MATP-4400 COVID-19 Final Notebook

Nicole Schwartzbard

May 4, 2020

Contents

Final Project: Submission Links

This should be the first section of your final project notebook. Fill out the following according to how you submitted your notebook!

- github repository: https://github.com/TheRensselaerIDEA/COVID-Notebooks
- My github ID: *nschwartzbard*
- github issues addressed by this work: `#15`, `#23`
- Github branch name of my submitted notebook: *Nicole_Schwartzbard_final* (example)
- link to merged notebook (post these to LMS!:
  - https://github.com/TheRensselaerIDEA/COVID-Notebooks/blob/master/MATP-4400-FINAL/COVID_FINAL_2020.Rmd (example; Rmd version)
  - https://github.com/TheRensselaerIDEA/COVID-Notebooks/blob/master/MATP-4400-FINAL/COVID_FINAL_2020.html (example; HTML version)

Overview & Problems Tackled

I focused on three main datasets for disparity index visualizations. The first was insurance data. This was for the US for the percentages of people uninsured, insured by Medicare, and insured by Medicaid. For raw percentages: Texas, New Mexico, and West Virginia had significantly higher percentages of people for each respective category. When looked at through Disparity Index, the differences were still apparent but less striking for those insured by Medicare. In terms of those insured by Medicaid for New York, by disparity index, the Bronx has significantly higher than the other counties.

The second problem I tackled was looking at obesity and access to healthy food in New York. For obesity there is a clear difference between downstate and upstate, with upstate having more obese adults, however the differences are not very large. For access for healthy food, there are large differences across New York, with NYC having low levels.

The third problem I tackled was studying asthma in New York. NYC, especially the Bronx, have significantly higher levels of asthma then other counties.

Data Description

Insurance data for the states is from *uninsured_by_state.csv* which was in the Github Repository, Dr. Erickson says it was from the Kaiser Family Foundation. The data composed of each state and the US. I used the percentages provided for Medicaid, Medicare, and uninsured to then calculate the disparity index. Note, that there was an option the states had to expand mediciad which some states adopted and some didn't, I did not separate the data into those two groups, but that is an option for future work.

Insurance data for New York is from www.health.ny.gov and is the March 2020 report for those enrolled in Medicaid, full link: https://www.health.ny.gov/health_care/managed_care/reports/enrollment/monthly/2020/docs/en03_20.pdf, the data is stored in *Medicaid_users.csv*. This Medicaid was split into parts for each county but I used the "Totals" row since I was looking at Medicaid overall not the specific options in New York, that is an option for future work. I then used the "Total Enrolled" to get the number of people, and used the population data from NY.data to calculate the percentages for each county.

Obesity data for New York was from the 2020 New York Data from www.countyhealthrankings.org, the data is included in the *more_county_data.csv*. The column I used was Adults with Obesity, this gave me the percentage. Limited Access to Healthy food was from the same data, with the column Limited Access to Healthy Foods.

Asthma data was from health.ny.gov, full link: https://health.ny.gov/statistics/ny_asthma/data/2016eh/a10.htm. The data is included in the *asthma.csv* data file. This data is asthma hospitalization rate per 10,000 from 2016, although it was reported in 2017. It was the best data I could find, even though it is three-four years old. There was a crude rate, and an adjusted rate based on the 2000 U.S. population. I chose to look at adjusted.

Results

Problem 1: Insurance

The question of insurance is a complicated one in the US. Many people are covered by their employeer, part of the benefits. However, for workers who work hourly jobs instead of jobs with a salary may not have those benefits. The main options that people have who are not ensured by their employer is Medicaid and Medicare. To be eligible for Medicare you must either be older than 65 or have some sort of disability, you also need to have paid taxes for medicare, which is between 1-2% of income. To be eligible for Medicaid, it is determined by the income and size of the family. It is intended for low-income people. Some states expanded Medicaid which makes it easier to qualify for Medicaid. Determining why some people are uninsured is more difficult. From the Kaiser Family foundation, https://www.kff.org/uninsured/issue-brief/key-facts-about-the-uninsured-population/, most people who are uninsured are low-income people of colors families who say the high cost of insurance is the main reason. Some of these people may not be aware of the Medicaid and Medicare options. Also, undocumented immigrants are not eligible for Medicaid or Medicare.

Thus, Medicaid, Medicare, and uninsured percentages can give insights into distribution of low-income families, families with disabilities, potentially undocumented immigrants, and so on. These are important groups who are being affected by Covid-19, thus studying the differences can lead to discovering issues in the containment of Covid-19.

Methods

First I made map visualizations of the raw percentage of people uninsured, insured by Medicaid and insured by Medicaid. For disparity indices, for a difference to be significant, it must be a difference of 0.2 or more. Thus, I used bin sizes of that size. If the differences had been larger, the traditional bin sizes from COVIDminder could have been used, however I wanted to provide as much detail as possible while the size bins retained value.

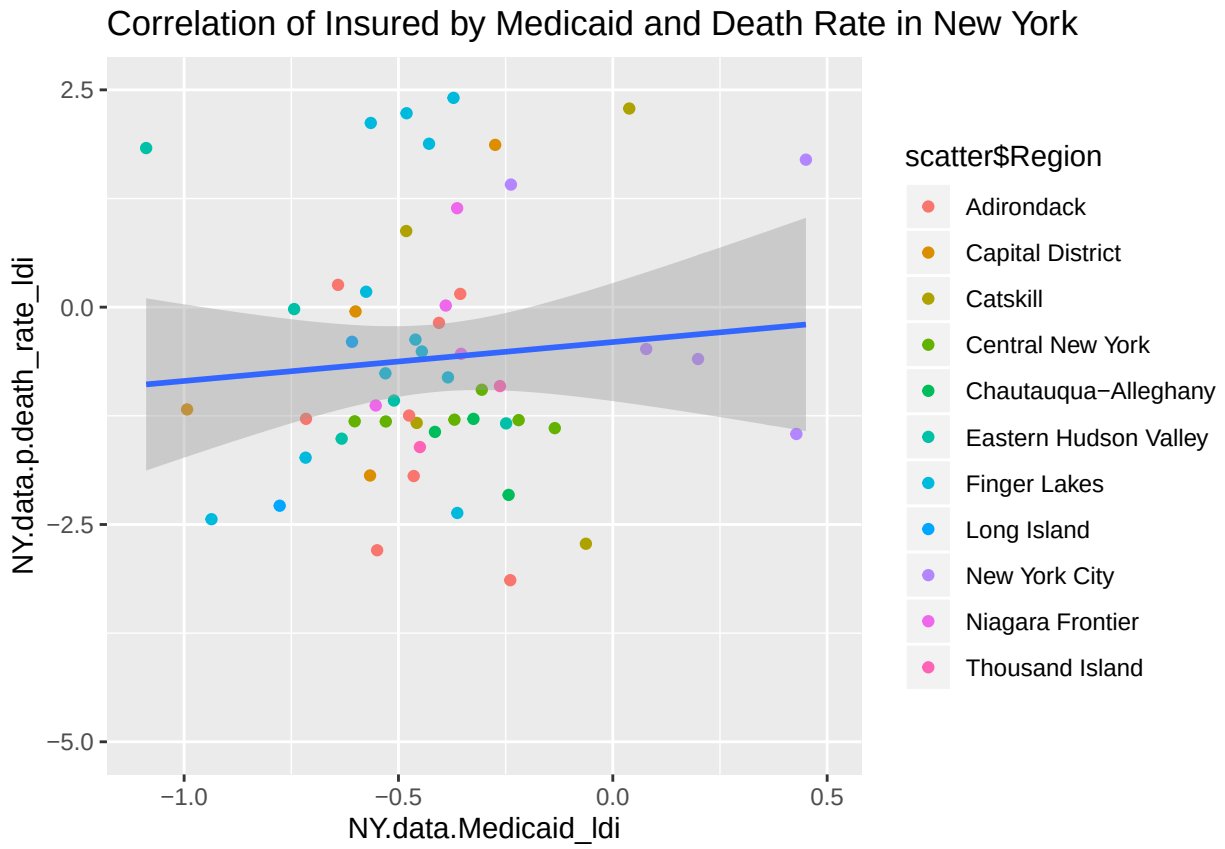
Then I remade the maps with a disparity index. I used the disparity index that COVIDminder uses, $ln(x/y)$ where x is the data from the part and y is the whole. So for US data, x is the state data and y is the total

US percentage. Then for New York, x is the county data and y is the New York state data. See branch *feature-15* for the code.

