Final Project

Rahul Malhotra

4/22/2020

Coronavirus Data Analysis by Various Regions

World Data

Country,Other	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	ActiveCases
USA	1212835	0	69921	0	188027	954887
Spain	248301	0	25428	0	151633	71240
Italy	211938	0	29079	0	82879	99980
UK	190584	0	28734	0	0	161506
France	169462	0	25201	0	51371	92890

United States Data

USAState	TotalCases	NewCases	TotalDeaths	NewDeaths	ActiveCases
New York	327374	0	24944	0	249085
New Jersey	129345	0	7951	0	120123
Massachusetts	69087	0	4090	0	56879
Illinois	63840	0	2662	0	60533
California	56089	0	2283	0	46258

New Jersey Data

County	TotalCases	NewCases	TotalDeaths	NewDeaths	ActiveCases
Hudson	16527	0	885	0	15642
Bergen	16334	0	1215	0	15119
Essex	14654	0	1292	0	13362
Passaic	14430	0	633	0	13797
Union	13357	0	738	0	12619

Bergen County, NJ Data

City/Town	TotalCases
Allendale	56
Alpine	21
Bergenfield	741
Bogota	163
Carlstadt	92

*Some columns are omitted from this output of the table.

Motivation

The four datasets I chose all have to due with coronavirus data and are subsets of each other. Each one narrows the data down to a location that is within the one before it. The reaosn I chose these datasets is because I live in Bergen County and wanted to explore the severity of the situation, from a data-driven perspective. Bergen County was actually the first county in New Jersey to report a positive case. The town I live in, Dumont, has not been as much of a hot spot as other towns, but the first few cases in New Jersey had been from towns very close to me, such as Teaneck. Bergen County has drastically lead the state for the number of cases and I want to see how things have or have not changed in the towns I live by.

Data Scraping and Cleaning

worldometers.info data

The first three datasets were scraped from the https://www.worldometers.info/coronavirus/ website. Luckily, this website offers the data in a tabular format, so scraping it wasn't too much of a challenge. However, the table it has for the world data does not just include countries. It also contains an entry for the Diamond Princess cruise ship, we well as a total count so I had to go through manually and look for entries I wanted to remove. A similar approach was used for the United States data.

The next major cleaning process involved converting the variables (other than location) from a character representation to a numeric. When initially scraped, these values contain commas and some contain plus signs (to indicate the new daily gain of cases and deaths). To fix these, I created a function called "numCleaner" which takes in a database, a column of that database, and a string. The function goes through each value in that column and replaces all instances of the passed in string to a blank space, removing it. I did this so I could reuse this function for each dataset and for both commas and plus signs. One difficulty I came across was passing to the function only the columns with numeric values, since if say, the country column, was passed into the function, it would change each row to NA. To remedy this, I created a vector of strings for each dataset with the column names that were to be passed into the function. Another problem I encountered was getting an error when trying to remove the plus signs. After some research, I found out that '+' is a special character in regex and needs to be escaped so instead of simply passing "+" into my function, I had to pass "\\+" instead.

The last bit of cleaning involved changing any numeric NA values to 0, which there were many of due to how the data is written on the website. For example, if a country has no new deaths, or the value has not yet been updated, the entry is left blank, instead of being 0. Luckily, the function I created to change the columns to numerics converts these blank entries to NA. This made it easy to then just convert every NA value to 0. However, I noticed that on the website, for the UK and Netherlands, the "TotalRecovered" column is always labeled "N/A". I believe this is because they have reported they will no longer report this statistic because they feel that the reports given by China are not accurate.

insidernj.com data

To scrape the data for the towns and cities within Bergen County was much more challenging compared to the other datasets. The original source, and several others, I planned on using had no html table or path available when using SelectorGadget. I found other sources that had the data in a table that could be scraped but had a "Show More" option at the end. So when I loaded the data into R, it only showed some of the cities and towns from the full table and I was unable to find a way to get all the data.

