Moving to London

Lakshan Aaron Samaraweera

April 3rd, 2021

1. Introduction

1.1. Background

London is one of the most historic capitals in Europe, originally founded by the Romans in 43AD. Further to this, in modern times, London is also one of the most multi-culturally diverse capitals in the world with 250 different languages spoken and 270 nationalities in the city alone. The total population of London stands at 8.9million and is split into 32 boroughs over a 1,572 km² distance. Due to its size and population, London can be a very daunting city to settle into, with no prior knowledge of which boroughs are safer than others, which boroughs meet the financial needs of the individual or family, which boroughs provide a higher chance of having access to qualified talent or increasing the chances of acquiring a high level of qualification and what are the most predominant venues you are most likely going to find in that borough. The decision to move to London can go wrong and can make the whole experience of living in this metropolitan city a very tough one. Not only can this analysis of London increase the chances of an individual or family settle in successfully, but also aid local councils and government to improve their boroughs.

1.2. Problem

The data gathered will aid in determining the sufficient borough for a family or individual to locate to. In addition, the data can also help contribute in helping businesses finding a borough to establish a new business or a new branch and also can influence decisions of the local authority in London.

1.3. Interest

The type of audience that would be interested in the data gathered and the analysis would be families and individuals relocating from other cities in the UK or the world. Furthermore, businesses and local councils and authorities would also find an interest in the analysis.

2. Data Description, Source and Cleaning

2.1. Data Description

To analyse London and to accomplish the objective, data will need to be collected. The data I have chosen to analyse the London boroughs are crime rates, average hourly pay of each borough, the level of qualification, average rent price and average house price. Furthermore, I will use the Foursquare API to find the top 10 most common venues in each borough in a 2km radius from the coordinates of each borough. The data I selected to analyse is based on data collected in 2019. I chose this period as this was the last year of 'normality' before the Covid-19 pandemic struck London and the globe which has affected many lives and in turn the data as well. Consequently, the covid-19 pandemic forced London to go on a city-wide lockdown. This has affected businesses which have affected the Foursquare venue data in 2020 and 2021 as businesses are forced to close due to restrictions enforced, there has been an effect on qualification levels as individuals are unable to take exams at the planned date due to lockdown restrictions enforced which has affected individual's ability to be qualified, also this has affected house prices and rent prices as demand to move into the city has been affected as in pre-lockdown, individuals would usually move closer to work if they worked in London, however due to lockdown restrictions on commuting and working in offices, individuals are forced to work at home and are in need of a more open living environment where some parts of London are unable to provide, some have worked remotely and have moved out of the city to less densely populated areas. Henceforth, data from 2020 and 2021, does not give an honest analysis of London as the city is in a period of lockdown.

2.1.1. Crime Rates

Data on crime rates will help the individual or family scope the safety of the borough. This data is based on crime rates in 2019 and is totalled over that period. The crime rate is split into Arson and Criminal Damage, Burglary, Drug Offences, Miscellaneous Crimes Against Society, Possession of Weapons, Public Order Offences, Robbery, Sexual Offences, Theft, Vehicle Offences and Violence Against the Person. The data was sourced from the London Datastore in the form of the CSV file. The data was split into major text and minor text which are a type of crime and a description of the crime, respectively. All I needed to process the data was the number of crimes committed in each borough. Therefore, I dropped the minor text column and grouped the boroughs by the major text and summed the number of crimes in each major text column. I proceeded to find the sum of all crime committed over the year 2019 for each borough and created a new column.

2.1.2. Average Hourly Earnings

This data is based on how much on average an individual living in a particular borough earns per hour. The analysis can help gauge an insight into how expensive a borough can be as this is how much an individual needs to make to live in that borough. Qualification level Qualification data is split into five parameters no qualification, level 1 qualified; which is basic/essential skills, level 2 qualified; which is five GCSE's A*-C (end of high-school qualification), level 3 qualified; 2+ A-levels or equivalent, level 4; degree level or above. If an individual was to establish a business that requires an employee with a certain level of qualification, they can use this data to find the borough with the highest level of qualified talent and establish their business in that borough as this will help individuals have access to employment more efficiently. Another use for this data is for families with children that need to look for boroughs were having a high level of qualified talent can increase the chances of their children attaining a level 4 qualification in today's intellectually driven society. The data sourced was from the London Datastore and was in the form of an excel sheet. The data sourced was from 2003 to 2020, I therefore dropped the unnecessary columns for the analysis which left 2019 available for analysis. Furthermore, I deleted rows that would skew the data for instance, rows with null values and rows that do not affect London e.g., Yorkshire and the City of London. The data gathered and cleaned was merged with the crime data frame.

2.1.3. Qualification level

Qualification data is split into five parameters no qualification, level 1 qualified; which is basic/essential skills, level 2 qualified; which is five GCSE's A*-C (end of high-school qualification), level 3 qualified; 2+ A-levels or equivalent, level 4; degree level or above. If an entity were to establish a business that requires an employee with a certain level of qualification, they can use this data to find the borough with the highest level of qualified talent and establish their business in that borough as this will help individuals have access to employment more efficiently. Another use for this data is for families with children that need to look for boroughs were having a high level of qualified talent can increase the chances of their children attaining a level 4 qualification in today's intellectually driven society. The data sourced was from the London Datastore and was based on the qualifications of the working age NVQ. The data was in the form of an Excel file. To process and clean the data, percentages were taken of the level of qualification from no qualification to level 4 qualification between the ages of 16-64 of each borough. The qualification date frame was merged with the main data frame which included crime and average hourly earnings.

2.1.4. Average house price

Average house price is based on house prices in each of the 32 boroughs in 2019. There are many different types of houses with a varying number of rooms. This data will be used to analyse the financial value of a borough which will aid the individual or family understand how high valued or undervalued a borough is. The data sourced was from the London Datastore and comprised of average house price data from 1995 to 2020 by the month which was a large dataset. To clean the data, the 'datetime' library was imported to extract the 2019 average house price for each month. An average was taken over the year and the new data frame was merged with the main data frame.

