Analysing FourSquare & Housing Prices Data in London

Nico Le Le Man

Cousera IBM Professional Data Science Certification

10th May 2020

Table of Contents

1. Introduction

1.1. Background

It goes without saying that the coronavirus (COVID-19) has had, is currently and will continue to have a significant impact on businesses and the economy worldwide. This is evident with stock market and oil prices crash, record breaking number of people filing for unemployment and major airlines on the brink of administration.

The Real Estate & Property market is no exception to the coronavirus impact, with the London property market coming to a halt back in March when the full lock down was announced to prevent the spread of the virus. Physical viewings were postponed, constructions were suspended and estate agents & mortgage lenders no longer able to value properties in person.

As a result, Zoopla has predicted that completed sales in the UK will be 50% lower in 2020 than in 2019 and Knight Frank has also predicted that the number of sales in Greater London will fall by 35%. However, despite the bleak outlook for property and housing prices this year, a large number of firms & their analysts believe that the housing market could make a very strong recovery by 2021, with an estimated range of 3% - 6%.

1.2. Business Problem

The best decisions are often backed up by insight and data, by utilising Machine Learning we can effectively and efficiently generate those insights in order to provide potential homebuyers and investors the best decision-making support as possible. This brings us to our business problem: How can we generate insight so home-buyers and investors can make well informed choices when purchasing properties in London, especially in this uncertain economic situation?

In order to solve this business problem, we will cluster the London areas based on the average sales price, local venues, and amenities, i.e. schools, supermarkets, coffee shops. We will then compare these clusters with the average property prices and rental prices for each borough, and also calculate the rental yield for each cluster for investors who are buying to let. This will provide valuable information on whether a property is a viable choice for homebuyers & investors.

2. Data Acquisition

2.1. Data sources

The Price Paid Data (property sales data) in London will be sourced from HM Land Registry, where the data is based on the raw data released each month. The dataset will include the following columns: Transaction unique identifier, Price, Date of Transfer, Postcode, Property Type, Old/New, Duration, PAON (Primary Addressable Object Name), SAON (Secondary Addressable Object Name), Street, Locality, Town/City, District, County and PPD Category Type.

The FourSquare API will be used to access and explore venues and amenities based on the Latitude and Longitude collected using the GeoCoder library, which will then be read into a dataframe for data wrangling and cleaning. This dataframe will be merged with the Price Paid Data from HM Land Registry and processed to be suitable for fitting the machine learning model.

The list of boroughs in London will be scrapped from the Wikipedia page and the average property and rental prices per borough will be scraped from Foxtons (A UK estate agency). This data will then be used to compare with the average property prices in a neighbourhood and calculate the rental yield. All of which will help homebuyers to decide whether they are over/underpaying for a property and investors to see how the property valuation & rental yield compare with the market average.

The clusters generated by the unsupervised learning model will be visualised using Plotly.

2.2. Data collection

Price Paid Data 2019:

The PPD (Price Paid Data) CSV file downloaded from the HM Land Registry website did not include headers, so I had to manually add those in after reading the CSV file into a dataframe. The resulting dataframe had over 960,000 rows and 16 columns.

As the PPD data tracks property sales in England and Wales, the rows that correspond to property sales in London had to be extracted. Most of the columns such as TUID, PAON, SAON, Locality and PPD_Cat_Type were not needed in my case therefore they were also dropped in the feature selection stage.

Property & rental prices:

I decided to collect the property and rental prices from Foxton, as they had a page that displayed the average property and rental prices for an area based on the postcode prefix. First a list of London postcode districts was scrapped from www.doogal.co.uk/london_postcodes.php, using those postcodes I scraped the corresponding property and rental prices from www.foxtons.co.uk/living-in, and finally the data was written into a CSV file.

Some of the postcode districts did not have data available, as a result those rows were dropped in the dataframe.

2.3. Features selection

As mentioned above, columns (or features) that were not needed for this project were dropped and they were:

- TUID (Transaction Unique Identifier)
- Duration
- PAON (Primary Addressable Object Name)
- SAON (Secondary Addressable Object Name)
- Locality
- PPD_Cat_Type
- Record_Status

I then extracted rows that represented property sales in London and also where the price of transaction was less than £2,000,000. I also dropped rows that contain 'NaN' which returned a dataframe consisting of 56318 rows and 10 columns. As I want to compare the property sales price with the data collected from the Foxton website, I decided to split the postcode and append the prefix to a new column in the dataframe.

Upon sorting the dataframe by the street column and inspecting it, there were a large number of rows where property sales had happened on the same street. This could skew the clusters as we would end up getting the same FourSquare venue data for each of these 'duplicate' rows. In order to overcome this, I grouped the dataframe by street, district and the postcode prefix, where the mean property prices were calculated for the 'duplicate' rows. The reason 3 columns were used in the group by process was because some street names are not unique and are used in different districts, therefore grouping by those 3 columns ensured that only the correct streets were grouped. An example of this would be 'Abbey Gardens', where City of Westminster, Hammersmith and Fulham and Southwark all have a street with that identical name:

	street	district	postcode_prefix	avg_price
0	ABBESS CLOSE	LAMBETH	SW2	296000.0
1	ABBEVILLE ROAD	LAMBETH	SW4	613870.0
2	ABBEY GARDENS	CITY OF WESTMINSTER	NW8	588750.0
3	ABBEY GARDENS	HAMMERSMITH AND FULHAM	W6	470750.0
4	ABBEY GARDENS	SOUTHWARK	SE16	330500.0

Figure 1. Preview of the dataframe containing the average property price for a street

The grouped dataframe consists of 15555 rows and 4 features, where each row is for a unique street in London. Using this dataframe and the GeoCoder library, I collected the latitude and longitude for each row and wrote the data to a CSV file. This process was done using a Python script as the computational time require was fairly longer at approximately 4 hours. Around 700 rows were dropped after reading the new CSV file into a dataframe as the GeoCoder did not return a latitude and longitude for them, therefore we have 14323 rows and 6 features.

	street	district	postcode_prefix	avg_price	latitude	longitude
0	ABBESS CLOSE	LAMBETH	SW2	296000	51.442879	-0.108249
1	ABBEVILLE ROAD	LAMBETH	SW4	613870	51.453304	-0.140988
2	ABBEY GARDENS	CITY OF WESTMINSTER	NW8	588750	51.533905	-0.179989
3	ABBEY GARDENS	HAMMERSMITH AND FULHAM	W6	470750	51.484844	-0.213365
4	ABBEY GARDENS	SOUTHWARK	SE16	330500	51.491653	-0.066099
...	...	...	...	...	...	...
14344	YUNUS KHAN CLOSE	WALTHAM FOREST	E17	275000	51.578888	-0.019688
14345	ZANGWILL ROAD	GREENWICH	SE3	406000	51.472534	0.042171
14346	ZEALAND ROAD	TOWER HAMLETS	E3	790000	51.531441	-0.037656
14347	ZENITH CLOSE	BARNET	NW9	375000	51.592243	-0.255944
14348	ZOFFANY STREET	ISLINGTON	N19	828000	51.566402	-0.127573

Figure 2. Overview of the dataframe containing the average property price for a street and its corresponding postcode

If I were to get the venue data using the FourSquare API using the dataframe above, the computational required will be significant and not viable. In addition, an application can only make a maximum of 5000 requests per hour to the venues endpoint. In order to reduce the dataset without introducing any data bias, I sampled 20% of the full dataframe which resulted in 2865 rows that will be much more manageable when carrying out EDA.

FourSquare venues data:

The FourSquare data was collected in 3.2. Exploring venues. The process was done using a custom function that looped over each row in the sample dataframe, sending a GET request which returned all the venues within a 300-meter radius. The venue data returned consists of the venue name, venue latitude, venue longitude and venue category. Finally, the new data is appended to the end of the sample dataframe and saved as a .pkl file. This will eliminate the need to re-run the FourSquare venue data collection step thus saving time between runs.

3. Exploratory Data Analysis (EDA)

3.1. Price paid data

The PPD sample data is visualised on a Plotly map. As excepted, neighbourhoods such as Kensington, Knightsbridge, Chelsea, and Belgravia have the highest average property prices. It is also clear that properties on the higher end of the price range are mostly located on the west side of Central London, in comparison to the east side where they are far fewer properties that exceed £1,000,000.

Niche neighbourhoods outside of Central London can also be seen in Figure 3, most notably Hampstead Garden Suburb, Cottenham Park and Dulwich where properties are often valued at £2 million and up to £15 million.

Out of the 2865 streets that were observed, 88% of the property paid prices were below £1,000,000 and 12% were above. From Table 1 it can be seen that the standard deviation for the average property prices between properties valued at above and below £1,000,000 only differs by approximately £55,000

For properties that are valued above £1,000,000, the average property prices remain under £1,500,000 within the upper quartile.

