Introduction to Data Science
Course Project – Final Report

Anna Godin and Katherine Thai

1 Data Collection

Originally, we were planning to collect data from the Rutgers bus API, and analyze the accuracy of the predicted wait times against a Google Maps API. However, due to the outbreak and the suspension of in-person instruction, we were not able to collect sufficient data to make any worthwhile conclusions. Therefore, we switched focus to the most prevalent issue, the novel coronavirus

The source of our data is from a Github repository that continuously updates COVID-19 data in the US. It has really detailed data, including cases for each county in the US. We will be focusing on harder hit areas of the U.S., as well as predict trajectories of states that have yet to have a large outbreak.

2 Data Format Description

Our data comes from the following Github repository: https://github.com/nytimes/covid-19-data. There are two files that we are importings and creating data frames from: *us-counties.csv* and *us-states.csv*. *us-counties.csv* has the following attributes: date, county, state, fips, cases, deaths. *us-states.csv* has the same attributes, but excludes 'county'. This repo is updated daily with new data for the previous day. The most relevant attributes are date, cases, deaths county, and the state. We are able to draw many conclusions from this data, from examining the rate of increase in cases to the distribution of all cases by state or by county in each state.

3 Descriptive Statistics

Stats for March 14, 2020

Cases:
Min: 1 (Alaska)
Max: 610 (New York)
Mean: 55.73076923076923
Std Dev: 125.37334291276663

Deaths:
Min: 0 (Alabama)
Max: 40 (Washington)
Mean: 1.1538461538461537
Std Dev: 5.570878817511352

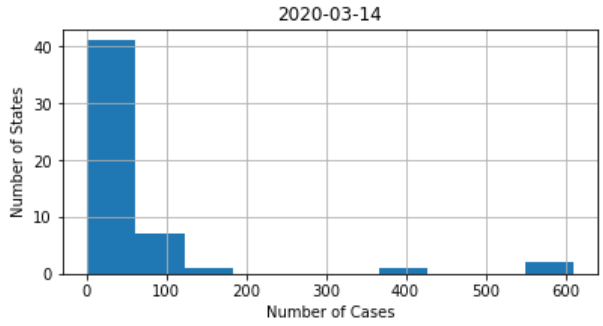
Stats for April 8, 2020

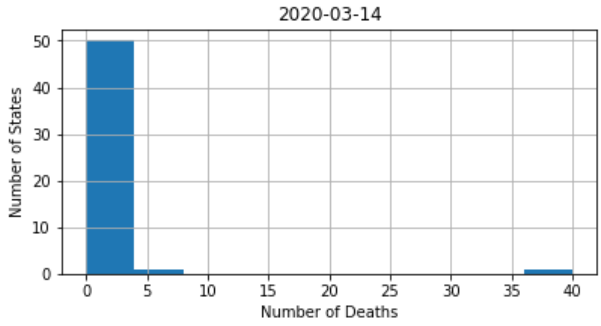
Cases:
Min: 224 (Alaska)
Max: 149401 (New York)
Mean: 8247.192307692309
Std Dev: 21567.39906509083

Deaths:
Min: 0 (Wyoming)
Max: 6268 (New York)
Mean: 285.0
Std Dev: 888.2489714568306

