Applied Data Science Capstone

The Battle of Neighborhoods

1. Introduction(Business Problem & Target Audience)

1) Business Problem

Queens is the easternmost of the five boroughs of New York City as well as having a large population of Koreans. According to the 2010 United States Census, the Korean population of Queens was 64,107, representing the largest municipality in the United States with a density of at least 500 Korean Americans per square mile. Due to the high percentage of Koreans in this area as well as the increasing popularity of Korean food, Queens, NY is an ideal location to open a Korean restaurant.

However, there are already so many Korean restaurants operating in this area and the market is highly competitive. As it is a highly developed city, the cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analysed carefully.

In accordance with this, the idea of this study is to help Koreans who are planning to open new Korean restaurants in Queens, NY to choose the right location by providing relevant data.

2) Target Audience

The target audience will be Koreans who are planning on opening a restaurant in Queens, so I will only focus on that borough during my analysis. The objective is to locate and recommend to the management which neighborhood of Queens will be the best in which to open a restaurant. The management should also be able to understand the rationale of the recommendations made.

2. Data

1) Data 1. Link to the dataset is :

https://geo.nyu.edu/catalog/nyu_2451_34572

New York city has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, I will need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighborhood. The link for this dataset is given above.

2) Data 2. Foursquare API_Korean Restaurant category ID: 4bf58dd8d48988d113941735

New York city geographical coordinates will be utilized as input for the Foursquare API, that will be leveraged to obtain venue information for each neighborhood.We will use the Foursquare API to explore neighborhoods in New York City. In addition, Korean Restaurant category Id 4bf58dd8d48988d113941735 is used for retrieving data from Foursquare API.

3. Methodology

In this project, I will use the basic methodology as taught in Week 3 lab.

First, I will convert addresses into their equivalent latitude and longitude values. Then I will use the Foursquare API to explore neighborhoods in Queens, NY. After that, I will obtain data on the most common venue categories in each neighborhood, and then use this information to group the neighborhoods into clusters K-means clustering algorithm will be used to complete this task. And also, I will use the Folium library to visualize the neighborhoods in Queens, NY.

```
neighborhoods.head()
```

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

```
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, longitude))
```

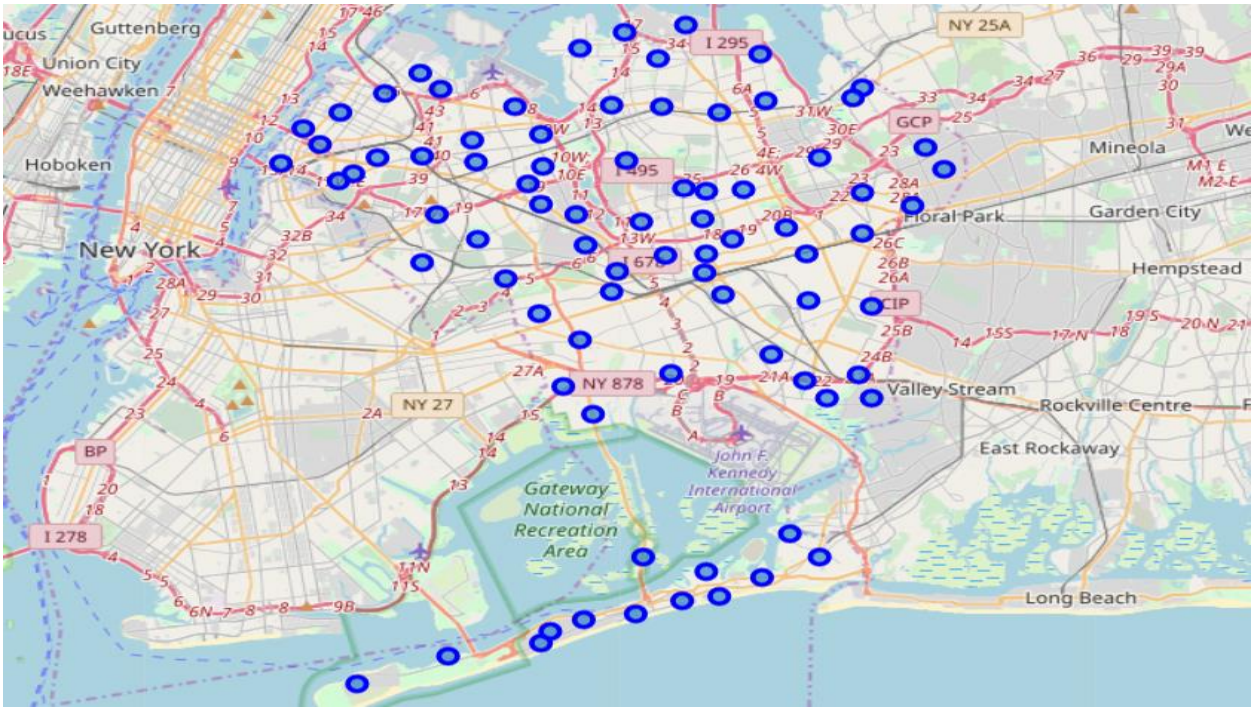
```
The geograpical coordinate of New York City are 40.7127281, -74.0060152.
```