2.1.5. Average rent price

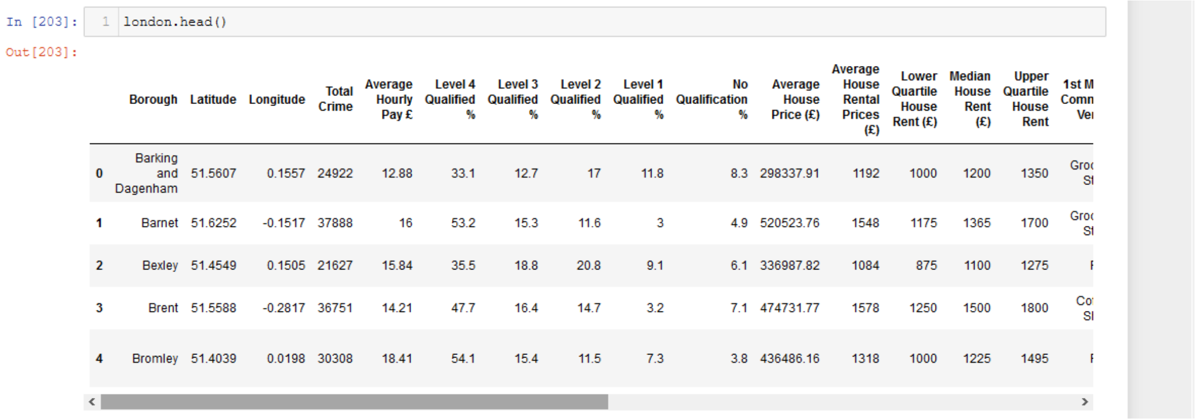
The average rent price is based on private rent and does not include local council rent. This data will help individuals looking to move into a borough of London on a short-term basis. The data was sourced from the London data store and ranged from 2011 to 2019. Only data on 2019 was extracted from the dataset and an average, lower quartile, median and upper quartile was taken for each borough and merged with the main data frame.

2.1.6. Venues

Venue data is taken from the Foursquare API. The top 100 venues are taken from each borough from a 2km distance from the coordinates of each borough. Out of the 100 venues a list of the top 10 common venues were taken from each borough. Using the venue data can give insight to the individual or family as to what is the most predominant venue in the borough based on how common those venues occur. So, if an individual is moving into London for a short-term period, they may want to experience the social life of London and focus on living in boroughs where pubs are the common venue in the borough. Or, if a family wants to move to London they would focus on parks and grocery shops. The venues were sourced from the Foursquare API. A 100 venues were taken from a 2km radius from the extracted coordinates of each borough. The data was counted and put into clusters and merged with the main data frame.

2.1.7. Coordinates

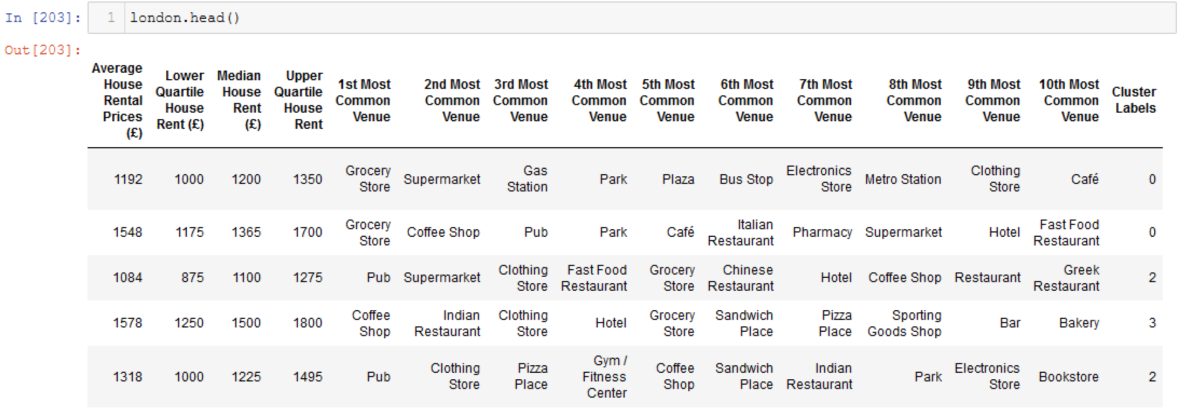
Will aid in mapping the borough on a map and in turn help with the venue data. The coordinates were sourced from Wikipedia's London borough page. I imported the library called 'BeautifulSoup' which allowed me to scrape the table data from the page and process and merged the data on to the main data frame.



```
In [203]:  london.head()
Out[203]:
```

	Borough	Latitude	Longitude	Total Crime	Average Hourly Pay £	Level 4 Qualified %	Level 3 Qualified %	Level 2 Qualified %	Level 1 Qualified %	No Qualification %	Average House Price (£)	Average House Rental Prices (£)	Lower Quartile House Rent (£)	Median House Rent (£)	Upper Quartile House Rent	1st M Comn Vei
0	Barking and Dagenham	51.5607	0.1557	24922	12.88	33.1	12.7	17	11.8	8.3	298337.91	1192	1000	1200	1350	Groc St
1	Barnet	51.6252	-0.1517	37888	16	53.2	15.3	11.6	3	4.9	520523.76	1548	1175	1365	1700	Groc St
2	Bexley	51.4549	0.1505	21627	15.84	35.5	18.8	20.8	9.1	6.1	336987.82	1084	875	1100	1275	F
3	Brent	51.5588	-0.2817	36751	14.21	47.7	16.4	14.7	3.2	7.1	474731.77	1578	1250	1500	1800	Co St
4	Bromley	51.4039	0.0198	30308	18.41	54.1	15.4	11.5	7.3	3.8	436486.16	1318	1000	1225	1495	F

Figure 1. Main London data frame part 1



```
In [203]:  london.head()
Out[203]:
```

Average House Rental Prices (£)	Lower Quartile House Rent (£)	Median House Rent (£)	Upper Quartile House Rent	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	Cluster Labels
1192	1000	1200	1350	Grocery Store	Supermarket	Gas Station	Park	Plaza	Bus Stop	Electronics Store	Metro Station	Clothing Store	Café	0
1548	1175	1365	1700	Grocery Store	Coffee Shop	Pub	Park	Café	Italian Restaurant	Pharmacy	Supermarket	Hotel	Fast Food Restaurant	0
1084	875	1100	1275	Pub	Supermarket	Clothing Store	Fast Food Restaurant	Grocery Store	Chinese Restaurant	Hotel	Coffee Shop	Restaurant	Greek Restaurant	2
1578	1250	1500	1800	Coffee Shop	Indian Restaurant	Clothing Store	Hotel	Grocery Store	Sandwich Place	Pizza Place	Sporting Goods Shop	Bar	Bakery	3
1318	1000	1225	1495	Pub	Clothing Store	Pizza Place	Gym / Fitness Center	Coffee Shop	Sandwich Place	Indian Restaurant	Park	Electronics Store	Bookstore	2

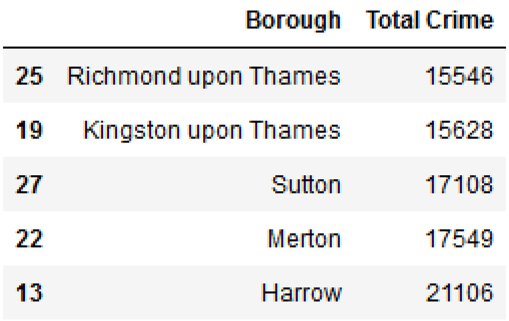
Figure 1.2. Main London data frame part 2

3. Methodology

The main components used for the data analysis was borough, total crime, Average hourly pay, level 1,2,3,4 and no qualification data, average house and rent prices of each borough.

