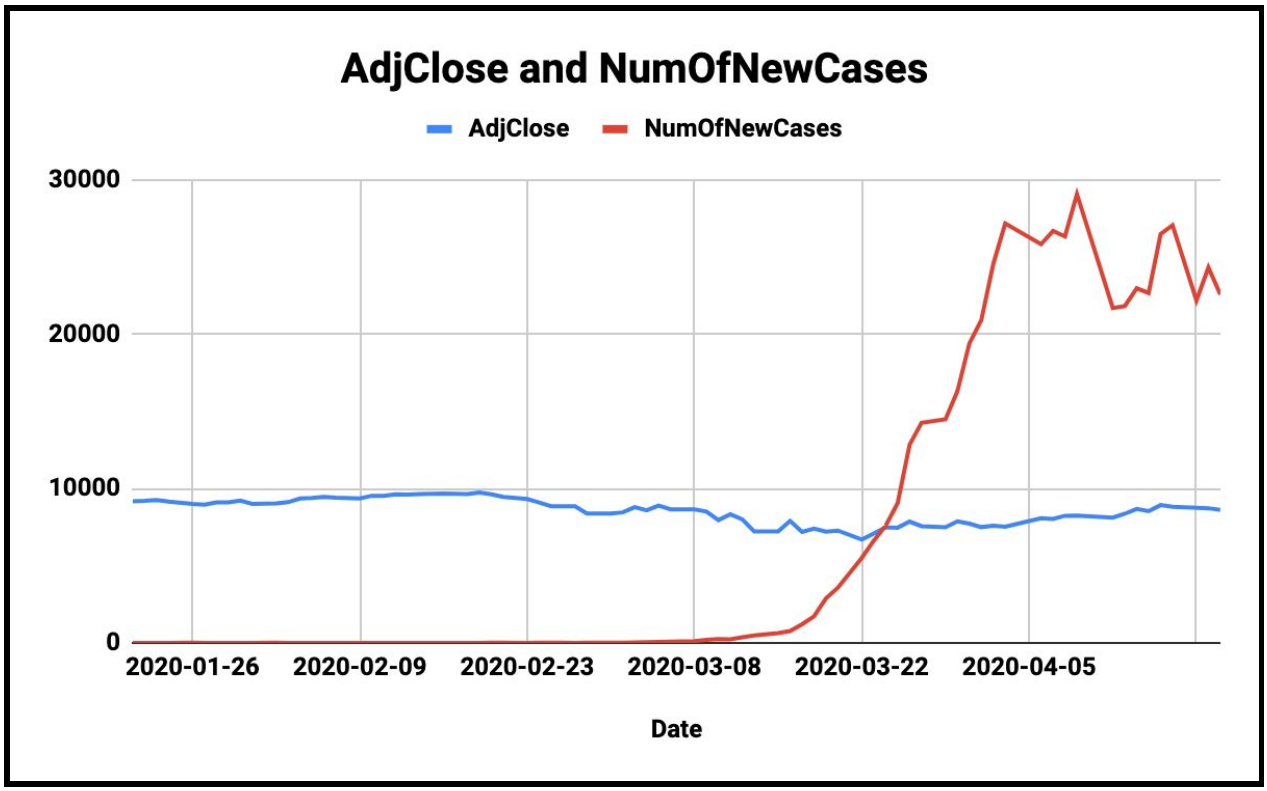
REAL Team Six PM5

External Data Source #1:
https://www.kaggle.com/nickbaynham/trading-indexes-apr-2019-to-apr-2020?select=nasdaq.csv

What information does this data contain: The data contains the opening, closing, high, low, and adjusted closing prices of the NASDAQ every day the stock market was open from 4/21/19 – 4/20/20.

How do we plan on using this data: We will compare the adjusted closing price of the NASDAQ to the number of new Covid-19 cases in the United States on that day to see how the two values relate to one another. We will also compare the adjusted close price value to the number of new Covid-19 deaths in the US on that day. We will view this data from 1/21/20 – 4/20/20 because that is when our data overlaps.

