Data Wrangling Final Report

Matt Dula

5/4/2020

I. Introduction

The COVID-19 pandemic has had a never-before seen impact on global society. It has now affected almost every corner of the globe one way or another. In times of catastrophe like the COVID-19 pandemic, easy access to information is vital to get a grasp on the situation. Having access to clean and easy to read COVID-19 data can ensure that individuals are well educated and informed on the topic. I wanted to focus my project on this pandemic because it is such a huge event that will affect society for years to come. My goal is to scrape coronavirus data from the internet to attempt to examine trends in death rate from COVID-19 across US states. In addition to scraping coronavirus data, I will scrape a series of health-related measurements for each state to compare with COVID-19 death rates. My data will include information on obesity, smoking, healthcare, and overall public health. I will then construct map plots using these metrics for each state to visualize COVID-19 death rate and health related information geographically in America. I aim to show that areas with worse health ratings (higher obesity ratings, higher smoker rating, worse healthcare, etc.) will have higher death rates than areas with better health.

While evidence has shown that individuals with preexisting conditions like obesity and smoking are more likely to die from the novel coronavirus, the virus has been affecting area with high population density. The virus can spread easily between individuals in densely populated areas, infecting thousands and overwhelming hospitals which may lead to higher death rates. A state's overall health may have little to do with its COVID-19 death rate. To account for this, I will also compare population density with COVID-19 death rate in New York and New Jersey counties.

Methods

Coronavirus Data

I began by scraping coronavirus data from https://www.worldometers.info/coronavirus/country/us/ which includes information on a state level. The table I scraped into R contained columns for total number of cases, total number of deaths, number of new cases, and number of new deaths as well as several other statistics. For the purpose of my project I only required the number of cases and the number of deaths for each state. Once the data were loaded into R, I selected for only the state name, total cases, and total death columns. The numeric columns were originally input into a new table as characters and not integers. To convert to an integer, I first had to remove any non-digits and then convert the columns to integers. This step was necessary for all tables and numeric columns that I scraped into R. From here on I will not mention this step, but it can be assumed that I included it each time I scraped a new table.

I then cleaned the state name column by removing all non-states from the table. Fortunately, all non-states included in the table were at the end of the table, so I removed them manually all at once. The first row also included the total for the entire country, so I removed that row as well. This left the table with 51 rows for each US state and Washington DC. To avoid having to manually remove non-states in future tables, I created a smaller table with only the names of the 50 states and Washington DC so I can use a join to remove unwanted rows. Lastly, I mutated the coronavirus table to add a new column that divides the total deaths by total cases to calculate a death rate for each state. The coronavirus table now includes the name of the state, the total number of cases, the number of deaths, and the death rate seen below.

```
## # A tibble: 10 x 12
##    USAState TotalCases NewCases TotalDeaths NewDeaths ActiveCases
##    <chr>         <int> <chr>          <int>     <int> <chr>
##  1 new york     326823 +2,940         24874       226 248,604
##  2 new jer~     129287 +1,849          7951        65 120,065
##  3 massach~      69087 +1,000          4090        86 56,879
##  4 illinois      63840 +2,341          2662        44 60,572
##  5 califor~      55694 +835            2253        41 47,845
##  6 pennsyl~      52817 +769            2845        13 49,056
```

```
##  7 michigan      43754 ""              4049        NA 24,046
##  8 florida       36897 +819            1399        20 34,812
##  9 texas         32894 +901             908        20 15,585
## 10 connect~      30173 +886            2556        61 27,552
## # ... with 6 more variables: `Tot Cases/1M pop` <chr>, `Deaths/1M
## #   pop` <int>, TotalTests <chr>, `Tests/1M pop` <chr>, Source <chr>,
## #   death_rate <dbl>
```