3.1. Crime Data Analysis

To visualise the data, total crime and borough information was taken from the main data frame and sorted from low counts of crime to high counts of crime. This table will be later used to visualise the crime count in each borough of London.



	Borough	Total Crime
25	Richmond upon Thames	15546
19	Kingston upon Thames	15628
27	Sutton	17108
22	Merton	17549
13	Harrow	21106

Figure 2. Crime data sorted from low to high.

After creating a new data frame from the main data frame, the data was used to plot a bar chart of each borough crime counts over 2019.

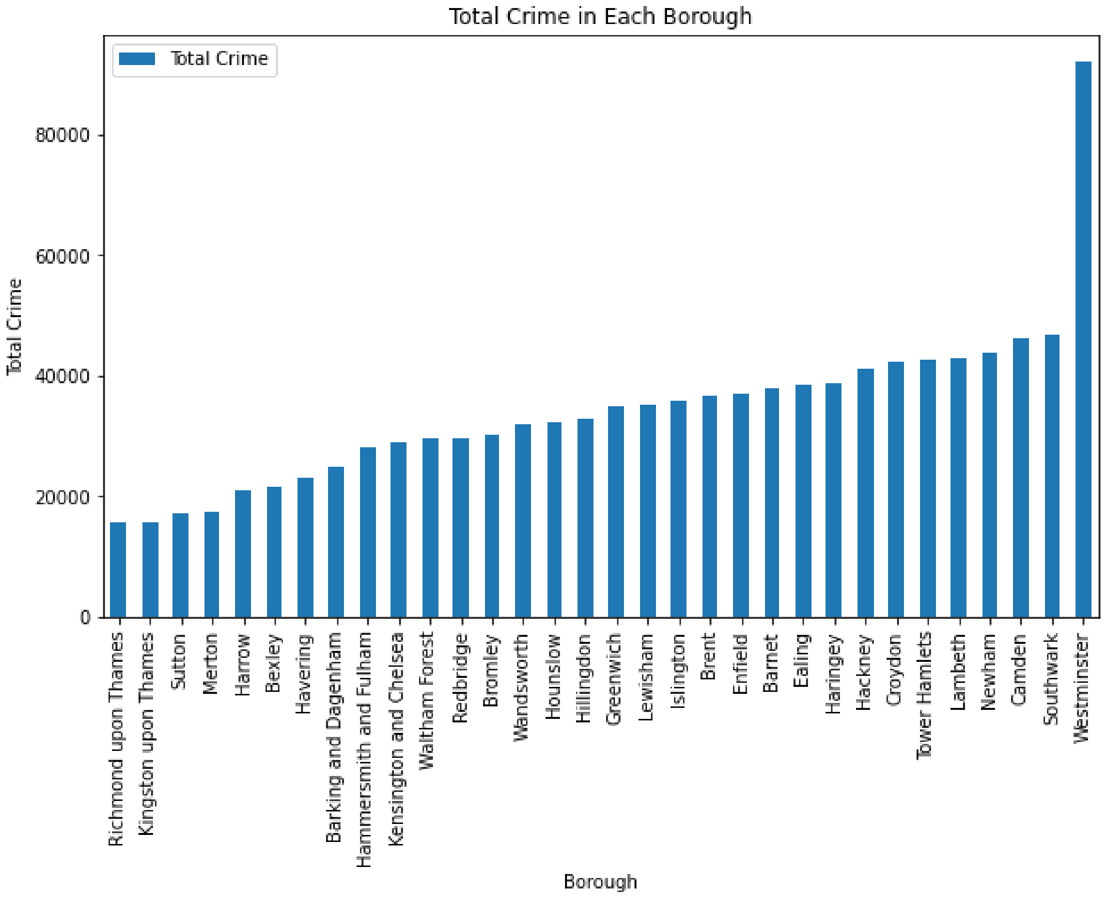


Figure 2.2. Bar chart of the total crime in each borough over 2019

In the above bar chart, you can clearly see that the borough of Westminster has the largest total crime than any borough in London. The second highest level of crime in London come from the London borough of Southwark and is just a little over half the total amount of crime that Westminster. The lowest level of crime come from Richmond upon Thames and Kingston upon Thames. Later in this report the crime data will be visualised further in the form of a Choropleth map to give an indication of the neighbourhoods that lie in the borough of Southwark.

3.2. Average Hourly Earnings, House Price and Rent Price

Extracting the average hourly earnings, house price and rent price, I came to the conclusion that these were financially driven parameters that would help an entity to make a financial based decision as to were they would like to locate. Therefore, I extracted the average hourly earnings, house price and rent price from the main data frame. I went further to sum up all the values for each borough and created a new column named 'financial value for each borough'. I proceeded to create bins by using a histogram to bin each borough into a certain financial value category.

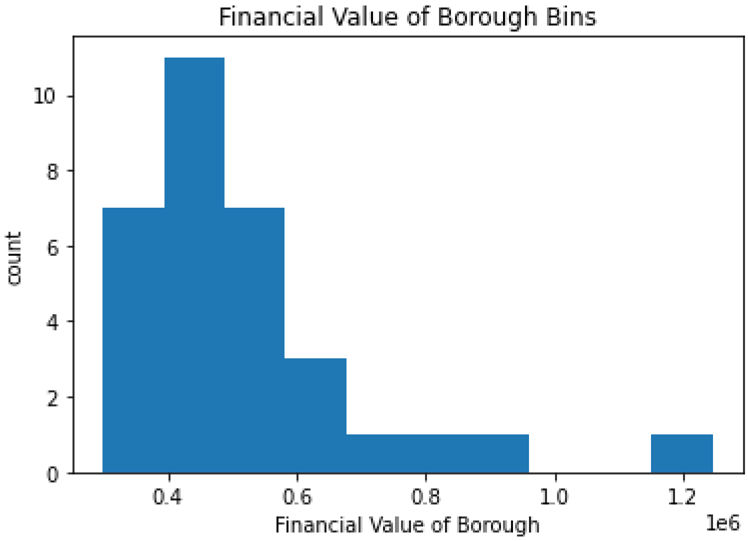


Figure 3. Histogram to count how many boroughs are in each categorical bin.

Using the histogram, I created the following categories:

- Low Borough FV (Financial Value): £250,000 - £450,000
- Low Tier 2 Borough FV: £451,000 - £600,000
- Medium Borough FV: £601,000 - £750,000
- Medium Tier 2 Borough FV: £751,000 - £850,000
- High Borough FV: £851,000 - £1,000,000
- High Tier 2 Borough FV: £1,000,000 +

I chose these categories and the size based on the data and the large disparities between some boroughs.