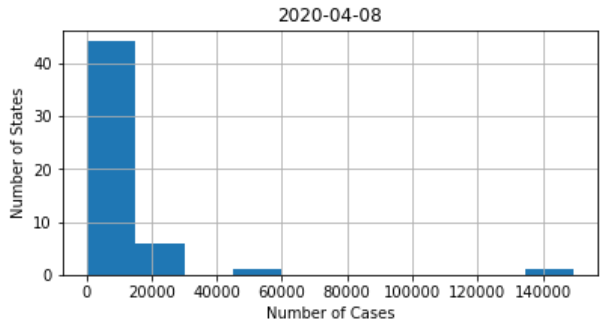
Histograms

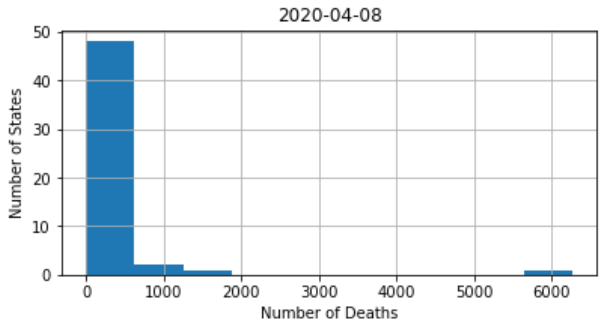
March 14, 2020





April 8, 2020



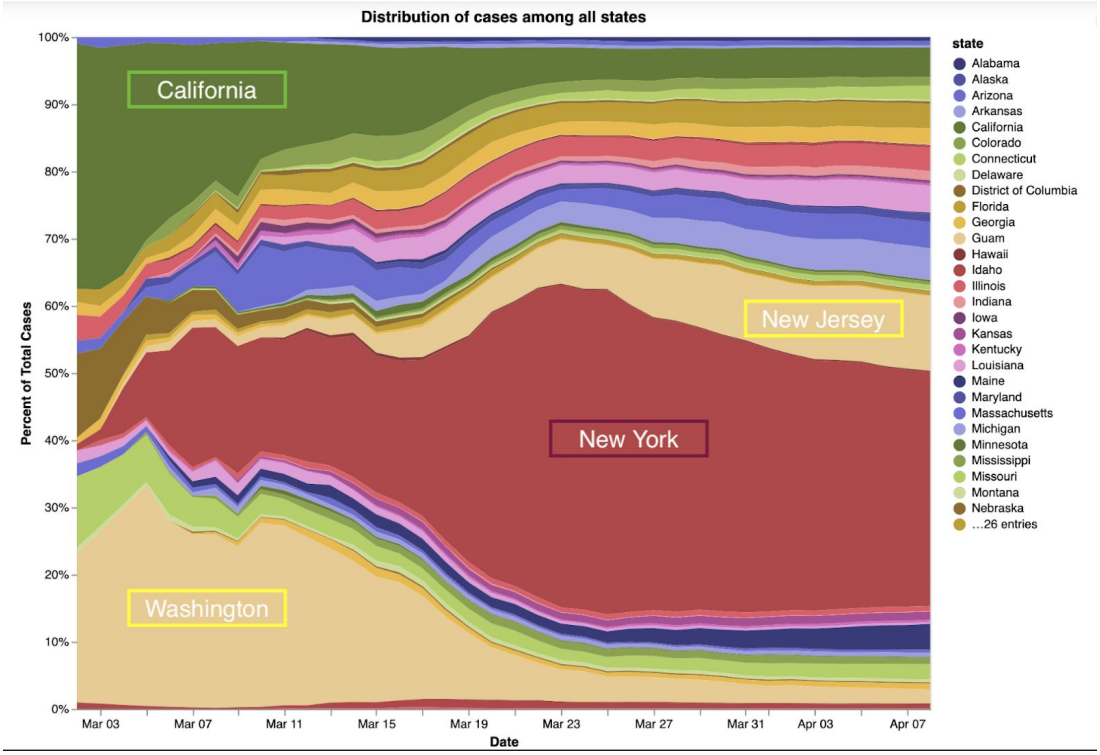


Range of dates

Variable - first date reported is 2020-01-22, and the latest date reported is updated daily, since the nature of this data changes frequently. We have been pulling new data every day to see how our charts change, and make new conclusions/adjust existing observations as needed