Health Data

Next, I added the each of the health variables. Each variable needed to be scraped from a different online source. I added obesity data from https://en.wikipedia.org/wiki/Obesity_in_the_United_States which gives the adult obesity rate for each US state. The smoking rate data came from https://worldpopulationreview.com/states/smoking-rates-by-state/ which gives the adult smoking rate by state. Next, https://wallethub.com/edu/states-with-best-health-care/23457/ gives an overall healthcare rating for each state. Lastly, https://www.usnews.com/news/best-states/rankings/health-care/public-health gives state by state rankings (1-50) for six different public health variables. To get a single public health statistic for each state I averaged each ranking for each state. Each of these four health variables that I imported required basic cleaning such as removing non-digits and converting to integer or double. The public health dataset required additional cleaning to make it compatible with the rest of the tables. Upon scraping the data, the state name column contained each name twice (New JerseyNew Jersey instead of New Jersey). To solve this problem, I arranged the data in alphabetic order by state as well as the state table. I then changed the public health state column with the state table to give the public health table state names in the correct format. Finally, with all health tables in the correct format, I joined each with the coronavirus table.

```
## # A tibble: 10 x 12
##    USAState death_rate Obesity_rate smoking_rate health_score
##    <chr>         <dbl>        <dbl>        <dbl>        <dbl>
##  1 new york     0.0761         25.7         14.1         52.3
##  2 new jer~     0.0615         27.3         13.7         53.4
##  3 massach~     0.0592         25.9         13.7         62.3
##  4 illinois     0.0417         31.1         15.5         52.2
##  5 califor~     0.0405         25.1         11.3         51.2
##  6 pennsyl~     0.0539         31.6         18.7         55.7
```

```
##  7 michigan     0.0925         32.3         19.3         55.1
##  8 florida      0.0379         28.4         16.1         47.4
##  9 texas        0.0276         33           15.7         45.9
## 10 connect~     0.0847         26.9         12.7         56.4
## # ... with 7 more variables: `Mental Health` <int>, `Low Mortality
## #   Rate` <int>, `Low Smoking Rate` <int>, `Low Infant Mortality
## #   Rate` <int>, `Low Obesity Rate` <int>, `Low Suicide Rate` <int>,
## #   average <dbl>
```

County Coronavirus Data

The second half of my project required coronavirus data on a county level. Due to time constraints I could not do county level data from the entire country. Instead I chose to focus on New York and New Jersey because of they offer a mix of dense and sparsely populated counties as well as being the two most infected states in America. I scraped New York county coronavirus data from https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_New_York_(state) and New Jersey county level data from https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_New_Jersey. Both state datasets required basic cleaning. The New York dataset did not include data from each of the five boroughs in New York City and instead combined them into one total. I had to separately scrape coronavirus data for each of the five boroughs from https://projects.thecity.nyc/2020_03_covid-19-tracker/. I then changed the name of each of the boroughs to their county names and added them to the previous table. (Note: the Wikipedia article I used to scrape NY county coronavirus data was updated to include NYC counties after I wrote this. I left it in the report because it was a step I took, but I took the code out because it was not necessary anymore.)

I then had to scrape population density data for New York (https://en.wikipedia.org/wiki/List_of_counties_in_New_York) and New Jersey (https://en.wikipedia.org/wiki/List_of_counties_in_New_Jersey) counties. Once again, the newly scraped data required cleaning as before. These density tables also required one additional step of cleaning. While the previous coronavirus county level data contained only the name of the county, the density data contained the county name as well as the word 'county' (ex. Middlesex vs. Middlesex County). This was an easy fix; I used str_replace

to ensure each table had the same county name format. In order to make map plots of the data using choropleth on the county level, each county needs to have a unique county code. Fortunately, there is a table including this information that can easily be joined by county name. Starting with the county.regions table including each county's unique code, I filtered by state (either New York or New Jersey) and then joined the data. This gave two tables, one for New York and one for New Jersey including county codes, county names, cases, deaths, death rate, and population density. Finally, I was able to combine the two state's county level data into one table seen below.

