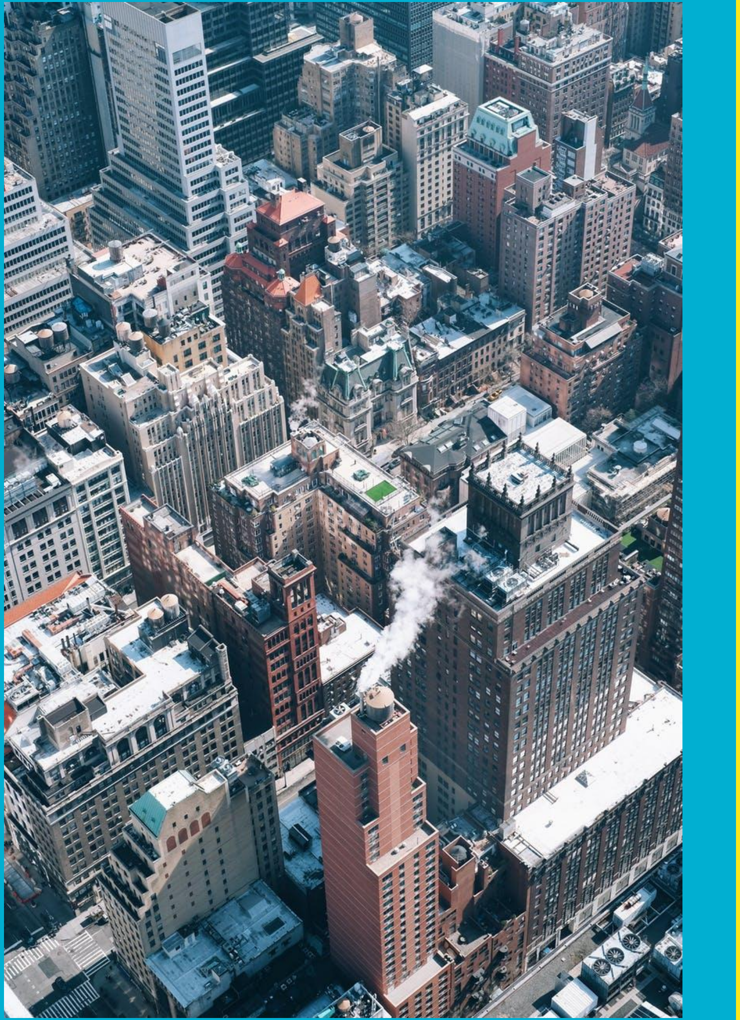
The Neighborhoods of NYC

-Abhishek Pandey



Introduction

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.

Problem / Analysis aim

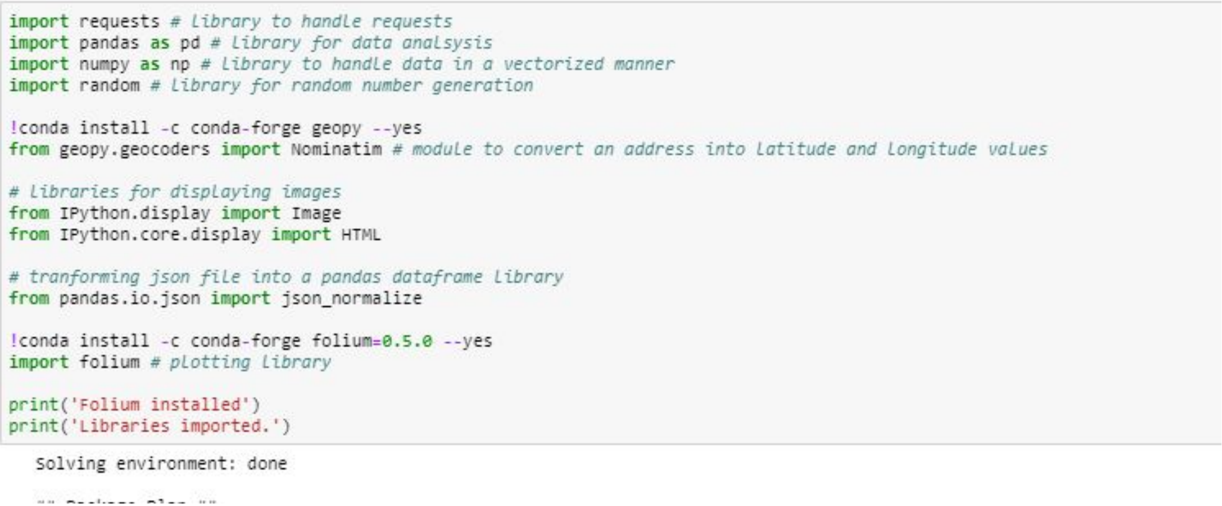
IIllustrated through the means of this business report is an approach of analysis for the various neighborhoods of the new york city and how one can approach towards buying of real-estate for the purposes of business, commercial, residential, etc.

