Group Project

Emily Gill, Ayla Rios, Matt Luo, Sahana Harikrishnan

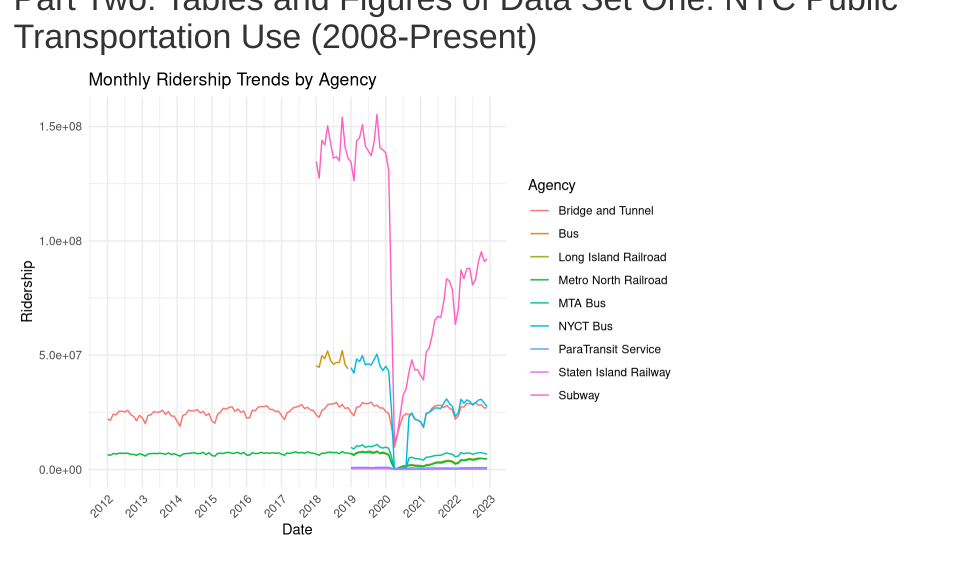
2024-12-11

Introduction: The American Lung Association ranked New York City (NYC) in the top 15 of cities with the worst air pollution. NYC is also known nationwide both for its expansive public transportation system and for large amounts of vehicle traffic. With this in mind, we decided to investigate possible correlations between ridership of public transportation and air pollution in the city to see if an increase in use of public transportation has an impact on air pollution levels. We used ridership data from the Metropolitan Transportation Authority (MTA), as it operates the bus, subway, and rail services for NYC. For air pollution, we used the NY Department of Health's air quality data from its New York City Community Air Survey, focusing on the measurements of fine particles and NO2. Both fine particles and NO2 are forms of air pollution released by cars – fine particles contribute to various health issues and NO2 leads to ozone and smog. Within both datasets, we chose to focus on data collected from 2012-2022, as we anticipate that COVID-19 policies in 2020-2021 may have also had an impact on air pollution and public transportation ridership in NYC. (Metropolitan Transportation Authority. MTA Monthly Ridership-Traffic Data: Beginning 2020. Accessed at https://data.ny.gov/Transportation/MTA-Monthly-Ridership-Traffic-Data-Beginning-Januar/xfre-bxip/about_data on 12/10/2024.) (New York City Department of Health, Environment & Health Data Portal. Air Quality data. Accessed at https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/air-quality/? on 12/10/2024.)

Part One: Tidy Version of Data Set One: NYC Public Transportation Use (2008-Present)

