Clustering and Segmentation of New York City Neighborhoods on Air Quality and Venues

RISHIKA RAJ AKULA

NOVEMBER 2020

Air Quality and Health

- Air quality is can be a major factor in the health issues that people face

- New York City is one of the largest and fastest-growing cities in the world and has to deal with significant air pollution.

- 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.

- This inforamtion 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.

Data Retrieval

- **Foursquare API:**
  - https://foursquare.com/
  - Venue Data.
- **New York City Environment & Health Data Portal:**
  - http://a816-dohbesp.nyc.gov/IndicatorPublic/
  - Provided inforamtion about **Environment & Health in NYC**
  - 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.
  - Provided in csv file

# Data Cleaning

- New York City Environment & Health Data
  - Contained information about 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)
  - Raw data was cleaned
    - Duplicate data, unnecessary id columns, redundant data columns were dropped
    - Certain values were converted from floats to integers.
    - Year values were formatted to be more uniform
    - Data was divided into 5 tables
      - One for each of the categories selected.
      - Ozone measurements
      - Fine Particulate Matter (PM 2.5) measurements
      - Asthma Hospitalizations (for Children age 0 to 4, Children Age 5 to 17, and Adults

Data Cleaning

- Location Data
  - Nominatim with Geocoder was used for longitude and latitude coordinates for each neighborhood
  - Neighborhood names were extracted from New York City Environment & Health Data

- Foursquare
  - With location data venue data for each neighborhood was extracted
  - 271 unique categories gathered
  - The one hot encoding method used to process data.
  - Then the top 10 most common venues for each neighborhood was found and added to a table.

Ozone and Fine Particulate Matter
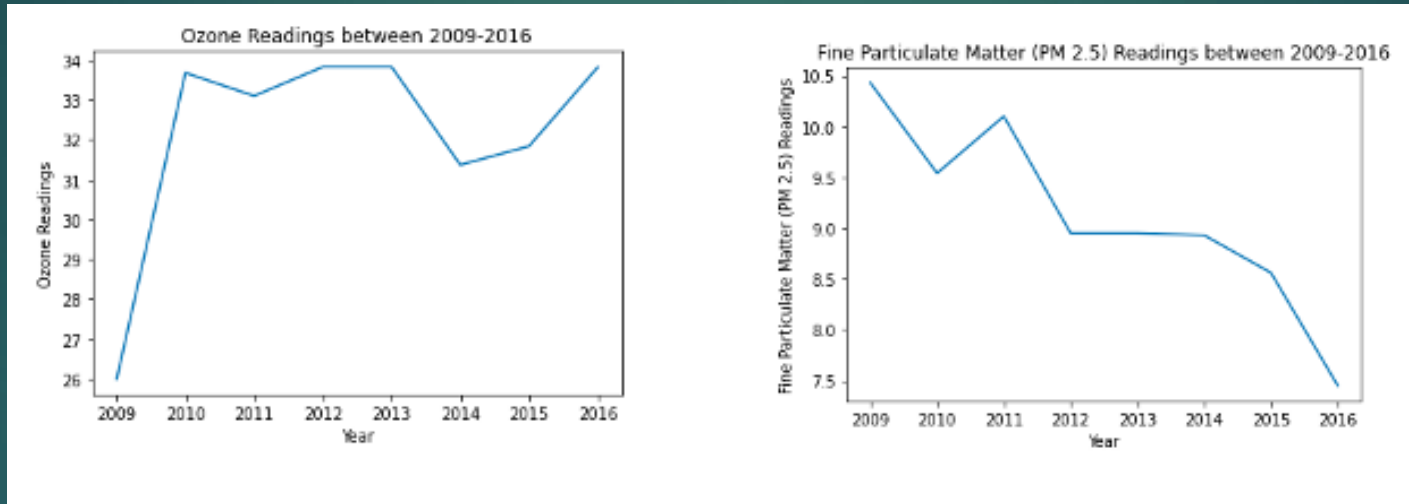
- ▶ Ozone and Fine Particulate Matter have an opposite trend
  - ▶ Fine Particulate Matter had a overall downward trend but Ozone had a sharp increase after 2009 and fluctuated at high levels for the continuing years.

