Yunmei Zheng
NYC and Polluants
Professor Burris
December 16, 2022

The population of New York City and the pollutants that surround it in each area are the focus of this Project. Just to give some history on New York City, it is a very multicultural city with residents from all over the world. Additionally, it is densely populated 8 million people in 2020. While everyone is exposed to air pollution, degrees of exposure and population vulnerability differ among neighborhoods, making it one of the most significant environmental risks to urban populations. Common air contaminants have been related to cancer, cardiovascular and respiratory disorders, as well as premature deaths. To more accurately describe the air quality and health in NYC, these metrics offer a perspective across time and geographic areas of the city. The American Lung Association claims that numerous research shows that low socioeconomic position and ethnicity have an effect on the populations that are frequently exposed to greater levels of pollution. Ozone (O3), nitrogen oxides, and fine particulate matter (PM2.5) are some of the most significant pollutants in NYC.

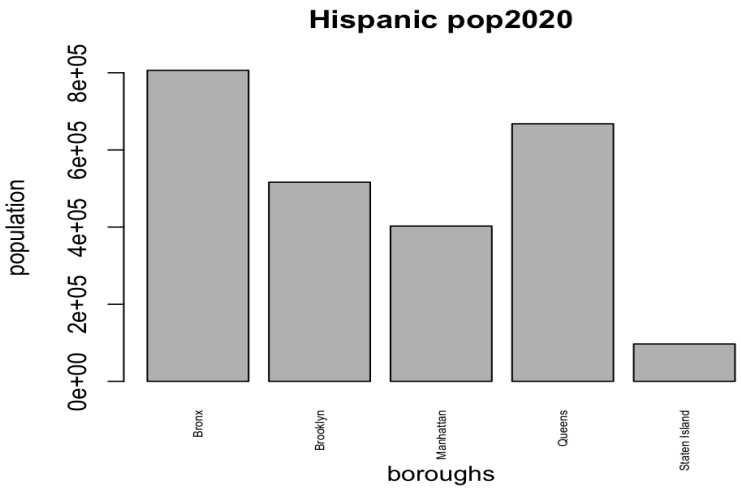
PM2.5 is a tiny, inhalable particle with a diameter of 2.5 micrometer or less. The small size allows it to enter the bloodstream and travel deep within the lungs. It raises the chance of death by causing lung function problems, asthma, and respiratory inflammation. Chronic lung disease may result from NO exposure over an extended period of time. Chronic inhalation of 03 might cause congestion, coughing, throat irritation, and chest pain. The majority of emissions are caused by the burning of diesel fuel in buildings, power plants, and construction machinery. Even while traffic is a big contributor to the city's pollution, it isn't dispersed equally around the metropolis. In this study, we'll examine which region and demographic are most exposed to pollutants.

Some of the question that needs to be answered are

- Which borough has the most minority?
- Does the minority in NYC have more exposure to pollutants?
- Should income be considered a factor?
- Which place has the most pollutants?

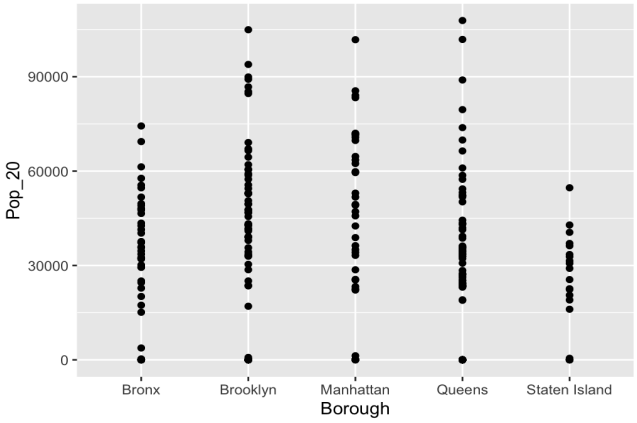
For this project, I used two datasets to help me answer the questions. The first one was from the 2020 Census and contained information on the neighborhood's racial demographics. The second dataset from the NYC open source database contains the city's pollutants between 2010 and 2020.
Some of the the cleaning methods I utilized were:
mutate + As.numeric: it converts the columns into numerics so i can use it.
Str_detect: pull out a string within a certain column into a new dataframe
as.Date: sort the date column into m/d/y so it's easier to call out a certain month or year
Filter: Pulls the values i need into a new dataframe

Some EDA:



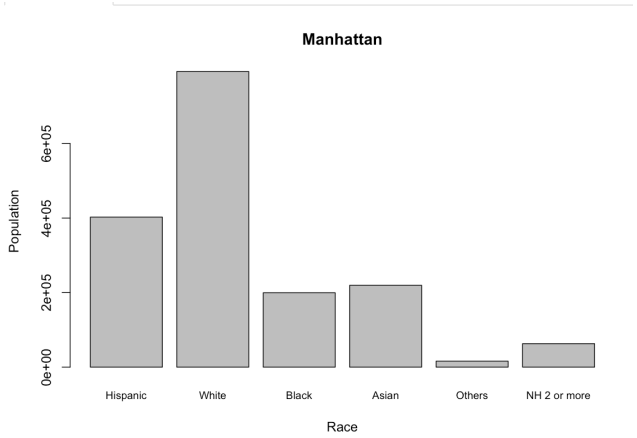
This bar graph displays the total number of Hispanic residents in each borough. We can observe from this barplot that the Bronx has the largest percentage of Hispanic residents.
With this knowledge, we can say that the majority of the minority group of Hispanics lives in the Bronx.

Figure 1: Hispanic Population



The three boroughs with the highest populations are Brooklyn, Manhattan, and Queens, as seen by this graph. When we look at Staten Island, we can observe that one of the neighborhoods has a population of up to 50,000 people.

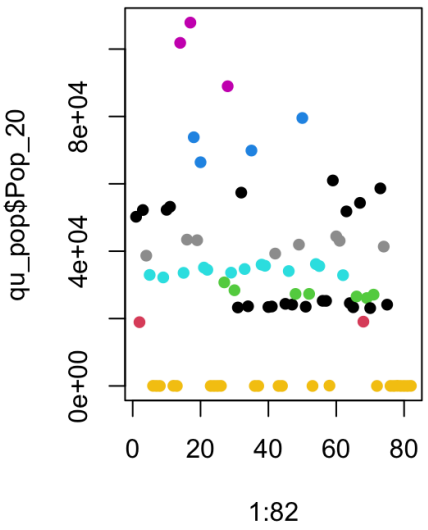
Figure 2: Population Distribution Among Boroughs



In contrast to Figure 1, this barplot depicts the racial makeup of the borough. This chart of Manhattan's racial makeup reveals that the majority of its residents are white. Hispanics make up half the population, while Asians and Blacks make up 4 times it.

Figure 3: Race in Manhattan

Some Models:

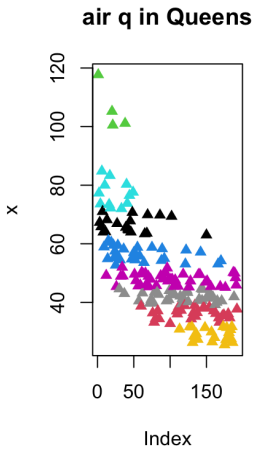


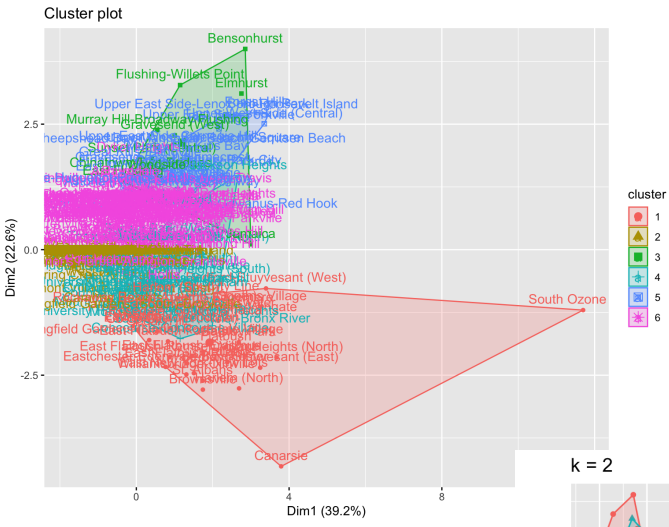
This is a k-mean that solely includes Queens residents. Given that there are 82 data in total, and that number's square root is roughly 9, I chose a k of 9. We can observe from this model that only a small number of values are quite high. The majority of the bottom yellows are zeros; I decided to leave them because they had no effect on the models.

Figure 4: K Means of Queens' population

This model is a collection of data about Queens' air quality from 2010 to 20220. Only a handful of them tend to be higher than 60, with the rest falling below that range.

