# THE BATTLE OF NEIGHBORHOODS – WEEK 2

# CHOICE OF NEIGHBORHOOD FOR AIRBNB STAY IN NEW YORK CITY

## INTRODUCTION

### Problem Background:

For this project, the chosen location is New York City, NY in United States. New York City has been described as the cultural, financial, media, and entertainment capital of the world. It boasts of several world renowned tourist attractions and restaurants which draws travelers from all across the world.

Airbnb is an online marketplace where members can use the service to arrange lodging or homestays for short or long durations. Airbnb properties exist in many countries and New York City has a lot of options for stay using Airbnb.

### Problem Description:

Domestic and International travelers who travel for work or vacation generally tend to stay closer to the places of interest and usually have little or no knowledge regarding the location and place of stay. Also, as with any tourist city in the world, the crime rate in Manhattan is usually high.

One of the major problems faced by travelers during the course of stay in a new city is their personal safety. While being in the proximity of the points of interest, is it possible to come up with a solution wherein a safer neighborhood can be determined within the available Airbnb properties using the past crime data?

With the available data, this project is attempting to find a safer neighborhood to stay at an Airbnb property by analyzing the locations of Crime, with the geo locations of Airbnb properties along with analyzing the density of Foursquare data of New York listing out the points of interest (Venues). This project examines and analyzes the data using tools available in Python to explore the following:

1. Demographics of Airbnb listings
2. Exploratory analysis with trends in crime in Manhattan
3. Access (distance) to nearby places of interest

Target Audience

To recommend a suitable location, a Travel company appointed me with an objective of locating and recommending the Management – a safer neighborhood in New York city which is closer to the Points of interest. . The Management understands the limitations in this project for the recommendations made.

Success Criteria

The success criteria of this project is a good recommendation of neighborhood to stay during the visit to New York City with lesser crime rate and easily accessible venues around the Airbnb property.

## DATA

For solving the above problem, following data is used:

1. http://tomslee.net/airbnb-data-collection-get-the-data - Details of Airbnb listings (in .csv format) with their location coordinates and ratings included for the City of New York where only data in 2017 (2017-07-12) -is used for this project.
   Limitations : Only first 5000 Listings are loaded which may result in the final recommendation being skewed

| | borough | neighborhood |
|---|---|---|
| 0 | Queens | Jackson Heights |
| 1 | Brooklyn | Cypress Hills |
| 2 | Brooklyn | Sheepshead Bay |
| 3 | Manhattan | Hell's Kitchen |
| 4 | Manhattan | Upper East Side |

2. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data - New York City crime data acquired from City of New York website (.csv file) which includes the type of crime, location of crime including the co-ordinates is used to overlay on the Map of New York using Folium Libraries in Python.

   Limitations: Due to the large size of dataset (464035 rows), the heat map generated may look misleading but is clarified using the Bar graphs

| | CMPLNT_NUM | ADDR_PCT_CD | BORO_NM | CMPLNT_FR_DT | CMPLNT_FR_TM | CMPLNT_TO_DT | CMPLNT_TO_TM | CRM_ATPT_CPTD_CD | HADEVELOPT | HOUSING_PSA | ... | SUSP_SEX | TRANSIT_DISTRICT | VIC_AGE_GROUP | VIC_RACE | VIC_SEX | X_COORD_CD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 651421035 | 41.0 | BRONX | 11/28/2018 | 00:00:00 | 11/28/2018 | 00:01:00 | COMPLETED | NaN | NaN | ... | U | NaN | 25-44 | BLACK | M | NaN |
| 1 | 149013323 | 14.0 | MANHATTAN | 12/31/2018 | 23:40:00 | 12/31/2018 | 23:50:00 | COMPLETED | NaN | NaN | ... | NaN | NaN | 45-64 | WHITE | F | 987866.0 |
| 2 | 642981531 | 73.0 | BROOKLYN | 12/31/2018 | 23:30:00 | 12/31/2018 | 23:40:00 | COMPLETED | NaN | NaN | ... | F | NaN | 18-24 | WHITE HISPANIC | F | 1006995.0 |
| 3 | 429685363 | 67.0 | BROOKLYN | 12/31/2018 | 23:20:00 | 12/31/2018 | 23:30:00 | COMPLETED | NaN | NaN | ... | M | NaN | 25-44 | BLACK | F | 999584.0 |
| 4 | 290330841 | 5.0 | MANHATTAN | 12/31/2018 | 23:15:00 | 12/31/2018 | 23:20:00 | COMPLETED | NaN | NaN | ... | U | NaN | 25-44 | WHITE | M | 986164.0 |

