# Covid-19 Data Analysis

## Overview

With the current state of the world regarding Covid-19 and the effects that it has had on every person across the world, I thought it would be fitting to take a look at Covid-19 cases in the US to analyze. The uncertainty of the virus has probably been the most alarming and fear-causing factor for many of us and that's why I wanted to look for trends in Covid-19 cases which could potentially uncover some unknown characteristics of the virus. The main question I wanted to explore was the spread of Covid-19 within each state and within the counties in each state. In further investigating my main question, more sub-questions came up that I wanted to answer. This included what states and counties had the most cases as well as what the positivity rate for each state and county.

## Data and model

### Finding a Dataset

I found my data, "Covid-19 in USA", on Kaggle, gathered by the NYTimes and a Covid-19 tracking project. The data described the number of cases of Covid-19 in the US. There were three csv files given: cases of Covid-19 by county, cases of Covid-19 by state, and cases of Covid-19 in the US. These three datasets all had the number of test cases that tested positive and the date for which the positive results occurred. The date column ranged from January to July and the states included all 50 states in addition to US territories. The state cases and US cases datasets had additional information such as the number of hospitalized patients, number of patients in the ICU, negative and pending cases, etc.

### Cleaning the Dataset

I referred back to the main questions I wanted to answer in the beginning to decide how to use and clean these datasets. The main exploration I wanted to do was about the spread of Covid-19 cases. Therefore, the columns that I was most interested in would be the date, state, county, cases in county, positive cases, negative cases, and pending cases.

Since these columns would be the main columns used in data exploration and analysis, I needed to ensure that these columns had no outliers, missing information, and had overall accurate information. Starting with the date column, I wanted to make sure that all the dates were formatted as a datetime type so it would be easier to plot and understand the data in relation to time. The original type of the date column was a float, so I transformed the date column of all three datasets to datetime format.

Next, I took care of missing data. I looked at what percentage of the data missing, which I found to be 26.8%. To take a closer look at what data was missing, I looked at how many of the rows in each column had data missing. First looking at the columns I was most interested in, date and state had no missing data. However, there was missing data in the columns positive, negative, and pending, which were some of the essential columns I needed for exploration and analysis. To fix this, I deleted any rows which had missing data for all 3 columns: positive, negative, and pending. This meant that there was no data at that date for that state and the row had no other useful information. After deleting those rows, the positive column had no more missing data but the negative and pending columns still did. However, I felt that as long as all the rows had information for the positive cases, that didn't necessarily mean that there had to have been negative or pending results for that day.

Data Exploration

To get a general idea of the changes in positive Covid-19 cases in each state, I used the state cases dataset and grouped the data by state. For each state, I gathered all the dates and positive test results for those days. This made it easier to plot.

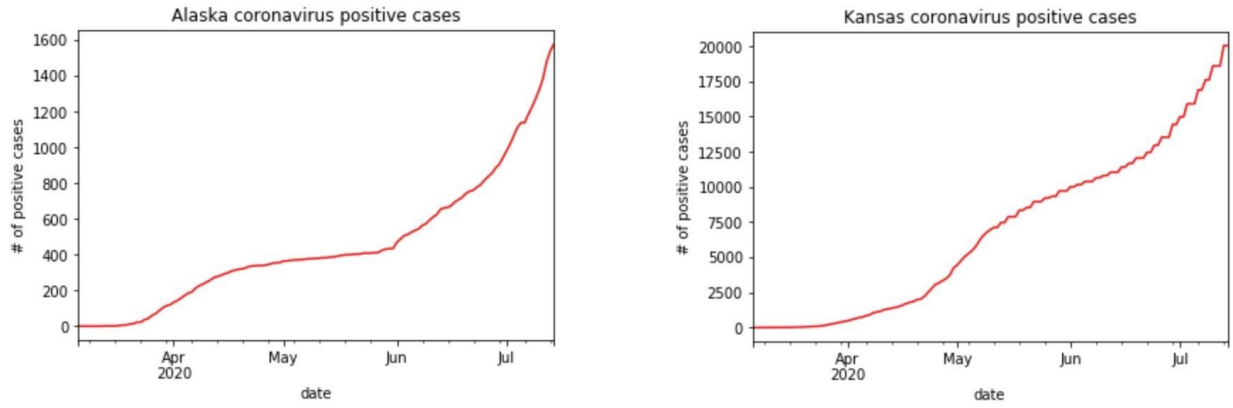



***Figure 1.1****: Graph showing the changes in positive Covid-19 cases since March 2020 in Alaska and Kansas*

At first glance, each state's graph showed an upward trend in Covid-19 cases. There was no easy way to compare each state's graph to another. Due to the difference in tests administered and state population, the number of positive cases differed drastically between each state. For example, when just glancing at the curves of figure 1.1, it seems like Alaska and Kansas have a similar number of cases and the cases are changing in a similar manner. However, when taking a closer look at the y-axis of the number of positive cases, the graph of Kansas had 12.5 times the number of positive cases as well as a different scale than figure 1.1. Therefore, the change in Covid-19 could not accurately be compared.

With the drastic differences between the high and low number of cases between each state, I decided to look at the actual numbers to get an idea of the total number of positive cases each state has had. To then answer my subquestion of which states have had the greatest number of Covid-19 positive cases, I grouped the state case dataset by state and summed up the total positive cases from each day. I sorted this data from highest total positive cases to lowest total positive cases as shown in figure 1.2. With this new table, I created a bar chart with each state and the total positive Covid-19 cases to get a better visualization of the difference in positive cases between each state. However, just like graphing the change in positive cases for each state, there was not a fair way to compare each state without understanding other factors like the state population or the density of the state.

| state_name | |
|---|---|
| New York | 34474680.0 |
| New Jersey | 14195355.0 |
| California | 11994093.0 |
| Illinois | 9775748.0 |
| Massachusetts | 8315627.0 |

***Figure 1.2:*** *total positive cases by state from highest to lowest*

Something that was given in the state cases dataset was a column of the total tests administered for that day. This was a good comparison that I could use to form a ratio between the total positive tests to total tests. This way, I could see what percentage of the total tests were actually positive which was a much more reliable comparison for each state.

Before calculating the percentages, I wanted to see if the total tests in each state really were as drastically different as were the total positive tests per state. I totaled up the tests in each state from each day recorded and sorted this table from highest total tests to least. I was proven right as the difference in the total test from the state with the most, New York, and the state with the least, American Samoa, was 213,600,805. Of course the state that had such a high number of total tests would have more positive test results.

| state_name | positive | total | positivePercentage ((positive/total)*100) |
|---|---|---|---|
| **New Jersey** | 14195355.0 | 75208416 | 18.874690 |
| **New York** | 34474680.0 | 213623603 | 16.138048 |
| **Massachusetts** | 8315627.0 | 53080639 | 15.666027 |
| **Maryland** | 4358772.0 | 29236794 | 14.908516 |
| **Connecticut** | 3608645.0 | 25000593 | 14.434238 |

***Figure 1.3:*** *table of top 5 states with the most positive cases (positive = positive cases, total = total tests given, positivePercentage = positivity rate)*

This further confirmed that finding the percentage of positive test cases out of the total tests would be beneficial due to the drastic differences in tests administered. Figure 1.3 explains the calculations to find the percentage of positive test results. Furthermore, the percentage found also described the positivity rate at which Covid-19 spread in each state which helped uncover which states had the worst spread rate of Covid-19 and which ones had the least.

One remaining issue I had was what to do with the pending tests. I discovered that 39 out of the 56 states had tests still pending. Therefore, I decided to look at the worst case scenario and see how the positivity rates would differ if all pending cases turned out to be positive. I added the total number of pending cases to the total number of positive cases to calculate the new worst case scenario positivity rate. However, to test if this was the right method to use to deal with the pending cases, I created a stacked bar chart (figure 1.4) of the original positive result percentages (red), the pending percentages (blue), and the negative percentages (yellow)out of total tests administered.