In order to define an instance of the geocoder, we need to define a user_agent. We will name our agent ny_explorer, as shown below.

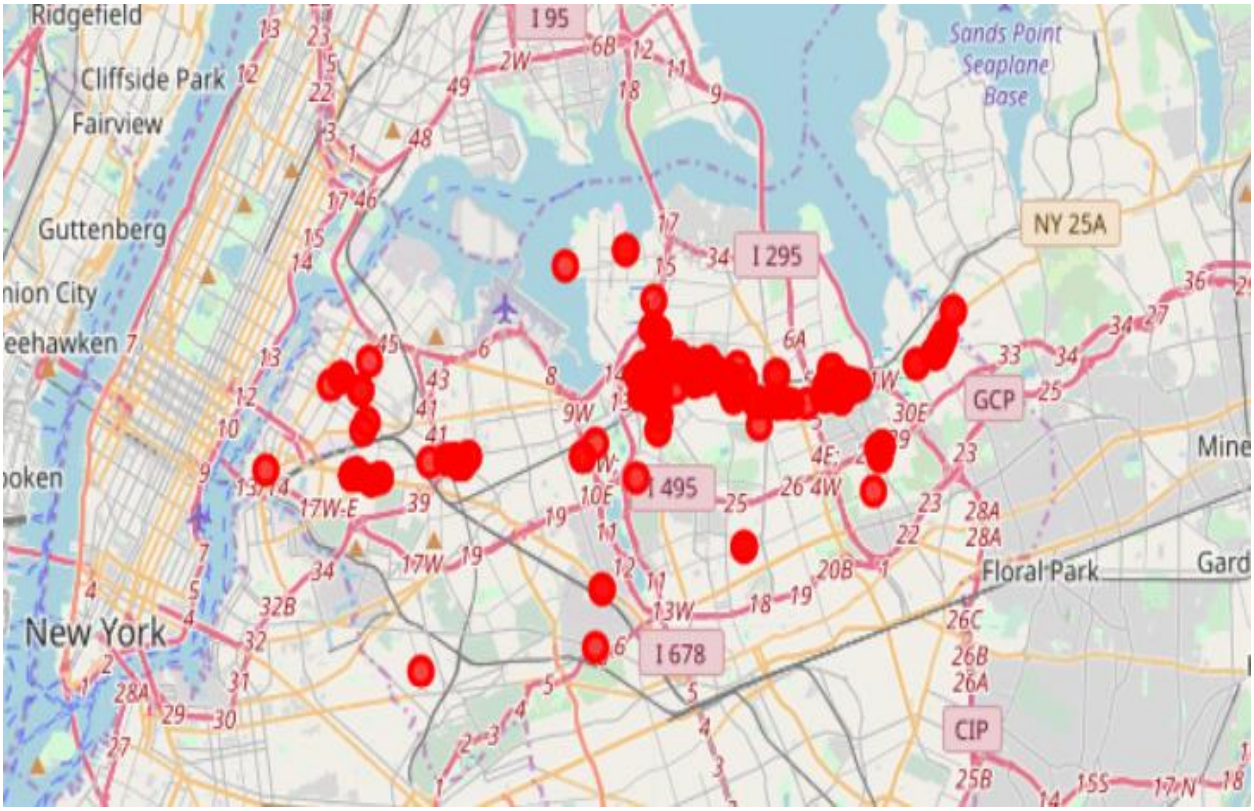
```
Queens_data = neighborhoods[neighborhoods['Borough'] == 'Queens'].reset_index(drop=True)
Queens_data.head()
```

