Data 115 Final: COVID-19 in NYC

2023-12-07

Why did I choose my data set?

For this assignment I really wanted to use something that I found intriguing. Most of my high school was during the COVID-19 pandemic and I have always wondered how effective the vaccine mandate was. After the mandate, it seemed like everything went back to our worlds sense of normal. During the pandemic there was a lot of people who didn't want to get the vaccine due to many reasons. A lot of people did not want to put something in their body that hadn't been tested for very long. I think it is important to note that in the newly stages of the vaccine mandate, it was hard for the government to enforce it. People had various reasons to why they were unwilling to get the vaccine which mainly consisted of religious reasons and concerns on a newly developed vaccine. I have always been curious on how the vaccine affected the death and hospitalization rates of highly populated cities like New York, New York. This led me to the motivating question I wanted to use to analyze my data with. How had the vaccine mandate on December 27, 2021 impacted the number of COVID-19 hospitalization and death counts in New York City.

Data Process

To find my data I decided to look up a data set on data.gov. I though data.gov was a good site to extract a data set from because it is sourced by the federal government. This being said, the data is reliable and more than likely accurate. After finding this data set, I looked at all of the different variables. I decided to depict which ones were relevant enough and could help me and my motivating question. There were 67 variables in which 61 of these were irrelevant to my analysis.

There were a few issues I ran into while trying to create visuals and extracting only certain data. As I mentioned, there were only six of sixty-seven variables that were relevant to my analysis. The way I decided to solve this was creating two separate variables. One variable represented data from February 29, 2020 up until the mandate on December 27, 2021. The other variable represented data from after the mandate on December 28th, 2021 up until the most recent data recorded on May 1, 2023. I then set both of my variables to contain only the variables from rows one to six in my data set. This allowed me to see all the variables I wanted to use. This cleaning helped narrow my data and make it easier to work with. Another issue I ran into with my data was the visuals. The graphs that I made were hard to understand for a reader who wasn't familiar with my data set. The way I fixed this was adding onto my code and adding a color to each of my variables. I made my variable that represented data prior the mandate blue and data that represented data after the mandate red. This made it easier to understand where each variable stopped and started.This also makes my graph easier to read and understand when looking at it. This is an important thing to have in a visual because readers don't want to have to dig for information when they come across a visual. They also want it to make sense as soon as they see it. My graph does a good job outputting the data for a reader to comprehend easily.

After I cleaned up the data set I was working with, it made the process of making my graphs much easier. I started off by plotting my graph in my ggplot function which worked well and made a good graph. Even though this made a good graph I wanted my data to lay on top of each other. To establish this I decided to put it in my geom_point function instead. This was extremely helpful because it made the visualization a lot easier to read and compare the data.

What have I learned?

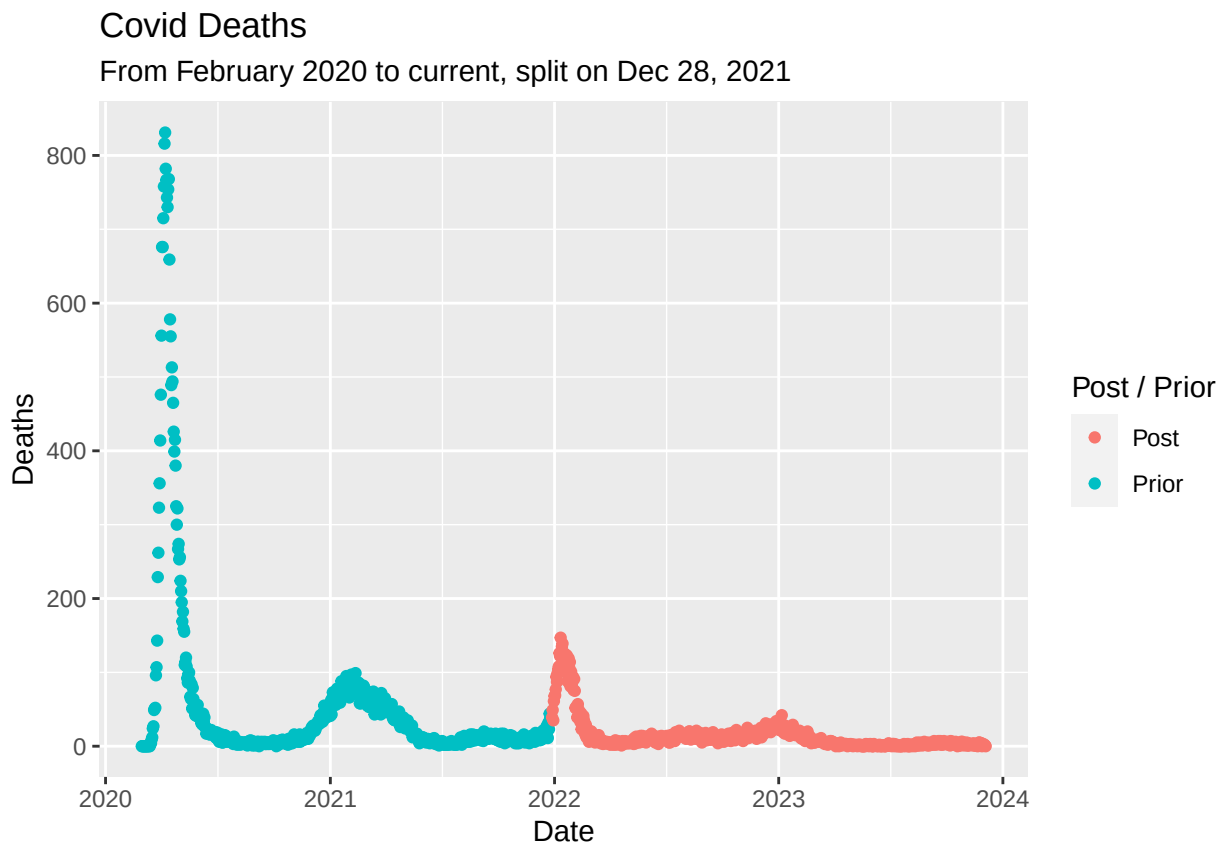
By creating scatter plots with my data and analyzing the trends, it is easily noticeable that the vaccine mandate caused a huge decrease in the amount of COVID-19 hospitalized and death counts in NYC. In both graphs, there is a short timeline of a few days where the numbers continue to increase before they started to decrease. I think it is important to note that it took a little time before the general public had access to the vaccines. It also takes time for a huge difference to be noticeable. Large gatherings also started once the mandate was in place. In a high populated city like New York, there will be a rise in cases while people are still in the process of getting all of their vaccines. This is a big reason why there was a slight increase in the data before it started going down.

