Final report for IBM Capstone Project

Segmenting and Clustering Districts in London

Author name: Helena Ferreira Pinto

Submitted in partial fulfilment of IBM's Professional Certificate in Data Science

April 2020

Feedback box for project markers:

WHAT I liked about the report:
WHAT could/should be improved:

1. Introduction

In this project we will explore London Districts. Specifically, this report will be targeted to stakeholders interested in moving to London, UK, who would like to choose a neighborhood based on the type of venues present at close proximity.

We will use data science tools to generate clusters of similar neighborhoods, therefore allowing our client to more easily investigate the different areas.

2. Data

The following data sources were used:

- List of London postcode areas was obtained from Wikipedia
  https://en.wikipedia.org/wiki/London_postal_district
- Coordinates of London and the different postcode districts was obtained using the Geocoder Python package
  https://pypi.org/project/geocoder/
- Number of venues and their category in every district area was obtained using Foursquare API
  https://foursquare.com/

3. Methodology

In the first step of our analysis, we have collected the required data: location and type (category) of the most common venues in London Districts.

The second step of our analysis was the calculation and exploration of 'venue frequency' across different district areas.

In the third and final step, we created clusters of locations that contain similar most common venues using Kmeans clustering.

This will allow our client to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal home location.

4. Analysis

4.1. Part 1- Web scraping

We import the data with London postcodes and respective District Names, from wikipedia page: https://en.wikipedia.org/wiki/London_postal_district

The data is processed and inserted in a pandas dataframe in way that facilitates its further use for display in maps, as shown in Figure 1.

	index	District	District Name	Postcode	Latitude	Longitude
0	0	Eastern	Head district	E1	51.52022	-0.05431
1	1	Eastern	Bethnal Green	E2	51.52669	-0.06257
2	2	Eastern	Bow	E3	51.52702	-0.02594
3	3	Eastern	Chingford	E4	51.61780	-0.00934
4	4	Eastern	Clapton	E5	51.55897	-0.05323
...	...	...	...	...	...	...
115	116	Paddington	Shepherds Bush	W12	51.50645	-0.23691
116	117	Paddington	West Ealing	W13	51.51453	-0.31951
117	118	Paddington	West Kensington	W14	51.49568	-0.20993
118	119	Western Central	Head district	WC1	51.52450	-0.12273
119	120	Western Central	Strand	WC2	51.51651	-0.11968

120 rows × 6 columns

Figure 1: DataFrame containing London District data

4.2. Part 2- Geospatial Coordinates

We use Python's Geocoder package to obtain geospatial coodinates of London's Postcodes. We then display that data using Folium. Refer to Figure 2 for a visualization of location of the different districts in London.

