The Battle of Neighbourhoods

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.

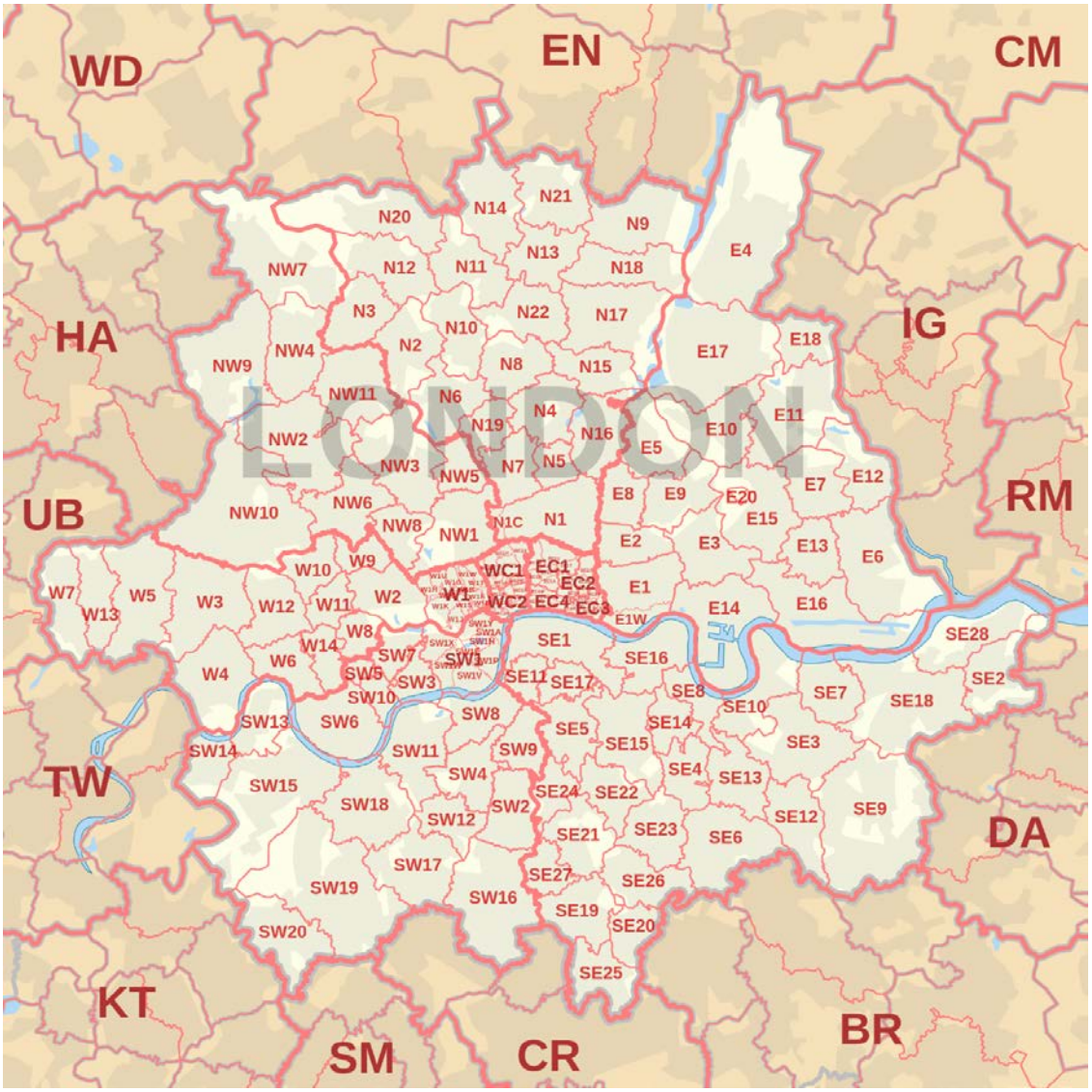
As a lifetime resident, analysing the distribution of venues in London's neighbourhoods is particularly interesting since it provides insight into the distinct character of each area. The project also intends to explore potential correlations between venue types and property prices. This information 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.

Data

1. London will be split into postcode regions in order to compare the distribution of venues and their correlation to property prices.
2. Second, a list of venues in range of each borough will be required for clustering. This will be done using the **Foursquare API**.
3. 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/.