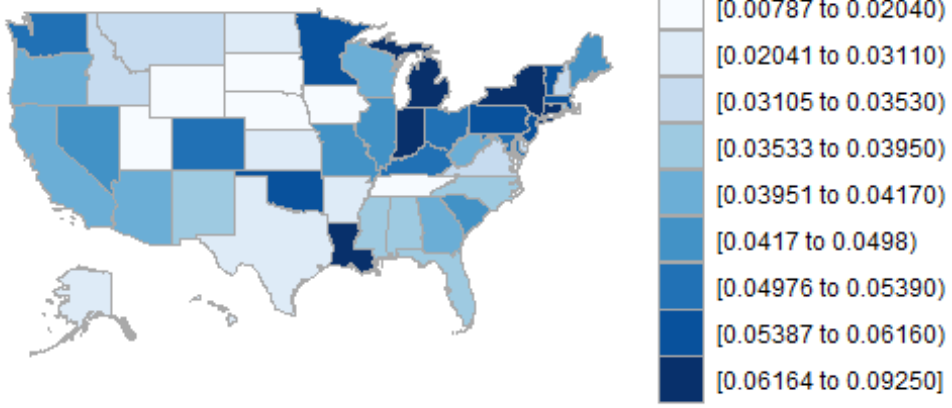
```
##    region county.name Cases Deaths death_rate  Density
## 1   36001      albany  1258     36 0.02861685   570.74
## 2   36003    allegany    35      2 0.05714286    47.34
## 3   36005       bronx 38916   2784 0.07153870 24118.20
## 4   36007      broome   328     22 0.06707317   280.56
## 5   36009 cattaraugus    53      2 0.03773585    61.31
## 6   36011      cayuga    52      1 0.01923077    92.62
## 7   36013  chautauqua    36      1 0.02777778    89.94
## 8   36015     chemung   125      1 0.00800000   216.23
## 9   36017    chenango    99      1 0.01010101    56.16
## 10  36019     clinton    62      4 0.06451613    73.46
```

Results

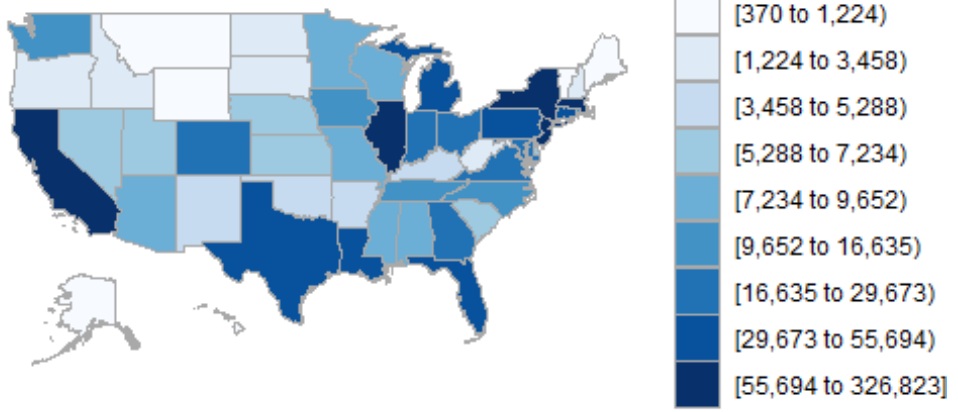
COVID-19 death rate and Health

First, I compared maps of two coronavirus statistics, death rate and total cases, to each of the health statistics, obesity rate, smoking rate, healthcare score, and public health score. Each of the maps are below.