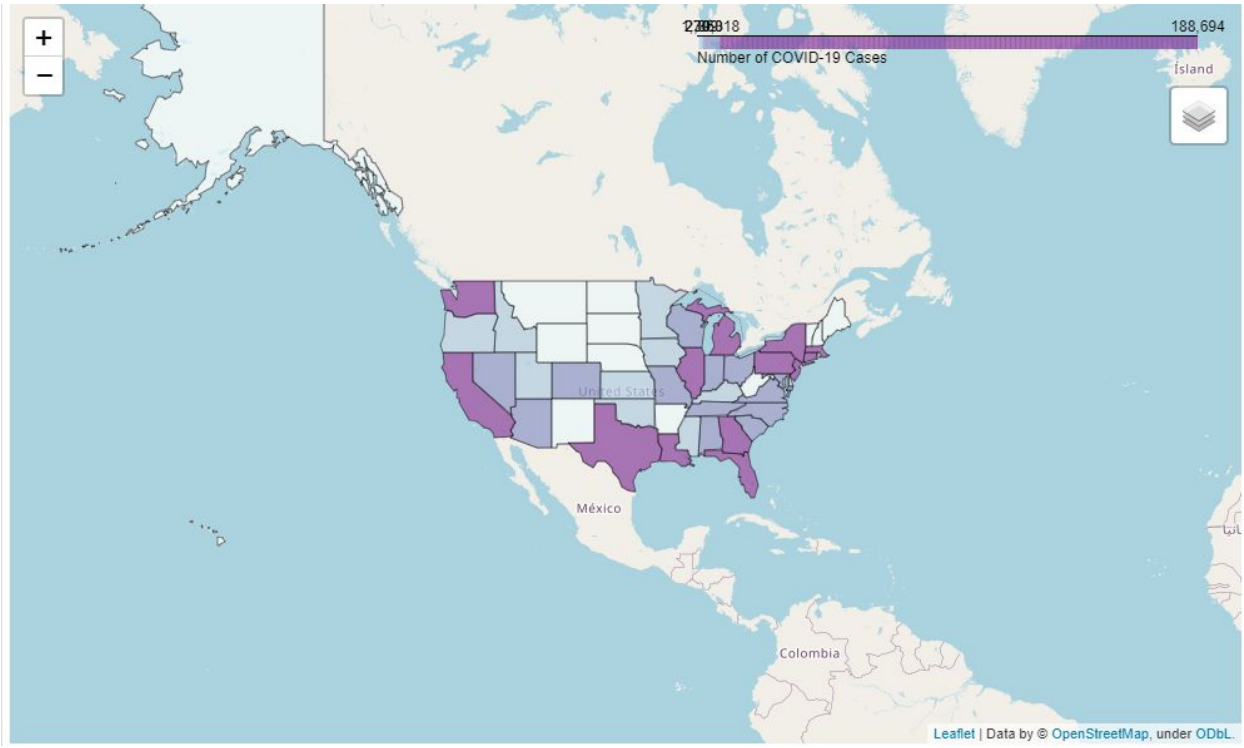
4 Data Analysis, Visualization, and Insights