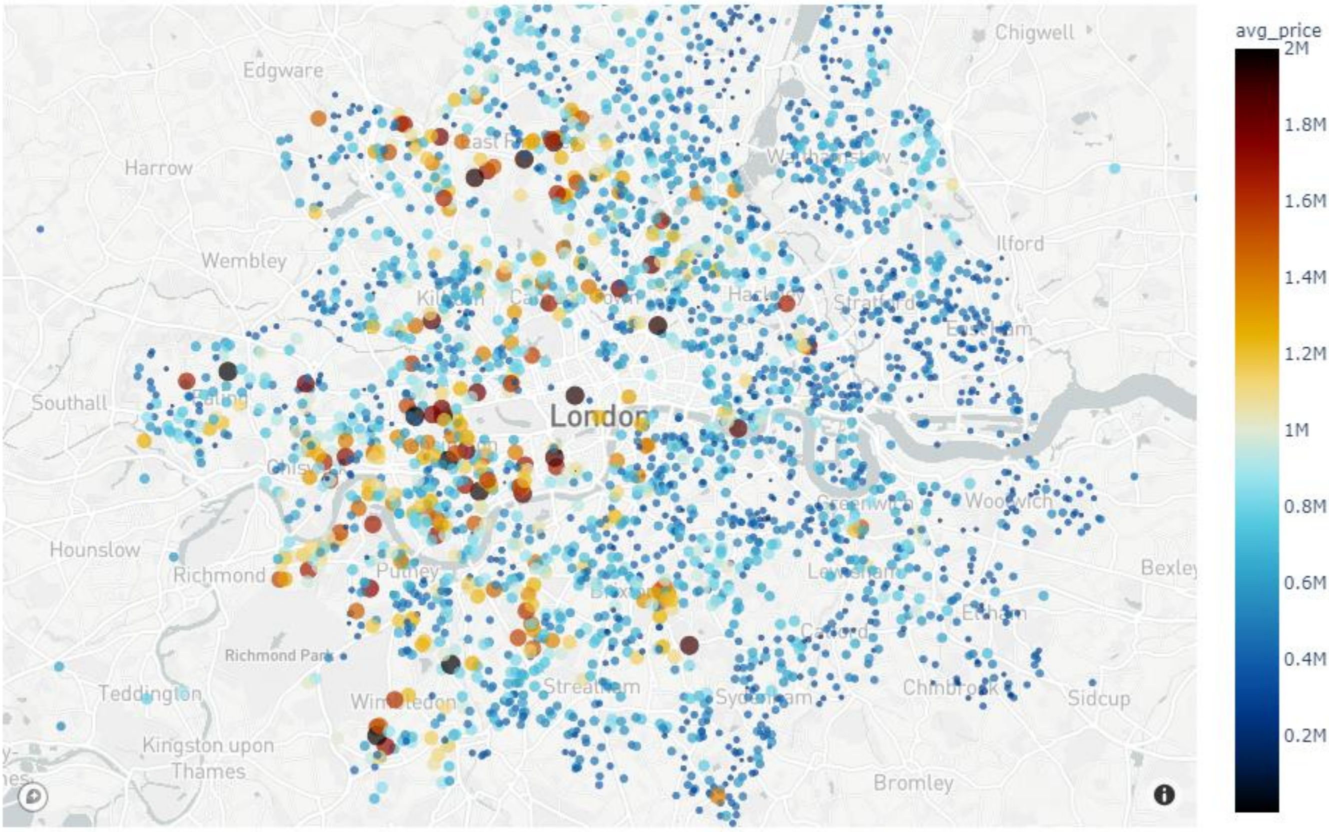

Figure 3. Map of property sales and their prices in London for 2019

	Average property price < £1,000,000	Average property price > £1,000,000
Count	2539	316
Mean	£509,862	£1,311,139
Standard deviation	£199,877	£254,028
25%	£365,175	£1,112,375
50%	£481,875	£1,235,714
75%	£644,646	£1,464,375

Table 1. Comparison of property prices above and below £1,000,000

Looking at the average property prices in each borough, the boroughs with neighbourhoods mentioned above (Kensington, Knightsbridge, Chelsea, and Belgravia) have the highest average property prices. Bromley, Hounslow, Kingston Upon Thames and Richmond Upon Thames all seem to have fairly high average property prices between £750k and £850k. However it is worth noting that all these boroughs have a low number of sales compared to City of Westminster or Kensington and Chelsea (Figure 5), therefore the value shown in Figure 4 might not reflect the true average property prices.

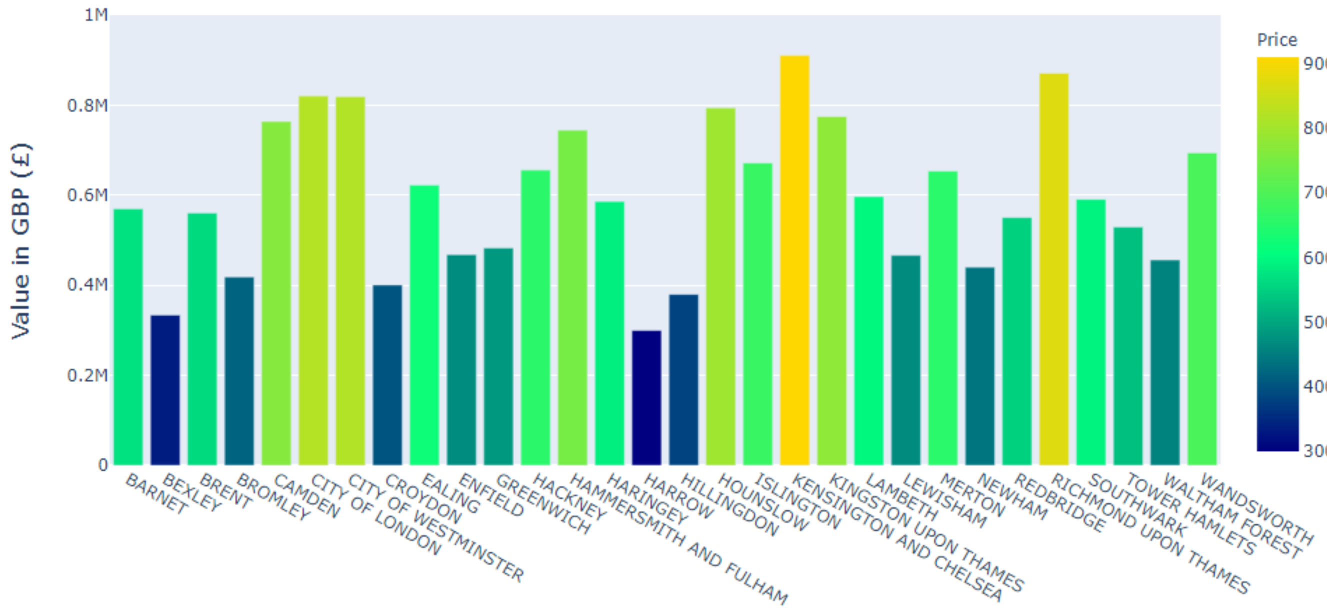

Figure 4. Average property prices in each borough for 2019

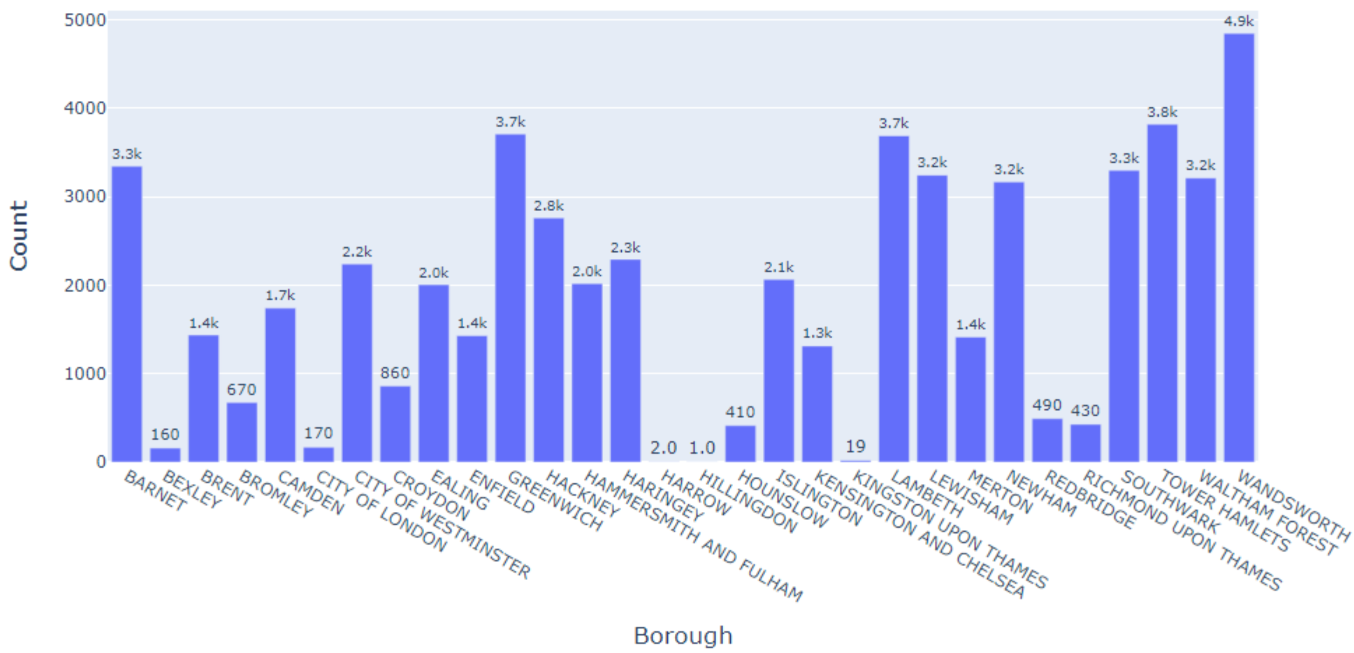

Figure 5. Number of property sales in each borough for 2019

3.2. Exploring venues

From the FourSquare venues data collected, it is clear that pubs are the most common venues which is unsurprising as there are over 3500 pubs in Central London. This is followed by cafes and coffee shops, with grocery stores and hotels being the 4th and 5th most common venues.

