Clustering stations in London based on Arts & Entertainment venues

Christos Papakonstantinou

25 October 2019

1. Introduction

1.1 Background

London is the most visited city in Europe with millions of tourists each year and the London railway system is the popular choice among visitors for exploring what this city has to offer. However, the railway system is quite complex, and it is confusing for someone who is interested in visiting specific types of venues. Therefore, there is a need to cluster stations based on the Arts & Entertainment venues near them so tourists can easily filter them based on their preferences and better plan their trip.

1.2 Problem

The purpose of this problem is to use the categories from the Foursquare API to cluster stations into groups based on the types of Arts & Entertainment venues near them (e.g. Museums, Art Galleries).

1.3 Interest

Clustering the stations like this and then visualising the clusters gives an interesting view of the city and points which sections are similar. This is also useful for tourists who might be, for example, art enthusiasts and are only interested in visiting all the Art Galleries.

2. Data Acquisition and Cleaning

2.1 Data Sources

The stations dataset, including names, zones, latitudes and longitudes was collected from the website doogal.co.uk and can be found here. The Foursquare API was also used to get the venues close to each station.

2.2 Data Cleaning

All columns except ‘Station’, ‘Latitude’, ‘Longitude’ and ‘Zone’ in the stations DataFrame where dropped. Next, stations outside zone 1 and 2 where removed as these zones constitute the part of London that most tourists will be in. After collecting the venue information, stations that did not have more than four unique categories of venues were removed. This is because it was decided that the stations would be clustered based on their top 4 venue categories for simplicity. Lastly, some categories were combined to improve clustering and because they were too similar (for example Movie Theatre and Indie Movie Theatre).