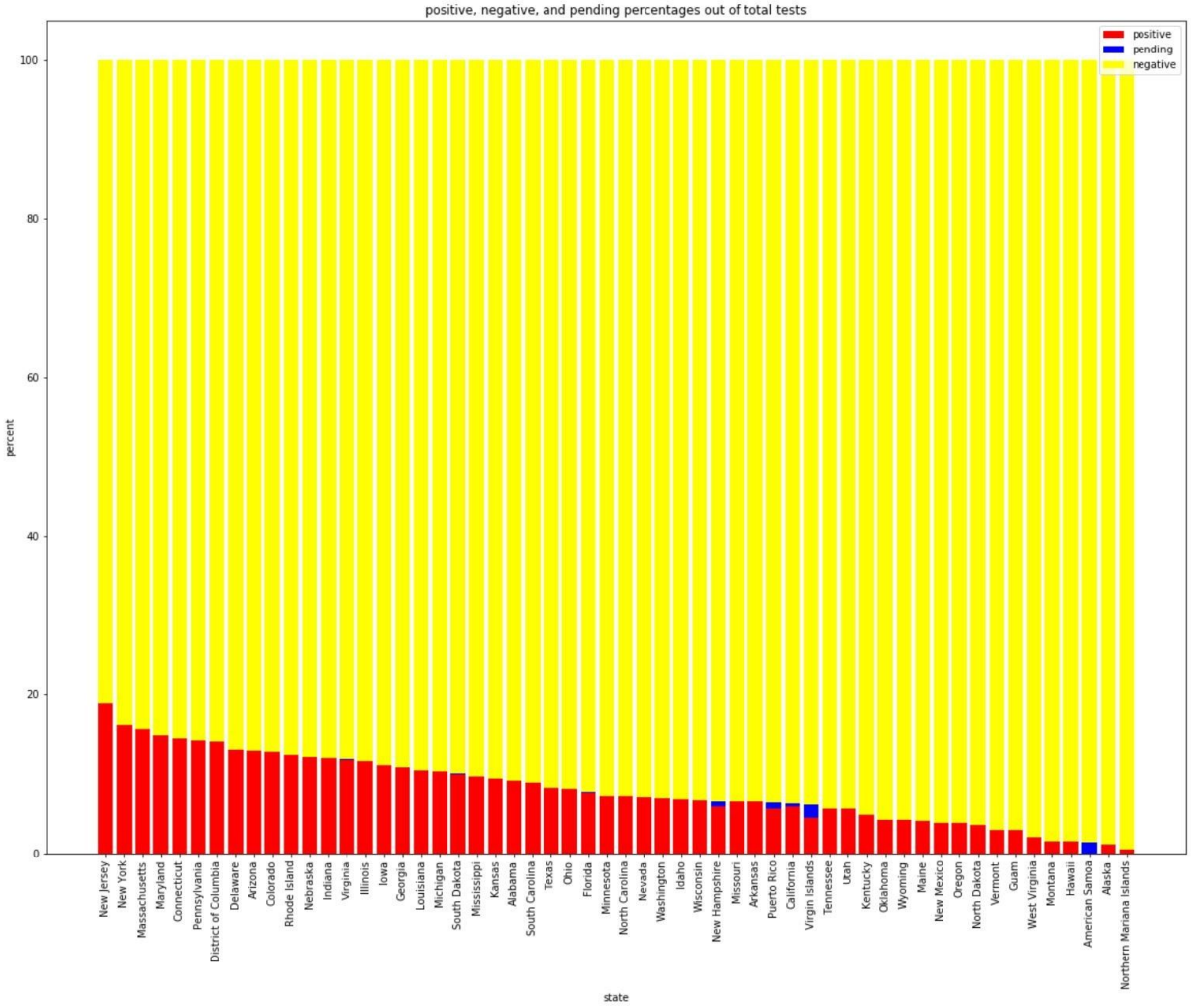



***Figure 1.4:*** *stacked bar chart of each state's positive, pending, and negative cases*

I was able to see how much converting the pending cases to positive cases would change the positivity rate for each state. In Figure 1.4, it can be seen that the pending cases assumed to be positive cases really only changed 5 of the state's positivity rate by a noticeable amount. This meant that pending cases were not a huge issue if they were to be ignored and I decided that just using the positive test results to explore the data would be okay.

The final exploration I did with the cases of state data was creating a line graph of the changes in positivity rate over time for the top 5 states with the highest overall positivity rate. I wanted to see if the higher the overall positivity rate, would there be a huge spike in positivity rates at one point or whether there was a common decrease in positivity rates around a certain time.

Next, I took a look at the county data per state. I organized the data first by state. For each state I made sure all the counties for the state were listed as well as all the list of dates in which Covid-19 results were collected. For each state, county, and date there was the positive number of cases for the columns specified. This allowed for an organized table that I could more easily use to make graphs and perform calculations.

To answer one of the original questions, which county in each state had the most total positive cases up so far, I summed up each county's total positive cases. Then, I filtered out the county with the largest total number of positive cases in each state.

| state | county | cases in county | total | county percentage of state |
|---|---|---|---|---|
| Northern Mariana Islands | Unknown | 2060 | 2060 | 100.000000 |
| Virgin Islands | Unknown | 7860 | 7860 | 100.000000 |
| Guam | Unknown | 107967 | 107967 | 100.000000 |
| District of Columbia | District of Columbia | 718628 | 718628 | 100.000000 |
| Nevada | Clark | 790886 | 993013 | 79.645080 |
| Rhode Island | Providence | 891122 | 1207527 | 73.797273 |
| Hawaii | Honolulu | 49945 | 71972 | 69.395043 |
| Illinois | Cook | 6224131 | 9642120 | 64.551478 |
| South Dakota | Minnehaha | 272059 | 427327 | 63.665296 |

***Figure 1.5:*** *table of state, county in state with the most cases,total cases in that county, total cases in the state, and the percentage of the maximum county case in the state to the total state cases ((cases in county/ total)*100)*

As I did with the state case dataset, I wanted to calculate a percentage to create an easy basis for comparison. The percentage I was most interested in was how much the county's with the most positive cases in its respective state accounted for the total positive state cases. To do this, I totaled up the positive cases, under the column total in figure 1.5, for each county and for each state to come to the total recorded positive cases for that state. Next, I took the county in each state with the largest number of positive cases and placed that number under the column, cases in county, in figure 1.5. Finally, under the column, county percentage of state, I took the column, cases in county, and divided that by the column, total. To compare these percentages between each state, I graphed these percentages in a bar chart.

Lastly, I wanted to better understand the changes in positive cases for each county which can be seen. So I looked at the 5 states whose counties had the largest number of cases accounting for the least amount of their state cases. I ignored the top 4 territories seen in figure 1.6. The counties accounted for 100% of the cases because they each only had 1 county. With the next top 5 states, I graphed a line graph of how the cases have changed over time in each county. In doing so, I was most curious about how the county with the largest number of positive cases per state fluctuated in positive cases per day compared to the rest of the counties.

| state | county | cases in county | total | county percentage of state |
|---|---|---|---|---|
| Northern Mariana Islands | Unknown | 2060 | 2060 | 100.000000 |
| Virgin Islands | Unknown | 7860 | 7860 | 100.000000 |
| Guam | Unknown | 107967 | 107967 | 100.000000 |
| District of Columbia | District of Columbia | 718628 | 718628 | 100.000000 |

***Figure 1.6**: All states/territories whose largest number of cases in the county accounted for 100% of its state/territory cases*

I then noticed that in these 5 graphs, there seemed to be a clumping of counties whose cases rose much faster than the rest of the counties. To further investigate, I created a more detailed visual representation for each state and its counties changes in cases over time. I started by looking at the 5 states in which the county who had the most cases were only a small percentage of the total state cases.

