CLUSTER ANALYSIS OF LONDON VENUES AND HOUSE PRICES

Nasrin Babanli

Problem description

- Most of you know that there is a continuous flow to big financial centers of the world, one of which is London.
- London is considered the second in the Global Financial Centres Index ranks of the world's top financial centers. Lots of people from different countries get job offers in this big city each year.
- Newcomers to this city are unfamiliar with house prices of each neighborhood of London and venues located nearby.
- In this project I am going to describe different neighborhoods of London and cluster each neighborhood by **venue type** and **average house prices** for London.

Data Acquisition and Cleaning

- In the first stage I have downloaded data through **wget** command in order to access the data. Then I dropped hyperlinks to Wikipedia references, brackets and duplicate string values from the data frames for postcodes.
- In property prices data frame I removed currency signs, decimal places and also converted average prices to numeric values.

```
In [8]: london_df.head()
```

Out[8]:

	Location	London_borough	Post town	Postcode district	Dial Code	OS grid ref
0	Abbey Wood	Bexley	LONDON	SE2	020	TQ465785
1	Acton	Ealing	LONDON	W3	020	TQ205805
2	Addington	Croydon	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY	DA5	020	TQ478728

	Area	Avg price
0	BR1	439284
1	BR2	456361
2	BR3	439013
3	BR4	555188
4	BR5	431399

Data Acquisition and Cleaning

- Then I merged the tables of London Boroughs and London House Prices into a unique table consisting of all relevant data to be used in clustering.
- After merging two tables, I found the Longitudes and Latitudes of each neighborhood by using geocoder library and added these data to final dataframe.

	Location	London_borough	Post town	Postcode district	Dial Code	OS grid ref	Avg price
0	Abbey Wood	Bexley	LONDON	SE2	020	TQ465785	340136
1	Crossness	Bexley	LONDON	SE2	020	TQ480800	340136
2	West Heath	Bexley	LONDON	SE2	020	TQ475775	340136
3	Acton	Ealing	LONDON	W3	020	TQ205805	531557
4	Addington	Croydon	CROYDON	CR0	020	TQ375645	347140

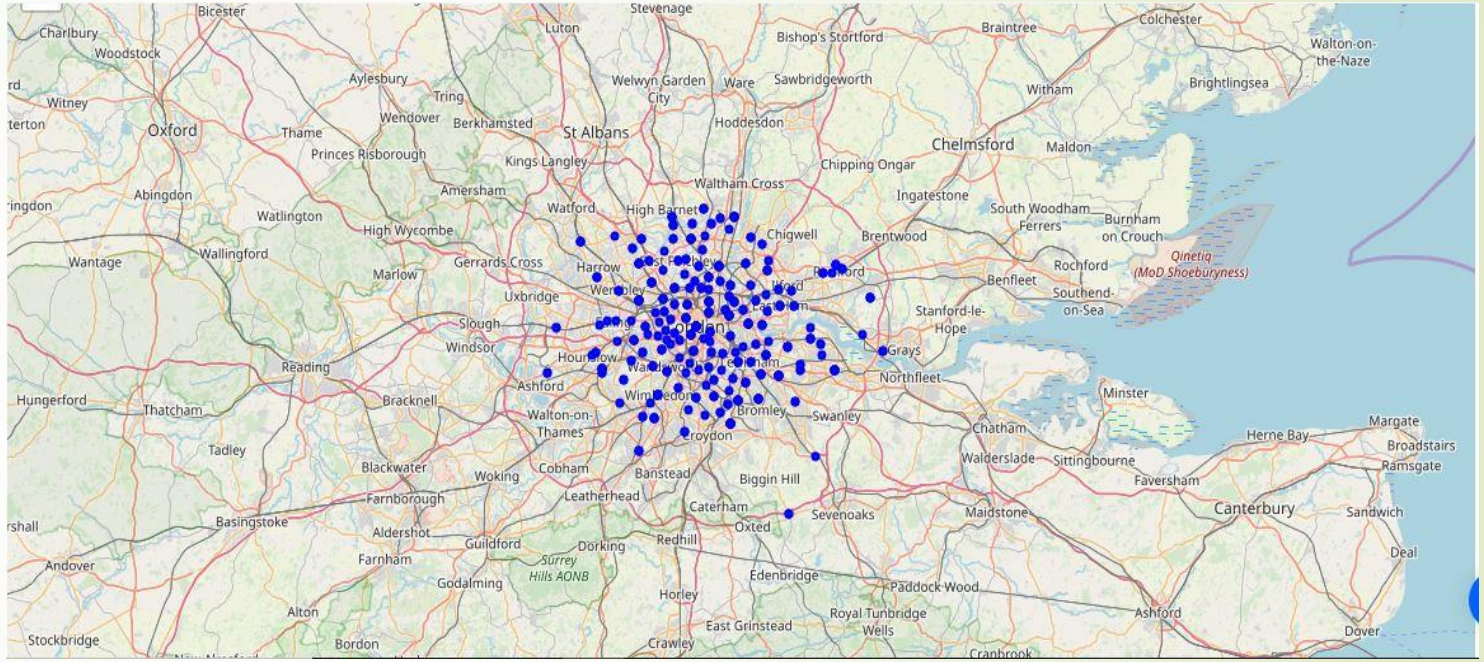
```
london_data[['Latitude','Longitude']]=pd.DataFrame(coords_list,columns=['Latitude', 'Longitude'])
london_data.head()
```

