Segmenting and Clustering New York City Neighborhoods by Air Quality and Venues

Rishika Raj Akula, November 2020

1. Introduction

1.1 Background

Air quality is can be a major factor in the health issues that people face. The World Health Organization reports that "ambient air pollution accounts for an estimated 4.2 million deaths per year."[1] Due to this, the pollution and air quality of an area could be a very significant deciding factor in where people want to live.

1.2 Problem

New York City is one of the largest and fastest-growing cities in the world and has to deal with significant air pollution. So, if someone had to move to NYC for school or work, and they had a pre-existing condition like asthma, they would benefit from having information about what levels of air pollution are present in different parts of the city. This way they could make an informed decision about where to live while taking their health into consideration.

This project will focus on categorizing and clustering neighborhoods in NYC based on the venues, to see which neighborhoods are similar, and Outdoor Air and Health information, to understand the air quality in each neighborhood.

1.3 Interest

The information from this type of analysis would be of interest to someone looking to move to New York City or looking to move to a different part of New York City for health reasons. This information could also be helpful to identify which neighborhoods and types of neighborhoods are having the most issues with air pollution.

One health issue that over 25 million Americans have is asthma. [2] “This is 7.7 percent of adults and 8.4 percent of children,” which is a significant portion of the population. [2] This is why this project will be interested in asthma hospitalizations in New York City. On the side of air quality, it is known that high levels of Ozone and Fine Particulate Matter (PM 2.5) trigger and exacerbate asthma, which is why these two characteristics will be focused on. [3] [4]

2. Data

For this project data was collected from two sources:

2.1 Foursquare API:

https://foursquare.com/

Foursquare was used to gather venue data about each neighborhood.

2.2 New York City Environment & Health Data Portal:

http://a816-dohbesp.nyc.gov/IndicatorPublic/

This database allows you to choose categories of what data you are interested in. So, for this project, Ozone and Fine Particulate Matter (PM 2.5) readings were chosen for the air quality data. For Public Health Data, Asthma Hospitalizations for three different age categories were chosen. The data from these categories were collected in a CSV file.

3. Methodology

3.1 New York City Environment & Health Data

Data Retrieval and Wrangling

As mentioned before this Data Portal allowed the selection of specific categories of data and for this project the categories chosen were: Ozone measurements, Fine Particulate Matter (PM 2.5) measurements, and Asthma Hospitalizations (for Children age 0 to 4, Children Age 5 to 17, and Adults). This data was provided in a CSV file.

The data was loaded into a Data frame and then cleaned up. Unnecessary id columns were dropped, columns were renamed to allow for easier referencing, and rows with NaN values were dropped. Each category had multiple different measures for the data, for this analysis only Number and Mean values were used so rows with Estimated Annual Rate, Age-Adjusted Rate, 10th Percentile, 90th Percentile values were dropped. Finally, in the description column, it was decided to only keep rows that were Borough, Citywide, and Neighborhood (UHF 42). So, rows which were Neighborhood (UHF 32), and Neighborhood (Community District) were dropped. This resulted in a data frame with 8 columns and 3221 rows.

This large data frame was then divided into separate data frames based on the data it was describing. This resulted in 5 new tables, one for each: Ozone (O3), Fine Particulate Matter (PM2.5), Asthma Hospitalizations (Children 0 to 4 Yrs. Old), Asthma Hospitalizations (Children 5 to 17 Yrs. Old), Asthma Hospitalizations (Adults).

For each table, the unique descriptors for the columns of Measure, Description, and year were checked. It was noticed that the three Asthma Hospitalizations tables consisted only of distinct years, whereas the Ozone and Fine Particulate tables consisted of Summer, Winter, and Annual Average values. For uniformity, Ozone and Fine Particulate tables were cleaned up to match the year format of the Asthma Hospitalizations tables. So now there were 5 uniform tables that could be used for analysis.

Visualization

To understand the importance or to see if there was any similar trend of Ozone and Fine Particulate Matter with Asthma Hospitalizations, data was plotted to see if any trends could be uncovered.

