Exploring areas around London Underground Stations

Edwin Nicolaas Asberg
October 10, 2019

1.   Introduction

Traffic jam is one of the biggest nightmares of people living in big cities. As the streets get busier and busier with cars, people tend to seek to buy or rent homes within a walking distance from metro stations.

In London, one of the largest cities in the world, this is no different. Londoners enjoy a wide offer of metro stations to commute and travel across the city. But instead they rather prefer to call it 'underground'.

In order to provide more information for people who are seeking to buy or rent their homes in the English capital, this work attempts to offer an analysis to categorize Underground stations according to the type of venues mostly offered around them.

2.   Data

The data used will be the location (latitude and longitude) of London Underground Stations, collected from
https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations#Tube_Stations_List

The venue offer will be retrieved using Foursquare API, which provides geo information and a short description about locations. Foursquare is one of the leading geotagging and review tool, providing rich and insightful data online.

3.   Methodology

As methodology of analysis, k-means clustering will be used. K-means clustering considers the similarity of individuals to for a cluster, minimizing the so-called 'intra cluster' distances and maximizing the 'inter-cluster' distances. Euclidean approach will be used.

First, to see the distribution of the Underground stations in the city of London, they were geolocated using their latitude and longitude information to form the map below.
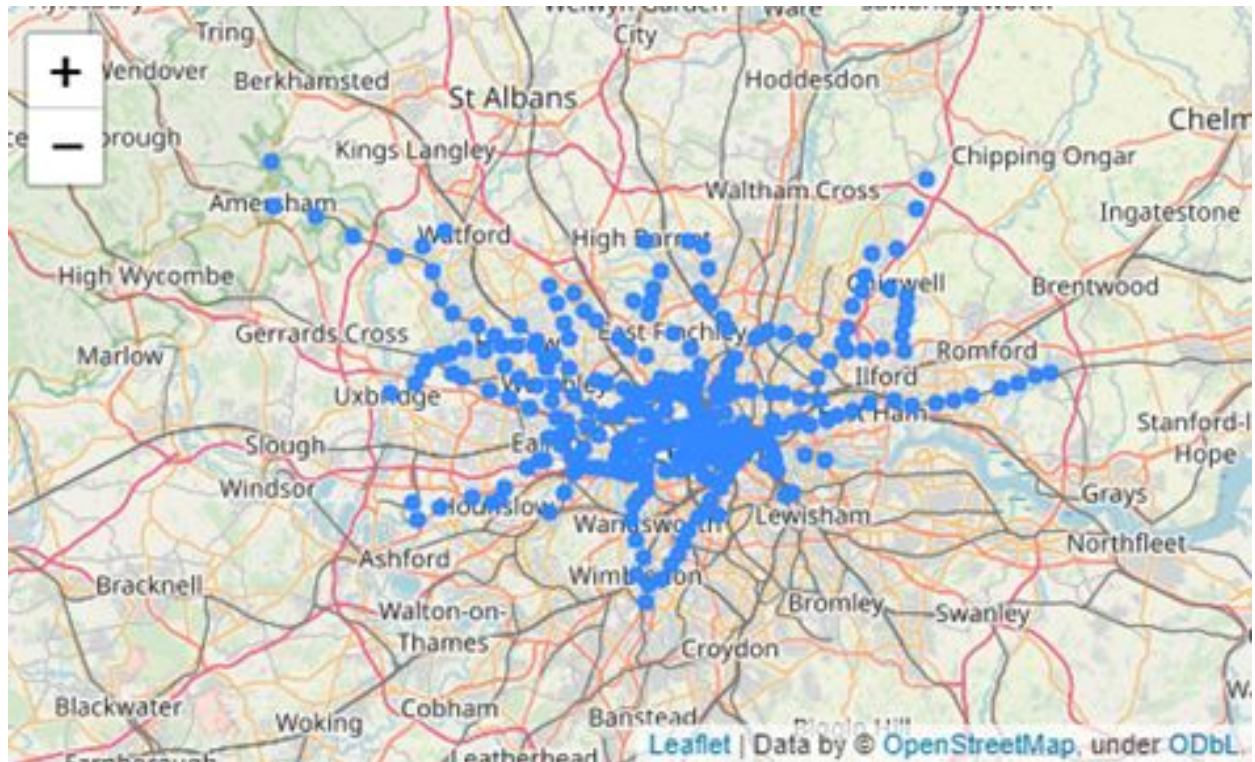

Then, through the Foursquare API, it was calculated the frequency of all venue types within a radius of 500 meters around the metro station. By doing so, we could list the top-5 venues for each station, having a final data set that has its first 5 rows looking as follows:

	Underground	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Acton Central	Gym / Fitness Center	Pub	Coffee Shop	Wine Bar	Bakery
1	Acton Town	Museum	Kebab Restaurant	Pizza Place	Coffee Shop	Fish & Chips Shop
2	Aldgate	Hotel	Coffee Shop	Cocktail Bar	Gym / Fitness Center	Salad Place
3	Aldgate East	Coffee Shop	Pub	Hotel	Indian Restaurant	Café
4	Alperton	Supermarket	Gym / Fitness Center	Indian Restaurant	Middle Eastern Restaurant	Asian Restaurant

