The Last Battle:

The final step in the Battle of the Neighborhoods.

&

A Coursera Capstone Project Report

Authored by

-Abhishek Pandey

Introduction

1.Description and Discussion of the Background

New York CIty is the Epicenter of the arts, and architectural beauty, a. trendsetter. New York City wears many crowns, and spreads an irresistible feast for all. Packed with eye candy of all sorts – architectural glories, Old World cafes, atmospheric booksellers within its compact size and streets – NYC is a wanderer's delight. Walking a few blocks in this jumbled city of 195-plus nationalities will give you the feeling of crossing continents. You can lose yourself in the crowds of Times Square amidst all the dazzling lights or stroll up to Central Park for a serene walk or a quick nap under the shady trees. Every neighborhood in New York City offers a dramatically different version of the city, from the 100-year-old Jewish delis of the Upper West Side to the meandering cobblestone lanes of Greenwich Village.

The New York City is a hub of 306 such neighborhoods that all fall under one of the 5 boroughs, namely - Manhattan, Bronx, Brooklyn, Queens, and Staten Island. The city covers a total area of 783.8 km^2 and is heavily populated with around 8.62 million people. This indicates that the neighborhoods are dense in nature filled with all sorts of retail outlets, restaurants, malls, street food shops, museums, libraries and much more.

Illustrated through the means of this business report is an approach of analysis for the various neighborhoods of the new york city and how one can approach towards buying of real-estate for the purposes of business, commercial, residential, etc.

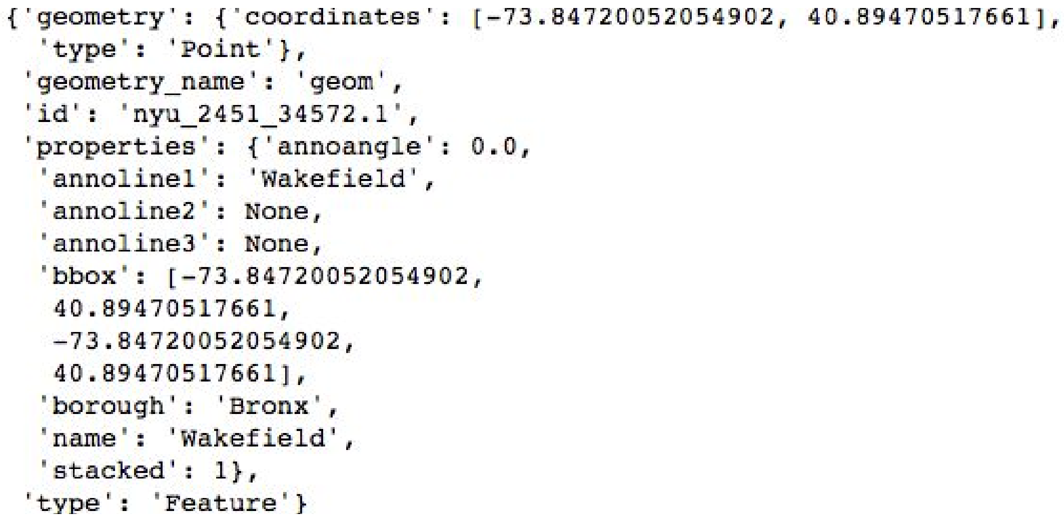
This kind of information would be very important to a variety of customer segments such as, people who are thinking of shifting to the new york city, business people who are looking for expansion of their office in the city, people who are trying to invest into the real-estate, real-estate selling websites or any franchisee owner looking to put out a new outlet. The aim of this kind of exploration of data is mainly to provide the user with a guide of the different measured features that can help in making an informed and quicker decision. So let's get on ahead with this exploration of NYC!

