IBM Coursera Applied Data Science Capstone Project

A BIRD'S VIEW OF LONDON

By:
Ankush Babbar

Introduction

Tourism has always been a booming section across the globe. No matter which country which you live in, you always come across a group of people, big or small, who always like to visit places. Tourism is not only an important aspect of a country’s economy but also for its global standing. It also generates more employment opportunities, revenues, and plays a significant role in development.

Also, for many people or tourists, shopping is one of the major things which people do. Visiting shopping malls is a great way to relax and enjoy themselves during the stay. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets. Watch movies and perform many more activities.

London is the capital and largest city of England and the United Kingdom. The city stands on the River Thames in the south-east of England, at the head of its 50-mile (80 km) estuary leading to the North Sea. London has been a major settlement for two millennia. Londinium was founded by the Romans. The City of London, London's ancient core and financial centre − an area of just 1.12 square miles (2.9 km2) and colloquially known as the Square Mile − retains boundaries that closely follow its medieval limits. The adjacent City of Westminster is an Inner London borough and has for centuries been the location of much of the national government. Thirty-one additional boroughs north and south of the river also comprise modern London. London is one of the world's most important global cities. It exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation. It is one of the largest financial centres.

Business Problem

The City is the financial capital of the country and is diverse in culture, as well as the city have many visitors every year. So, aim of the project is to analyse the interests of the people living in the city and determine the localities that are popular for some specific venue categories. Also, while many immigrants come and search for a house to live, we would look into what localities and neighbourhoods differ from other.

This city is very popular and have been experiencing increase in visitors overall over years. Those who are planning or interested in visiting the city might be interested what are good localities and local people interest overall. The interested audience in the project will be the visitors, tourists, and all the immigrants who are travelling to London to explore their area of interests.

Data

To Solve the problem, we need the following data:

1.) List of boroughs and neighbourhoods in London, this defines the scope of this project which is confined to the city of London, the capital city of the country of England.

2.) Latitude and Longitude co-ordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

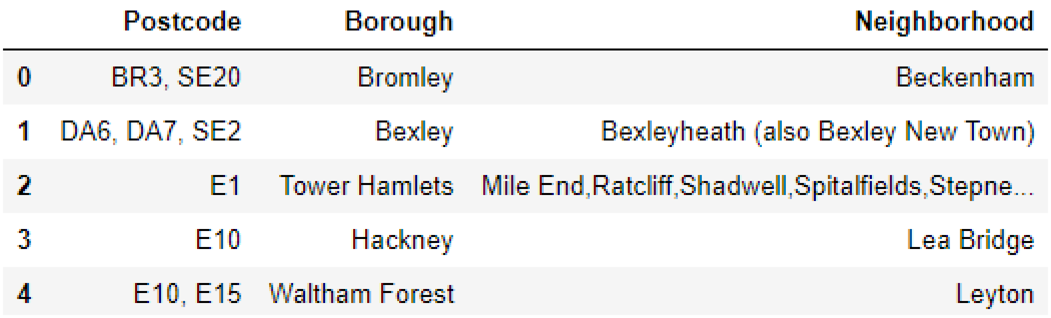
3.) Venue data, particularly related to trending venues, we will use this data to perform clustering on the neighbourhoods.

The Wikipedia page - https://en.wikipedia.org/wiki/List_of_areas_of_London is the major source of data that is being used to obtain all the Boroughs and neighbourhoods in London. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests. Then we will get the geographical co-ordinates of the neighbourhoods using Python geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,00 developers. Foursquare API will provide many categories of the venue data. This project that will make use of many data science skills, from web scraping (Wikipedia), working with API (foursquare), data cleaning, data wrangling, to machine learning (k-means clustering) and map visualization (Folium). In the next section, we will present the methodology section where will discuss the steps taken in the project, that data analysis that we did and the machine learning techniques that we used.

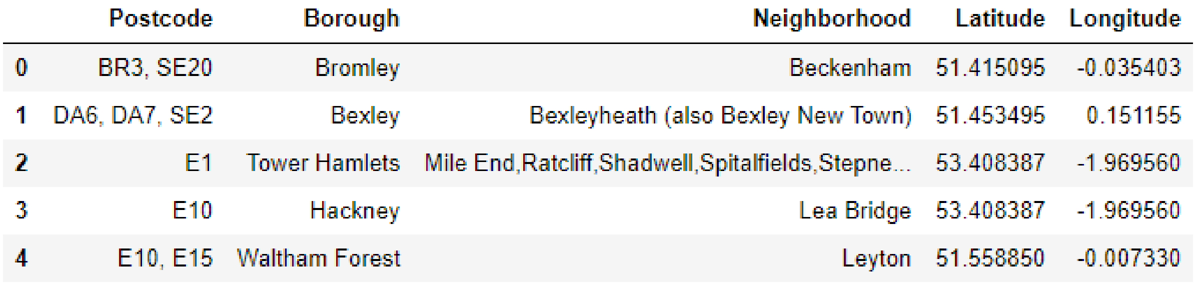
Methodology

The first step is to collect the data, this is done by scrapping the Wikipedia page, after cleansing the data we are left with 161 unique Postal Codes for London and we could see 307 of them , thus grouping them so as to the data with same Postal Codes clubs together . Then We have around 192 unique combinations of London Postcode and Borough. The Dataframe looks like the below mentioned table :