The patterns in the graph led me to wonder if there was a pattern of all states having only a handful of counties who accounted for most of the state cases. I began by going through each state and sorting the total cases in each county from largest to smallest. For each state, starting with the county who has the most cases in the state, and going down the list, I tried to find how many of the top counties would need to be summed together to make a combined total of 50% or more of the state cases.

| state | counties | number of counties | total counties | case percentage of state | county percentage of state |
|---|---|---|---|---|---|
| Delaware | [Sussex, New Castle] | 2 | 4 | 84.441160 | 50.000000 |
| New Jersey | [Bergen, Hudson, Essex, Passaic, Union] | 5 | 22 | 54.593395 | 22.727273 |
| Connecticut | [Fairfield, New Haven] | 2 | 9 | 64.583061 | 22.222222 |
| Massachusetts | [Middlesex, Suffolk, Essex] | 3 | 15 | 55.565820 | 20.000000 |
| Hawaii | [Honolulu] | 1 | 5 | 69.395043 | 20.000000 |

***Figure 1.7:*** *Top 5 states with the highest proportion of the number of counties whose cases summed together makes up 50% or more of state cases out of total counties in the state*
*State, counties = counties whose cases summed together make up 50% or more of the overall state cases, number of counties = number of counties whose cases summed together make up 50% or more of overall state cases, total counties = total number of counties in the specified state, case percentage of state = the percentage that the cases summed together make up out of the overall state case, county percentage of state = number of counties/total counties*100*

There were some states who ended up having a combined total of their county cases be much larger than 50%, such as Delaware and Hawaii in figure 1.7. However taking one county out of the equation would have meant that the combined total was less than 50%. Therefore, I set 50% as the minimum amount for the total county cases. After totalling up how many of the top counties it would take to count for 50% or more of the state cases, I calculated the percent that the counties that were totalled up accounted for in terms of the total counties in the state to really see if there was a small cluster of counties responsible for the state cases. This calculation can be seen in the county percentage of state column in figure 1.7.

## Results

I began looking at the number of positive Covid-19 cases there have been in each state up until this point. New York had the largest number of cases. In figure 2.1 it can be seen that even the state with the second largest number of cases, New Jersey, couldn't compare to New York who had about 2.25 times more cases. The top 5 states that have the most cases are New York, New Jersey, California, Illinois, Massachusetts. This was interesting because I expected the states that had the largest populations would be the states with the most cases. The states with the largest population, however, are California, Texas, Florida, New York, and Pennsylvania (U.S. Census Bureau, 2019). Only two of the most populated states made the top 5 Covid-19 cases.

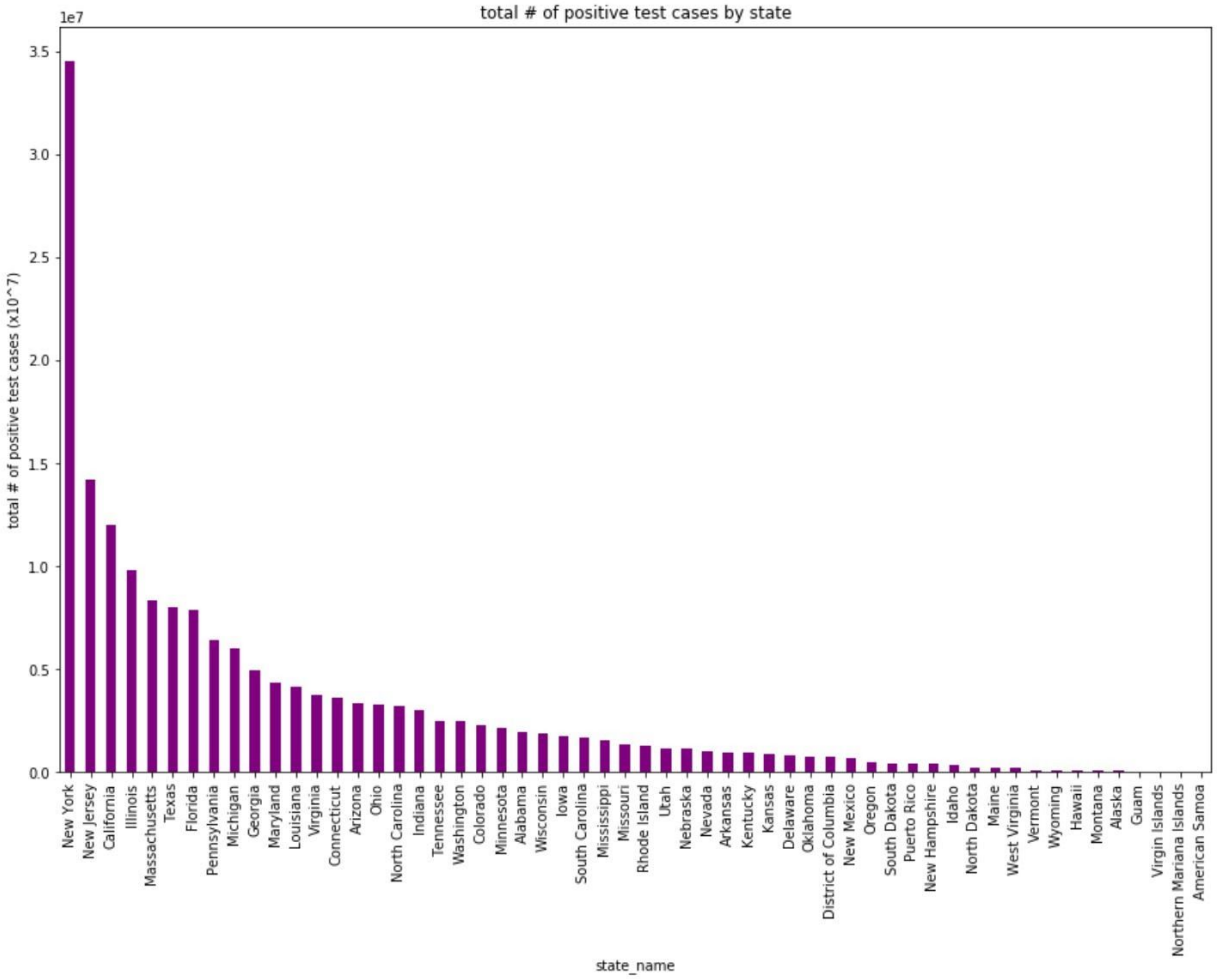



***Figure 2.1:*** *Graph of the total number of positive test cases in each state, ordered by the most to least positive cases*

Additionally, New York having the most number of cases only has half of California's population
yet had 3.5 times the amount of cases. As there seemed to be no correlation between how populated a state was and the amount of positive cases, I decided to take a look at population density. The most densely populated states found are New Jersey, Rhode Island, Massachusetts, Connecticut, Maryland (U.S. Census Bureau, 2010). Again 2 out of the top 5 most densely populated states were found to have the most positive cases.