	Borough	Neighborhood	Latitude	Longitude
0	Queens	Astoria	40.768509	-73.915654
1	Queens	Woodside	40.746349	-73.901842
2	Queens	Jackson Heights	40.751981	-73.882821
3	Queens	Elmhurst	40.744049	-73.881656
4	Queens	Howard Beach	40.654225	-73.838138

A map of Queens, NY with neighborhoods superimposed on top.



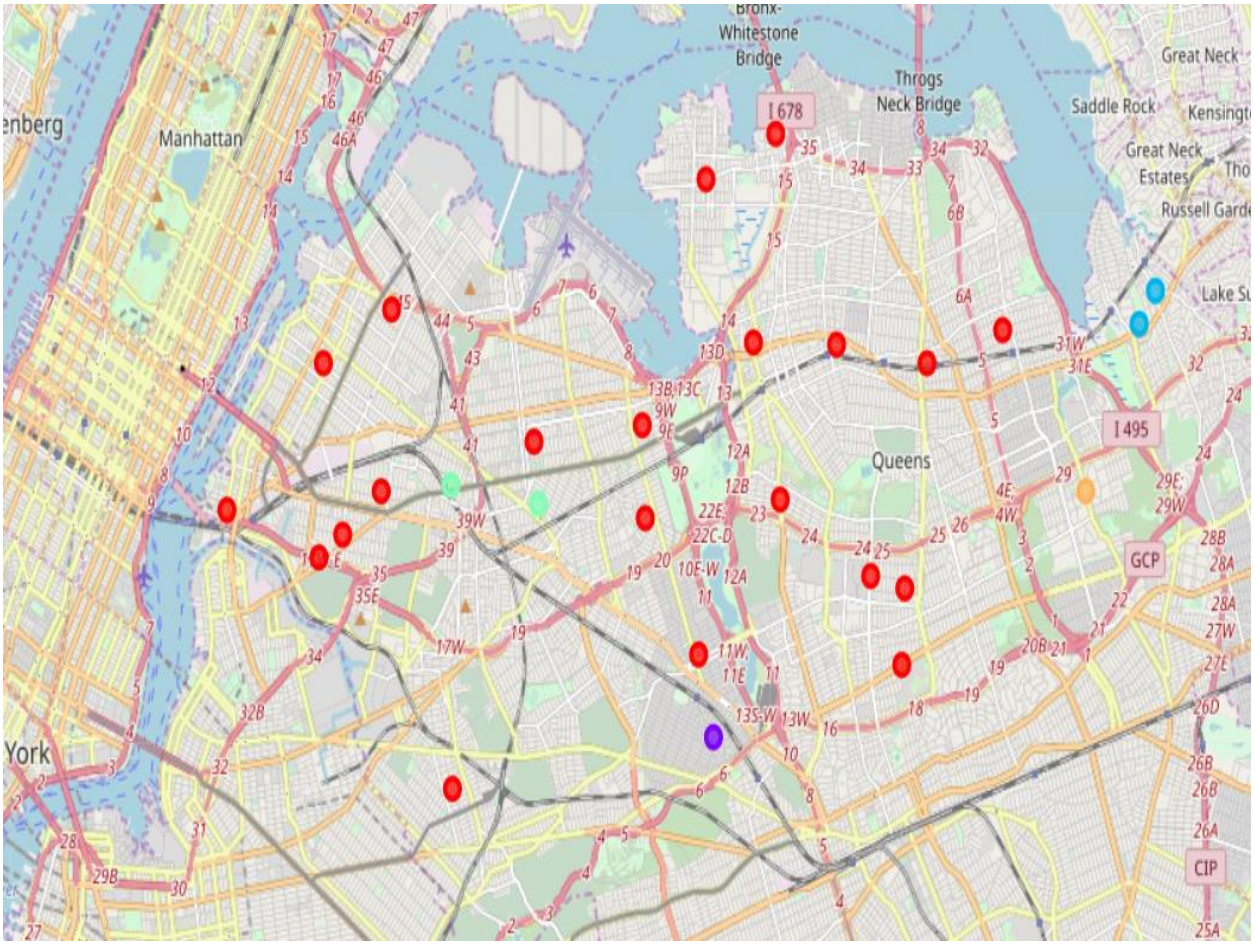
A map of Korean Restaurant in Queens, NY