	Postcode	Borough	Neighborhood
0	BR3, SE20	Bromley	Beckenham
1	DA6, DA7, SE2	Bexley	Bexleyheath (also Bexley New Town)
2	E1	Tower Hamlets	Mile End,Ratcliff,Shadwell,Spitalfields,Stepne...
3	E10	Hackney	Lea Bridge
4	E10, E15	Waltham Forest	Leyton

After which, we use geopy API to get the latitude and longitude of all the areas of London. Missing values were removed from the dataset. The data of co-ordinates with their respective areas are joined in the main Dataframe. With all the cleaning and cleansing, we have the following table

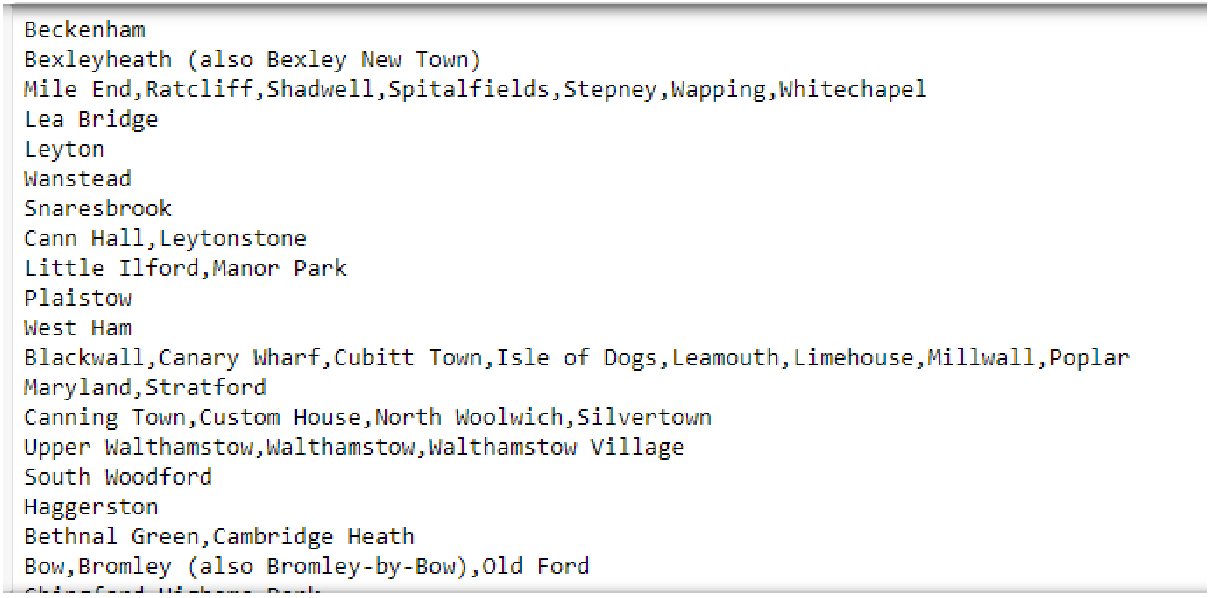


	Postcode	Borough	Neighborhood	Latitude	Longitude
0	BR3, SE20	Bromley	Beckenham	51.415095	-0.035403
1	DA6, DA7, SE2	Bexley	Bexleyheath (also Bexley New Town)	51.453495	0.151155
2	E1	Tower Hamlets	Mile End,Ratcliff,Shadwell,Spitalfields,Stepne...	53.408387	-1.969560
3	E10	Hackney	Lea Bridge	53.408387	-1.969560
4	E10, E15	Waltham Forest	Leyton	51.558850	-0.007330

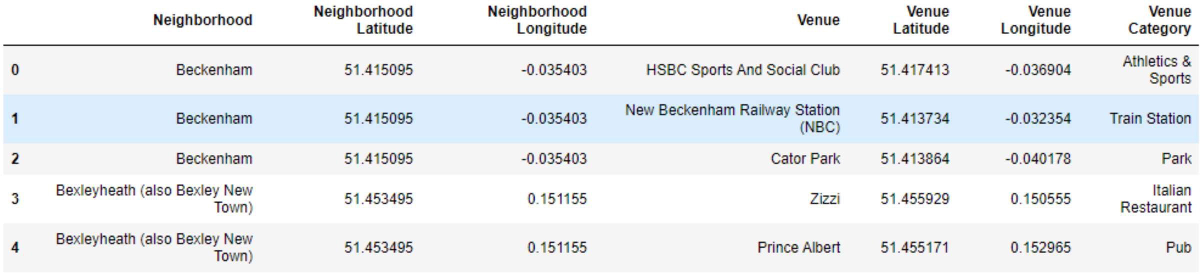
Since, we have got Co-ordinates of various Boroughs of London, we would plot the same, in the Map of London, which we have plot using the Folium Module.