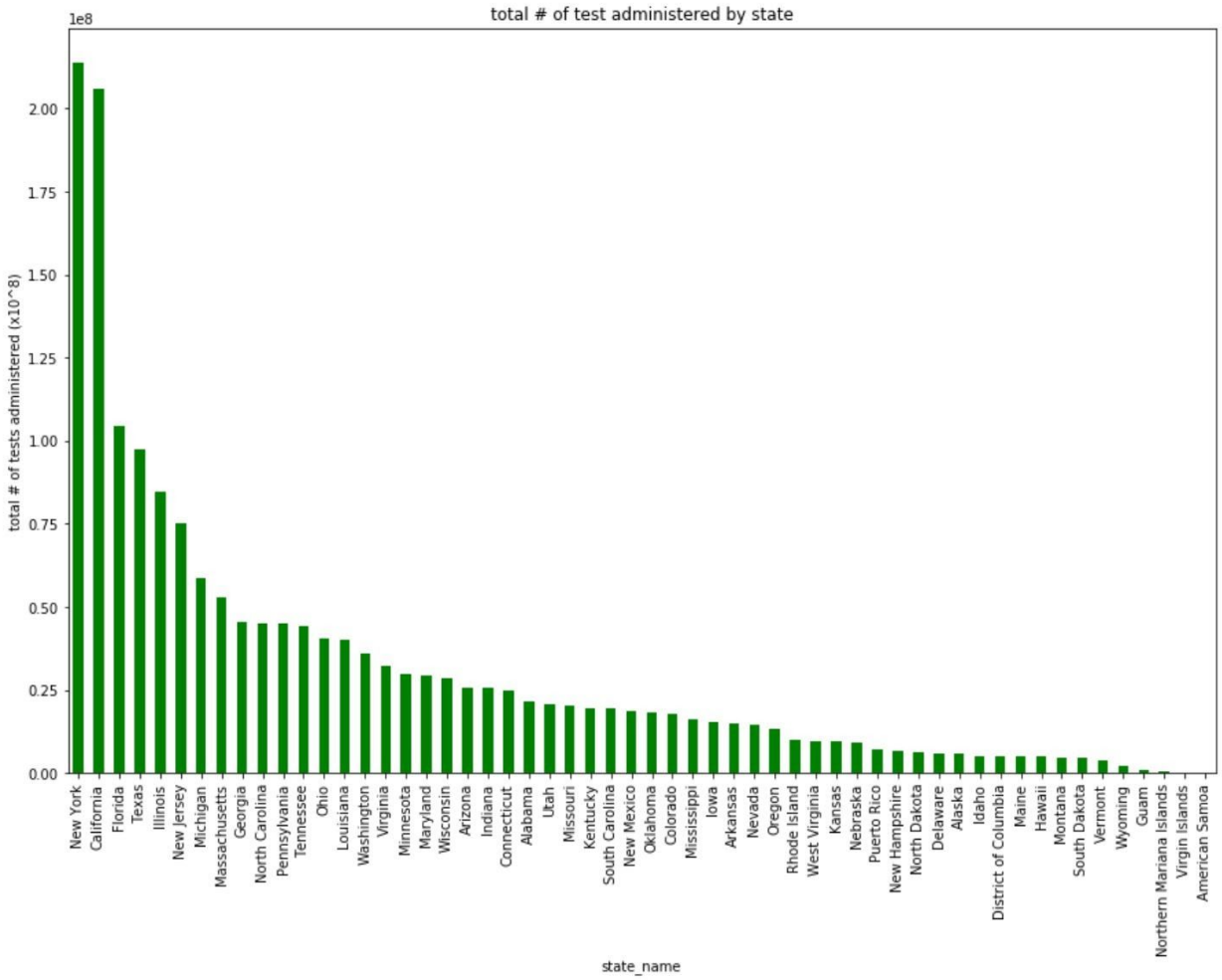



***Figure 2.2:*** *Graph of the total number of tests administered in each state, ordered from states who administered the most amount of tests to the least amount of administered tests*

Because no correlation was found between the top 5 states with the most positive Covid-19 cases and the population statistics found, I begin to take a look at the total tests administered. In doing so, I found that New York, California, Florida, Texas, and Illinois were found to be the states who had the most tests given (figure 2.2). This made sense with what I found earlier about the states with the largest population. These top 5 states who gave the most tests, aside from Illinois, were also the states with the largest populations. This logically made sense as the greater the population, the more people there are to want to take a test.

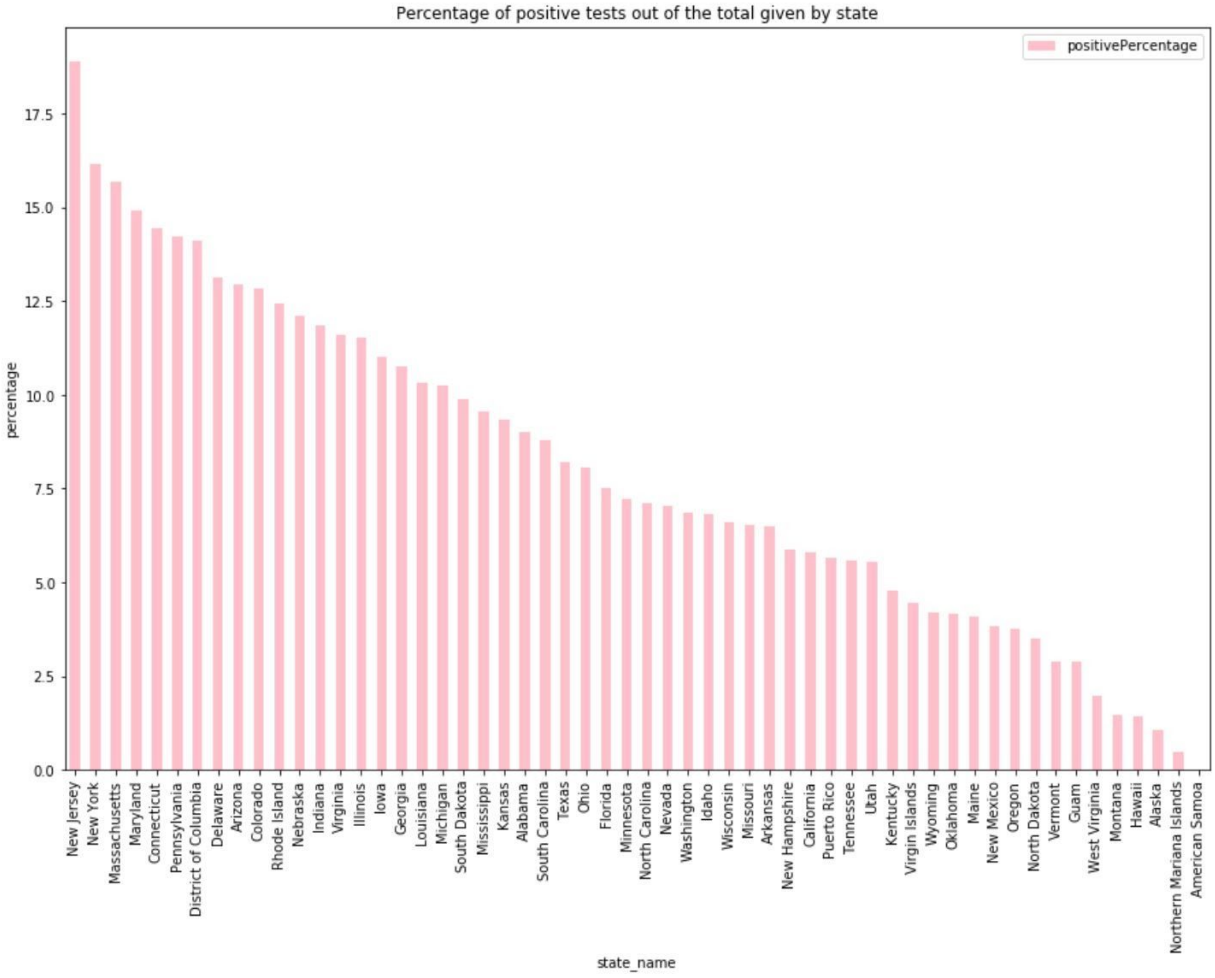



***Figure 2.3:*** *Graph of the percentage positive test results out of the total tests administered*

Diving deeper into another measurement that I could look at to find a pattern, I calculated the positivity rate using the total positive number of cases in each state and the total tests administered in each state (Figure 2.3). The results closely matched up with the top 5 most densely populated states. The top 5 states that appeared to have the highest positivity rates included New Jersey, New York, Massachusetts, Maryland, and Connecticut. With the exception of New York, the other 4 states were also in the top 5 states of most densely populated. This again logically made sense as the closer people lived next to each other, the more chances the virus would be able to spread as people were living their daily lives.