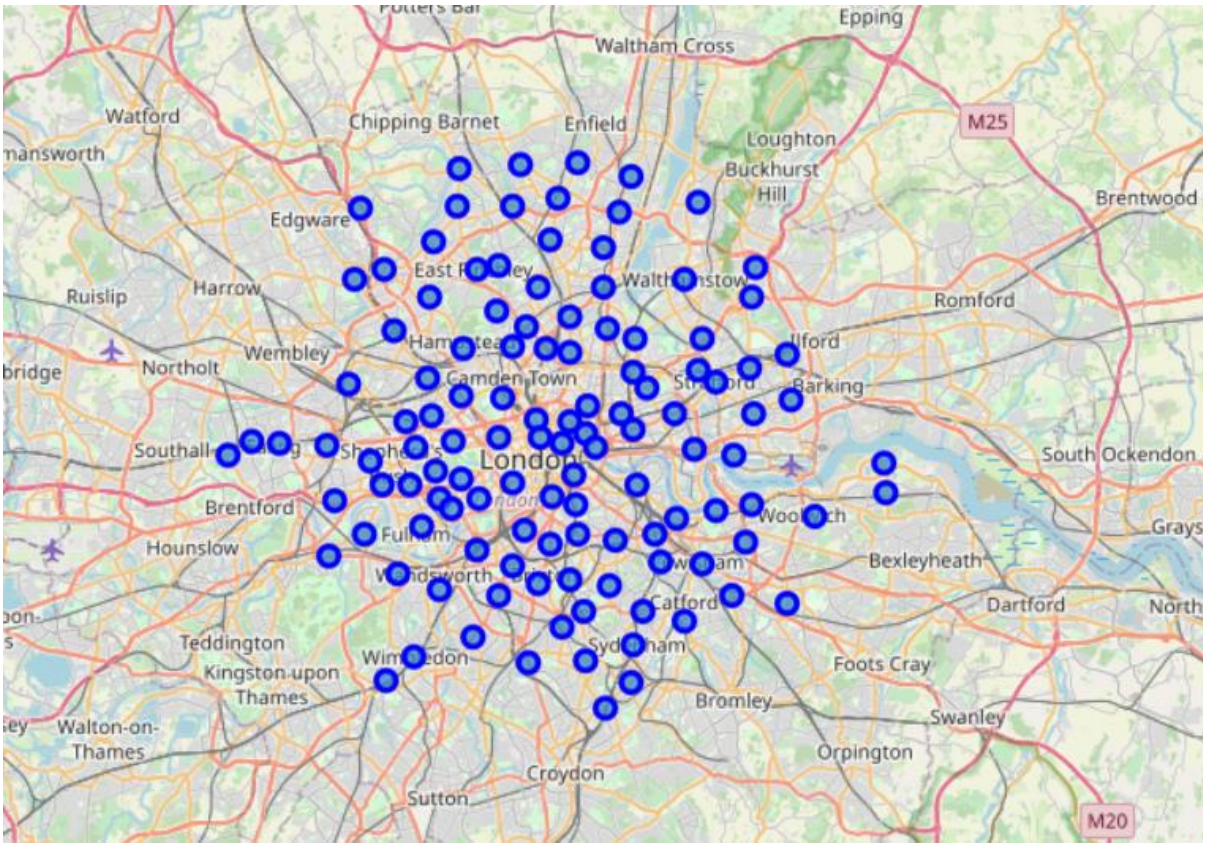

Figure 2: Map of London showing different postal Districts

4.3. Part 3- Explore and cluster the districts in London

In this part, we used a machine learning algorithm, which is part of unsupervised learning: k-means clustering.

Firstly, we calculate the frequency of different venue categories in each location. We then determine the 10 most common venues in each location. Locations are then clustered based on the types of most common locations. From the Elbow Method for k-means, we choose an optimal value of 10 clusters (cf. Figure 3). Districts within the same cluster should ideally have similar most common venues.

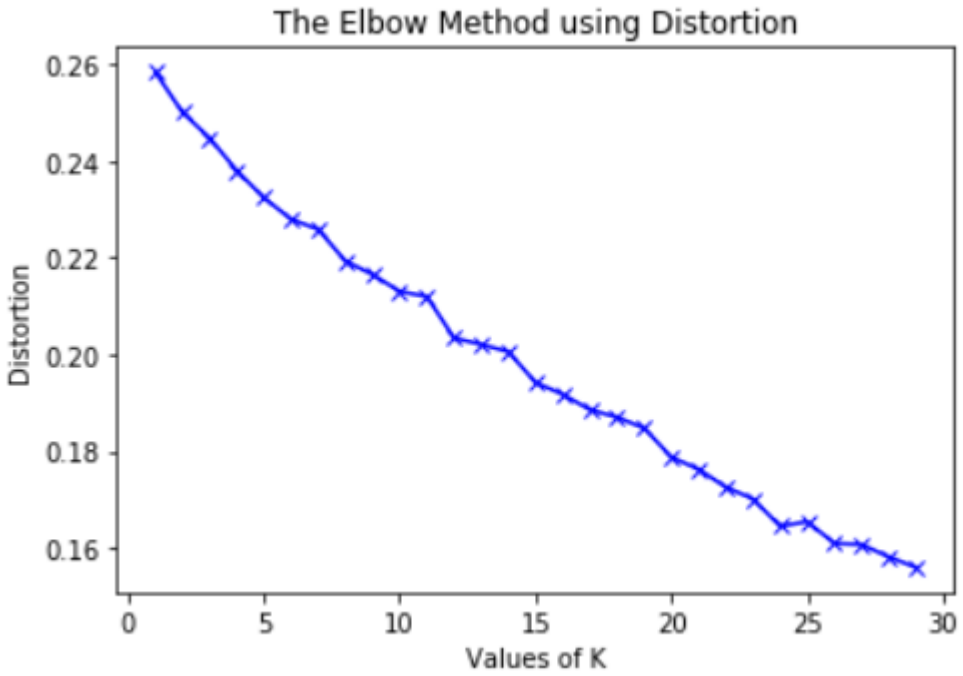

Figure 3: Elbow method for choice of best cluster coefficient

After clustering, we display the results in a map, shown in Figure 4.

