CLUSTER ANALYSIS OF LONDON VENUES AND HOUSE PRICES

*

Nasrin Babanli

*

1. Introduction

Problem Desrciption

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. That is why in order to facilitate decision making process on making best choice in neighborhood I have prepared a comprehensive analysis with map in order to better vizualize the variety of choices.</p>

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.

In the frist stage of my notebook in github repository I will describe libraries I used for my data analysis and source for postal codes and neighborhoods of London. In stage 2 I will show sources for property data of London and cleansing this data. In stage 3 I will show how I got longitude and latitude data for each neighborhood and merged this data frame with average house prices data frame created in the second stage. In stage 4 I retrieved datas for London venues from Foursquare API by defining Foursquare credentials. In stage 5 I will show one of the most popular Machine Learning tools called K-means algorithm to cluster the neighborhoods by venue type, vizualized the clusters with Folium map, binned average house prices in 7 distinct categories and grouped the venues in various categories within 6 clusters.

2. Data Acquisition and Cleaning

2.1. Data Sources

The data for postcodes and neighborhoods of London was taken from Wikipedia website. Average house prices were taken from website showing property data for London.

The main task for the first and second stages was to scrape source websites and wrangle the needed data, clean it, and then read it into a pandas dataframe so that it is in a structured format.

2.2 Data Cleaning

Firstly, I have imported main libraries to be used in my research project. I have used numpy library for handling vectorized data, pandas library for data analysis, json library for handling files in JSON format, Nominatim for converting address into latitude and longitude values, Folium for creating maps, Matplotlib for plotting modules and other libraries used for different needs.

In the first stage I have downloaded data through **wget** command in orfer to access the data. Then I dropped hyperlinks to Wikipedia references, brackets and duplicate string values from the data frames for postcodes.

```
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

In property prices data frame I removed currency signs, decimal places and aldo converted average prices to numeric values.

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

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.

	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

After merging two tables, I found the Longitudes and Latitudes of each neighborhood by using geocoder library and added these data to final dataframe.

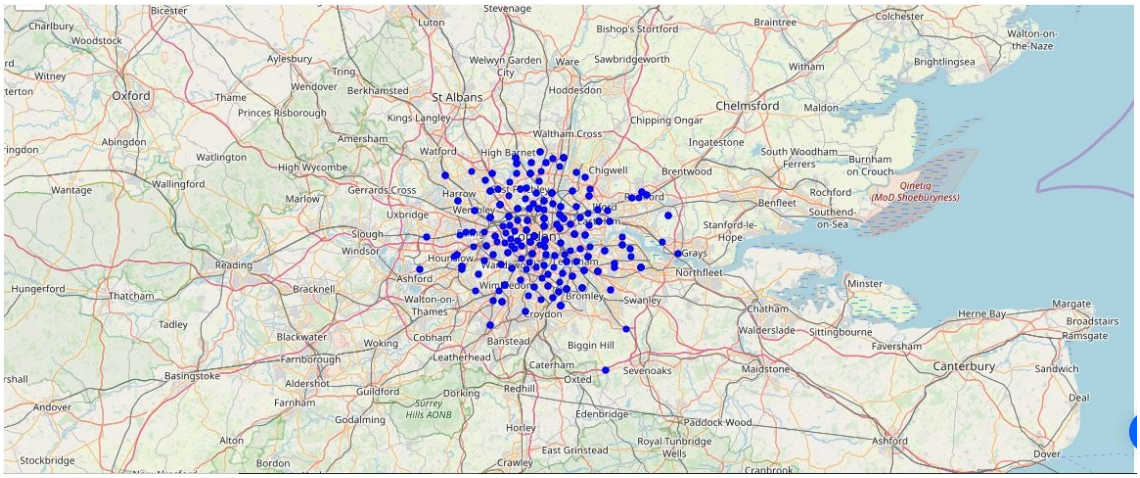
```
london_data[['Latitude','Longitude']]=pd.DataFrame(coords_list,columns=['Latitude', 'Longitude'])
london_data.head()
```

