Exploring Airbnb's price for various listings in the neighborhood of New York city

Amrutha Veena | 06-23-2020

INTRODUCTION

Airbnb, Inc. is an American online marketplace company based in San Francisco, California, United States. Airbnb offers arrangement for lodging, primarily homestays, or tourism experiences. It is common nowadays for people to book stays in different cities for many purposes including holidays, business meetings etc.

Often tourists and travelers have problem in identifying the best stay that accommodates their budget as well as a place that is near to many popular venues in the city they are visiting. This project aims at classifying the different price listings of Airbnb in the neighborhood of New York city. New York city has a total of five boroughs and at least 150 neighborhoods. We will be using the New York city's geospatial data and the Foursquare API to explore the most popular venues in New York city and the corresponding Airbnb's price listings in that neighborhood and categorize the prices as LOW, MID-1, MID-2 and HIGH.

We will also explore the borough with the highest number of venues and cluster the neighborhoods of that borough using K-means algorithm and understand the different clusters to which the neighborhoods and their top 10 venues belong to.

DATA

For this project we will be using the FourSquare API data along with the Airbnb's [1] open data of it's price listings in the neighborhood of New York city. The data consists of 16 columns describing the listings, price, neighborhood, number of reviewers, ratings and so on. Using the above data along with the FourSquare API , we can explore the various venues near a particular listing thus helping the user to decide as to whether the price for the listing is worth or not.

METHODOLOGY

The first step is to read the Airbnb New York listings' data that is downloaded from [1], then examine the data.

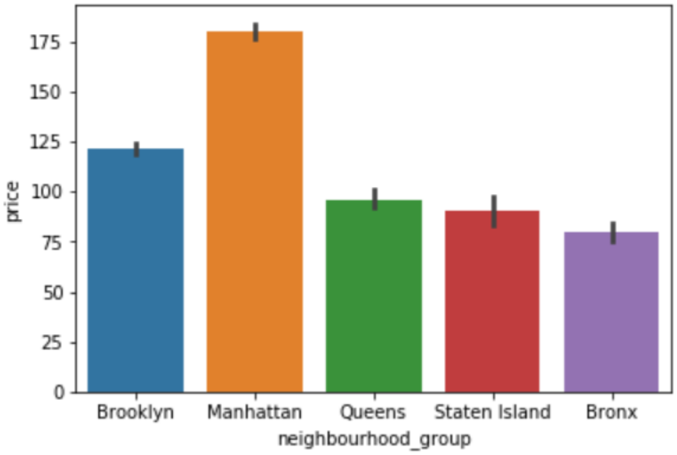
A. Data Wrangling

The above data set needs to be cleaned and all the missing values have to be accounted for. First we examine the type of each column in the data set and we see if there are any anomalies and change it to a corresponding data type. Then find the number of missing values in each column and account for the missing value by either removing the rows containing NaN values or replacing the Nan value with the mean of that column. Based on the data we can adopt any method to handle missing values.

B. Exploratory Data Analysis and Visualization

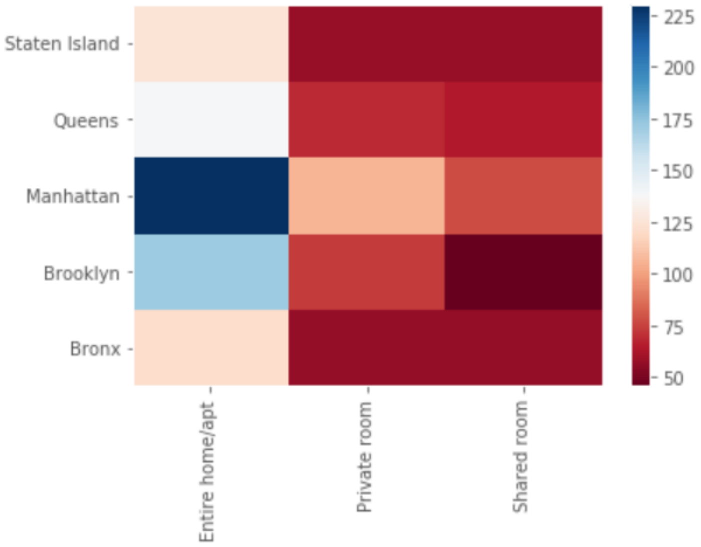
In this step let us answer the following questions through visualizations and EDA.

Question 1: Which neighborhood group or borough of New York city has the highest price?