1.
Hypothesis:
As the number of new Covid-19 cases increased in the United States, the lower the NASDAQ went as more new cases would lead to more uncertainty over how the US is handling the pandemic resulting in more uncertainty over the US economy.
Chart/Graph:



Summary:

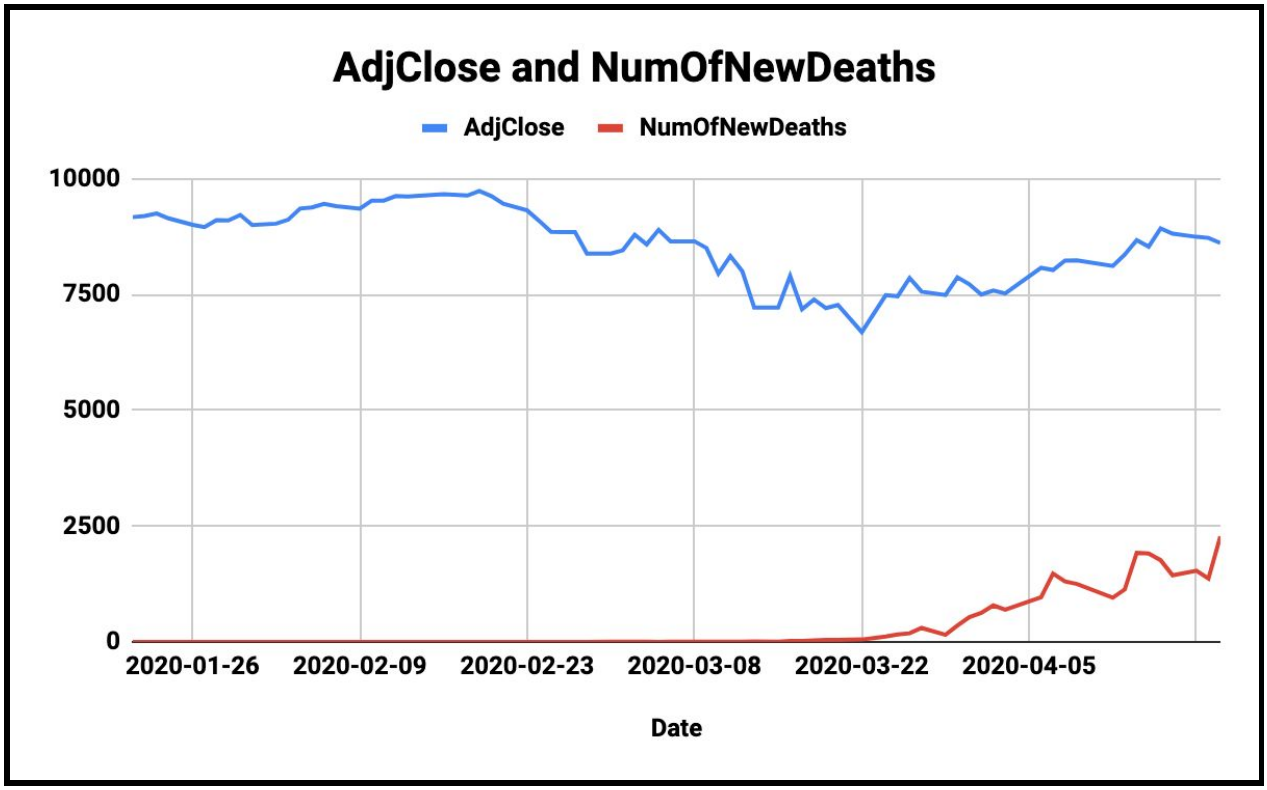
The data invalidates our hypothesis because there doesn't seem to be a direct relationship between a high number of new cases and the stock market going down. When the rate of change of the number of new cases was highest, that is when the stock market took the biggest dip. This relates to our application in that we cannot create a direct relationship between the stock market and the number of new Covid-19 cases which we are tracking. However, there might be a relationship between the rate of change of number of new cases and the closing price of the stock market. This would be very interesting to track if we had more data for May, June, and July because those months saw a period of lull and then recently another spike in the number of new cases as well as to possibly track the stock market price versus the rate of new cases instead of just the number of new cases per day.

2.

Hypothesis

As the number of new Covid-19 deaths increased in the United States, the lower the NASDAQ went as more new deaths would lead to more uncertainty over how the US is handling the pandemic resulting in more uncertainty over the US economy.

Chart/Graph



Summary:

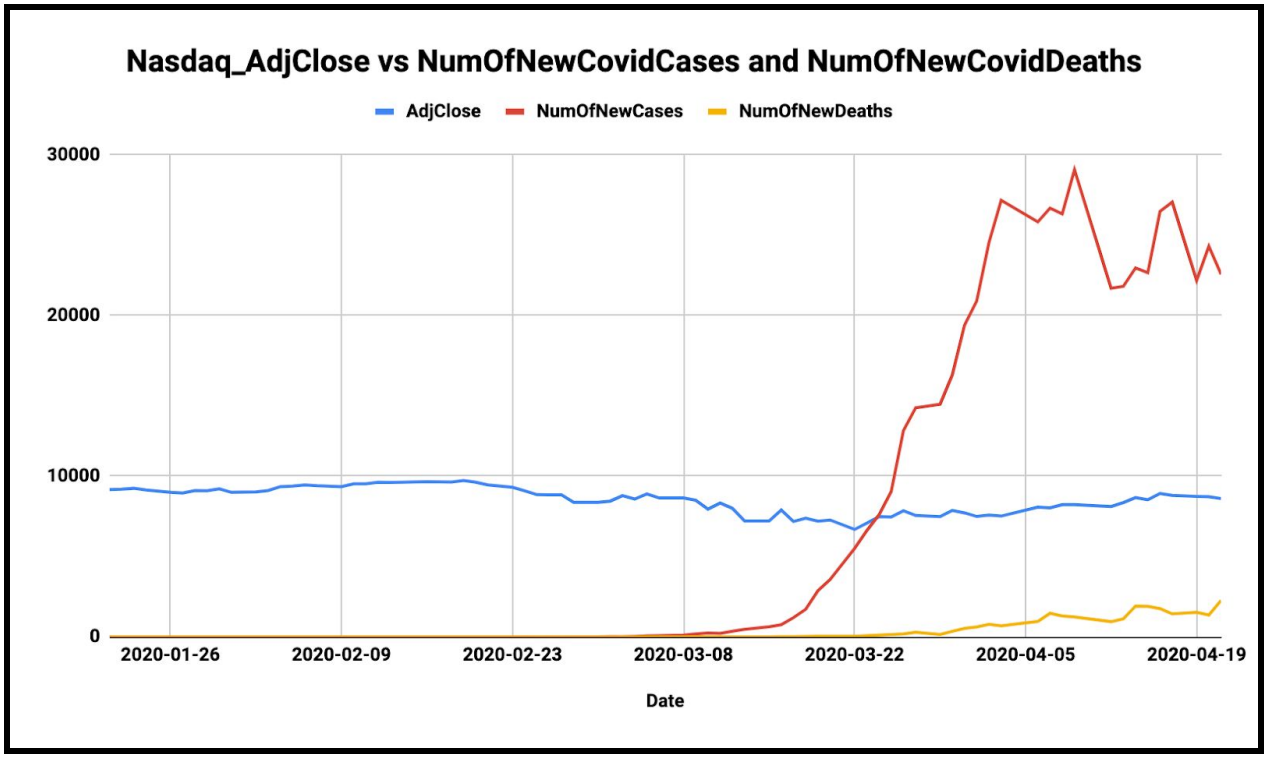
The data invalidates our hypothesis because the stock market started to decrease before there was a rise in the number of new deaths. Deaths trail the number of new cases so the rate of change in the number of deaths will always lag behind the rate of change of new cases. The death rate is not predictive of the stock market price, so we can't necessarily use our application to track the death rate to know when to buy and sell. It would be interesting to see this data with the recent spikes in the number of new deaths and see how the stock market has responded to that. We can check whether the stock market took a dive again before the rate of new deaths increased.

3.

Hypothesis:

The number of new Covid-19 cases would have a more profound relationship to the adjusted closing price of the NASDAQ than the number of new Covid-19 deaths. This is because the number of new cases would lead to more uncertainty about when things can reopen and how the US is handling the pandemic than the number of new deaths. An increase in the rate of new deaths is also most likely caused in the rate of new cases, so the number of new deaths will probably lag behind.

Chart/Graph:



Summary:

The data validates our hypothesis as there seems to be more a direct relationship between the stock market and the number of new cases than the number of new deaths. It can be seen that the number of deaths lags behind both other values, as predicted. One way we could possibly incorporate the data into our hypothesis is track the stock market price versus the rate of change of new cases and new deaths and see if there is any relationship. It would be interesting to track that data especially over the past couple of months which have seen a sudden spike in the number of new cases and deaths. Also, there was a better then expected jobs report that came out, so it would be interesting to see if that report affected the stock market more than the number of new cases.

External Data Source #2:

https://www.kaggle.com/suhailsh7/state-unemployment-rate-change-march-to-april-2020

What information does this data contain: The data contains the unemployment rate of each state in March and April of 2020 as well as how much the unemployment rate increased from March to April per state.

