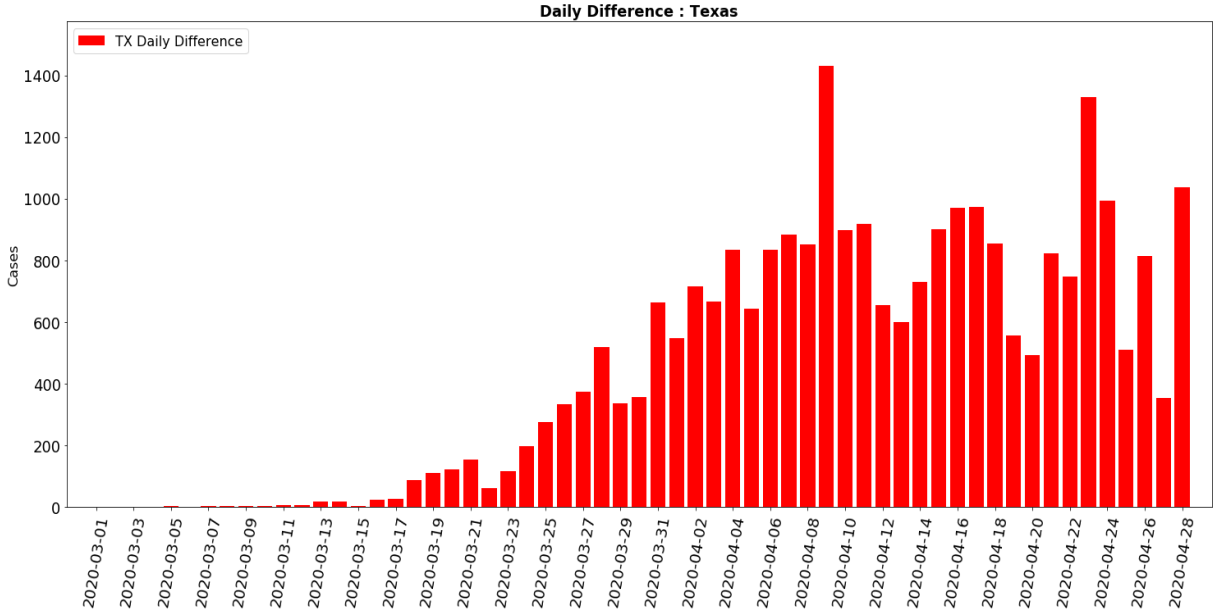
Key Takeaways

Total number of confirmed cases for each zip code appears to be driven by the population.

- This is illustrated by the table to the left.
  - Furthermore, we can conclude that these cities are much more densely populated than other locations in the dataset.

United States and Texas share similar peaks and valleys when comparing daily differences





All States (daily difference)
Max : 36,188
Mean : 17,159

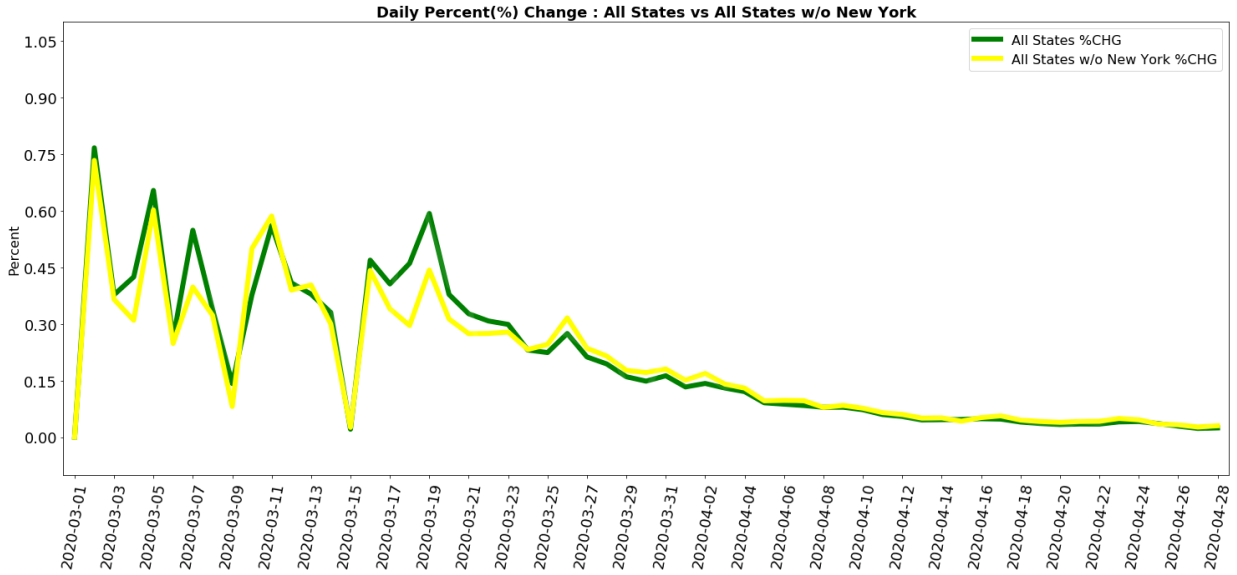
All States w/o New York (daily difference)
Max : 28,782
Mean : 12,157