2.**Data Description**

With respect to the above data analysis problem I have taken the help of the following data sources :

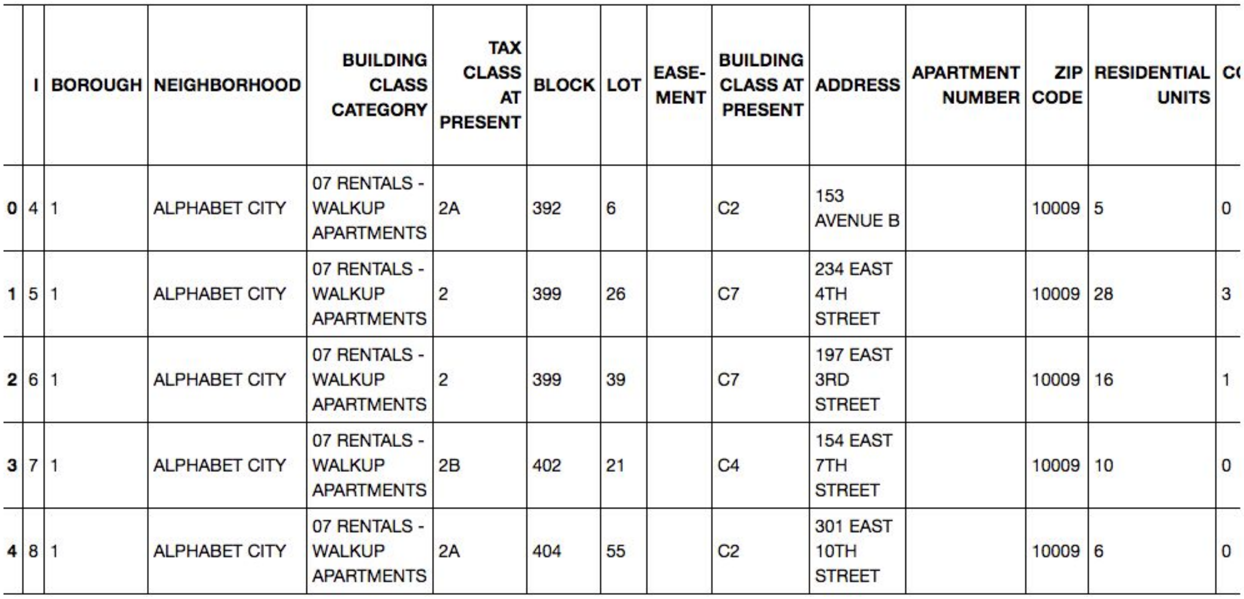
- **Foursquare API -** I have extracted data from the Foursquare API, which is one of the most trusted platforms for getting location data from. This entailed in making use of the Developer Tools provided by Foursquare after creating an app instance on their dashboard. Using the Foursquare API we will be able to access the latitude, the longitude, the postal codes of the neighborhoods, various venues that are nearby, their geographical coordinates, and much more.

- **NYU Spatial Data Repository -** The data of the various neighborhoods and the boroughs of the city of new york are present at the repository for free public use. I have downloaded the dataset from a link wherein the same dataset was uploaded. The result of reading the dataset is this. As you can see, our primary focus is to get the names and coordinates of the neighborhoods and the boroughs to which they belong.



```
{'geometry': {'coordinates': [-73.84720052054902, 40.89470517661],
  'type': 'Point'},
 'geometry_name': 'geom',
 'id': 'nyu_2451_34572.1',
 'properties': {'annoangle': 0.0,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661],
  'borough': 'Bronx',
  'name': 'Wakefield',
  'stacked': 1},
 'type': 'Feature'}
```