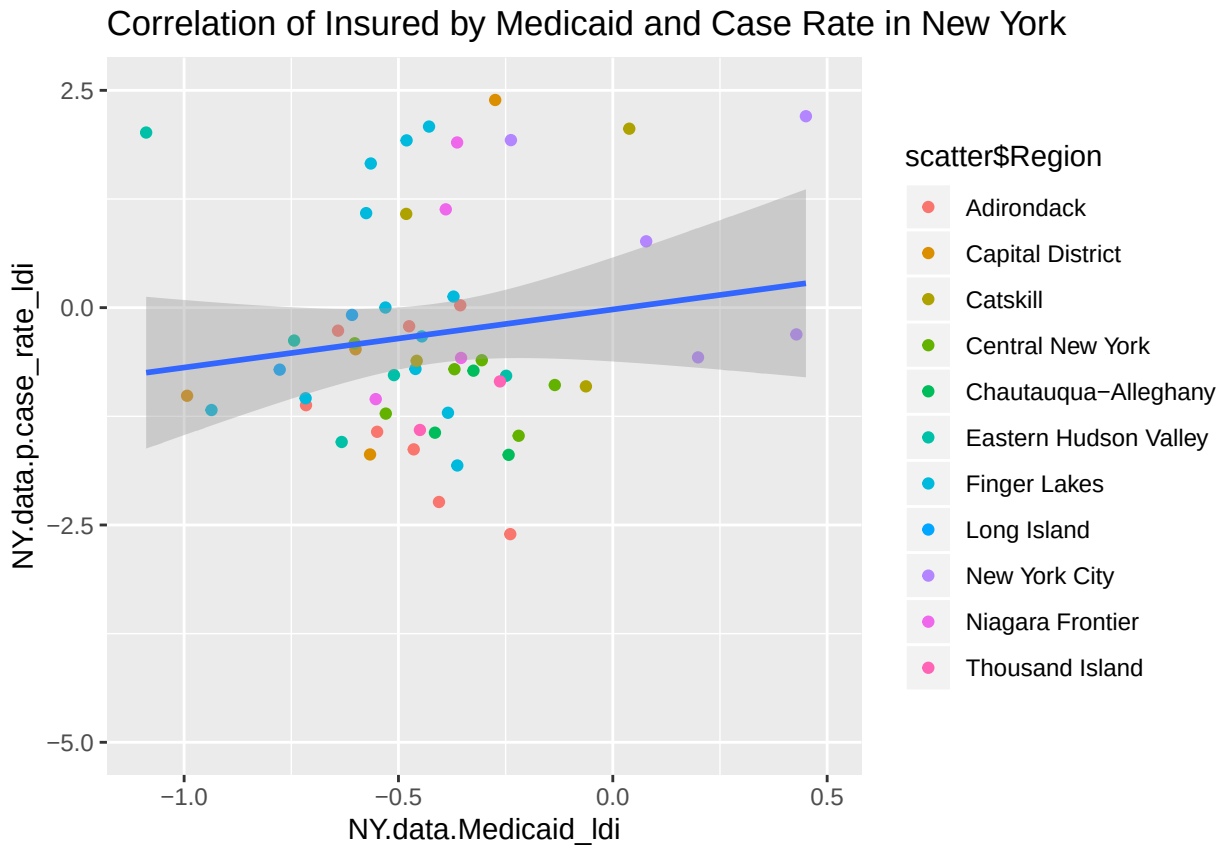
Finally, I wanted to see if there was any correlations for New York state, for the disparity index for insured by Medicaid and mortality rate from Covid-19.

```
#Get Uninsured data and format column names correctly
Uninsured = read.csv('../data/csv/uninsured_by_state.csv', row.names = NULL)
colnames(Uninsured) = c("NAME", "Status", "Employer", "Non-group", "Medicaid", "Medicare", "Military",

#Get America's Average Medicare
Avg = Uninsured[Uninsured$NAME == "United States",]$Medicare

#match order with states in state data so that the state name is accurate on the map
Uninsured = Uninsured[match(states$NAME, Uninsured$NAME),]

#read in data
Medicaid_users = read.csv('../data/csv/ny_Medicaid_03_20.csv')

#head(Medicaid_users)

#get just totals
Medicaid_users = Medicaid_users[Medicaid_users$Plan.Name == "TOTALS:",]

#match data to NY.data order
Medicaid_users = Medicaid_users[match(NY.data$County, Medicaid_users$County), ]

Medicaid_users$Xnum = as.numeric(paste(Medicaid_users$X))
Medicaid_users$County = as.character(paste(Medicaid_users$County))
#Calculate percentages of Medicaid users

Medicaid_users$percent = Medicaid_users$Xnum/NY.data$Population

#Calculate Disparity Index
Avg = Uninsured[Uninsured$NAME == "New York",]$Medicaid

NY.data$Medicaid_ldi = log(Medicaid_users$percent/Avg)

scatter = data.frame(NY.data$County, NY.data.p$death_rate_ldi, NY.data.p$case_rate_ldi, NY.data$Medicai


#get regions
NY_counties_regions = read.csv('../data/csv/NY_counties_regions.csv')

NY_counties_regions = NY_counties_regions[match(NY.data$County, NY_counties_regions$County),]

scatter$Region = NY_counties_regions$Region

#Remove Inf and -Inf for plots,
scatter = scatter[scatter$NY.data.p.death_rate_ldi > -5,]
scatter = scatter[scatter$NY.data.p.death_rate_ldi < 5,]
scatter = scatter[scatter$NY.data.p.case_rate_ldi > -5,]
scatter = scatter[scatter$NY.data.p.case_rate_ldi < 5,]

#Plot graphs
```

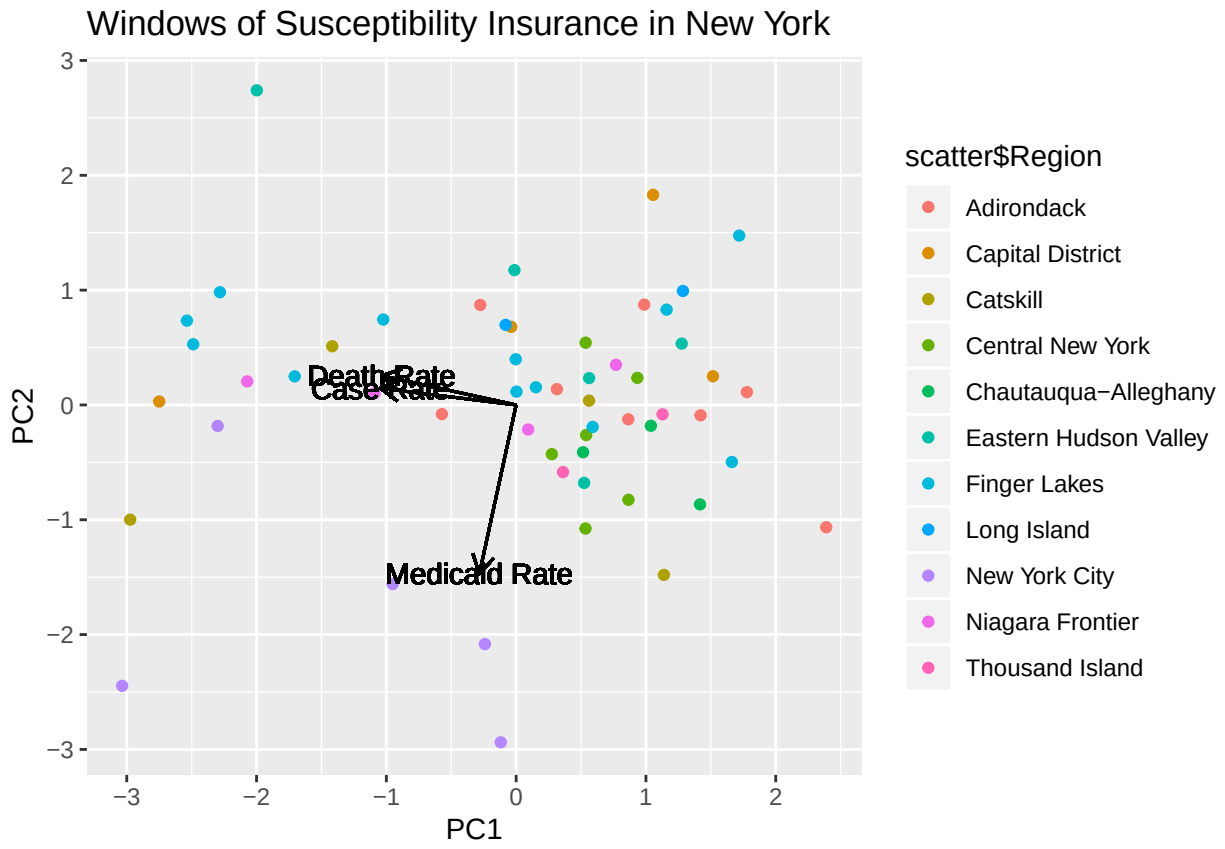
```
ggplot(scatter, aes(x = NY.data.Medicaid_ldi, y = NY.data.p.death_rate_ldi)) + geom_point(aes(color = s
```




```
ggplot(scatter, aes(x = NY.data.Medicaid_ldi, y = NY.data.p.case_rate_ldi)) + geom_point(aes(color = sca
```




```
#PCA analysis
insurance_pca = prcomp(as.matrix(scatter[,2:4]), retx=TRUE, center=TRUE, scale=TRUE)

#Add regions to projected data
plot.df = cbind.data.frame(insurance_pca$x, scatter[,4])

#scaling factor for arrows in biplot graph
s = 1.5

ggplot(plot.df, aes(PC1, PC2)) + geom_point(aes(color = scatter$Region)) + ggtitle('Windows of Susceptil
  geom_segment(x = 0, y = 0, xend = s*insurance_pca$rotation[1, 1], yend = s*insurance_pca$rotation[1, 1
  geom_segment(x = 0, y = 0, xend = s*insurance_pca$rotation[2, 1], yend = s*insurance_pca$rotation[2, 1
  geom_segment(x = 0, y = 0, xend = s*insurance_pca$rotation[3, 1], yend = s*insurance_pca$rotation[3, 1
  geom_text(x = s*insurance_pca$rotation[1, 1], y = s*insurance_pca$rotation[1, 2], label = "Death Rate
  geom_text(x = s*insurance_pca$rotation[2, 1], y = s*insurance_pca$rotation[2, 2], label = "Case Rate"
  geom_text(x = s*insurance_pca$rotation[3, 1], y = s*insurance_pca$rotation[3, 2], label = "Medicaid Ra
```