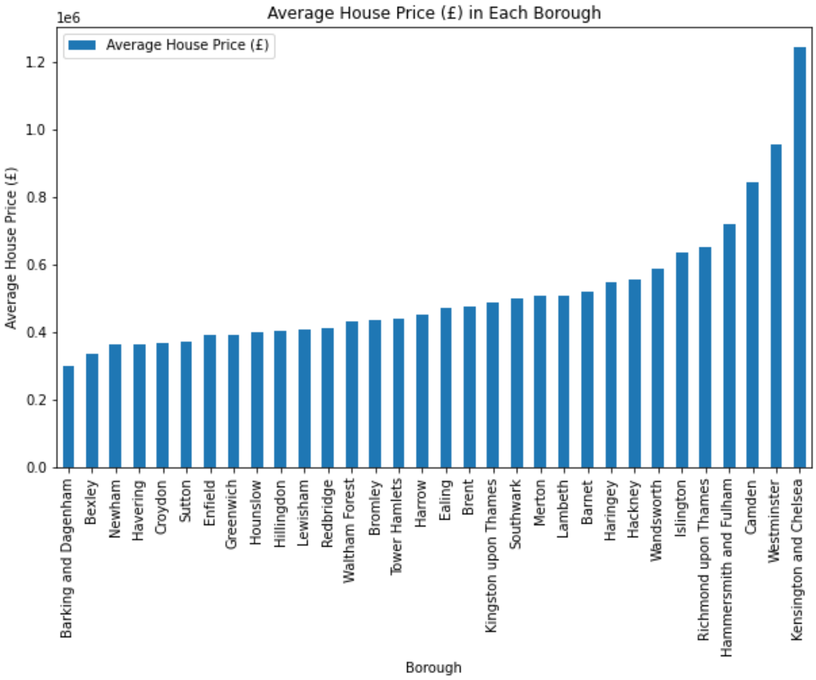


Figure 3,1. Bar Chart on Average House Prices

As you can see there is massive exponential increase of house prices from Wandsworth to Kensington and Chelsea which is the cause for there irregular ranges for the bins. The same also applies to average rent prices as well.

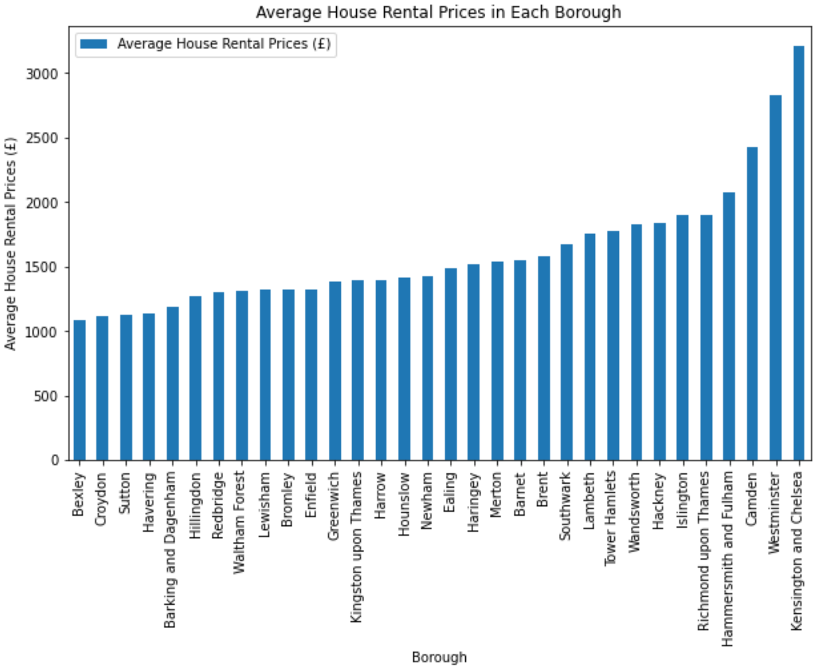
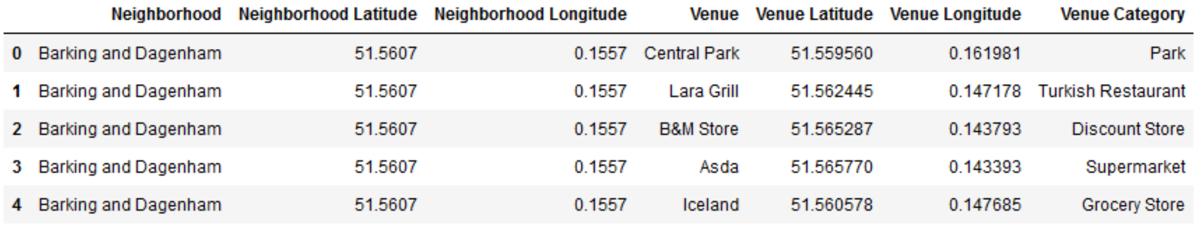


Figure 3.2. Bar Chart on Average Rent Prices

3.3. Venues from Foursquare

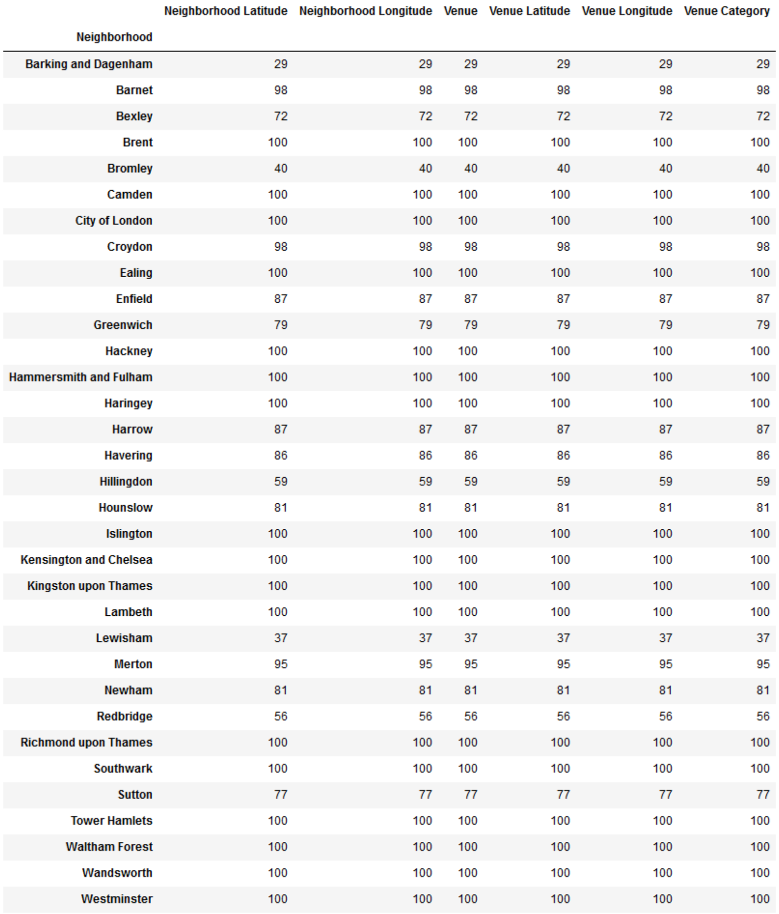
Using the Foursquare API, I extracted up to 100 venues from each borough over a 2km distance using the coordinates scraped from the London Boroughs Wikipedia page. This provided a new data frame which consisted of the borough, latitude and longitude points, the venue, the venue's latitude and longitude points and the category of the venue.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barking and Dagenham	51.5607	0.1557	Central Park	51.559560	0.161981	Park
1	Barking and Dagenham	51.5607	0.1557	Lara Grill	51.562445	0.147178	Turkish Restaurant
2	Barking and Dagenham	51.5607	0.1557	B&M Store	51.565287	0.143793	Discount Store
3	Barking and Dagenham	51.5607	0.1557	Asda	51.565770	0.143393	Supermarket
4	Barking and Dagenham	51.5607	0.1557	Iceland	51.560578	0.147685	Grocery Store

