The Battle of Neighbourhoods

Clustering London Venues and Analysing Correlations to Property Prices

1. Introduction

London is one of the largest metropolises in the world. With a population of over 8 million people living in a space of 1572km^2, this bustling city is full of venues ranging from restaurants to night clubs. As a result, each neighbourhood has grown a distinct character which often drives the types of venues that dominate the area. For instance, we would expect to find more night clubs and bars in vibrant Eastern locations, such as Shoreditch, when compared to the family friendly areas of the West, like Hammersmith and Fulham.

Venues often reflect the lifestyle of local residents and can dictate the status of tenants that Landlords might be exposed to. This is crucial for property owners looking to live in areas which complement their personality or landlords targeting specific demographics (e.g. professionals in Canary Wharf). In turn, this drives the business decisions of commercial properties who are trying to provide services in line with the needs of the local neighbourhood. Therefore, the distribution of venue types across different regions of London is particularly important for citizens looking to live or invest in London.

As a lifetime resident, analysing the distribution of venues in London's neighbourhoods is particularly interesting since it provides insight into the distinct character of each area. The project also intends to explore potential correlations between venue types and property prices. This information can be used to establish which venue distributions are associated with higher property prices as well as providing a basis for more accurate property price predictions.

2. Data

In order to complete the analysis, three main steps and datasets will be required.

First, London needs to be split into relevant regions in order to compare the distribution of venues and their correlation to property prices. The city is already conveniently sectioned into postcodes with clearly defined boundaries and addresses. A map and csv list of postcodes, including their comprised neighbourhoods, can be found at: https://www.doogal.co.uk/london_postcodes.php. The Data will be turned into a pandas Dataframe and then cleaned to display boroughs using the first 3 characters of each postcode, the longitude and latitude values of their centres, and the neighbourhoods contained within each borough.

Second, a list of venues in range of each borough will be required for clustering. This can be done using the **Foursquare API** and defining an appropriate radius within which to query. Once a list of venues has been obtained, exploratory analysis can be performed to see what kind of venues are most abundant in different regions.

Third, property prices will be needed in order to explore potential correlations. Data on average property prices divided by postcodes is publicly published by the UK government and readily available from many sources, for example zoopla at the following link: https://www.zoopla.co.uk/house-prices/. This information can be combined with the previously clustered groups to perform an analysis and see whether (i) the property prices of the postcodes grouped by cluster are comparable to each other and (ii) whether there is a correlation between the types of venues and property prices.

3. Methodology

3.1. Separating London into postcode districts

Data downloaded from https://www.doogal.co.uk/PostcodeDistricts.php was used to prepare a Dataframe of London's postcode districts. The original data has 3,112 rows and includes all postcode districts in the whole of the UK. This means that the data for London will need to be extracted and cleaned. However, the dataset does not include a label for postcodes in London so a different approach is required.

It can been seen from **figure 1** that London can be split using the first few letters and numbers of the postcode system in the UK. Furthermore, all postcodes in London start with N, E, SE, SW or W followed by a number. This information can be used to separate London's postcode districts from the rest of the UK.