I finally was able to find data at https://www.insidernj.com/bergen-county-town-covid-19-list-15982-cases-total-friday/ which is a news article containing the number of cases for each town and city in Bergen County. However, the data was not stored as a table, however I found a node labeled "p" using SelectorGadget which gave each sentence in the article as a entry into a character vector. So now, what I had was a vector where the

first few entries were some text, followed by lines containing the information I wanted as: “City/Town: Cases”, and then some more text. What I noticed was that the reported cases was given in alphabetical order, so my approach was to find the index of the first city or town, which happened to be Allendale. Then, I got the index of the last city or town by adding 69 to this value (since there are a total of 70 cities and towns in Bergen County).

I now had a vector of the cities and towns but needed to get this into a data frame. To do so, I made use of the separate function, passing in the vector as a data frame, since the city or town is separated from the number of cases in each by a colon, “:”. Having the data separated into columns, I was mostly done. The last couple of steps involved using the “str_replace_all” to get rid of the "*" which sometimes appears in the “City/Town” column, and then converting the cases column from a character type to a numeric.

Top 10 Recovery Rate (by Country)

Country,Other	RecoveryRate	TotalCases
Iceland	0.9577543	1799
China	0.9393347	82881
Thailand	0.9173083	2987
Luxembourg	0.8894984	3828
New Zealand	0.8761777	1486
Hong Kong	0.8645533	1041
Australia	0.8596466	6847
S. Korea	0.8592188	10804
Austria	0.8524422	15621
Switzerland	0.8405323	29981

We see China is among the countries with the best recovery rate, which we may expect due to their rapid response and strict restrictions. However, we see Iceland with the greatest recovery rate which is surprising since not many people talk about its success at containing the virus.

*Note: Only countries with more than 500 total cases were considered.

Top 10 Death Rate (by Country)

Country,Other	DeathRate	TotalCases
Belgium	0.1576382	50267
UK	0.1507682	190584
France	0.1487118	169462
Italy	0.1372052	211938
Netherlands	0.1246505	40770
Sweden	0.1218696	22721
Hungary	0.1156507	3035
Spain	0.1024080	248301
Algeria	0.1000430	4648
Mexico	0.0911865	24905

Among the top countries with the highest death rates, we see many european countries, including Italy and France. We expect to see them as they have been getting a lot of attention in the news due to the severity of the virus in their respective countries. However, interestingly, we do not need see the United

States among the countries with the highest death rates, despite having both the most deaths and cases.

*Note: Only countries with more than 500 total cases were considered.

Relating Serious Cases to Death Rate

The dataset for the countries has a column which gives the number of serious or critical cases. What I want to explore is if the proportion of cases that are classified as serious or critical cases is related to the death rate. There will be some cases that are considered serious or critical where people will recover, and there will also be non-serious or non-critical cases where people will not recover, so it is not excactly the same. However, I want to take a look at the countries with the highest death rates and see if the same countries are the ones with the highest proportion of serious or critical cases.

To do this, we'll create a new data frame which contains the country and this proportion, which will be defined as the number of serious or critical cases divided by the number of active cases. Then, we can join this data frame with the death rate data frame and compare the two values.

Country,Other	DeathRate	PropOfSerious
Belgium	0.1576382	0.0218588
UK	0.1507682	0.0096529
France	0.1487118	0.0397890
Italy	0.1372052	0.0147930
Netherlands	0.1246505	0.0192731
Sweden	0.1218696	0.0286560
Hungary	0.1156507	0.0267770
Spain	0.1024080	0.0316395
Algeria	0.1000430	0.0100686
Mexico	0.0911865	0.0564516
Channel Islands	0.0753676	0.0000000
Indonesia	0.0745663	0.0000000
Honduras	0.0704584	0.0102775
San Marino	0.0704467	0.0109890
Burkina Faso	0.0684524	0.0000000