Results

The full results are in the *feature-15* branch for the interactive maps. I've included some clips here for ease of viewing.

Discussion

First looking at the Uninsured Disparity Across the US. The Midwest and Northeast have less people uninured. Texas by far has the most followed by Georgia and Oklahoma. This does not correlate with the Covid-19 pandemic, which is good, however it is possible that the people who are uninsured would not go to the hospital or report if they had the virus.

Turning to look at the Medicaid Disparity across the US, Now New Mexico is the worst, followed by Louisiana, New York, and others. For Louisiana and New York, these are some of the states with the highest mortality and case rate for the US. Thus, this will be a high cost for the government to pay for this care. It will take longer for this cost to be recognized then the other immediate economic effects of Covid-19.

For Medicaid disparity across New York, New York has significantly higher rates, especially New York and the Bronx. Unfortunately I could not find data for those uninsured and insured by Medicare for New York. Thus, that limits the value of this visualization. This data appears to look closer to the case rate and mortality rate for New York, hence I looked at the correlation. The graphs are shown above colored based on the regions in New York. There is not a strong correlation for either, there is a slight upward positive relationship. New York City is somewhat of a cluster, as is the Finger Lakes for those Insured by Medicaid. But other than that the regions are not clusters.

Looking at the biplot, naturally the higher case rate and mortality rate have some clusters, with five counties that have high case rate and death rate. However looking at the projection for the Medicaid rate, there are counties with high Medicaid rate and low Medicaid rate so there is not additional correlations.