]:

	Location	London_borough	Post town	Postcode district	Dial Code	OS grid ref	Avg price	Latitude	Longitude
0	Abbey Wood	Bexley	LONDON	SE2	020	TQ465785	340136	51.492450	0.121270
1	Crossness	Bexley	LONDON	SE2	020	TQ480800	340136	51.492450	0.121270
2	West Heath	Bexley	LONDON	SE2	020	TQ475775	340136	51.492450	0.121270
3	Acton	Ealing	LONDON	W3	020	TQ205805	531557	51.513240	-0.267460
4	Addington	Croydon	CROYDON	CR0	020	TQ375645	347140	51.384755	-0.051499

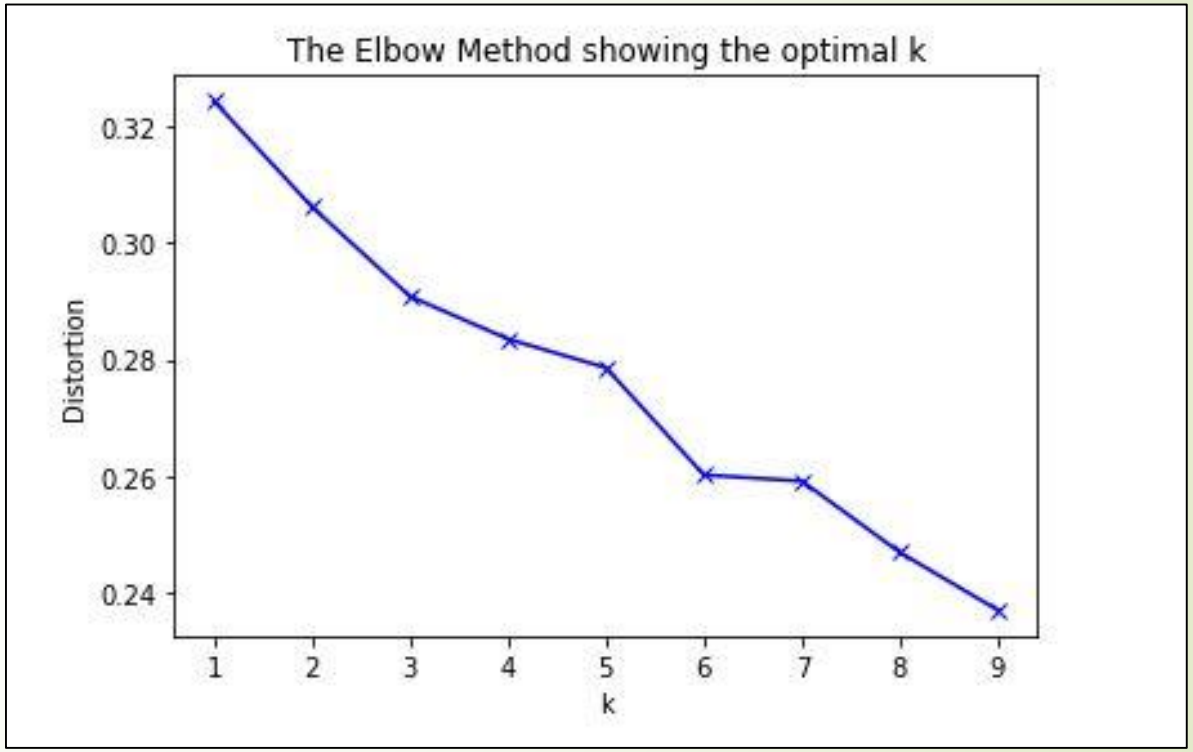
Data Vizualization

- Then I used python folium library to visualize neighborhoods and boroughs of London in a single map, where latitude and longitude data retrieved in previous step helped me to visualize it.