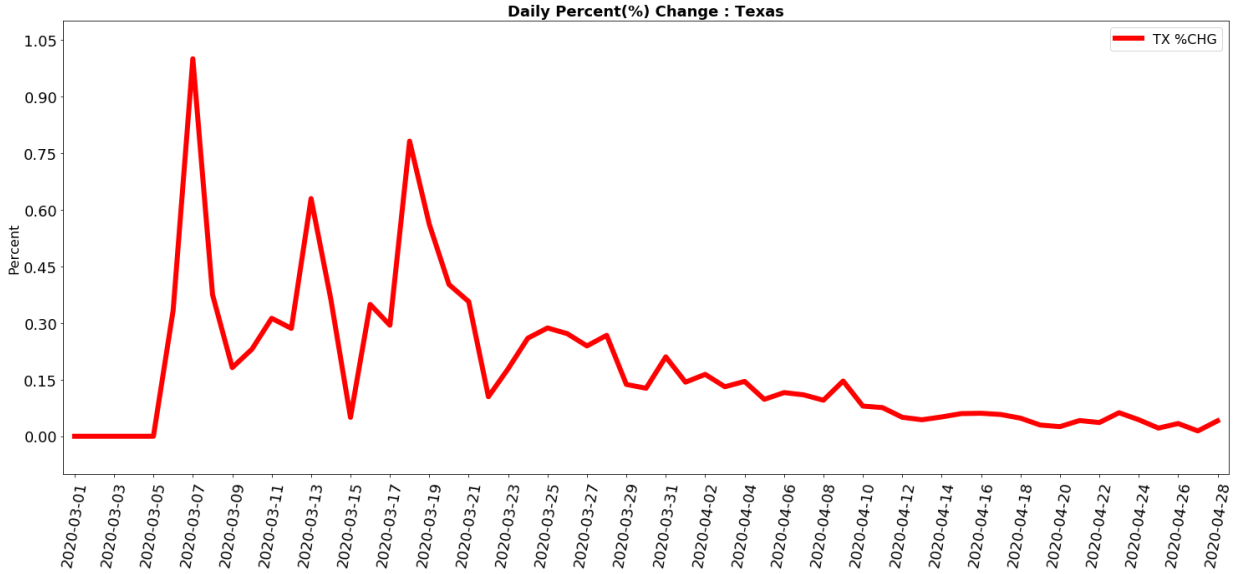
Texas (daily difference)
Max : 1,431
Mean : 467

All States vs TX
Regression and Correlation analysis on daily difference
r – squared = 0.87
correlation = 0.93

All States w/o New York vs TX
Regression and Correlation analysis on daily difference
r – squared = 0.89
correlation = 0.94

Confirmed cases see a steady decline as stay at home orders are put in place at the end of March