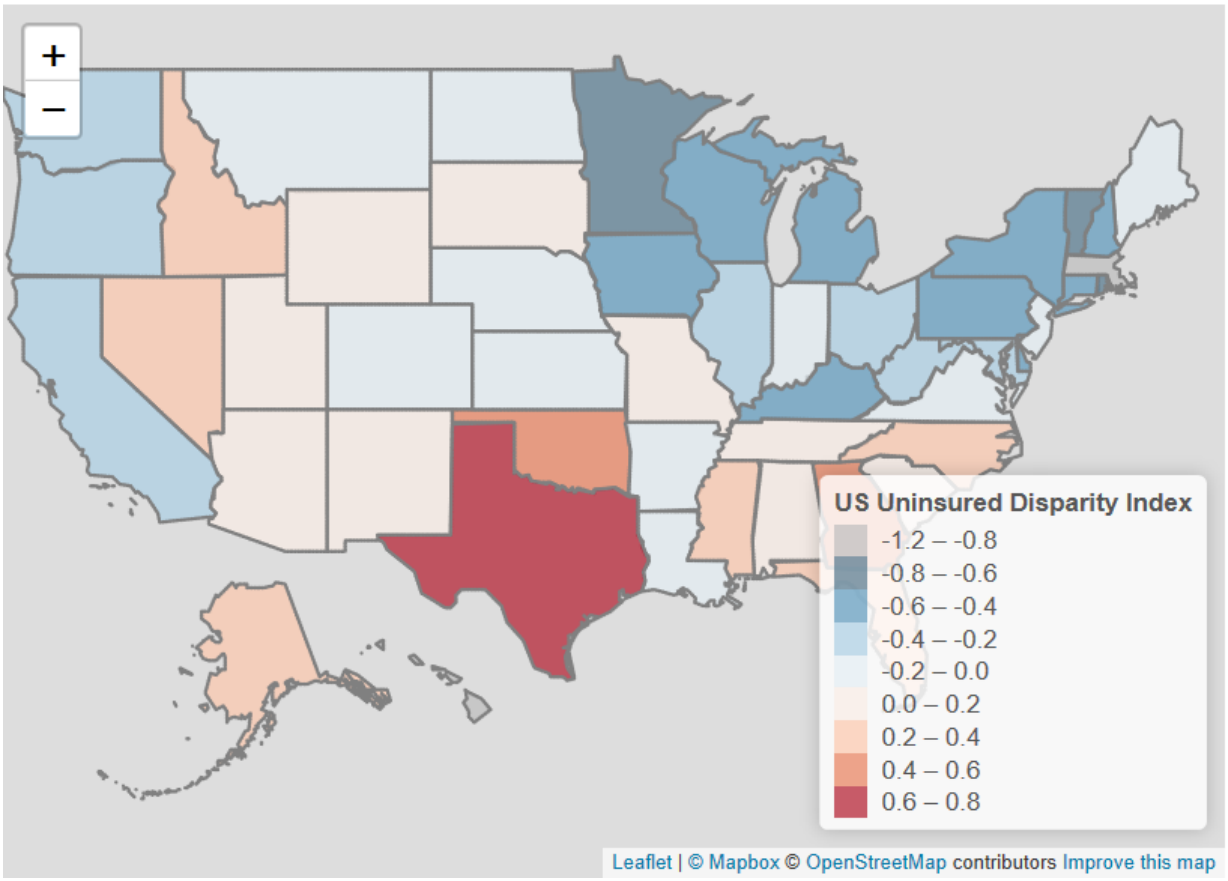


Figure 1: Uninsured Disparity Across the US

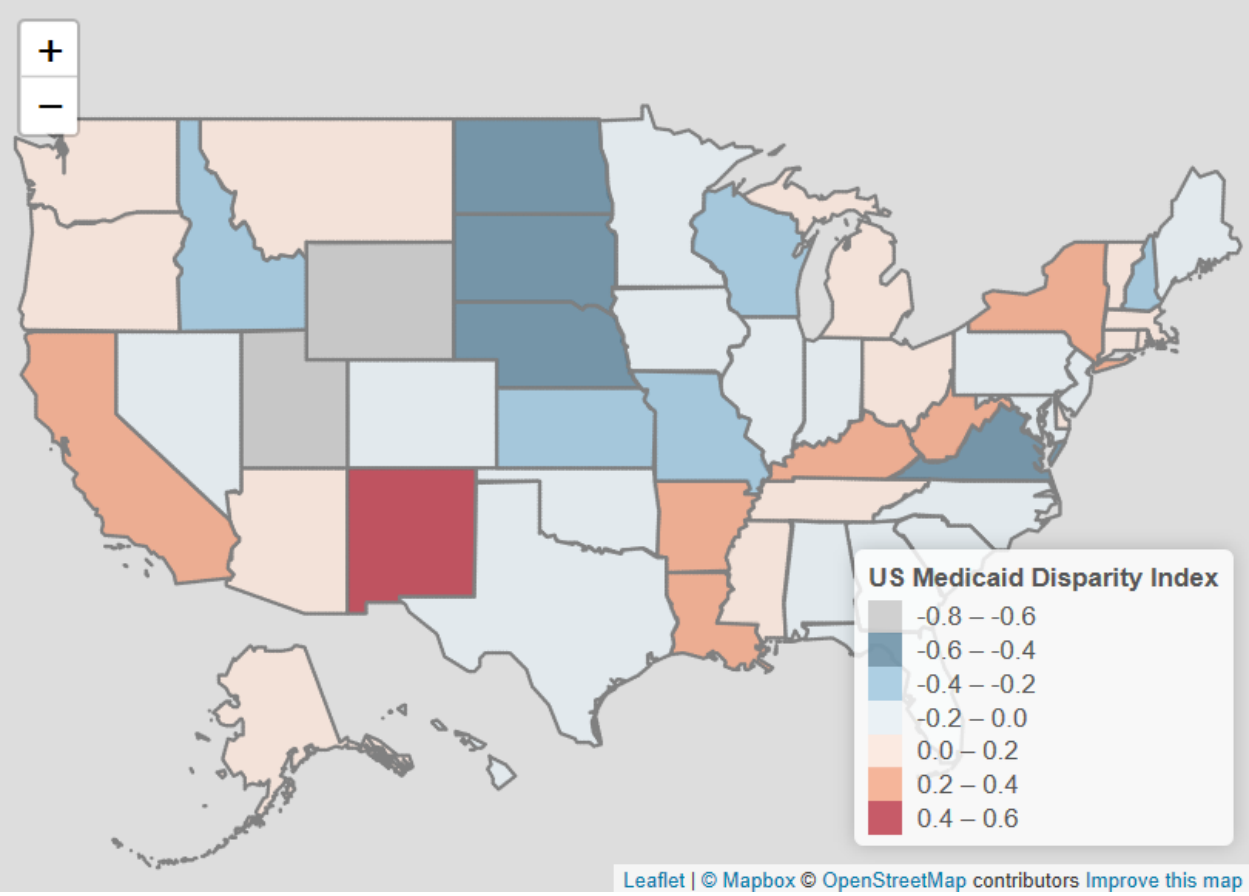


Figure 2: Medicaid Disparity Across the US

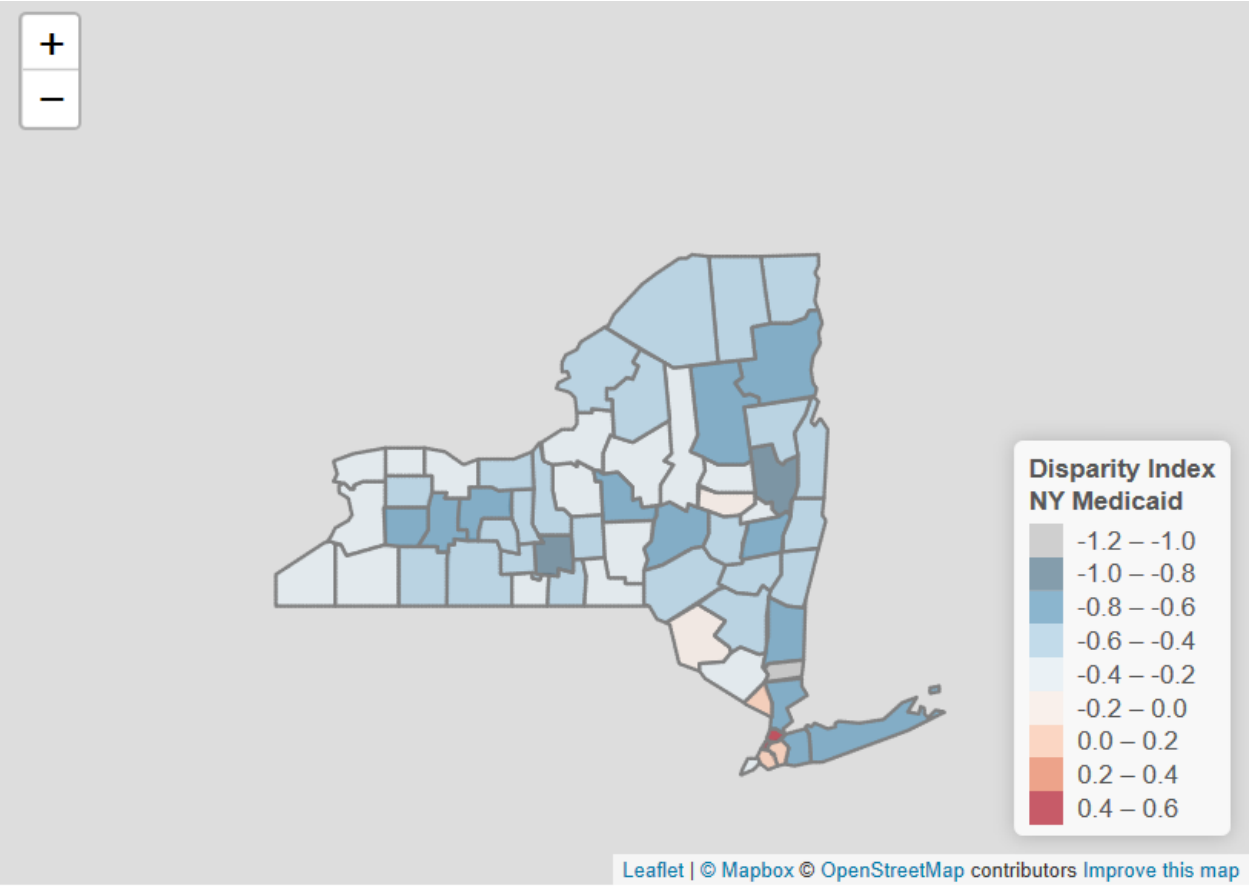


Figure 3: Medicaid Disparity Across New York

Problem 2: Obesity

The CDC believes that those who have severe obesity are at higher risks from complications from Covid-19. Since Covid-19 can cause breathing difficulties, those who are obese often already have breathing issues as well as other underlying health conditions that can make care more difficult and problems be exacerbated. There are many causes of obesity, many are genetic however the other parts are generally exercise and healthy food. Depending where you live can greatly impact the access to healthy food, hence why I also studied that. Studying this problem can lead to discoveries to give information to those who are obese and officials to help protect those who are obese.