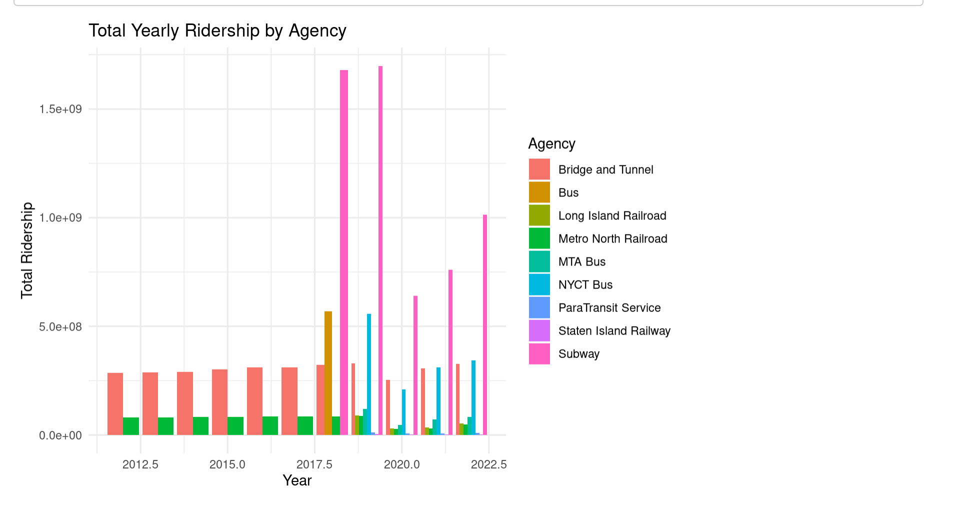
Though the MTA Monthly Ridership/Traffic data was relatively simple and did not have any n/a values to deal with, we still took several steps to clean the data to make it easier to create the figures. First, we cleaned the column names so they were all lowercase and any spaces were converted to underscores, so naming conventions between the two datasets were uniform. Then, we renamed the acronyms in the 'agency' column to the full agency names to make it easier for a layperson without knowledge of MTA agencies to understand at a glance. We also added two new columns: a 'date' and 'year' column. The 'date' column took the character values from the original 'month' column and converted them into a date-class variable for uniformity with the air pollution dataset. The 'year' column extracted the year value from the 'date' column to make it easier to take the mean of ridership per year for the figures. We did not encounter many difficulties, but in our research of the agencies we found that the "Bridge and Toll" agency reported toll counts of cars going through the bridges and tunnels. Since this agency reported ridership of cars and not public transportation, we decided to exclude those values when taking the mean of public transportation ridership per year.

Part Two: Tables and Figures of Data Set One: NYC Public Transportation Use (2008-Present)



In this graph we see that in general, most of the modes of transportation are used at a consistent rate over time, with regular dips in the winter. The bus data beginning in 2018 has an interesting pattern where it appears to regularly fluctuate month to month. This oscillation is most likely because of how people build habits. As expected, there was a significant dip in the use of all transportation during COVID-19, and no agency has recovered to pre-pandemic levels. This is most likely due to employers keeping work-from-home policies instituted during the pandemic, meaning that less people need to use transportation at all day-to-day. From the data, we see that subways see the most use by far over any other agency. Subways are a cheap way of travel throughout the city, and the agency has seen the highest ridership after the COVID-19 pandemic. Unfortunately, the data did not include the data on subways before 2018, but it is clear that they are the dominant form of transportation in the city.

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```



The next graph that we made was to look at the total yearly ridership by agency. This further supports the graph from above as it shows that subways are the dominant mode of transportation. This is likely because of the NYC subway system's relative affordability compared to driving and taking the train, as well as the fact that it serves every borough in the city. Further, the graph shows that, after the COVID-19 pandemic, the use of every form of transportation, besides subway transportation, was not able to rebound back to its previous position. The importance of this graph to compare ridership between agencies is also to show the trend of how people have adapted to different types of transportation after the COVID-19 pandemic.

Part Three: Tidy Version of Data Set Two: NYC Air Pollution Data Since (2008-2022)

These methods were chosen to create a consistent, accurate, and usable dataset for creating figures. First, columns were renamed to a uniform format, and the two data files were merged to allow straightforward comparisons of NO2 and PM2.5 over the same time periods. The data was then filtered to the years 2012 to 2022 to ensure that only relevant information was retained. We made sure key columns were numeric values to ensure compatibility, especially when creating tables and figures. We removed outliers using the IQR method which aimed to improve data quality by reducing the influence of extreme values. Extracting the year information and adjusting column names further enhanced the dataset. We made sure to make column names all lower case with "_" as the separator so that each group member had an easier process. Finally, the cleaned and summarized data ensured that the process was documented and the results could be easily shared and used in further manipulation.