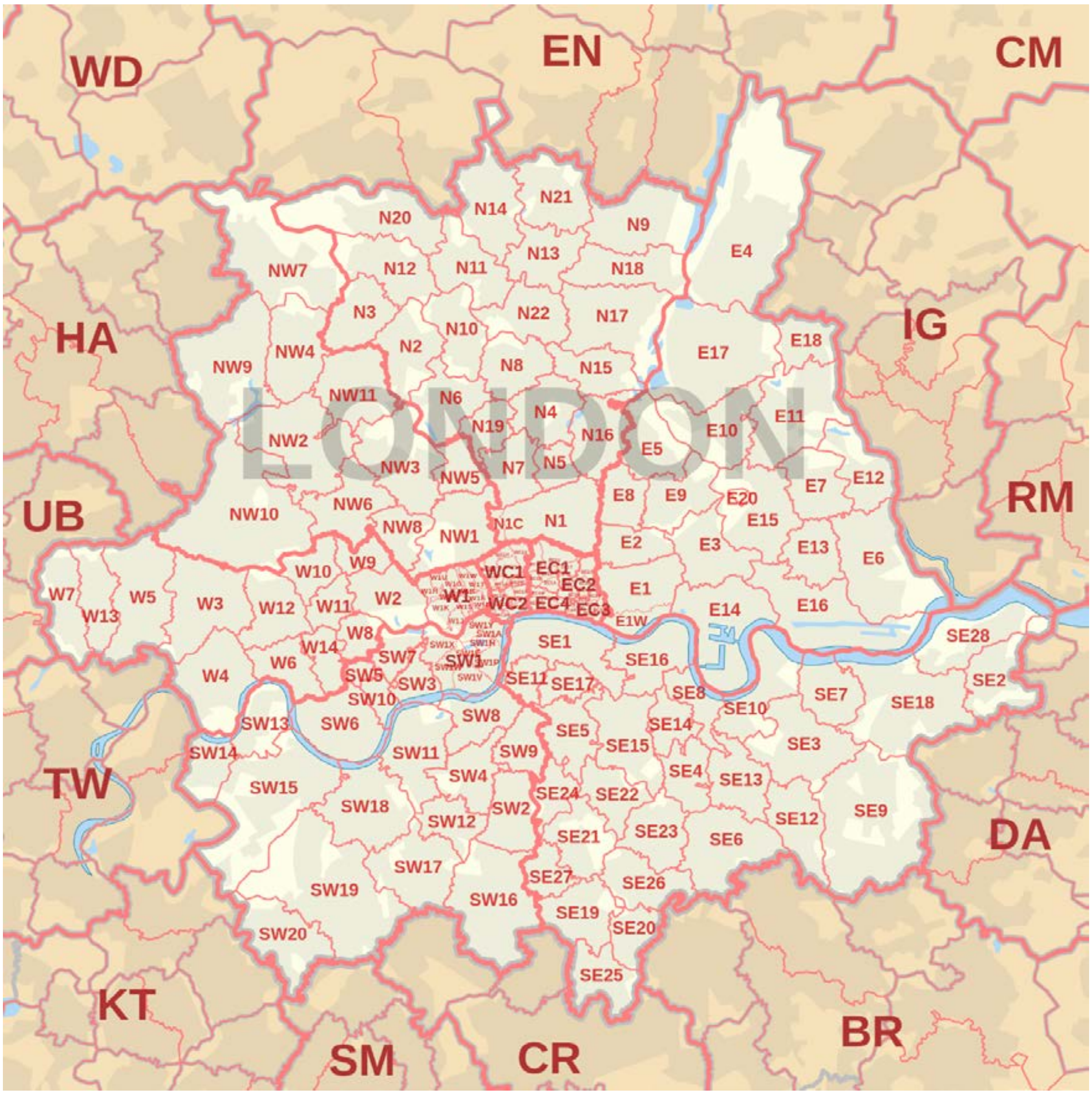


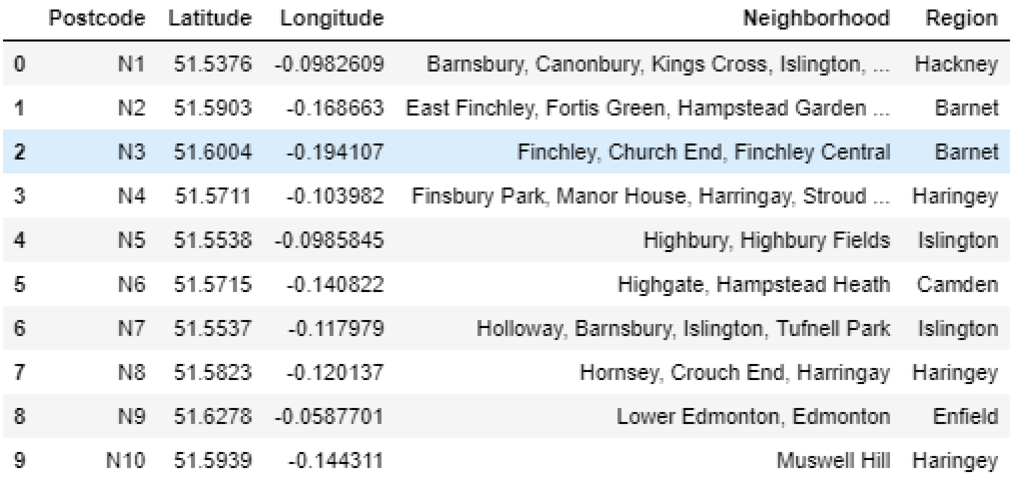
Figure 1: Map of London's postcode districts