Asthma Hospitalizations
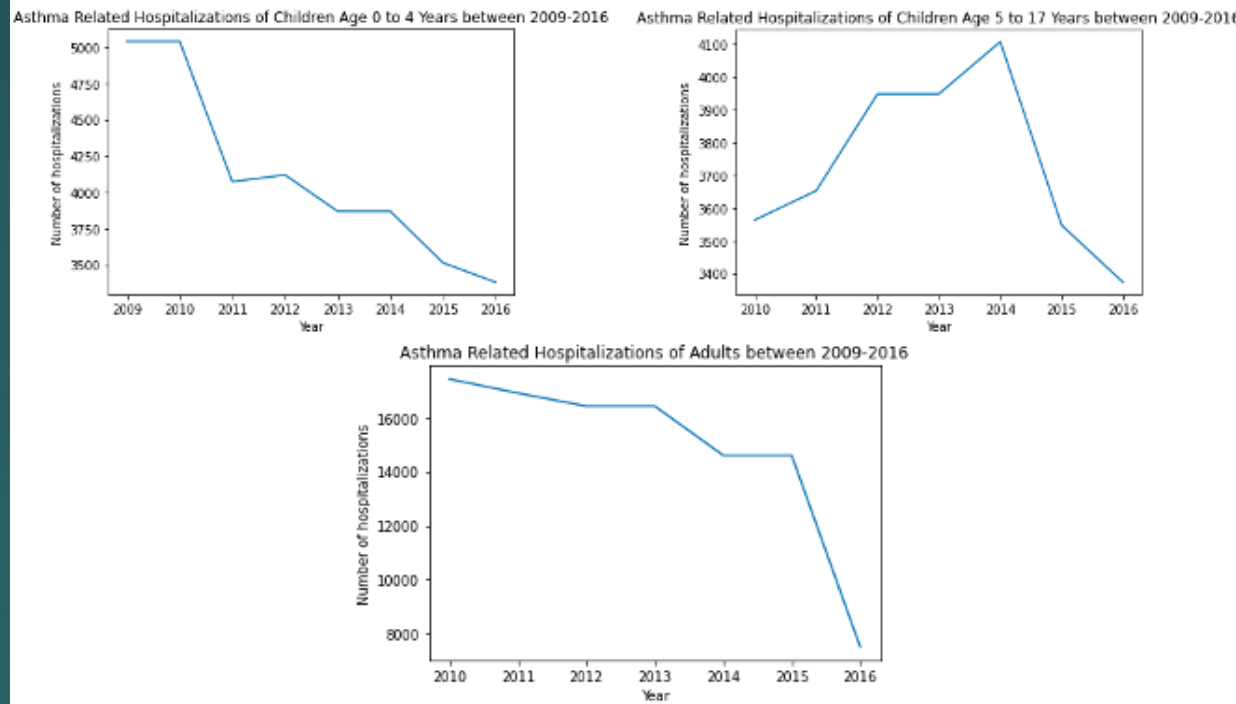
- Asthma Hospitalizations for Children Age 0 to 4 and Adults had conitunous declines.

- Asthma Hospitalizations for Children Age 5 to 17 increased until 2014, when they began declining significantly.

  - No clear indication was found as to why this happened by looking at Ozone and Fine Particulate plots

Air Quality and Asthma Hospitalizations

- Was no similar pattern between Ozone and Asthma Hospitalizations

- Fine Particulate Matter trended downwards so did Asthma Hospitalizations (at least after 2014).