Now, we have plotted the neighbourhoods of London, we would explore the neighbourhoods of London from Foursquare. We would explore the top 100 venues in this area with their categories within 500m radius.

Below are the top 100 venues :

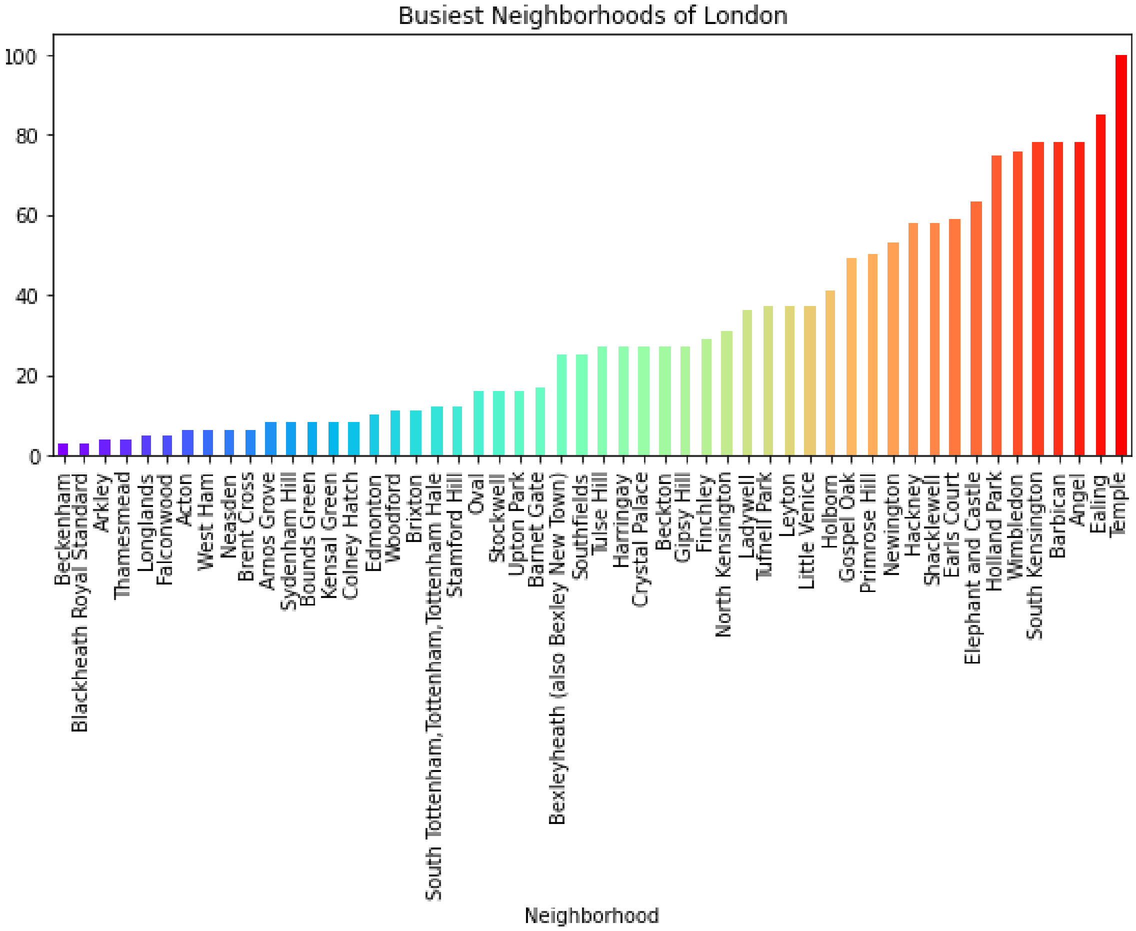


```
Beckenham
Bexleyheath (also Bexley New Town)
Mile End,Ratcliff,Shadwell,Spitalfields,Stepney,Wapping,Whitechapel
Lea Bridge
Leyton
Wanstead
Snaresbrook
Cann Hall,Leytonstone
Little Ilford,Manor Park
Plaistow
West Ham
Blackwall,Canary Wharf,Cubitt Town,Isle of Dogs,Leamouth,Limehouse,Millwall,Poplar
Maryland,Stratford
Canning Town,Custom House,North Woolwich,Silvertown
Upper Walthamstow,Walthamstow,Walthamstow Village
South Woodford
Haggerston
Bethnal Green,Cambridge Heath
Bow,Bromley (also Bromley-by-Bow),Old Ford
```