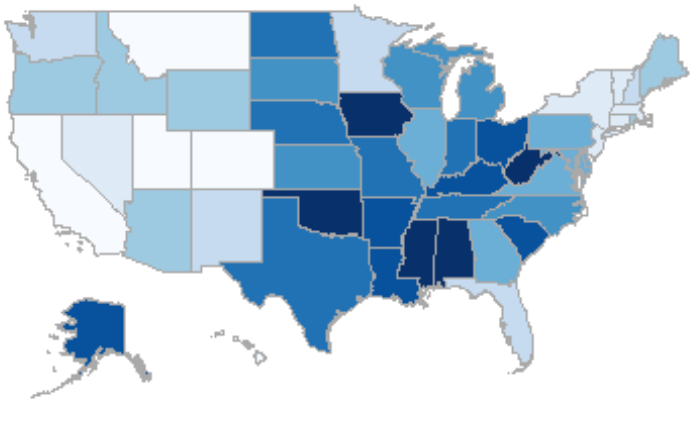
Coronavirus Death Rate by US State



Total Coronavirus Cases by US State



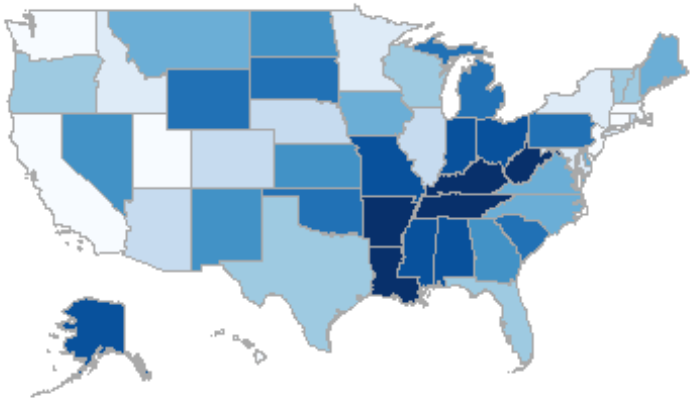
Adult Obesity Rate by US State



Obesity Rate

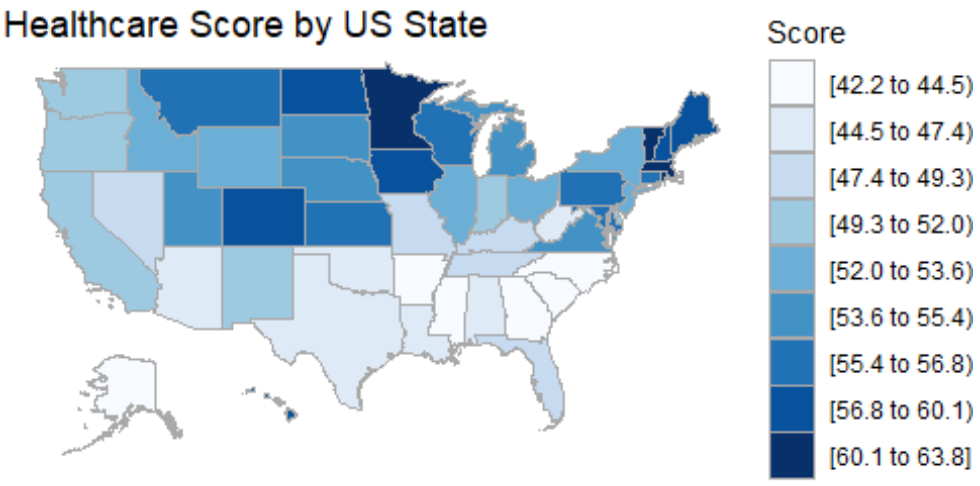
- [22.6 to 25.7)
- [25.7 to 27.7)
- [27.7 to 28.8)
- [28.8 to 30.1)
- [30.1 to 31.9)
- [31.9 to 32.5)
- [32.5 to 33.8)
- [33.8 to 36.3)
- [36.3 to 38.1]

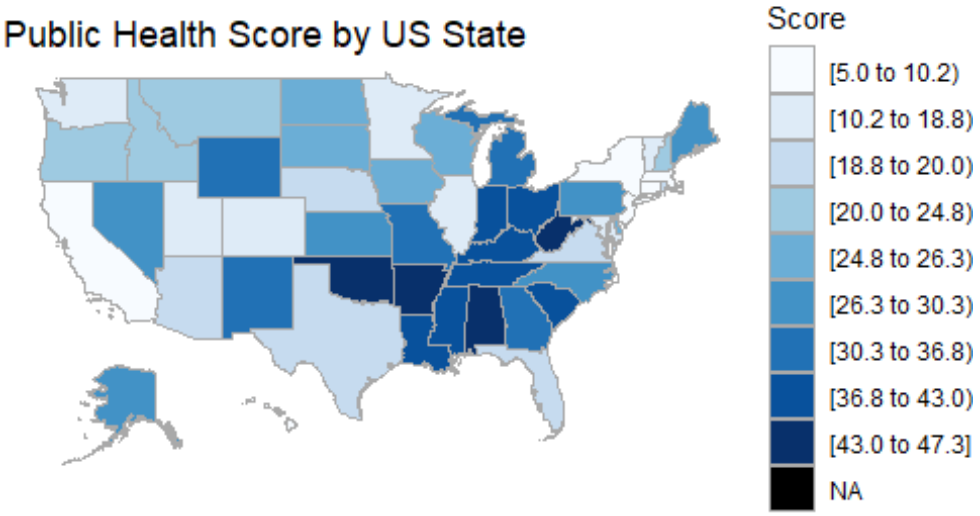
Adult Smoking Rate by US State



Smoking Rate

- [8.9 to 13.8)
- [13.8 to 14.6)
- [14.6 to 15.7)
- [15.7 to 16.4)
- [16.4 to 17.4)
- [17.4 to 18.7)
- [18.7 to 20.8)
- [20.8 to 22.3)
- [22.3 to 26.0]