Defining cluster size by venue type

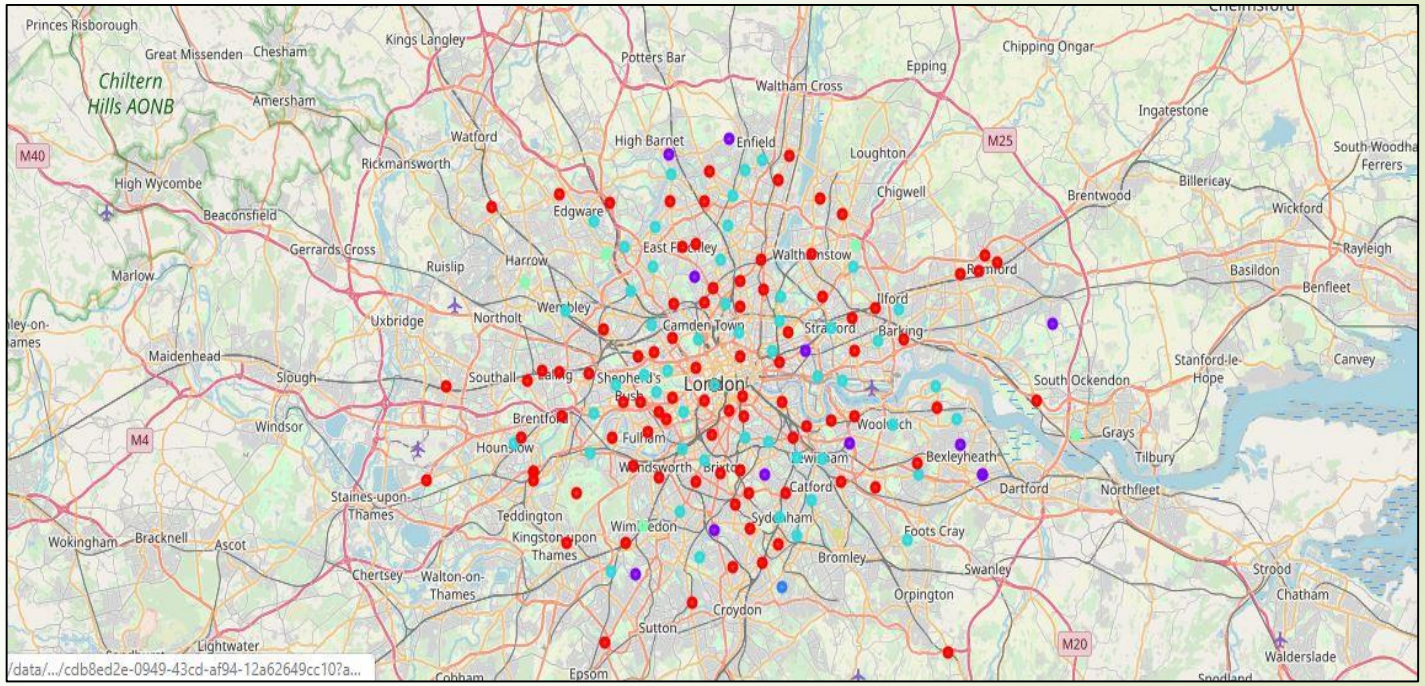
- I have firstly tried K-Means algorithm with 7 clusters and then visualized most optimal cluster size with K-Means elbow method in order to get optimal amount of k.
- I have found that optimal amount for clusters is 6 in this analysis and therefore divided venues into 6 categories and labelled them as follows:
  - *Hotels and Various Social Venues*
  - *Stores and seafood restaurants*
  - *Pubs and Historic venues*
  - *Fitness centers*
  - *Restaurants and Bars*