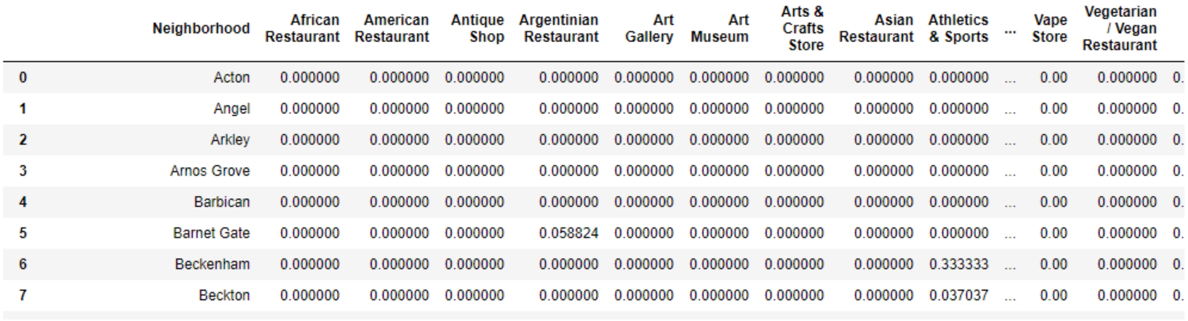


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beckenham	51.415095	-0.035403	HSBC Sports And Social Club	51.417413	-0.036904	Athletics & Sports
1	Beckenham	51.415095	-0.035403	New Beckenham Railway Station (NBC)	51.413734	-0.032354	Train Station
2	Beckenham	51.415095	-0.035403	Cator Park	51.413864	-0.040178	Park
3	Bexleyheath (also Bexley New Town)	51.453495	0.151155	Zizzi	51.455929	0.150555	Italian Restaurant
4	Bexleyheath (also Bexley New Town)	51.453495	0.151155	Prince Albert	51.455171	0.152965	Pub

We also, have plotted a graph showing the popular neighbourhoods based on count of venues received from foursquare the most popular neighbourhoods.

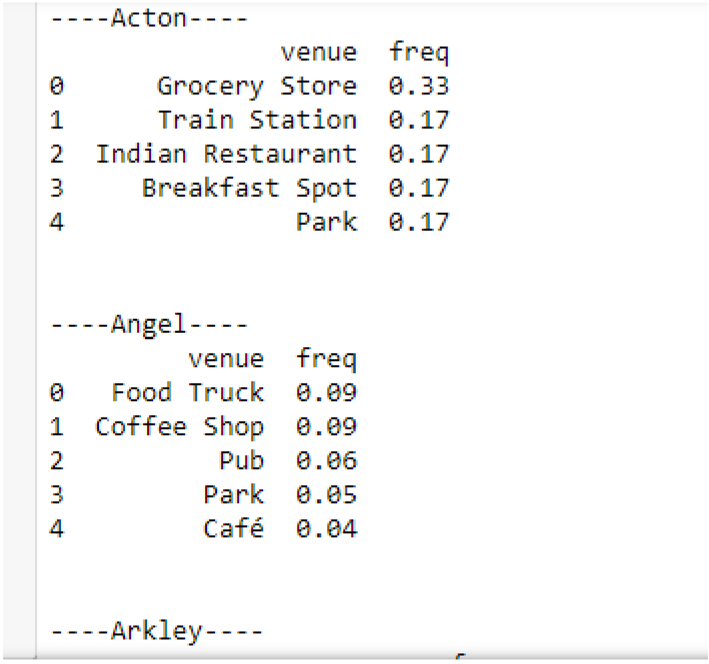


We would then analyse the neighbourhoods by using Machine Learning techniques such as K-Means Clustering. Next, we group rows by neighbourhood and by taking the mean of the frequency of occurrence of each category.



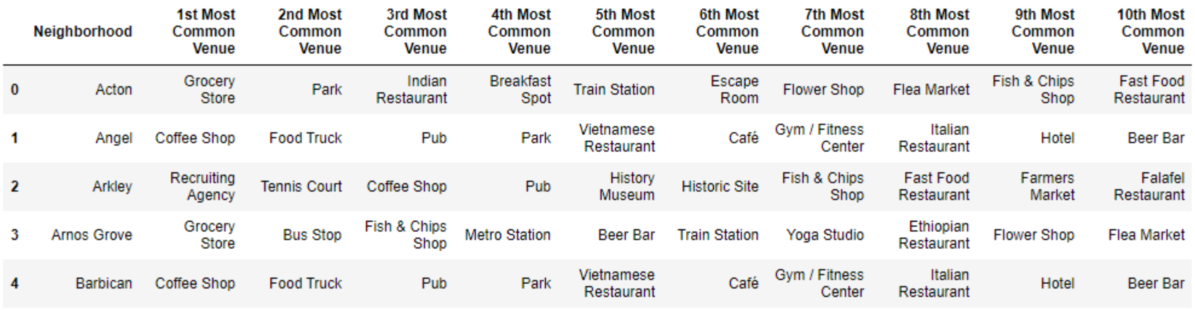
	Neighborhood	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Vape Store	Vegetarian / Vegan Restaurant	
0	Acton	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.
1	Angel	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.
2	Arkley	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.
3	Arnos Grove	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.
4	Barbican	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.
5	Barnet Gate	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.000000	0.
6	Beckenham	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	...	0.00	0.000000	0.
7	Beckton	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.037037	...	0.00	0.000000	0.