Part Four: Tables and Figures of Data Set Two NYC Air Pollution Data Since (2008-2022)

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
## # A tibble: [illegible]
[illegible]
```

```
## # A tibble: [illegible]
[illegible]
```

```
## # A tibble: [illegible]
[illegible]
```

```
## # A tibble: [illegible]
[illegible]
```

First, we started off with a summary table to show the key pollution statistics such as mean_ppb and mean_mcg_m3, which are used to measure the 2 pollutants, No2 (measured in ppb) and pm2.5 (measured in mcg/m3) respectively. Viewing this summary by geotypes by year helped illustrate larger patterns in the observed pollutants between different areas over time. For example, I see that in 2012 in a Borough, the average mean ppb and average mean mcg m3 were ~ 21 and 10 respectively, and decreased to ~ 16 and 6 in 2022. This is very similar to the trend Citywide, which started off as ~ 20 and 10 in 2012 and decreased to ~ 15 and 6 in 2022.

```
## # A tibble: 107 × 2
##    geography                                mean_ppb
##    <chr>                                       <dbl>
##  1 Midtown (CD5)                                29.7
##  2 Gramercy Park - Murray Hill                  27.5
##  3 Clinton and Chelsea (CD4)                    26.7
##  4 Financial District (CD1)                     26.7
##  5 Chelsea - Clinton                            26.6
##  6 Stuyvesant Town and Turtle Bay (CD6)         26.6
##  7 Greenwich Village and Soho (CD2)             26.3
##  8 Upper East Side (CD8)                        25.3
##  9 Lower Manhattan                              24.4
## 10 Manhattan                                    24.3
## # ℹ 97 more rows
```

```
## # A tibble: 107 × 2
##    geography                                mean_mcg
##    <chr>                                       <dbl>
##  1 Midtown (CD5)                               11.2
##  2 Greenwich Village and Soho (CD2)            10.5
##  3 Clinton and Chelsea (CD4)                   10.4
##  4 Gramercy Park - Murray Hill                 10.2
##  5 Chelsea - Clinton                           10.2
##  6 Financial District (CD1)                    10.1
##  7 Stuyvesant Town and Turtle Bay (CD6)        10.0
##  8 Greenpoint and Williamsburg (CD1)            9.91
##  9 Upper East Side (CD8)                        9.87
## 10 Woodside and Sunnyside (CD2)                 9.85
## # ℹ 97 more rows
```

Next, we wanted to rank the mean ppb and mcg for different smaller geographies in the NYC area to see which areas had higher pollutants for both No2 and pm2.5. We noticed that Midtown was the highest for both which makes sense given its high tourist activity, office presence, and residential areas as well. South Beach - Tottenville had the least mean ppb and second to last mean mcg, which also makes sense because of the lack of vehicles present.