London is one of the most diverse, multicultural food scenes in the world, with over 70 Michelin star restaurants. Italian is among the more popular cuisines ranking 6th & 7th on the most common venues chart. This is followed by Indian, Chinese, Thai, Middle Eastern and Turkish restaurants in descending order.

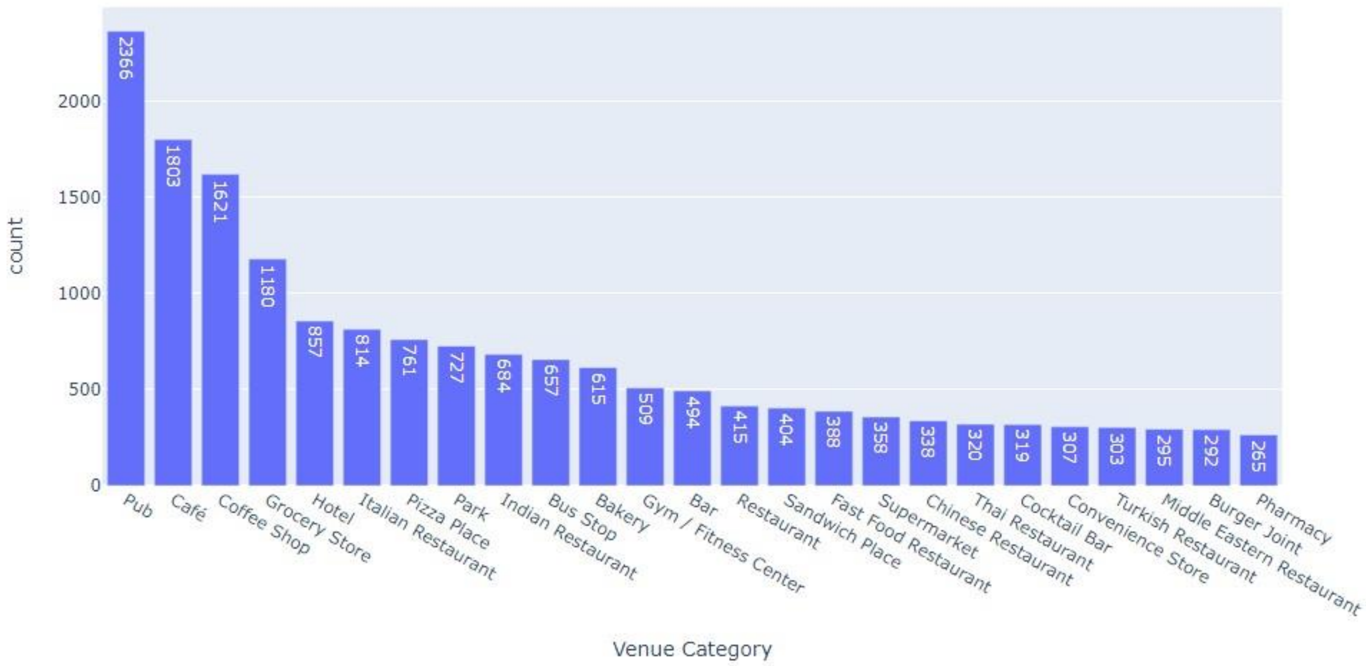

Figure 6. The top 25 most common venues across all boroughs

Figure 7 shows that for 12 out of the 28 boroughs, the most common venue in those boroughs were pubs.

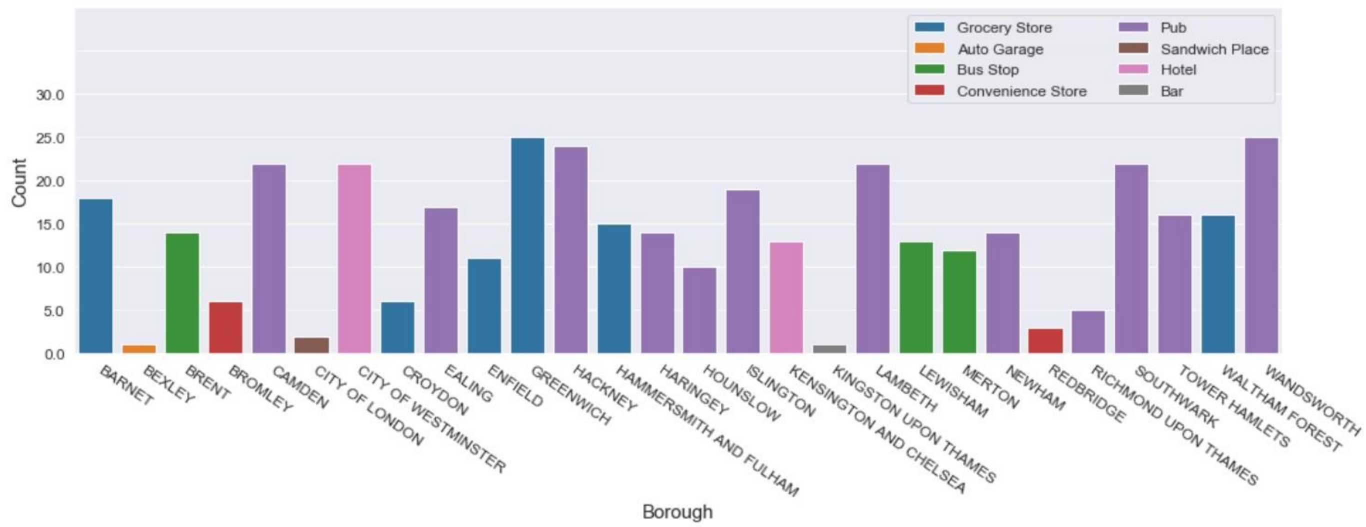

Figure 7. The most common venue in each borough

Let us take a look at the relationship between property prices and the impact of certain venues within the neighbourhood.

Comparing neighbourhoods where there is a pub within a 300-meter radius versus neighbourhoods where there is not, the average property price is slightly higher for the former by approximately £40,000. However, this is an overly broad and generalised comparison as other factors such as borough, types of property, age of property, number of bedrooms will also have an impact on the average property prices. The pattern is similar if we look at neighbourhoods with café & coffee shops vs without. Interestingly, the pattern is reversed when we look at grocery stores, neighbourhoods that don't have a grocery store nearby have a higher average property price.

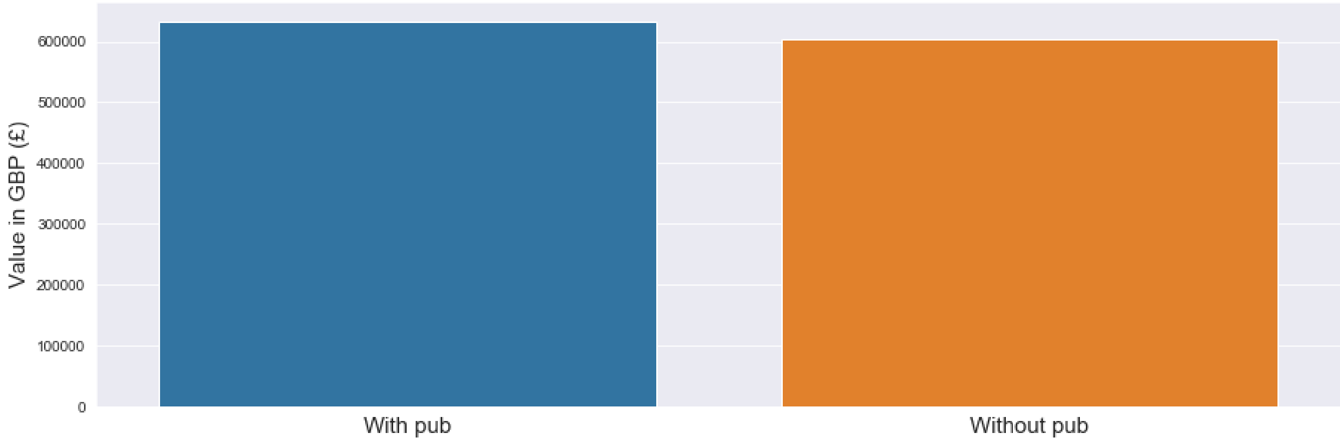

Figure 8. Average property price in neighbourhoods within 300-meters of a pub vs without

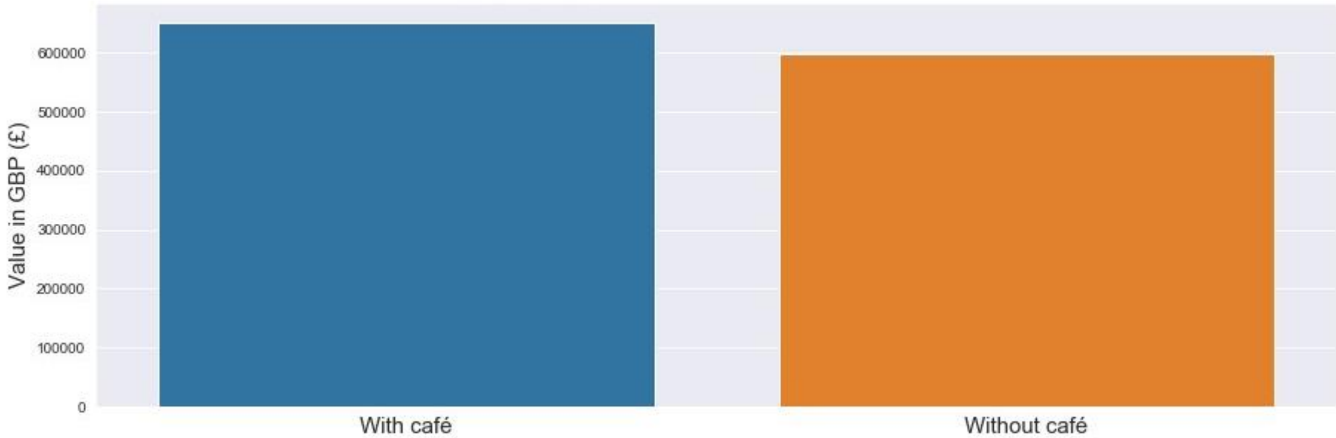

Figure 9. Average property price in neighbourhoods within 300-meters of a cafe vs without