Cluster map of London

Then I visualized clusters in a separate London cluster map, where

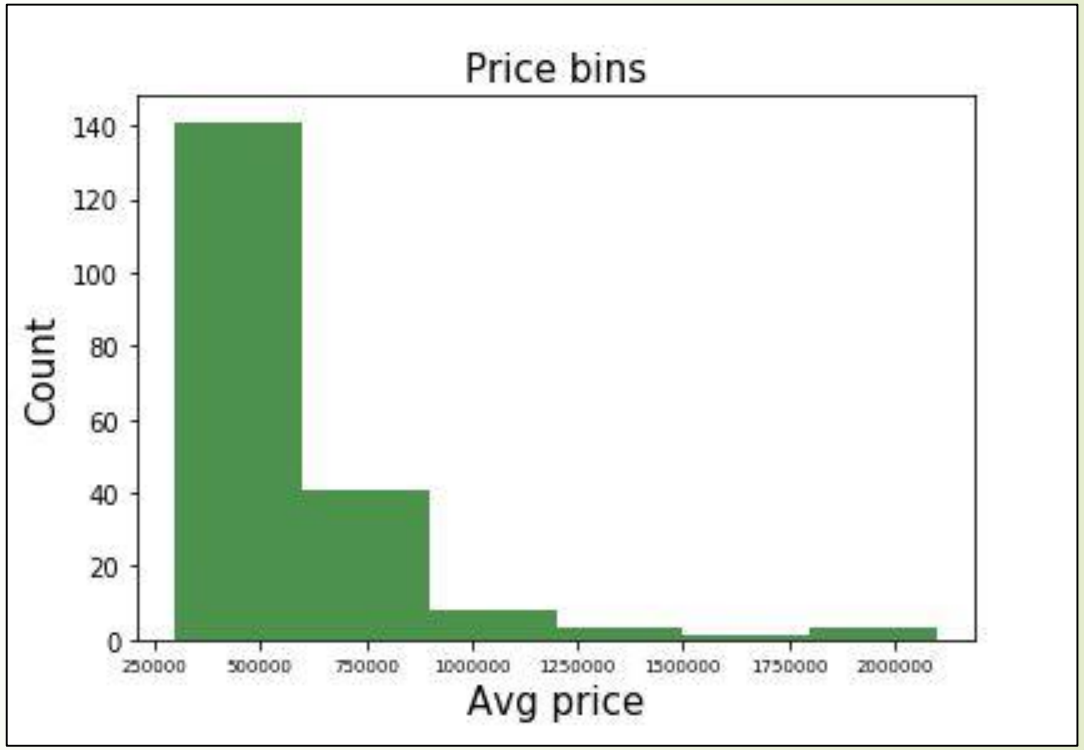
- **red points** indicate *cluster 0*,
- **purple points** indicate *cluster 1*,
- **blue points** indicate *cluster 2*,
- **aqua points** indicate *cluster 3*,
- **green points** indicate *cluster 4*, and
- **khaki points** indicate *cluster 5*.



Dividing house prices into bins

I have divided house prices into 7 bins as follows:

- Very low price
- Low price
- Lower average price
- Average price
- Higher average price
- High price
- Very high price



Results

- **The most expensive** regions in London are places located closer to Hotels and Various Social venues, pubs and historic places in downtown which correspond to clusters 0 and 3, which are **Kensington, Chelsea and Westminster** areas.
- **Houses with average prices** (around 1 mln GBP) are also located mainly in **Richmond, Camden, Kensington and Chelsea** regions with great variety of historic and various social venues.
- Lo**w price houses** are located in other clusters of London, mainly in **Croydon, Bexley and Barnet** boroughs.