Predicting London Rental Locations

A. De Rose, August 2020

1 Introduction

1.1 Background

London is a vibrant city bustling with life. In 2019, London hit an approximate population of 9 million. With a limit on building space and such a large volume of people in the city, there is no wonder why house prices are sky high. For many individuals in the city, purchasing property is off the books and as such, they have to resort to renting. There are a variety of different factors that come into play when deciding the perfect location. When the list of determining factors increases in size, it makes that decision more complicated. With lots of other stressors in life, wouldn't it be brilliant if we could get a machine to do all the hard work for us?

1.2 The problem

In this project, the goal is to determine a list of boroughs in London which satisfy the following list of criteria:

- Reasonably safe to live in (a big city is never going to be perfectly safe!).
- Affordable for a possible first time renter.
- Provides a selection of essential public services close by.

A renter can then use this information to take a more guided approach to search for the perfect rental.

1.3 Target Audience

The core audience for this information is new renters who have not lived in London previously. However, this analysis will be applicable to, and hopefully useful for, any individual who is looking to rent in London.

2 The Data

2.1 Requirements

In order to satisfy the three criteria outlined above there will be a variety of data sets required. These sets of data will stand as a basis for the full analysis in this report. The focus will be on

data sourced in 2019, such that all data sets are fully populated and are not impacted by COVID-19. The list of data sets that will be required are detailed below:

1. Reported crime by London borough.
2. Estimated population by London borough.
3. Average monthly rent by London borough.
4. Public service venues by London borough.

2.2 Sources

These sets of data will be sourced from a variety of locations.

1. **Crime**: Reported by the Metropolitan Police, provided by the London Datastore.
2. **Population**: Thomas Brinkhoff: City Population, http://www.citypopulation.de.
3. **Rent**: Reported by the Valuation Office Agency, provided by the London Datastore.
4. **Public service venues**: Provided by the FourSquare places API.

3 Methodology

3.1 Data preparation

Each of the data sources detailed above provide data in different formats. As such, a slightly different procedure was required to gather them into manageable data structures.

3.1.1 Collection

1. The crime data was the easiest to deal with, given that it was provided in a csv format. Upon loading the data set, it is noted that there are a total of 27 features and 1,569 records. The features included were; the type of crime committed, the borough it was committed in, and the number of crime committed per month for the last two years.
2. The population data was provided in a table on a html webpage. As such, a simple web scrape was required. This data set contains 7 features and 34 records. The table contained the location name, the location type (borough / city), and 5 different population estimates for various years including 2019.
3. The rent data, also provided by the London Datastore, has an xls file format. When loaded, this data set had 9 features and 6,160 records. The features contained in this set were the borough, various date features, and a selection of descriptive statistics.

4. Finally, the venue data was provided through an API call via the FourSquare places API. The call yielded a JSON response. After filtering the JSON response, the resulting data set contained 5 features and 2,361 records. After sifting through the JSON response, this data set contained features such as the borough, venue name and coordinates, and the venue type.

3.1.2 Cleaning and feature selection

Once each data source was loaded up into a pandas dataframe, it became clear that there were a range of issues present in the data sets that had to be corrected.

Two years' worth of data was provided in the crime data set, and this was broken down by month. A new 'Total crimes' column was developed from the sum of crimes committed during 2019. All features, excluding the borough and total number of crimes, were discarded. Furthermore, the type of crime committed was also detailed, in this analysis the assumption is that the renter would deem any type of crime an issue, as such, the sum of each type of crime that occurred in each borough was utilised. Finally, there was a borough titled 'London Heathrow and London City Airports' which was excluded from the data set, given that living in or close to an airport wouldn't be ideal.

Borough	Total Crimes
Richmond upon Thames	12932
Kingston upon Thames	13216
Sutton	13951
Merton	14513
Harrow	17365

Figure 1. Total Crimes by Borough (2019)

The population data was scraped from the CityPopulation website in a relatively manageable format. During the scraping process, only the borough and 2019 population per borough was carried through into the dataframe. After this process, both the City of London and Greater London records were discarded as they were not present in the crimes data set. After a minor renaming of the Westminster borough, both the crime and population data sets were merged into one, detailing the number of crimes and population for each borough during 2019. An additional column was included in this data set which contained the 'Crime Rate', calculated as the total number of crimes as a percentage of the population size during 2019.