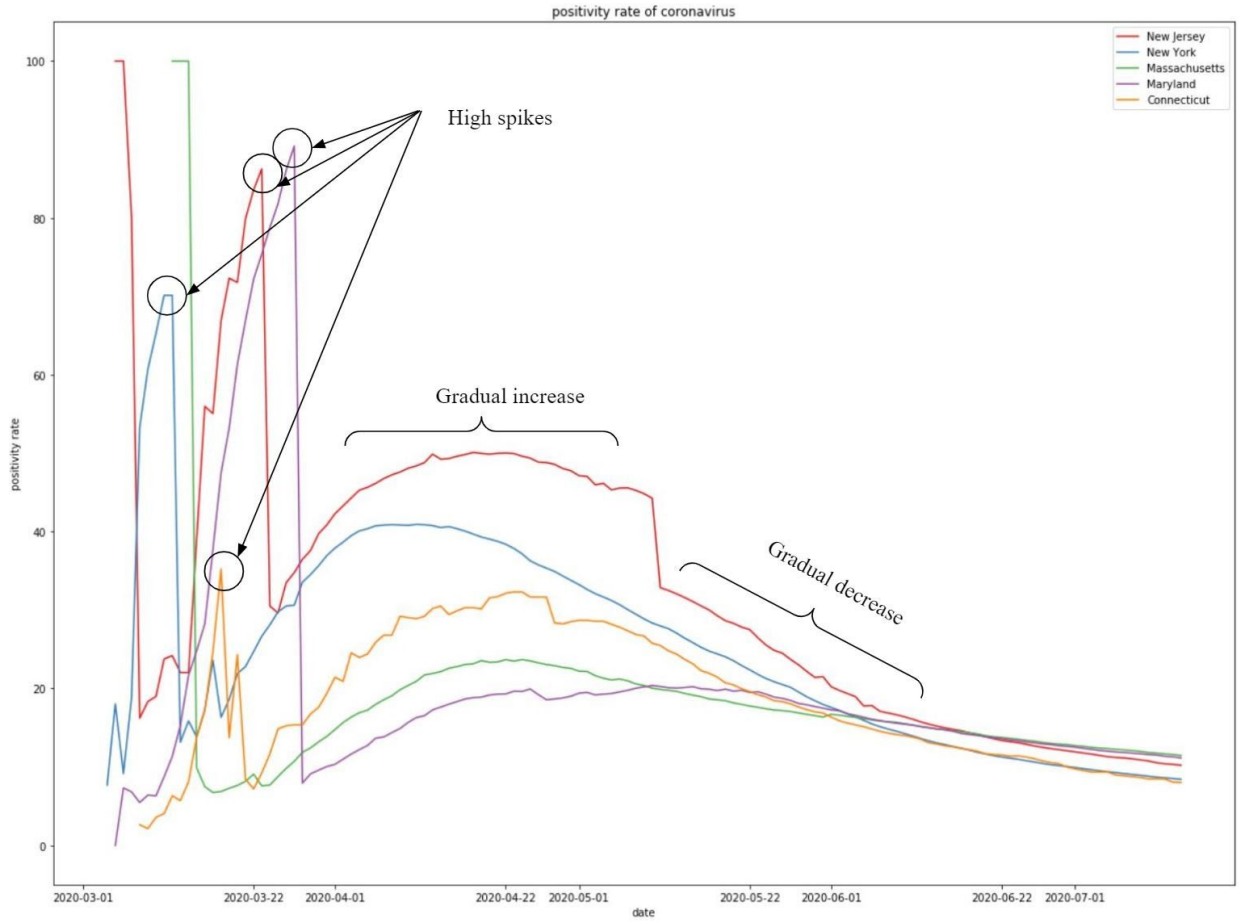



***Figure 2.4:*** *Graph of the change in positivity rate over time for the top 5 states who had the highest overall positivity rate*

Finally, I took a look at the positivity rate day by day in the top 5 states seen to have the highest positivity rates. In figure 2.4, New Jersey and Massachusetts were shown to start off with 100% positivity rate. However, this was due to the fact that only 1 or 2 tests had been administered and it just so happened that those 1 or 2 tests came back positive. In comparing these 5 states, they all seemed to have a pretty similar pattern of having a huge spike, dropping drastically, having a steady increase which later led to a steady decrease. What differed from state to state was the actual positivity rates. Although all on a similar trajectory, some states had higher positivity rates overall compared to the other states. Additionally, some states appeared to be behind in terms of when they first saw their spike and dip. This was probably due to when the virus was first introduced into a state. Because of the difference in introduction, the pattern of how the virus would affect a state was probably already known looking at states who had an earlier introduction to the virus.