The following steps were taken to extract London's postcodes and clean the data:

(i) Removed irrelevant columns.
(ii) Extracted postcodes in the UK that start with N, E, SE, SW or W.
(iii) Deleted extracted postcodes that were not in London e.g. 'WN' for Wigan.
(iv) Deleted non-geographic postcodes e.g. E20.

In our resulting data set, postcodes in Central London are further subdivided into subsections. For example W1 is subdivided into 14 sub-districts (W1A, W1B, W1C, etc...). This results in geographical divisions that are much smaller than intended. Since the Foursquare API (that will be used to obtain venue data) queries venues within a defined radius, these small subdivisions could be problematic when defining the size of the radii. Therefore, accuracy can be improved by reversing the subdivisions into their larger Central London districts i.e. by grouping the subdivisions into WC1, WC2, WC3, WC4, W1, SW1, EC1 and EC2. For the same reason, the small subdivided districts of WC99, N1C and E1W were also dropped.

The resulting Dataframe includes 120 postcode districts in London, with their comprised Neighbourhoods as well as the Longitude and Latitude co-ordinates of their centres, as shown in **Figure 2** demonstrating the first 10 rows.



	Postcode	Latitude	Longitude	Neighborhood	Region
0	N1	51.5376	-0.0982609	Barnsbury, Canonbury, Kings Cross, Islington, ...	Hackney
1	N2	51.5903	-0.168663	East Finchley, Fortis Green, Hampstead Garden ...	Barnet
2	N3	51.6004	-0.194107	Finchley, Church End, Finchley Central	Barnet
3	N4	51.5711	-0.103982	Finsbury Park, Manor House, Harringay, Stroud ...	Haringey
4	N5	51.5538	-0.0985845	Highbury, Highbury Fields	Islington
5	N6	51.5715	-0.140822	Highgate, Hampstead Heath	Camden
6	N7	51.5537	-0.117979	Holloway, Barnsbury, Islington, Tufnell Park	Islington
7	N8	51.5823	-0.120137	Hornsey, Crouch End, Harringay	Haringey
8	N9	51.6278	-0.0587701	Lower Edmonton, Edmonton	Enfield
9	N10	51.5939	-0.144311	Muswell Hill	Haringey

Figure 2: First 10 rows of London's districts Dataframe

3.2. Obtaining venue data

The foursquare API was used to explore the districts and find their venues. The queries were designed with a limit of 200 within a radius of 1km from the centre of each postcode district. As a result, 8048 venues were returned. One hot encoding and normalisation was then used to produce a Dataframe showing the top 10 venues present in each district. Since many of the districts share common venues (e.g. Pubs and Cafes), the boroughs can now be clustered based on venue categories.

3.3. Clustering districts based on venue categories

Since the dataset is relatively small, the most common type of unsupervised machine learning algorithm can be used; K-means clustering. The elbow method was used to help determine the optimum number of k-clusters, as shown in **figure 3**. The result was ambiguous, but trial and error showed that k = 3 is the optimum number of clusters.