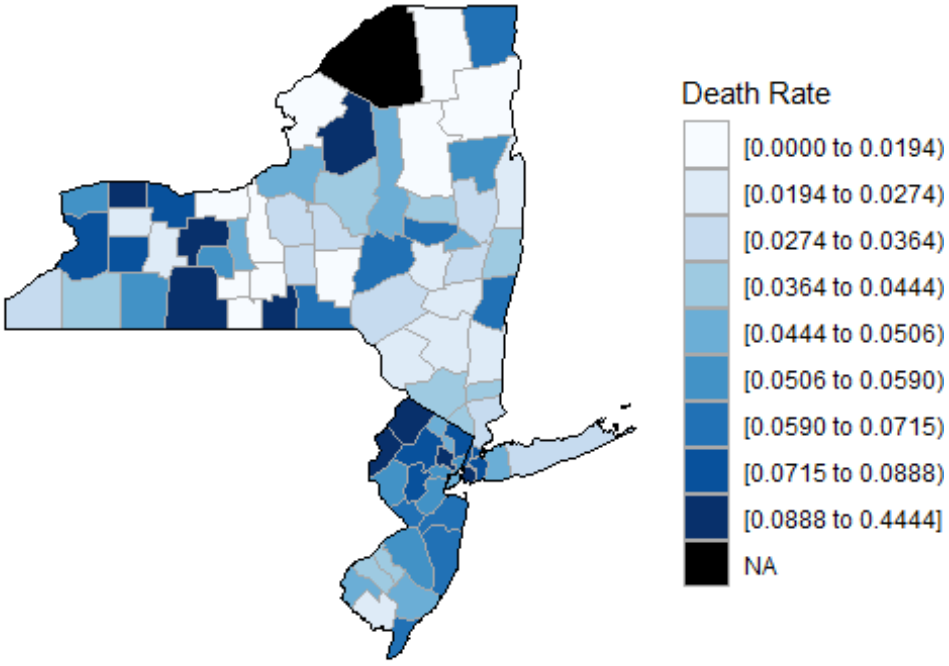
From looking at each of the maps there does not seem to be relationship between states that have high death rates and states that have worse overall health. The four health related maps show a clear trend of worse health statistics in the Southeast (Mississippi,

Alabama, etc.). However, the death rate and total cases related to coronavirus seem to more negatively effect states with better health like New York, Illinois, and California. Notice that these three states have the three most populous cities in the US: New York City, Chicago, and Los Angeles. More urban areas tend to have better health than rural areas in America and coronavirus spreads faster in urban areas than rural area. Coronavirus may be more deadly to people in poor health, but a larger cause of increased death rate may be densely populated areas where hospitals can be overpacked leading to poorer treatment.