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

K-means clustering

▶ With Venue data combined with Ozone and Fine Particlate Matter Readings (from 2019) k-means clustering was performed

▶ K = 4 clusters

▶ Each cluster had a unique number of neighborhoods segmented to it based on Ozone readings, Fine Particle Matter readings, and venue data.

   ▶ The result was segmentation of neighborhoods that were similar in Ozone and Fine Particle Matter levels and surrounding venues.
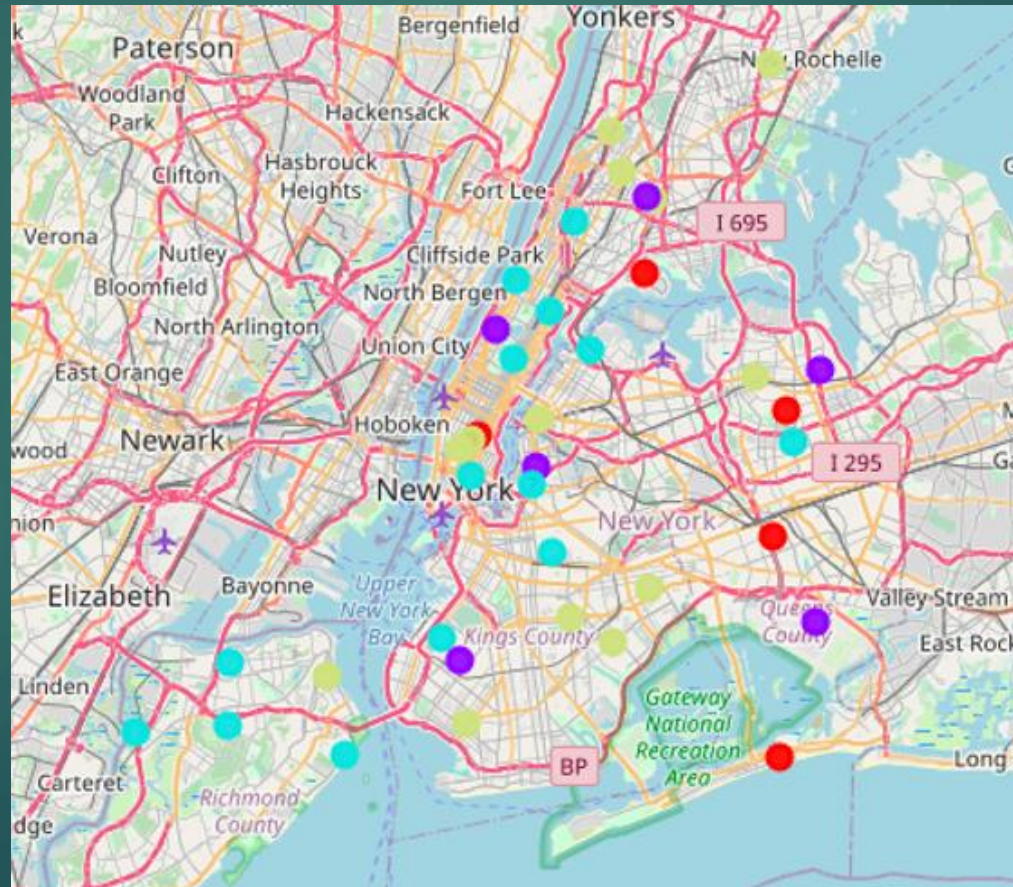
New York City Visualization