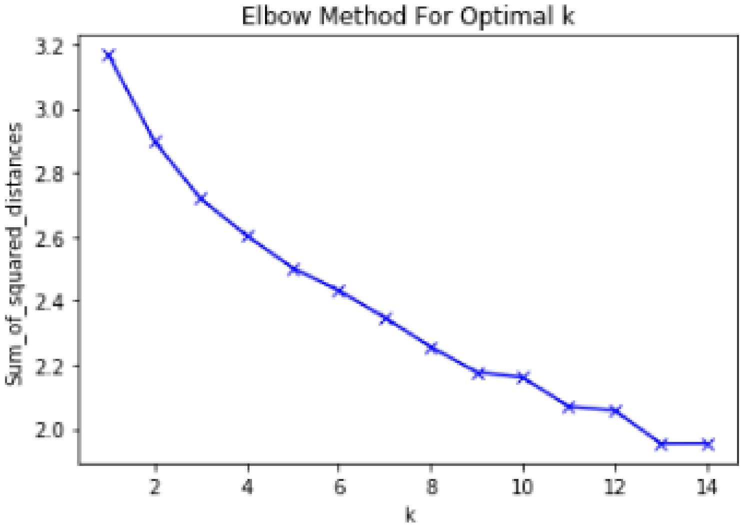


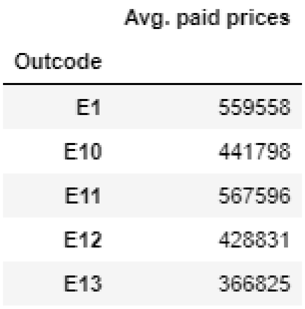
Figure 3: Graph of elbow method to find optimal K

After running the K-means algorithm, the information was merged with the top 10 venues list previously sorted to produce a table with cluster labels. Using the Folium package, a map of London can now be visualised with markers displaying the postcodes, neighbourhoods and cluster labels (see **figure 5** in the results section below).

Finally, venue information for each cluster was extracted from the merged table to produce 3 separate Dataframes. This allows for analysis to be performed on the venue distribution of individual clusters. For example, by summing the venue categories of a given cluster, we can see which venue types are most abundant and compare their distributions (see **figure 6** in the results section below).

3.4. Average property prices by postcode

The data was extracted from https://www.zoopla.co.uk/house-prices/, which provides average property prices sorted by postcode. The information was compiled in Excel, uploaded to a GitHub repository and read into a pandas Dataframe, the first 5 rows of which are shown in **figure 4**.



	Avg. paid prices
Outcode	
E1	559558
E10	441798
E11	567596
E12	428831
E13	366825

Figure 4: First five rows of Average Property prices by Postcode

One of the main objectives of this project is to see how the distribution of venues affects property prices. In order to do this, we need to determine the value of each cluster. This was done using an algorithm which extracts the prices in figure 4 for postcodes contained in a given cluster. The result is 3 separate Dataframes displaying average property prices for each cluster. Analysis can now be performed to compare the mean, median, minimum and maximum values of each cluster, as discussed further in the results section below.

4. Results

Figure 5 shows the geographical centres of each district marked with their respective postcode, neighbourhoods and cluster label. The clusters have been colour coded to visualize their spatial distribution. The resulting radial pattern suggests that there is a correlation between a given cluster and their distance from the centre.

