The Battle of Neighbourhoods: Coursera Capstone Project

Opening a new coffee shop in Queens, NY

Introduction of the Business problem

Intro

I will explore the best location for a new coffee shop in Queens, NY. Queens is an up-and-coming borough with an emerging coffee culture.

Target group

Entrepreneurs looking to invest to a coffee house. Data scientists looking for common applications of foursquare data.

Data section

For this project we need the following data:

1. New York City data that contains Borough, Neighborhoods along with there latitudes and longitudes Data Source: https://cocl.us/new_york_dataset Description: This data set contains the required information. And we will use this data set to explore various neighborhoods of new york city.
2. Coffee shops in Queens neighborhood of new york city. Data Source: Foursquare API Description: By using this API we will get all the venues in the Queens neighborhood. We can filter these venues to get only coffee shops.

Approach

Collect the new york city data from https://cocl.us/new_york_dataset.
Using Foursquare API we will get all venues for each neighborhood.
Filter out all venues which are coffee shops.
Data Visualization and some statistical analysis.
Analyzing using Clustering (especially K-Means): Find the best value of K and Visualize the neighborhood with a number of coffee shops.
Compare the Neighborhoods to Find the Best Place for Starting up a cafe.
Inference From these Results and related Conclusions

Data Preparation

We get the data of the New York boroughs and neighbourhoods together with their coordinates.

```
with open('newyork_data.json') as json_data:
    newyork_data = json.load(json_data)
neighborhoods_data = newyork_data['features']

# define the dataframe columns
column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']

# instantiate the dataframe
neighborhoods = pd.DataFrame(columns=column_names)
```

And after filtering for Queens, we get the data frame:

	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

Using Foursquare Location Data

For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the Venue, Venue category with their longitudes and latitudes.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Astoria	40.768509	-73.915654	Favela Grill	40.767348	-73.917897	Brazilian Restaurant
1	Astoria	40.768509	-73.915654	Orange Blossom	40.769856	-73.917012	Gourmet Shop
2	Astoria	40.768509	-73.915654	Titan Foods Inc.	40.769198	-73.919253	Gourmet Shop
3	Astoria	40.768509	-73.915654	Off The Hook	40.767200	-73.918104	Seafood Restaurant
4	Astoria	40.768509	-73.915654	CrossFit Queens	40.769404	-73.918977	Gym

Exploratory Data Analysis

There are 271 unique categories in which Coffee shop and Café are two of them. We will do one hot encoding for getting dummies of the venue category. Then we will calculate the mean of all venue groups by their neighbourhoods, which is interpreted as the relative frequency of this category.

I will add together Cafe and Coffee Shop columns because they are the same thing.

```
queens_grouped['Coffee place'] = queens_grouped['Café'] + queens_grouped['Coffee Shop']
```

```
queens_cafe = queens_grouped[['Neighborhood', 'Coffee place']]
queens_cafe.head()
```

	Neighborhood	Coffee place
0	Arverne	0.100000
1	Astoria	0.050000
2	Astoria Heights	0.000000
3	Auburndale	0.000000
4	Bay Terrace	0.026316

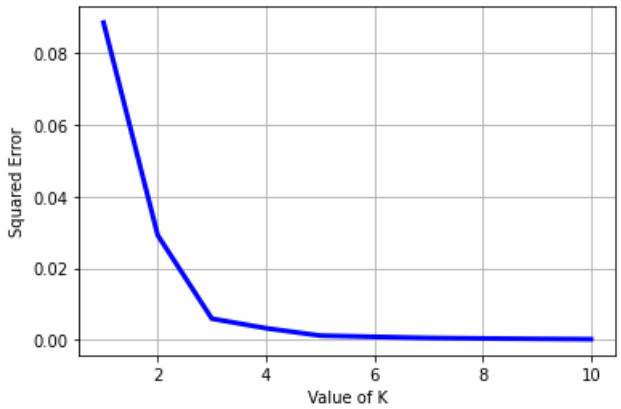
Clustering the Neighbourhoods

We will extract Coffee place data from the above table and fit this into the code for finding the best value of K.

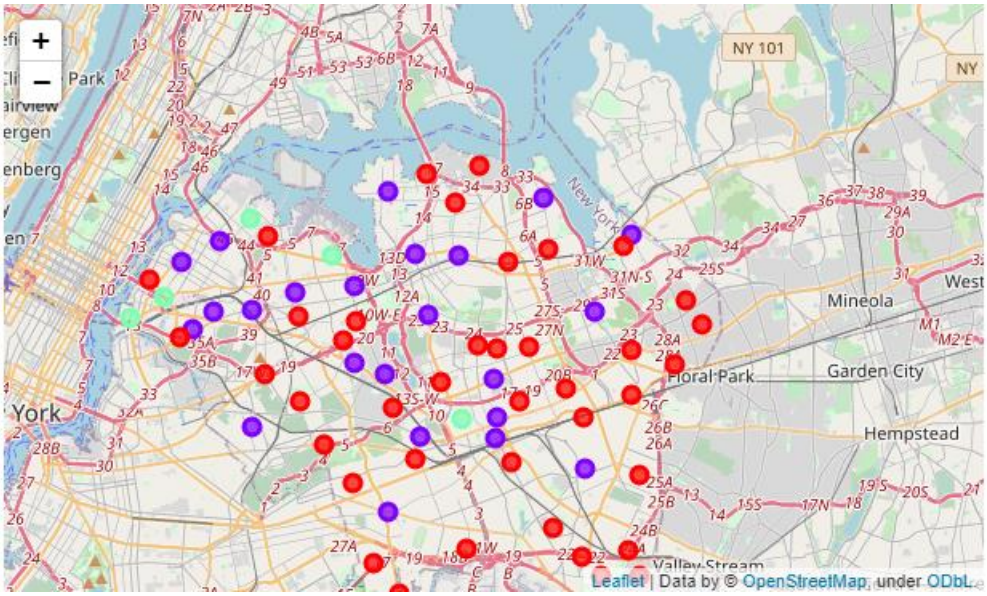
Select optimal number of clusters using the Elbow method

```
cost = []
for i in range(1,11):
    KM = KMeans(n_clusters = i, max_iter = 500)
    KM.fit(queens_cafe.drop('Neighborhood', 1))
    # calculate squared error
    cost.append(KM.inertia_)

# plot the cost against K values
plt.plot(range(1,11), cost, color = 'b', linewidth='3')
plt.xlabel('Value of K')
plt.ylabel('Squared Error')
plt.grid()
plt.show()
```



We split in 3 clusters and visualise using Folium:



The red cluster has the lowest frequency of coffee places, the purple has medium density and the light green the highest one.

Discussion and Conclusion

Cluster 0 has the lowest frequency of coffee places. I will pick Queensbridge because it is the one closest to Manhattan.