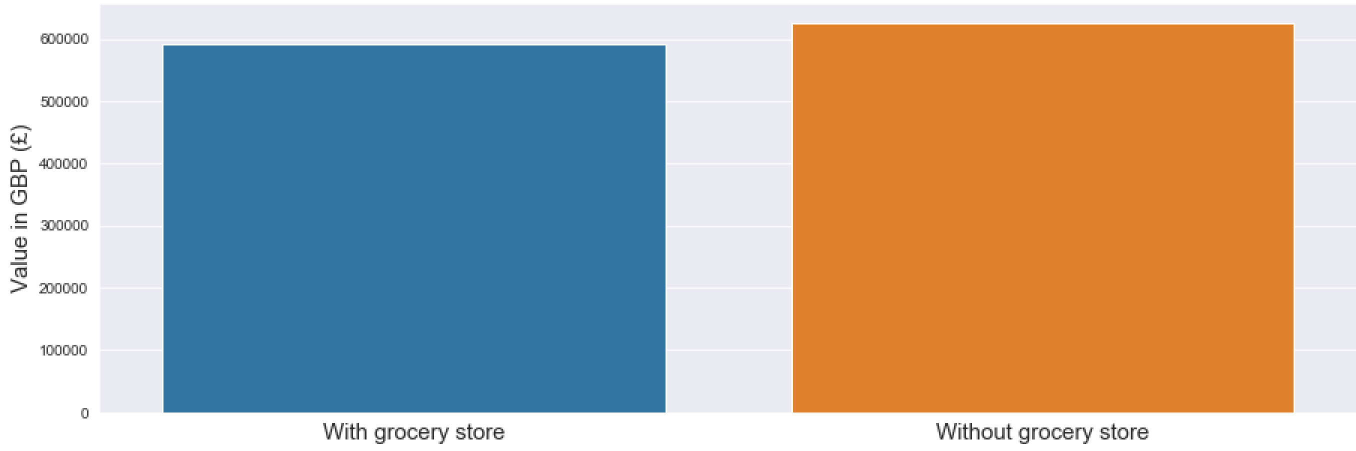

Figure 10. Average property price in neighbourhoods within 300-meters of a grocery store vs without

3.3. Analyse each neighbourhood

As mentioned in 2.3. Features selection, the data collected from the FourSquare API consists of Venue Name, Venue Category, Venue Latitude and Venue Longitude. The data is pre-processed by using one-hot encoding to convert the categorical variable, 'Venue Category', into a new binary value for each unique variable. The encoder output is cleaned, grouped and aggregated, resulting in a dataframe where each column is a feature and each row is a unique street/neighbourhood.

	street	district	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	...	Winery	Wings Joint	Women's Store	Xinjiang Restaurant	Yakitori Restaurant	Yoga Studio	Zoo Exhibit
0	ABBEY GARDENS	SOUTHWARK	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
1	ABBEY GROVE	GREENWICH	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2	ABBEY PARADE	MERTON	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.017241	0.0	0.0	0.0	0.0
3	ABBEY ROAD	BEXLEY	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
4	ABBEYFIELD ROAD	SOUTHWARK	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2626	WYTHFIELD ROAD	GREENWICH	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2627	YEATE STREET	ISLINGTON	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2628	YEOMAN STREET	LEWISHAM	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2629	YORK AVENUE	RICHMOND UPON THAMES	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2630	YORK WAY ESTATE	ISLINGTON	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

Figure 11. Dataframe of one-hot encoded venue category data for each neighbourhood

	street	district	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 GARDENS	SOUTHWARK	Grocery Store	Bus Stop	Food & Drink Shop	Pub	Plaza	Pizza Place	Pharmacy	Park	Burger Joint	Breakfast Spot
1	ABBEY GROVE	GREENWICH	Convenience Store	Train Station	Coffee Shop	Platform	Farm	Electronics Store	Empanada Restaurant	English Restaurant	Entertainment Service	Ethiopian Restaurant
2	ABBEY PARADE	MERTON	Clothing Store	Coffee Shop	Tea Room	Bar	Café	Pub	Bakery	Vegetarian / Vegan Restaurant	Sandwich Place	Thai Restaurant
3	ABBEY ROAD	BEXLEY	Convenience Store	Train Station	Coffee Shop	Platform	Farm	Electronics Store	Empanada Restaurant	English Restaurant	Entertainment Service	Ethiopian Restaurant
4	ABBEYFIELD ROAD	SOUTHWARK	Pub	Boarding House	Brewery	Bus Stop	Farmers Market	Farm	Falafel Restaurant	Factory	Fabric Shop	Zoo Exhibit

Figure 12. Dataframe of the top 10 most common venues in each neighbourhood

4. Modelling

I decided to use 2 types of clustering algorithm, k – means and k – modes.

k – means is one of the most commonly used unsupervised clustering algorithm, which creates k number of clusters of data points aggregated based on similarities. The algorithm minimises the within-cluster variance (squared distances from the mean) by calculating the Euclidean distance between points and assigns each data point to the closest cluster centroid.

The mathematical condition for the K clusters C_k and the K centroids μ_k can be expressed as:

$$\text{Minimize} \sum_{k=1}^{K} \sum_{\mathrm{x}_n \in C_k} ||\mathrm{x}_n - \mu_k||^2 \text{ with respect to } C_k, \mu_k.$$

Figure 13. K clusters and K centroids mathematical expression

k – modes differs to k – means as the distance metric used it the Hamming distance or dissimilarity, meaning the smaller the number of total mismatches between 2 objects/rows/data points, the more similar they are.

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

where

$$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j. \end{cases}$$

Figure 14. Hamming distance mathematical expression

4.1. K - Means clustering

The Elbow Method is used to determine the optimal value of k as this is one of the most popular methods. I used 2 metric values calculated from a range of k values in order to determine the 'elbow point', i.e. the point after which the metrics starts decreasing linearly. Those 2 metric values are:

Distortion: Calculated as the average of the squared distances from the cluster centres of the respective clusters where typically the Euclidean distance is used.

Inertia: The sum of squared distances of samples to their closest cluster centre.

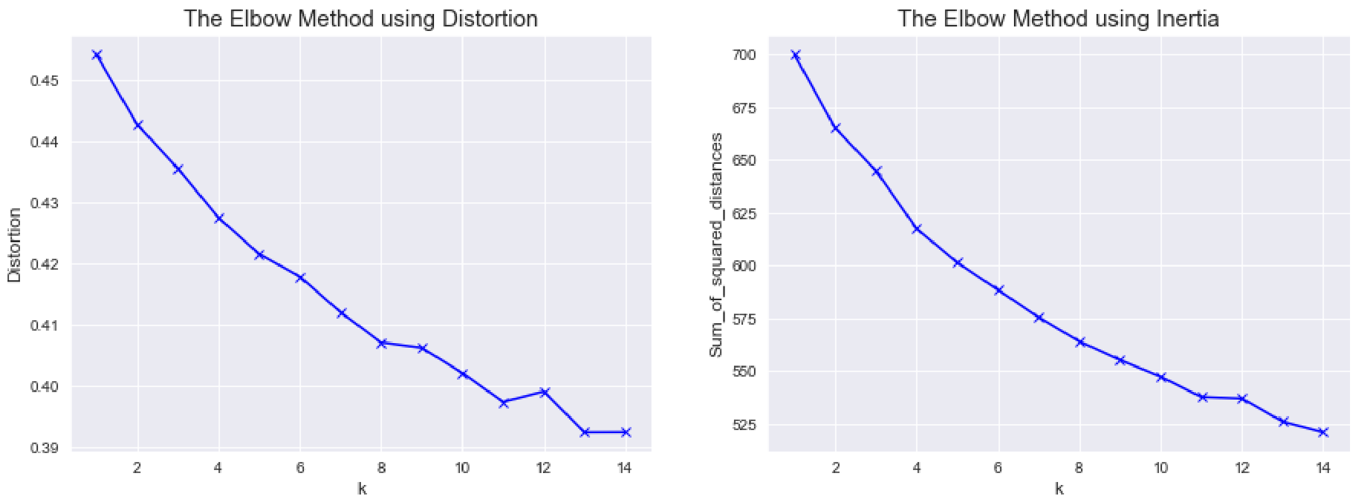

Figure 15. Line plot of The Elbow Method using distortion and inertia to the optimal k

As shown in Figure 15, the optimal k value is 5 as increasing it results in a smaller change in the distortion and inertia values in comparison to decreasing k.

The graph also shows that k = 10 & 12 could be potential k – values. However, as we increase the cluster numbers it could result in artificial boundaries being created within real data clusters, causing inaccuracies in our results therefore were not considered.