Borough	Crime Rate
Westminster	31.75
Kensington and Chelsea	16.17
Camden	14.73
Hammersmith and Fulham	12.90
Southwark	12.56

Figure 2. Crime Rate by Borough (2019)

The rent data provided by the London Datastore was provided in an easy to manage format. After loading this up into a dataframe, it was required to filter to specific subsections of the initial data set. Firstly, the data set was cut down to 2019 data only. Secondly, the decision was made to isolate to one bedroom flats for brevity, however, the analysis can be adjusted to include more scenarios in the future given the data availability. While this data set provided a variety of interesting descriptive statistics for each borough, average (mean) rent was the core focus and as such, other columns were discarded. Subsequently, after casting the rent data from a string to an integer, this data set was merged with the crime rate data set.

	Average rent
Kensington and Chelsea	2062
Camden	1659
Hammersmith and Fulham	1454
Southwark	1419
Islington	1558

Figure 3. Average Rent by Borough (2019)

Finally, after loading the public service venues into a dataframe one hot encoding was used to help build an understanding of the frequency of each venue type for each borough. This was done by taking the mean frequency of each venue type per borough and the results were displayed in a new dataframe.

Borough	Afghan Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Austrian Restaurant	...	Vietnamese Restaurant	Warehouse Store	Waterfront
Barking and Dagenham	0.0	0.00	0.0	0.0	0.0	0.00	0.000000	0.00	0.0	...	0.0	0.00	0.0
Barnet	0.0	0.00	0.0	0.0	0.0	0.00	0.000000	0.00	0.0	...	0.0	0.00	0.0
Bexley	0.0	0.01	0.0	0.0	0.0	0.01	0.000000	0.01	0.0	...	0.0	0.00	0.0

Figure 4. Mean public service venue frequency by Borough

After doing this, it was also possible to identify the top 10 venues per borough. This information was also stored in a separate dataframe. In this analysis, the top 10 boroughs are focused on to avoid issues with boroughs having an uneven number of venues provided by the API call.

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
Barking and Dagenham	Park	Bus Stop	River	Grocery Store	Supermarket	Cosmetics Shop	Automotive Shop	Plaza	Recreation Center	Pub
Barnet	Pub	Coffee Shop	Grocery Store	Indian Restaurant	Movie Theater	Bookstore	Pharmacy	Pizza Place	Restaurant	Park
Bexley	Pub	Grocery Store	Coffee Shop	Supermarket	Park	Hotel	Gym / Fitness Center	Clothing Store	Pharmacy	Garden Center
Brent	Coffee Shop	Indian Restaurant	Park	Clothing Store	Hotel	Pub	Hookah Bar	Gym / Fitness Center	Café	Golf Course
Bromley	Clothing Store	Pub	Coffee Shop	Indian Restaurant	Café	Portuguese Restaurant	Gym / Fitness Center	Pizza Place	Burger Joint	Bar

Figure 5. Top 10 most common venues by Borough

3.2 Exploratory Data Analysis

3.2.1 Total crimes and crime rate

The decision was made to use the crime rate over the total number of crimes to reduce the impact borough population had on reality. A borough with a higher than average population will likely have a higher number of crimes due to the increased volume of people.