3. Data set for New York Neighborhood derived from earlier assignment - https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json - .json file is used to map the Neighborhood with its coordinates using folium libraries

| | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| **301** | Manhattan | Hudson Yards | 40.756658 | -74.000111 |
| **302** | Queens | Hammels | 40.587338 | -73.805530 |
| **303** | Queens | Bayswater | 40.611322 | -73.765968 |
| **304** | Queens | Queensbridge | 40.756091 | -73.945631 |
| **305** | Staten Island | Fox Hills | 40.617311 | -74.081740 |

4. Foursquare APIs – New York city geographical coordinates are used as input, Foursquare API calls are made to acquire venues in the prescribed radius

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **0** | Wakefield | 40.730862 | -73.987156 | Sake Bar Decibel | 40.729416 | -73.987745 | Sake Bar |
| **1** | Wakefield | 40.730862 | -73.987156 | Momofuku Ssäm Bar | 40.731711 | -73.985571 | Asian Restaurant |
| **2** | Wakefield | 40.730862 | -73.987156 | Han Dynasty | 40.732130 | -73.988090 | Chinese Restaurant |
| **3** | Wakefield | 40.730862 | -73.987156 | Hi-Collar - ハイカラ (Hi-Collar) | 40.729449 | -73.985918 | Coffee Shop |
| **4** | Wakefield | 40.730862 | -73.987156 | Casey Rubber Stamp | 40.729962 | -73.985098 | Arts & Crafts Store |

# METHODOLOGY

## Analytical Approach

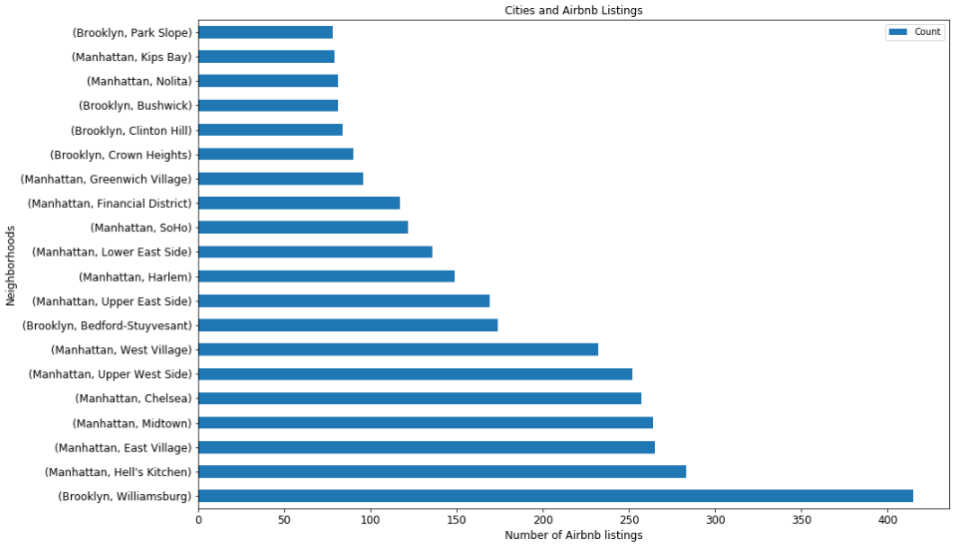
New York City is divided into 5 boroughs and 36 neighborhoods. In this project, the first part is cleaning up and analyzing Airbnb datasets to understand and visualize the density of properties in the above 5 boroughs. The second part carries out exploratory analyses of crime data available to ascertain the most and least crime neighborhoods. The Third part extracts the 'Venues' in the New York neighborhoods using data available and clustering by means of K-means to visualize and conclude the density of Venues in neighborhoods. Combining the above, the conclusions were drawn and recommended.