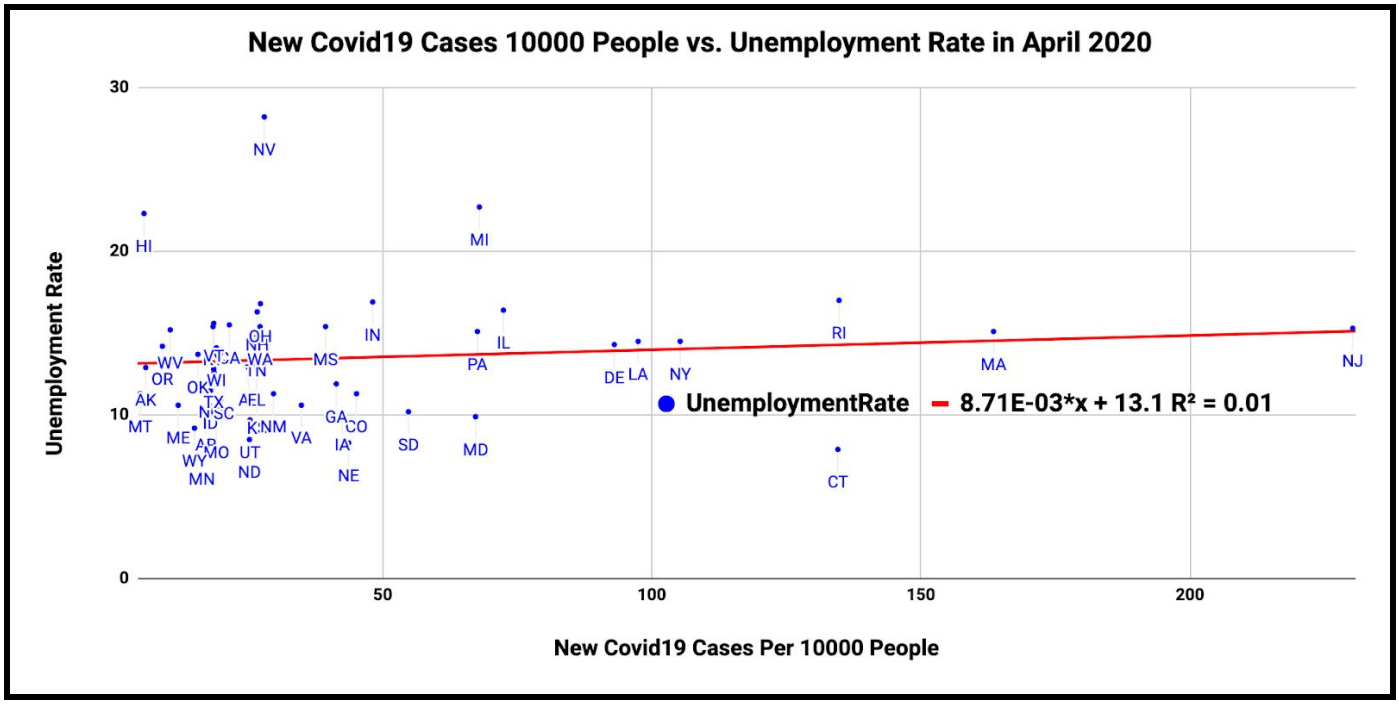
How do we plan on using this data: We will compare the unemployment rate of each state in April to the number of total Covid-19 cases and deaths in that state during that month. Also, we will compare the unemployment rate increase of each state to how many more Covid-19 cases each state had in April than in March to see if there is any relationship between the data.

1.

Hypothesis:

The states that have the highest unemployment rate, have the highest number of Covid-19 cases per 10000 because these states have the most uncertain future as to when the state will reopen.

Chart/Graph:



Summary:

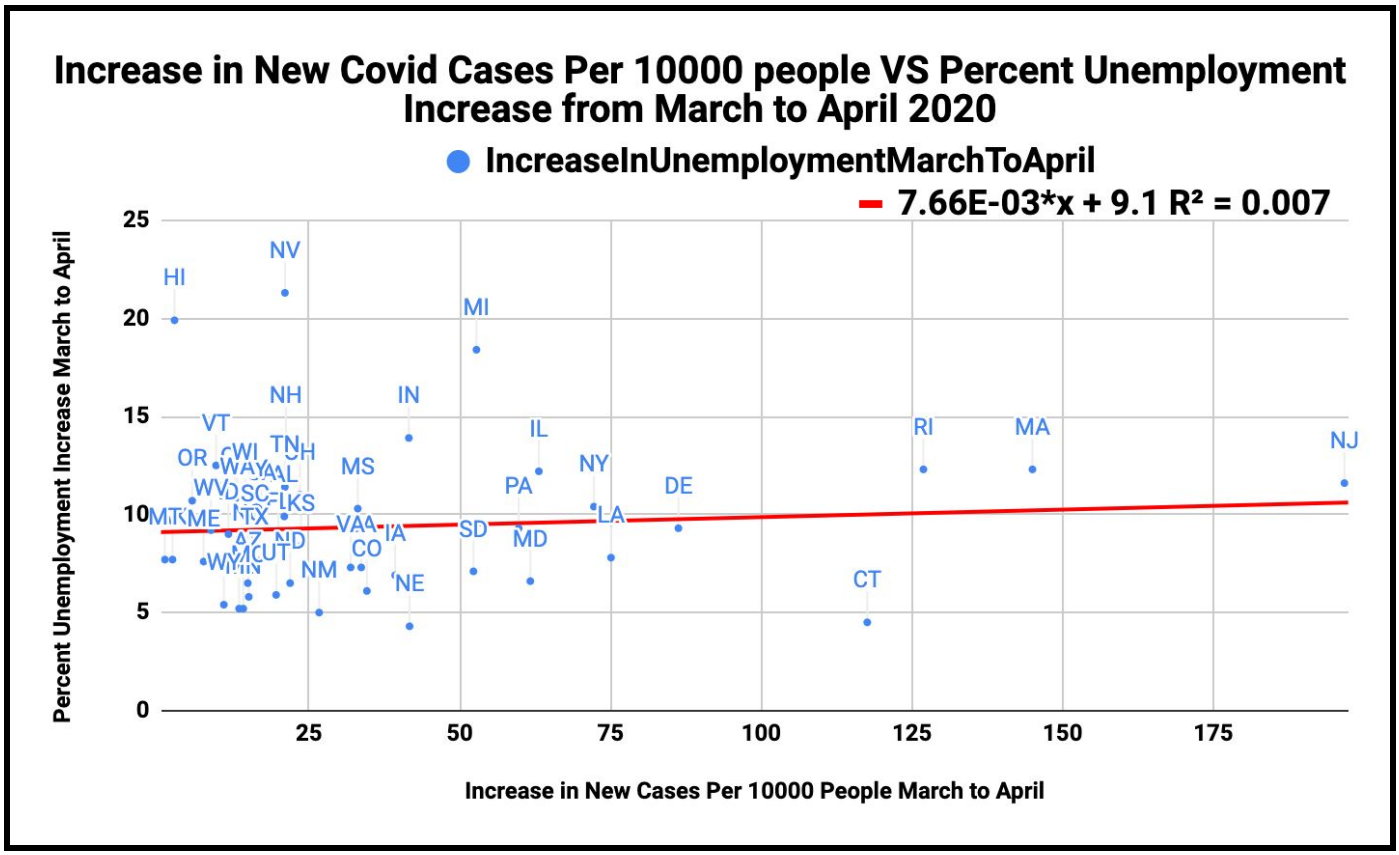
The data invalidates our hypothesis because while the slope is positive, it is extremely small (less than .01) and R^2 value is 0.01 which means that there is basically no correlation between the data. Nevada which has the highest unemployment rate, is in the smaller half of the number of new cases and New Jersey which has the highest number of new cases, has an about average unemployment rate. This relates to our application because we thought that maybe we could use tracking the number of new cases as a way of being able to accurately predict the job market. It would be interesting to see this trend now with states opening back up and people going back to work, but the number of new cases is rising again.

2.

Hypothesis:

The more that the unemployment rate increased in a state from March to April, the larger the number of new Covid-19 cases per 10000 in April than in March for that state. This is because more cases leads to a later reopening and a larger struggle on businesses who can't fully open because of the pandemic.

Chart/Graph:



Summary:

This data invalidates our hypothesis because once again while the slope is positive, it is less than 0.01. Also, the R^2 value is 0.007 which means that there really is no correlation between the data. Nevada and Hawaii had the highest unemployment rate increase from March to April, but they were on the low end of the increase in the number of new cases. This might be because those states rely so heavily on tourism, so other states closing might have a larger effect. This relates to our application in that we thought we might be able to predict whether the unemployment rates would greatly change from one month to the next if there was a large change in the number of new cases over the same time. It would be interesting to track this data as the number of new decreases from one month to the next and track whether the unemployment rate drops over that time and whether the states that saw the largest decrease in the number of new cases, saw the largest drop in their unemployment rate.