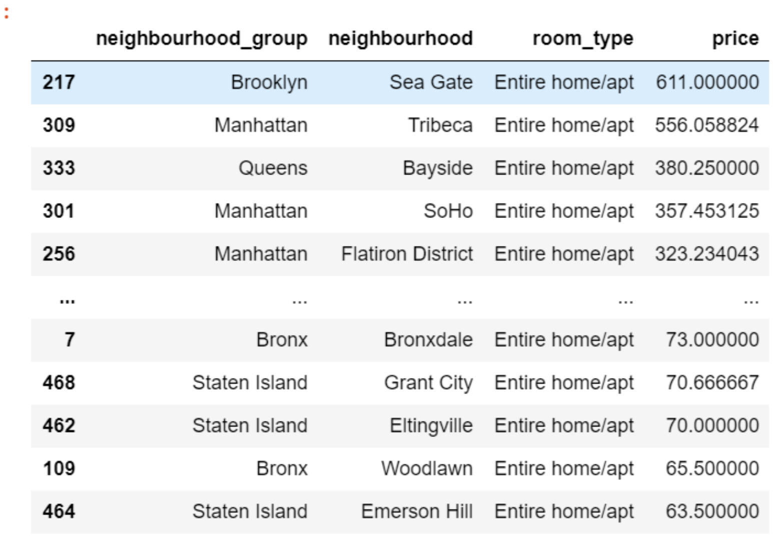
From the above plot we see that Manhattan has the highest average price of all the listings of Airbnb in New York.

Question 2: How does the price vary in different boroughs by room_type?

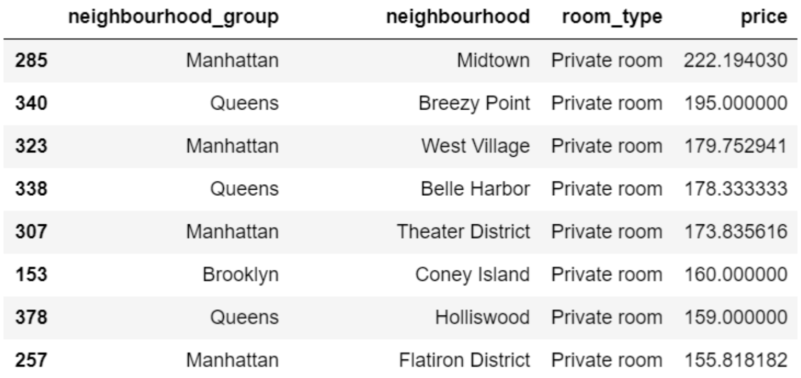


We see that Entire home/apt in Manhattan has the highest price compared to other boroughs

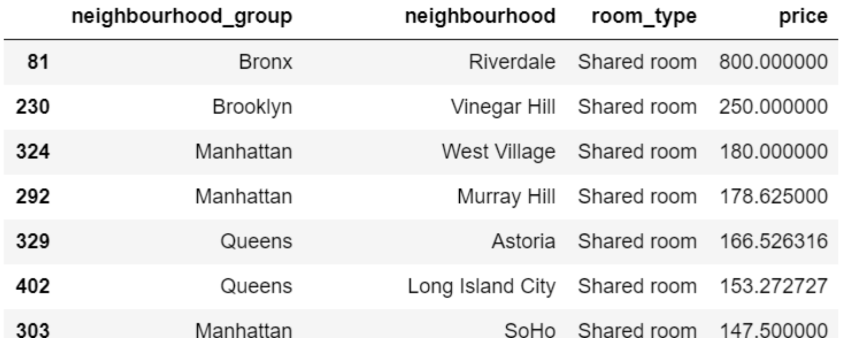
Question 3: Which neighborhood has the highest average price for different room types?



	neighbourhood_group	neighbourhood	room_type	price
217	Brooklyn	Sea Gate	Entire home/apt	611.000000
309	Manhattan	Tribeca	Entire home/apt	556.058824
333	Queens	Bayside	Entire home/apt	380.250000
301	Manhattan	SoHo	Entire home/apt	357.453125
256	Manhattan	Flatiron District	Entire home/apt	323.234043
...	...	...	...	...
7	Bronx	Bronxdale	Entire home/apt	73.000000
468	Staten Island	Grant City	Entire home/apt	70.666667
462	Staten Island	Eltingville	Entire home/apt	70.000000
109	Bronx	Woodlawn	Entire home/apt	65.500000
464	Staten Island	Emerson Hill	Entire home/apt	63.500000



	neighbourhood_group	neighbourhood	room_type	price
285	Manhattan	Midtown	Private room	222.194030
340	Queens	Breezy Point	Private room	195.000000
323	Manhattan	West Village	Private room	179.752941
338	Queens	Belle Harbor	Private room	178.333333
307	Manhattan	Theater District	Private room	173.835616
153	Brooklyn	Coney Island	Private room	160.000000
378	Queens	Holliswood	Private room	159.000000
257	Manhattan	Flatiron District	Private room	155.818182