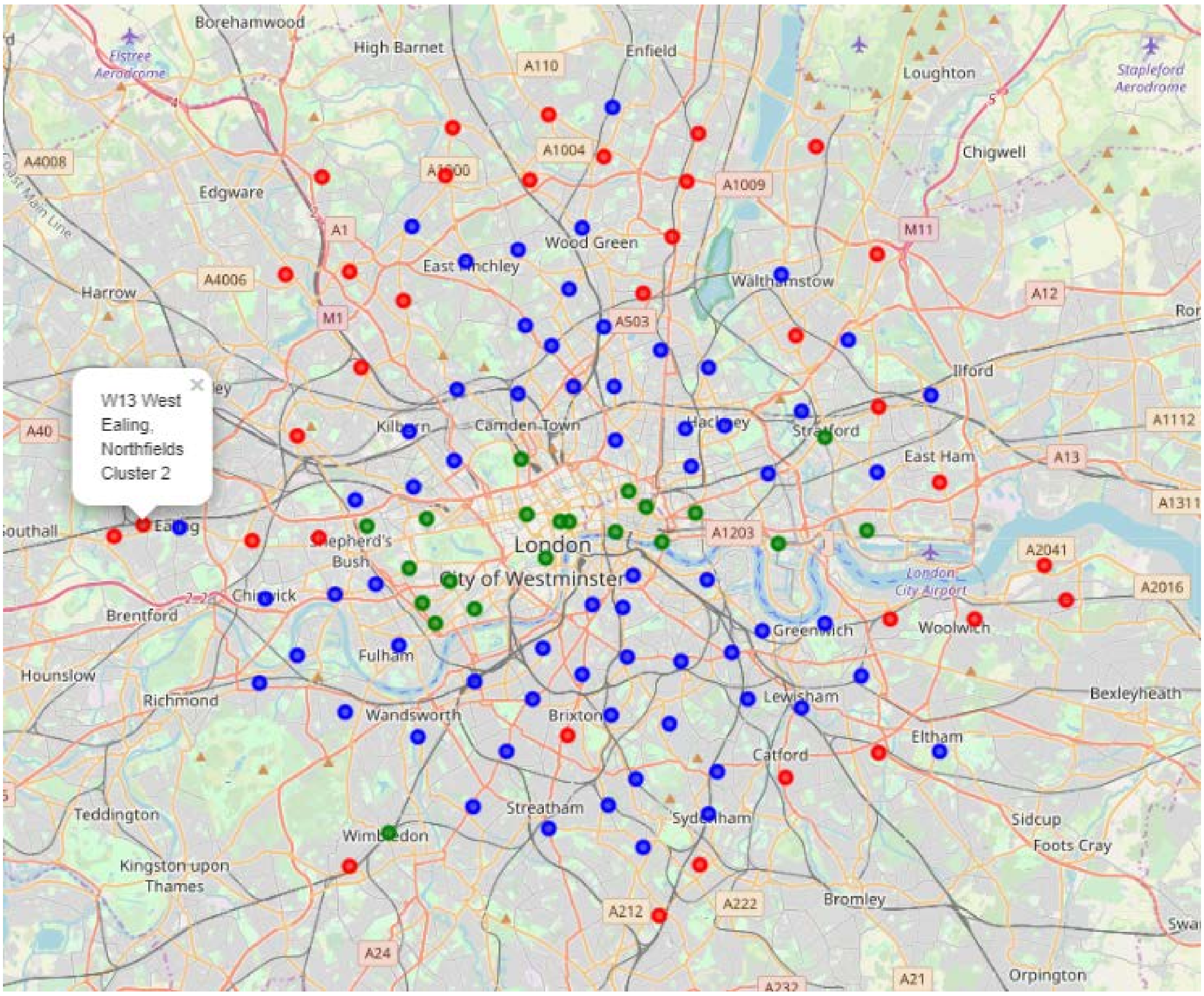


Figure 5: Folium map of London Clusters

The distributions of the top 20 venue categories for each of the three clusters can been seen in **figures 6, 7 and 8**. When considered with the above, this implies that the abundance of venue categories changes based on a given postcode’s distance from the centre. For instance, fast food restaurants are most common in cluster 2 and postcodes in cluster 2 are usually the furthest out. Therefore, the data suggests that fast food restaurants are more likely the further you are from the centre of London.