External Data Source #3:
https://www.kff.org/coronavirus-covid-19/issue-brief/state-data-and-policy-actions-to-address-coronavirus/

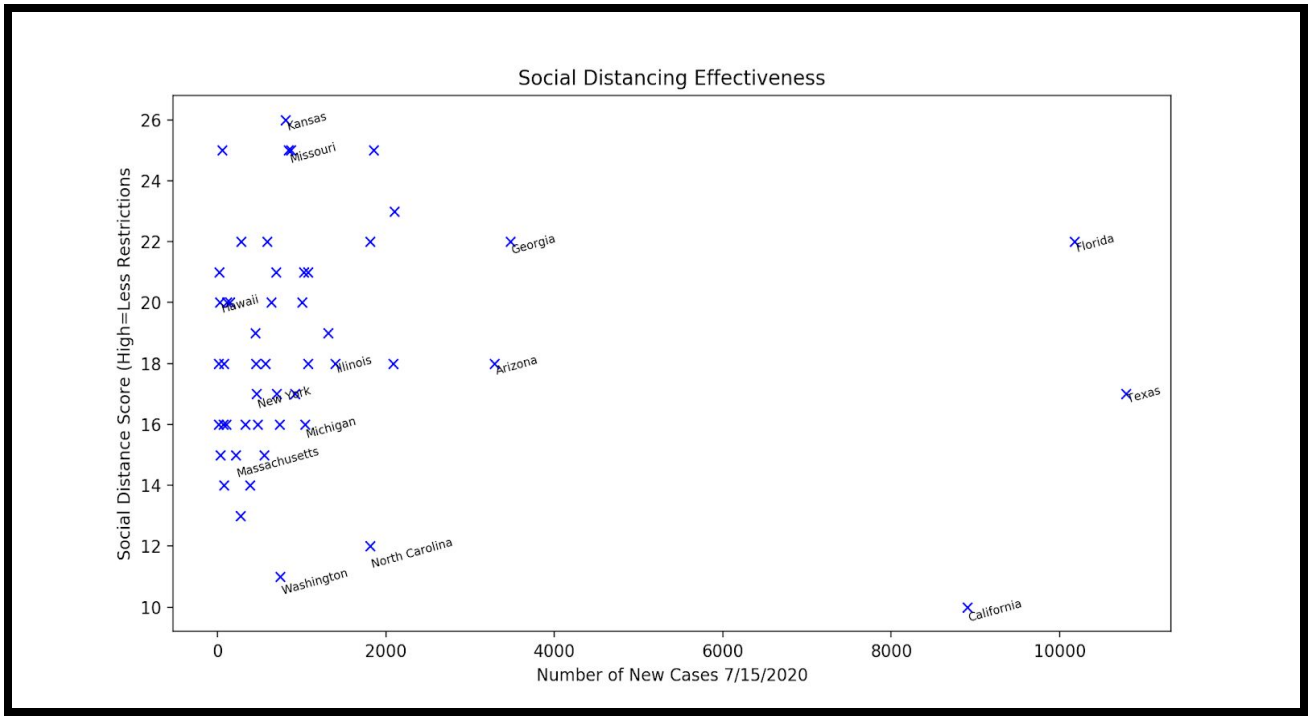
What information does this data contain: This dataset contains social distancing and closure policy actions by State. The policies considered included: status of reopening, stay at home orders, mandatory quarantine for travelers, non-essential business closures, large gatherings ban, restaurant limits, bar closures, mask requirements, primary election postponement, and emergency declaration.

How do we plan on using this data: We will create a 'social distancing score' using the information contained in this dataset. Per each field, we assign a number where low values mean more restrictive measures and high values are less restrictive. We translate this data to a total score that will be a sum of these values. We will then compare the total score by state against recent cases (last date available in Covidify database). Our table Covid By Date reports cases cumulatively, so we need to take a day by day difference aggregated by State (sum all counties) to calculate new cases per day. We expect that the states with recent surges will have higher scores.

Hypothesis:

1. The states that have seen surges within recent weeks will have had less restrictions in place. The states that have less cases will have lower scores (more restrictions in place).

Chart:



Summary:

This analysis shows our hypothesis to be true for some states, but not for all. The 3 states with big recent increases in the number of cases per day include Florida, Texas, and California. The structure of our database easily allowed us to link the social distance law by state to recent covid cases. We see for Florida and Texas, the social distance score is relatively high, meaning they have not been strict with enforcing social distancing measures by law. We also see states with really high scores (i.e. Kansas, Missouri) have had not as intense of a surge as Texas and Florida. We do confirm that states with more restrictions (i.e. Massachusetts, New York) have had less cases as of current. Ultimately, this graph fuels more hypotheses: is the high number of recent cases in Florida but not Kansas correlated to population density or population age? We did make some assumptions in creating this social distance score and experienced epidemiologists would likely weigh additional data in our Covidify database to create a more encompassing social distance score than just law restrictions.