The algorithm is deployed using k = 5, the cluster labels generated are added to the dataframe in Figure 12, along with the postcode prefix, average price, latitude and longitude. The clusters are visualised on a map of London in Figure 17 and classified based on the most common venues for the corresponding cluster.

street	district	postcode_prefix	avg_price	latitude	longitude	Cluster Labels	...	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mo Comm Ven
ABBEY GARDENS	SOUTHWARK	SE16	330500.0	51.491653	-0.066099	4.0	...	Pharmacy	Food & Drink Shop	Café	Pub	Burger Joint	P
ABBEY PARADE	MERTON	SW19	242750.0	51.531393	-0.292546	0.0	...	Indian Restaurant	Mediterranean Restaurant	Metro Station	Fish & Chips Shop	Pharmacy	Engl Restaur
ABBEY ROAD	BRENT	NW10	950000.0	51.530067	-0.269922	3.0	...	Lebanese Restaurant	Movie Theater	Indian Restaurant	Supermarket	Sandwich Place	Z Exh
ABBEY ROAD	CAMDEN	NW6	396429.0	51.540987	-0.189608	0.0	...	Turkish Restaurant	Gym	Grocery Store	Bus Stop	Financial or Legal Service	Fab Sh
ABBOTS PARK	LAMBETH	SW2	489000.0	51.442994	-0.113085	0.0	...	Exhibit	Fabric Shop	Factory	Falafel Restaurant	Farm	Farm Mar
...	...	...	...	...	...	...	...	...	...	...	...	...	
YORK ROAD	EALING	W3	462000.0	51.518474	-0.264443	0.0	...	Clothing Store	Breakfast Spot	Cycle Studio	Czech Restaurant	Exhibit	Fab Sh
YORK WAY	CAMDEN	N1C	357350.0	51.536473	-0.122328	0.0	...	Italian Restaurant	Plaza	Pizza Place	Breakfast Spot	Market	Mob Pho Sh
YORK WAY ESTATE	ISLINGTON	N7	275625.0	51.545192	-0.125491	0.0	...	Café	Soccer Field	Tennis Court	Supermarket	Music Venue	Brew
YOUNG STREET	KENSINGTON AND CHELSEA	W8	1275735.0	51.501156	-0.189701	0.0	...	Juice Bar	Garden	French Restaurant	Pub	English Restaurant	Bak
YUKON ROAD	WANDSWORTH	SW12	648000.0	51.449287	-0.145819	4.0	...	Wine Shop	Bed & Breakfast	Gastropub	Zoo Exhibit	Exhibit	Fab Sh

Figure 16. Dataframe of venue data and cluster label for each neighbourhood

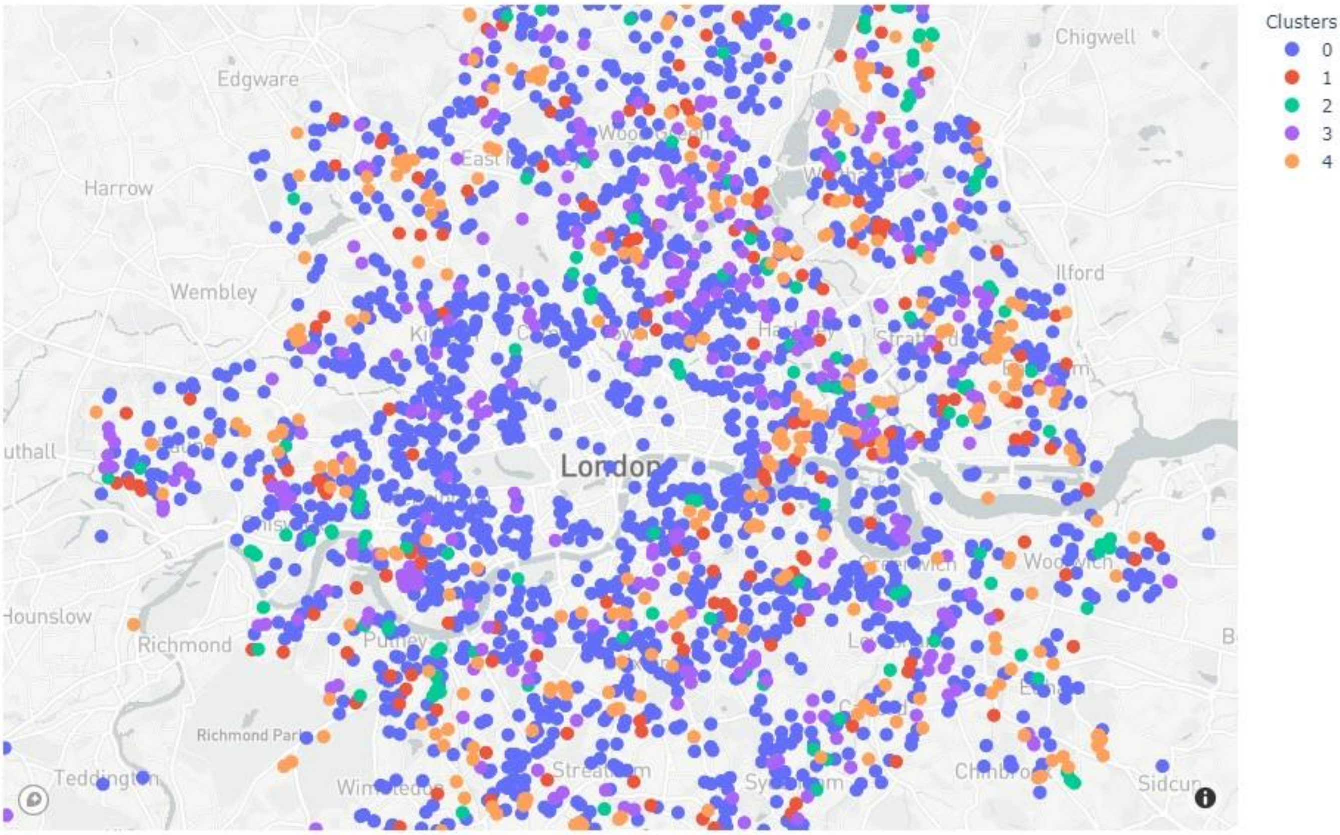

Figure 17. Map of London neighbourhood clusters using k - means

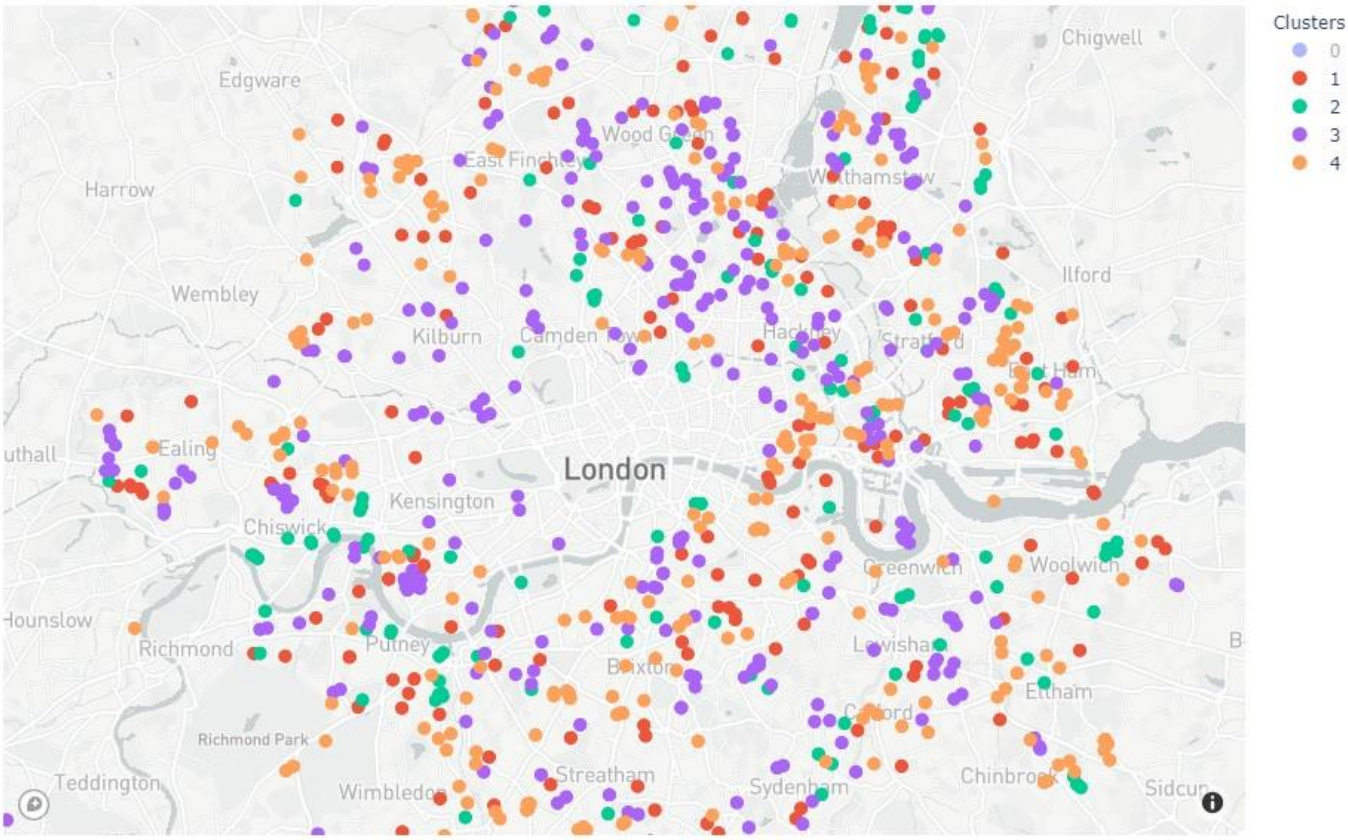

Figure 18. Map of London neighbourhood clusters using k – means (without Cluster 0)

Cluster – 0

Cluster 0 is the largest cluster by a considerable margin, making up 64% of the total number of data points. In comparison the next largest cluster is Cluster 3, making up only 13% of the total number of data points. Pubs and Coffee Shops make up the majority with 208 and 170 of them respectively in Cluster 0. Followed by Bus Stops (103 counts), Italian Restaurants (100 counts) and Hotels (84 counts). From Figure 17 we can see that the cluster is a lot denser in West London, most noticeably around Kensington, Kilburn, and Westminster.