All States (daily % change)
Max : 76%
Mean : 21%

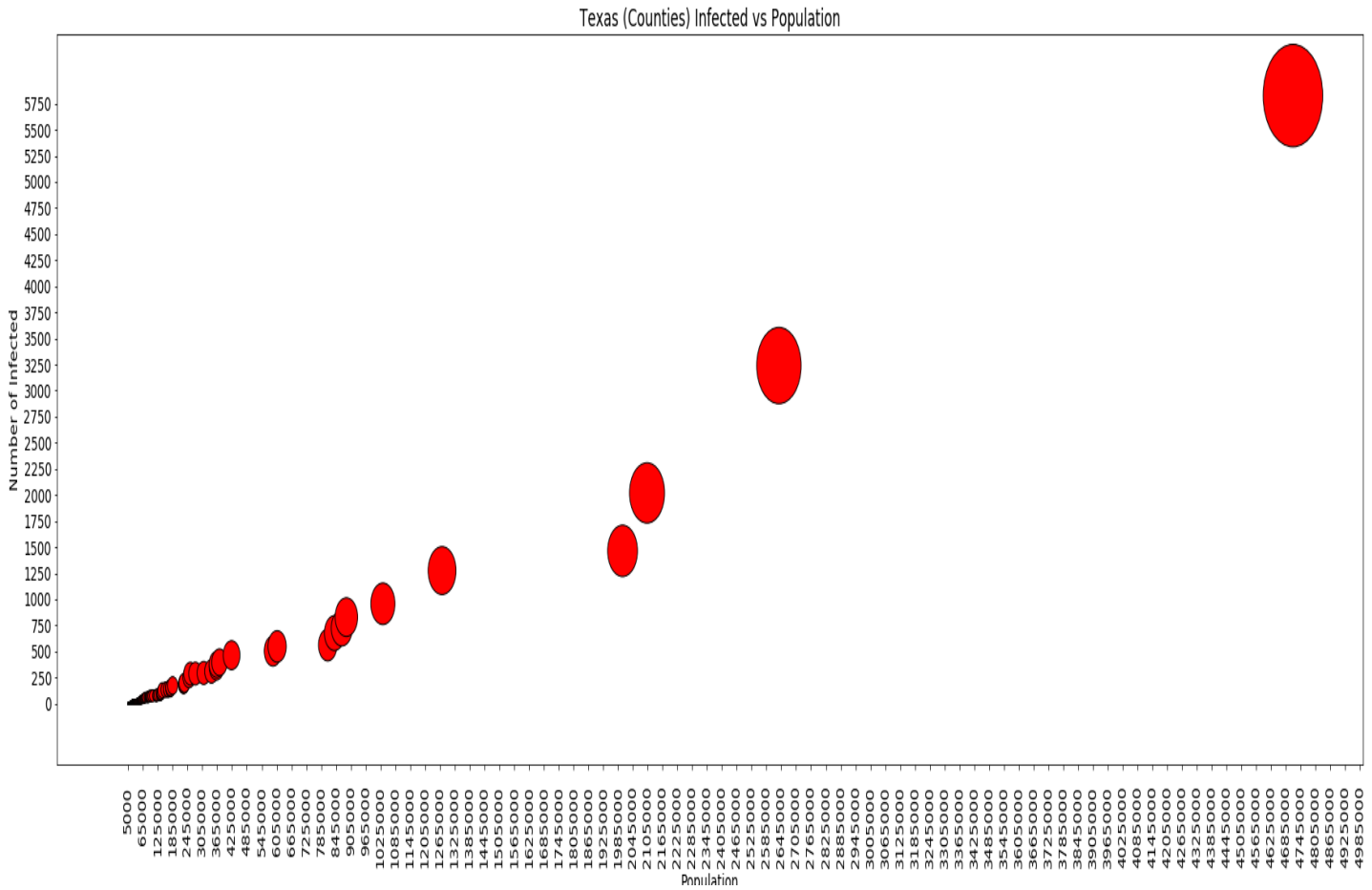
All States w/o New York (daily % change)
Max : 73%
Mean : 20%

Texas (daily % change)
Max : 100%
Mean : 18%

All States vs TX
Regression and Correlation analysis on daily difference
r – squared = 0.32
correlation = 0.56

All States w/o New York vs TX
Regression and Correlation analysis on daily difference
r – squared = 0.23
correlation = 0.48

Number of infected vs population (Texas counties)