```
## `summarise()` has grouped output by 'season'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 22 × 4
## # Groups:   season [2]
##    season  year avg_mean_ppb avg_mean_mcg_m3
##    <chr>  <dbl>        <dbl>           <dbl>
##  1 Summer  2012         17.9           10.6
##  2 Summer  2013         17.5           10.5
##  3 Summer  2014         16.4            9.01
##  4 Summer  2015         15.7            9.75
##  5 Summer  2016         15.1            8.44
##  6 Summer  2017         15.3            9.36
##  7 Summer  2018         14.1            8.66
##  8 Summer  2019         14.6            8.28
##  9 Summer  2020         12.1            7.15
## 10 Summer  2021         12.6            8.60
## # ℹ 12 more rows
```

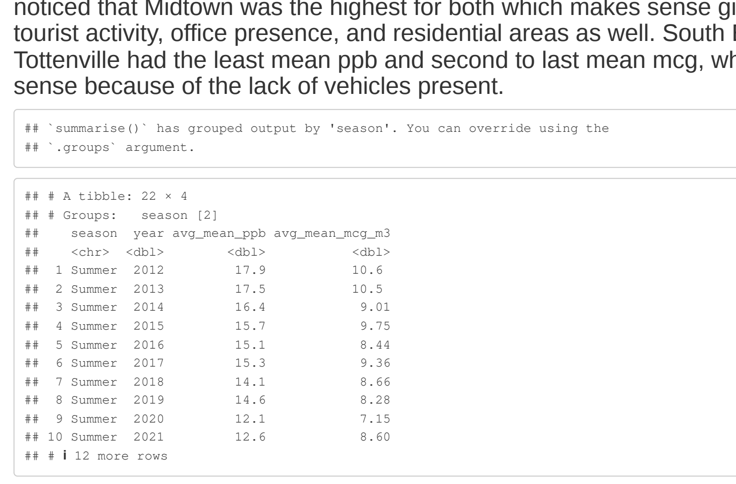
We the wanted to check for seasonality and filtered by Summer and Winter to see how the pollution data would differ. In the summer, the most recent values (2022) were ~12 and 7 for ppb and mcg, and in the summer, the most recent values were ~ 16 and 6 for ppb and mcg respectively. The reason that there is a more drastic difference between the ppb values which measure No2, is because these are the particles that are usually released from vehicles, and people will more likely take vehicles in the winter rather than walking.

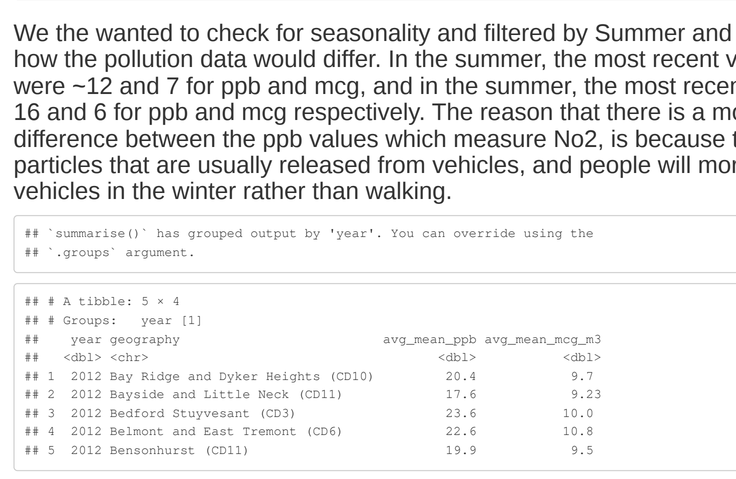
```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 5 × 4
## # Groups:   year [1]
##    year geography                              avg_mean_ppb avg_mean_mcg_m3
##   <dbl> <chr>                                         <dbl>           <dbl>
## 1  2012 Bay Ridge and Dyker Heights (CD10)             20.4            9.7
## 2  2012 Bayside and Little Neck (CD11)                 17.6            9.23
## 3  2012 Bedford Stuyvesant (CD3)                       23.6           10.0
## 4  2012 Belmont and East Tremont (CD6)                 22.6           10.8
## 5  2012 Bensonhurst (CD11)                             19.9            9.5
```

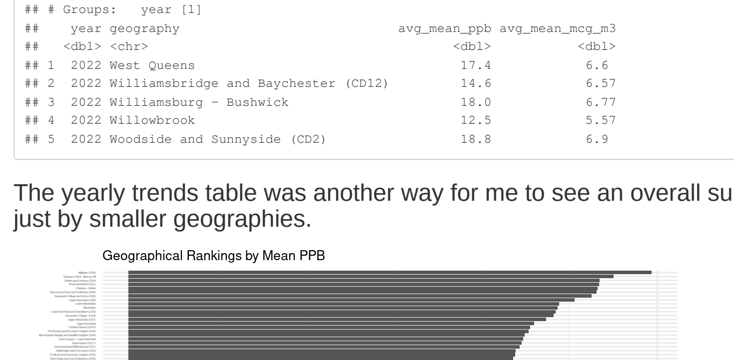
```