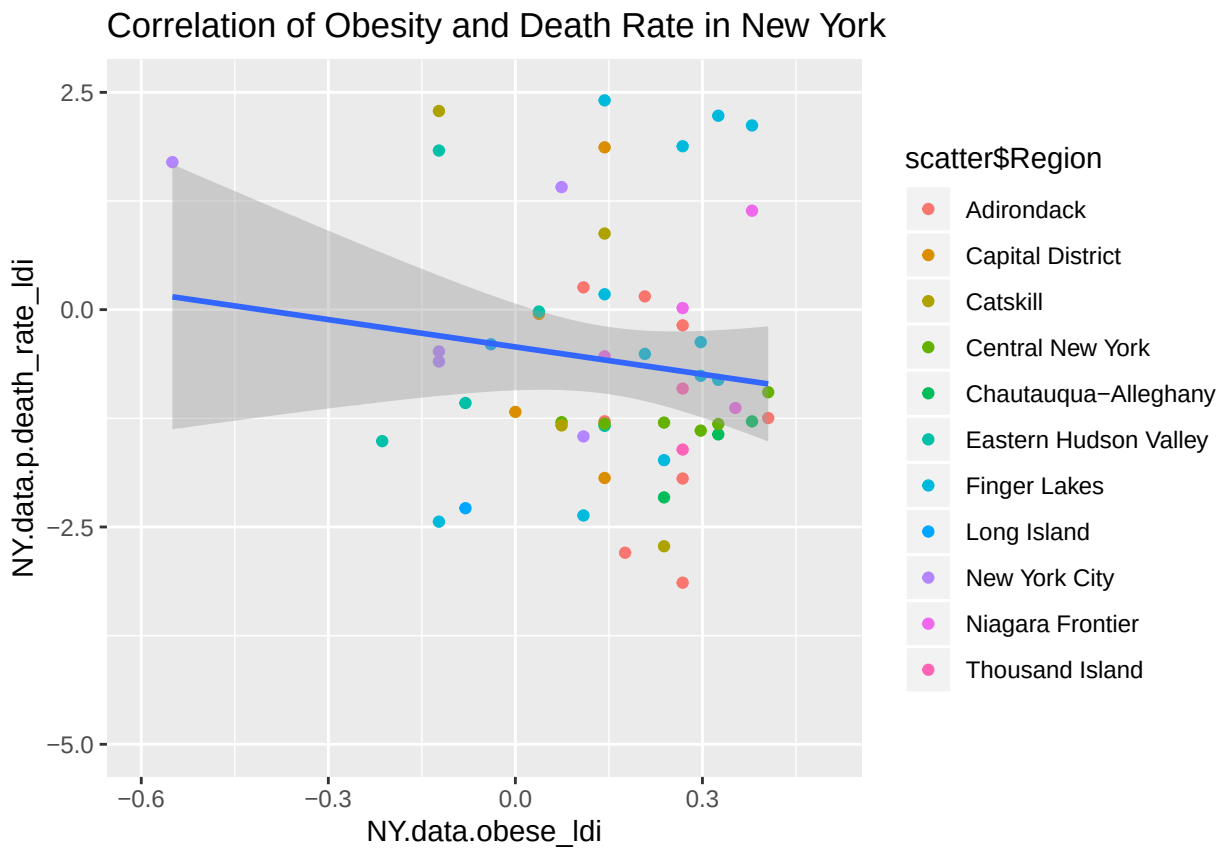
Methods

First I made map visualizations of the raw percentage of people who are obese and healthy food insecure. For disparity indices, for a difference to be significant, it must be a difference of 0.2 or more. Thus, I used bin sizes of that size. If the differences had been larger, the traditional bin sizes from COVIDminder could have been used, however I wanted to provide as much detail as possible while the size bins retained value.

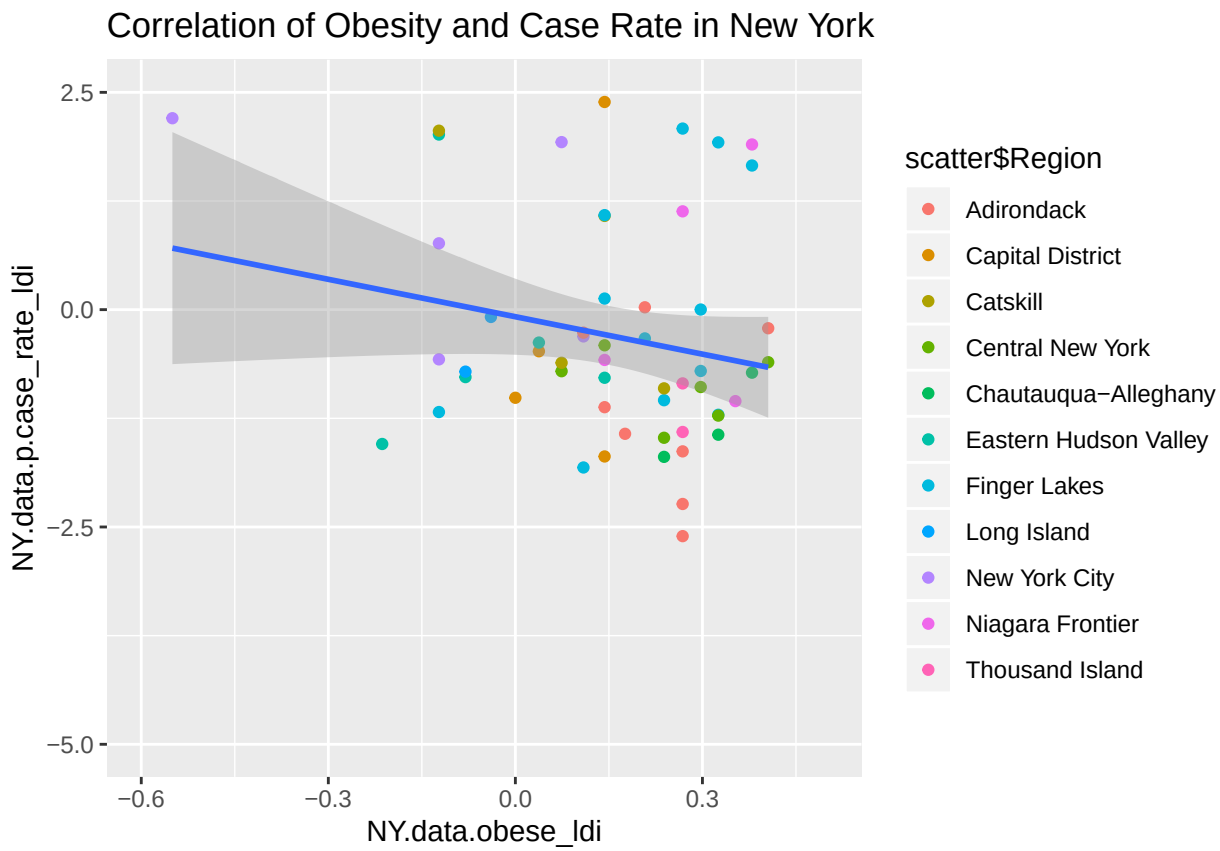
Then I remade the maps with a disparity index. I used the disparity index that COVIDminder uses, $ln(x/y)$ where x is the data from the part and y is the whole.x is the county data and y is the New York state data. See branch *feature-23* for the code.

Finally, I wanted to see if there was any correlations for New York state, for the disparity index for obesity and mortality rate from Covid-19.

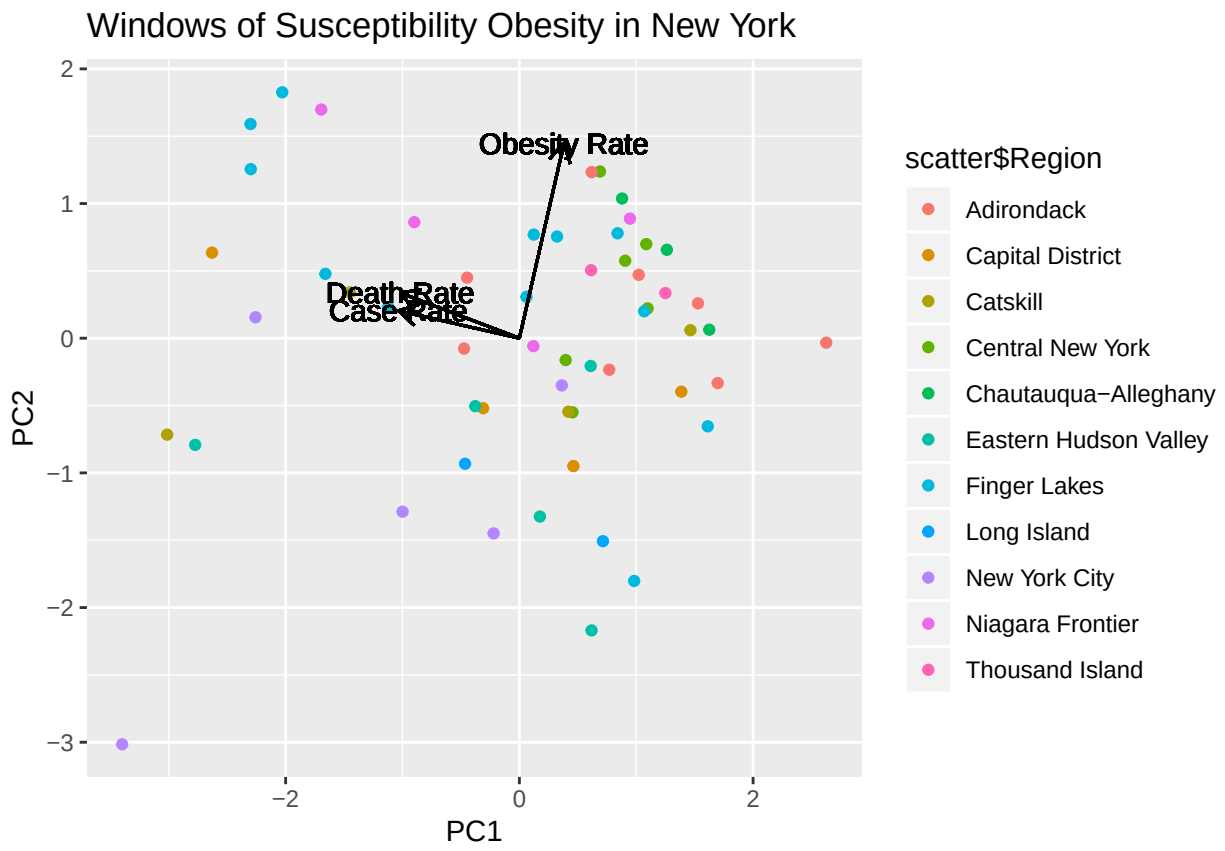
```
#set up obesity data
obese = read.csv('../data/csv/more_county_data.csv')
obese$percent = as.numeric(paste(obese$X..Adults.with.Obesity))

Avg = obese[obese$County == "",]$percent

obese = obese[match(NY.data$County, obese$County), ]


NY.data$obese_ldi = log(obese$percent/Avg)

scatter = data.frame(NY.data$County, NY.data.p$death_rate_ldi, NY.data.p$case_rate_ldi, NY.data$obese_l


#get regions
NY_counties_regions = read.csv('../data/csv/NY_counties_regions.csv')

NY_counties_regions = NY_counties_regions[match(NY.data$County, NY_counties_regions$County),]

scatter$Region = NY_counties_regions$Region

#Remove Inf and -Inf for plots,
scatter = scatter[scatter$NY.data.p.death_rate_ldi > -5,]
scatter = scatter[scatter$NY.data.p.death_rate_ldi < 5,]
scatter = scatter[scatter$NY.data.p.case_rate_ldi > -5,]
scatter = scatter[scatter$NY.data.p.case_rate_ldi < 5,]

#Plot graphs
ggplot(scatter, aes(x = NY.data.obese_ldi, y = NY.data.p.death_rate_ldi)) + geom_point(aes(color = scat
```




```
ggplot(scatter, aes(x = NY.data.obese_ldi, y = NY.data.p.case_rate_ldi)) + geom_point(aes(color = scatt
```




```
#perform pca for obesity
obesity_pca = prcomp(as.matrix(scatter[,2:4]), retx=TRUE, center=TRUE, scale=TRUE)

plot.df = cbind.data.frame(obesity_pca$x, scatter[,4])
s = 1.5

ggplot(plot.df, aes(PC1, PC2)) + geom_point(aes(color = scatter$Region)) + ggtitle('Windows of Susceptil
  geom_segment(x = 0, y = 0, xend = s*obesity_pca$rotation[1, 1], yend = s*obesity_pca$rotation[1, 2], a
  geom_segment(x = 0, y = 0, xend = s*obesity_pca$rotation[2, 1], yend = s*obesity_pca$rotation[2, 2], a
  geom_segment(x = 0, y = 0, xend = s*obesity_pca$rotation[3, 1], yend = s*obesity_pca$rotation[3, 2], a
  geom_text(x = s*obesity_pca$rotation[1, 1], y = s*obesity_pca$rotation[1, 2], label = "Death Rate") +
  geom_text(x = s*obesity_pca$rotation[2, 1], y = s*obesity_pca$rotation[2, 2], label = "Case Rate") +
  geom_text(x = s*obesity_pca$rotation[3, 1], y = s*obesity_pca$rotation[3, 2], label = "Obesity Rate")
```