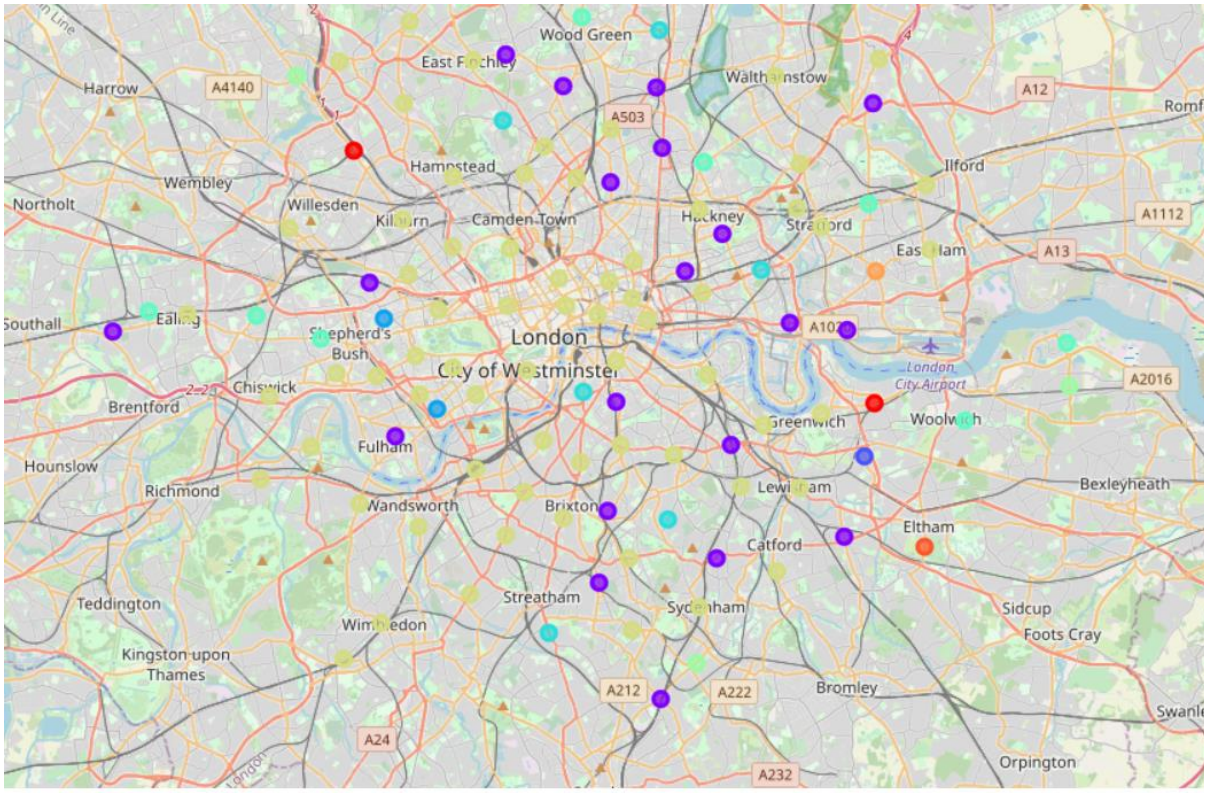

Figure 4: Map of London showing the clustered districts. Each cluster has its own marker colour.