Methodology

The dataset was cleaned and re-organised to split London into 120 postcode regions. These regions can be seen in the following map:



Map of London's postcode districts

The resulting Dataframe includes, Neighbourhoods as well as the Longitude and Latitude co-ordinates of their centres.

	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

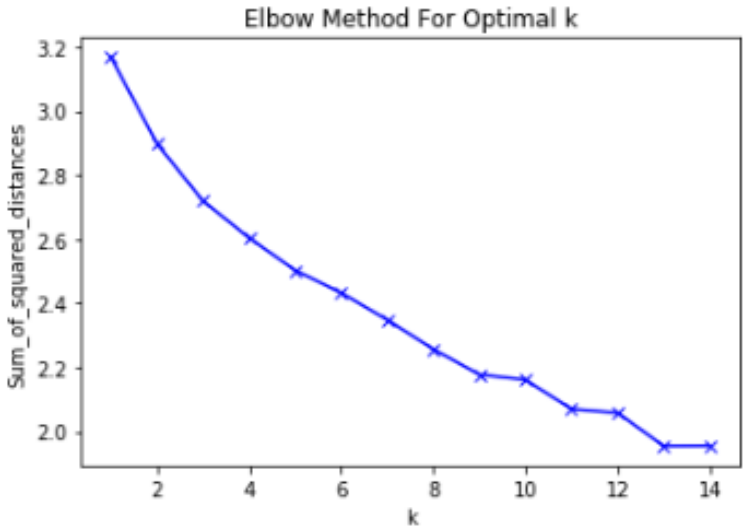
First 10 rows of London's districts Dataframe

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.

Clustering districts based on venue categories

K-means clustering was used to analyse the data. The elbow method helped determine the optimum number of k-clusters. Although the result was ambiguous, trial and error showed that k = 3 is the optimum number of clusters.



Graph of elbow method to find optimal K

Average property prices by postcode

The information was compiled in Excel, uploaded to a GitHub repository and read into a pandas Dataframe.

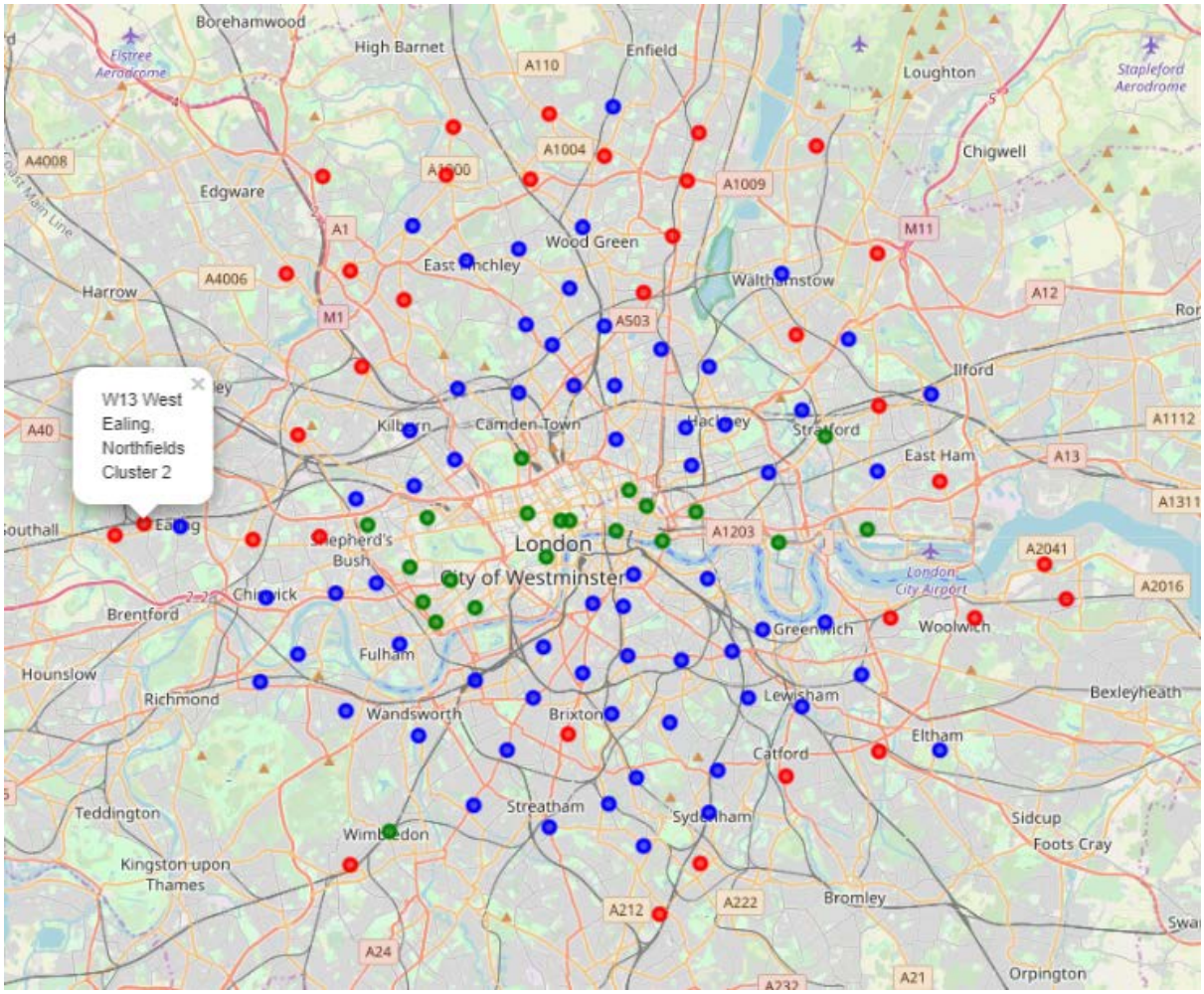
Outcode	Avg. paid prices
E1	559558
E10	441798
E11	567596
E12	428831
E13	366825