First data for all of New York City (Citywide) was aggregated onto one table for easy retrieval. I was noticed that only Asthma Hospitalizations had data from years before 2009, so this data was excluded. It was also noticed that only Ozone and Fine Particulate Matter had data past year 2016, so data with years 2017-2019 were excluded. There were five resulting plots showing data from 2009 – 2016.

3.2 Foursquare and Clustering

Data Retrieval and Wrangling

First Geocoder was used to get longitude and latitude coordinates for each of the neighborhoods. The neighborhood names were retrieved from the original raw data. Then using

Nominatim with Geocoder, location data was retrieved, from which longitude and latitude coordinates were extracted and added to a table.

With the longitude and latitude coordinates of each neighborhood, Foursquare was used to get venue data. In total there were 271 unique categories gathered. The one hot encoding method was used to process this venue data. Then the top 10 most common venues for each neighborhood were found and added to a table.

K-means clustering and neighborhood segmentation

To begin k-means clustering, Ozone and Fine Particle Matter readings for 2019 were added to the venue data. The readings for 2019 were chosen to ensure the most recent and accurate data was used. Once this information was combined, preliminary clustering was conducted.

Based on this it was discovered that k = 4 clusters were most optimal due to neighborhood division and layout. K-means clustering was ran using k =4 and the cluster labels were added to the venue data.

Each cluster had a unique number of neighborhoods segmented to it based on Ozone readings, Fine Particle Matter readings, and venue data. The result would be the segmentation of neighborhoods that were similar in Ozone and Fine Particle Matter levels and surrounding venues.

The minimum, maximum, and mean Ozone and Fine Particle Matter data were calculated for each cluster.

Visualization

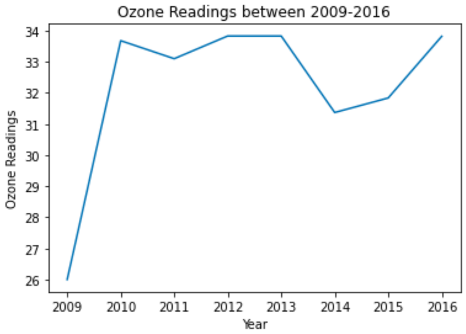
With the cluster labels added to the neighborhood venue data, Folium was utilized to visualize the spread of the neighborhood segmentation. Each cluster was assigned a different color and was plotted on the map with longitude and latitude coordinate data.

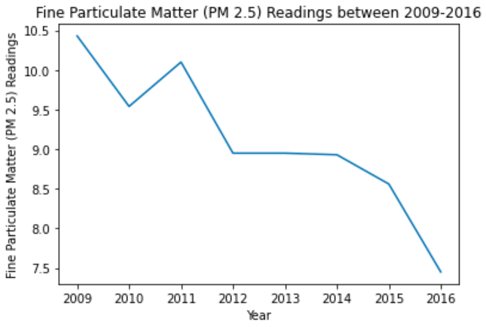
4. Results

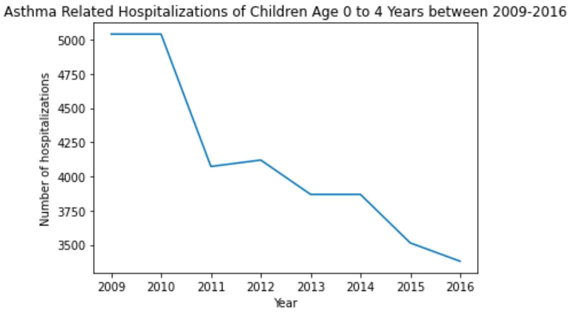
The results from all the data retrieval, data wrangling, k-means clustering, neighborhood segmentation, and visualization were analyzed.