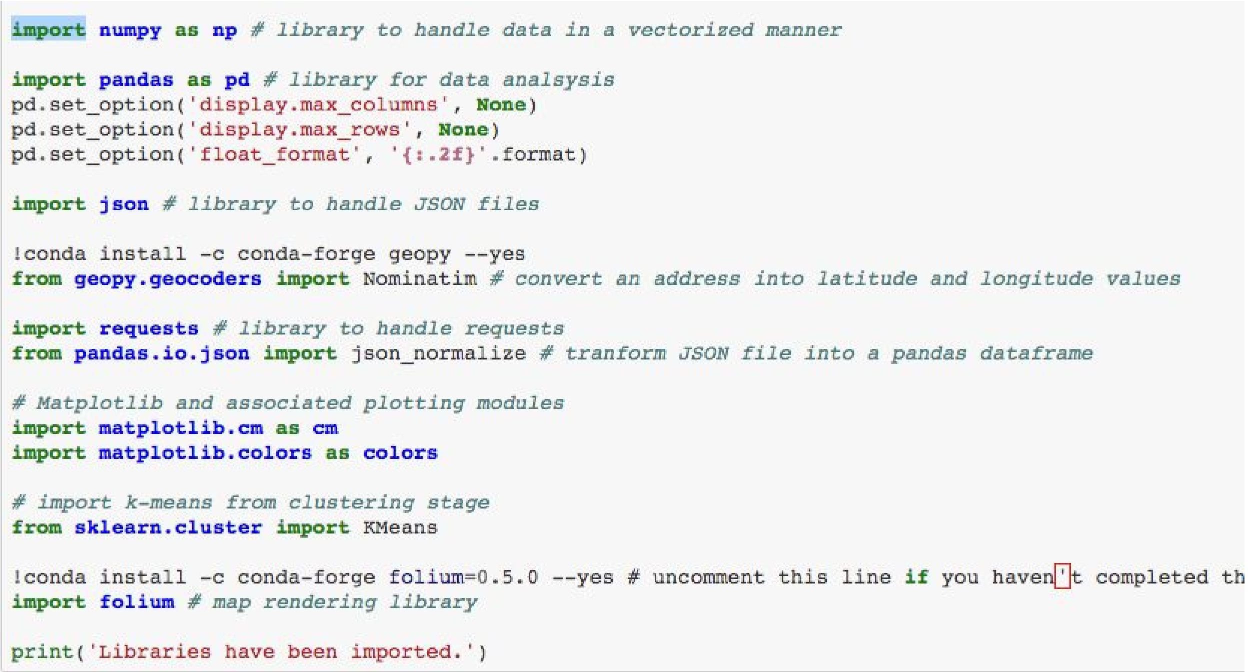
- **Kaggle dataset -** There are not too many public datasets available for demographics of the city of new york. Hence, in an effort to provide more accessible analysis of the neighborhoods, I collected The Sales Price of Housing for each neighborhood of New York City from a Kaggle dataset. We extract specifics in this dataset, namely the neighborhood, borough, total units sold, sales price for these transactions.



	I	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASE-MENT	BUILDING CLASS AT PRESENT	ADDRESS	APARTMENT NUMBER	ZIP CODE	RESIDENTIAL UNITS	CO
0	4	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	392	6		C2	153 AVENUE B		10009	5	0
1	5	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	26		C7	234 EAST 4TH STREET		10009	28	3
2	6	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	39		C7	197 EAST 3RD STREET		10009	16	1
3	7	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2B	402	21		C4	154 EAST 7TH STREET		10009	10	0
4	8	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	404	55		C2	301 EAST 10TH STREET		10009	6	0

Along with the datasets, there were many libraries that were required for this analysis, as it is the first step in quicker and correct code compilation for the analysis such as:

1. Requests
2. Folium
3. Pandas
4. Numpy
5. Geopy.geocoders
6. Json
7. Matplotlib
8.Sklearn
9. Wordcloud

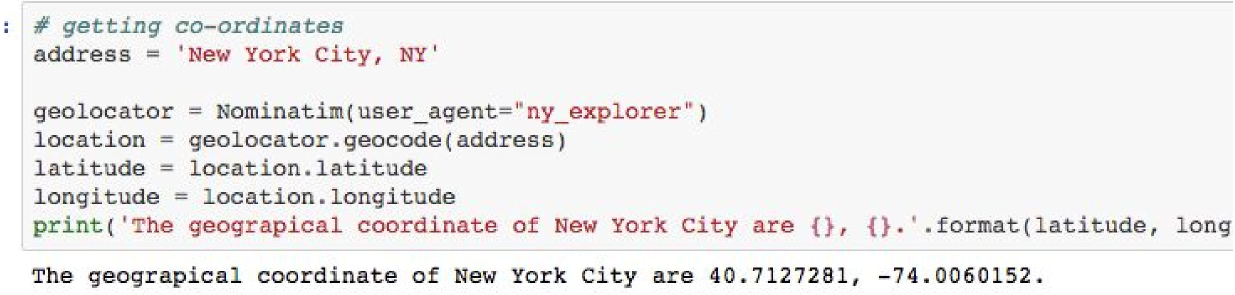


```
import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
pd.set_option('float_format', '{:.2f}'.format)

import json # library to handle JSON files

!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors

# import k-means from clustering stage
from sklearn.cluster import KMeans

!conda install -c conda-forge folium=0.5.0 --yes # uncomment this line if you haven't completed th
import folium # map rendering library

print('Libraries have been imported.')
```

3.Methodology

So the focus was to create a centralized dataframe which would be able to tell us about all the details required to make an informed decision for the problem statement mentioned. The first step was to get the data of the neighborhoods and the boroughs of the city of new york, and visualize the many places (neighbourhoods) offered by the dream city.