## # A tibble: 5 × 4
## # Groups:   year [1]
##    year geography                                avg_mean_ppb avg_mean_mcg_m3
##   <dbl> <chr>                                           <dbl>           <dbl>
## 1  2022 West Queens                                      17.4            6.6
## 2  2022 Williamsbridge and Baychester (CD12)             14.6            6.57
## 3  2022 Williamsburg - Bushwick                          18.0            6.77
## 4  2022 Willowbrook                                      12.5            5.57
## 5  2022 Woodside and Sunnyside (CD2)                     18.8            6.9
```

The yearly trends table was another way for me to see an overall summary over time, just by smaller geographies.

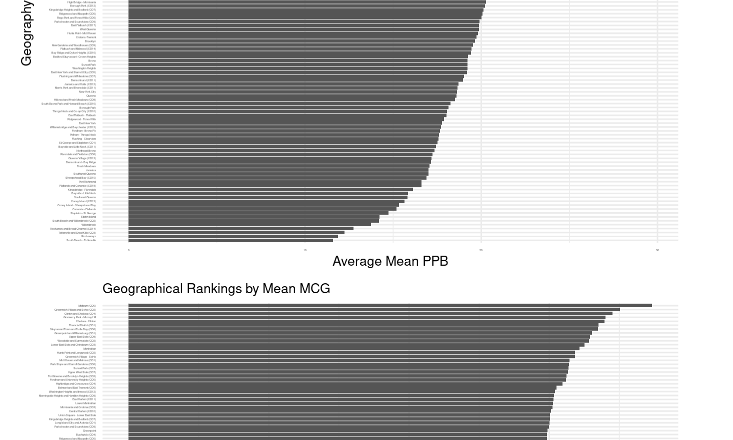




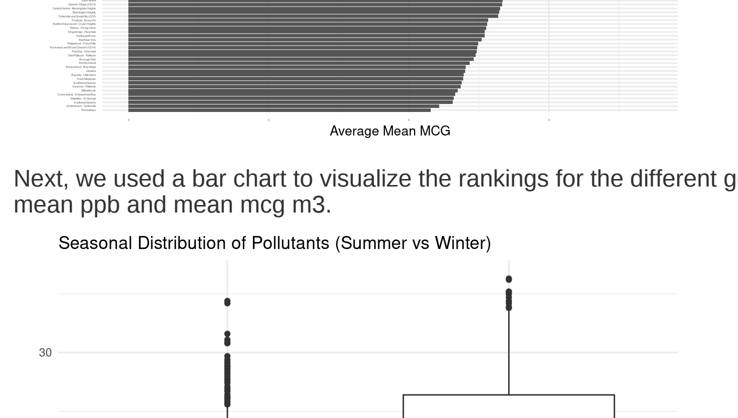
Next, we used a bar chart to visualize the rankings for the different geographies by mean ppb and mean mcg m3.



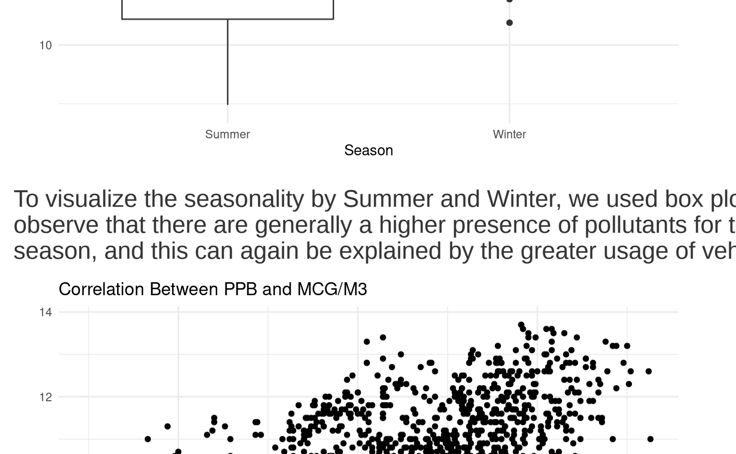
To visualize the seasonality by Summer and Winter, we used box plots. We can observe that there are generally a higher presence of pollutants for the Winter season, and this can again be explained by the greater usage of vehicles.



Next, we used a scatterplot to show the correlation between the No2 pollutants (ppb) and pm2.5 (mcg m3). The scatterplot shows a string positive relationship between Ean No2 (ppb) and pm2 (mcg m3). This could be because they share many similar sources such as vehicles, fossil fuels, etc. Additionally, it is clear that in an urban environment like NYC, the presence of both of these pollutants. Also, oftentimes, No2 can lead to other fine particles such as PM2 due to certain chemical reactions.



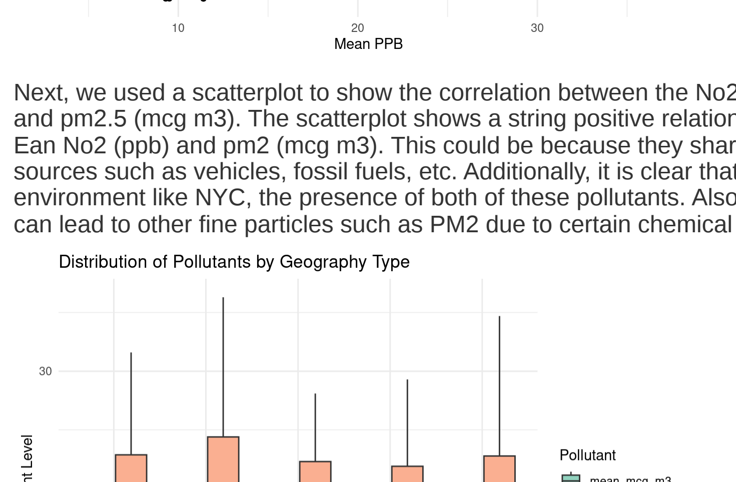
Next, we used box plots to see the distribution of values for ppb and mcg by genotype. As we can seem the distribution for ppb which measures No2 has a much more wide spread distribution with higher maxes, lower mins, higher quartiles, and higher medians.



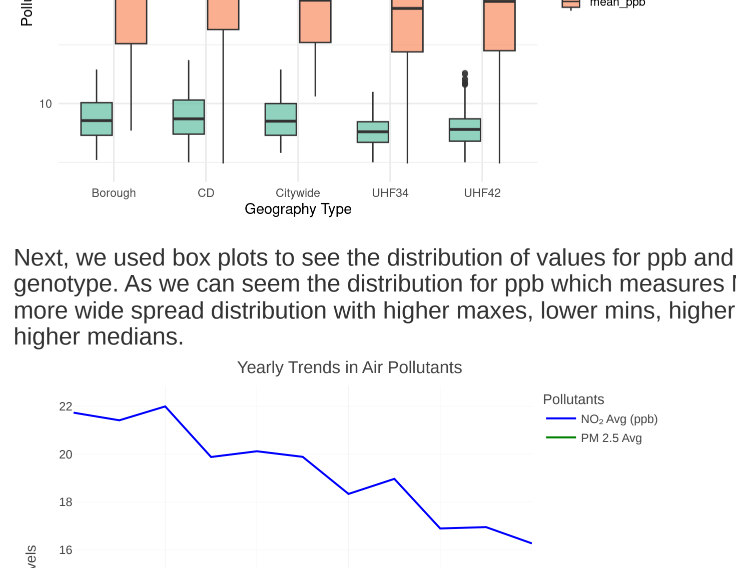
Lastly, we created an interactive line plot to show the two pollutants. It is clear that there is a smaller presence in the fine particles than the N02, but they both decrease overtime. There are also less extreme peaks for the fine particles rather than No2.

Part Five: Tables and Figures of Both Data Sets

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