Country,Other	PropOfSerious	DeathRate
Thailand	0.3160622	0.0180783
Iran	0.2059888	0.0636309
Brazil	0.1500415	0.0678236
Moldova	0.0880059	0.0310734
Lebanon	0.0834951	0.0337838
Andorra	0.0776699	0.0600000
Germany	0.0736611	0.0420880
China	0.0734177	0.0558994
Austria	0.0651026	0.0384098
Luxembourg	0.0642202	0.0250784
Mexico	0.0564516	0.0911865
Kyrgyzstan	0.0530612	0.0120482
Uruguay	0.0518135	0.0258752
Argentina	0.0492936	0.0532024
North Macedonia	0.0476190	0.0559947

The above data frames show the top 15 countries, first by death rate and then by proportion of serious or critical cases. We can see that the only country that shows up in both is Mexico, which suggests that the proportio of serious or critical cases does not tell us much about a country's death rate. This is a bit unexpected since we would assume that the more serious or critical cases a country has, the more people will die. However, it could be that there is some lag. As the coronavirus reaches its peak, we will have a large number of cases that may not exactly be in line with the number of deaths that there "should be", that is, the deaths must come after the cases. So, it could be that there will be a time will the two will be more strongly correlated.

Also, another factor to consider is that classifying a case as serious or critical can be a bit subjective, especially when each country and hospital has their own guidelines. Unless there is an objective and universal way to classify a case as such, this may not be the best comparison to make.

Predicting Number of Deaths

All Countries

```
## TotalCases
## 0.06351195
```

Using a naive approach, we construct a univariate linear model where the response is the total number of deaths and the predictor is the total number of cases. Looking at the coefficient for the total cases, we see it is about 0.064. That is, for every new case, we expect the number of deaths to increase by 0.064. In other words, we can say that the death rate is around 6.4%. However, let's get the death rate for each country, and compare its average to this value.

```
## The average death rate among all countries is 0.040606714345245
```

Taking the average of the death rate among all countries, we see that is about 0.04, which is about two-thirds the value of the coefficient we got in our linear model. We should expect a disparity in the two values since the linear model only considers the number of cases as a predictor. There are likely several other characteristics within a country that play a role in determining the number of deaths a country will have, namely those having to do with that country's health care and ability to treat those who have been affected.

United States

```
## TotalCases
## 0.07329621
```

We get the coefficient for the total cases to be about 0.074 which implies that for every new case in the United States, we expect the number of deaths to increase by about 0.074, that is a death rate of 7.4%. This value is about 0.01 greater than the coefficient for the model for all countries. Now, let's compare this value to the United State's actual death rate.

```
## The average death rate among all countries is 0.0576508758404894
```

Again, we see that the actual death rate is lower than the estimated one we obtained from the linear regression model, likely due to other covariates playing a role in the number of cases that result in death.

Conclusion

While we found that the United States is not among the top 10 countries with the highest death rate, we do see that it does have a higher death rate, both in actuality and by our regression model, than compared to the world average. This shows the severity of the pandemic in the country, as well as the country's lack of preparedness for the virus, despite being a superpower.

Predicting Number of Cases

All Countries

```
## TotalTests
##  0.1210073
```

Again, we use a naive approach, but this time to predict the number of cases based on the number of tests. We see that the coefficient is about 0.12. That is, for each new test, it will result in about 0.12 new cases. We can think of this as a positivity test rate. So, based on the linear model, there is about a 12% global positivity rate. Now, let's see how this compares if we take the average of this rate, which will be defined as the number of cases divided by the number of tests.

```
## The average positivity rate among all countries is 0.0828182189024995
```

We see that the average of the actual positivity rate is about 0.08, which is about two-thirds of the value we obtained from the linear model. A lower value for this rate is actually "better" since it implies that there are less positive cases per test. Thus, this likely suggests that there are other factors, other than the number of tests, which are resulting in a lesser value. One major factor, that is not available in the dataset and is hard to quantify, is the degree of social distancing a country is practicing. Some others are the overall level of health and age of the population.