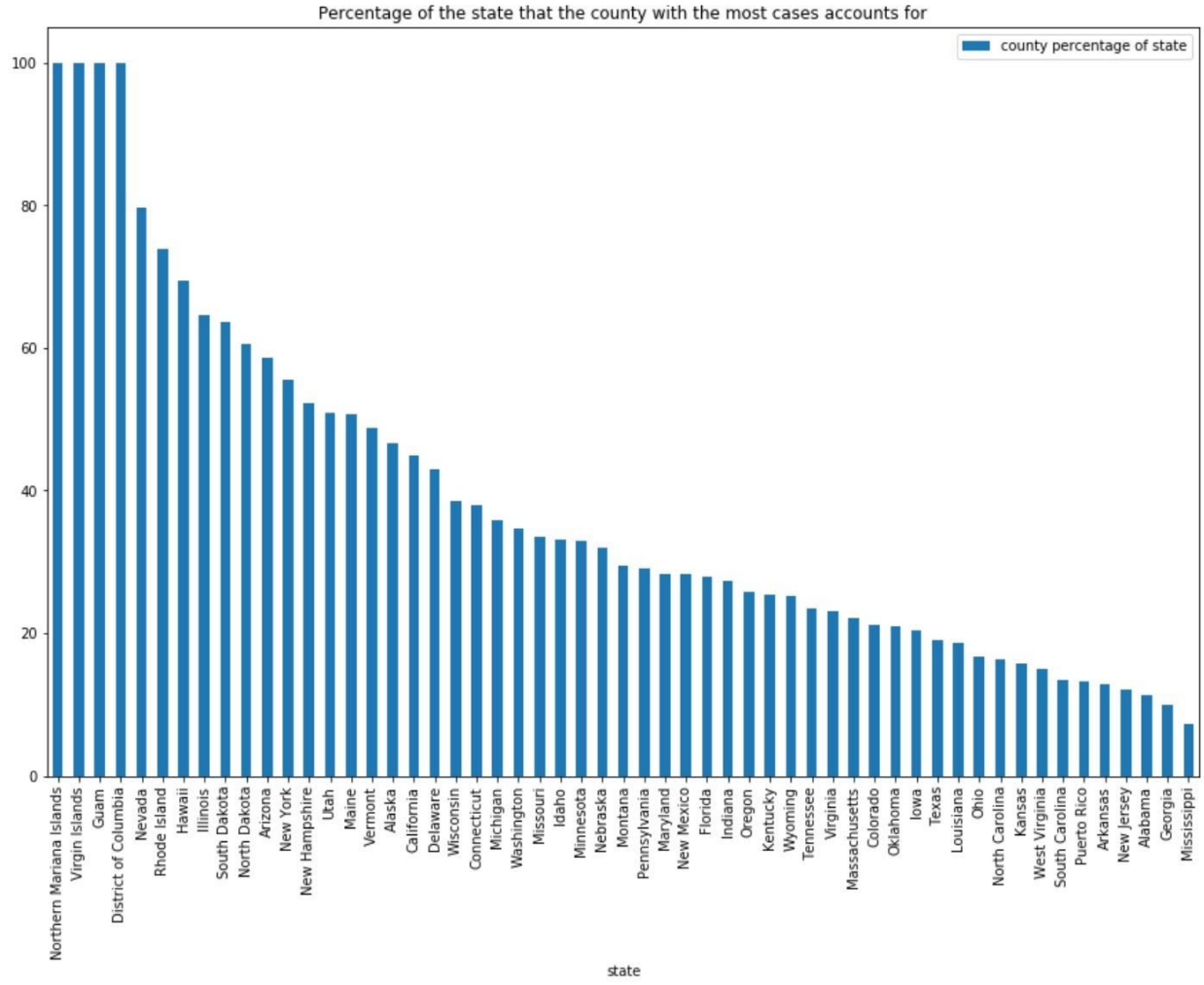
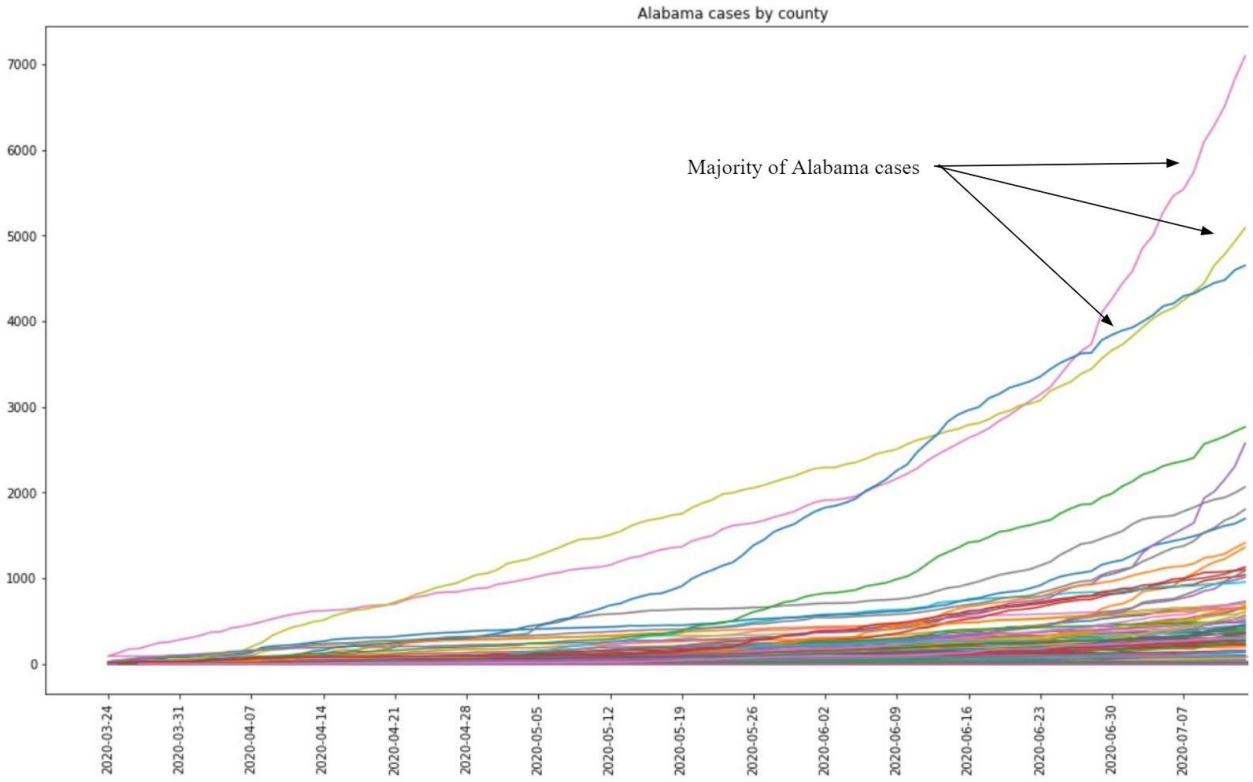
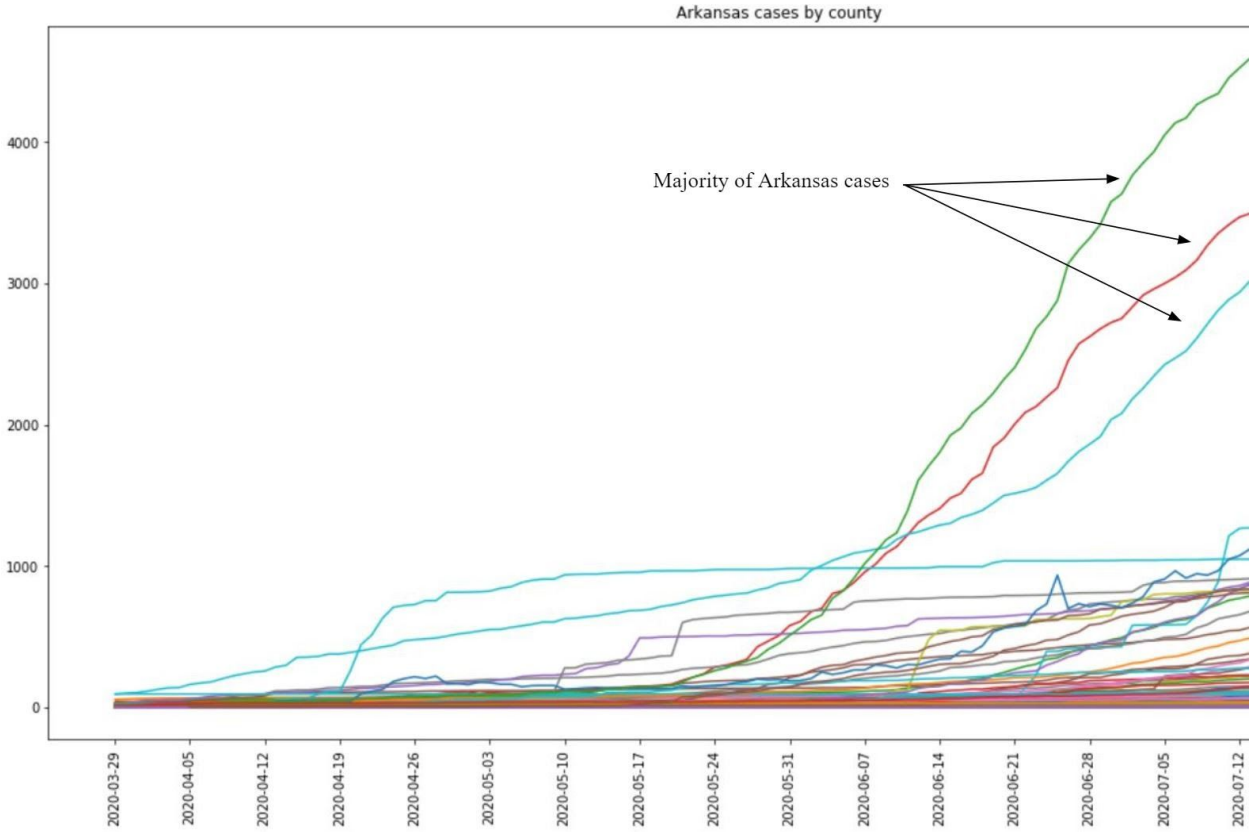
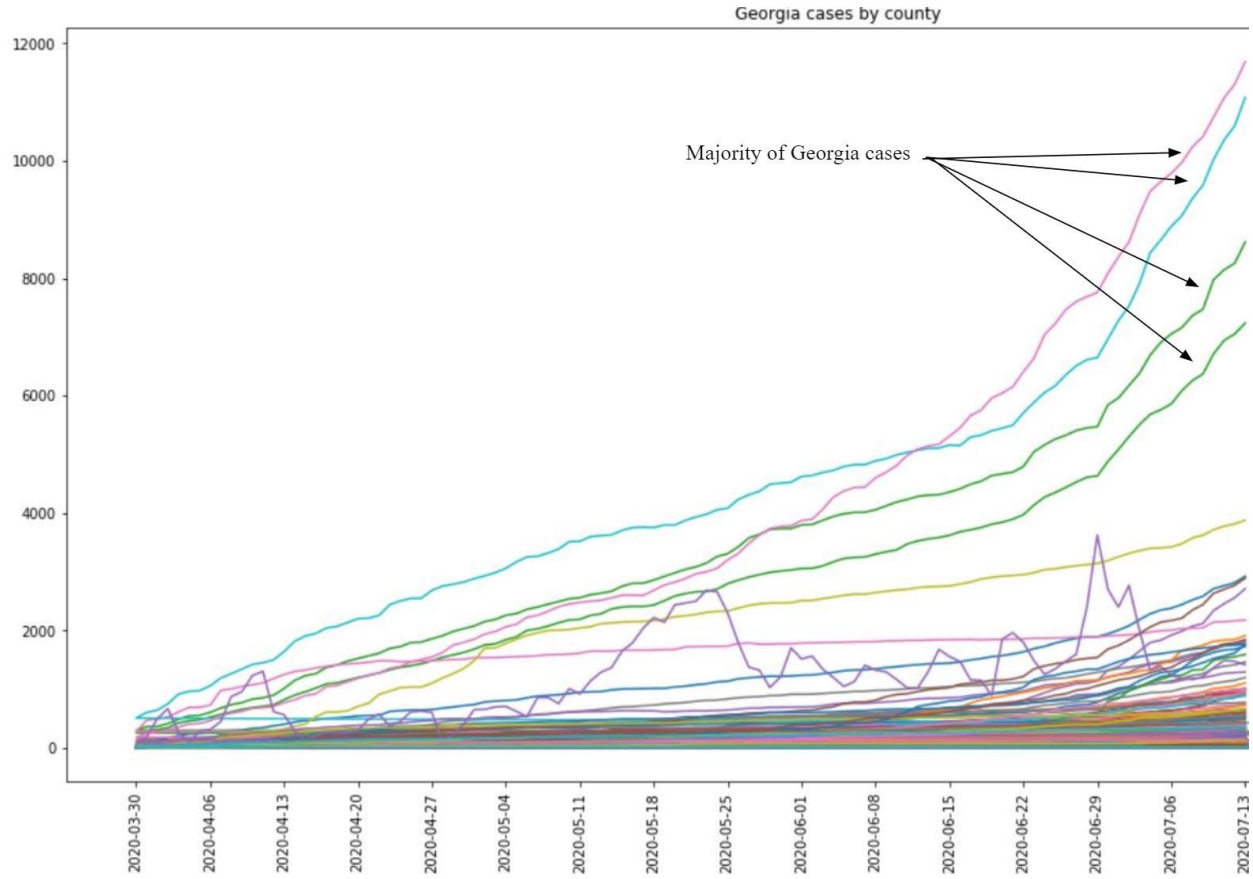