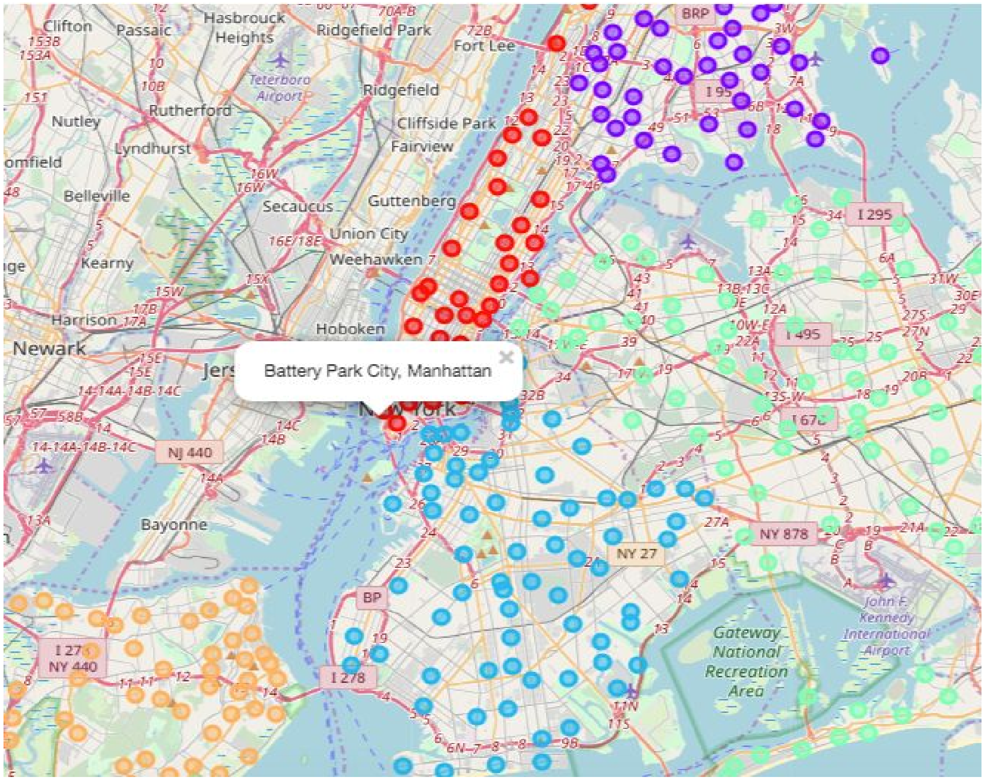
So I first got the geographical coordinates of NYC through geopy library.



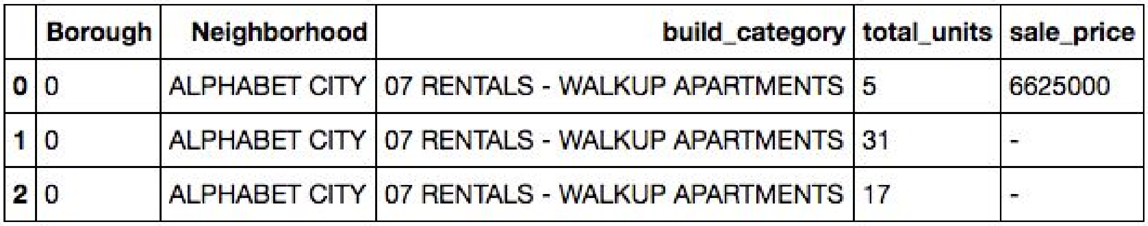
```
: # getting co-ordinates
  address = 'New York City, NY'

  geolocator = Nominatim(user_agent="ny_explorer")
  location = geolocator.geocode(address)
  latitude = location.latitude
  longitude = location.longitude
  print('The geograpical coordinate of New York City are {}, {}.'.format(latitude, long
```

```
The geograpical coordinate of New York City are 40.7127281, -74.0060152.
```