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!

Data Description

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

- **Foursquare API**
- **NYU Spatial Data Repository**
- **Kaggle Dataset**



```
import requests # library to handle requests
import pandas as pd # library for data analsysis
import numpy as np # library to handle data in a vectorized manner
import random # library for random number generation

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

# libraries for displaying images
from IPython.display import Image
from IPython.core.display import HTML

# tranforming json file into a pandas dataframe library
from pandas.io.json import json_normalize

!conda install -c conda-forge folium=0.5.0 --yes
import folium # plotting library

print('Folium installed')
print('Libraries imported.')
```

```
Solving environment: done

## Package Plan ##
```

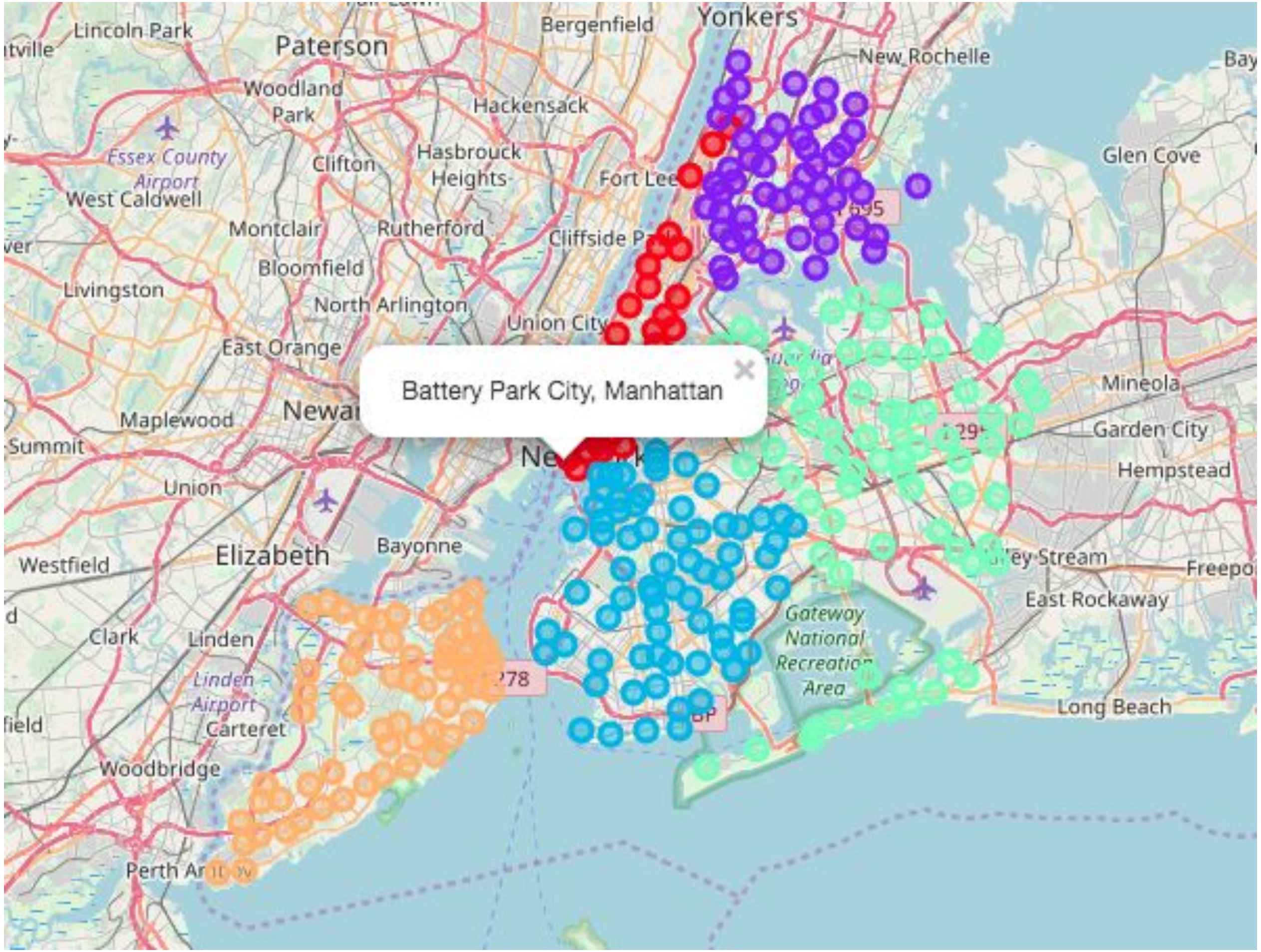
Libraries Used

The first step was to install the required libraries. Following are the different libraries that have come into play throughout the course of this analysis:

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

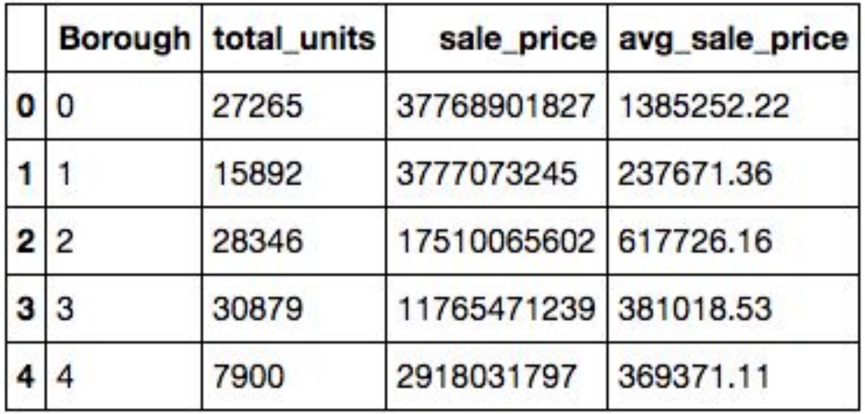
Methodology

- So I first got the geographical coordinates of NYC through geopy library.
- Got the neighborhoods and boroughs data from the NYU Data repository and merged them to create a final dataframe that is used along with Folium library to create the map.


Battery Park City, Manhattan
Yonkers
Bergenfield
Lincoln Park
Paterson
New Rochelle
Woodland Park
Hackensack
Essex County Airport
Clifton
Hasbrouck Heights
Fort Lee
Glen Cove
West Caldwell
Montclair
Rutherford
Cliffside Park
Bloomfield
Livingston
North Arlington
Union City
East Orange
Mineola
Garden City
Hempstead
Maplewood
Summit
Union
Elizabeth
Bayonne
Westfield
Clark
Linden
Linden Airport
Carteret
Woodbridge
Perth Amboy
Gateway National Recreation Area
East Rockaway
Long Beach
278
Freeport

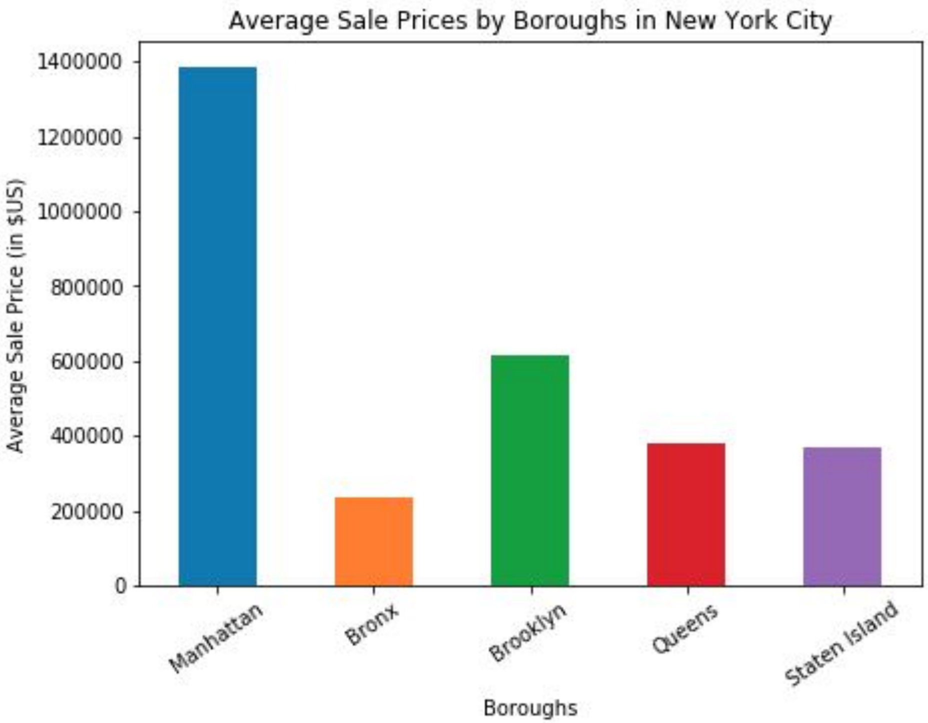
Merging of Sales Price

- I collected and transformed the housing sales price data from the kaggle datast to be shown by boroughs & individual neighborhoods
- The average sales price of the housing by boroughs



	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

contd.