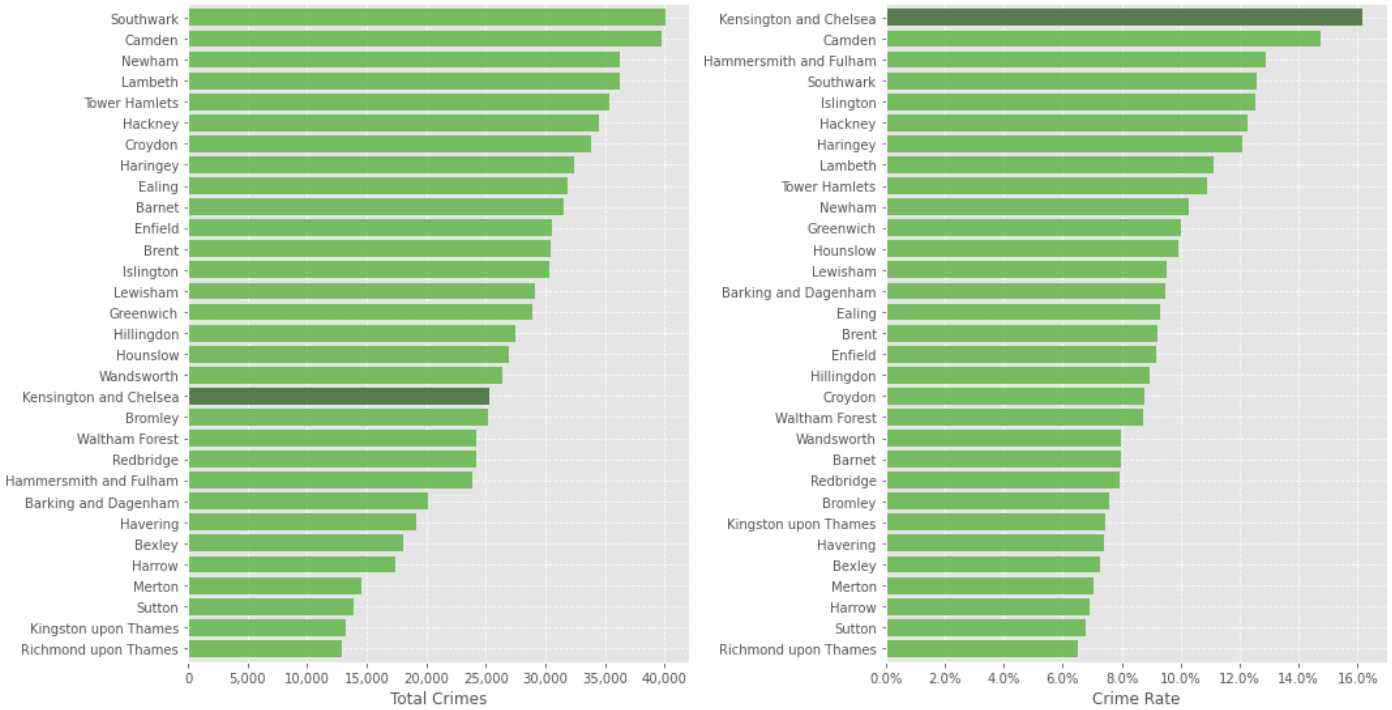

Figure 6. Total crimes by borough (left) and Crime rate by borough (right).

Without taking the total number of crimes as a proportion of population, there is a risk of suggesting a borough which, on the surface, seems to have a low number of crimes, but in reality has a large crime rate. A perfect example of this is Kensington and Chelsea, highlighted in the chart above. While the total volume of crimes is on the lower end, the borough tops the list in terms of crime rate.

3.2.2 Map of crime rate by Borough

Whilst looking at the crime data in a bar chart allows us to determine an order, it is relatively difficult to visualise in terms of location so is it useful to plot this data on a map. Fortunately, the Folium package allows us to do this relatively simply. By providing the Folium the borough boundaries in the form of a geoJSON file (this was sourced from Stuart Grange's GitHub), along with coordinates to plot the labels, a clean map image is developed and colour coded by crime rate for each borough respectively.