United States

```
## TotalTests
##   0.225176
```

Looking at just the United States, the coefficient for the total tests is much higher. This model implies that for every new test, it will result in 0.22 cases. Or, in other words, the United States has a 22% positvity test rate. Again, let's compare this to the average rate among all states.

```
## The average positivity rate in the United State is 0.162525455846761
```

Similar to when we looked at all countries, the actual positivity rate is about two-thirds the value of the positivity rate predicted by the linear model. Since it is similar to all countries, it suggests that the other factors involved in the positivity rate are not much different for the United States compared to the rest of the world. This may suggest that the country's social distancing could be on a similar level to others, but again, there are other factors, such as health and age to consider.

Top 10 Positivity Rate (by Country)

Country,Other	PositivityRate
Algeria	0.7150769
Andorra	0.4482965
Ecuador	0.3976625
Brazil	0.3198921
Mali	0.2670350
Afghanistan	0.2614745
Mexico	0.2489479
Dominican Republic	0.2467268
San Marino	0.2368742
Honduras	0.2246805

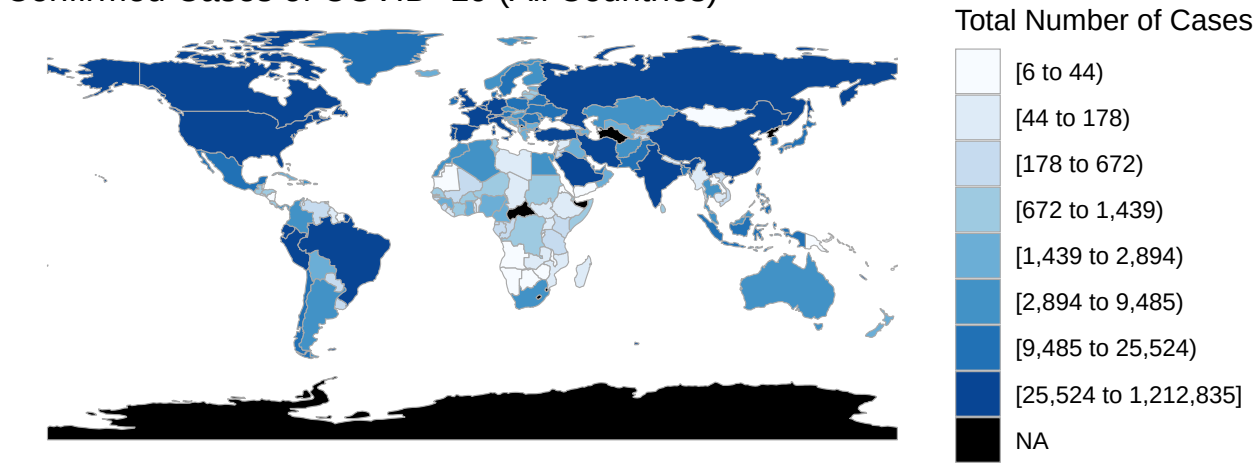
Looking at the countries with the greatest positivity rates, I am quite surprised. We see that Algeria is an outlier with a positivity rate of about 72% and the rest of the top 10 countries having a rate of at least 22%. I am also surprised to not see the United States make the top 10, as they have the most cases. However, it could be that there are not within the top 10 since they may be conducting a lot of tests.

Another factor to consider is the restrictions countries have placed on testing. Since it is expected that many people have the virus, countries are not able to test everyone. Thus, only those who qualify to get tested are getting actually getting tested. These qualifications tend to be showing clear symptoms and the severity of one's condition. So, it could be the case that the countries with higher positivity rates are also those with stricter testing restrictions, since that would mean they only normally test people who are much more likely to have the virus.

*Note: Only countries with more than 500 total cases were considered.