4.   Results

By running the K-mean clustering algorithm, the stations where grouped in the 7 clusters, as requested. Geographically, they were distributed as the map below:
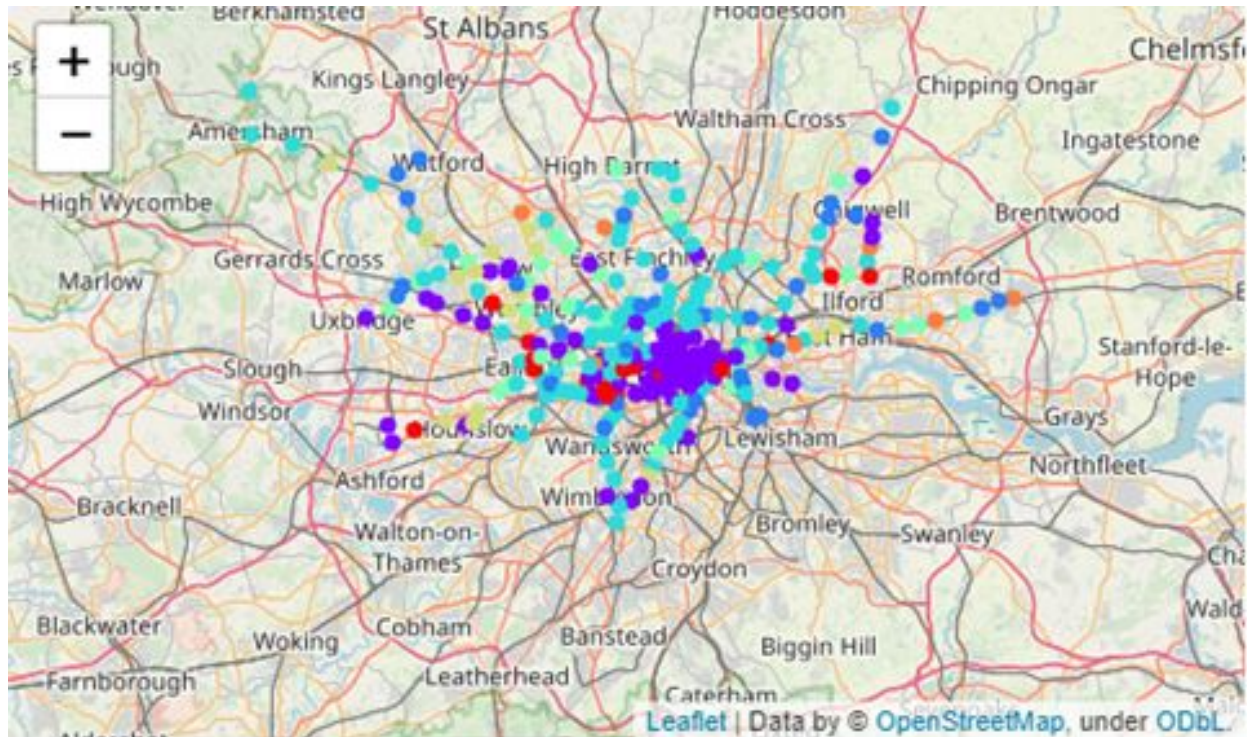
In the map above, we can see that central stations were part of the same cluster (purple dots), and in suburbs another type of cluster was more prevalent (light blue). By seeing the frequency of top venues in each of the cluster, conclusions could be drawn with regards to the type of places offered around the stations (the tables can be found at https://github.com/enasberg/IBM_Data-Science_Certificate/blob/master/Capstone%20Project%20-%20Week%202.ipynb).

- **Cluster 0**: Stations within this cluster are mainly surrounded by Hotels. Pubs are also important venues;
- **Cluster 1**: Around the stations in this cluster, there are a variety of venues with emphasis on Coffee Shops;
- **Cluster 2**: This cluster has Pubs and Cafés as the most common venues;
- **Cluster 3**: This is the largest cluster, meaning the most heterogeneous as well. Coffee Shops and Pubs are well offered around the stations here;
- **Cluster 4**: In this cluster, stations are surrounded by Grocery Stores;
- **Cluster 5**: If you are into Indian food, these stations are the place to live. Indian Restaurants are the most common venues here;
- **Cluster 6**: Gym and Fitness Centre and Parks are well offered in this small cluster (only 7 stations).

5.   Discussion

As an unsupervised machine learning algorithm, k-means clustering outcomes are not always easy to interpret. This technique basically reads out the data and retrieves meaning without any human input to teach it.

Even though this might be the case in some instances, the results from the present analysis were quite straightforward. Most of the clusters had well-defined characteristics, while one or another were to large to have more meaningful information.

To minimize the latter, the script was ran forcing the algorithm to group more clusters, but the end-result was that very small clusters were created. After some trial and error, the number 7 was found to be the best choice.

In the end of the process, the main objective of the analysis was achieved: people aiming to buy or rent a house near an Underground station in London are empowered with more information to choose the areas that best fit their aspirations,

6.   Conclusion

In the end of the day, the k-means clustering algorithm was able to bring insightful information that it was aimed. Therefore, this study showed an innovative way to provide a solution to a problem affecting millions of people around the world.