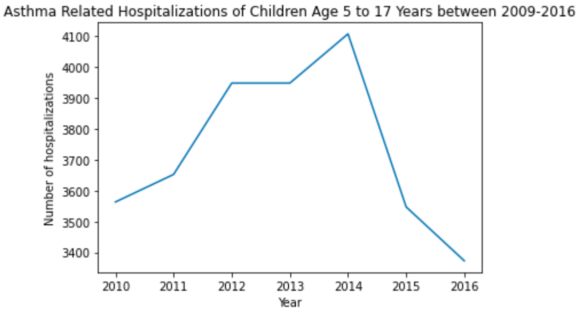
4.1 New York City Environment & Health Data

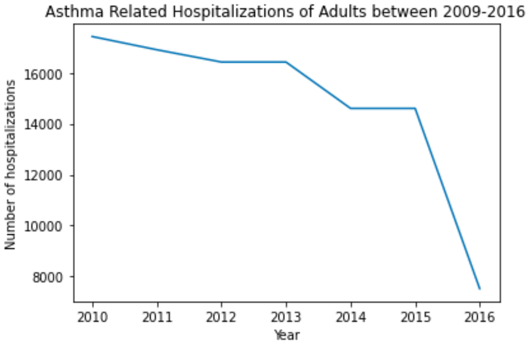
From the Environment and Health Data, Citywide Ozone readings, Fine Particulate Matter readings, and Asthma Hospitalizations were plotted. These plots can be seen below.











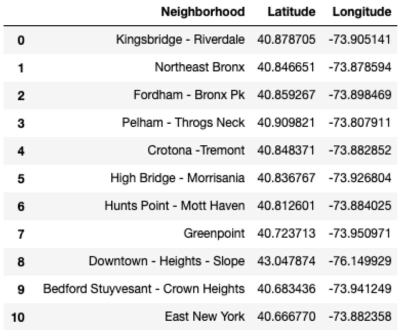
From the plots, it can be seen that Ozone and Fine Particulate matter have an opposite trend where Fine Particulate Matter had an overall downward trend, but Ozone had a sharp increase after 2009 and fluctuated at high levels for the continuing years. Asthma Hospitalizations for Children Age 0 to 4 and Adults also had continuous declines. However, Asthma Hospitalizations for Children Age 5 to 17 was a different story. The number of hospitalizations increased until 2014 when they began declining significantly. Unfortunately looking at only the Ozone and Fine Particulate there was no indication as to why this occurred.

However, it can be observed that although Ozone did not have a distinct pattern, as Fine Particulate Matter trended downwards so did Asthma Hospitalizations (at least after 2014).

This shows why Air Quality should be taken into serious account when considering moving to a new place, as it is shown to have a relationship with Asthma Hospitalizations, and possibly with other types of Hospitalizations and doctor visits as well.

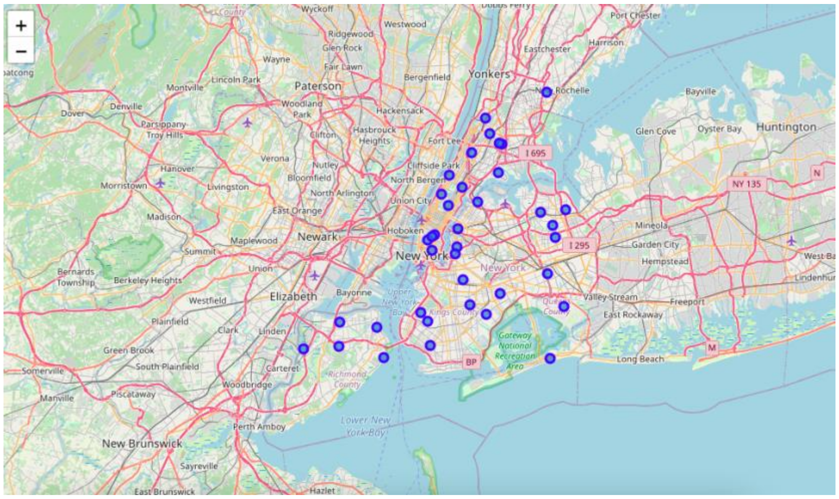
4.2 Neighborhood Clustering and Segmentation

Neighborhood names, and longitude and latitude coordinates for each neighborhood were added to a table.