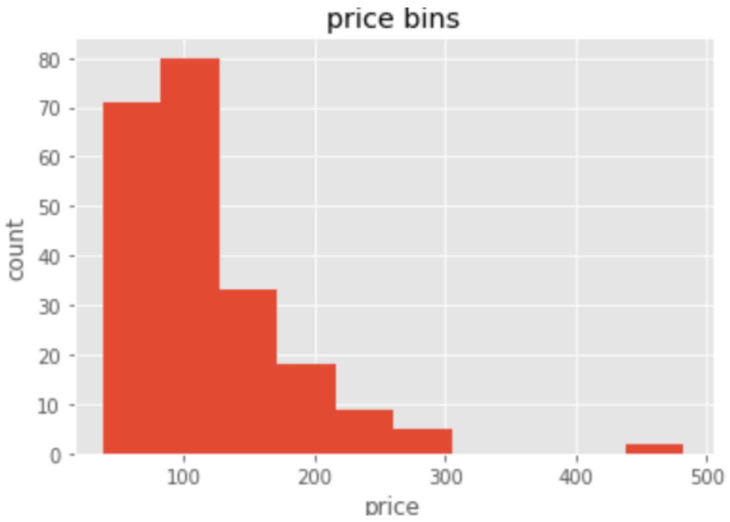
	neighbourhood_group	neighbourhood	room_type	price
81	Bronx	Riverdale	Shared room	800.000000
230	Brooklyn	Vinegar Hill	Shared room	250.000000
324	Manhattan	West Village	Shared room	180.000000
292	Manhattan	Murray Hill	Shared room	178.625000
329	Queens	Astoria	Shared room	166.526316
402	Queens	Long Island City	Shared room	153.272727
303	Manhattan	SoHo	Shared room	147.500000

Riverdale of Bronx has highest average price for a shared room and Midtown of Manhattan private room, while Sea Gate of Brooklyn has the highest average price for Entire home/apt room type

C. Segmenting and clustering the neighborhoods of New York city

Now that we have done basic EDA, let us focus on analyzing the popular venues around these boroughs. Before that let us analyze the frequency of prices in different ranges.

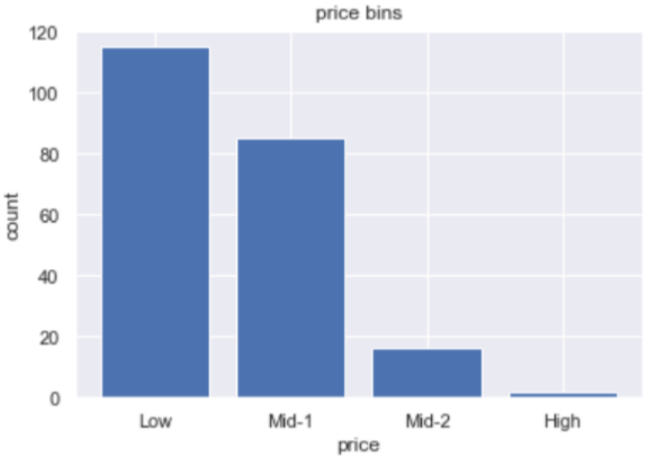
Let us do data binning for the price table and divide the price ranges to four bins as shown below.



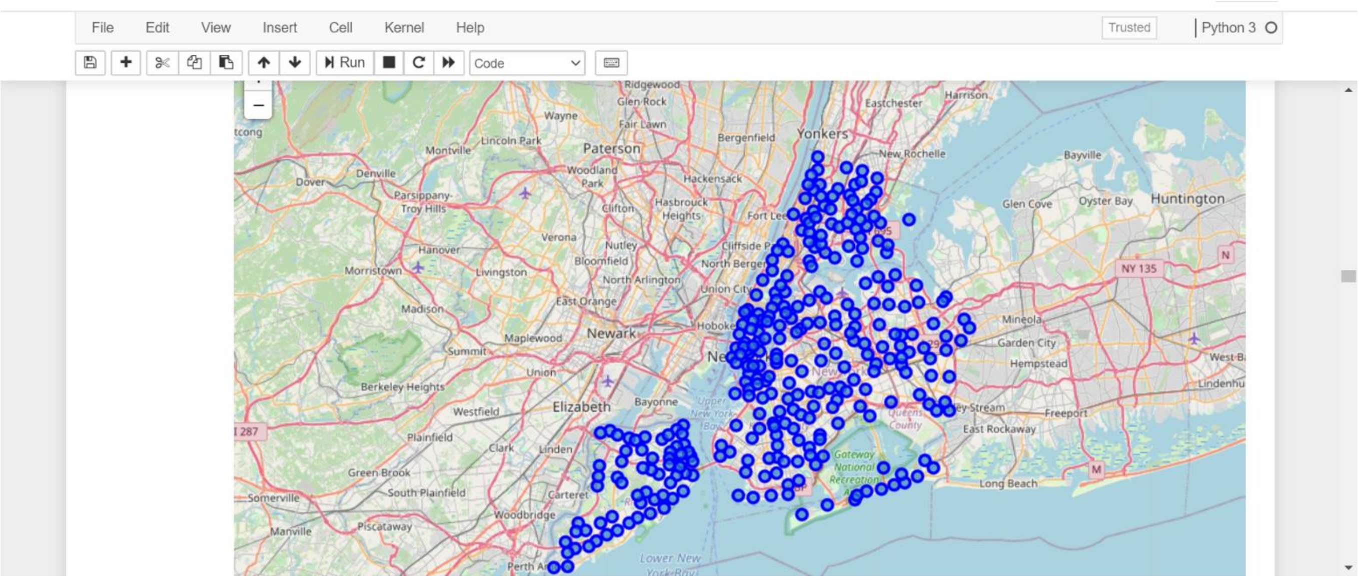
So the ranges are as follows:

1. Low : < 100
2. Mid-1 : >=100 and , <200
3. Mid-2 : >=200 and <300
4. High : >=300

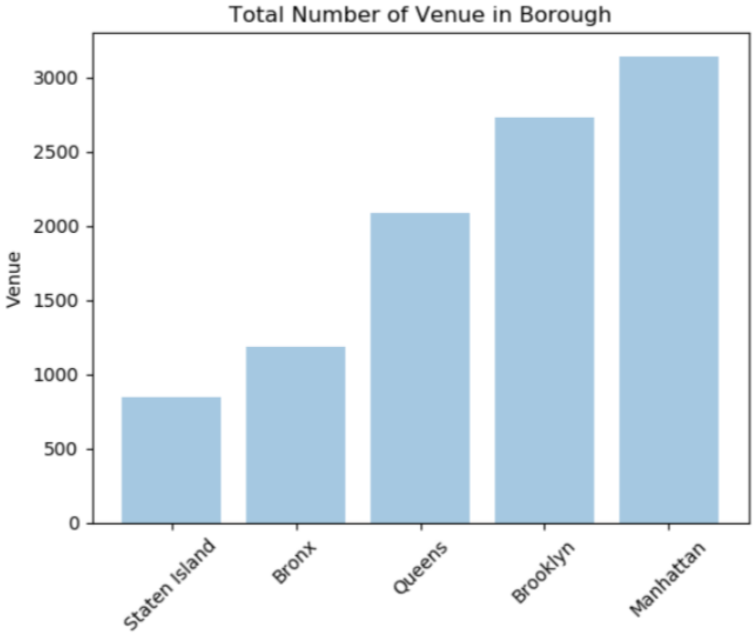
the corresponding price bins are shown below.



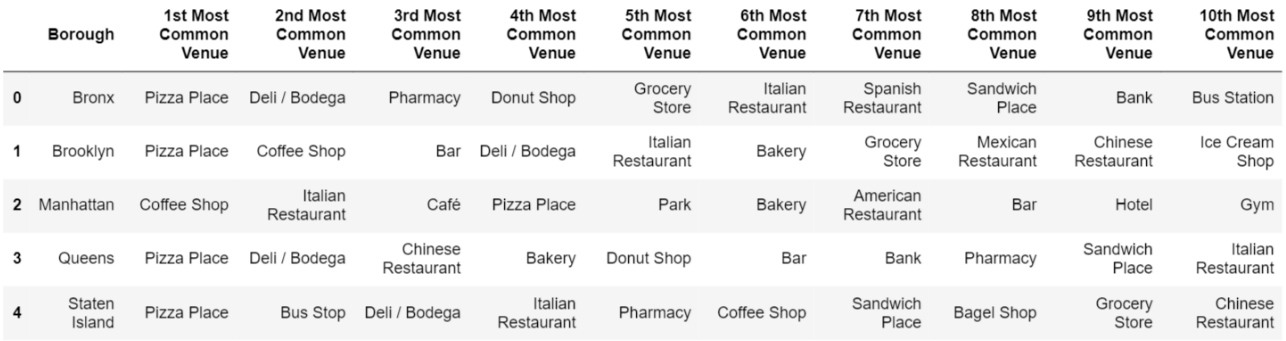
Now lets us use the geopy package to obtain the latitude and longitude of New York city and plot the folium map as shown below.



Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them. Using the FourSquare API we can get the nearby venues of each neighborhood and/or the borough. The below plot shows the number of venues for each borough.



Now let us find the top 10 most common venue in each borough as shown below.



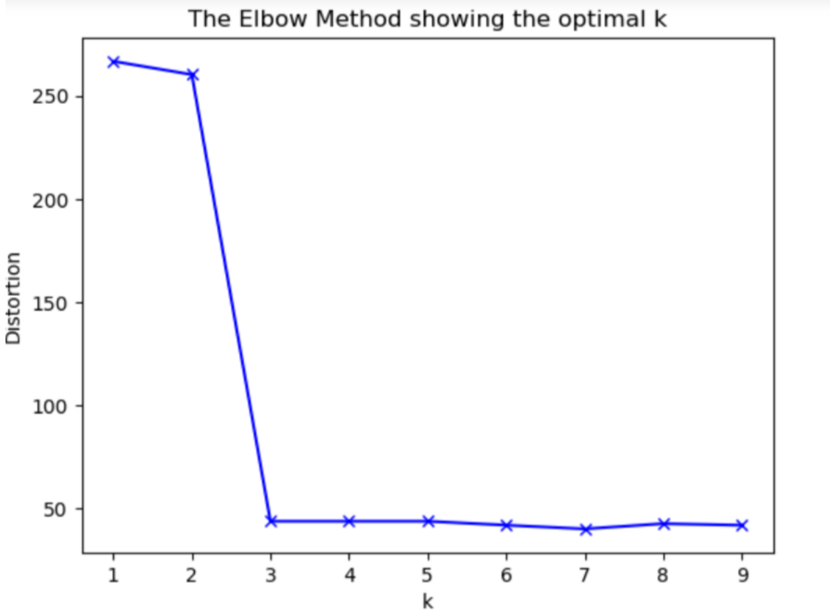
	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	Pizza Place	Deli / Bodega	Pharmacy	Donut Shop	Grocery Store	Italian Restaurant	Spanish Restaurant	Sandwich Place	Bank	Bus Station
1	Brooklyn	Pizza Place	Coffee Shop	Bar	Deli / Bodega	Italian Restaurant	Bakery	Grocery Store	Mexican Restaurant	Chinese Restaurant	Ice Cream Shop
2	Manhattan	Coffee Shop	Italian Restaurant	Café	Pizza Place	Park	Bakery	American Restaurant	Bar	Hotel	Gym
3	Queens	Pizza Place	Deli / Bodega	Chinese Restaurant	Bakery	Donut Shop	Bar	Bank	Pharmacy	Sandwich Place	Italian Restaurant
4	Staten Island	Pizza Place	Bus Stop	Deli / Bodega	Italian Restaurant	Pharmacy	Coffee Shop	Sandwich Place	Bagel Shop	Grocery Store	Chinese Restaurant