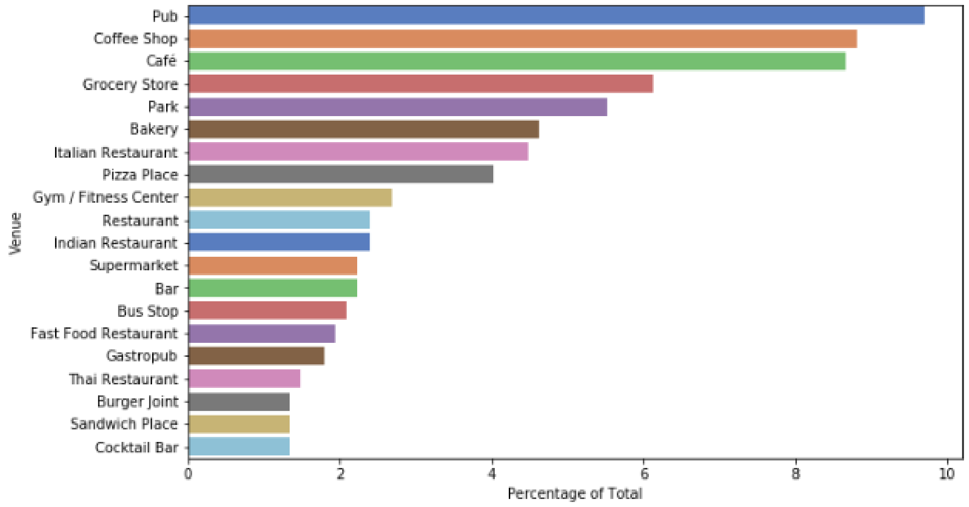


Figure 6: Venue distribution of cluster 0

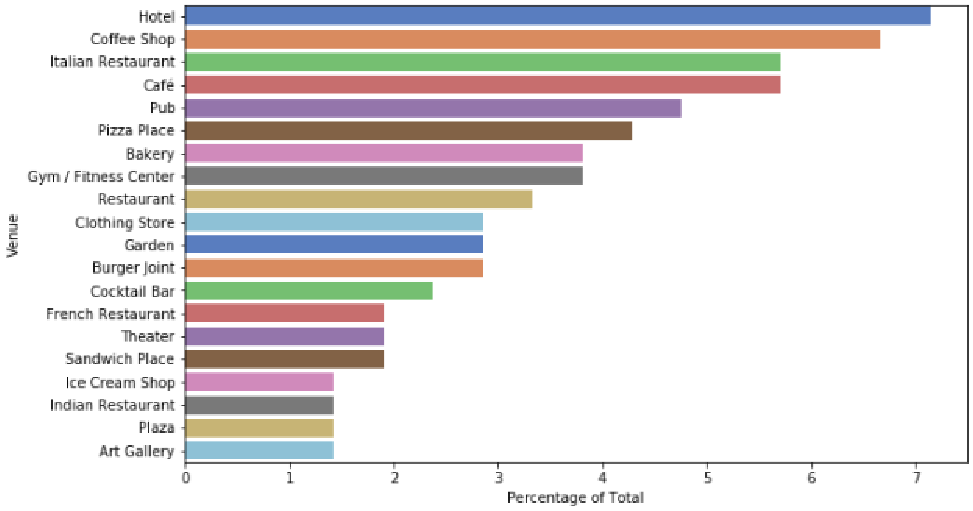


Figure 7: Venue distribution of cluster 1

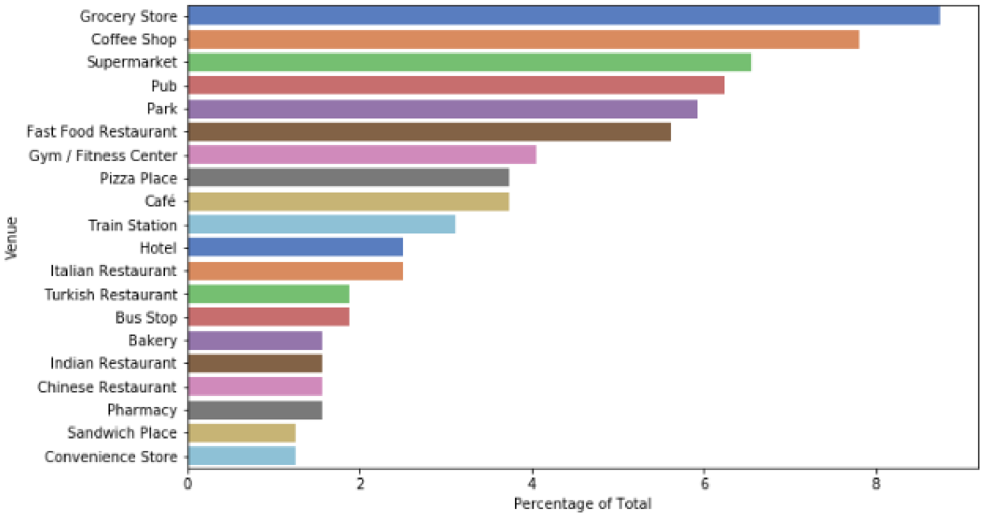


Figure 8: Venue distribution of cluster 2

In order to visualize the effect this distribution has on property prices, **figure 9** compares box plots of each cluster based on the average value of their districts. We would expect to see that the overall value decreases the further a postcode is from the centre of London. In accordance with figures 5 to 8 above, it is clear that as expected, cluster 0 has the highest median, mean and overall value; followed by cluster 1 then cluster 2.