## Exploratory Data Analysis

### Data 1 –

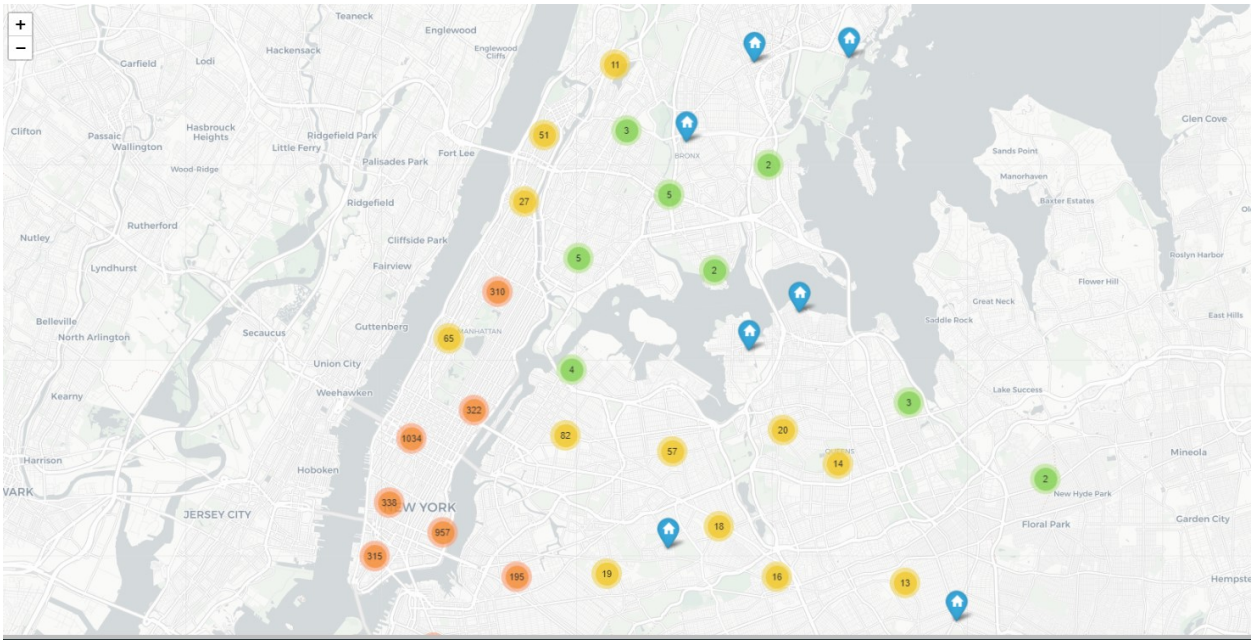
- Airbnb properties in NYC data acquired from [http://tomslee.net/airbnb-data-collection-get-the-data](http://tomslee.net/airbnb-data-collection-get-the-data)
- Above dataset is in the form of a .csv file.
- As the dataset was huge, only the first 5000 rows were loaded
- Transformed the above .csv file into a dataframe using Pandas
- Data cleansing carried out by dropping irrelevant columns and grouping property details by borough and neighborhood to acquire statistics

With the available condensed statistical data (condensed to top 20 listings), a horizontal graph is plotted showing the number of properties on X axis and Borough, Neighborhoods on Y Axis.