D. Cluster the neighborhoods with the highest number of venues

As we saw above, Manhattan has the highest number of venues, with 282 unique venue categories. So let us cluster the neighborhoods of Manhattan and find the top 10 venues each of its neighborhood.

We will be using the K-means algorithm to cluster the neighborhoods of Manhattan. *k*-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. It is an unsupervised learning algorithm.

In order to find the optimal value of K in K-means algorithm we will use the elbow technique. The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

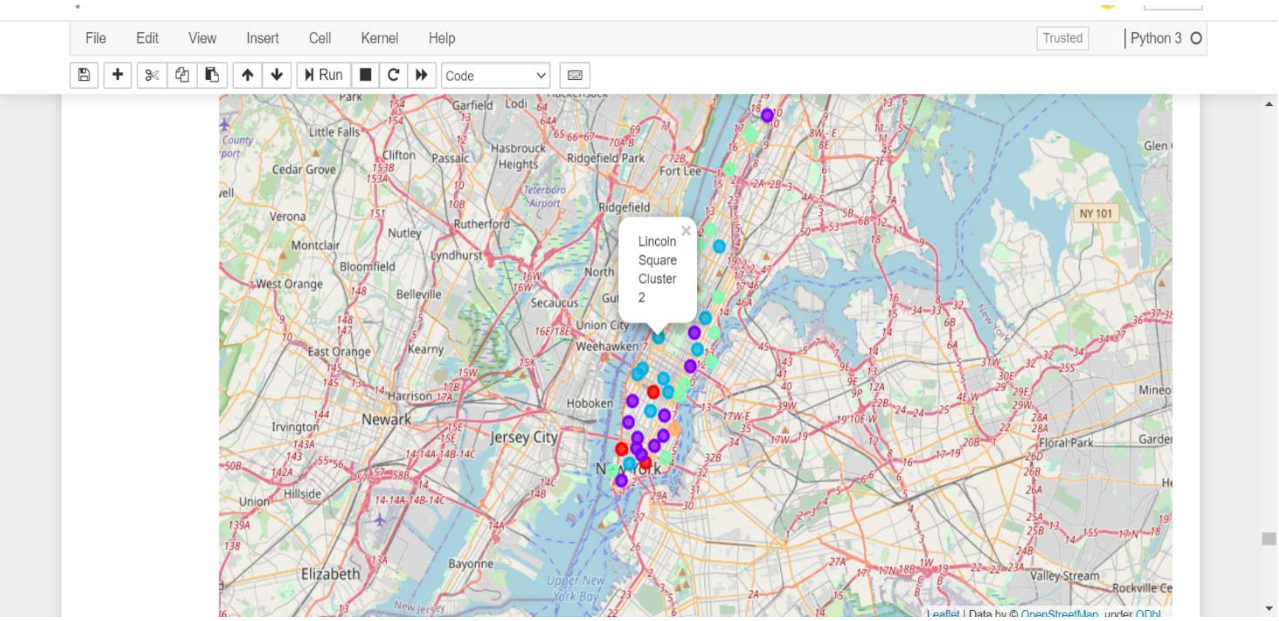


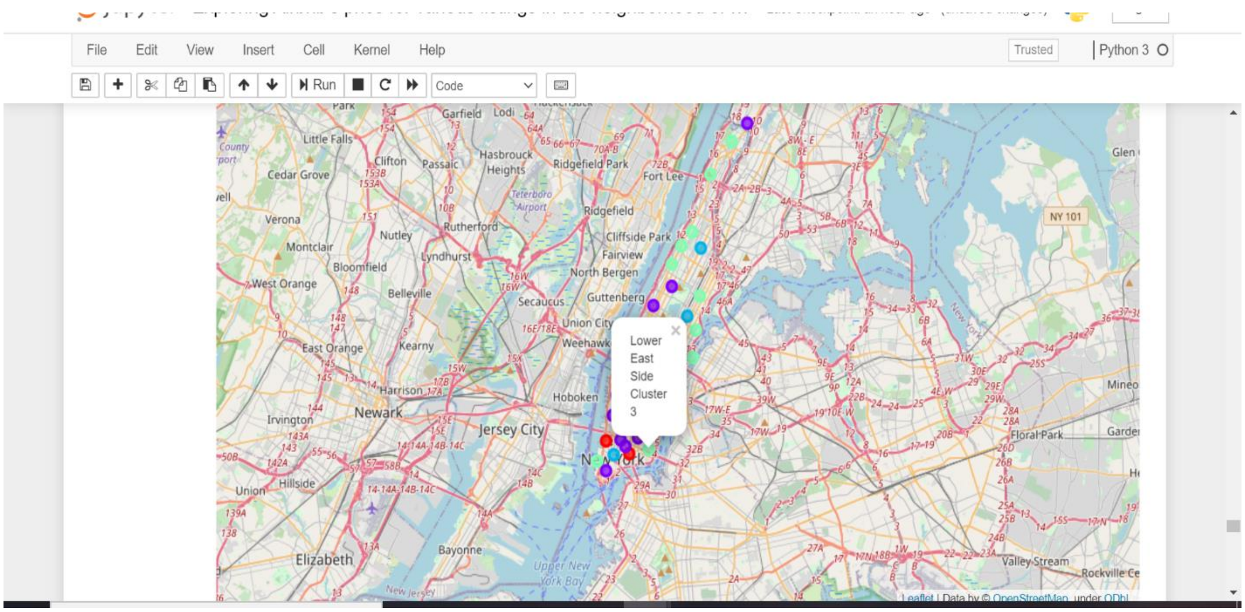
To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 5.