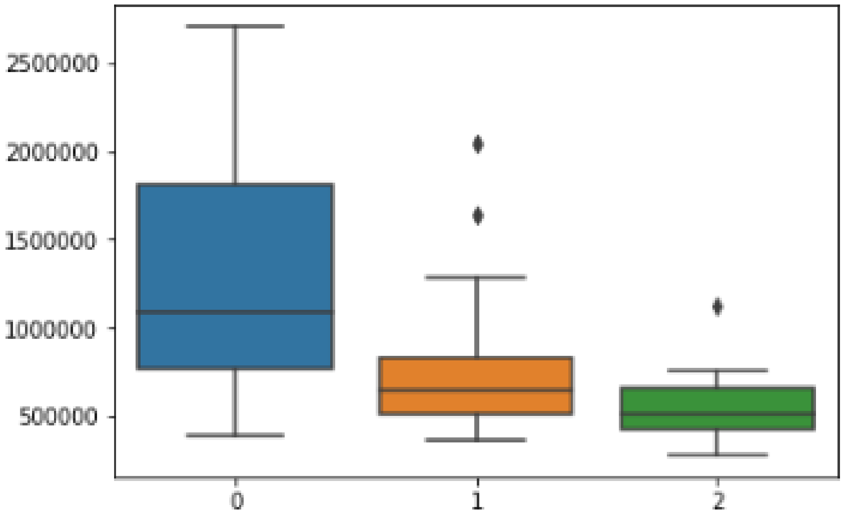


Figure 9: Average property values of each cluster

5. Discussion

This project has established that:

(i) Districts of lower or higher value have different distributions of venues;

(ii) There is a positive correlation between a district's distance from the centre of London and the likelihood of particular venue categories;

(iii) There is a positive correlation between distance from the centre of London and property value; and

(iv) There is a correlation between venue distribution and property value.

There is insufficient data to determine causal links between the three factors. For instance, we cannot conclude whether district value is lower because of the distribution of venues or if venue distribution is a result of the value of the area. Furthermore, venue distribution and cluster value are both affected by distance from the centre meaning that distance could be the only defining factor affecting the change in both variables. A further study, which factors in the distance, would be required in order to establish a definite correlation between venues and value.

The list of venues was generated by defining an arbitrary distance from the centre of each postcode. However, London's boroughs all have different sizes, and some districts are much smaller than others. This means that the radii overlap with some zones and leave gaps with others resulting in some venues being counted multiple times while others may not be counted at all. In order to improve the accuracy of the distribution, a list of venues queried within the clearly defined postcode boundaries should be used. This would improve the precision of the clustering algorithm coupled with the price correlations since the property data is also confined to the boundaries of each postcode.

Finally, the data can also be used for further exploration to establish correlations outside of the scope of this project. For example, one could predict the likelihood of finding a given venue based on distance from the centre of London, or determine the most optimal place to open a pub based on the value of the area's properties, or use the data as a factor in predicting the value of property prices in London.

6. Conclusion

This project shows that there is a correlation between the value of a given postcode and the distribution of its venue types. There is also a positive correlation between both factors and the postcode's distance from the centre of London. Causal links cannot be established, but further analysis could be done by considering these distances and factoring out the difference. The data suggests that the composition of venues in cluster 0 is more desirable in terms of property value than the distribution of clusters 1 and 2.

References

1. https://www.doogal.co.uk/london_postcodes.php
2. https://developer.foursquare.com/
3. https://www.zoopla.co.uk/house-prices/
4. https://en.wikipedia.org/wiki/