I then used the folium library to display the New York neighborhoods by their boroughs on the map

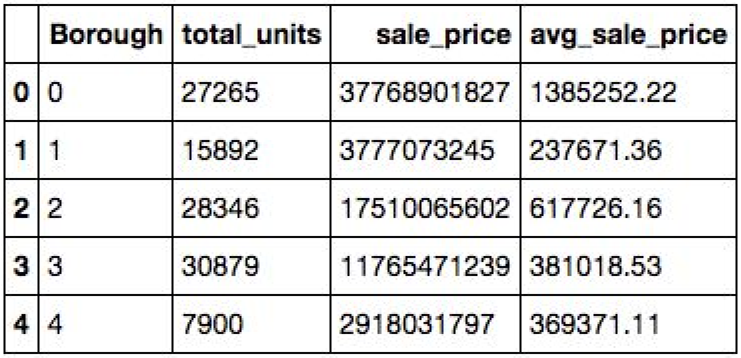


Having imported the data from the Kaggle dataset, I stored it in the dataframe for further cleaning and transformation of features, suitable to our needs for finding the average sales price of every neighborhood and each borough.



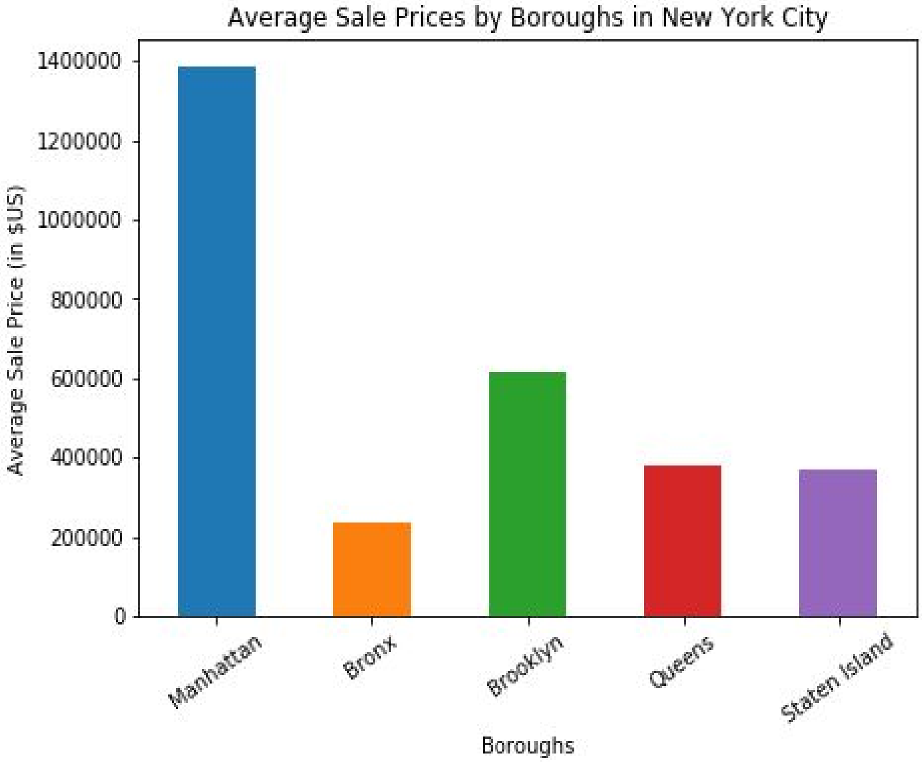
	Borough	Neighborhood	build_category	total_units	sale_price
0	0	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	5	6625000
1	0	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	31	-
2	0	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	17	-

