Capstone Project - The Battle of Neighborhoods

Introduction

London is one of the world's leading tourism destinations, and the city is home to an array of famous tourist attractions. Moreover, London is the capital city of England.
For this project, we will explore the surrounding of each metro stations and classify them. This project will be the benefit for the city planner and government to see the distribution of important venues. It will provide the visualization to show place where should be developed.

Problem

The distribution of important places does not decentralize in London. Consequently, it is crucial to see and plan the distribution of places to make every station accessing to all necessary venues within their surroundings.

Target audience

City planner and government of England is our target audience because they can see the venue distribution. To make sure that every neighborhood has the similar group and number of venues.

Data Description

In this project, city which is analyzed is London. The data acquired for this project is a combination of data from 2 sources.

Firstly, list of stations and their geographical coordinates is scraped from Wikipedia.
https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations
There are 53 metro stations in London. We will essentially need a dataset that the metro stations' name in London as well as the latitude and longitude coordinates of each station.

	Name	Latitude	Longitude
0	Acton Town	51.502500	-0.278126
1	Acton Central	51.50883531	-0.263033174
2	Acton Central	51.50856013	-0.262879534
3	Aldgate	51.51394	-0.07537
4	Aldgate East	51.51514	-0.07178

The category id will be derived from the Foursquare API. There are 9 venue categories.

```
Arts & Entertainment (4d4b7104d754a06370d81259)
College & University (4d4b7105d754a06372d81259)
Event (4d4b7105d754a06373d81259)
Food (4d4b7105d754a06374d81259)
Nightlife Spot (4d4b7105d754a06376d81259)
Outdoors & Recreation (4d4b7105d754a06377d81259)
Professional & Other Places (4d4b7105d754a06375d81259)
Residence (4e67e38e036454776db1fb3a)
Shop & Service (4d4b7105d754a06378d81259)
Travel & Transport (4d4b7105d754a06379d81259)
```

The metro st ta will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each station and to explore the metro stations' surroundings with the radius **1000 meter** for each station from their given latitude and longitude information.

```
def get_venues_count(lat,long, radius, categoryId):
    explore_url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&
                CLIENT_ID,
                CLIENT_SECRET,
                VERSION,
                lat,
                long,
                radius,
                categoryId)
    # make the GET request
    return requests.get(explore_url).json()['response']['totalResults']
```

The full CSV will merge the stations' name, geographical coordinates, and a number of each venue categories in each station. The CSV is the following figure below.

	Name	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Acton Town	51.502500	-0.278126	3	3	0	33	15	18	51	4	28	13
1	Acton Central	51.50883531	-0.263033174	4	4	0	26	12	17	0	3	31	9
2	Acton Central	51.50856013	-0.262879534	4	4	0	26	12	17	29	3	32	9
3	Aldgate	51.51394	-0.07537	36	61	2	185	145	107	74	20	69	94
4	Aldgate East	51.51514	-0.07178	38	62	1	183	144	103	67	22	81	94