Average Sale Prices by Boroughs in New York City
Average Sale Price (in $US)
1400000
1200000
1000000
800000
600000
400000
200000
0
Manhattan
Bronx
Brooklyn
Queens
Staten Island
Boroughs

Clustering the neighborhoods

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.

K-Means Clustering

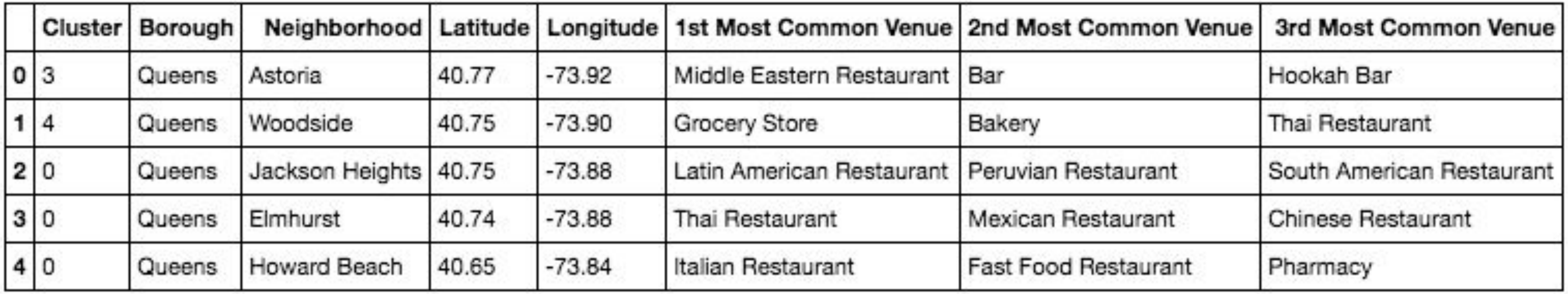
Clustered the neighborhoods of the Queens borough with respect to the following features:

- Average Sales Price of neighborhood
- Geographical Coordinates
- Venue categories nearby

Final Dataset

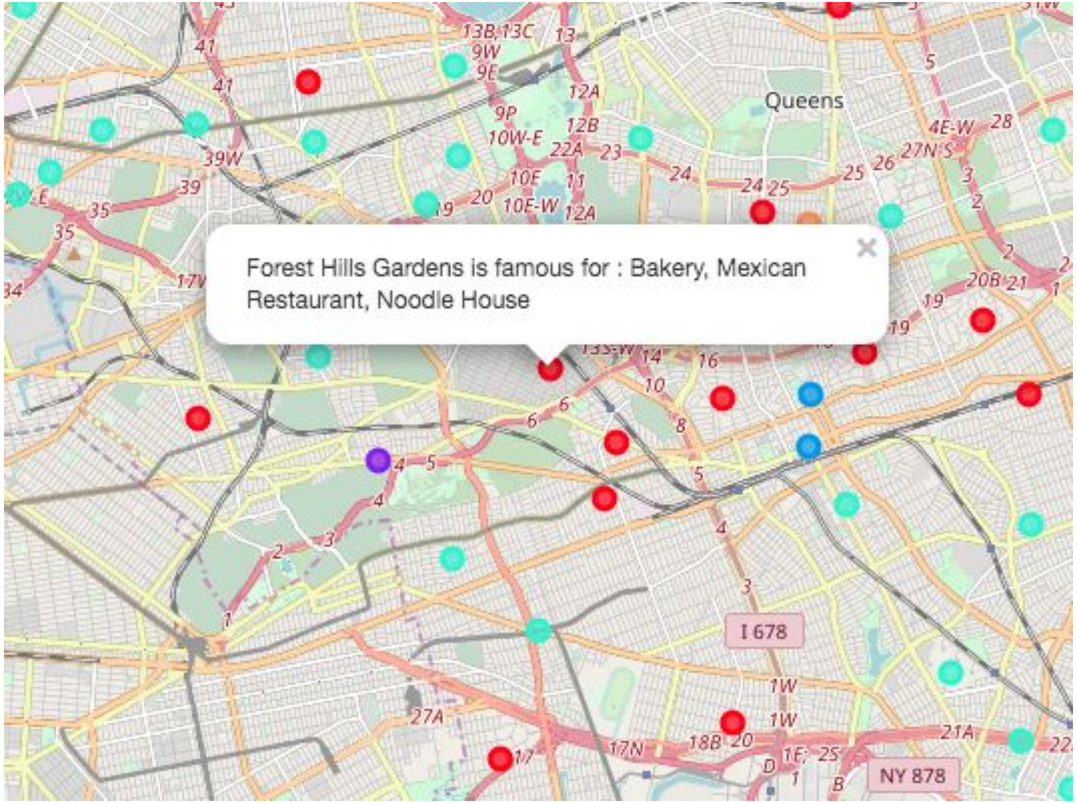
- This dataset provides the complete data of the various neighborhoods of Queens and their most common venues.