Figure 4. Data frame of venues in each borough

To confirm how many venues were given by the Foursquare API, the Boroughs were grouped, and each venue provided by the foursquare API was counted.

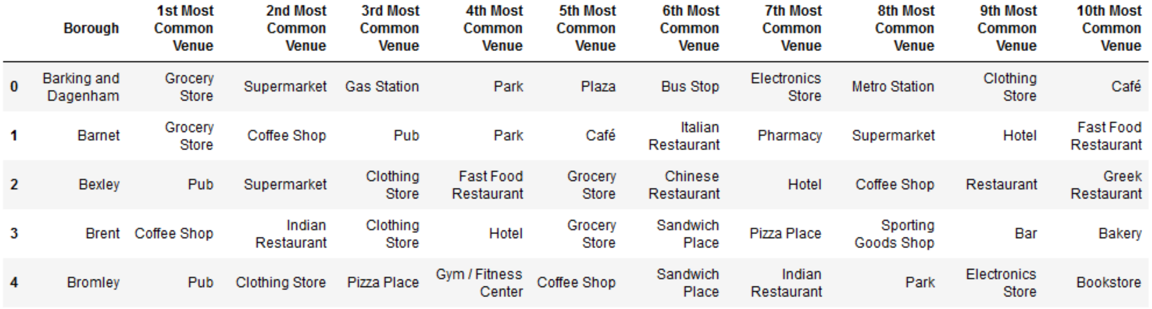


	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Barking and Dagenham	29	29	29	29	29	29
Barnet	98	98	98	98	98	98
Bexley	72	72	72	72	72	72
Brent	100	100	100	100	100	100
Bromley	40	40	40	40	40	40
Camden	100	100	100	100	100	100
City of London	100	100	100	100	100	100
Croydon	98	98	98	98	98	98
Ealing	100	100	100	100	100	100
Enfield	87	87	87	87	87	87
Greenwich	79	79	79	79	79	79
Hackney	100	100	100	100	100	100
Hammersmith and Fulham	100	100	100	100	100	100
Haringey	100	100	100	100	100	100
Harrow	87	87	87	87	87	87
Havering	86	86	86	86	86	86
Hillingdon	59	59	59	59	59	59
Hounslow	81	81	81	81	81	81
Islington	100	100	100	100	100	100
Kensington and Chelsea	100	100	100	100	100	100
Kingston upon Thames	100	100	100	100	100	100
Lambeth	100	100	100	100	100	100
Lewisham	37	37	37	37	37	37
Merton	95	95	95	95	95	95
Newham	81	81	81	81	81	81
Redbridge	56	56	56	56	56	56
Richmond upon Thames	100	100	100	100	100	100
Southwark	100	100	100	100	100	100
Sutton	77	77	77	77	77	77
Tower Hamlets	100	100	100	100	100	100
Waltham Forest	100	100	100	100	100	100
Wandsworth	100	100	100	100	100	100
Westminster	100	100	100	100	100	100

Figure 4.1. Data frame of the counted venues in each borough

As displayed most boroughs have up to 100 venues over a 2km distance. However, some boroughs do not this is down to the size of the borough and how less dense the borough is.

Once the venue data was gathered and counted the venues were grouped by the top ten most common venues in each borough.



	Borough	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	Barking and Dagenham	Grocery Store	Supermarket	Gas Station	Park	Plaza	Bus Stop	Electronics Store	Metro Station	Clothing Store	Café
1	Barnet	Grocery Store	Coffee Shop	Pub	Park	Café	Italian Restaurant	Pharmacy	Supermarket	Hotel	Fast Food Restaurant
2	Bexley	Pub	Supermarket	Clothing Store	Fast Food Restaurant	Grocery Store	Chinese Restaurant	Hotel	Coffee Shop	Restaurant	Greek Restaurant
3	Brent	Coffee Shop	Indian Restaurant	Clothing Store	Hotel	Grocery Store	Sandwich Place	Pizza Place	Sporting Goods Shop	Bar	Bakery
4	Bromley	Pub	Clothing Store	Pizza Place	Gym / Fitness Center	Coffee Shop	Sandwich Place	Indian Restaurant	Park	Electronics Store	Bookstore

Figure 4.2. Data frame on top ten most common venues

Due to the number of common venues in each borough, cluster segmentation of venues for each borough was implemented. Cluster segmentation will enable the audience to understand which borough has a predominant type of venue. To cluster the venues, K-Means machine learning algorithm was used, this was an unsupervised model. To determine the number of clusters to be used, the elbow method was implemented.

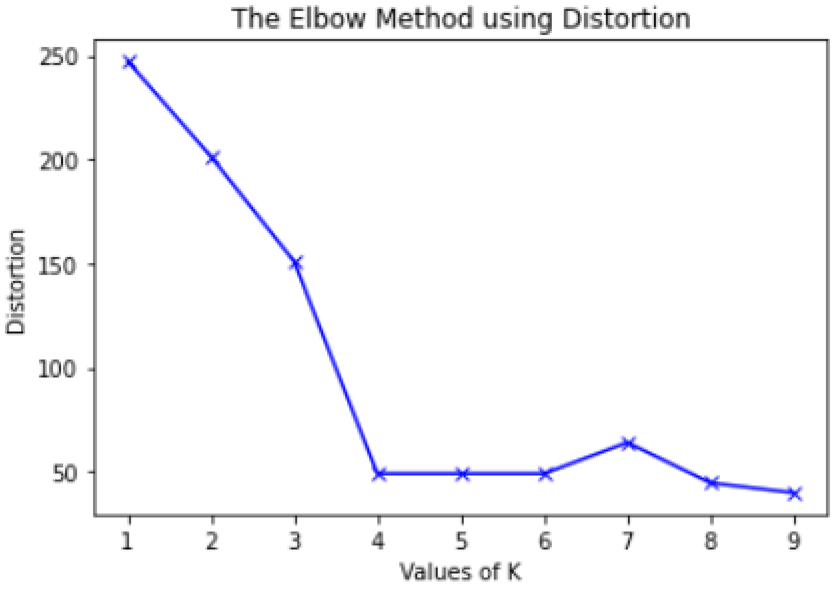


Figure 5. Chart displaying the elbow method.