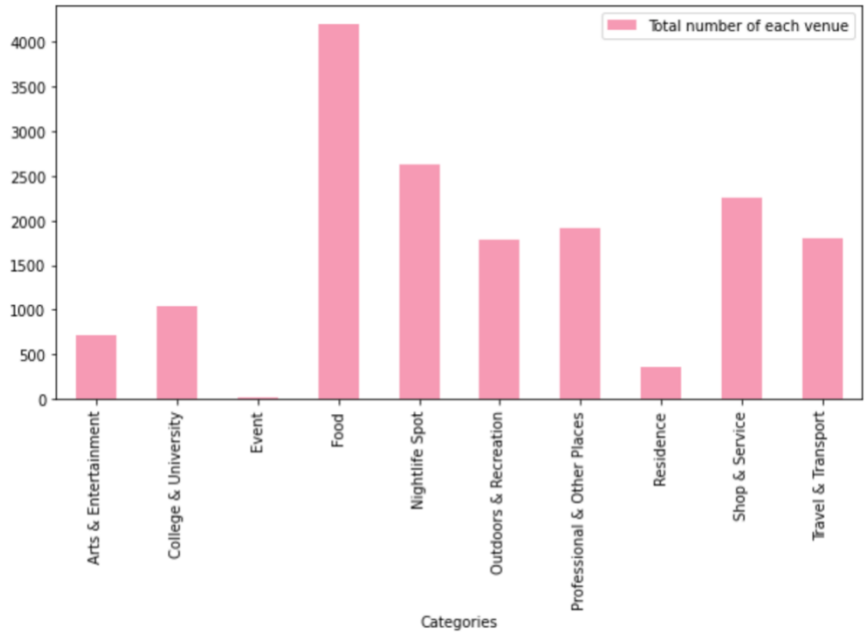
Methodology

I used pandas to make my master CSV containing 4 main components which are "Name", "Latitude", and "Longitude".

	Name	Latitude	Longitude
0	Acton Town	51.502500	-0.278126
1	Acton Central	51.50883531	-0.263033174
2	Acton Central	51.50856013	-0.262879534
3	Aldgate	51.51394	-0.07537
4	Aldgate East	51.51514	-0.07178

I utilized the Foursquare API to explore the boroughs and segment them. I designed the radius **10000 meter** for each station from their given latitude and longitude information.

```
#Request number of venues, store result as CSV
for i, row in stations_venues_df.iterrows():
    print(f"Now index is {i}")
    for c in categories_list:
        try:
            stations_venues_df.loc[i, c[0]] = get_venues_count(stations_venues_df.Latitude.iloc[i], stations_venues_
        except:
            pass
#         import pdb; pdb.set_trace()
    stations_venues_df.to_csv('stations_venues.csv')
```

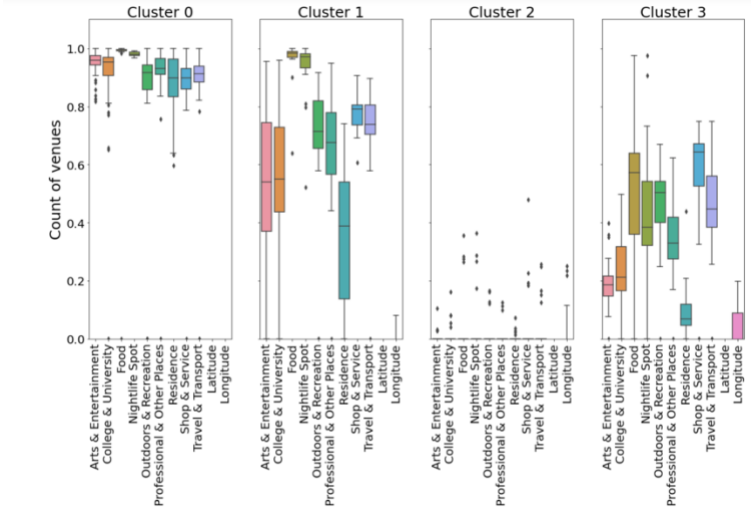


In explanatory analysis, I imported scikit-learn to use MinMaxScalar to normalize data (scale count of venues from 0 to 1 where 0 is the lowest value in a set and 1 is highest)