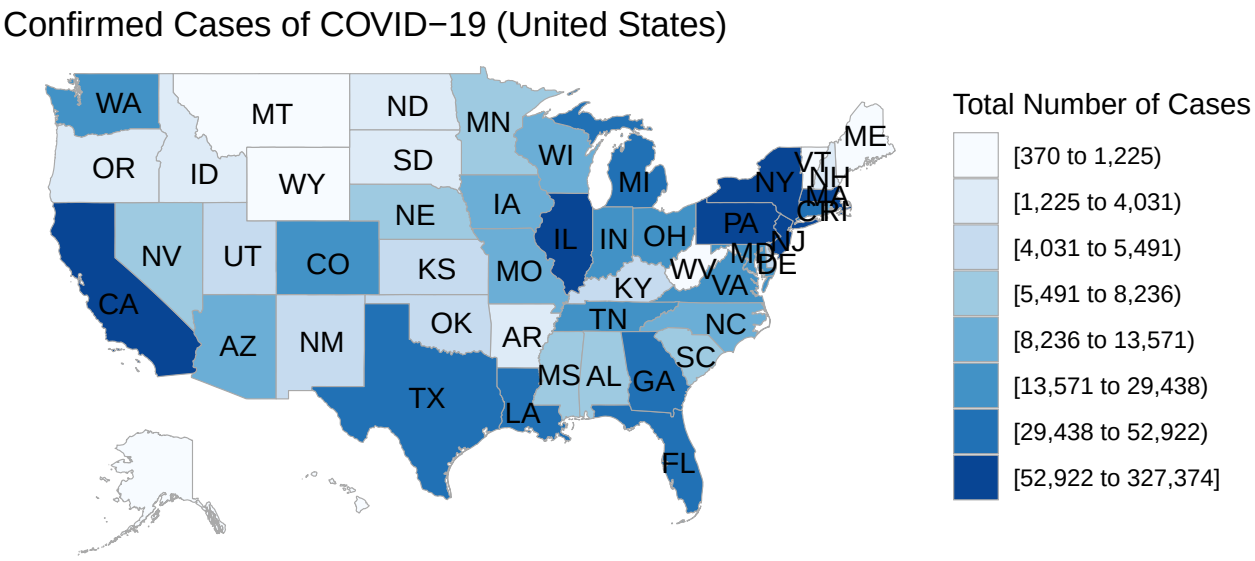
Creating Choroplethr Maps

All Countries

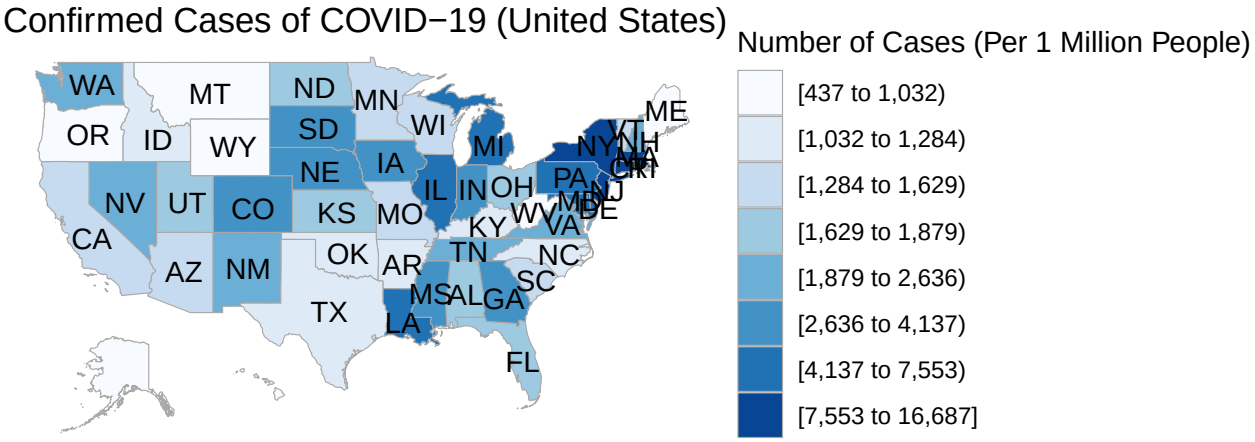


One interesting thing I noticed from looking at this map is that the number of cases in the continent of Africa are very low. This is quite the contrast from the Ebola outbreak of 2014 where Africa was the most affected.