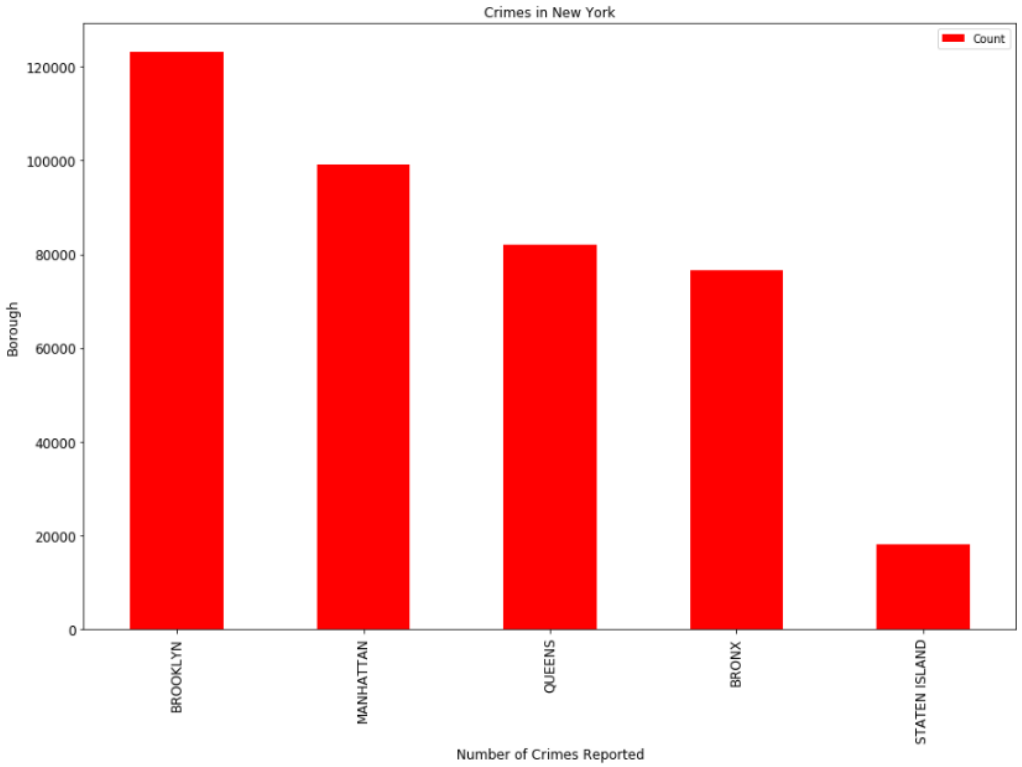

It is evident from the above graph that after Williamsburg neighborhood in Brooklyn, several neighborhoods in Manhattan (Hell's Kitchen, East Village, Midtown, Chelsea, Upper West Side etc) have a number of properties listed. By overlaying the coordinates on the map using FastMarkerCluster, the density of Airbnb properties are observed much greater in (lower) Manhattan Area.