External Data Source #4:
https://github.com/owid/covid-19-data/blob/master/public/data/ecdc/total_cases.csv

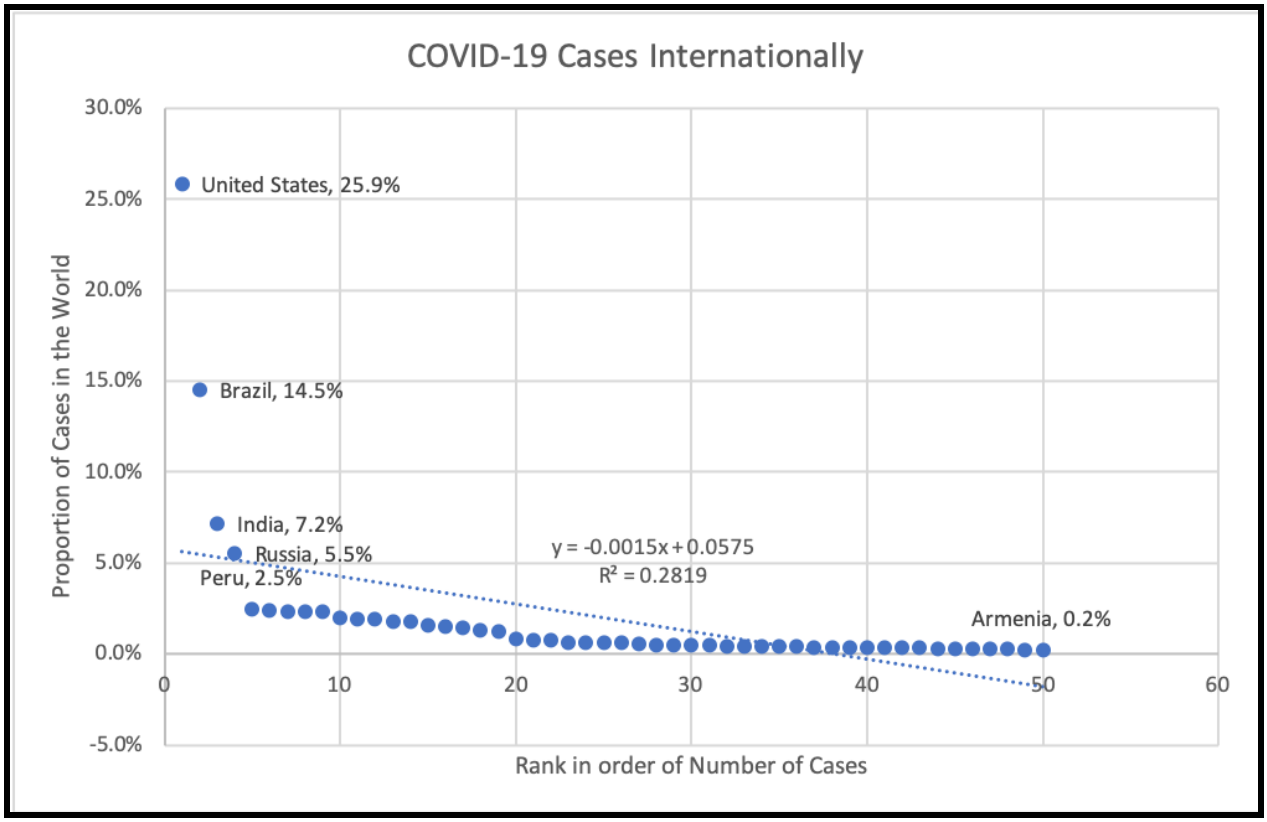
What information does this data contain: This dataset contains the number of COVID-19 cases per country per day as well as the global aggregate of cases.