United States



Looking at the total number of cases in the United States, we can see that some of the states with the most cases are those near the coasts. This could make sense since when people travel into America from foreign countries, they usually tend to go to these coasts. If we looked at an earlier version of this map, say in March, we might see this pattern be more apparent and would indicate where the virus started in the US. However, now, it is more likely that the states with the most cases are likely those with the highest populations. Let's take a look at this by looking at the number of cases per capita.

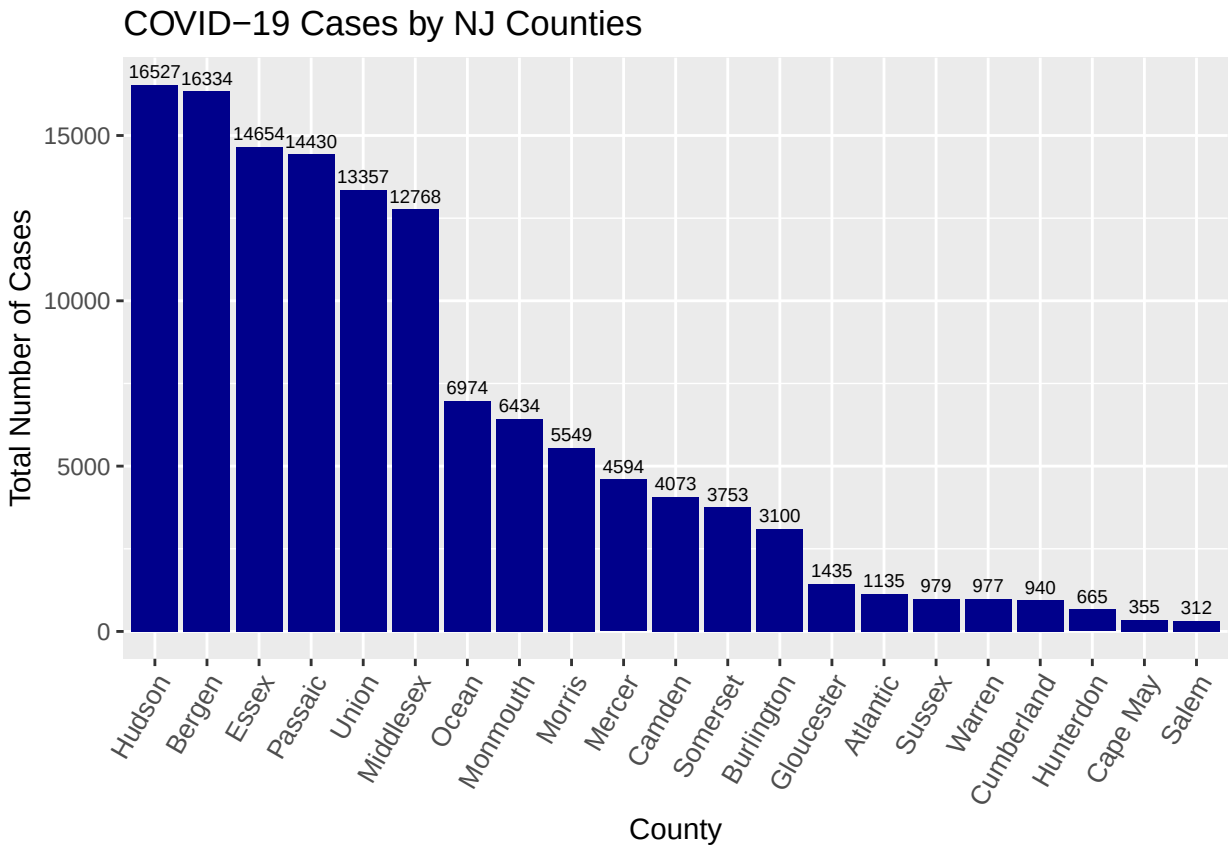


When we adjust the number of cases to be per capita we can see that it is the highest in the northeastern region of the country. Even states like New York, which have a relatively large population, still show the highest number of cases per capita. Also, it is interesting to see how many of the states with a higher number of cases per capita are concentrated in this area. This might suggest how the virus does spread exponentially, especially in a region where there is a lot of travel and interaction.

One difficulty I had with creating this choroplethr, that seemed more of an issue with R, was the name of the column for the number of cases per 1 million people. Originally, the name of the column was "Tot Cases/1M pop" but when I tried to pass this name into the "dplyr" rename function, it would not work. I tried 'Tot Cases/1M pop' and 'Tot Cases/1M pop' as well and neither worked. I think this may have happened because the website used some kind of formatting for the space in between "Tot" and "Cases/1M", so to fix this, I renamed the column to "CasesPerMillion" in the choroplethr version of the dataset.