Results

Discussion

As shown in the visualization for Adults with Obesity, the Regions, Long Island, New York City, and Hudson Valley have lower levels then the rest of the state, but not by a large margin. In terms of Limited Access to Healthy Food, there does not appear to be a strong correlation between that and adults with obesity which is interesting, indicating that exercise and genetic are perhaps more significant than anticipated. Access to healthy food does not look similar to New York Covid-19 data. Obesity levels look more similar. However when studying the scatterplots looking for correlation, there is not a correlation. Similarly for the biplot, as for above, there is not a correlation for the obesity rate.

Problem 3: Asthma

The CDC believes that moderate-to-severe asthma can put people at higher risks from complications from Covid-19. Again, since Covid-19 can cause breathing difficulties, and asthma makes it harder to breathe. Many people have varying levels of asthma, some genetic and some from health conditions that hvae arose over the course of their life. Studying this problem can lead to discoveries to give information to those who have asthma and officials to help protect those who have asthma.

Methods

I followed the same procedure as above. First I made map visualizations of the raw percentage of people hospitalized from Asthma, for a difference to be significant, it must be a difference of 0.2 or more. Thus, I used bin sizes of that size. If the differences had been larger, the traditional bin sizes from COVIDminder could have been used, however I wanted to provide as much detail as possible while the size bins retained value.

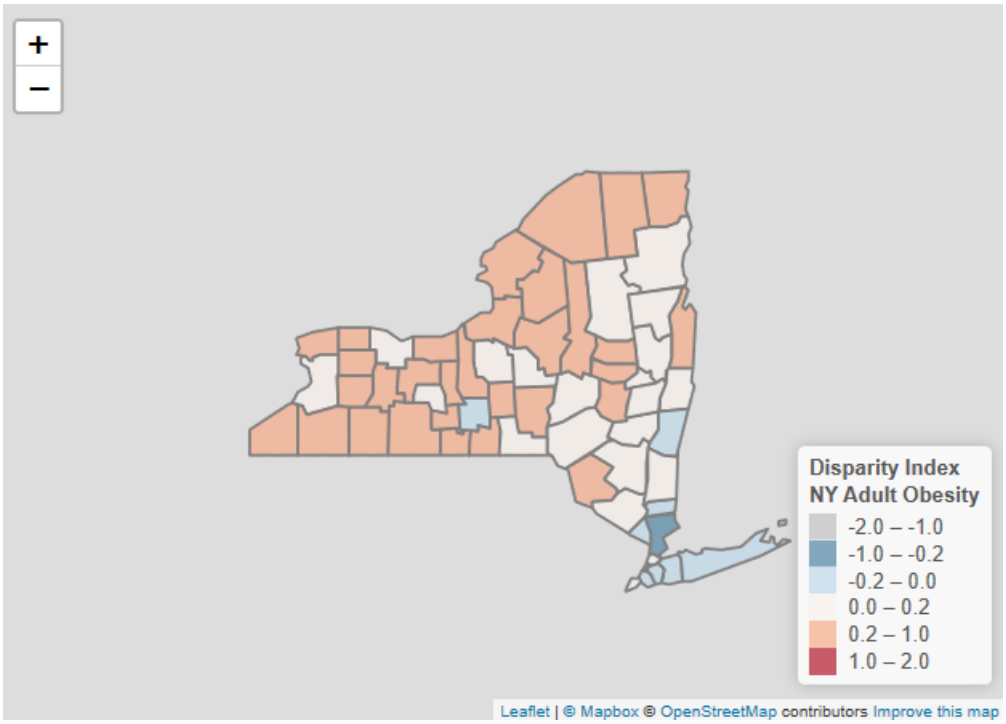


Figure 4: Adults with Obesity Disparity Across New York

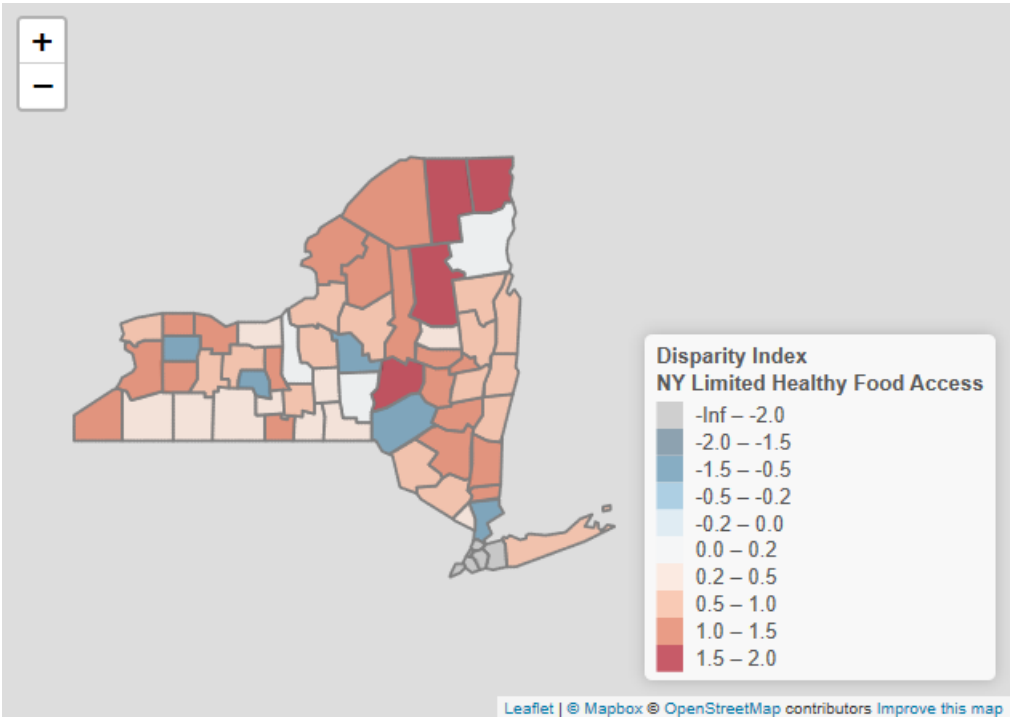


Figure 5: Limited Access to Healthy Food Across New York

Then I remade the maps with a disparity index. I used the disparity index that COVIDminder uses, *ln(x/y)* where x is the data from the part and y is the whole. x is the county data and y is the New York state data. See branch *feature-23* for the code.