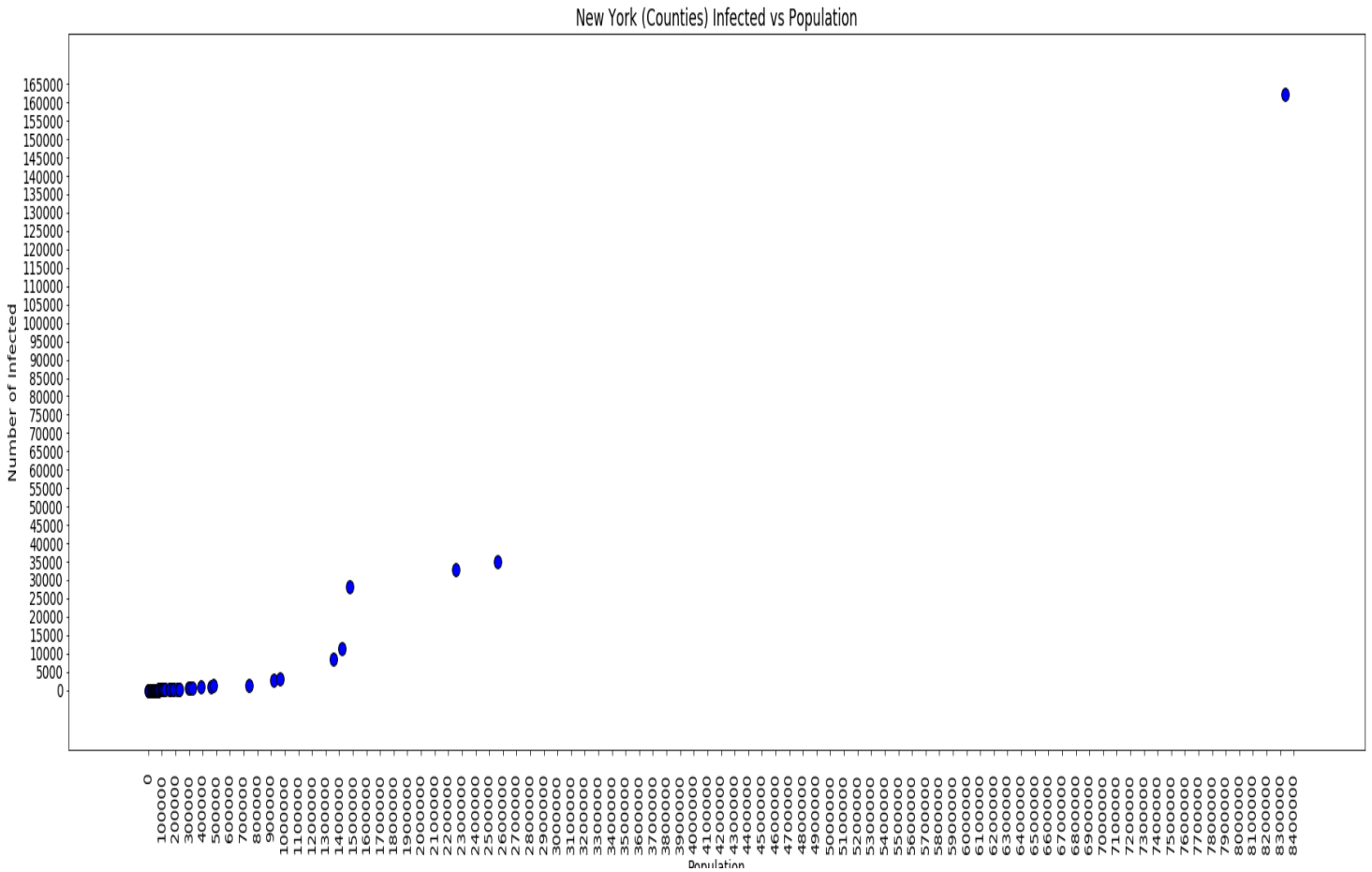


This chart shows the relationship between the number of people infected in Texas (Counties) vs the population.

- The more people in each respective county (population) the more people that are positive for COVID seem to be.

-The leading county in Texas is Harris County roughly with 5,500 to 5,800 confirmed cases.

Number of infected vs population (New York counties)



This chart shows the relationship between the number of people infected in New York (Counties) vs the population.

- The more people in each respective county (population) the more people that are positive for COVID seem to be.
- The leading county in New York is the New York County (Manhattan) roughly with 163-165 thousand confirmed cases.