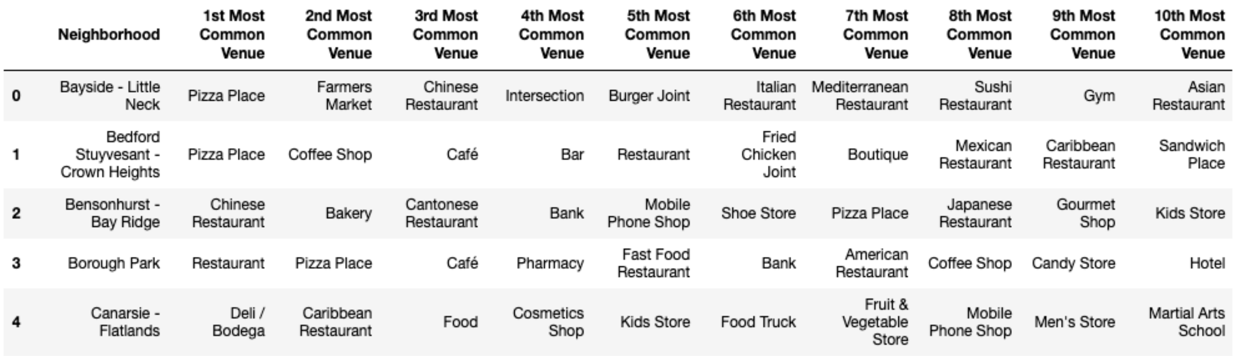


	Neighborhood	Latitude	Longitude
0	Kingsbridge - Riverdale	40.878705	-73.905141
1	Northeast Bronx	40.846651	-73.878594
2	Fordham - Bronx Pk	40.859267	-73.898469
3	Pelham - Throgs Neck	40.909821	-73.807911
4	Crotona -Tremont	40.848371	-73.882852
5	High Bridge - Morrisania	40.836767	-73.926804
6	Hunts Point - Mott Haven	40.812601	-73.884025
7	Greenpoint	40.723713	-73.950971
8	Downtown - Heights - Slope	43.047874	-76.149929
9	Bedford Stuyvesant - Crown Heights	40.683436	-73.941249
10	East New York	40.666770	-73.882358

These coordinates were used to plot a general map of the neighborhoods in New York City without any clustering or segmentation done. This was the resulting map.

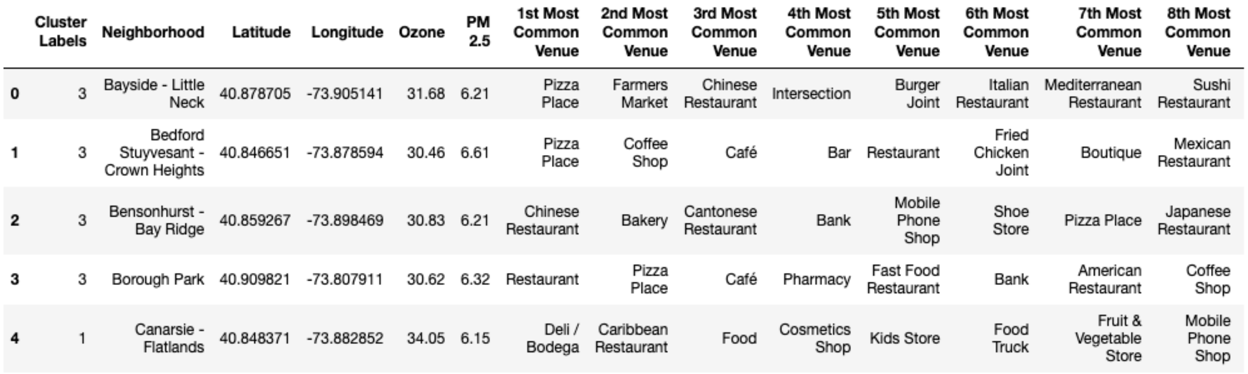


Then with the Foursquare data collected with the coordinates. The venue data was extracted, and the top 10 common venues were displayed. A snapshot of this table can be seen below.