We too found out the top 5 common venues of each neighbourhood.



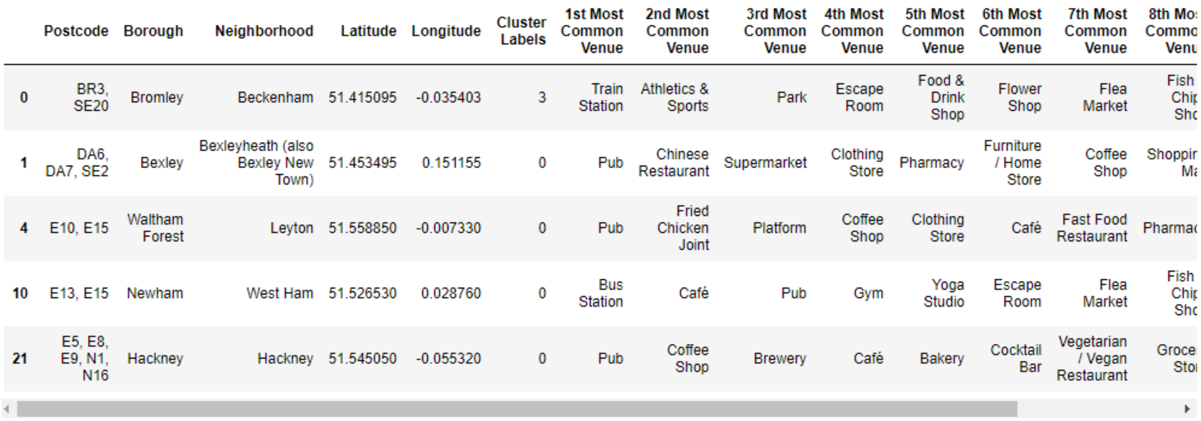
```
----Acton----
               venue  freq
0      Grocery Store  0.33
1      Train Station  0.17
2  Indian Restaurant  0.17
3     Breakfast Spot  0.17
4               Park  0.17


----Angel----
         venue  freq
0   Food Truck  0.09
1  Coffee Shop  0.09
2          Pub  0.06
3         Park  0.05
4         Café  0.04


----Arkley----
```

And , regrouped and created a new dataframe and display the top 10 venues for each neighbourhood.



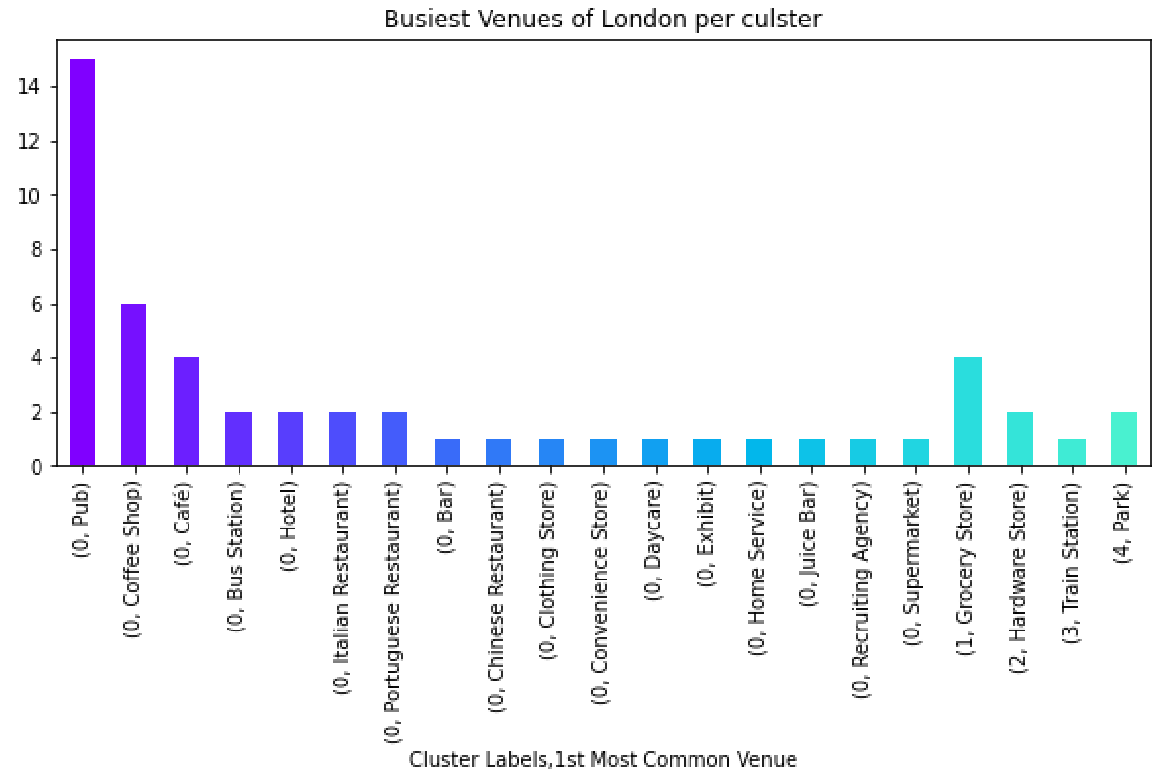
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Acton	Grocery Store	Park	Indian Restaurant	Breakfast Spot	Train Station	Escape Room	Flower Shop	Flea Market	Fish & Chips Shop	Fast Food Restaurant
1	Angel	Coffee Shop	Food Truck	Pub	Park	Vietnamese Restaurant	Café	Gym / Fitness Center	Italian Restaurant	Hotel	Beer Bar
2	Arkley	Recruiting Agency	Tennis Court	Coffee Shop	Pub	History Museum	Historic Site	Fish & Chips Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant
3	Arnos Grove	Grocery Store	Bus Stop	Fish & Chips Shop	Metro Station	Beer Bar	Train Station	Yoga Studio	Ethiopian Restaurant	Flower Shop	Flea Market
4	Barbican	Coffee Shop	Food Truck	Pub	Park	Vietnamese Restaurant	Café	Gym / Fitness Center	Italian Restaurant	Hotel	Beer Bar