The distance metric used was Canberra. I used the Canberra distance metric instead of Cosine or Euclidian distance metric as the elbow was more prominent and thus efficient enough to judge the number of clusters required to cluster the venue data. As the graph indicates, the number of clusters that are appropriate to clusters the venue data are 4. Each cluster will be assigned to the main data frame from 0,1,2 and 3.

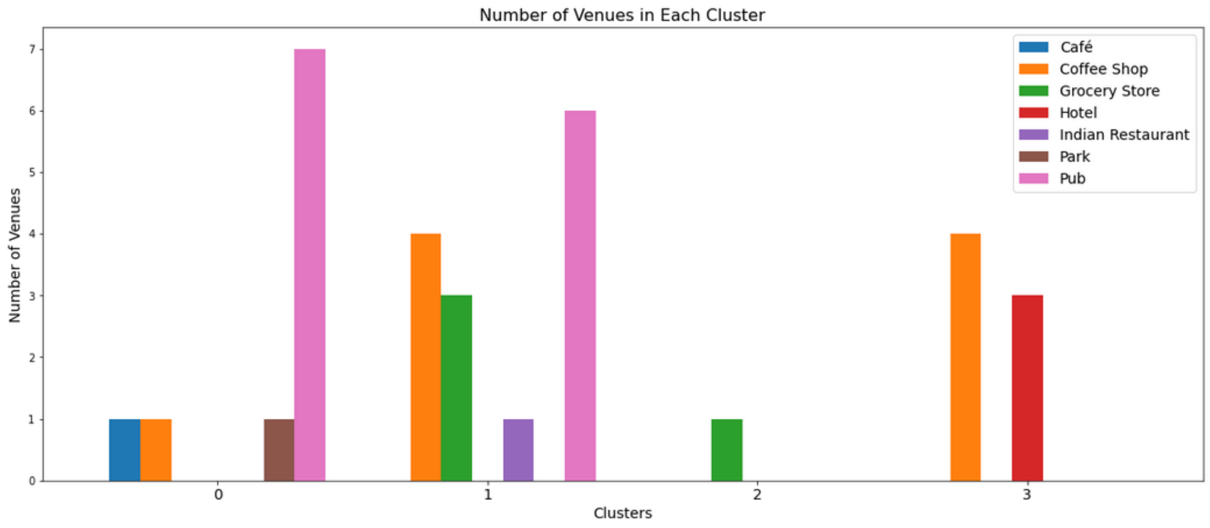
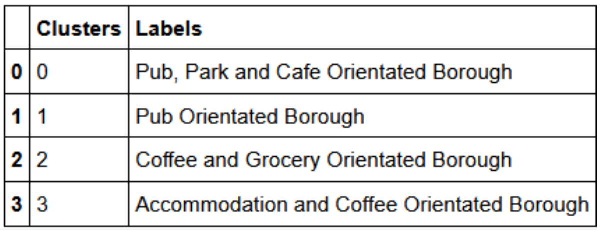


Figure 5.1. Bar chart of the number of venues in each cluster

The bar chart above shows the clusters that have been assigned to each borough and how many venues are there of each type in each cluster. Based on the analysis of the bar chart, the clusters were named as follows:



	Clusters	Labels
0	0	Pub, Park and Cafe Orientated Borough
1	1	Pub Orientated Borough
2	2	Coffee and Grocery Orientated Borough
3	3	Accommodation and Coffee Orientated Borough

Figure 5.2. Cluster Labels

Each label was assigned to the borough. This will give the audience an indication what the most predominate type of venue in that borough is.

3.4. Percentage of Borough Qualified

From the main data frame, the borough, level 4 and no qualification data were extracted and a new data frame was created. To visualise the disparities between the percentage of the borough that was fully qualified and the percentage of the borough population that had no qualification, a horizontal bar chart was created.

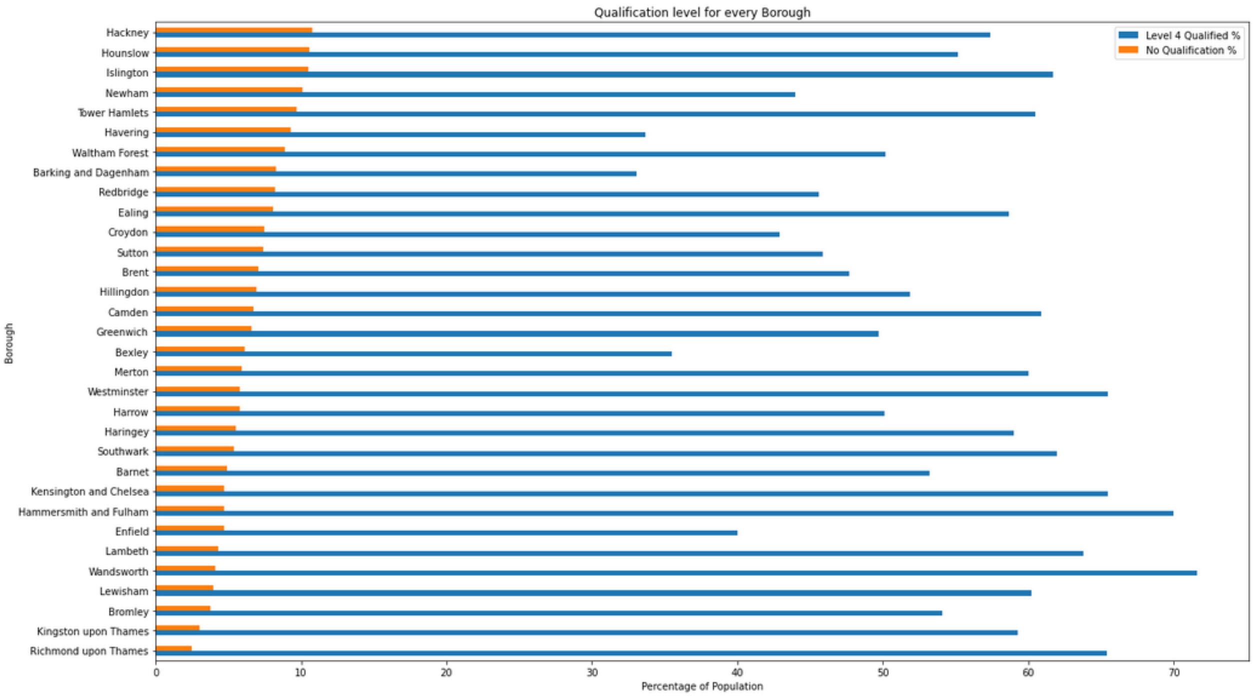


Figure 5.3. Horizontal bar chart of level 4 qualified talent and no qualified talent

According to the chart in figure 5.3, Richmond Upon Thames has the lowest percentage of the borough population that have no qualification. Furthermore, when viewing the level 4 qualified bar in blue of Richmond Upon Thames they have one of the highest percentages of the borough population having a level 4 qualification. This may suggest that educational facilities are at a high standard. However, on the other end of the spectrum the London borough of Hackney has the highest rate of their population with no qualifications, but surprisingly do not have a low rate of level 4 qualified talent. With Barking and Dagenham having a high no qualification percentage of the borough having no qualifications and having a low level 4 qualified percentage, this may suggest that Barking and Dagenham have a low standard of educational facilities.