Fig: Kaggle Dataset - original with flawed transactions



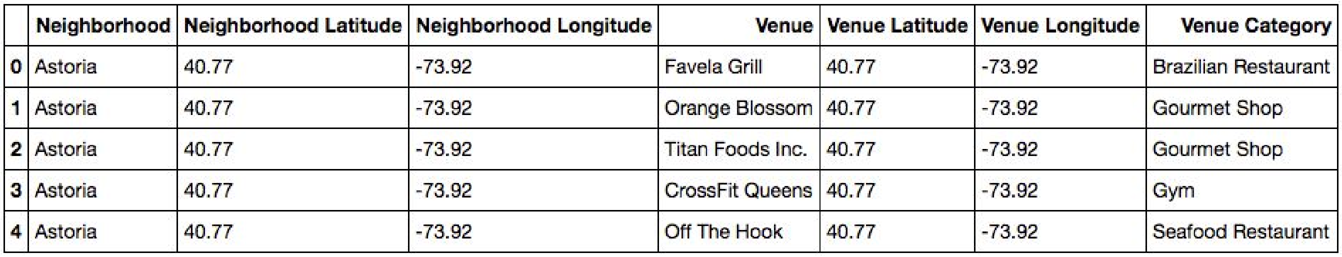
	Borough	total_units	sale_price	avg_sale_price
0	0	27265	37768901827	1385252.22
1	1	15892	3777073245	237671.36
2	2	28346	17510065602	617726.16
3	3	30879	11765471239	381018.53
4	4	7900	2918031797	369371.11

Fig: Dataframe nyc_sales_borough - after transformations



The above given table informs us about the average sales price of each borough which acts as a budget classifier for the customers. It is visualized as given here.

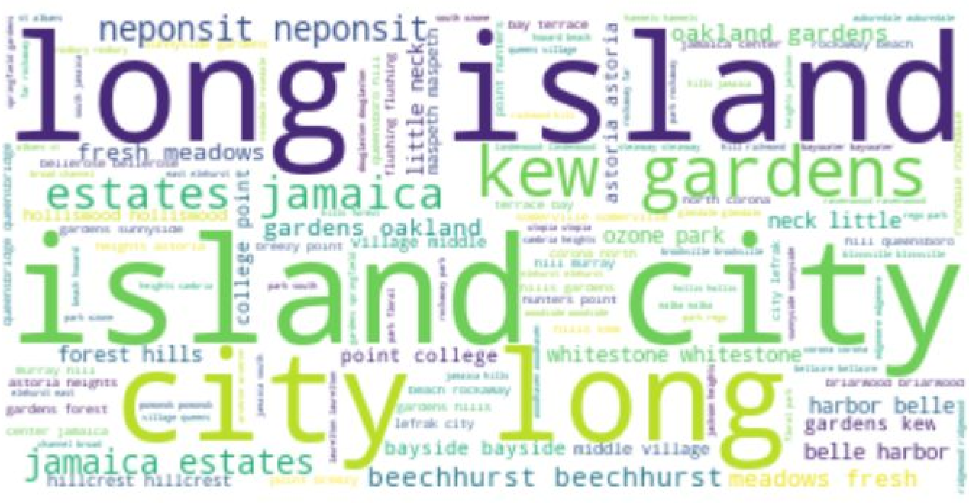
Having chosen a particular borough to explore further in, the customer would need the neighborhood information such as of nearby venues and attractions. So we extract the top venues from the data we collect for every neighborhood, using the Foursquare API.



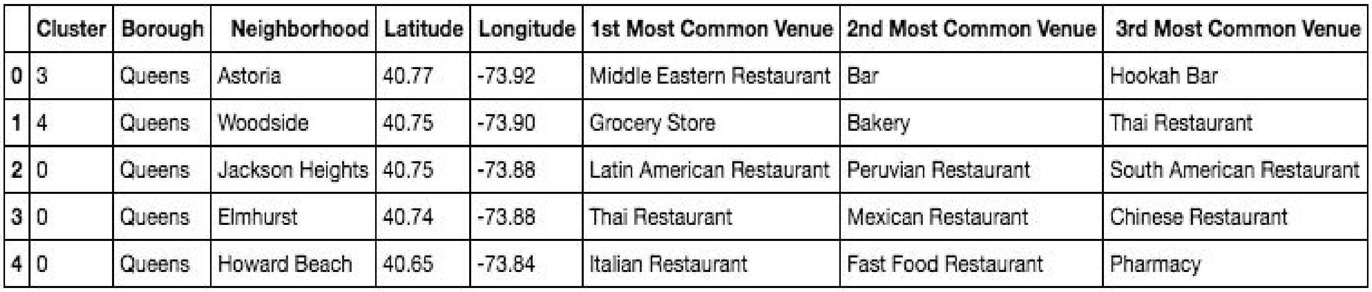
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Astoria	40.77	-73.92	Favela Grill	40.77	-73.92	Brazilian Restaurant
1	Astoria	40.77	-73.92	Orange Blossom	40.77	-73.92	Gourmet Shop
2	Astoria	40.77	-73.92	Titan Foods Inc.	40.77	-73.92	Gourmet Shop
3	Astoria	40.77	-73.92	CrossFit Queens	40.77	-73.92	Gym
4	Astoria	40.77	-73.92	Off The Hook	40.77	-73.92	Seafood Restaurant