Additional Analysis

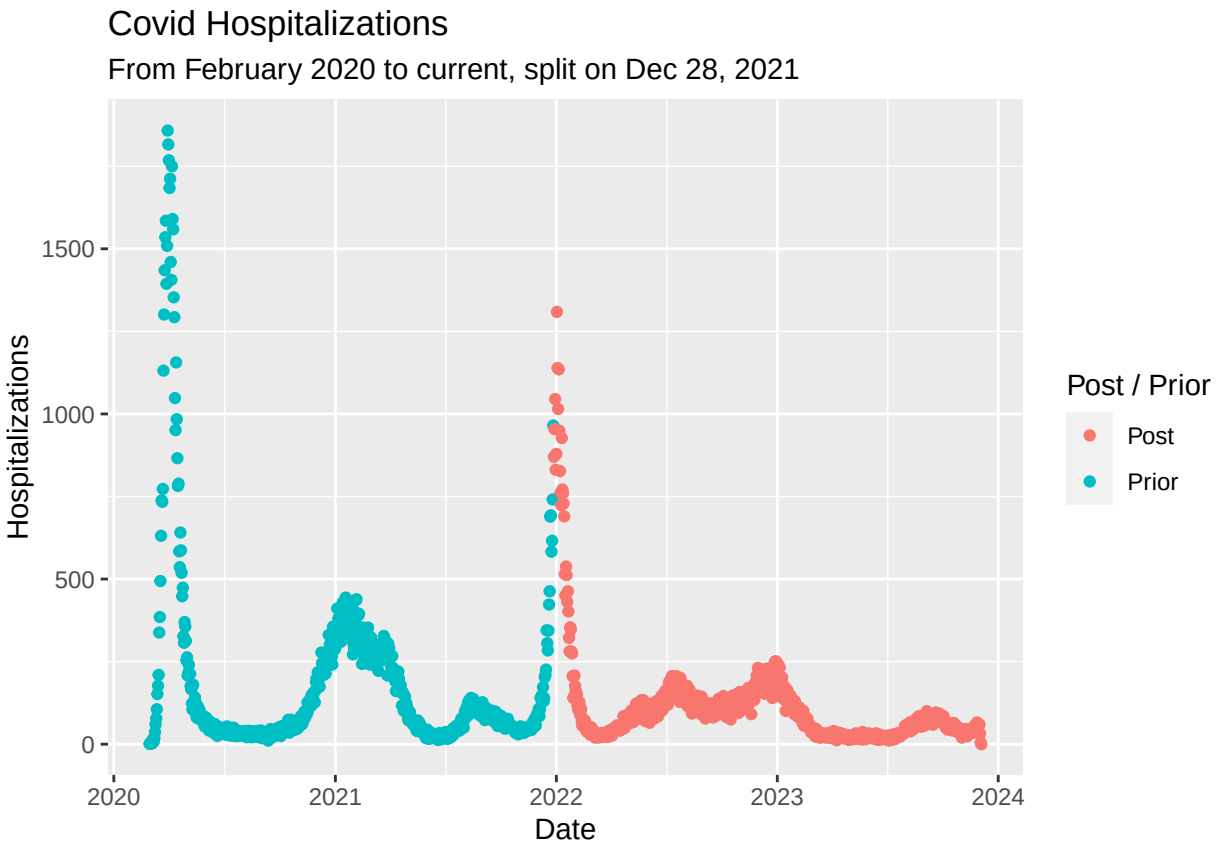
Something I would love to do additional analysis on is looking to see if there is any correlation between the ages of people that were associated with each hospitalization and death count. I know that COVID-19 effects age groups differently. Some are more at risk than others depending on their health. I think it is important to look at the ages and health risks each person has. These are important factors that would help dig deeper into my data analysis. This could also lead to a better understanding of what type of people the vaccine helped. To do this analysis I would need the age of each case along with any health risks the individual might have which would make them more susceptible to contracting COVID-19.

Code and Visuals

```
ggplot(data, aes(x = date_of_interest, y = DEATH_COUNT, color = post_prior)) +
  geom_point() +
  labs(title = "Covid Deaths",
       subtitle = "From February 2020 to current, split on Dec 28, 2021",
       x = "Date",
       y = "Deaths",
       color = "Post / Prior")
```




```
ggplot(data, aes(x = date_of_interest, y = HOSPITALIZED_COUNT, color = post_prior)) +
  geom_point() +
  labs(title = "Covid Hospitalizations",
       subtitle = "From February 2020 to current, split on Dec 28, 2021",
       x = "Date",
       y = "Hospitalizations",
       color = "Post / Prior")
```


Covid Hospitalizations
From February 2020 to current, split on Dec 28, 2021
Hospitalizations
1500
1000
500
0
2020
2021
2022
2023
2024
Date
Post / Prior
Post
Prior