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 for postcodes contained in a given cluster.

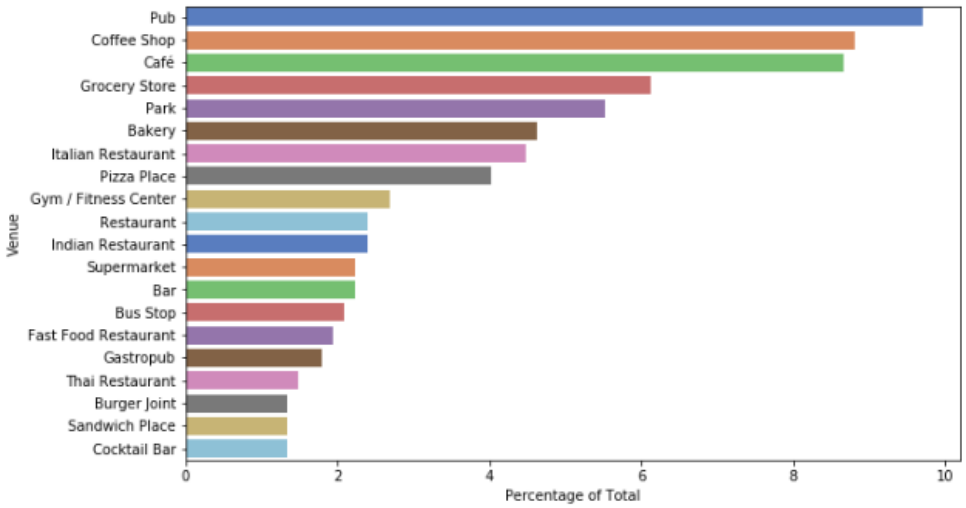
Results

A folium map was generates to show 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.