Then use this feature to group the neighborhoods into clusters K-means clustering algorithm will be use to complete this task. And also, the Folium library to visualize the neighborhoods in Manhattan and its emerging clusters.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
0	Queens	Astoria	40.768509	-73.915654	0.0	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
1	Queens	Woodside	40.746349	-73.901842	3.0	Korean Restaurant	Fried Chicken Joint	Sushi Restaurant	New American Restaurant	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
2	Queens	Jackson Heights	40.751981	-73.882821	0.0	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
3	Queens	Elmhurst	40.744049	-73.881656	3.0	Korean Restaurant	Fried Chicken Joint	Sushi Restaurant	New American Restaurant	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
4	Queens	Howard Beach	40.654225	-73.838138	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Group the neighborhoods into clusters K-means clustering algorithm



4. Result

K-mean Cluster: Using K-mean to cluster data areas with less number of Korean restaurants

Based on dataframe analysis above Cluster 1 and Cluster 4 areas are the best places to open new Korean restaurants.

- Clust 0

	Neighborhood	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
0	Astoria	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
2	Jackson Heights	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
5	Corona	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
6	Forest Hills	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
9	Flushing	Korean Restaurant	Chinese Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Cajun / Creole Restaurant	Bakery
11	Sunnyside	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
14	Ridgewood	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery

- Clust 1

```
Queens_merged.loc[Queens_merged['Cluster Labels'] == 1, Queens_merged.columns[[1] + list(range(5, Queens_merged.shape[1]))]]
```

	Neighborhood	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
67	Forest Hills Gardens	New American Restaurant	Korean Restaurant	Sushi Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery

- Clust 2

- Clust 2

```
Queens_merged.loc[Queens_merged['Cluster Labels'] == 2, Queens_merged.columns[[1] + list(range(5, Queens_merged.shape[1]))]]
```

	Neighborhood	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
24	Little Neck	Korean Restaurant	Cajun / Creole Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Bakery
25	Douglaston	Korean Restaurant	Cajun / Creole Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Bakery

- Clust 3

- Clust 3

```
Queens_merged.loc[Queens_merged['Cluster Labels'] == 3, Queens_merged.columns[[1] + list(range(5, Queens_merged.shape[1]))]]
```