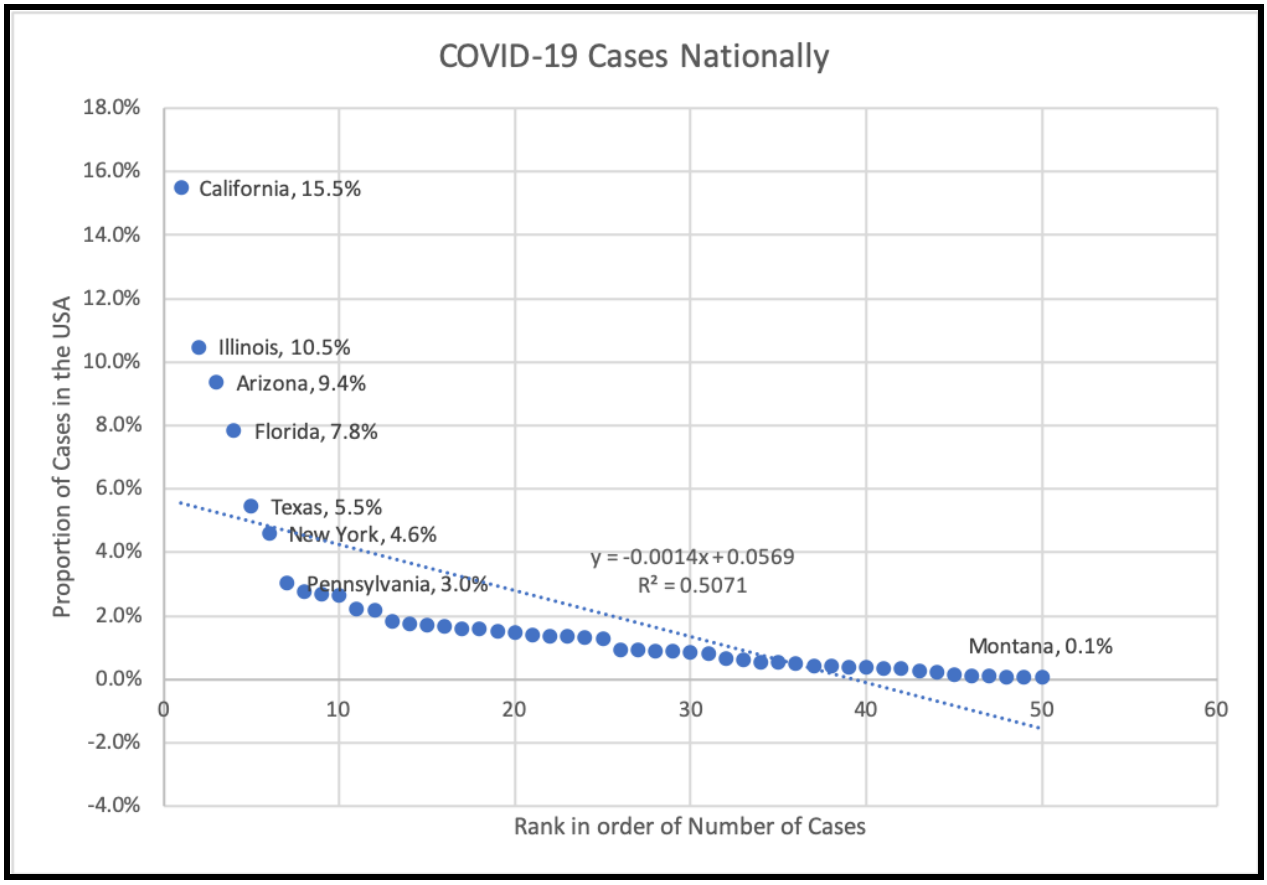
How do we plan on using this data: We will compare the number of cases in other Countries with high Covid-19 recent daily case rates to individual states.

Hypothesis:

1. The states of the US are governed at a federal and local state level regarding their policies. We hypothesize that the slope of the top 50 countries with the most COVID-19 cases will be similar to the slope of the COVID-19 cases for the 50 states due to the varying interpretation of the virus and thereby policy and precautions in each state. We would also expect the absolute value of the slope of the international plot to be greater than that of the national plot because of greater differences between countries than between states.

Charts: *(from output CSV 04HighestCovidCases.csv)*





Summary:

The slope of the best fit linear line through the top 50 countries with the most COVID-19 cases most recently have a slope of -0.0015 (i.e. international plot) and the slope for the 50 states of America is -0.0014 (i.e. national plot). Therefore, to some extent, our hypothesis is true and we can conclude there are some similarities in the COVID-19 case count across the countries with the COVID-19 case count across the US states. The R^2 (R squared) values of both plots do not indicate a strong effect or strong correlation. The absolute value of the slope of the international plot is only slightly greater by 0.001 than that of the national plot, but due to the low to moderate R squared values these plots are not indicative of a strong correlation or informative enough to draw a definitive conclusion from.

From observation of our plots, we can see that the international and national plots are similar in that there are about 4-6 outliers that hold a higher proportion of the case count than the remaining countries/states. The bottom 30 countries of the international plot seem to have a slope of ~0 whereas this observation can't be made for the national plot until we reach the bottom 10-6 countries. Therefore, this may indicate that internationally countries are more similar in count of COVID-19 cases compared to states within the USA. For our application, continually updating our model with the most recent data is important to track any significant changes or discrepancies in the USA's handle on COVID-19 compared to the world.