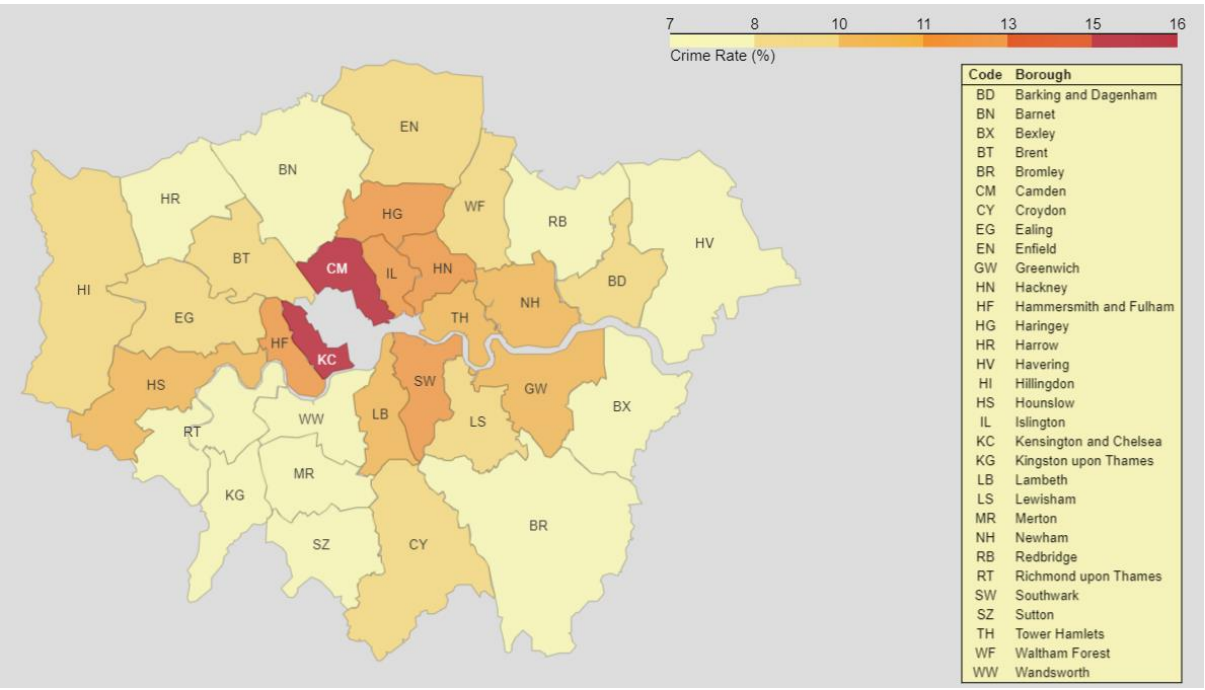

Figure 7. Crime rate by London Borough (2019)

By reviewing this map image, it is clear that generally the closer you are to central London, the more dangerous it becomes, however, there are pockets of safety such as Wandsworth, Lambeth and Tower Hamlets. The south west and far east are the safest areas to be, as the crime rate is on average below 8%.

3.2.3 Average monthly rent by borough

An extremely large range in the average rent can be observed, with Kensington and Chelsea topping the list at over £2,000 and Bexley trailing at just over £750 per month. There is an extremely large (approximately £400) difference between the first and second position which gives us insight into the wealth distribution through each London borough.

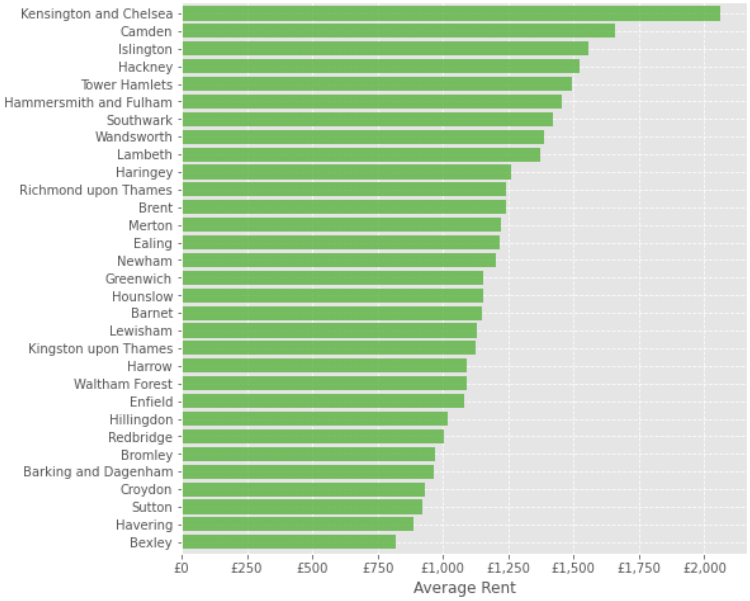

Figure 8. Average monthly rent by borough (2019)

3.2.4 Map of average monthly rent by borough

Similarly to the above scenario, it is much easier to process this information by viewing it on a map. Using a similar process to above, the rent data can be visualised through the Folium library.

Again, there is a similar reduction in average monthly rent as the distance to central London increases. In fact, the rent price of a borough is highly dependent on the travel time by public transport into central London.

It is also interesting to see the south west generally has a higher average monthly rent even though it is further away from the city. This may have a relationship with the lower crime rate in the area.