- Helps in making informed and faster decision.



	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

Visualization

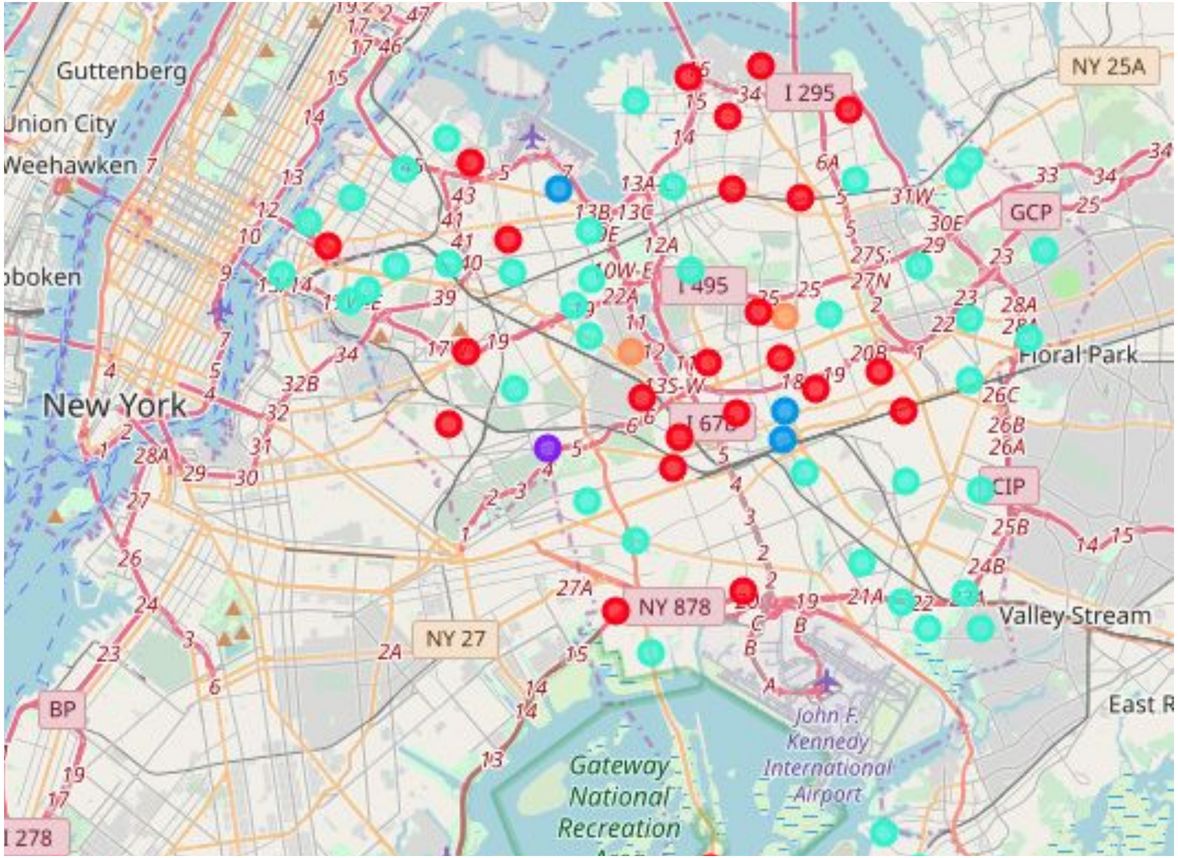


Forest Hills Gardens is famous for : Bakery, Mexican Restaurant, Noodle House

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.



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.

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.