Visualization of the distribution of cases plotted as a function of time



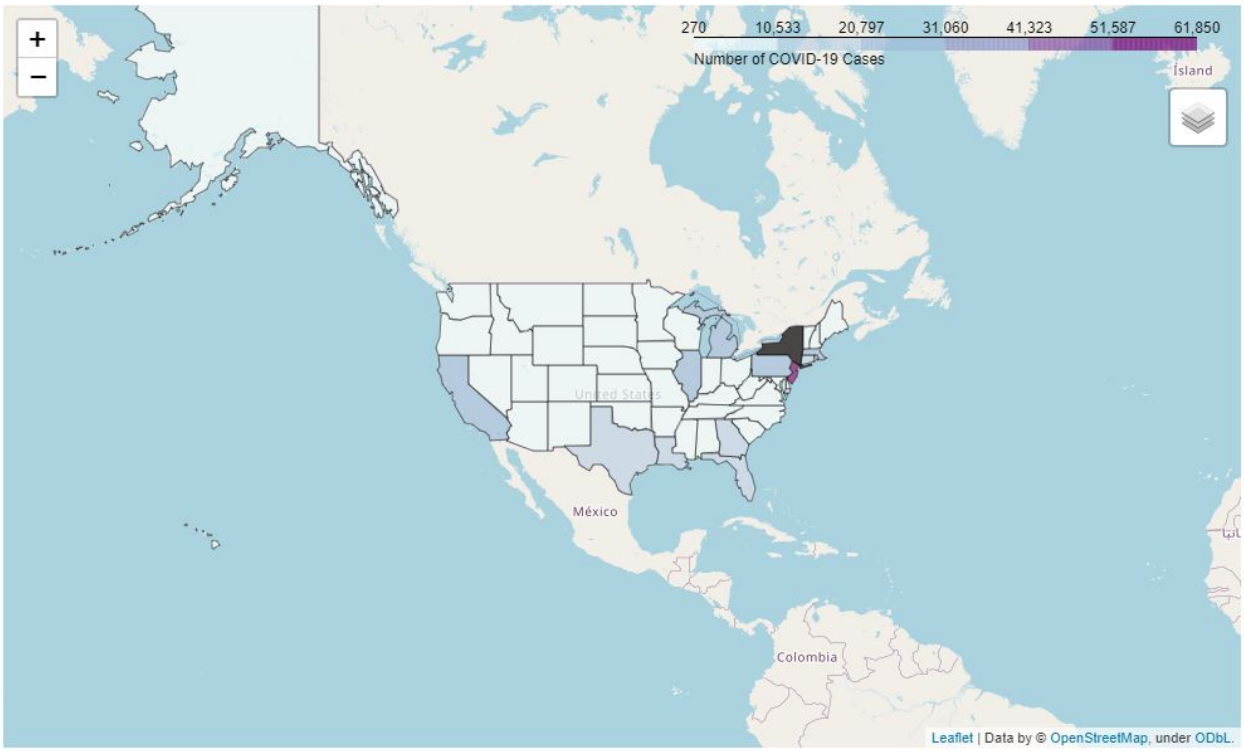
The above is a normalized area chart of the distribution of cases plotted as a function of time. This graph visualizes which states were the initial outbreak epicenters and how New York quickly exploded as the new epicenter of the outbreak in the US. We can see that Washington, where the first case was recorded, had ~25% of the country's cases on March 2nd, and California had 40% on the same date. As March went by, New York began to take up more than half of the country's cases around March 24th. We can also see how New Jersey's cases steadily grew with time, to become the state with the 2nd most cases in the country.

Choropleth map of cases across America with/without NY



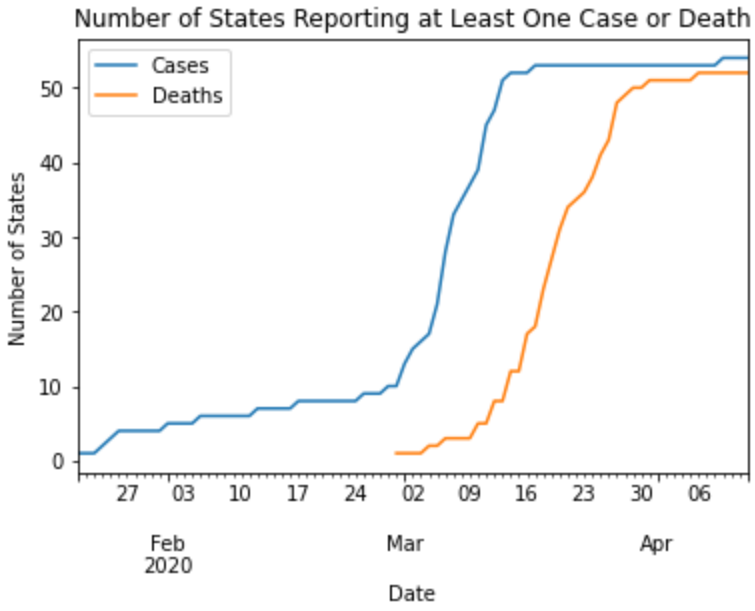
In the map above, each state is in one of four bins depending on the number of cases it has reported. These bins are defined by quartiles with the following ranges: 270 - 1308, 1309 - 2859, 2860 - 9317, 9318 - 188694. This method of choosing bins attempts to place the same number of states into each bin, but with New York and New Jersey having tens of thousands more cases than the other states, the last bin's range is enormous.