Focusing on New Jersey

All Counties

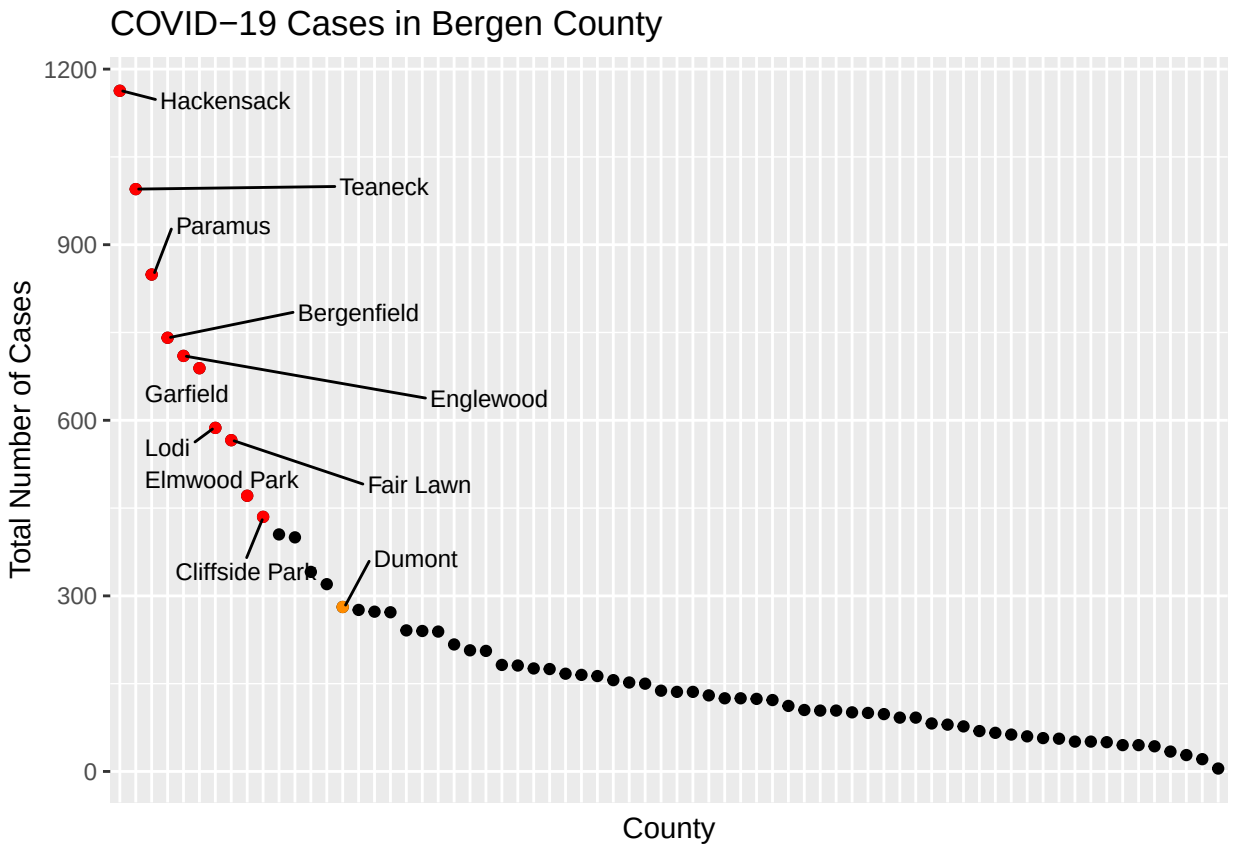


Looking at this graph, we can see the New Jersey counties with the highest number of cases. We see that the top 6 counties range from about 12,700 (Middlesex) to 16,500 (Hudson) cases. Then, we see a sharp drop to the county with the 7th most cases, which is Ocean county with only about 7,000 cases. The graph is fairly skewed to the right showing the relatively large disparity within the state. Bergen county, which is where I live, has been leading the state for the number of cases ever since the initial outbreak and was very far ahead of the other counties. However, as of the day I am writing this, Hudson counted has surpassed Bergen county for the most number of cases.

Initially, I wanted to make a choroplethr map for New Jersey. My plan was to use the "county_choroplethr" function with this dataset, and then zoom in on New Jersey to get a county map of just the state. However, when I ran the code, I got the following error:
"Error in approx(cum, xx, xout = (1:g) * nnm/g, method ="constant", rule = 2, : zero non-NA points"

I checked my data again and there were no NA values. I tried searching online for a solution and came across a couple threads where people had a similar issue but nothing having to do with choroplethr maps, so I went with this graph, which I feel does a good job at exploring the number of cases by county in the state.

Bergen County



Similar to the cases by NJ county graph, we saw a very positively skewed graph for the cases in Bergen County. For the top 10 or so city/towns, we see a relatively large gap between the number of cases. After that, we see the cases between the rest of the city/towns is much closer to one another. This really shows where the "hot-spots" are and how the virus can become so widespread in a single area.