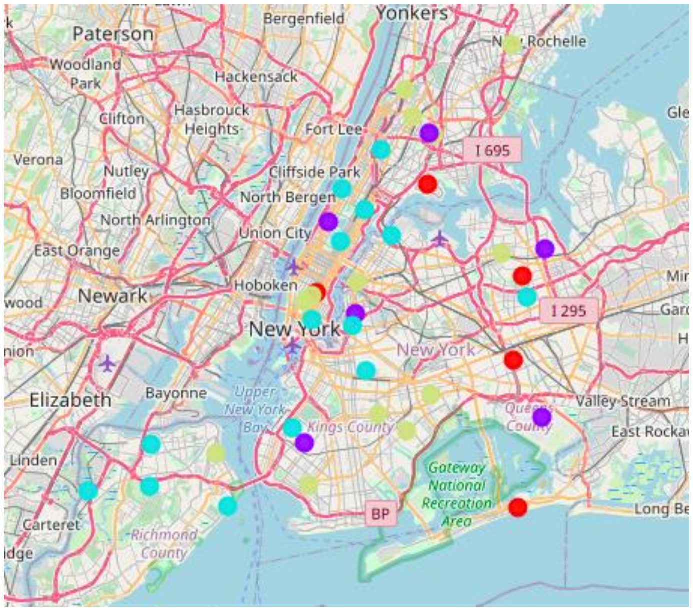
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bayside - Little Neck	Pizza Place	Farmers Market	Chinese Restaurant	Intersection	Burger Joint	Italian Restaurant	Mediterranean Restaurant	Sushi Restaurant	Gym	Asian Restaurant
1	Bedford Stuyvesant - Crown Heights	Pizza Place	Coffee Shop	Café	Bar	Restaurant	Fried Chicken Joint	Boutique	Mexican Restaurant	Caribbean Restaurant	Sandwich Place
2	Bensonhurst - Bay Ridge	Chinese Restaurant	Bakery	Cantonese Restaurant	Bank	Mobile Phone Shop	Shoe Store	Pizza Place	Japanese Restaurant	Gourmet Shop	Kids Store
3	Borough Park	Restaurant	Pizza Place	Café	Pharmacy	Fast Food Restaurant	Bank	American Restaurant	Coffee Shop	Candy Store	Hotel
4	Canarsie - Flatlands	Deli / Bodega	Caribbean Restaurant	Food	Cosmetics Shop	Kids Store	Food Truck	Fruit & Vegetable Store	Mobile Phone Shop	Men's Store	Martial Arts School

Then k-means clustering was conducted, and the resulting data table looked like this:



	Cluster Labels	Neighborhood	Latitude	Longitude	Ozone	PM 2.5	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	3	Bayside - Little Neck	40.878705	-73.905141	31.68	6.21	Pizza Place	Farmers Market	Chinese Restaurant	Intersection	Burger Joint	Italian Restaurant	Mediterranean Restaurant	Sushi Restaurant
1	3	Bedford Stuyvesant - Crown Heights	40.846651	-73.878594	30.46	6.61	Pizza Place	Coffee Shop	Café	Bar	Restaurant	Fried Chicken Joint	Boutique	Mexican Restaurant
2	3	Bensonhurst - Bay Ridge	40.859267	-73.898469	30.83	6.21	Chinese Restaurant	Bakery	Cantonese Restaurant	Bank	Mobile Phone Shop	Shoe Store	Pizza Place	Japanese Restaurant
3	3	Borough Park	40.909821	-73.807911	30.62	6.32	Restaurant	Pizza Place	Café	Pharmacy	Fast Food Restaurant	Bank	American Restaurant	Coffee Shop
4	1	Canarsie - Flatlands	40.848371	-73.882852	34.05	6.15	Deli / Bodega	Caribbean Restaurant	Food	Cosmetics Shop	Kids Store	Food Truck	Fruit & Vegetable Store	Mobile Phone Shop

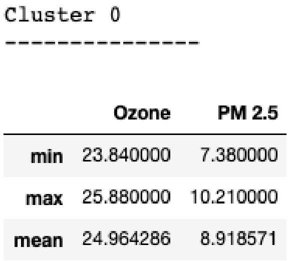
This data was then plotted on a map using Folium, where each cluster was distinguished with a different color.



The following were the number of neighborhoods in each cluster:

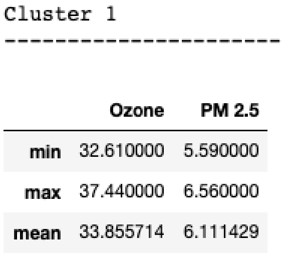
- Cluster 0 (Red): 7 neighborhoods
- Cluster 1(Blue): 7 neighborhoods
- Cluster 2 (Purple): 15 neighborhoods
- Cluster 3 (Yellow): 13 neighborhoods

The minimum, maximum, and mean Ozone and Fine Particle Matter data were calculated for each cluster.