The boroughs to the far east of London have a much lower rent as transport links are relatively poor and travel times from other means of transport are extremely long.

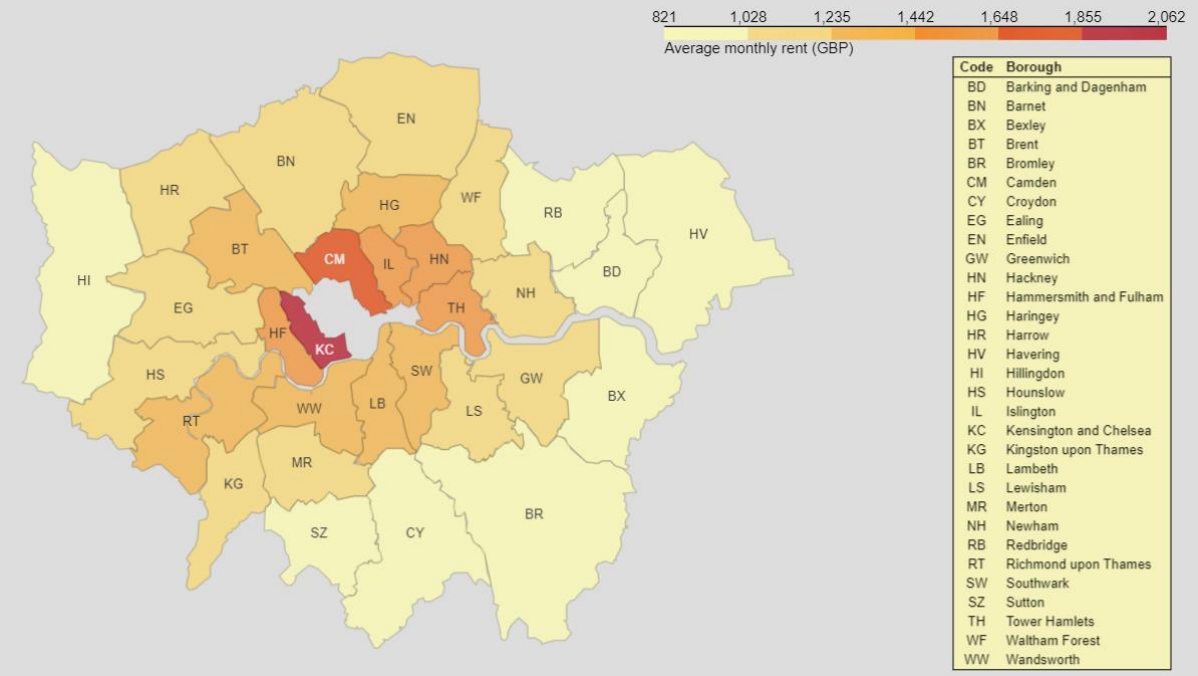

Figure 9. Average monthly rent by borough (2019)

3.3 Predictive Modelling

3.3.1 Feature selection for clustering

Using the data sets detailed above, a total of 265 features were selected for clustering. These include the calculated crime rate, the average rent per borough, and the mean frequency of all the venue types.

3.3.2 Algorithm selection

There are a number of both supervised and unsupervised models that could be used to group the boroughs, in this analysis, the k-means clustering algorithm was utilised to group the boroughs into a total of 6 clusters ($k = 6$).

Using the k-means clustering algorithm on the feature set developed above, each borough was categorised into its respective cluster based on its similarity to the features of other boroughs.

The results provided by the clustering algorithm were then grouped with the crime rate, average monthly rent and top 10 most common venues for each borough.

3.3.3 Quantifying the results

Whilst the crime rate and average rent are easy to evaluate, the top 10 most common venue types are part of a categorical data set and it is not easy to gather meaningful insights from a list of categories.

To gain some meaningful insights, a personal data set was developed to quantify the value of each public service venue. By exporting a list of the unique venue types, it was then possible to separate the “essential” public service venues from the “non-essential”. The logic used to separate these venue types was based on a blog article by Strutt & Parker. The key venues such as grocery shops, travel links, green spaces, gyms and churches were classified as essential, whilst others were not.

By using this data set, the percentage of the top 10 most common venue types that were classed as “essential” could be calculated. This column was added to the results dataframe and titled the “Essential Venue Proportion”.