New York Airbnb property listings visualization

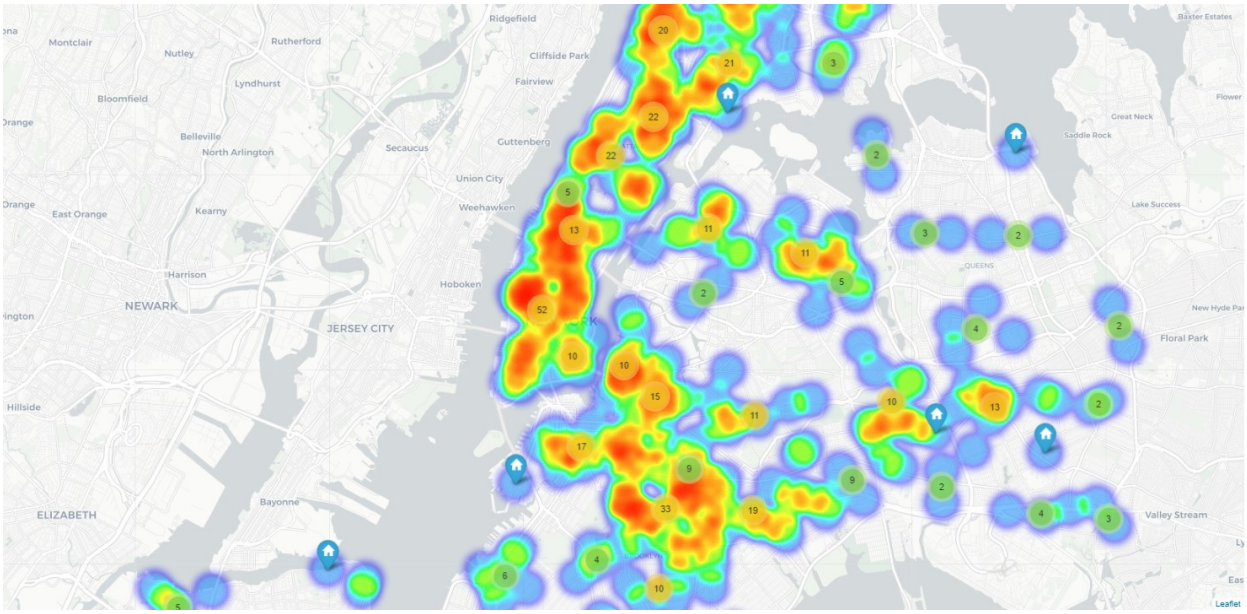



Data 2 –

- New York City crime data acquired from City of New York website (.csv file) which includes the type of crime, location of crime including the co-ordinates
- https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data
- The dataset is a .csv file which was loaded and converted into a dataframe using Pandas
- The dataset contains 464035 rows which were not condensed except for removal of NA values.
- By cleaning the dataset of irrelevant columns and removing NaN values, the dataframe is grouped by borough and neighborhood counting the number of crimes reported in each neighborhood.
- For visualization and statistical inferencing, total count of crimes in each neighborhood is calculated and plotted on a bar graph.




- With the above graph, it is evident that the maximum reported crimes were in Brooklyn followed by Manhattan and Queens Borough.
- A heat map is generated overlaying the coordinates of crimes reported in the 5 boroughs with the color depicting the intensity/count of number of crimes.




Heat map of New York City showing the areas of crime reported

Please note that due to the large size of dataset (464035 rows), the heat map generated may look misleading but is clarified using the above bar graph.

## Data 3

- Data set for New York Neighborhood derived from earlier assignment - https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json - .json file is used to map the Neighborhood with its coordinates using folium libraries
- Ny_neighborhoods dataframe created with assigned columns
- Populated the above dataframe with the imported json data and saving it as .csv file for further use.
- Borough names with location coordinates dataframe generated with above data.

## Data 4

Foursquare APIs – New York city geographical coordinates are used as input, Foursquare API calls are made to acquire venues in the prescribed radius

- Using Nominatim, the geo coordinates of New York were retrieved.
- The URL was defined for pulling venue details from Foursquare API by passing credentials
- Neighborhoods and nearby venue details were populated in the dataframe

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Wakefield | 40.730862 | -73.987156 | Sake Bar Decibel | 40.729416 | -73.987745 | Sake Bar |
| 1 | Wakefield | 40.730862 | -73.987156 | Momofuku Ssäm Bar | 40.731711 | -73.985571 | Asian Restaurant |
| 2 | Wakefield | 40.730862 | -73.987156 | Han Dynasty | 40.732130 | -73.988090 | Chinese Restaurant |
| 3 | Wakefield | 40.730862 | -73.987156 | Hi-Collar - ハイカラ (Hi-Collar) | 40.729449 | -73.985918 | Coffee Shop |
| 4 | Wakefield | 40.730862 | -73.987156 | Casey Rubber Stamp | 40.729962 | -73.985098 | Arts & Crafts Store |