After, seeing the common venues, we applied K-means to cluster the neighbourhoods into 5 clusters, and assign cluster label to each neighbourhood.



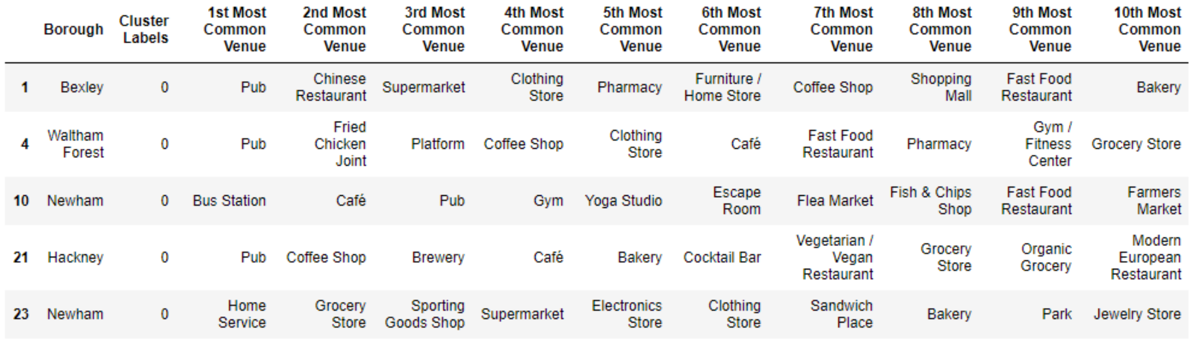
	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mos Commo Venu
0	BR3, SE20	Bromley	Beckenham	51.415095	-0.035403	3	Train Station	Athletics & Sports	Park	Escape Room	Food & Drink Shop	Flower Shop	Flea Market	Fish Chip Sho
1	DA6, DA7, SE2	Bexley	Bexleyheath (also Bexley New Town)	51.453495	0.151155	0	Pub	Chinese Restaurant	Supermarket	Clothing Store	Pharmacy	Furniture / Home Store	Coffee Shop	Shoppir Ma
4	E10, E15	Waltham Forest	Leyton	51.558850	-0.007330	0	Pub	Fried Chicken Joint	Platform	Coffee Shop	Clothing Store	Café	Fast Food Restaurant	Pharmac
10	E13, E15	Newham	West Ham	51.526530	0.028760	0	Bus Station	Café	Pub	Gym	Yoga Studio	Escape Room	Flea Market	Fish Chip Sho
21	E5, E8, E9, N1, N16	Hackney	Hackney	51.545050	-0.055320	0	Pub	Coffee Shop	Brewery	Café	Bakery	Cocktail Bar	Vegetarian / Vegan Restaurant	Groce Stor

We too plotted the busiest venues of London using clusters.