Cluster Labels	1st Most Common Venue	count
0	Pub	208
0	Coffee Shop	170
0	Bus Stop	103
0	Italian Restaurant	100
0	Hotel	84

Cluster 0 can be classified as a 'Pub and Coffee Shop' dominate cluster.

Cluster – 1

Cluster 1 has the lowest number of data points (neighbourhoods) out of the 5 clusters. Park type venues are the most common venues in this cluster with 65 counts.

Cluster Labels	1st Most Common Venue	count
1	Park	65
1	Bus Stop	10
1	Convenience Store	7
1	Gym / Fitness Center	6
1	Trail	5

Cluster 1 can be classified as a 'Park' dominate cluster.

Cluster – 2

Similar to Cluster 0, pubs are also the most common venues with a count of 96. However, the 2nd to 5th most common venues do not have a count remotely comparable to those from Cluster 0 coming in at a total of 14 counts.

Cluster Labels	1st Most Common Venue	count
2	Pub	96
2	Bus Stop	4
2	Convenience Store	4
2	Hotel	3
2	Supermarket	3

Cluster 2 can be classified as a 'Pub' dominate cluster.

Cluster – 3

There are 151 cafes in this cluster, followed by 15 pubs, 11 bus stops, 11 coffee shops and 11 Italian restaurants. We can also see that North East London has a higher number of neighbourhoods that are in this cluster.

Cluster Labels	1st Most Common Venue	count
3	Café	151
3	Pub	15
3	Bus Stop	11
3	Coffee Shop	11
3	Italian Restaurant	11

Cluster 3 can be classified as a 'Café' dominate cluster.

Cluster – 4

The most common venue for Cluster 4 is grocery stores with a count of 116. Similar to Cluster 3 the other venues have a relatively low count with none exceeding 15.

Cluster Labels	1st Most Common Venue	count
4	Grocery Store	116
4	Coffee Shop	15
4	Bus Stop	12
4	Train Station	11
4	Italian Restaurant	10

Cluster 4 can be classified as a 'Grocery Store' dominate cluster.

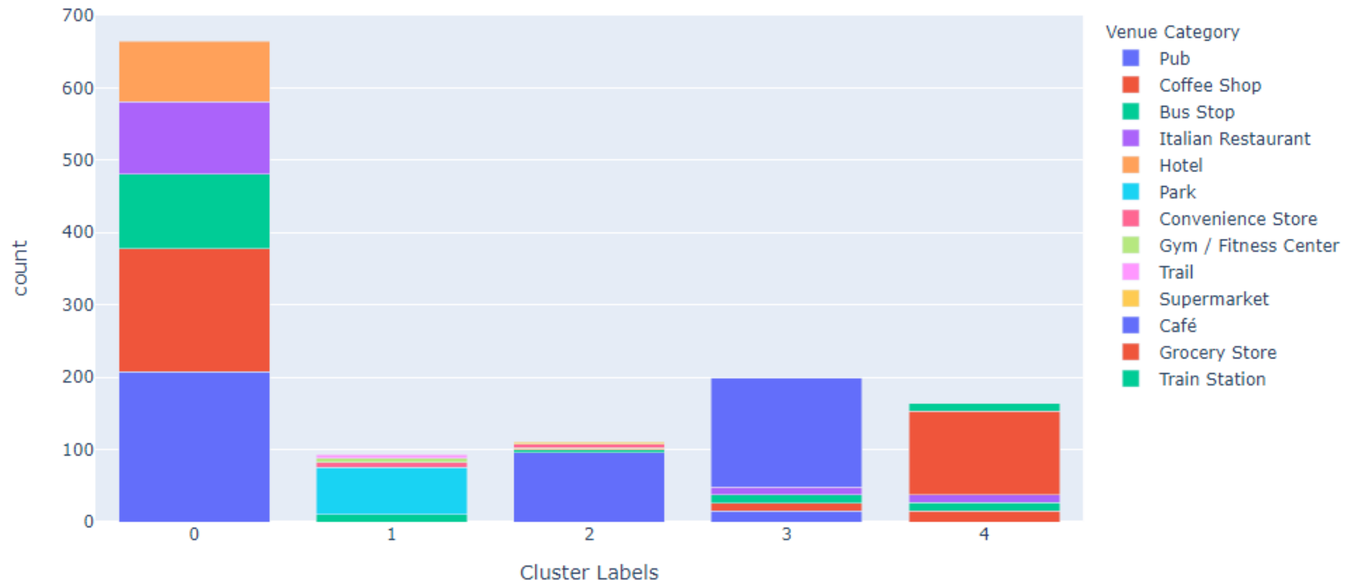

Figure 19. Top 5 most common venues in each cluster using k – means

Using k – means to cluster the neighbourhoods based on nearby venues resulted in Cluster 0 which contained more than 50% of the data points. This is most likely due to noise in the data, for example neighbourhoods where there were a small number of uncommon venues such as Automotive Store, Airport Terminal, and Laser Tag in combination with some of the more common venues such as Pubs.

4.2. K – Modes

The k – modes algorithm is used on the same data with the same k value selected for the k – means algorithm. From Figure 20 it is clear that the data points in clusters labelled using k – modes are more evenly distributed compared to those labelled using k – means. Cluster 2, represented by the green dots are generally located towards the edge of London, which I have determined as a 'Pub and Park' dominate cluster down below. This is different to Cluster 1, 3 and 4 where the data points are primarily located around central London, and these clusters are 'Coffee Shop & Café', 'Italian Restaurant' and 'Café' dominate clusters, respectively.

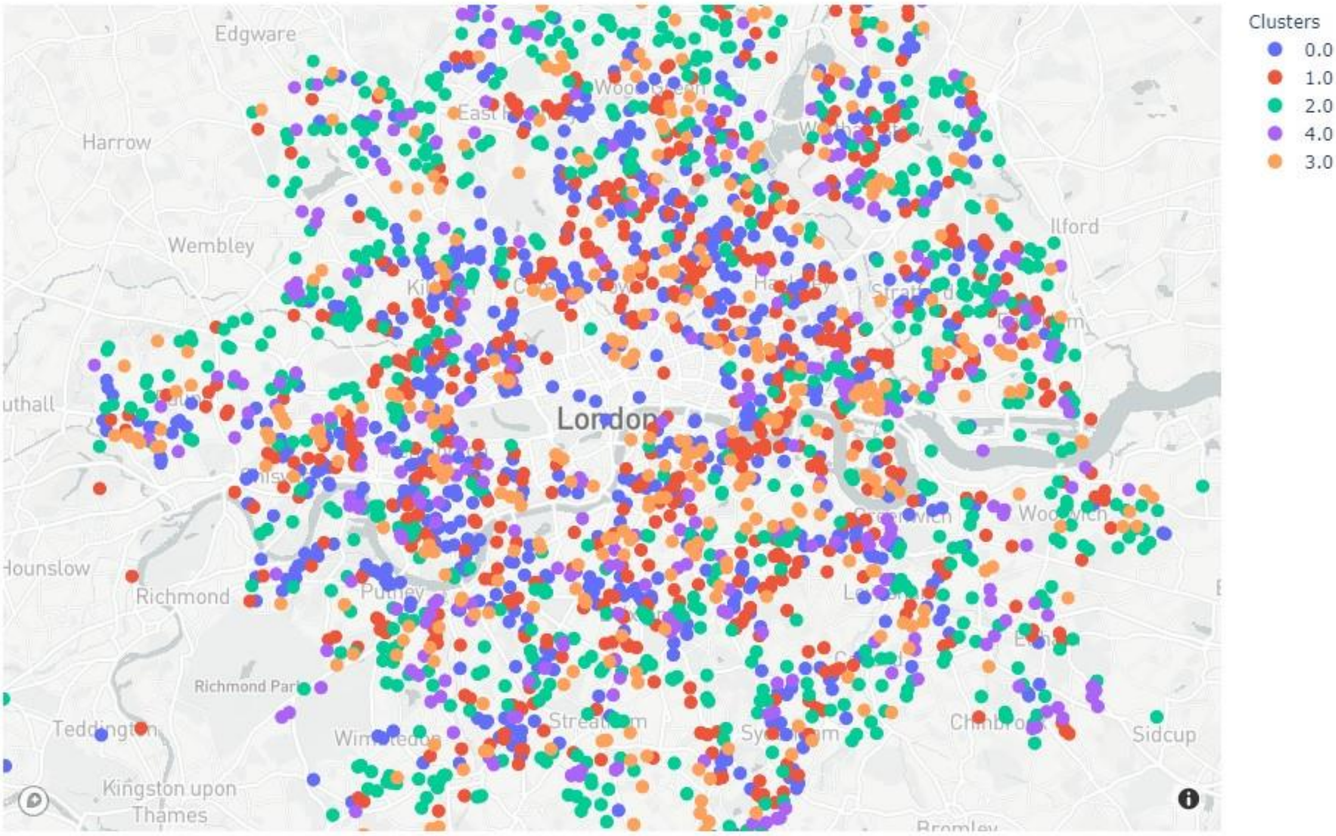

Figure 20. Map of London neighbourhood clusters using k – modes

However, the primary venue category for each cluster is less distinct when comparing Figure 21 with Figure 19, where the spread of the count of each venue categories within each cluster is relatively small.