```
]:
```

	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

2.3. Data Vizualization

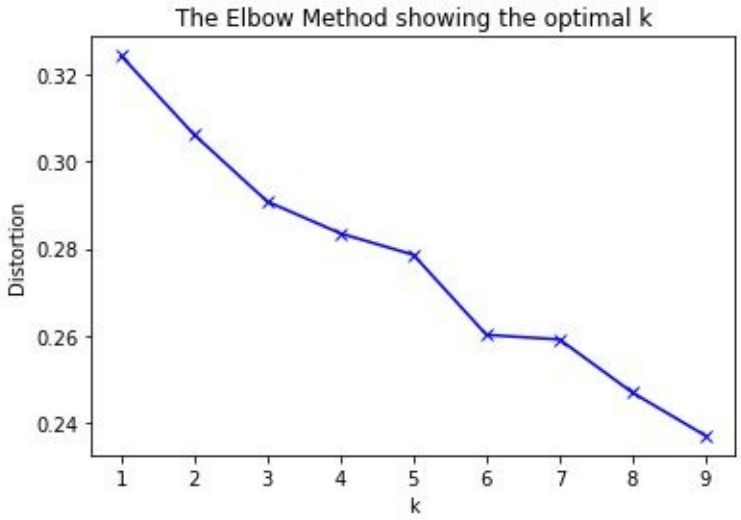
Then I used python folium library to visualize neighborhoods and boroughs of London in a single map, where latitude and longitude data retrieved in previous step helped me to vizualize it.



3. Methodology

3.1. Defining cluster size by venue type

I got some common categories for venues in London boroughs and in order to cluster them I have used one of the most popular machine learning algorithm called K-means to cluster London boroughs. I have firstly tried K-Means algorithm with 7 clusters and then vizualized 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

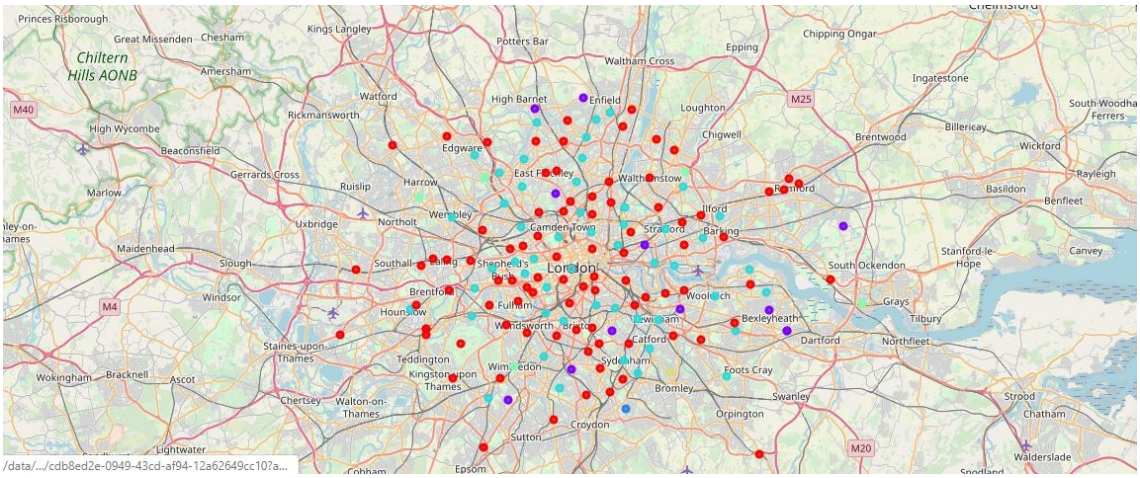
Stores and seafood restaurants

Pubs and Historic venues

Fitness centers

Restaurants and Bars

Then I vizualized 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*.

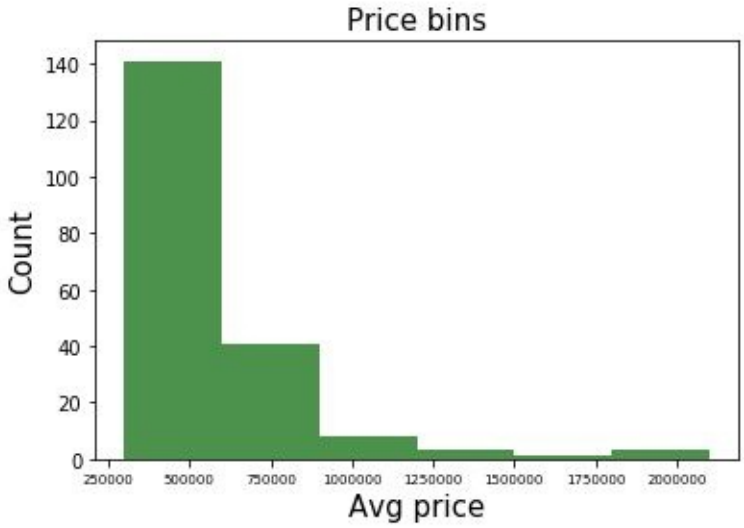


3.2. Dividing house prices into bins

I have divided house prices into 7 bins as follows:

1. Very low price
2. Low price
3. Lower average price
4. Average price
5. Higher average price
6. High price
7. Very high price

I vizualized bin sizes in the histogram as follows:



4. Results

According to analysis of venues and average house prices in London I have concluded that 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.

Low price houses are located in other clusters of london, mainly in Croydon, Bexley and Barnet boroughs.

5. Conclusion

In conclusion I can say that my project about clustering London boroughs into different categories based on house prices and venue types and vizualizing them in a single map can help newcomers to London for orientation in this big city.

You can read full analysis of the project in this following github repository as a pdf file.