Daily Difference: Comparison of All States vs All States with New York exclusion

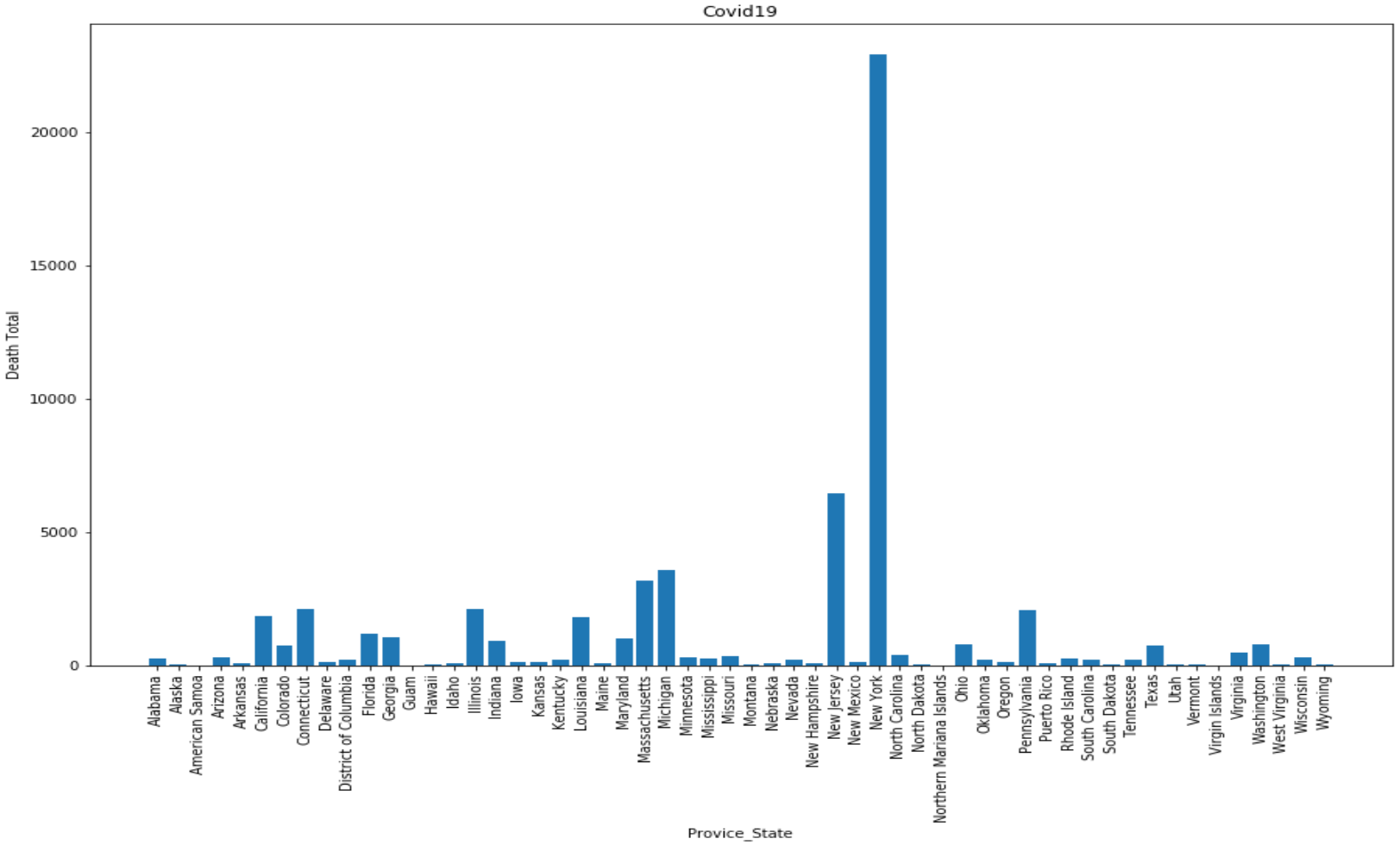
Independent T-test:

Texas and New York Counties – Infected Count vs Population Count

p-value = 1.3210399000455797e-05

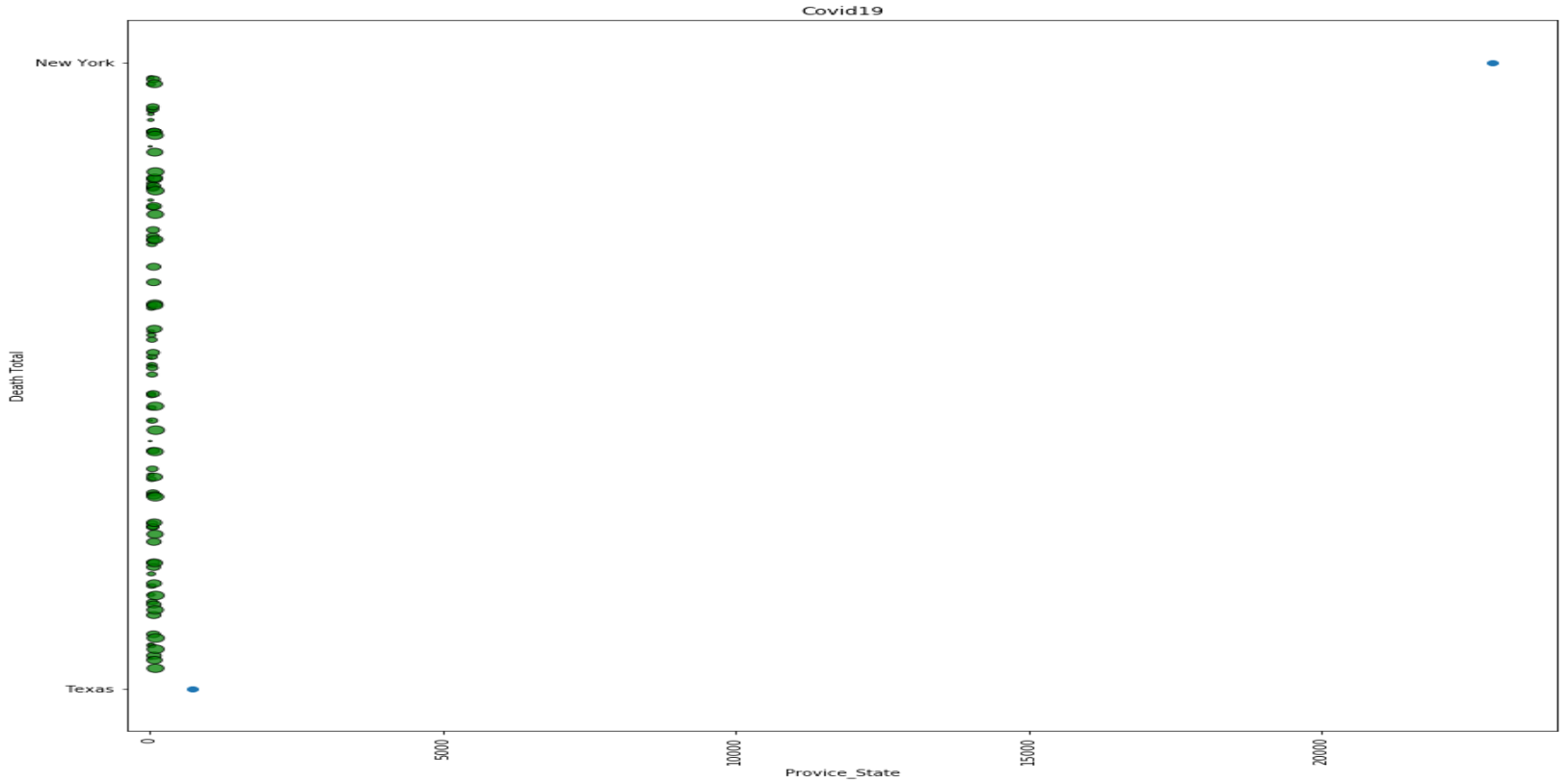
The p-value is < 0.05. This proves that the correlation between the population of each county and the amount of infected residents is strong and significant.

New York charts a considerable difference when compared to other states in the U.S.



- This bar chart display accurate numbers per state and the all Death Total.
- The difference between New York and other states since the density is much higher when making the comparison

Variance between New York and Texas are substantial



Compared on the scatter chart between New York & Texas. New York has 22,912 death from Covid19 while Texas has 719 during the outbreak. The Death Total is a huge difference from New York to Texas based off the number of people that continue to stay home.