Figure 5: Cluster of Queens' Polluants





These two models represent the neighborhood in New York City using cluster and k means. It is clustered by their race and the neighbors. The names of each neighborhood are displayed in Figure 6. The second graph displays each of the k values and the associated graphs.

Figure 6: Cluster of all neighbors (above)

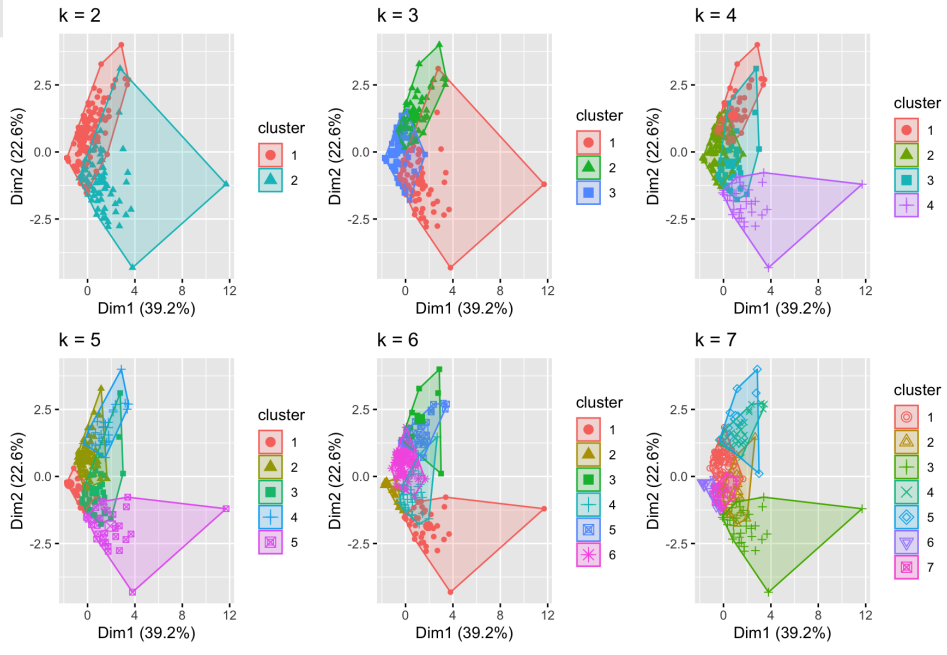


Figure 7: K Means of all neighborhoods

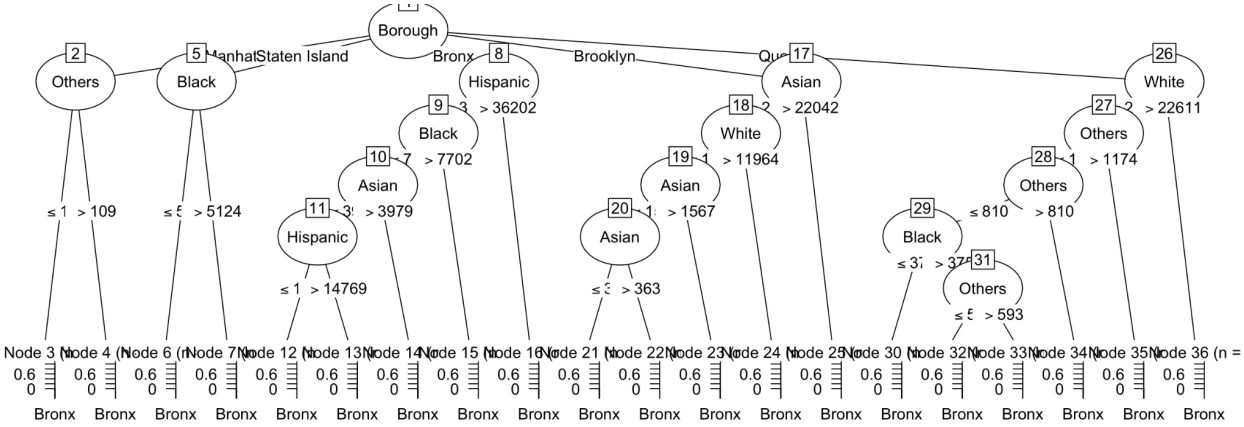


Figure 8: C4.5 Model

This model was forecast by dividing the boroughs into different ethnic groups before assigning them. I obtained an error of 57.6%, after performing the calculations using the confusion matrix, I obtained an accuracy of 40.66% from this model.