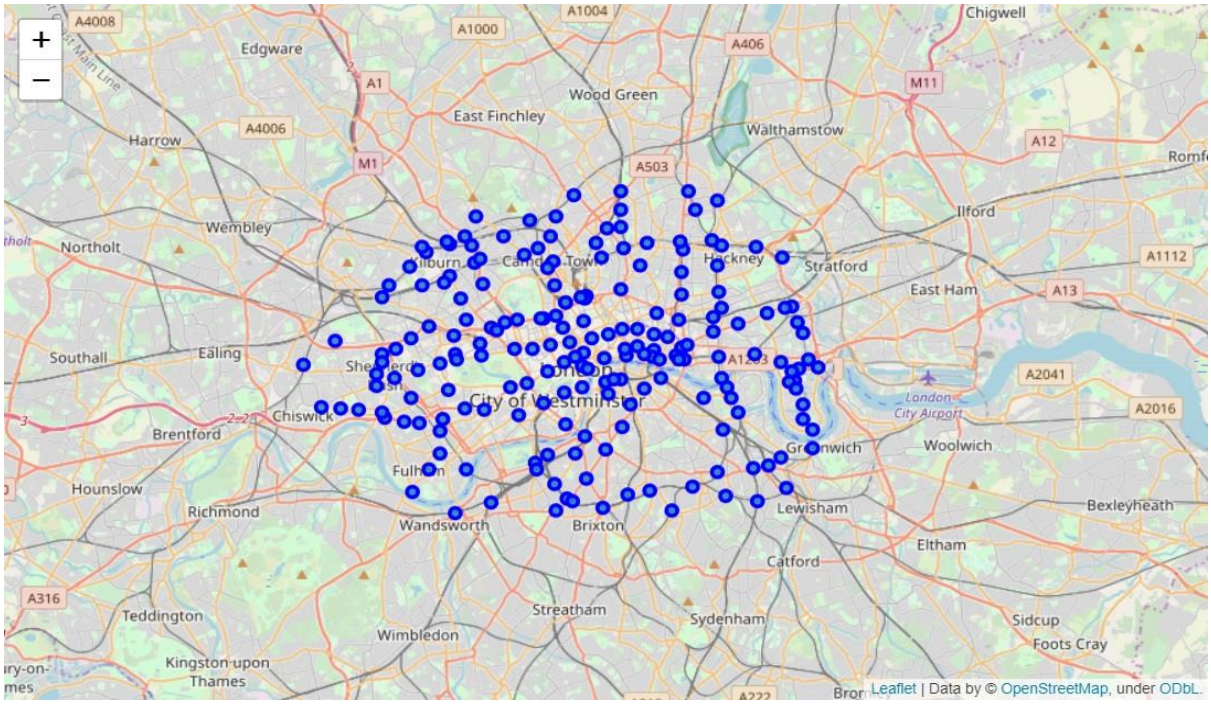

Figure 1: Visualisation of London stations using Folium

3. Methodology

3.1 Preparing Data for K-Means

After cleaning the data there are 117 stations left with four or more unique venue types. Since we are dealing with categorical values, One Hot Encoding was performed on the dataset by using the get_dummies() function. Following this, the DataFrame was grouped by 'Station' getting the mean (frequency) of the venue categories. A DataFrame with the top 4 categories for each venue was created and the data was ready for clustering.

3.2 Machine Learning with K-Means

The 'Elbow Method' was used to find the optimal number of clusters (Figure 2) by calculating the inertia for values of k from 2 to 9. There was no distinct curve however, so the number of clusters was decided intuitively and was given the value of 5. K-Means was run with the 'k-means++' initialisation method which selects initial centroids in a smart way for faster convergence. The algorithm was run a total of 12 times with different centroid seeds (n_init = 12) to increase the chance of getting quality clusters. After the algorithm ran and cluster labels were added to the DataFrame containing the top 4 venues categories for each station.

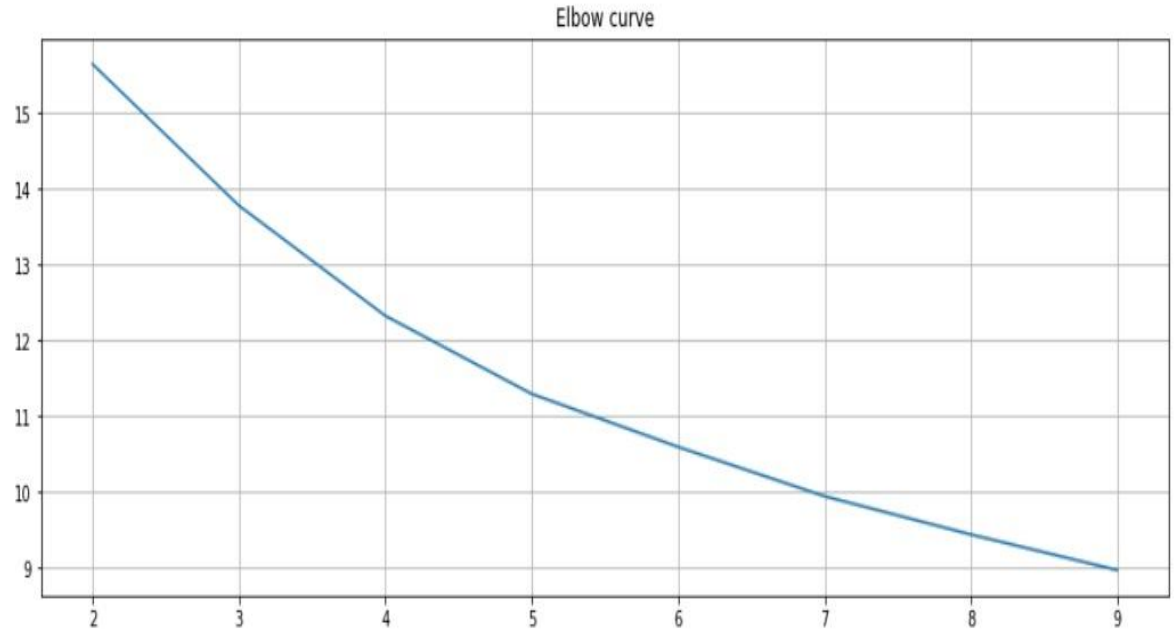

Figure 2. Inertia for different cluster numbers

4. Results

The resulting clusters (Figure 3) are distinct enough without one being significantly larger than the rest. The five main categories for each cluster are ‘Music Venues’, ‘Museums’, ‘Art Galleries’, ‘Movie Theatres’ and ‘Theatres’.