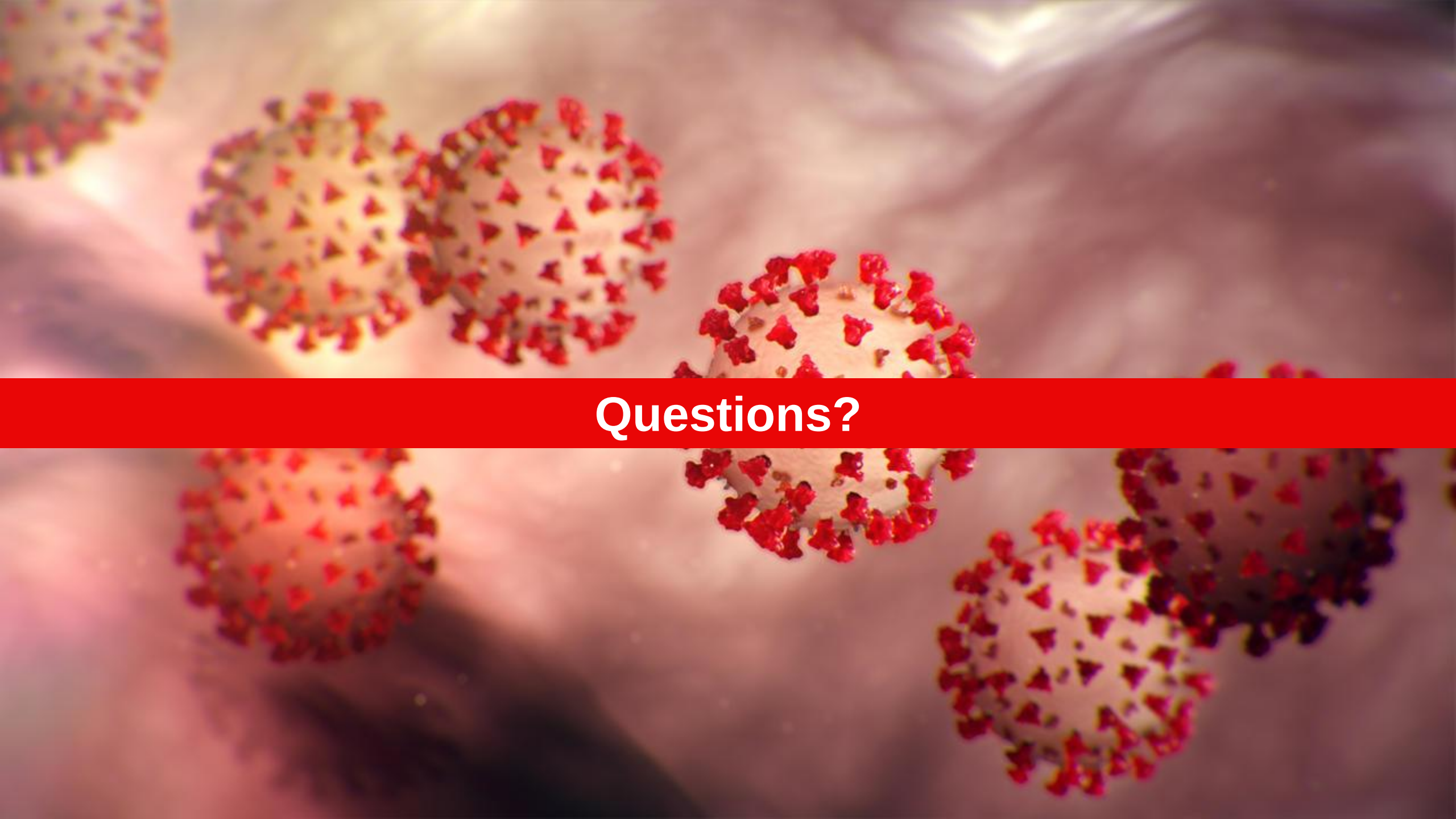
Post Mortem

Challenges

- Combining datasets that use different geographic identifiers
- Formatting
- Collaborating outside of class
- Working virtually
- Time
- Troubleshooting pandas

Additional Questions and Further Research

- Compare states that have lifted the stay at home ban vs states that have the ban on-going
- Further investigate population density hypothesis
- The affects of COVID-19 and ethnicity
- Forecasting future cases and deaths


Questions?