New York has been aggressively testing and therefore has over 100,000 more cases than the state with the second highest number of cases. Let's disregard New York to get a better look at how the other states compare.



Note the difference between putting the values into 6 evenly spaced bins versus putting them into bins based on quartiles. Because states like NY and NJ have such high relative case values (and have more deaths in general because the outbreak is so bad there), it's hard to pick bins that make both the visualization (the actual map) AND the legend valuable.

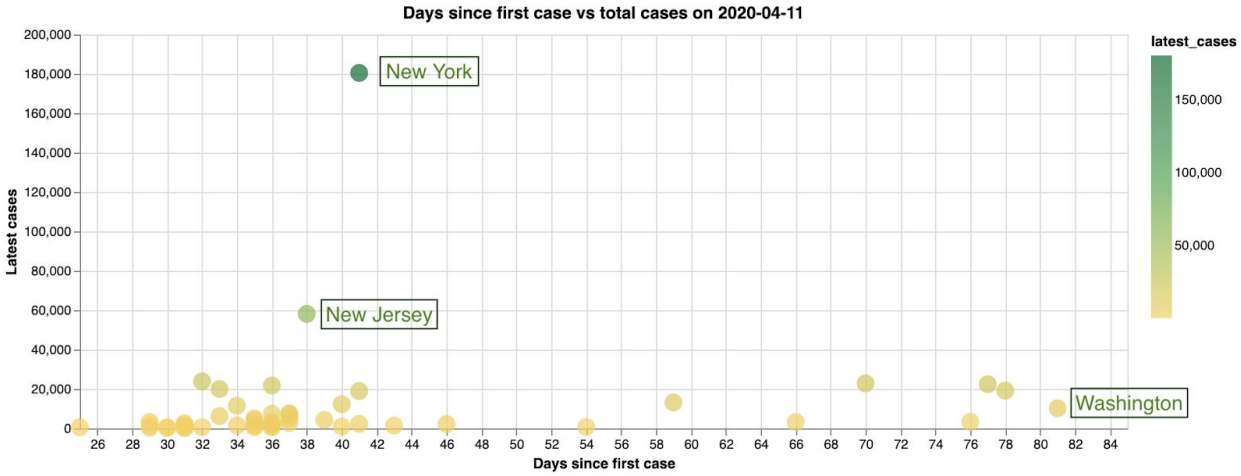
Examining the trend of when states began reporting cases in the US



This plot shows when states began reporting cases, beginning with the first recorded case in Washington on January 22, 2020. We can see that the first death occurred in late February (February 29 to be exact). By mid-March, all states were reporting at least once case of coronavirus. We can also see from this plot that not every state that has reported a case has reported a death yet.

Days since first reported case for each state, plotted against total cases

The following chart looks at how each states' cases have grown since the first reported case. The chart shows a few evident outliers in the data, visualizing how fast New York and New Jersey spiked in cases despite only reporting their first cases 41 and 38 days ago, respectively. On the other hand, Washington was the first to report a case in the country, but only has a small fraction of the cases that New York and New Jersey do. This shows that Washington and other early reporters of COVID-19 cases may have effectively slowed the spread of the virus.



5 Result of machine learning experiments

Our motive with our machine learning experiment was to determine if we can predict the number of cases for a county based on the number of days since the first case and the population density of that county. This is based on the belief that the more dense a county is, the higher the rate of spread will be, and therefore a higher number of cases vs a county that isn't as dense with people. Preliminary findings of this exploration showed that there is not a sufficient correlation between population density and number of cases. The MSE was quite high, so we decided to forgo further exploration of this, and focused on other aspects of our data that a ML model could more accurately predict.