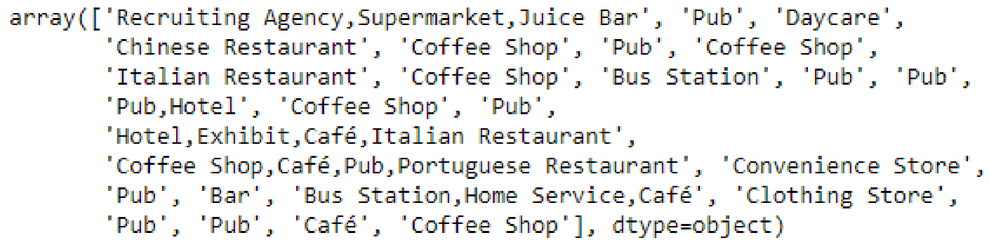
Following are the five clusters with made and their observations :

CLUSTER 1 :



	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Bexley	0	Pub	Chinese Restaurant	Supermarket	Clothing Store	Pharmacy	Furniture / Home Store	Coffee Shop	Shopping Mall	Fast Food Restaurant	Bakery
4	Waltham Forest	0	Pub	Fried Chicken Joint	Platform	Coffee Shop	Clothing Store	Café	Fast Food Restaurant	Pharmacy	Gym / Fitness Center	Grocery Store
10	Newham	0	Bus Station	Café	Pub	Gym	Yoga Studio	Escape Room	Flea Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market
21	Hackney	0	Pub	Coffee Shop	Brewery	Café	Bakery	Cocktail Bar	Vegetarian / Vegan Restaurant	Grocery Store	Organic Grocery	Modern European Restaurant
23	Newham	0	Home Service	Grocery Store	Sporting Goods Shop	Supermarket	Electronics Store	Clothing Store	Sandwich Place	Bakery	Park	Jewelry Store

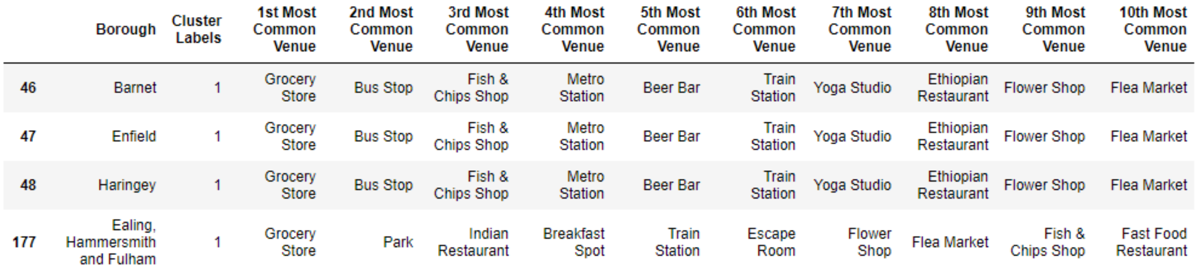
Observation :



```
array(['Recruiting Agency,Supermarket,Juice Bar', 'Pub', 'Daycare',
       'Chinese Restaurant', 'Coffee Shop', 'Pub', 'Coffee Shop',
       'Italian Restaurant', 'Coffee Shop', 'Bus Station', 'Pub', 'Pub',
       'Pub,Hotel', 'Coffee Shop', 'Pub',
       'Hotel,Exhibit,Café,Italian Restaurant',
       'Coffee Shop,Café,Pub,Portuguese Restaurant', 'Convenience Store',
       'Pub', 'Bar', 'Bus Station,Home Service,Café', 'Clothing Store',
       'Pub', 'Pub', 'Café', 'Coffee Shop'], dtype=object)
```

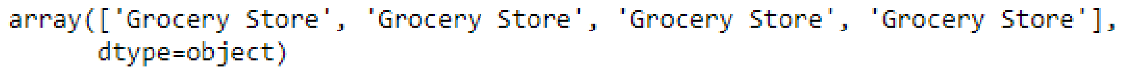
We can see that this cluster is famous for pubs, Restaurants and cafe seems like totaly go out place.

CLUSTER 2 :



	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
46	Barnet	1	Grocery Store	Bus Stop	Fish & Chips Shop	Metro Station	Beer Bar	Train Station	Yoga Studio	Ethiopian Restaurant	Flower Shop	Flea Market
47	Enfield	1	Grocery Store	Bus Stop	Fish & Chips Shop	Metro Station	Beer Bar	Train Station	Yoga Studio	Ethiopian Restaurant	Flower Shop	Flea Market
48	Haringey	1	Grocery Store	Bus Stop	Fish & Chips Shop	Metro Station	Beer Bar	Train Station	Yoga Studio	Ethiopian Restaurant	Flower Shop	Flea Market
177	Ealing, Hammersmith and Fulham	1	Grocery Store	Park	Indian Restaurant	Breakfast Spot	Train Station	Escape Room	Flower Shop	Flea Market	Fish & Chips Shop	Fast Food Restaurant