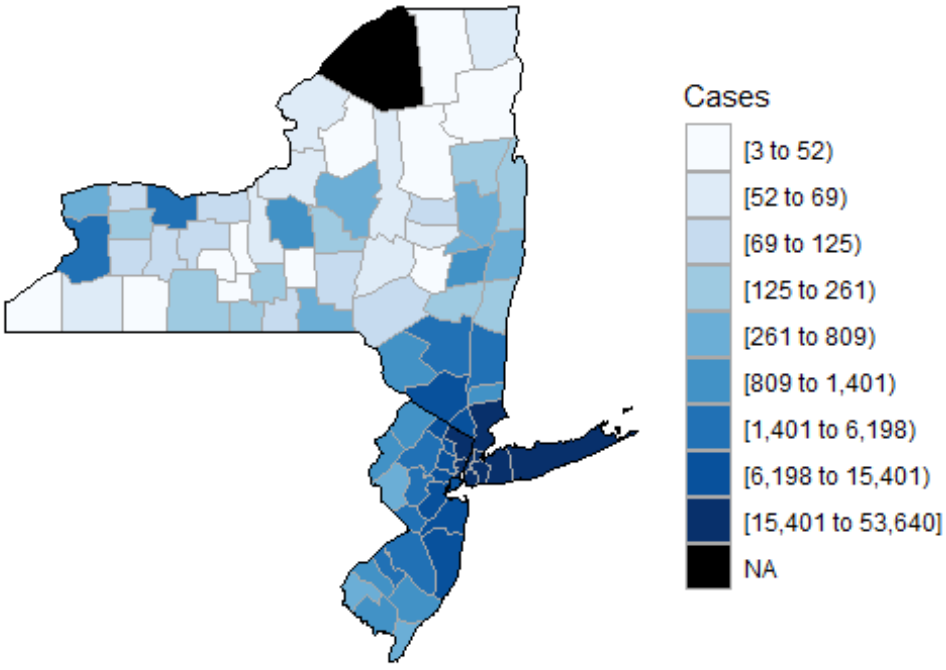
COVID-19 and Population Density

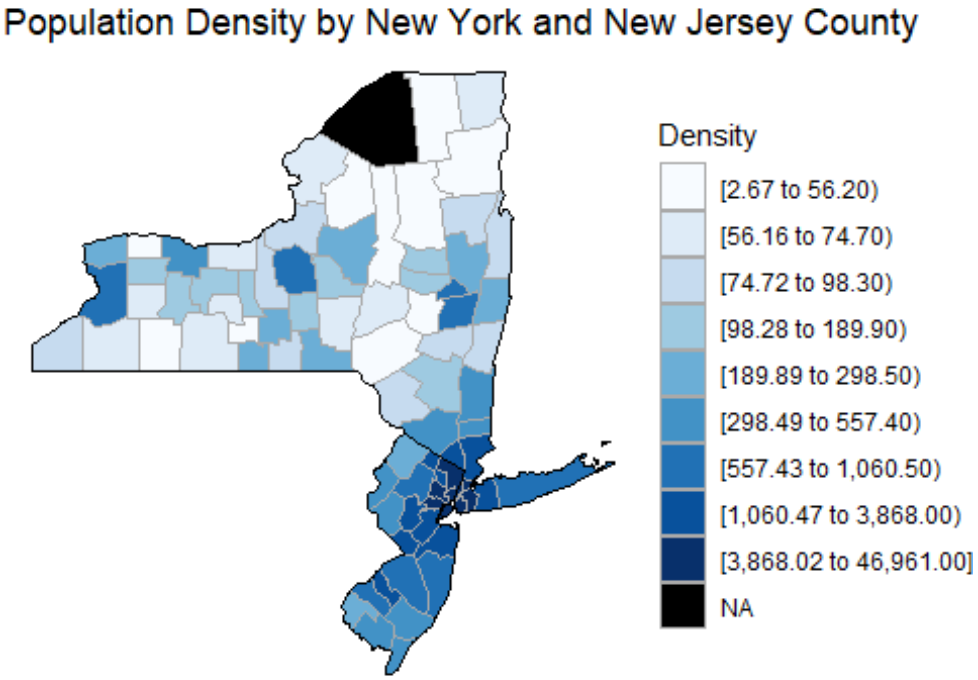
Next, I compared maps of COVID-19 death rate and total cases and population density in New York and New Jersey Counties.

Death Rate by New York and New Jersey County



Total Cases by New York and New Jersey County





The map of total cases and population density shows a clear positive relationship between the two. There is not a clear trend in the death rate map plot, but the death rate is high in several rural counties with low density, while the death rate is moderate for the highly dense counties in and around New York City.

Conclusion

COVID-19 is known to negatively effect individuals in poor health. However, an area's overall health may not be a good indicator of its Covid-19 related death rate. There are obviously many factors that will affect how coronavirus effects a certain area. This is still an ongoing issue that this county is only a few months into. The virus has reached certain states and regions before others. While overall health may not be a good indicator of a state's death rate right now, it may be at the end of this pandemic. I would be interested to revisit this a year or more down the road to see if anything has changed. For now, population density may be a good statistic to use for coronavirus. However, I only used two states to examine their relationship. If I had more time, I think it would be very interesting to look at this relationship throughout all counties of the US.

Overall, this project successfully helped me combine all aspects of course. My project was intensive on web scraping. It has helped me understand and become more proficient in web scrapping. I also had to clean every table that I scraped from the web. The project was also helpful in practice with table manipulation. This was also a good way to work with and become more comfortable with choropleth. The only aspect of this project that I was not able to do was aligning the maps better. I attempted to use grid.arrange to combine the plots, but they overlapped when I tried. The plots and the reports would have been more aesthetically pleasing if I was able to do that.