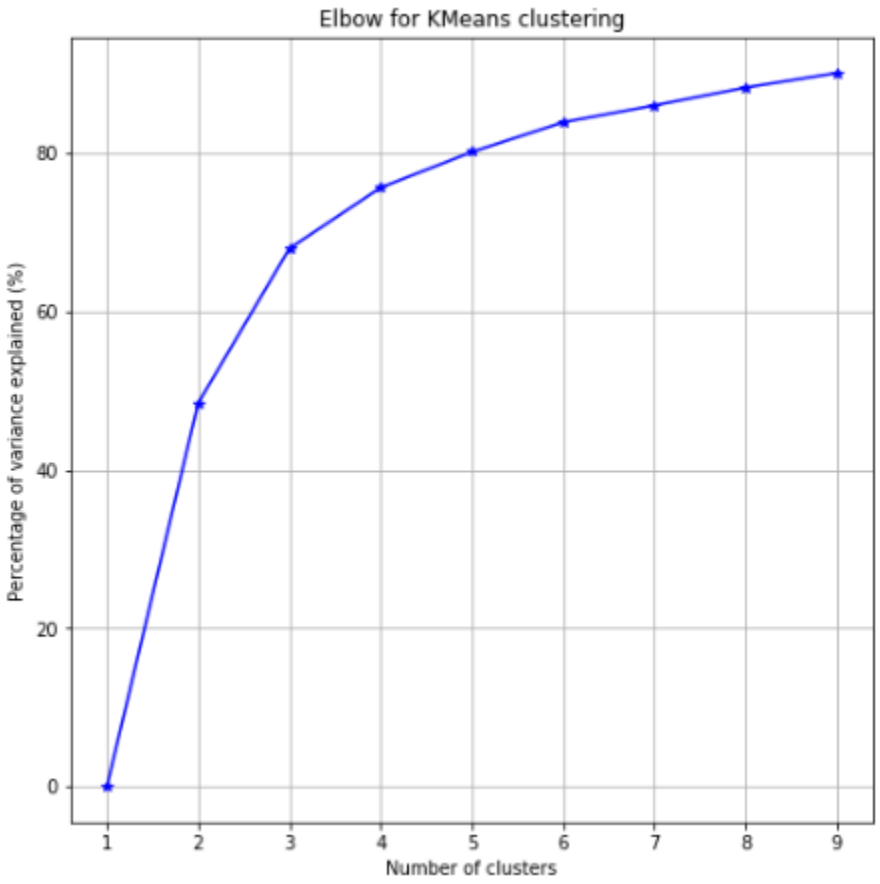
- Cleanup of dataset carried out by dropping irrelevant columns and carried out one hot encoding

For clustering the Venues in New York, the restaurants in the venues list were taken. One hot encoding was carried out to the restaurant dataframe pulled from the original New York venues dataframe

| | Neighborhood | American Restaurant | Arepa Restaurant | Asian Restaurant | Chinese Restaurant | Greek Restaurant | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Mediterranean Restaurant | ... | Ramen Restaurant | Seafood Restaurant | Soba Restaurant | Spanish Restaurant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Wakefield | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 |
| 1 | Wakefield | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 |
| 2 | Wakefield | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 |
| 3 | Wakefield | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 |
| 4 | Wakefield | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 |

5 rows × 22 columns

To find the number of clusters, the elbow method is used wherein the % variance drops drastically at k=2. This is also confirmed by the Silhouette analysis by using the silhouette score (which is not included in the notebook).




Elbow method for determining the optimum number of clusters

## RESULTS

From the venues data, we filtered restaurant data for all 5 Boroughs clustering. This can be extended to other points of interest but the scope is beyond the means of this project.