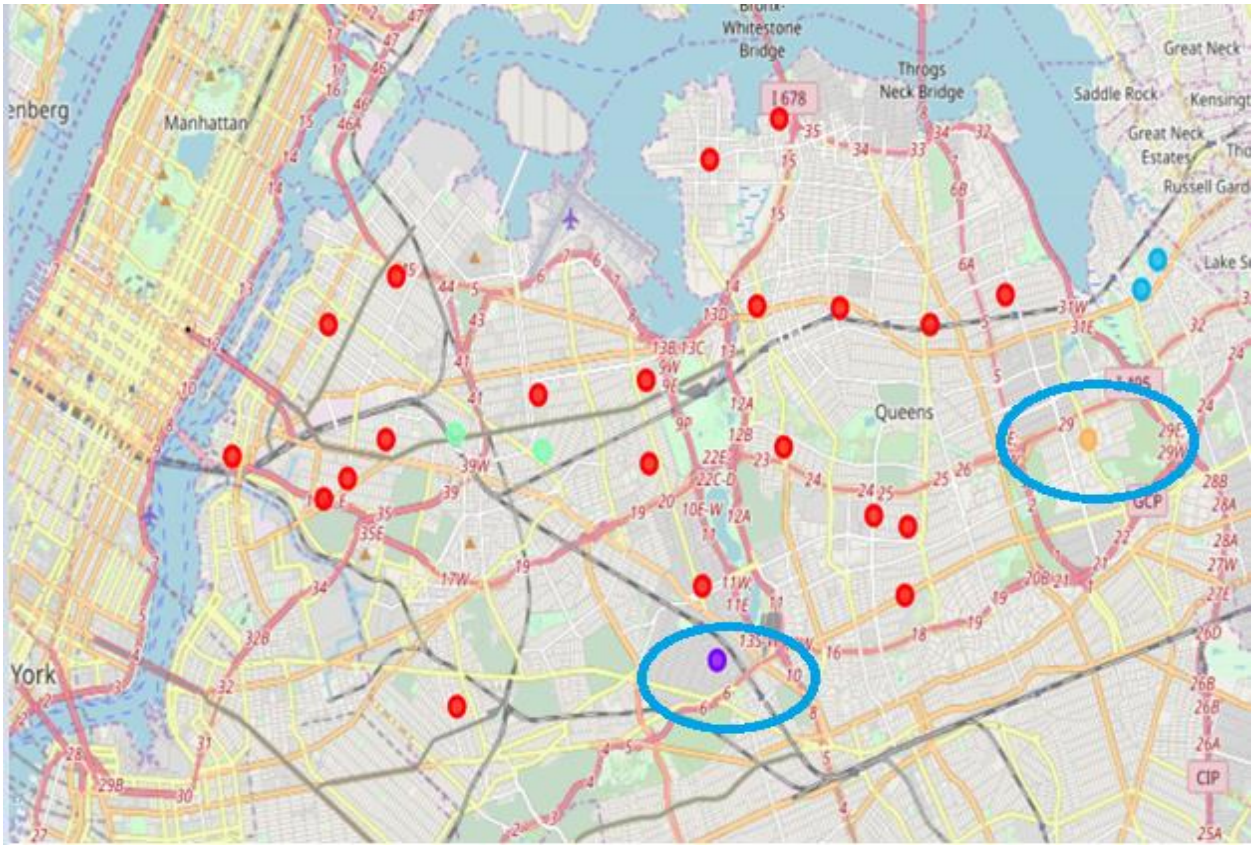
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
1	Woodside	Korean Restaurant	Fried Chicken Joint	Sushi Restaurant	New American Restaurant	Chinese Restaurant	Cajun / Creole Restaurant	Bakery
3	Elmhurst	Korean Restaurant	Fried Chicken Joint	Sushi Restaurant	New American Restaurant	Chinese Restaurant	Cajun / Creole Restaurant	Bakery

- Clust 4

- Clust 4

```
Queens_merged.loc[Queens_merged['Cluster Labels'] == 4, Queens_merged.columns[[1] + list(range(5, Queens_merged.shape[1]))]]
```

	Neighborhood	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
32	Oakland Gardens	Korean Restaurant	Sushi Restaurant	New American Restaurant	Fried Chicken Joint	Chinese Restaurant	Cajun / Creole Restaurant	Bakery



5. Discussion

In this section, I will be discussing the observations I have noted and the recommendation that I can make based on the results.

This analysis is performed on limited data. A sufficient amount of data increase the likelihood of achieving accurate results.

- There is high competition in Murray Hill, Flushing, Auburndale, and Bayside so it is very risky to open a business in these areas.
- There is low competition in Corona, Forest Hills, Hunters Point, Pomonok, Ridgewood, and Utopia so it is not risky to open a business in these areas.

- A more detailed analysis could be done by considering other factors such as transportation, demographics of inhabitants.

Finally, FourSquare proved to be a good source of data but frustrating at times. Despite having a Developer account I regularly exceeded my hourly limit locking me out for the day.

6. Conclusion

Although all of the goals of this project were met there is definitely room for further improvement and development as noted below. However, the goals of the project were met and, with some more work, could easily be devleoped into a fully phledged application that could support opening a business in an unknown location.

As per the neighborhood or restaurant type mentioned like Korean restaurants analysis can be checked. A venue with lowest risk and competition can be identified.