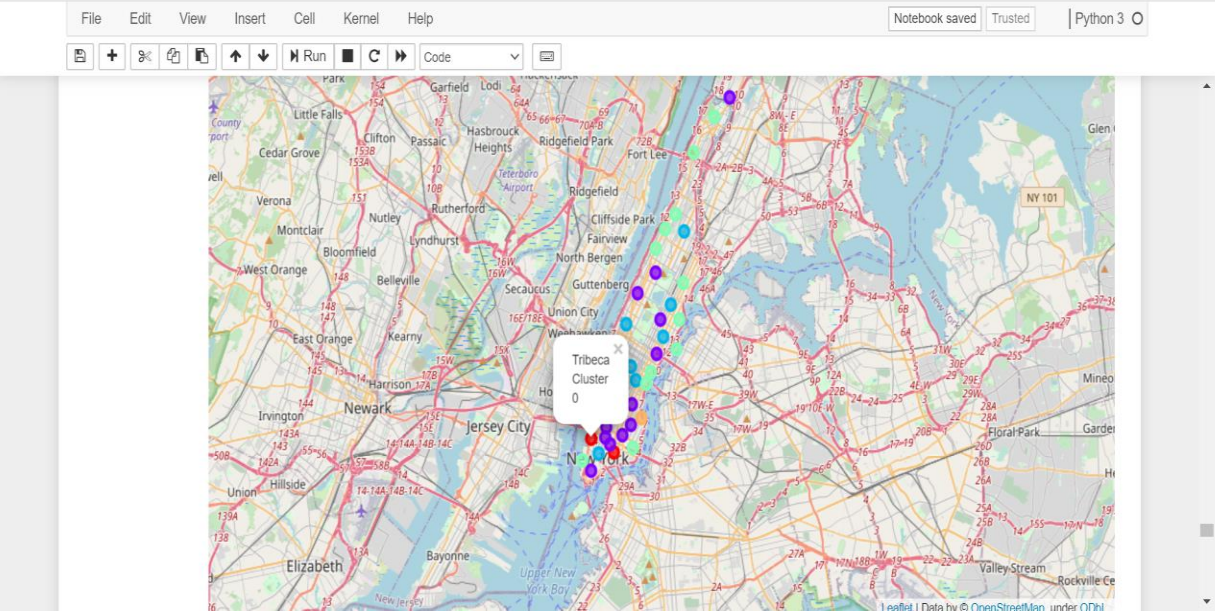
RESULTS

A. Clustered neighborhoods of Manhattan

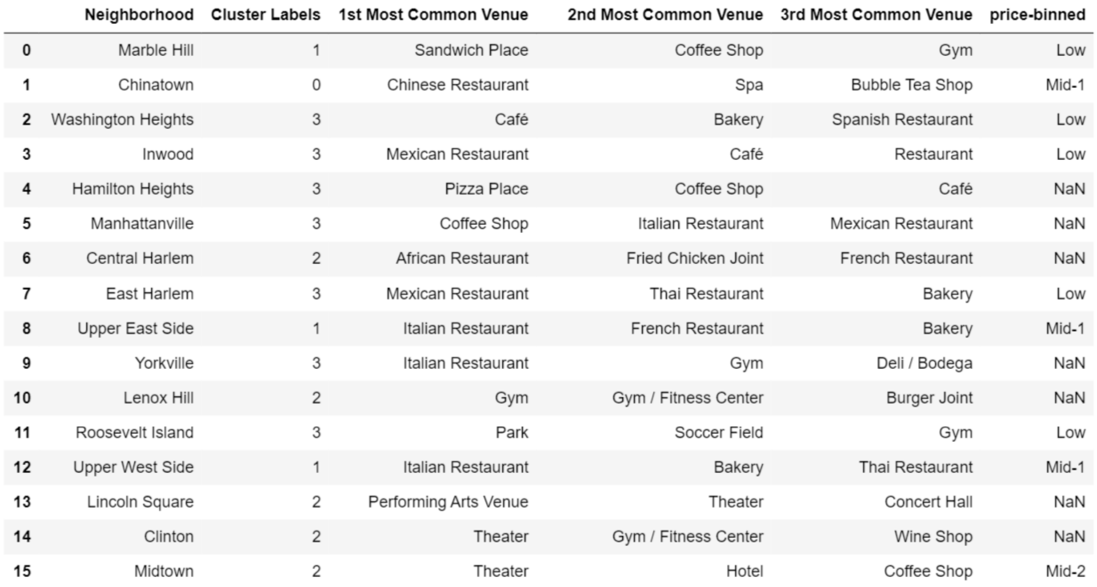
The below figures show some of the clusters to which the neighborhoods of Manhattan belong.




File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Python 3
Run
Code
Lower East Side Cluster 3
Leaflet | Data by © OpenStreetMap, under ODbL



B. **The price bin to which each of the neighborhood listings in Airbnb belong to along with their cluster labels and their top 3 venues.**



	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	price-binned
0	Marble Hill	1	Sandwich Place	Coffee Shop	Gym	Low
1	Chinatown	0	Chinese Restaurant	Spa	Bubble Tea Shop	Mid-1
2	Washington Heights	3	Café	Bakery	Spanish Restaurant	Low
3	Inwood	3	Mexican Restaurant	Café	Restaurant	Low
4	Hamilton Heights	3	Pizza Place	Coffee Shop	Café	NaN
5	Manhattanville	3	Coffee Shop	Italian Restaurant	Mexican Restaurant	NaN
6	Central Harlem	2	African Restaurant	Fried Chicken Joint	French Restaurant	NaN
7	East Harlem	3	Mexican Restaurant	Thai Restaurant	Bakery	Low
8	Upper East Side	1	Italian Restaurant	French Restaurant	Bakery	Mid-1
9	Yorkville	3	Italian Restaurant	Gym	Deli / Bodega	NaN
10	Lenox Hill	2	Gym	Gym / Fitness Center	Burger Joint	NaN
11	Roosevelt Island	3	Park	Soccer Field	Gym	Low
12	Upper West Side	1	Italian Restaurant	Bakery	Thai Restaurant	Mid-1
13	Lincoln Square	2	Performing Arts Venue	Theater	Concert Hall	NaN
14	Clinton	2	Theater	Gym / Fitness Center	Wine Shop	NaN
15	Midtown	2	Theater	Hotel	Coffee Shop	Mid-2

Note: Refer the jupyter notebook for all the listings.

Also we can call the above 5 clusters as:

Cluser 1: ASIAN FOOD

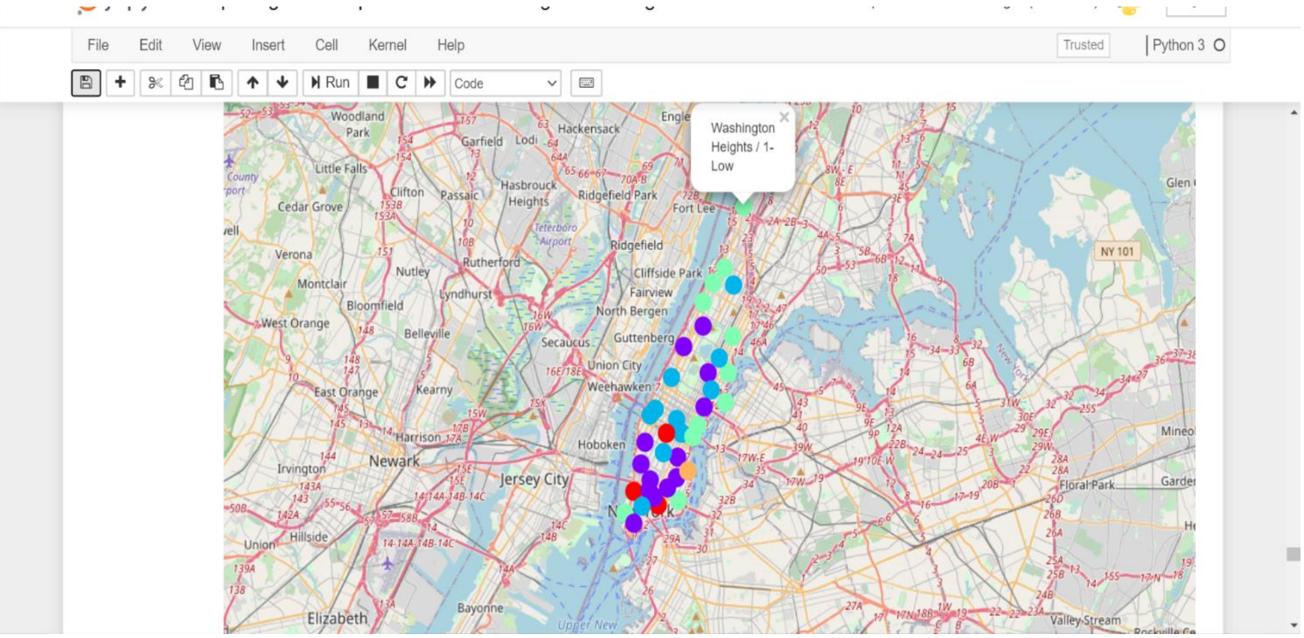
Cluster 2: ITALIAN & CAFE

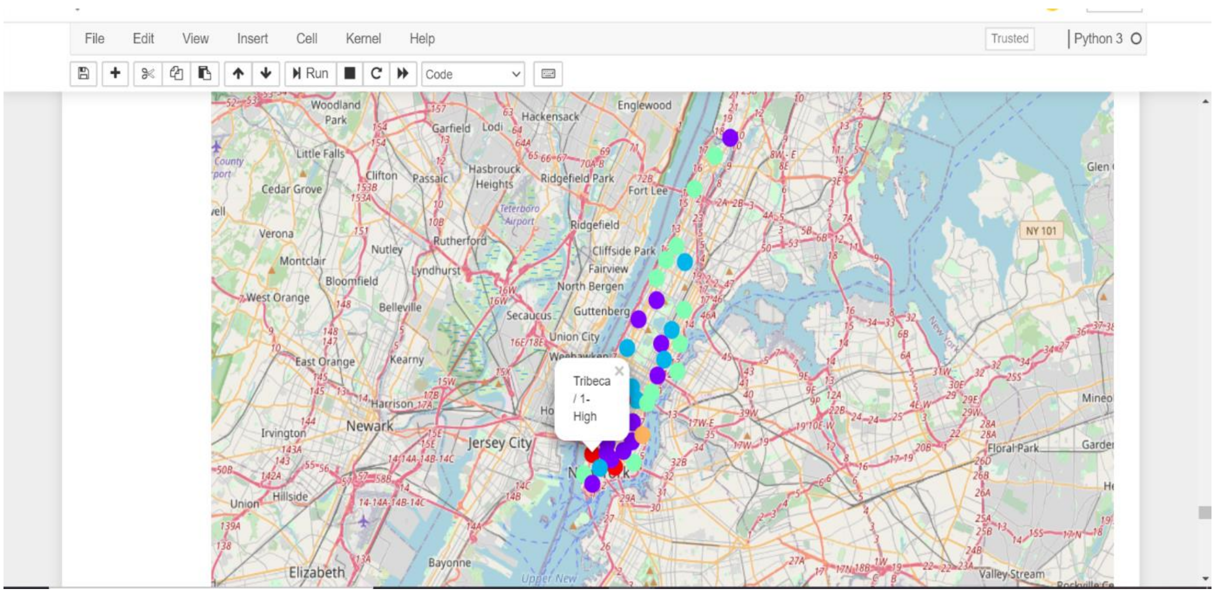
Cluster 3: RECREATION & FOOD

Cluster 4: RECREATION & CAFE

Cluster 5: RECREATION

C. Visualize clusters with price label





Tribeca / 1-High

DISCUSSION

New York City is the most populous city in the United States. New York City comprises 5 boroughs sitting where the Hudson River meets the Atlantic Ocean. At its core is Manhattan, a densely populated borough that's among the world's major commercial, financial and cultural centers.

This project aims at visualizing the popular venues in the city of New York and the average price of Airbnb's service in the neighborhood of interest. I choose Manhattan as it is the borough that has the highest number of venues. This project gives a sense for travelers to easily choose an Airbnb stay in the neighborhood of their interest and also provides a comparison of price with other neighborhoods.

The project can be extended to provide more details of each venue using the premium calls of FourSquare API and much more.

CONCLUSION

From the above observations we see that Tribeca neighborhood of Manhattan has the highest price compared to other neighborhoods. This analysis helps travelers to choose a stay from Airbnb that best suites their budget and also that is near to many popular venues in that area. Having many popular venues near the stay is huge advantage especially in a city like New York, where commute takes quite a lot of time.

Also we can call the above 5 clusters as:

Cluser 1: ASIAN FOOD

Cluster 2: ITALIAN & CAFE

Cluster 3: RECREATION & FOOD

Cluster 4: RECREATION & CAFE

Cluster 5: RECREATION

REFERENCES

[1] https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

[2] FourSquare

[3] New York Geo Json: https://geo.nyu.edu/catalog/nyu_2451_34572