We can choose any of the boroughs and perform similar exploration of the venues, so for now we have chosen to explore the Queens borough and its neighborhoods.

On further analysis of the data we collected from the Foursquare API, we come to know that there are more than 260 types of venues available in the vicinity of the Queens borough. A number of which were available in the neighborhoods of Long Island City, Astoria, Jamaica estates and so on. We plotted the neighborhoods according to their average sales price using wordcloud library. So the bigger the size of the word, the higher is the price of that neighborhood.

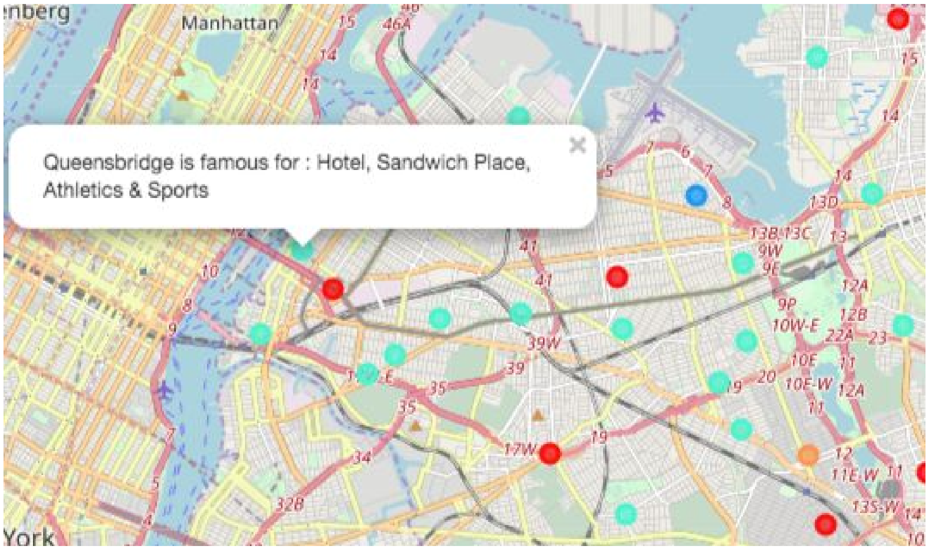


A K-means clustering model was used in order to group the different venues based on their venue categories, their location coordinates and the average sales price of the neighborhood they belonged to. Each venue was assigned a neighborhood based on the above mentioned features. With the value of k = 6, a k-means algorithm was run on the normalised set of input features. This is shown in the dataframe below.



	Cluster	Borough	Neighborhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	3	Queens	Astoria	40.77	-73.92	Middle Eastern Restaurant	Bar	Hookah Bar
1	4	Queens	Woodside	40.75	-73.90	Grocery Store	Bakery	Thai Restaurant
2	0	Queens	Jackson Heights	40.75	-73.88	Latin American Restaurant	Peruvian Restaurant	South American Restaurant
3	0	Queens	Elmhurst	40.74	-73.88	Thai Restaurant	Mexican Restaurant	Chinese Restaurant
4	0	Queens	Howard Beach	40.65	-73.84	Italian Restaurant	Fast Food Restaurant	Pharmacy

Following this, the different venues were plotted on a map with the help of the Folium library, each neighborhood displayed with a different coloured circular marker belonging to a single category. All the markers show the top 3 venues of the neighborhood as pictured in the data solution strategy.