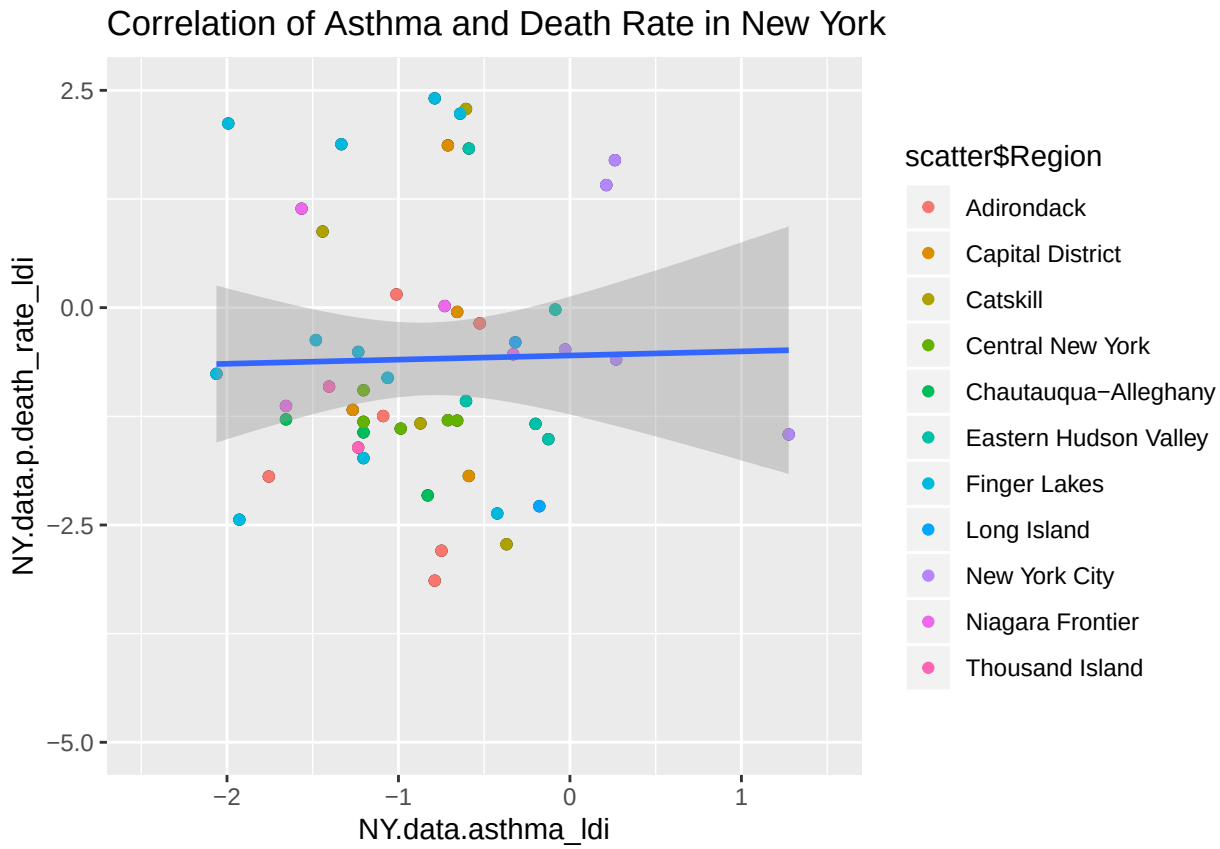
Finally, I wanted to see if there was any correlations for New York state, for the disparity index for asthma hospitalizations and mortality rate from Covid-19.

```
#read in data
asthma = read.csv('../data/csv/asthma.csv')

#get numbers instead of factors
asthma$percent = as.numeric(paste(asthma$Adjusted))
```

```
## Warning: NAs introduced by coercion
```

```
Avg = asthma[asthma$X == "New York State",]$percent

#match to ny.data
asthma = asthma[match(NY.data$County, asthma$X), ]


#caluclate disparity index
NY.data$asthma_ldi = log(asthma$percent/Avg)

scatter = data.frame(NY.data.p$death_rate_ldi, NY.data.p$case_rate_ldi, NY.data$asthma_ldi)

#get regions
NY_counties_regions = read.csv('../data/csv/NY_counties_regions.csv')

NY_counties_regions = NY_counties_regions[match(NY.data$County, NY_counties_regions$County),]

scatter$Region = NY_counties_regions$Region

#remove NA data
scatter = scatter[is.na(scatter$NY.data.asthma_ldi) == FALSE,]
```

```
#Remove Inf and -Inf for plots,
scatter = scatter[scatter$NY.data.p.death_rate_ldi > -5,]
scatter = scatter[scatter$NY.data.p.death_rate_ldi < 5,]
scatter = scatter[scatter$NY.data.p.case_rate_ldi > -5,]
scatter = scatter[scatter$NY.data.p.case_rate_ldi < 5,]

scatter$NY.data.asthma_ldi = ifelse(scatter$NY.data.asthma_ldi == -Inf, -5, ifelse(scatter$NY.data.asthm

ggplot(scatter, aes(x = NY.data.asthma_ldi, y = NY.data.p.death_rate_ldi)) + geom_point() + ggtitle("Co
```

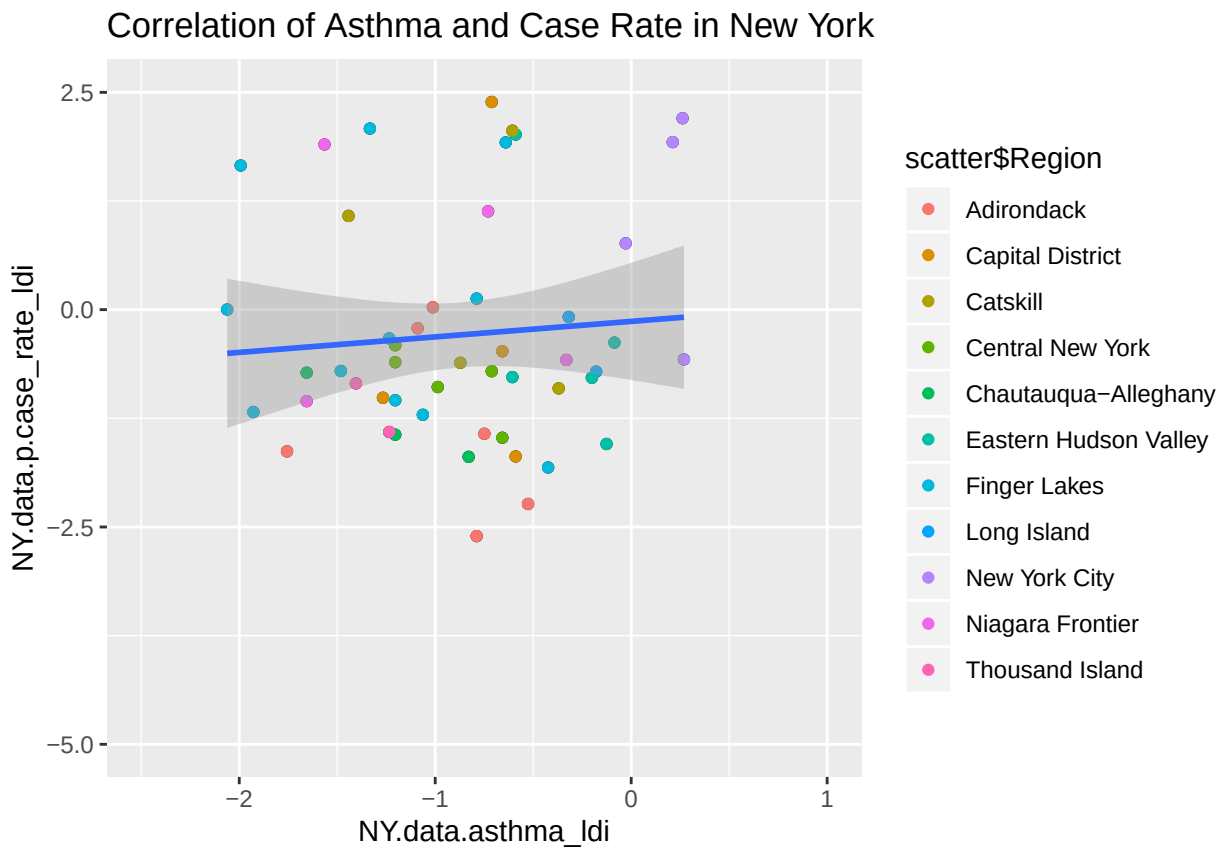



```
ggplot(scatter, aes(x = NY.data.asthma_ldi, y = NY.data.p.case_rate_ldi)) + geom_point() + ggtitle("Cor
```