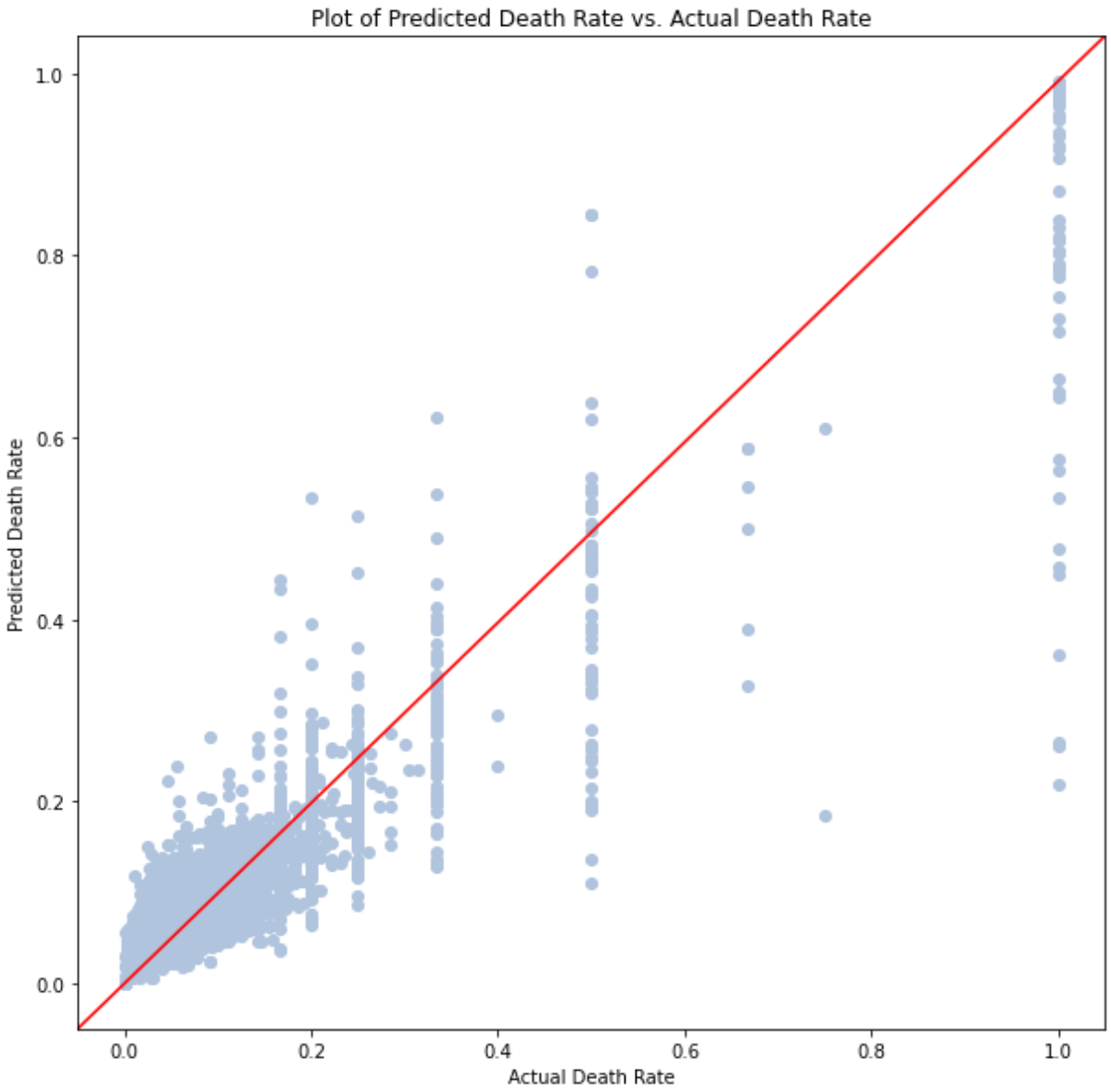
Another machine learning exploration we did was try to predict the death rate in a county based on the number of deaths, population density, population size, and days since the first case in that county. Our findings showed that there is indeed a high correlation between deaths,

population size, population density, days since the first case, and the case death rate for each county. The XGBoost model performed the best. After running the experiment and training/testing the model, the MSE was found to be 0.003, or 0.3% and the RMSE was 0.054, or 5.4%. While we tried multiple regression and SVR, we found that neither of these methods was predicting a death rate over 20%, regardless of the input. Please note that the values below may be different than those in the jupyter notebook because the jupyter notebook takes into account the new data that is added daily.