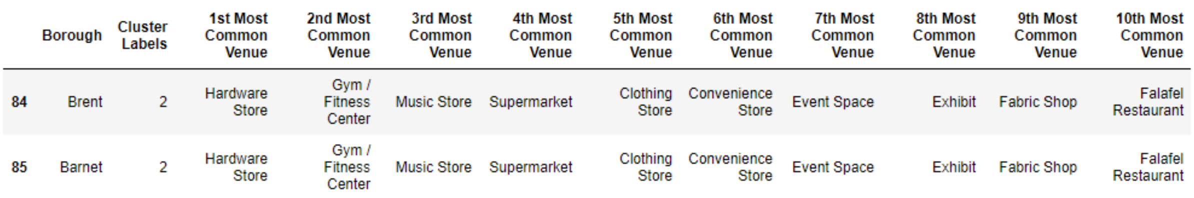
OBSERVATION :



```
array(['Grocery Store', 'Grocery Store', 'Grocery Store', 'Grocery Store'],
      dtype=object)
```

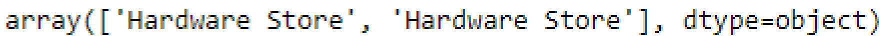
We can see this place is famous for Grocery stores

CLUSTER 3 :



	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
84	Brent	2	Hardware Store	Gym / Fitness Center	Music Store	Supermarket	Clothing Store	Convenience Store	Event Space	Exhibit	Fabric Shop	Falafel Restaurant
85	Barnet	2	Hardware Store	Gym / Fitness Center	Music Store	Supermarket	Clothing Store	Convenience Store	Event Space	Exhibit	Fabric Shop	Falafel Restaurant

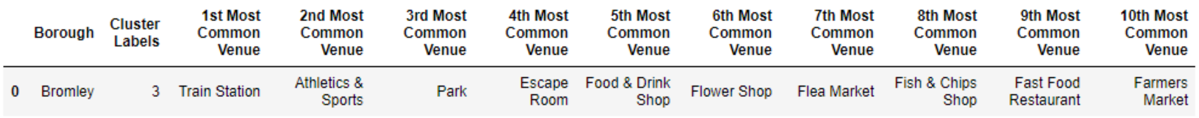
OBSERVATION :



```
array(['Hardware Store', 'Hardware Store'], dtype=object)
```

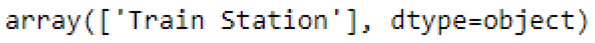
This cluster is all Hardware Stores

CLUSTER 4:



	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bromley	3	Train Station	Athletics & Sports	Park	Escape Room	Food & Drink Shop	Flower Shop	Flea Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market

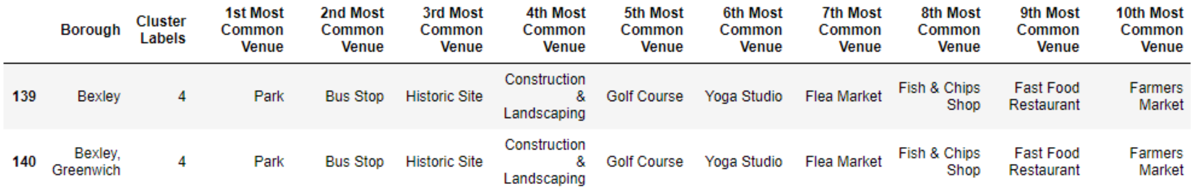
OBSERVATION :



```
array(['Train Station'], dtype=object)
```

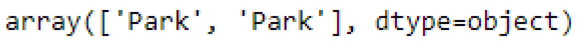
This cluser is more about the public transit

CLUSTER 5 :

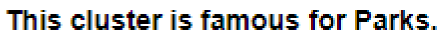


	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
139	Bexley	4	Park	Bus Stop	Historic Site	Construction & Landscaping	Golf Course	Yoga Studio	Flea Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market
140	Bexley, Greenwich	4	Park	Bus Stop	Historic Site	Construction & Landscaping	Golf Course	Yoga Studio	Flea Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market

OBSERVATION:



```
array(['Park', 'Park'], dtype=object)
```



This cluster is famous for Parks.

Results and Discussion :

The following are the highlights of the 5 clusters above:

If someone is deciding to buy a house in London we can see the neighborhood clusters venues and based on venues decide where he should buy the house basically analysing the locality.

1.If the person is party animal he should go for cluster1.

2.If the person is not fan of outdoor eats then he needs groceries a lot and cluster 2 has popular grocerey stores.

3.If the person wants some hardware, cluster 3 has famous hardware stores.

4.If he has travelling as his prefrence then public transit is famous in cluster 4 area's.

5.If the person want to spend some time in Parks, then it is in found in Cluster 5.

As I mentioned before, London is a big city with a high population density in a narrow area. The total number of measurements and population densities of the 33 districts in total can vary. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this metropol.

I used the Kmeans algorithm as part of this clustering study. I set the optimum k value to 5. However, only 33 district coordinates were used. For more detailed and accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled.

I also performed data analysis through this information by adding the coordinates of districts as static data on GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and clustering information on the London map. In future studies, web or telephone applications can be carried out to direct investors.

CONCLUSION :

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.