C4.5 creates a tree with a more diverse structure and is not just restricted to binary splits. C4.5 chooses the feature or attribute to divide based on an entropy-based metric. Until no more splits are allowed, the C4.5 algorithm generates each decision node by iteratively choosing the best split. There are 36 nodes altogether. The divide looks relatively uniform based on the nodes, which indicates a poor categorization.

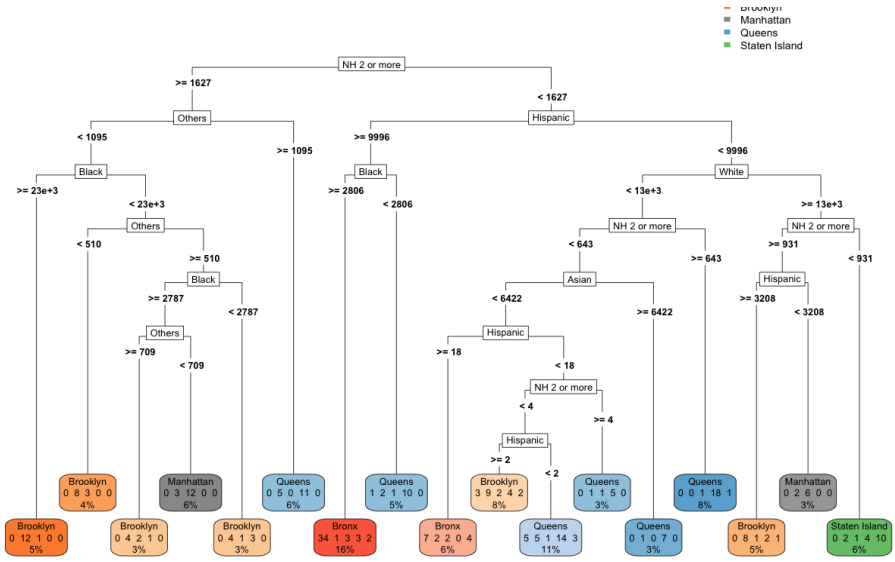


Figure 9: CART Model

The second model made use of CART, which divides data into subsets with comparable values for the target variables. It aims to provide as "pure" a set of leaf nodes as it can. A leaf node is said to be pure if all of its records share the same categorization or label for the target variable. I determined an accuracy of 72.5% using the provided confusion matrix.

Results and Takeaways

- Which borough has the most minority?
- Highest demographic in each borough
- Queens, Bronx: Hispanic
- Brooklyn, Manhattan. Staten Island: White
- Does the minority in NYC have more exposure to pollutants?

	Queens	Bronx	Brooklyn	Staten Island	Manhattan
Nitrogen Dioxide	10.00	13.60	10.82	7.83	15.97
Fine Particulate Matter	6.87	7.24	6.94	6.58	7.80
Ozone (O3)	30.52	30.85	29.69	28.61	28.77

Figure 8: Table of Boroughs and Polluants

All three pollaunts' summer 2020 data are included in the table above. Since there was no annual data value for O3, I used the summer data throughout to maintain consistency. We can infer from the above table that Bronx has the second-highest concentrations of all three pollutants, whereas Manhattan has the highest concentrations of NO and PM2.5. Given that White people make up the majority of Manhattan's population, it is clear that this pattern does not correspond with minority groups experiencing higher pollution levels. However, since Manhattan is the center of New York City and is so crowded and congested, we can ignore it. We might infer that Bronx, which is home to a majority of Hispanics and is a minority group, ranks second.

- Which place has the most pollutants?

	Bronx	Manhattan	Brooklyn	Staten Island	Queens
	Soundview	UWS	Borough Park	Great Kills	Elmhurst
NO	12.72	15.07	10.74	7.32	12.36
PM2.5	7.2	7.24	6.74	6.52	7.12
O3	31.72	28.97	29.38	29.67	30.98
	Hispanic	White	White	White	Asian

According Figure 9, Manhattan continues to have the highest level for two of the pollutants; but, if we overlook that once more, we can see that Soundview in the Bronx and Elmhurst in Queens have the highest levels. Minority groups make up the largest demographic in such neighborhoods.

Figure 9: Table of Neighborhood and Pollutants

We may infer from the two datasets that the minority in NYC does experience the highest levels of pollution to some degree, although fortunately none of the boroughs exceeded the standard established by the Environmental Protection Agency (EPA). Manhattan typically has the highest rates, due to its large visitation volume.