	Crime Rate	Average rent	Cluster Labels	Essential Venue Proportion
Barking and Dagenham	9.48	965	1	60.0
Barnet	7.98	1147	4	40.0
Bexley	7.27	821	1	50.0
Brent	9.23	1241	0	30.0
Bromley	7.58	972	1	20.0

Figure 10. Clustered boroughs and their essential venue proportion

3.3.4 Assessing the results

As it was decided to use 6 different clusters, each clusters were assigned a colour such that they could be distinguished visually. The results of the clustering are detailed in the table below.

Cluster	Mean crime rate	Mean rent	Mean essential venue proportion
Blue	16.170000	2062.000	10.000000
Purple	12.600000	1558.500	17.500000
Green	11.140000	1407.750	12.500000
Yellow	9.076667	1230.000	26.666667
Cyan	8.715000	1121.625	33.750000
Red	8.011250	940.250	42.500000

Figure 11. Clustering results

Each cluster contained a varying number of boroughs, with only the blue cluster isolating Kensington and Chelsea as an individual. The joint largest clusters were Red and Cyan, both containing a total of 8 separate boroughs.

3.3.5 Visualising the results

The results in the above table can be presented visually using a simple bar chart as seen in the following figure. This allows for a clearer separation between the “good” and “bad” boroughs based on the criteria detailed in the introduction.

Figure 12. Clustering results in a visual format

Observing these charts provides a clearer picture for which cluster should be pursued in more depth during a rental search. For the criteria defined above, this would likely be the yellow, cyan or red clusters.

3.3.6 Visualising the clusters

It may be also useful to visualise these clusters on a map to understand their geography. Using the borough maps, the clusters can also be presented in their respective colours.

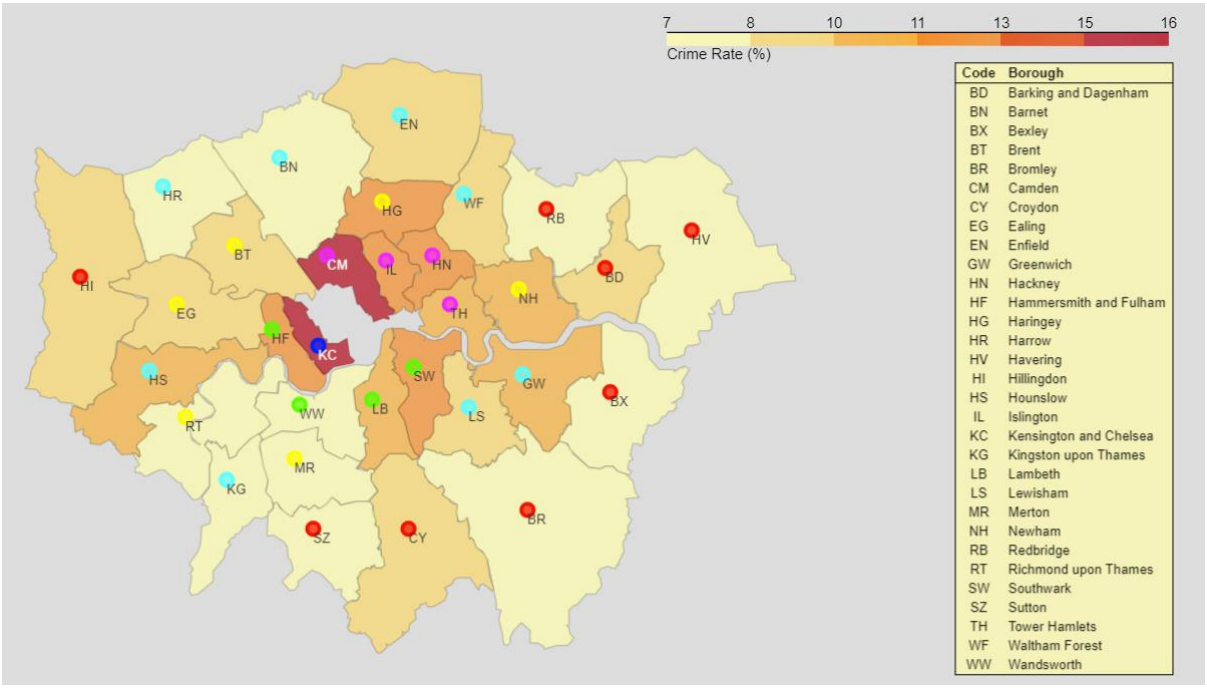

Figure 13. Borough clusters superimposed onto the Crime Rate map.