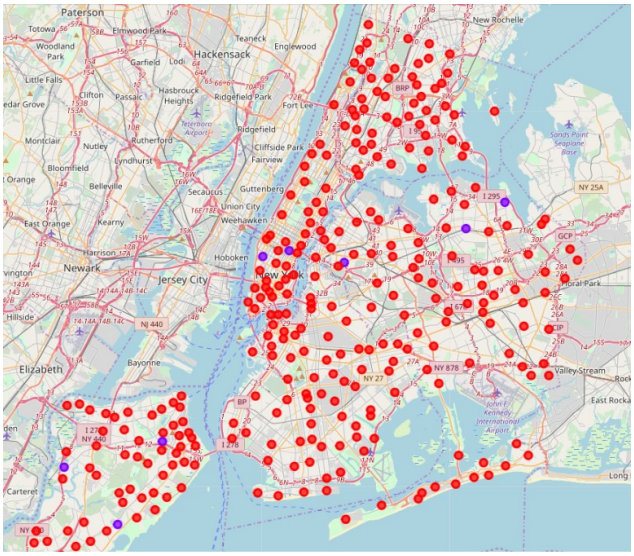
### Neighborhood K-Means clustering based on mean occurrence of Venue Category

To cluster neighborhoods into two clusters (arrived at 2 clusters as explained above), K-Means clustering Algorithm is used. K-Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with nearest mean . It uses iterative refinement approach.

The cluster total and cluster sum were arrived at after clustering as below  with Cluster 0 having less total and sum than Cluster 1

| | American Restaurant | Arepa Restaurant | Asian Restaurant | Chinese Restaurant | Greek Restaurant | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Mediterranean Restaurant | Mexican Restaurant | ... | Soba Restaurant | Spanish Restaurant | Szechuan Restaurant | Tapas Restaurant | Udon Restaurant | Ukrainian Restaurant | Vegetarian / Vegan Restaurant | Vietnamese Restaurant | Total | Total Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cluster0 | 1.0 | 1.0 | 1.0 | 3.0 | 1.0 | 1.0 | 3.0 | 2.0 | 1.0 | 2.0 | ... | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 3.0 | 2.0 | 33.0 | 66.0 |
| cluster1 | 2.0 | 2.0 | 2.0 | 6.0 | 2.0 | 2.0 | 6.0 | 4.0 | 2.0 | 4.0 | ... | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 6.0 | 4.0 | 66.0 | 132.0 |

In the below cluster visualization, we can see the types of clusters created using K-Means



Cluster 0: The Total and Total Sum of Cluster 0 is smaller (smallest of 2 clusters) which shows there is a shortage of Venues to explore (in comparison with Cluster 1)

Cluster 1: The Total and Total Sum of Cluster 0 has the highest value which shows the number of restaurants are very high

## DISCUSSION

There is a scope for improvement of above model by refining the following:

- Considering more points of interest (e.g. Concert halls, Museums and carrying out clustering) in developing the model
- Carrying out clustering of Crime data which may give more refined clusters than currently available visualizations and statistics
- Linear regression relationship between Airbnb property locations and Venues
- Multilinear or polynomial regression relationship between
    - Airbnb property listing
    - Crime in the neighborhood
    - Venues (points of interest)

### Assumptions/Limitations

- Dataset of Airbnb listings was restricted to 5000 rows which may have given us skewed results
- It is considered that all Airbnb properties are available for occupancy for this project
- Foursquare API radius limited to 1000 and results limited to 500
- Due to large crime dataset, random 500 rows were taken for visualization which may have skewed the result
- Maps for Airbnb, Crime listings and Venues created separately as- due to high number of data points, the Folium maps are not rendering in the Jupyter notebook for overlay of the above data points in a single map

## CONCLUSION

The analysis is performed on limited/restricted data. Brooklyn and Manhattan were close in the statistics of Airbnb listings although Manhattan has higher density of listings per sq. mile. The maximum reported crimes as per analysis were in Brooklyn followed by Manhattan. The clustering of restaurants has Cluster 1 spread over Brooklyn and Manhattan almost evenly. In view of the above statistics, Lower Manhattan is the desired location for staying at an Airbnb with relatively lower crime statistics and more Venues to explore.