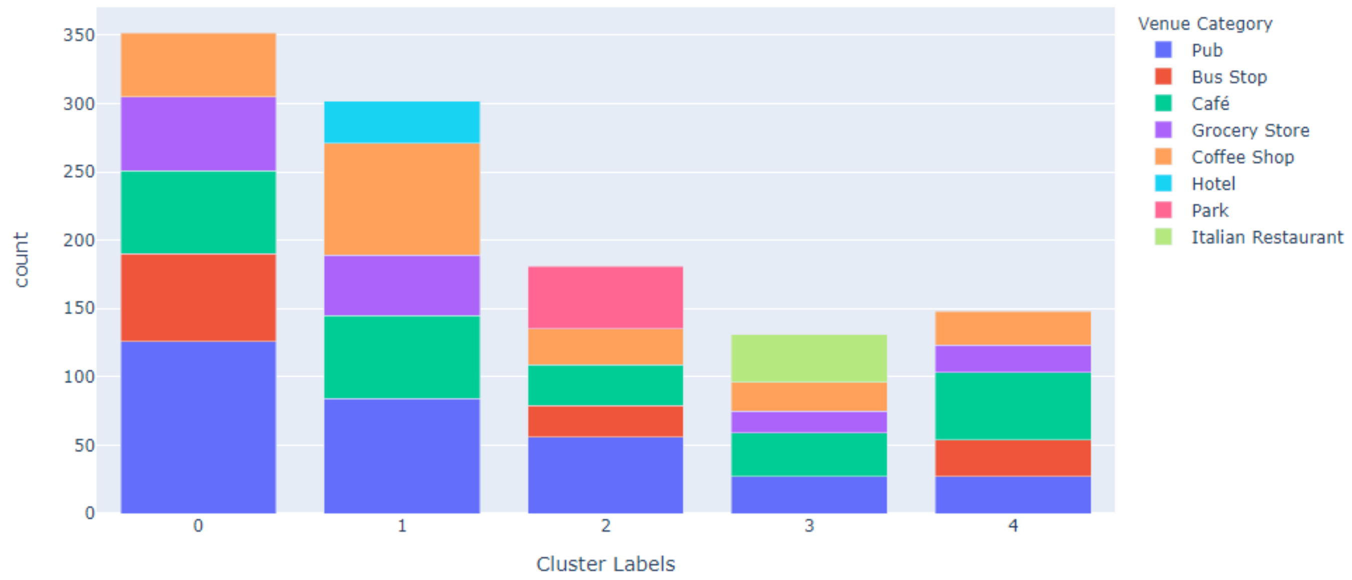

Figure 21. Top 5 most common venues in each cluster using k – modes

Cluster – 0

	Cluster Labels	1st Most Common Venue	count
0	0.0	Pub	126
1	0.0	Bus Stop	64
2	0.0	Café	61
3	0.0	Grocery Store	54
4	0.0	Coffee Shop	47

Cluster 0 can be classified as a 'Pub dominate cluster.

Cluster – 1

	Cluster Labels	1st Most Common Venue	count
5	1.0	Pub	84
6	1.0	Coffee Shop	82
7	1.0	Café	61
8	1.0	Grocery Store	44
9	1.0	Hotel	31

Cluster 1 can be classified as a 'Coffee Shop & Café' dominate cluster.

Cluster – 2

	Cluster Labels	1st Most Common Venue	count
10	2.0	Pub	56
11	2.0	Park	46
12	2.0	Café	30
13	2.0	Coffee Shop	26
14	2.0	Bus Stop	23

Cluster 2 can be classified as a 'Pub & Park' dominate cluster.

Cluster – 3

	Cluster Labels	1st Most Common Venue	count
15	3.0	Italian Restaurant	35
16	3.0	Café	32
17	3.0	Pub	27
18	3.0	Coffee Shop	21
19	3.0	Grocery Store	16

Cluster 3 can be classified as a 'Italian Restaurant' dominate cluster.

Cluster – 4

	Cluster Labels	1st Most Common Venue	count
20	4.0	Café	49
21	4.0	Bus Stop	27
22	4.0	Pub	27
23	4.0	Coffee Shop	25
24	4.0	Grocery Store	20

Cluster 4 can be classified as a 'Café' dominate cluster.

5. Discussion

Comparing property prices with the neighbourhood average property prices. Figure 22 shows that a majority of properties are valued below the neighbourhood average property prices. Based on the clusters labelled using k – means, Cluster 0 – Pub and Coffee Shop and Cluster 4 – Grocery Store both have a higher number of properties that are valued below the neighbourhood average property prices. This along with Figure 23 could indicate neighbourhoods where homebuyers and investors can potentially negotiate for a discounts and purchase a property for below the average asking price.

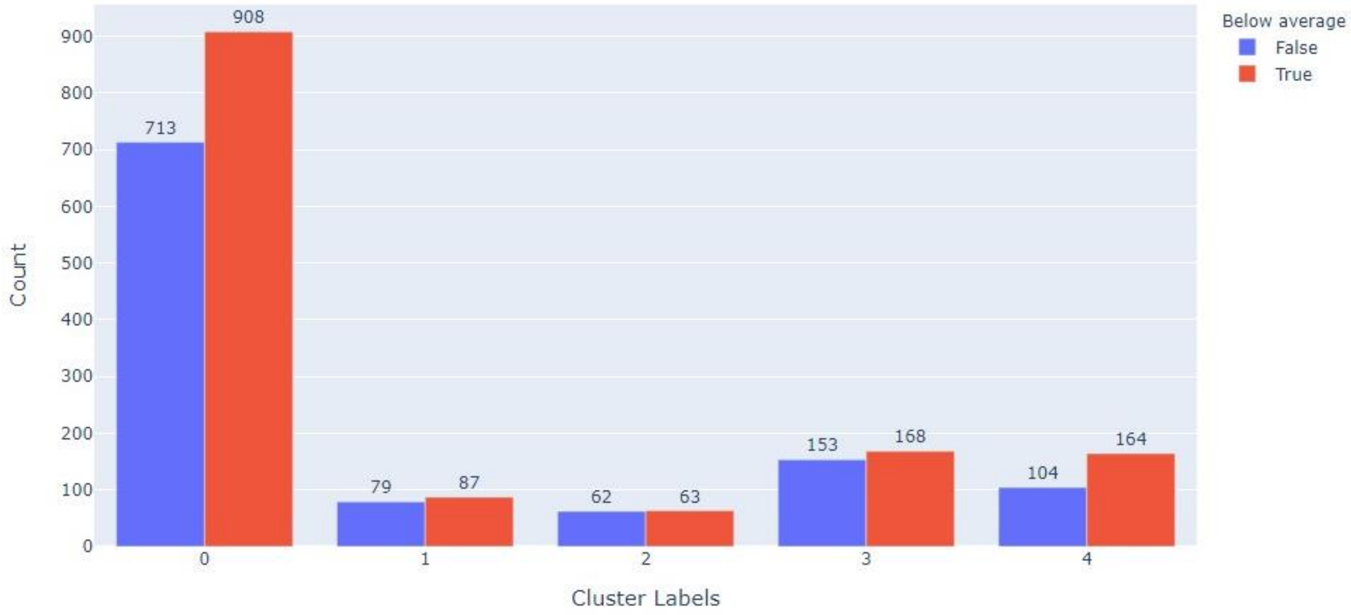

Figure 22. Number of properties where the value is below the average vs above the average for each cluster

Figure 23 shows the same comparison per borough, most of the boroughs have a higher number of properties that are valued below the neighbourhood average property prices with the exceptions of Bexley, City of London, Kingston and Chelsea, Newham, Redbridge and Richmond Upon Thames.

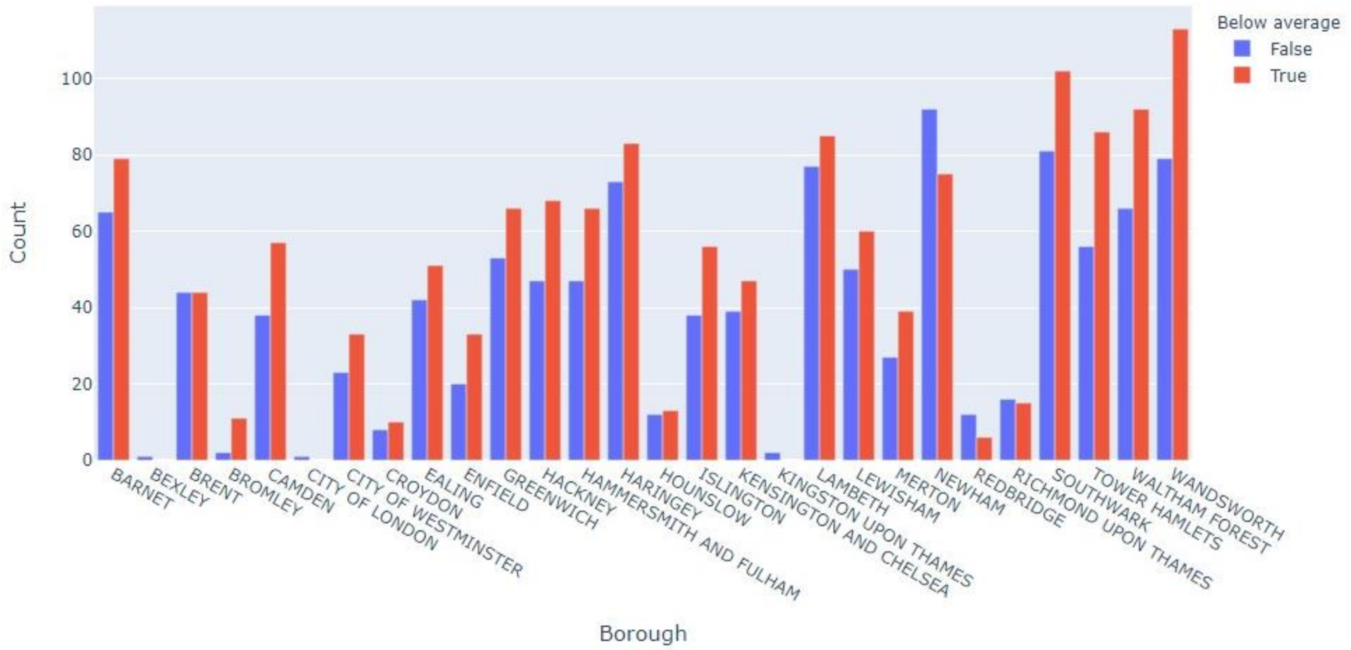

Figure 23. Number of properties where the value is below the average vs above the average for each borough