NO2 levels appear to show slightly lower values post-COVID (red points), possibly due to reduced vehicle emissions when ridership was low in the scatterplot. This chart shows how ridership patterns shifted significantly during the pandemic, and this may have influenced air quality (NO2 levels). During pre-COVID, higher ridership levels correlate with higher NO2 levels, reflecting increased vehicular traffic and emissions in conjunction with public transit usage. During post-COVID, ridership was lower, and NO2 levels seem to decrease slightly, potentially due to reduced traffic and fewer emissions. In the line chart, red lines indicates NO2 levels (ppb), which are relatively stable pre-COVID, despite increasing ridership. It exhibits a slight decline post-COVID, likely due to reduced emissions as traffic decreased. The green dotted line represents PM2.5 levels (μg/m³), which remain relatively low and stable across the years, with minimal fluctuations compared to NO2 levels. This suggests that while ridership and NO2 seem correlated, PM2.5 levels may be less directly influenced by public transportation trends.

Conclusion: After visualizing the datasets both individually and comparatively, we were able to identify some general trends. We noticed a decrease in the average pollution levels in NYC from 2012-2022, with what looks like a small increase in both forms of pollutants we looked at right after 2020. This may be because caution around the pandemic led people to favor personal vehicles over crowded public transit, which would mean more pollution generated by cars, trucks, etc. In the MTA data, we see a huge dip in ridership volume in all public transit areas in 2020, which we attribute to the pandemic. These levels do not appear to recover fully after 2020, likely due to a higher number of people working from home. When we looked at the two sets of data together by year, beginning in 2018 when the MTA began collecting data for the bus and subway, we saw that the NO2 levels rose along with ridership levels and the PM2.5 levels remained relatively stable – indicating that the source of the PM2.5 pollutant may not be linked directly to transportation. The data seems to suggest that before 2020 and the pandemic, higher average ridership is linked to higher NO2 levels, and that after the 2020 pandemic, both the level of ridership and general pollutant levels decreased.