- New York City Map with all nieghborhoods displayed without clustering
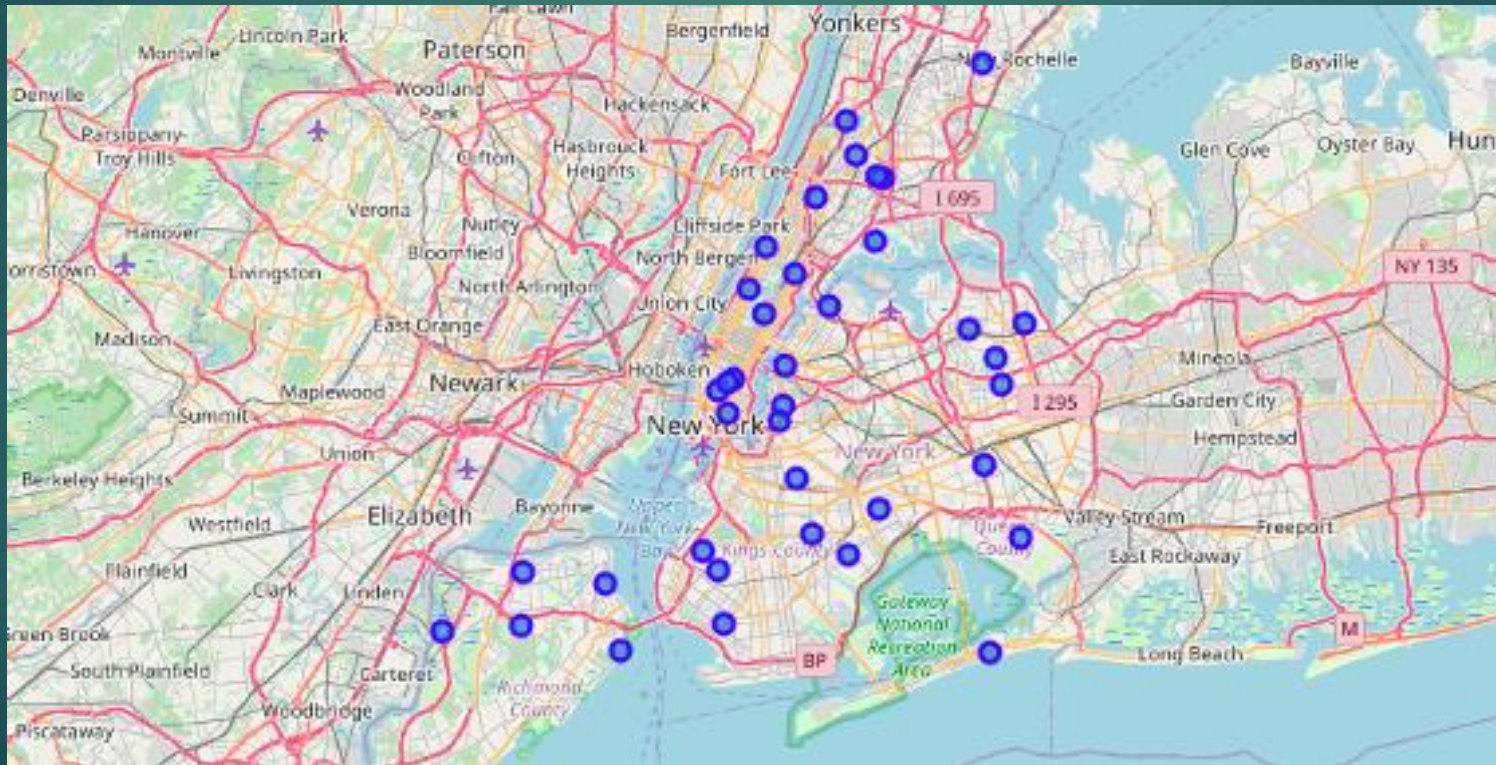
New York City Visualization

▶ New York City Map with all nieghborhoods displayed WITH clustering

Neighborhoods in Each Cluster

- 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

Min, Max, Mean of each Cluster (Ozone and Fine Particulate Matter)

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

- Looking at Mean values for each cluster:
  - 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

Conclusion and Future

- Using Air Quality and Health Data with Foursquare venue data k-means clustering was performed.
- There were 4 clusters
  - These clusters of neighborhoods shared similar air quality levels as well as similar surrounding venues.
- This information would be helpful to someone moving to New York City or a different part of the city based on air quality.
  - They would have a set of options where they could compare air quality and 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.