FINAL REPORT

Capstone Project - The Battle of Neighbourhoods

Introduction:

New York City's demographics show that it is a large and ethnically diverse metropolis. It is the largest city in the United States with a long history of international immigration. New York City was home to nearly 8.5 million people in 2014, accounting for over 40% of the population of New York State and a slightly lower percentage of the New York metropolitan area, home to approximately 23.6 million. Over the last decade the city has been growing faster than the region. The New York region continues to be by far the leading metropolitan gateway for legal immigrants admitted into the United States.

New York City has also been a major point of entry for immigrants; the term "melting pot" was coined to describe densely populated immigrant neighbourhoods on the Lower East Side. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. English remains the most widely spoken language, although there are areas in the outer boroughs in which up to 25% of people speak English as an alternate language, and/or have limited or no English language fluency. English is least spoken in neighbourhoods such as Flushing, Sunset Park, and Corona.

With its diverse culture, comes diverse food items. There are many restaurants in New York City, each belonging to different categories like Chinese, Indian, and French etc.

Problem:

To find the answers to the following questions:

The idea of this project is to categorically segment the neighborhoods of New York City into major clusters and examine their cuisines.

What cuisine is more prefers on the basis of region like Chinese,Indian etc?

This project will help to understand the diversity of a neighborhood by leveraging venue data from Foursquare’s ‘Places API’ and ‘k-means clustering’ unsupervised machine learning algorithm.

Data Section:

New York City's demographics show that it is a large and ethnically diverse metropolis. With its diverse culture, comes diverse food items. There are many restaurants in New York City, each belonging to different categories like Chinese, Indian, and French etc.

For this project we need the following data:

- New York City data that contains list Boroughs, Neighbourhoods along with their latitude and longitude.
  - 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 neighbourhoods of New York City

- GeoSpace data
  - Data source : https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm
  - Description: By using this geo space data we will get the New York Borough boundaries that will help us visualize choropleth map.

Methodology:

1. We begin by collecting the New York city data from the following link "https://cocl.us/new_york_dataset"
2. We will find all venues for each neighbourhood using Foursquare API

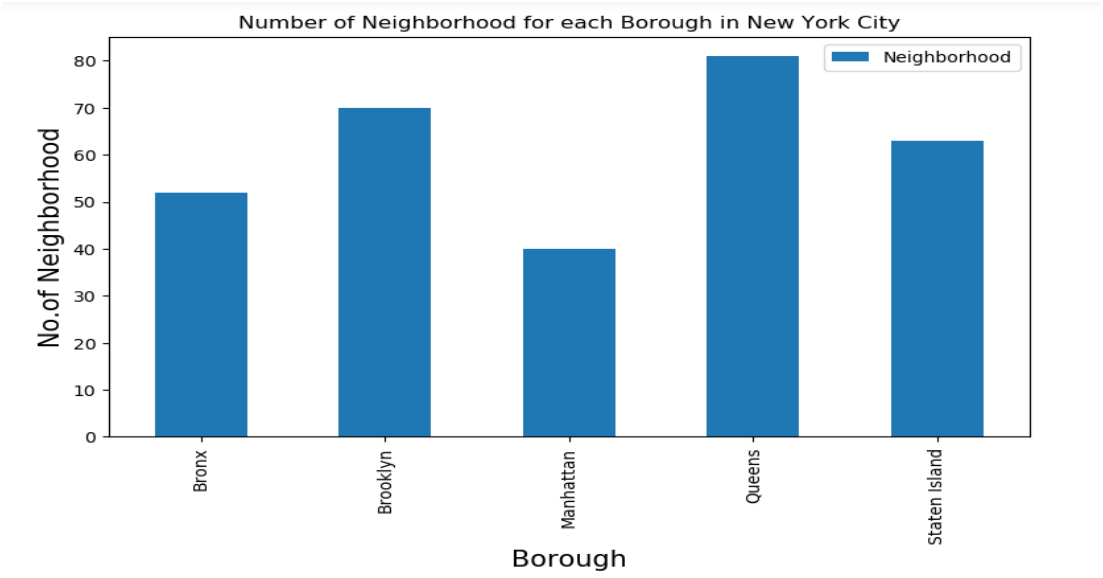
```
In [6]: new_york_data=get_new_york_data()

In [7]: new_york_data.head()
```

Out[7]:

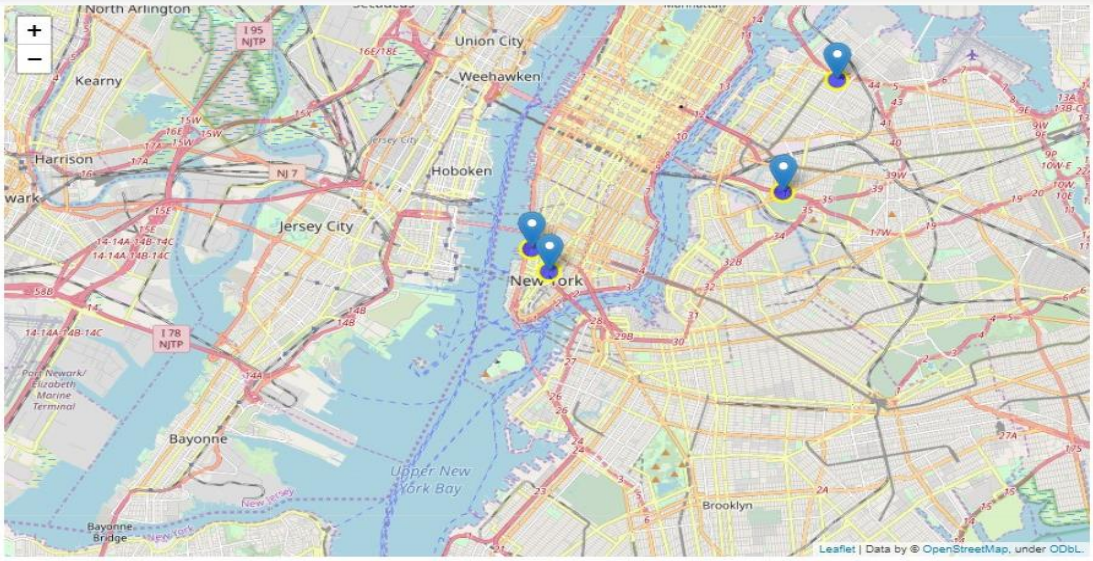
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

```
In [8]: new_york_data.shape
Out[8]: (306, 4)
```

The above result shows that there are 306 different Neighborhoods in New York.



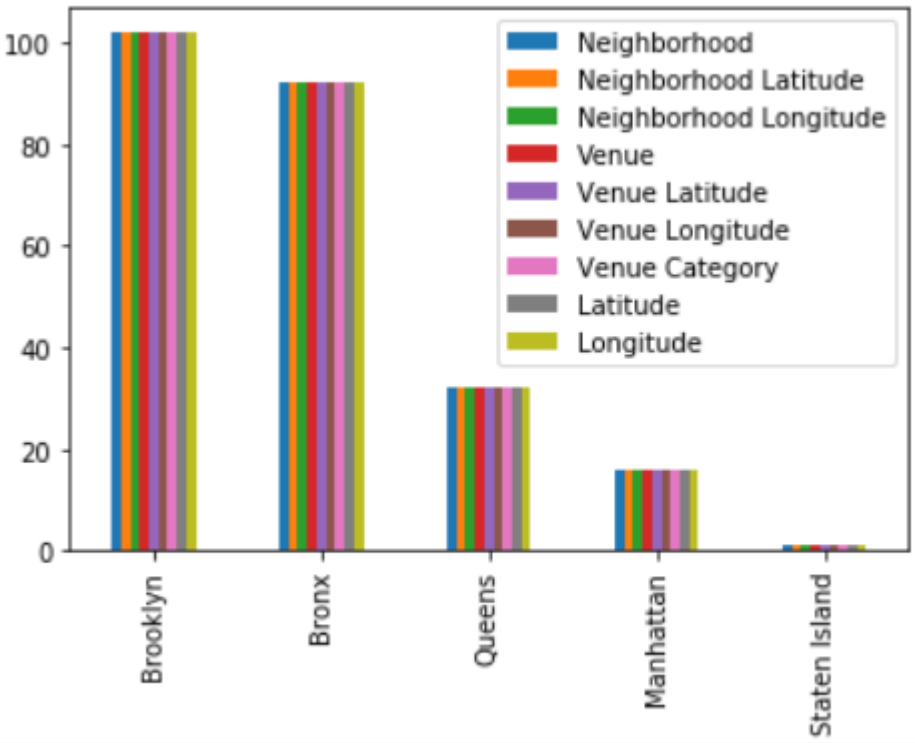
Neighbourhoods based on average rating:



Let's find out how many unique categories can be curated from all the returned venues

```
In [28]: print('There are {} uniques categories.'.format(len(nyc_venues['Venue Category'].unique())))
         nyc_venues.groupby('Venue Category')['Venue Category'].count().sort_values(ascending=False)

         There are 174 uniques categories.

Out[28]: Venue Category
         Coffee Shop                 581
         Pizza Place                 503
         Deli / Bodega               492
         Donut Shop                  333
         Bakery                      323
         Fast Food Restaurant        320
         Chinese Restaurant          242
         Italian Restaurant          222
         Café                        219
         Mexican Restaurant          182
         American Restaurant         176
         Fried Chicken Joint         165
         Bagel Shop                  163
```

Number of Chinese Resturants for each Borough in New York City



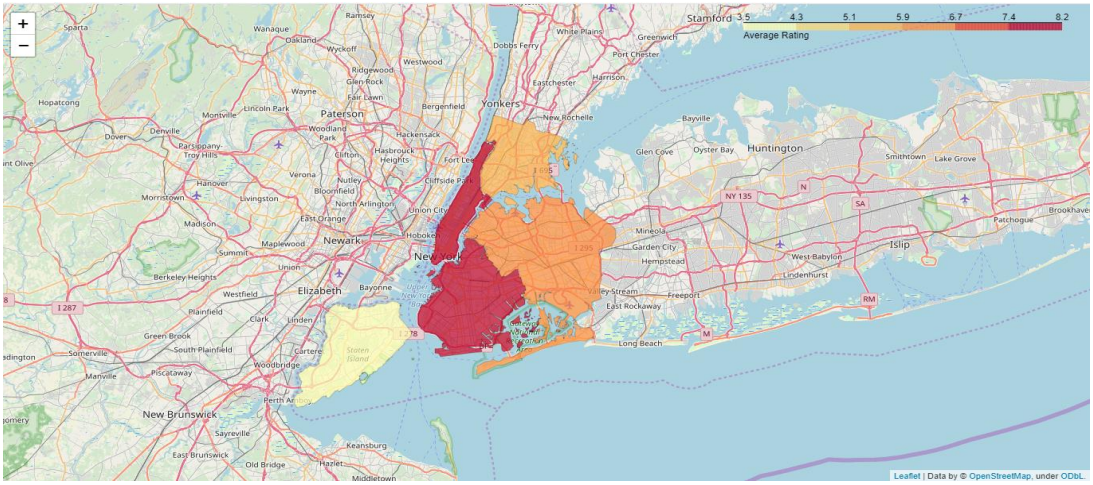
Let's count venues of each category in each neighbourhood

```
]: venue_counts = nyc_onehot.groupby('Neighborhood').sum()
venue_counts.head(5)
```

]:

Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	Brazil Restaur
Allerton	0	0	0	0	0	0	0	0	0	
Annadale	0	0	3	0	0	0	0	0	0	
Arden Heights	0	0	3	0	0	0	0	0	1	
Arlington	0	0	2	0	0	1	0	0	0	
Arrochar	0	0	0	0	0	0	0	0	0	

Borough based on average rating:



Most common venue

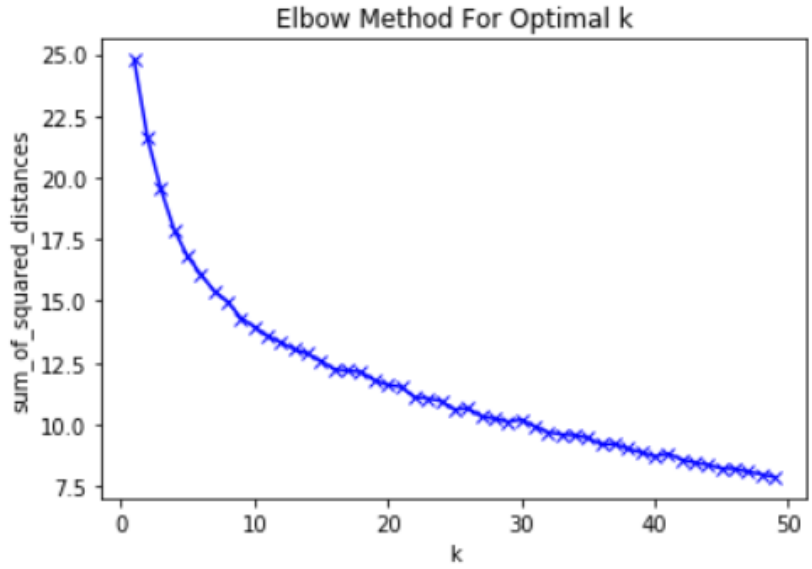
```
50]: for ind in np.arange(nyc_grouped.shape[0]):
         neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(nyc_grouped.iloc[ind, :], num

     neighborhoods_venues_sorted.head()
```

50]:

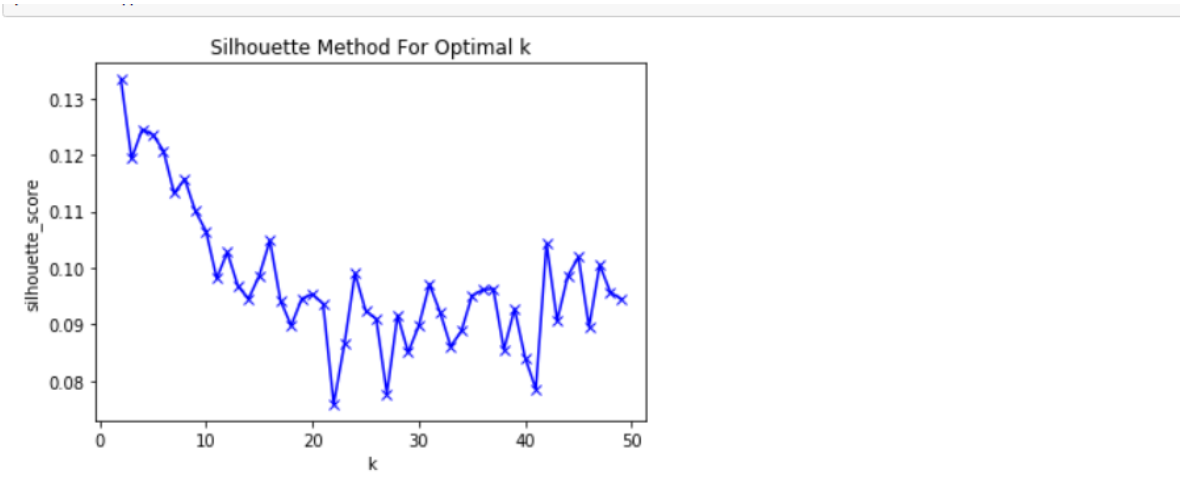
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Mexican Restaurant	Fried Chicken Joint	Pizza Place	Chinese Restaurant	Fast Food Restaurant
1	Annadale	Pizza Place	American Restaurant	Sushi Restaurant	Italian Restaurant	Japanese Restaurant
2	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Mexican Restaurant	Chinese Restaurant
3	Arlington	Pizza Place	Fast Food Restaurant	American Restaurant	Peruvian Restaurant	Spanish Restaurant
4	Arrochar	Italian Restaurant	Pizza Place	Steakhouse	Middle Eastern Restaurant	Chinese Restaurant

Clustering

```
plt.xlabel( K )
plt.ylabel('sum_of_squared_distances')
plt.title('Elbow Method For Optimal k');
```



Silhouette method for optimal K



There is a peak at k = 2, k = 4 and k = 8. Two and four clusters will give a very broad classification of the venues.

Cluster 0(implementation of K means)

Cluster 0

```
[62]: cluster_0 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 0, nyc_merged.columns[1:12]]
      cluster_0.head(5)
```

Out[62]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
36	Brookville	Fried Chicken Joint	Caribbean Restaurant	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Queens	40.660003	-73.751753
41	Cambria Heights	Caribbean Restaurant	Chinese Restaurant	Latin American Restaurant	Pizza Place	Fast Food Restaurant	Queens	40.692775	-73.735269
68	Crown Heights	Caribbean Restaurant	Fast Food Restaurant	Pizza Place	French Restaurant	Mexican Restaurant	Brooklyn	40.670829	-73.943291
77	East Flatbush	Caribbean Restaurant	Pizza Place	Fried Chicken Joint	Chinese Restaurant	Fast Food Restaurant	Brooklyn	40.641718	-73.936103
83	Eastchester	Caribbean Restaurant	Pizza Place	Fast Food Restaurant	Asian Restaurant	Chinese Restaurant	Bronx	40.887556	-73.827806

Result:

So now we can answer the questions asked above in the Questions section:

Manhattan is the best place to stay if you prefer Indian Cuisine.
Cluster wise which cuisine is preferred and is most popular

Cluster 0 — Caribbean

· Cluster 1 — Chinese

· Cluster 2 — Italian

· Cluster 3 — Italian American

· Cluster 4 — Pizza

· Cluster 5 — Mix of Cuisines

· Cluster 6 — Fast Food

· Cluster 7 — American

Conclusion:

Along with American cuisine, Italian and Chinese are very dominant in New York City and so is the diversity statistics**.**

This can be very useful for vendors who are willing to open shop and are confused to look in which area and which type of cuisines for there resturants.

There is always room for improvement and hence the above solution I have provided can also be improved for best results depending upon the data we have.