Interestingly for Cluster 4 – grocery store, in addition to having a higher number of properties valued below the neighbourhood average property prices, it also has the highest average rental yield percentage compared to the rest of the clusters at 9.1%. Cluster 1 – parks, have the lowest average rental yield percentage at 5.2%

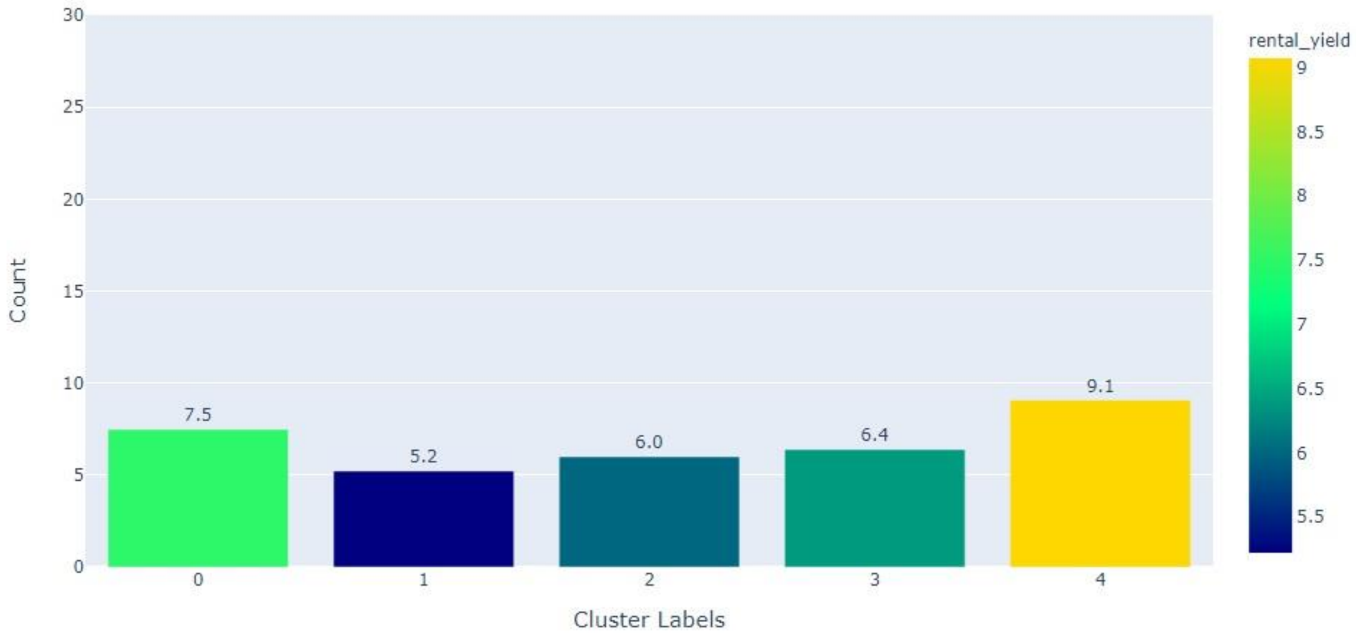

Figure 24. Average rental yield percentage for each cluster

The same data is visualised per borough as it is clear that Hackney has the highest average rental yield percentage at 17%, followed by Camden at 13%, on the other hand Kingston Upon Thames has the lowest at 2.5%.

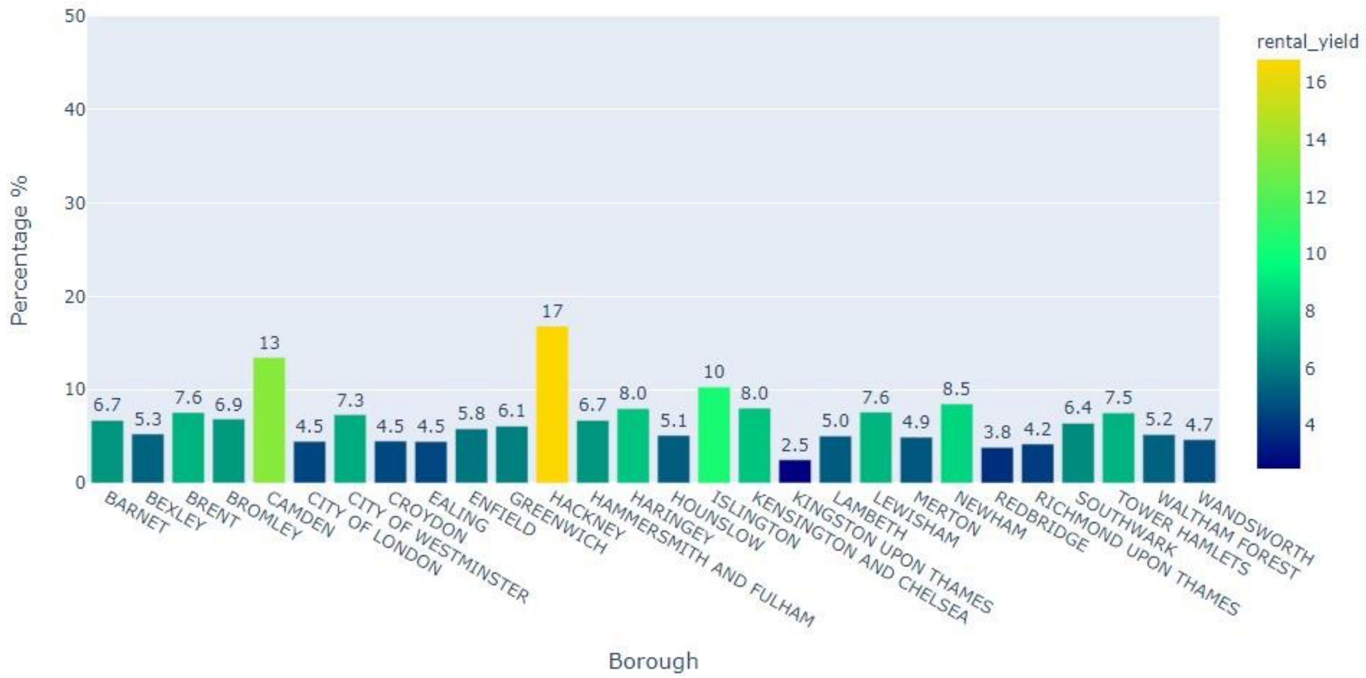

Figure 25. Average rental yield percentage for each borough

6. Conclusion

The objective of this report was to find out how can we generate insight from venue data so home-buyers and investors can make well informed choices when purchasing properties in London. Using k – means clustering I was able to cluster neighbourhoods based on their nearby venues, in combination with analysing property price paid data I was able to see the correlation between the two.

Neighbourhoods that have grocery stores nearby tend to have a higher number of property where the value of the property is below the neighbourhood average, and they tend to have a higher average rental yield percentage (9.1%). Homebuyers and investors can focus on properties located in Camden, Islington, Tower Hamlets and Kensington and Chelsea.

Alternatively, neighbourhoods that have pubs and coffee shops nearby also follow a similar trend as the one discussed above, with an average rental yield at 7.5%. Although some homebuyers might not like being in a close proximity of a pub due to noise levels at the evening, especially during the weekends.

7. Future directions

As mentioned in 4.1. K - Means clustering, the model could benefit from some improvements on tweaking the parameters as well as eliminating noise from the input data. Other clustering models that are designed for categorical data could also be looked into, such as k – prototype and hierarchical clustering.

This report was focused primarily on housing prices and nearby venue data; however, the scope of the report can be expanded to include other features such as number of bedrooms, type of property and age of property. Historical housing from 1995 – 2017 can also be investigated and analysed further.

An interactive dashboard can also be built using Plotly Dash to display all the relevant charts & plots, allowing a user to select and visualise data for a particular cluster or borough.

8. References

How to access HM Land Registry Price Paid Data: https://www.gov.uk/guidance/about-the-price-paid-data

Price Paid Data - HM Land Registry: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

Average private rental prices per borough:https://data.london.gov.uk/dataset/average-private-rents-borough

Borough property and rental prices - Foxtons: https://www.foxtons.co.uk/living-in/bermondsey

List of London boroughs : https://en.wikipedia.org/wiki/List_of_London_boroughs

London Borough GeoJSON: https://joshuaboyd1.carto.com/tables/london_boroughs_proper/public