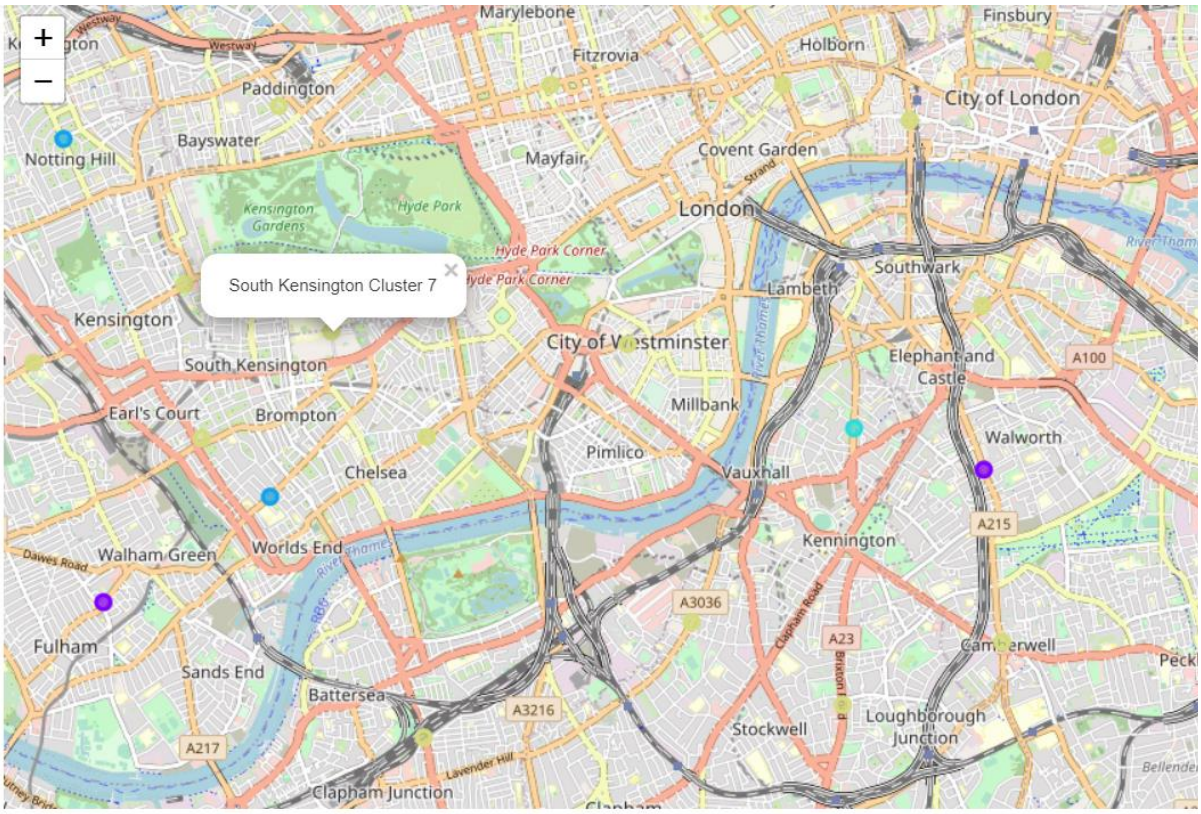

Figure 5: Zoomed in version of Fig 4
By clicking on a marker, we can read the district name and cluster number.

Now, the client can examine each cluster and determine the discriminating venue categories that distinguish them. Based on their preferences, they can them make a more informed decision about which areas suit them best as new location to live.

5. Results & Discussion

Our analysis shows the 10 most common venues in each district area in London (cf. Figure 6).

	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	Abbey Wood	Supermarket	Historic Site	Train Station	Coffee Shop	Platform	Zoo Exhibit	Exhibit	Falafel Restaurant	Farmers Market	Fast Food Restaurant
1	Acton	Grocery Store	Indian Restaurant	Train Station	Breakfast Spot	Park	Fish Market	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
2	Anerley	Supermarket	Convenience Store	Fast Food Restaurant	Hotel	Grocery Store	Zoo Exhibit	Fish Market	Exhibit	Falafel Restaurant	Farmers Market
3	Balham	Coffee Shop	Pub	Bakery	Grocery Store	Burger Joint	Caucasian Restaurant	Shop & Service	Breakfast Spot	Fish & Chips Shop	Gastropub
4	Barnes	Pub	Farmers Market	Park	Italian Restaurant	Community Center	French Restaurant	Movie Theater	Nature Preserve	Food & Drink Shop	Breakfast Spot

Figure 6: First 5 rows of the Dataframe containing the 10 most common venues of each district

10 clusters of relatively similar locations were established, based on these common venues. Cluster 7, containing 73 Districts, is by far the cluster with more locations. Common venues include: coffee shops, pubs, cafés, bakeries, gyms and restaurants.

At this stage, the data should be presented to the client to help inform their decision in choosing their home location. Further study of the results can also be done after receiving feedback from the client. For example, if the presence of a garden is a must for the client, our search can be refined. A Heatmap could also facilitate visualisation of frequencies of certain venues in London.

6. Conclusion

The purpose of this project was to perform an exploratory data analysis of the most common venues in London districts, in order to aid stakeholders in narrowing down the search for the optimal home location.

By calculating venue frequency (using data from Foursquare), we have first identified most common venue categories for each district. We have then created clusters of districts with similar common venues. Clustering was performed in order to create major zones of interest.

Final decision on optimal district will be made by stakeholders based on specific characteristics of neighborhoods and locations in every zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.