An observation from this chart is that the red and cyan clusters have mostly taken the boroughs around the edge of London. Given that distance to the centre was not a fundamental criterion of this analysis, the decision to classify these points as red and cyan was likely heavily weighted by the rental price in those areas, in other words, as they are further away from the centre, their rental price was lower.

4 Conclusion

In this report, the relationship between crime rate, monthly rental price and essential public service venue availability within each of the London boroughs was analysed. These key features were used within a k-means clustering model to group similar boroughs into 6 distinct clusters. This information can be used by an individual in the future to frame a plan on which boroughs to focus on when searching for a one-bedroom flat rental in London. The model helped identify specific groups of boroughs based on the following criteria:

- Reasonably safe to live in (a big city is never going to be perfectly safe!).
- Affordable for a possible first time renter.
- Provides a selection of essential public services close by.

By reviewing the results provided by the k-means model, the following clusters were isolated as the most appropriate given the initial criteria and would warrant a more in-depth search.

Cluster	Mean crime rate	Mean rent	Mean essential venue proportion
Yellow	9.076667	1230.000	26.666667
Cyan	8.715000	1121.625	33.750000
Red	8.011250	940.250	42.500000

Figure 14. Most appropriate clusters given the provided criteria.

This selection of clusters accounts for a sizeable portion of London so the individual can still experience the cultural variety available when attending viewings.

5 Evaluation

There were several assumptions made during the process of this analysis which are detailed below.

- The renter would not like to live in an airport.
- The renter would not like to live in Westminster due to the high crime rate.
- The renter is looking for a one bedroom flat.
- The essential venues the renter has align with those detailed in the report.

While these are logical, they must suit the user that is using this data to frame their search. However, this analysis can be adjusted quite simply to modify these assumptions to align with the end user.

There are some additional shortfalls in this analysis which must also be highlighted.

Firstly, doing this analysis at borough level is extremely broad, each borough of London is comprised of many smaller wards which each have their own microcosm. Limitations on the number of queries that could be made to the FourSquare API using a free account prevented a more granular analysis. Re-running this analysis at ward level would likely highlight hidden pockets within the boroughs that were excluded in the results section that would be habitable based on the criteria. Additionally, there are likely wards in the suggested clusters that do not adhere to the defined criteria.

Secondly, the borough's essential public service venue proportion was determined by the top 10 venues for each borough. One issue with this is that a whole borough's venue quality is determined by an extremely small percentage of its actual venues. The second issue is that there could have been less common essential venues available in the initial list provided by the API response. Similarly to the first shortfall, the number of venues that could be gathered in a single API call was limited to 100. However, the analysis could be updated to take into account all 100 venues for each borough, rather than focusing on the top 10, however, as discussed in the data collection section, the top 10 were isolated to in order to manage the impact of varying number of venues provided by the API call.

Within the modelling section, the k-means model was used and assumed as the best fit model for the provided features and records. Furthermore, the number of clusters $k = 6$ was also assumed to provide the best results. When replicating this analysis, this model should be used

in conjunction with other models and the results should be compared in order to find the best fit. Furthermore, the cluster number *k* should also be varied in order to understand the optimal number of clusters.

While the defined criteria in this report help paint part of the picture, they certainly do not provide definitive results. One fundamental criterion missing from this analysis is the distance from central London (by public transport) as alluded to multiple times in the report. The reason this was excluded was due to a lack of data availability, given that the borough is such a large area, travel times from one side of a borough compared to another vary greatly. If this analysis were to be repeated, the mean travel time across each ward within a borough could be used. This was decided against as the rent price fluctuated in line with the distance from central London. As such, rent price was used as a proxy for distance. Evidently, this is a significant assumption and should be reviewed in more depth.

6 Final point

There are a variety of improvements that can be made to improve the analysis of this problem, but this is my first sizeable project and it provides me with a basis to further develop my understanding of the data science landscape.

Thank you for reading this report! 😊