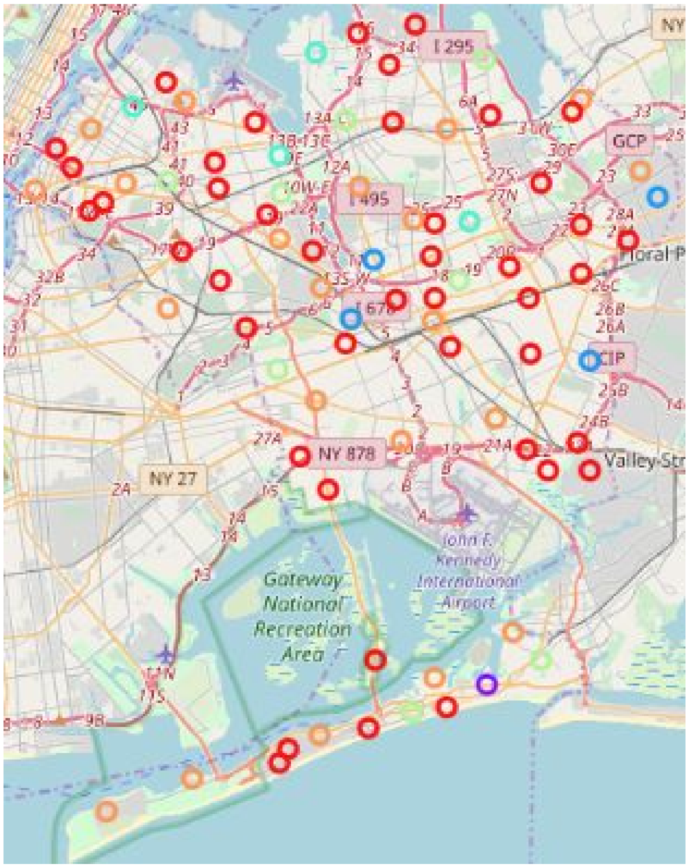


Queensbridge is famous for : Hotel, Sandwich Place, Athletics & Sports

4.**Results**

The side map is the one that I plotted after carrying out k-means clustering and segregation of the original dataset into six different clusters depending on the average sales price of the neighborhood, the geographical coordinates and the various types of venues that are found nearby the neighborhood.

It gives us the insight that a majority part of the clusters are dispersed all over the Queens borough and are not limited just by their geographical positioning. This points out, that the neighborhoods that might be far apart might still be more similar than the neighborhoods that are close to them.


I 295
I 495
I 678
NY 878
NY 27
GCP
CIP
NY
Floral P
Valley Str
Gateway National Recreation Area
John F. Kennedy International Airport

In case the user demands a venue detail he can simply click on it and get the top three most famous venue category nearby, which is where this map would definitely be handy.

5.**Discussion**

As noted throughout the analysis report, we concluded that the maximum number of neighborhoods in a cluster are not limited by their geographical coordinates. The clusters with characteristics that seem more like an outlier have small number of neighborhoods in them. There are more than 260 type of venues in the Queens borough, which seem to have very similar characteristics in the upper north and the lower west side of the Queens borough. Similar pattern of cluster distribution showcases low diversity in venue types and high geographical dissimilarity.

The machine learning model of k-means clustering was used in order to cluster the neighborhoods with respect to various types of venue categories, which of them are the most popular ones, and what is the average sales price of the neighborhood housing.

I ended the study by visualizing the data and clustering information on the neighborhoods of the Queens borough. In future studies, further analysis of various neighborhoods of other boroughs and dynamic inputs from the user can help take the analysis, a dynamic form on web based portals of the property dealers. We can get a detailed housing sales price for each neighborhood and provide a regression model which will be available for the customer, to predict in real-time, the neighborhood that will best suit his taste for establishing a business or building a home.

6.**Conclusion**

As a result of this analysis, our customer segment is, thus, made accessible to large amount of previously unstructured information into simple and easy to understand format of knowledge. The customer segment can now make an informed decision about the different neighborhoods and which one will fit his budget, his lifestyle, or increase the sales of an outlet. From this data, the user can also look up. Thus, through this analysis, I have successfully made full use of the knowledge I gained from completing the IBM Data Science Professional Certificate course.