For this graph, I chose to highlight the top 10 city/towns, since there are a total of 70 in Bergen County. The top 5 city/towns are actually very close to me (within 15 minutes) so it's shocking and a bit scary to see that the virus is so close. I also chose to highlight my town, Dumont. Although it's not in the top 10, it is definitely up there, being at the top of the non "hot-spots". Seeing that our town has almost 300 cases (as I'm writing this) is surprising, since I remember not too long when we had got our first case. With a population of under 18,000, I am also surprised I personally have not heard of anyone in Dumont testing positive, but I'm sure that will change.

Creating this graph actually took a fair bit of time. Plotting the data itself was not hard, but one challenge was finding out how to highlight specific points, such as the top 10 city/towns. I researched online and found out that you can add multiple "geom_point" and "geom_text" calls together, which I was unaware of prior. The next challenge was finding out how to add text labels to these points without having them overlap with each other and the points too much. What I did was play around with the size and directional adjustmentsf fielsd of the "geom_text" function until I got a plot that I was satisfied with. I then remembered that there is a function, "ggrepel", which actually takes care of this so I used that instead. However, even with this, there was some overlap, so I made some more manual adjustments with the sizing and padding values to get the final graph.

Appendix

Sources

Dadax. 2020. *worldometers.info* .https://www.worldometers.info/coronavirus/
Insider NJ. 2020, May 2. *insidernj.com.* https://www.insidernj.com/bergen-county-town-covid-19-list-15982-cases-total-frida

Github

https://github.com/RMalhotraGit/Data-Wrangling-Final-Project