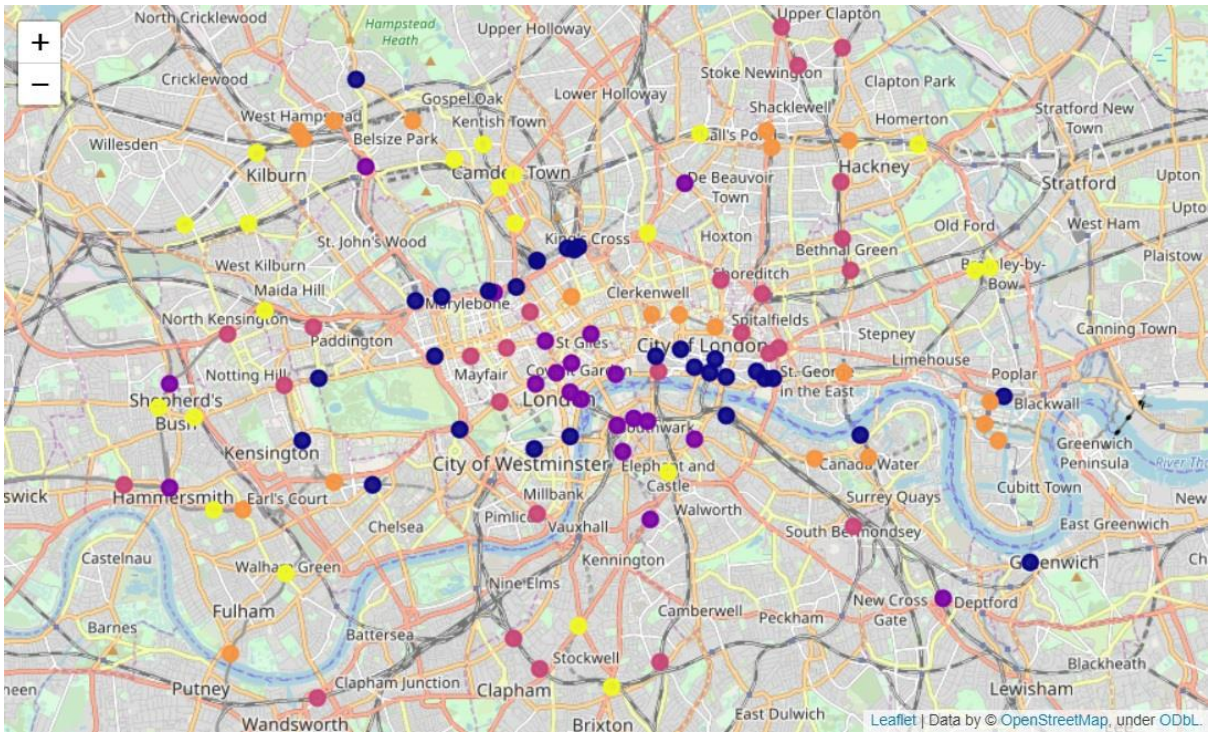

Figure 3. Stations in London clustered

5. Discussion

We notice that in the centre of London the main types of venues are Museums and Theatres which makes sense. Stations that are close together geographically also share the same cluster. This is because they share a similar distance from the venues. Although this is expected, an improvement could be made by removing stations that are too close together. However, since stations that are close to each other are sometimes used by different lines it would work against the goal of being an easy way for tourists to plan their trips. Considering that tourists will probably not care about visiting every single venue that falls under their preferences, consulting the map can help them choose the location of their hotel to be close to a cluster of stations. This will make it significantly easier for them to visit their favourite venues through the durations of their trip.

6. Conclusion

After using the Foursquare API and publicly available data, implementing the K-Means algorithm and visualising the results, we can conclude that Arts & Entertainment venues are an effective way of clustering the stations in London. The resulting clusters are distinct and constitute useful information for tourists looking to plan their trip. Further analysis and improvements could be made, but this project is a good starting point and fulfils its original goal.