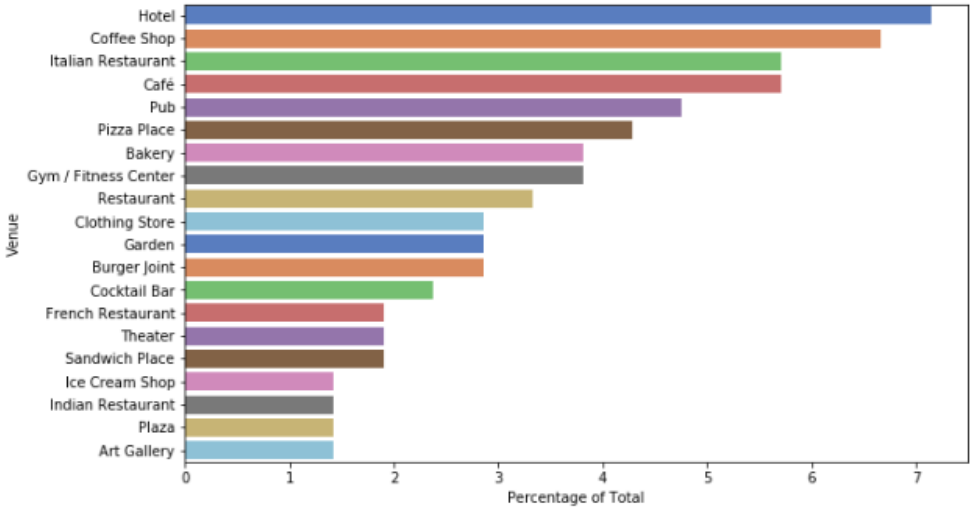


Folium map of London Clusters

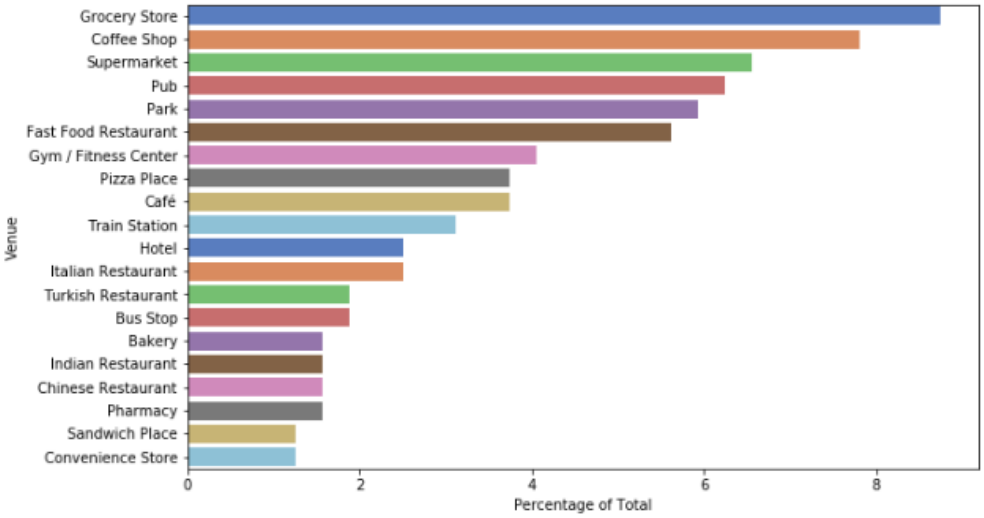
Bar charts were used to visualize the distribution of the top 20 venue categories for each of the three clusters. When considered with the above, it can be seen that the abundance of venue categories changes based on a given postcode's distance from the centre. For instance, fast food restaurants are 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.



Venue distribution of cluster 0

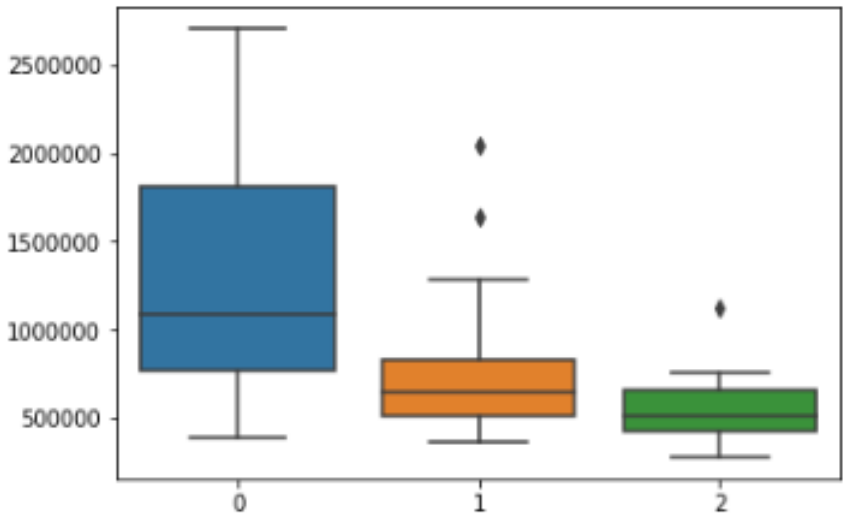


Venue distribution of cluster 1



Venue distribution of cluster 2

In order to understand the effect this distribution has on property prices, box plots of each cluster were produced 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. It is clear that as expected, cluster 0 has the highest median, mean and overall value; followed by cluster 1 then cluster 2.



Average property values of each cluster

Discussion

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.