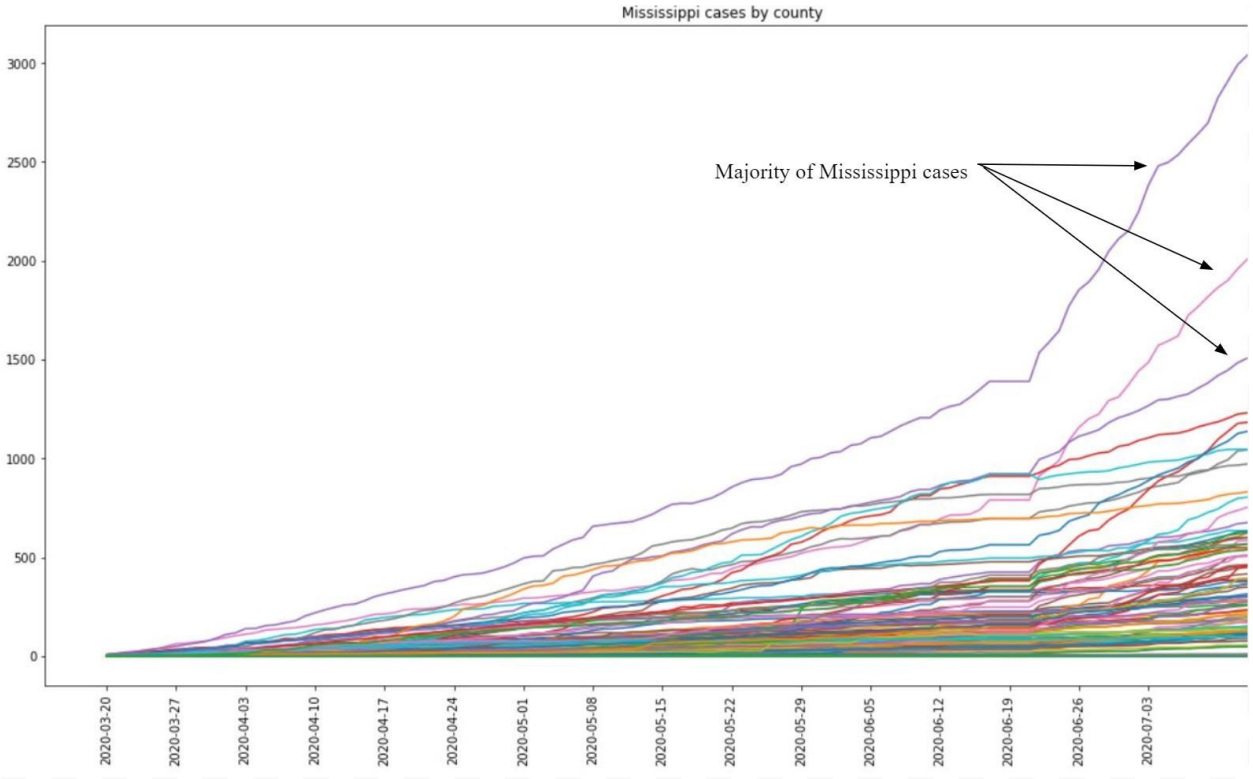

***Figure 2.5:*** *Graph of the percentage that the county who makes up the most cases in each state makes up out of the total state cases, ordered from the state's whose county cases makes up a lot of the state cases to the cases that makes up less of the state cases*

Taking a closer look at each state, I begin to explore how differences in county cases may affect the virus in its respective state (figure 2.5). In looking into the county's with the greatest number of cases for each state, I wanted to see how much of that county's cases accounted for the total state cases recorded. Using the total number of cases in the county with the most cases in a certain state and the total number of cases in the state, there was no trend that was found. The percentages that these county's made up of their respective states ranged from 7-80%. However, this could be due to the fact that bigger states have more county's or that a small number of counties are more densely populated than the rest of the county's in the state. This could be a reason why there is no pattern of one certain county taking up a majority of its' state cases across all the states.