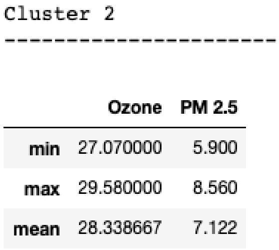
```
Cluster 0
---------------
```

	Ozone	PM 2.5
min	23.840000	7.380000
max	25.880000	10.210000
mean	24.964286	8.918571



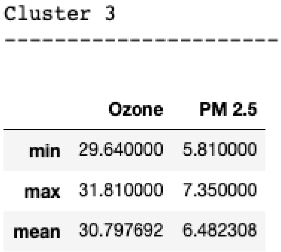
```
Cluster 1
----------------------
```

	Ozone	PM 2.5
min	32.610000	5.590000
max	37.440000	6.560000
mean	33.855714	6.111429



```
Cluster 2
----------------------
```

	Ozone	PM 2.5
min	27.070000	5.900
max	29.580000	8.560
mean	28.338667	7.122



```
Cluster 3
----------------------
```

	Ozone	PM 2.5
min	29.640000	5.810000
max	31.810000	7.350000
mean	30.797692	6.482308

Based on the Mean data it can be said that:

- Cluster 0: Lowest Average Ozone, Highest Average Fine Particulate Matter
- Cluster 1: Highest Average Ozone, Lowest Average Fine Particulate Matter

It can be noticed here that Ozone and Fine Particulate Matter have an inverse relationship.

5. Discussion

This analysis allowed us to show that on a basic level that air quality and asthma hospitalizations have a similar trend/relationship, especially Fine Particulate Matter. This shows the importance of considering air quality when moving to a new city or to a new part of a city, especially if you have a preexisting health condition.

K-means clustering was used to segment the neighborhoods into 4 clusters. These clusters of neighborhoods shared similar air quality levels as well as similar surrounding venues. By looking at the simple statistics of these clusters, it was found that Ozone and Fine Particulate Matter levels had an inverse relationship.

This information would be really helpful to someone living in New York City, who would like to move to a new part of the city based on air quality. They would have a set of options where they could compare venues of those neighborhoods to choose one where they would be most comfortable.

For further analysis on this topic, it would be recommended to add more air quality components to the clustering process. More neighborhoods and neighborhood data should also be added. This way you could analyze more areas and more qualities allowing for a much diverse and comprehensive data set for segmentation.

6. Conclusion

In this project, data from the New York City Environment & Health Data Portal and Foursquare were aggregated and used to cluster and segment neighborhoods. From the New York City Environment & Health Data Portal data about air quality and asthma hospitalizations were collected. From Foursquare, data about venues in each neighborhood in New York City were collected. With all this information, the (42) neighborhoods were clustered into 4 groups which shared similarities based on Ozone readings, Fine Particulate Matter readings, and venue data.

To improve clustering and segmentation accuracy, more specialized data about air quality and neighborhoods should be added to the analysis.

7. References

[1] World Health Organization
https://www.who.int/airpollution/ambient/en/#:~:text=Ambient%20air%20pollution%20accounts%20for,quality%20levels%20exceed%20WHO%20limits.

[2] Asthma and Allergy Foundation of America
https://www.aafa.org/asthma-facts/#:~:text=More%20than%2025%20million%20Americans,age%2C%20sex%20and%20racial%20groups.

[3] Asthma and Allergy Foundation of America
https://www.aafa.org/air-pollution-smog-asthma/#:~:text=Ozone%20triggers%20asthma%20because%20it,Ozone%20can%20reduce%20lung%20function.

[4] American Academy of Allergy, Asthma & Immunology.
https://www.aaaai.org/global/latest-research-summaries/Current-JACI-Research/particulate#:~:text=Fine%20particulate%20matter%2C%20also%20known,examined%20the%20associations%20of%20PM2.&text=For%20elucidating%20the%20plausible%20mechanism%20of%20PM2.