For the purposes of the analysis being efficient to read and understand a data frame with level 1,2,3 and 4 was created. Using the level 1,2,3 and 4 qualification percentages, an average of the population that was qualified was created in a new column. Using the binning technique, I placed each borough into categorical bins based on the percentage of the borough population that was qualified.

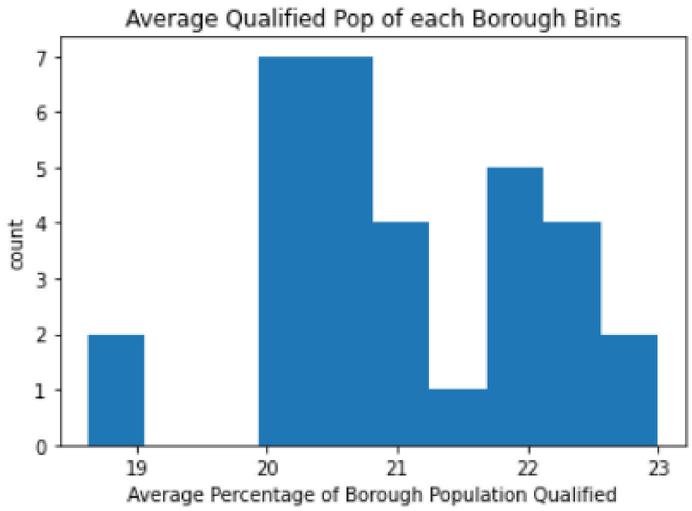


Figure 5.4. Histogram of the average percentage of borough population qualified.

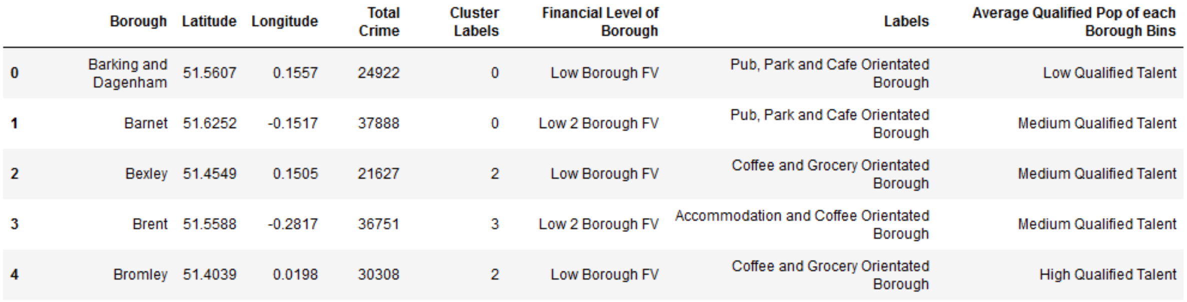
Using the histogram, I developed three bin categories, these are as follows:

- Low qualified talent: < 20%
- Medium Qualified talent: 20%-21%
- High Qualified talent >21%

Based on the average percentage of qualified talent in the borough, each category was assigned to a borough and merged with the main data frame.

3.5. Creating a New Data Frame

Once the results were taken from each component of the data, the results were merged into a new data frame.



	Borough	Latitude	Longitude	Total Crime	Cluster Labels	Financial Level of Borough	Labels	Average Qualified Pop of each Borough Bins
0	Barking and Dagenham	51.5607	0.1557	24922	0	Low Borough FV	Pub, Park and Cafe Orientated Borough	Low Qualified Talent
1	Barnet	51.6252	-0.1517	37888	0	Low 2 Borough FV	Pub, Park and Cafe Orientated Borough	Medium Qualified Talent
2	Bexley	51.4549	0.1505	21627	2	Low Borough FV	Coffee and Grocery Orientated Borough	Medium Qualified Talent
3	Brent	51.5588	-0.2817	36751	3	Low 2 Borough FV	Accommodation and Coffee Orientated Borough	Medium Qualified Talent
4	Bromley	51.4039	0.0198	30308	2	Low Borough FV	Coffee and Grocery Orientated Borough	High Qualified Talent

Figure 6. new data frame created to be enabled on a map.

4. **Results**

Using the results from the data frame, the information will be mapped on a map using the python library folium.