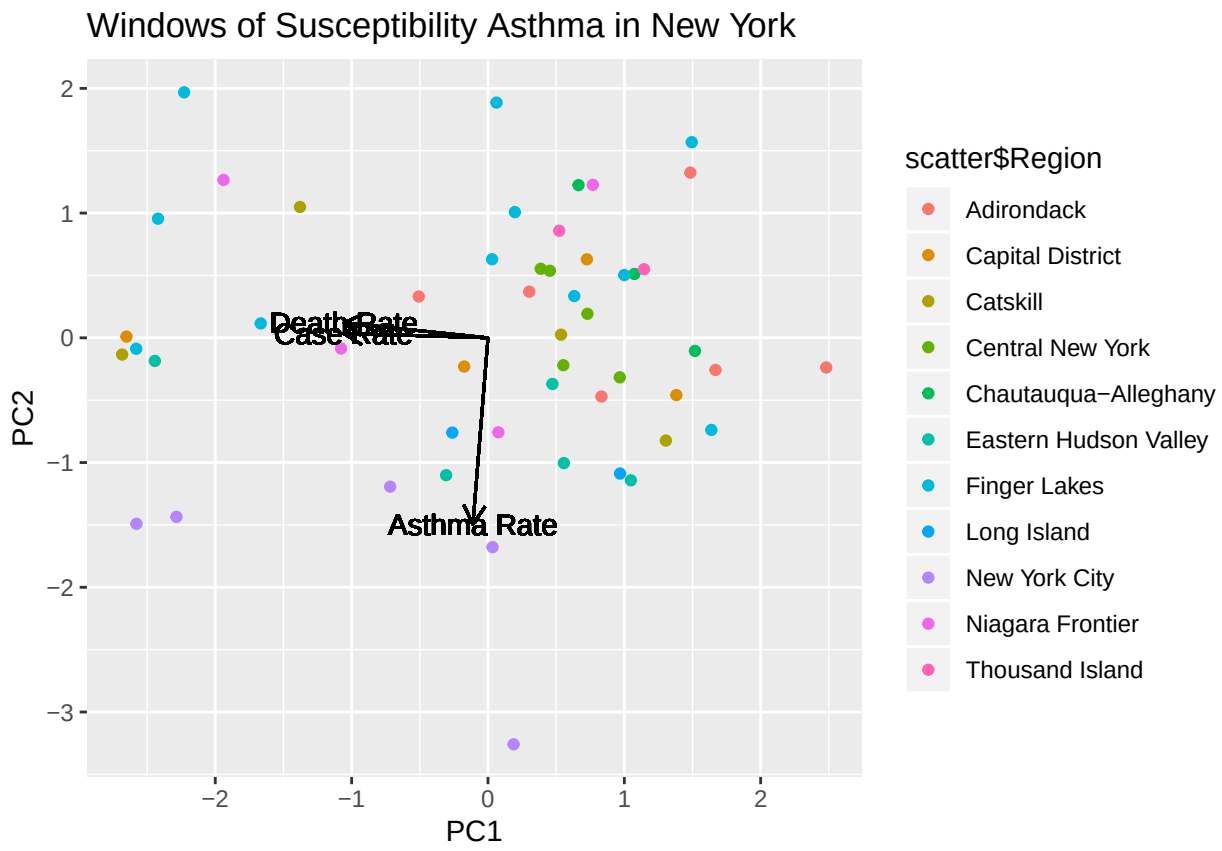
```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
#perform pca analysis
asthma_pca = prcomp(as.matrix(scatter[,1:3]), retx=TRUE, center=TRUE, scale=TRUE)

plot.df = cbind.data.frame(asthma_pca$x, scatter[,4])

s = 1.5

ggplot(plot.df, aes(PC1, PC2)) + geom_point(aes(color = scatter$Region)) + ggtitle('Windows of Susceptil
  geom_segment(x = 0, y = 0, xend = s*asthma_pca$rotation[1, 1], yend = s*asthma_pca$rotation[1, 2], ar
  geom_segment(x = 0, y = 0, xend = s*asthma_pca$rotation[2, 1], yend = s*asthma_pca$rotation[2, 2], ar
  geom_segment(x = 0, y = 0, xend = s*asthma_pca$rotation[3, 1], yend = s*asthma_pca$rotation[3, 2], ar
  geom_text(x = s*asthma_pca$rotation[1, 1], y = s*asthma_pca$rotation[1, 2], label = "Death Rate") +
  geom_text(x = s*asthma_pca$rotation[2, 1], y = s*asthma_pca$rotation[2, 2], label = "Case Rate") +
  geom_text(x = s*asthma_pca$rotation[3, 1], y = s*asthma_pca$rotation[3, 2], label = "Asthma Rate")
```



Results

Discussion

For New York, New York City, especially the Bronx and New York counties have much higher rates of asthma hospitalziations then the rest. However, this data is older, as mentioned above. This is also asthma hospitalizations, which does correlation to moderate-to-severe asthma but it is still not ideal. In terms of correlation, there is none. But, New York City does act like a cluster unlike the rest of the counties. In regards to the biplot, there is a large range for the asthma rate in regards to the case and death rate, so there does not appear to be a correlation there either.

Summary and COVIDMINDER Recommendations

One of the big questions about Covid-19 is why it affects people so differently, so many are asymptomatic and don't even notice, while others must go to the hospital, need a ventilator, and die. One of the theories is that this is due to differences in immune systems as well as other health conditions. Immune systems are difficult to measure, but health conditions such as obesity and asthma, both suggested by the CDC to be potentially impactful are measureable. I've shown how there are significant differences, and where you live does matter. However, I wasn't able to link the results of Covid-19 to these factors. I still think Obesity and Asthma should be added to COVIDminder as determinants since CDC views them as potential determinants.

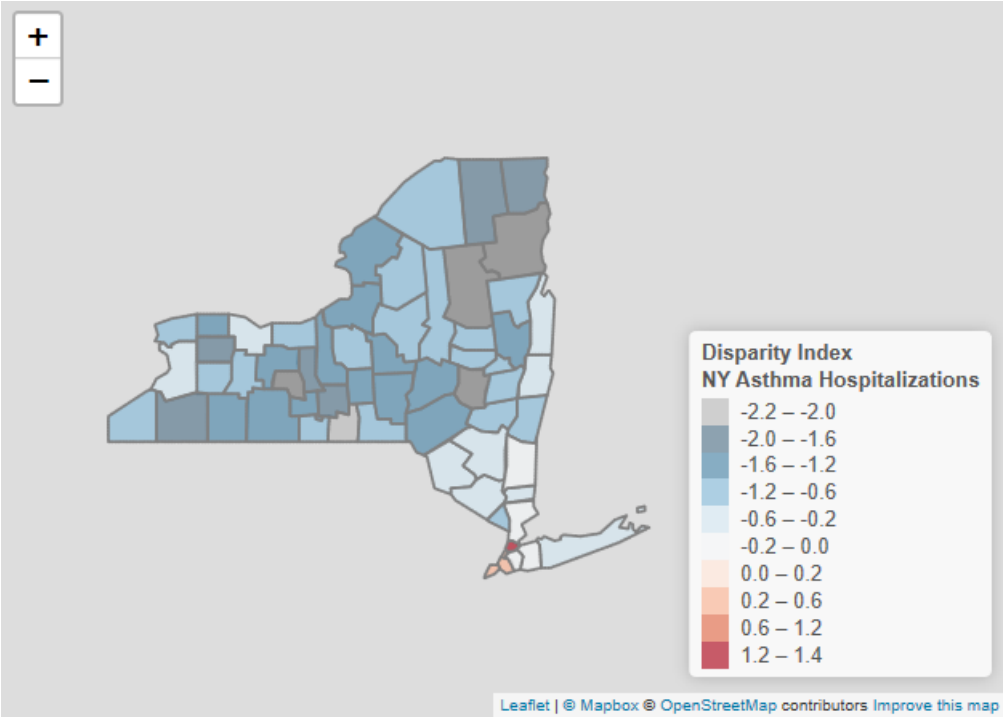


Figure 6: Asthma Hospitalizations Across New York