Method	**MSE**	**RMSE**
XGBoost	0.003	0.054
Support Vector Regression*	0.035	0.187
Multiple Regression	0.006	0.079

*Using an rbf kernel, C = 1, epsilon = 0.2. There were too many parameters for us to be able to explore SVR thoroughly. It's possible that some combination of parameters would yield an excellent model, but we were too limited in time to discover it.

Another way to look at the prediction performance is to consider the plot of the predicted vs the actual values. When values fall on the red line, x = y, it means that the prediction equals the actual value. We can see that for a low death rate (the area of the plot in the bottom left), the model predicts much better than it does as the death rate rises. Note the points on the plot at the far right with an actual value of 1.0. This was a problem inherent to our data: some counties did not report a case until they had a death, so the death rate was 100%.



The main challenge with performing machine learning on this dataset, is that there is not enough data to come to concrete conclusions. The data set we are using gets updated every 24 hours, to update case counts and deaths for each county in the country. This type of data is the first of its kind, and changing daily, especially the rate of new cases/deaths. This means that a model that works well for a set of data one day, may not work when new data is added, due to rapidly changing numbers and environments.

6 Related Work

Evidently, there is a lot of related work being done regarding the coronavirus pandemic at the moment. Many studies have been focusing on CT image analysis and classification to create tools for detection, quantification, and tracking of COVID-19. This research can possibly be a

game changer when it comes to diagnosing a patient with coronavirus. This data science approach is focused on a clinical understanding of the virus, and a lot of work is being put in to aid the fight in coronavirus. There is another study that focuses on using a neural network to predict the risk category of a country. It is proposed that this tool can be used to predict future outbreaks and propose preventative steps early on.

While these approaches involve analyzing very specific data that is not readily available online, we took a more broad approach, and analyzed the number of cases/deaths for states and their counties. We mainly focused on visualizing this data for New York and New Jersey, because these two states make up a large portion of the country's cases. The goal of our exploration was to look at what factors are at play when trying to analyze why New York and New Jersey have such high case counts. We explore the possibility of population density being a cause for this, due to how densely populated these two states are. We also explore what could cause death rates to be higher in certain counties compared to others. The main features we looked at to be possible causes are population density, population size, and days since the first case, to see if these features affect the case/death counts in each county.

7 Conclusion

The main insight that we gained was how much the distribution of our data affected our analysis and the machine learning experiments. We suspect that the machine learning did not work particularly well because of the data points with a 100% death rate. The counties that had a 100% death rate had no specific trend in population or population density; they were often just counties in which the only case of coronavirus was also the only death. The way that states approached testing also meant that our data wasn't exactly representative of how coronavirus is spreading and interacting with the population. Some states, like New York, aggressively tested and saw a huge increase in cases early on. Some states, like New Jersey, had a county-by-county approach to testing, so certain counties were being tested more than others, and those counties therefore had higher case counts. The way that deaths and cases were reported by state also differed: some states had a backlog of tests whose results they would report sometimes weeks after the tests had been conducted and the patients had already recovered.

Overall, XGBoost worked the best, and linear regression and SVR did not work at all. If we had more time, we would explore more SVR parameters and other machine learning methods, possibly neural networks and deep learning. Unfortunately, this was not our original project plan, but the pandemic forced us to come up with an entirely new project halfway through the semester.

8 Acknowledgements

We used several Python libraries to conduct our research: matplotlib, matplotlib.pyplot, pandas, altair, Numpy, Folium, and XGBoost/sklearn for our machine learning portion.

We obtained our data from the following sources:

- https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv
- https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv
- https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-total.html
- https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html#LND

To gain familiarity with different machine learning methods and python, we consulted the following tutorials:

- https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f
- https://www.datatechnotes.com/2019/06/regression-example-with-xgbregressor-in.html