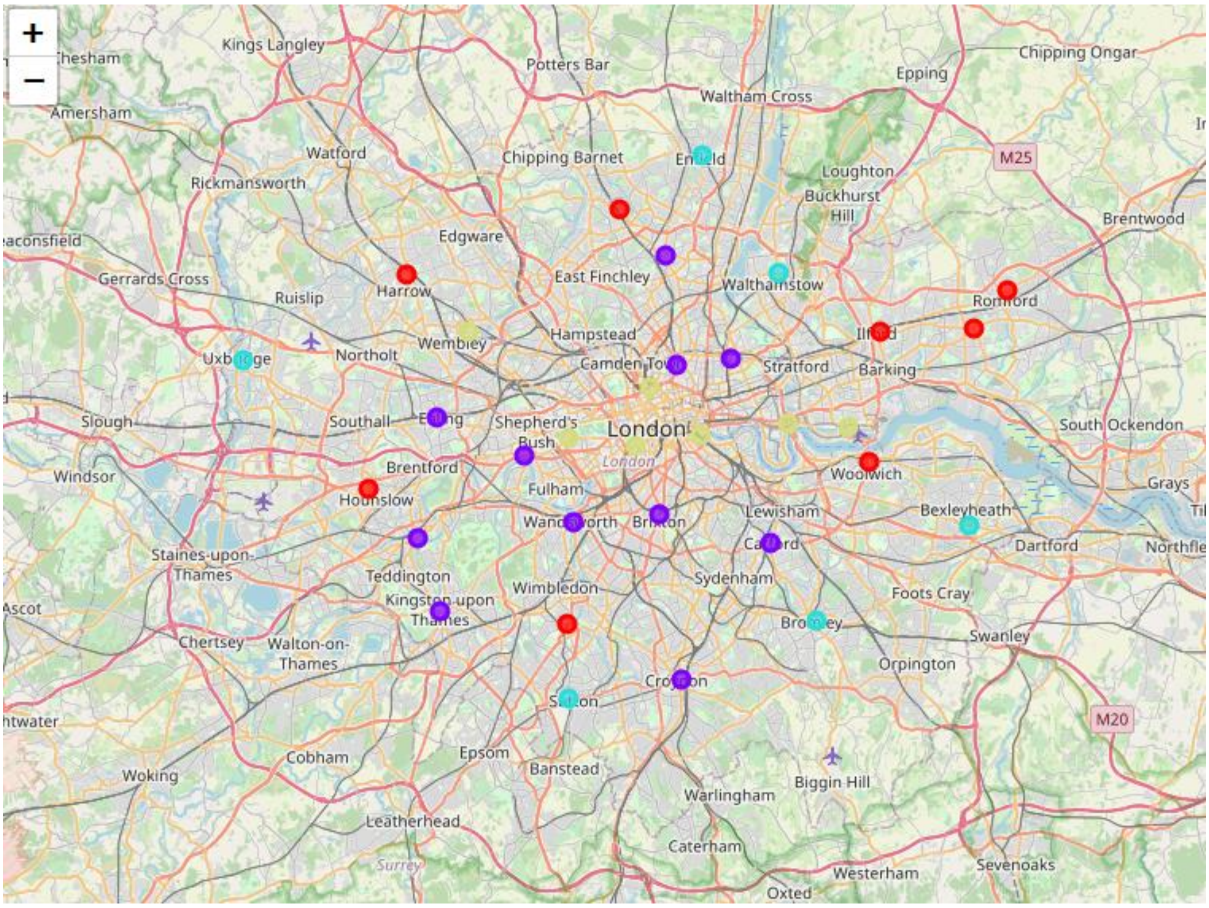


Figure 6. Map of London with the clusters superimposed on the map.

Each pin that is superimposed on the map of London has a particular colour from red, purple, turquoise and yellow. These colours represent the cluster and the type of venues the borough predominantly has. However, the audience would not understand how much crime is in each borough. Therefore, a choropleth map would need to be created to indicate the level of crime in each borough.

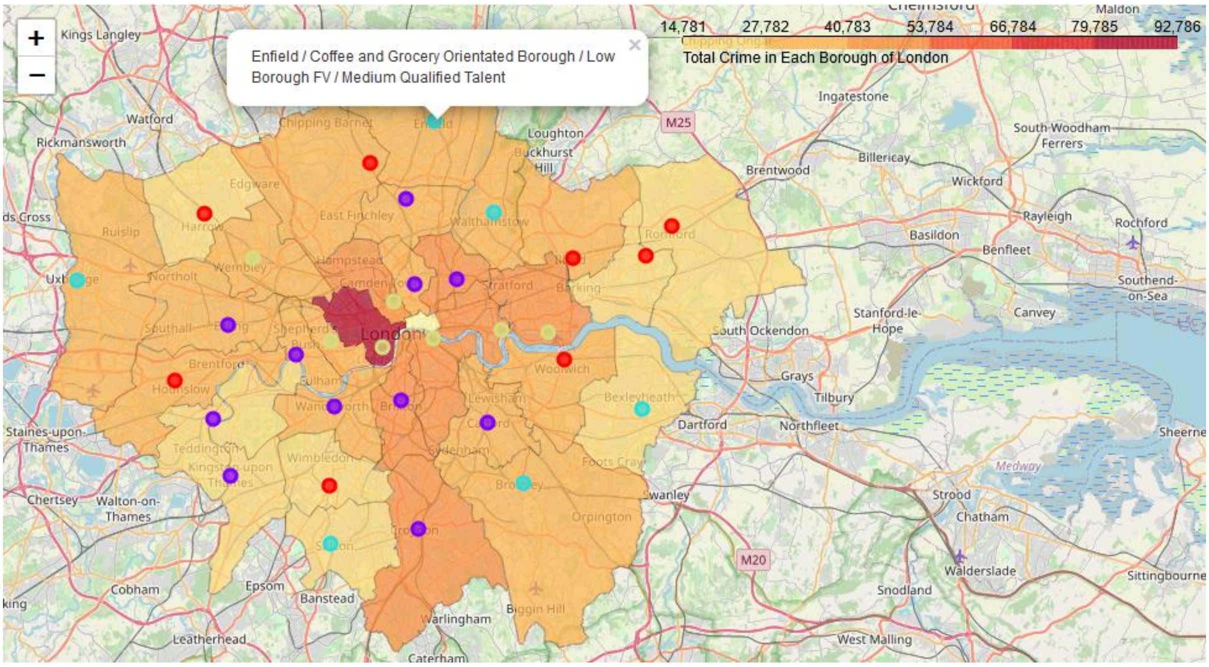


Figure 6.1. Choropleth map of London

In figure 6.1, the choropleth map of London is displayed with the result. Each pinpoint of the borough will provide the information about the borough's financial value toward the entity that is planning to locate to a specific borough, the most predominate venues that can be found in the borough, the level of qualified talent the borough has to offer and the borough shaded can give an indication on the level of crime in the borough.

5. Discussion

Over the analysis the observation that was most intriguing were the borough of Westminster having the highest crime rate in the whole city, however comprising of highly qualified individuals and has a high financial value. Further analysis would be needed to delve deeper into why this borough, having all the attributes it does has the highest level of crime. Furthermore, when considering crime rates, it seems that the outer borough of London is ridden with less crime than other boroughs. When considering the cluster venues, the outer borough of London has more parks, pub, and cafes than the inner boroughs, which may suggest that there is more open space due to parks and a less hectic environment which normally comes with big cities. As for the inner boroughs of London most of them are purple which indicates that they mostly have pubs. This may cater to the office work force in the inner-city areas, however more data is needed to come to that conclusion. Going closer to central London the boroughs mostly have hotels which may suggest that they try to cater for those visiting the central part of the capital. Lastly on the outskirts the turquoise pinpoint indicates that there are more coffee and grocery shops in that area, which could suggest that there may be a lower population in those boroughs and low level of visitors. However more data would be needed to come to that conclusion.

There are many short comings of the data, for instance extracting information on specific neighbourhoods of each borough would delve in deep into the level of crime and where the most predominant venues are in that borough. In addition, discovering the foot traffic of Westminster would also help understand the level of crime in that borough. The data gathered and analysed was for the purpose of the study to cluster the venue data and map the results on the map of London.

6. Conclusion

Using this analysis will help families settle into neighbourhoods that have a predominance of parks for children, also be safe in know which boroughs have high crime rates, increase the chances of the children of the family receiving a secure education and knowing whether a family will be able to settle into the borough without breaking the bank to live there.
Using this analysis, a further analysis can be built on top and can aid local authorities in funding projects to decrease crime, fund educational facilities and provide incentives to families or individuals to move to specific boroughs.