Alabama cases by county
Majority of Alabama cases
7000
6000
5000
4000
3000
2000
1000
0
2020-03-24
2020-03-31
2020-04-07
2020-04-14
2020-04-21
2020-04-28
2020-05-05
2020-05-12
2020-05-19
2020-05-26
2020-06-02
2020-06-09
2020-06-16
2020-06-23
2020-06-30
2020-07-07


Arkansas cases by county
Majority of Arkansas cases
4000
3000
2000
1000
0
2020-03-29
2020-04-05
2020-04-12
2020-04-19
2020-04-26
2020-05-03
2020-05-10
2020-05-17
2020-05-24
2020-05-31
2020-06-07
2020-06-14
2020-06-21
2020-06-28
2020-07-05
2020-07-12


Georgia cases by county
Majority of Georgia cases
12000
10000
8000
6000
4000
2000
0
2020-03-30
2020-04-06
2020-04-13
2020-04-20
2020-04-27
2020-05-04
2020-05-11
2020-05-18
2020-05-25
2020-06-01
2020-06-08
2020-06-15
2020-06-22
2020-06-29
2020-07-06
2020-07-13


Mississippi cases by county
Majority of Mississippi cases
3000
2500
2000
1500
1000
500
0
2020-03-20
2020-03-27
2020-04-03
2020-04-10
2020-04-17
2020-04-24
2020-05-01
2020-05-08
2020-05-15
2020-05-22
2020-05-29
2020-06-05
2020-06-12
2020-06-19
2020-06-26
2020-07-03

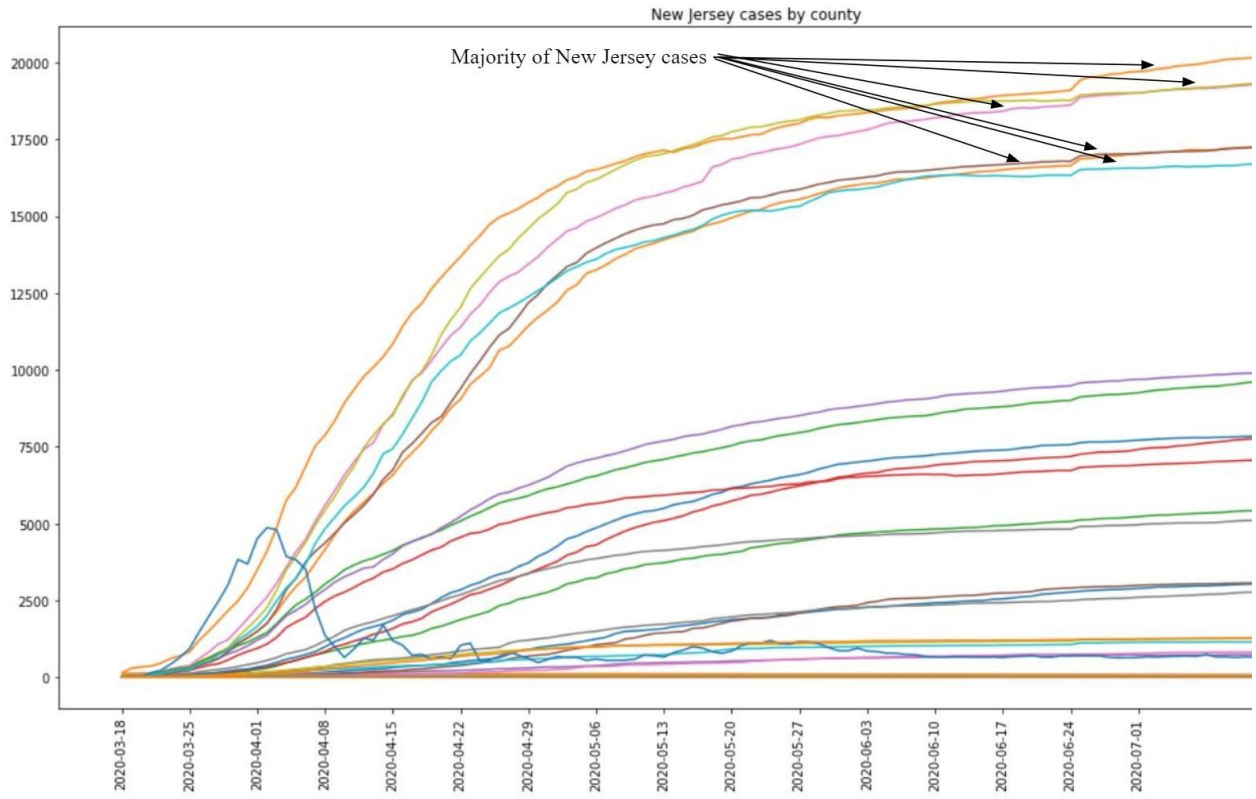



***Figure 2.6:*** *5 states whose county with the highest amount of cases accounted for the least amount of their state's overall cases*

Through taking a closer look at the changes in cases in each county, I found that the majority of the counties in these states seemed to have a low amount of cases over the 6 month course of the data. There was the exception of 3-5 counties who really saw a spike in cases in comparison. This led me to wonder whether there were a handful of counties in each state that were responsible for the majority of state cases. For example, in figure 2.6, the graph of the cases in counties of Arkansas, there are 3 distinct counties whose cases are increasing at a much higher rate than the rest. Similarly, there are 6 distinct counties for New Jersey, 3 distinct counties for Alabama, 4 distinct counties for Georgia, and 3 distinct counties for Mississippi in which these county cases are rising significantly higher than the other counties in the state.

Looking into how many counties in each state made up 50% of the state cases, a range of 1-16 counties were found. In figure 1.7, the range in the percentages of these counties in the state who made up roughly 50% of the state cases is 0.97-50%. However, only the state of Delaware who has 4 counties and the combination of the 2 of counties was found to make up 50% of the cases. The next percentage level of the counties who make up the majority of cases to the total counties in the state was significantly less, at 22%. Therefore, this percentage was so high for Delaware because it does not have that many counties in the state. For the remaining states, it could be seen that 20% or less of the state made up a majority of the cases. This still is a decent

chunk of the state, however, still manageable in which the state could put more restrictions on 20% of the state.

## Conclusion

In conclusion, population density was a clear factor that contributed to higher positivity rates in a state. In predicting a second wave and what that might look like, this could be useful information to further prevent another large wave. States with a denser population may want to consider holding back on certain businesses reopening. For example, indoor businesses such as hair salons, movie theaters, nail salons, could all either be banned from opening or only allowed to open outdoors. Since indoor activities increase the chance of one contracting the virus, a step to control the reopening of indoor businesses will minimize the chance of a large second wave. Furthermore, because a trend in how the virus hits a state has been seen, it can help better states understand how to prevent that huge spike in cases that many have seen in the first wave.

Looking more closely at the state, more precautions can also be taken within the counties of each state. 50% or more of the state's overall cases can be linked to a handful of counties. In identifying these handful of counties in figure 1.7, the state can put more restrictions on these counties to again further prevent a large second wave. Additionally, because the majority of these handful of counties only account for about 20% or less of the state, having stricter restrictions on these counties will not adversely affect the remainder of the state. Businesses in counties of the remaining 80% of the state would not have as harsh restrictions. This would mean that 91.24% of the country could be operating almost normally, under the condition that they follow all Covid-19 guidelines for reopening. In the second wave, the economy would not have to undergo a major shutdown again, but now with the information of densely populated states as well as the counties who account for the most cases in the state, the country will now be able to hopefully get the positivity rate even lower than it was before.

Works Cited

U.S. Census Bureau (2019). State Population Totals: 2010-2019. [Table]. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html#par_textimage_1574439295.

U.S. Census Bureau (2010). 2010 Census: Population Density Data (Text Version). [Table]. https://www.census.gov/data/tables/2010/dec/density-data-text.html