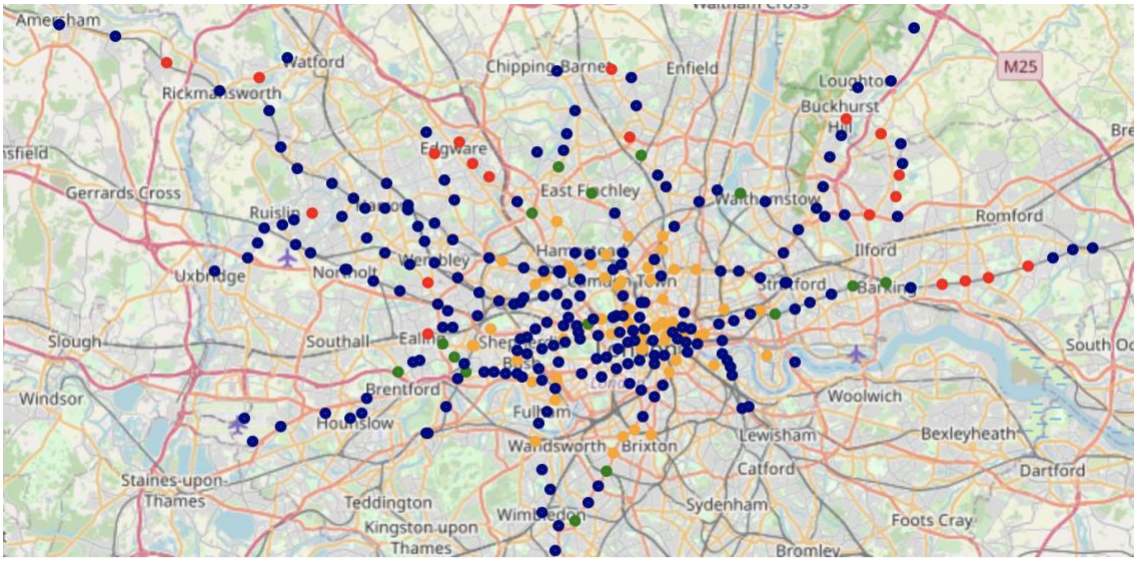
```
from sklearn.preprocessing import MinMaxScaler

X = new_df.values[:,3:]
cluster_dataset = MinMaxScaler().fit_transform(X)
```

I used KMeans to cluster data into 4 groups and used boxplot to visualize each cluster.

```
from sklearn.cluster import KMeans
#set number of clusters
kclusters = 4

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(cluster_df)

kmeans_labels = kmeans.labels_
# Change label numbers so they go from highest scores to lowest
replace_labels = {0:2, 1:0, 2:3, 3:1}
for i in range(len(kmeans_labels)):
    kmeans_labels[i] = replace_labels[kmeans_labels[i]]

stations_clusters_df = stations_venues_df.copy()
stations_clusters_df['Cluster'] = kmeans_labels
stations_clusters_minmax_df = cluster_df.copy()
stations_clusters_minmax_df['Cluster'] = kmeans_labels
stations_clusters_minmax_df['Name'] = new_df['Name']
stations_clusters_minmax_df['Latitude'] = new_df['Latitude']
stations_clusters_minmax_df['Longitude'] = new_df['Longitude']
```



Lastly, I plot them on map by using Folium. Each station will display the 3 largest number of venue categories with normalized scores.



Result

There are 4 clusters. Here is how we can characterize the clusters by looking at venue scores:

- Cluster 0 (Orange) : it is consistency high scores in all venue categories. This is mostly located in the middle of London.

- Cluster 1 (Green) : food and nightlife spot are largely situated in this cluster, however, there is lower mark and distribute surrounding the middle of the city.
- Cluster 2 (Navy) : it is the cluster which each station contains a few of the determined venue categories. There are lower scores in all categories and a large number of marks in London.
- Cluster 3 (Red) : it contains moderate scores of all categories and had in some stations in suburb.

Discussion

Nowadays, the city of London is the capital city of England and it is the one of the developed city of the world. Moreover, the London plan will develop in e-commerce, tourism, and environment in the future. As you can see, most parts of London are financial and business services. A state-of-the-art city should be combined and improved in technology to leverage the city.

According to KMeans, a large number of business services are located in the center of city, while the suburb contains differently some venues in each station. Nonetheless, data does not take into account a venue's size.

Conclusion

Although Foursqaure data is limited, it can provide some insight which is the distribution of venues between the center and the suburb in London with the up-to-date data.