CLUSTERING STUDY LONDON

Investigating Best Rental Location for College Student

Contents

1 Introduction

This investigation will look into the clustering of neighborhoods in London to find the optimum location to live in for a college student.

1.1 Background

Finding a place for an appropriate accommodation is a very difficult decision for many university students. Although the distance from the University place a significant role other factors need to be considered to suit the students lifestyle. As somebody starting university after my gap year I decided to explore a possible university that I applied to and the optimum location to get housing.

1.2 Problem

I have decided to investigate the problem of finding the best location for a college student to rent an apartment. The following investigation will be for a college student that is looking to go to Imperial College London and will investigate the following features as the criteria:

- Social Criteria
  - Crime Rate (Safety)
  - Distance from University
  - Venues:
    - Gym
    - Nightclubs
    - Supermarkets
    - Public Transport Stops
- Economic Criteria:
  - Average Price of Rental

Source: https://unsplash.com/photos/iP8ElEhqHeY

Based on the venues I will create a set of clusters within the city and then preform analysis to determined which locations fit using the features of "average price of rental", "crime rate" and "Distance from University".

2 Data Collection

The following section will explain the methedology as well as the data sources.

2.1 Methodology

The data collected will be based on the different boroughs in london. This data set will be manipulated so that it can be used in conjunction with the foursquare api to find the geolocations of Gyms, Nightclubs and Supermarkets to determine the centroids of the clusters which will be used as the optimum location to rent an apartment.

2.2 Data Sources

In this section I will discuss where I will be taking my data from and where the data what type of data will be collected for the investigation.

The data for the boroughs will be found on the following website:

- https://en.wikipedia.org/wiki/List_of_London_boroughs

The following data will be parsed from the website:

- Borough:
  - Name
  - Latitude
  - Longitude
- Venue
  - Name
  - Latitude
  - Longitude

Following this the features "average price of rental", "crime rate" and "Distance from the University" will be used to determine which cluster is the best. The data for the following can be found here:

- Distance from the University: Google Maps
- Crime Rate: https://www.finder.com/uk/london-crime-statistics
- Average Price of Rental: https://www.mylondon.news/news/property/ london-rent-prices-cheapest-borough-17525488

After finding the best cluster the results and methodology will be evaluated and the validity of the investigation provided.

2.3 Feature Selection

The features where selected based on factors that I would prefer in the location of an accommodation. Although this would not accurately represent the average university student I decided to undertake this project to give me a better understanding of how I could find an optimum location. Hence for me important factors include venues such as Gyms, Nightclubs and Supermarkets alongside factors such as distance from the university, crime rate as well as the average price for rental. Personally the distance from the university has the highest significance hence I would like to investigate how it affects Venue Number, Crime and Rental cost.

2.4 Data Preparation

Parsing the Data on wikipedia required to read the document using pandas and remove any unwanted artifacts and dropping excess data. Leading to the following data table:

	Borough	Latitude	Longitude
0	Barking and Dagenham	51.5607	0.1557
1	Barnet	51.6252	-0.1517
2	Bexley	51.4549	0.1505
3	Brent	51.5588	-0.2817
4	Bromley	51.4039	0.0198

Figure 1: Table Showing Boroughs with Geocordinates in London

Next the data for the distance, rent and crime is added. The distance was calculated based on the euclidean distance and the data for crime and rent was parsed through two data sources *finder.com* and *mylondon.news* respectively. Next the data is normalized using the scikit-learn normalize function:

	Borough	Latitude	Longitude	Rent	Crime	Distance
0	Barking and Dagenham	51.5607	0.1557	0.597858	0.287882	0.913043
1	Barnet	51.6252	-0.1517	0.638754	0.433556	0.347826
2	Bexley	51.4549	0.1505	0.574489	0.248944	0.891304
3	Brent	51.5588	-0.2817	0.702045	0.431029	0.334239
4	Bromley	51.4039	0.0198	0.601753	0.350556	0.586957

Figure 2: Final Data Frame for the Parsed Data

Next using the foursquare api the venue locations are retrieved and a new data frame is created which will be used for clustering:

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude
0	Barking and Dagenham	51.5607	0.1557	Y's Gym	51.565280	0.184793
1	Barking and Dagenham	51.5607	0.1557	The Gym London Chadwell Heath	51.567393	0.117979
2	Barnet	51.6252	-0.1517	Pumping Iron Fitness Gym	51.617391	-0.143078
3	Barnet	51.6252	-0.1517	Nuffield Health Fitness & Wellbeing Gym	51.613354	-0.147670
4	Bexley	51.4549	0.1505	Better Gym Bexleyheath	51.458329	0.132591

Figure 3: Data Retrieved from Foursquare API

Using this Data the Clusters of the Boroughs can be found however first the influence of distance on the factors crime and rent will be investigated.

3 Exploratory Data Analysis

Through this section the factor of interest namely distance from the university will be investigated and how it affects the factor crime and rent.

3.1 Borough Location

Using the geolocation data parsed from the wikipedia table, the boroughs in London can be plotted using the folium library. This gives a good visualisation of how spread out the boroughs are within London:

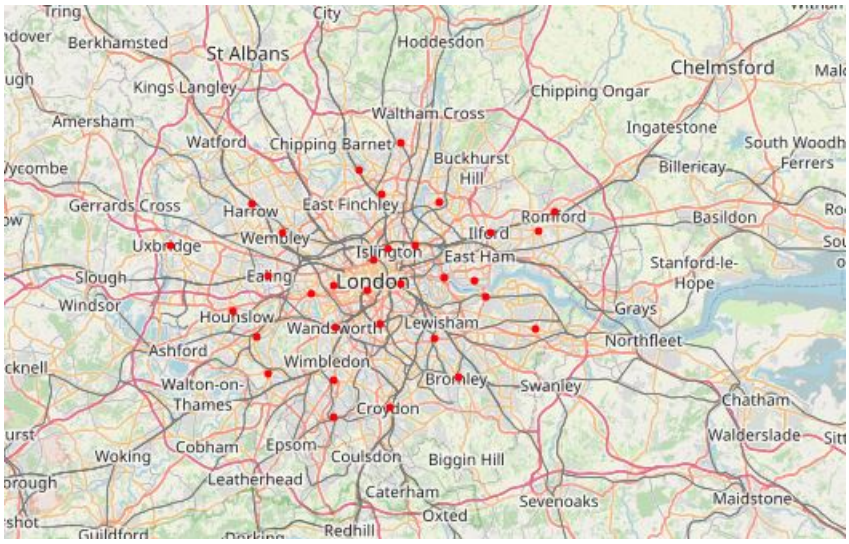

Figure 4: Locations of Boroughs in London

From the positions of the boroughs it can be seen that they are spaced out throughout London. These will be used as the the center points to find the venues within a give radius which in this case will be 500m. Now for each of the Boroughs I will investigate how their distance from the university affect the Rent and Crime.

3.2 Distance from University and Crime

The data that was previously collected and normalized from *finder.com* is now plotted using the regplot function of Seaborn to produce the following graph:

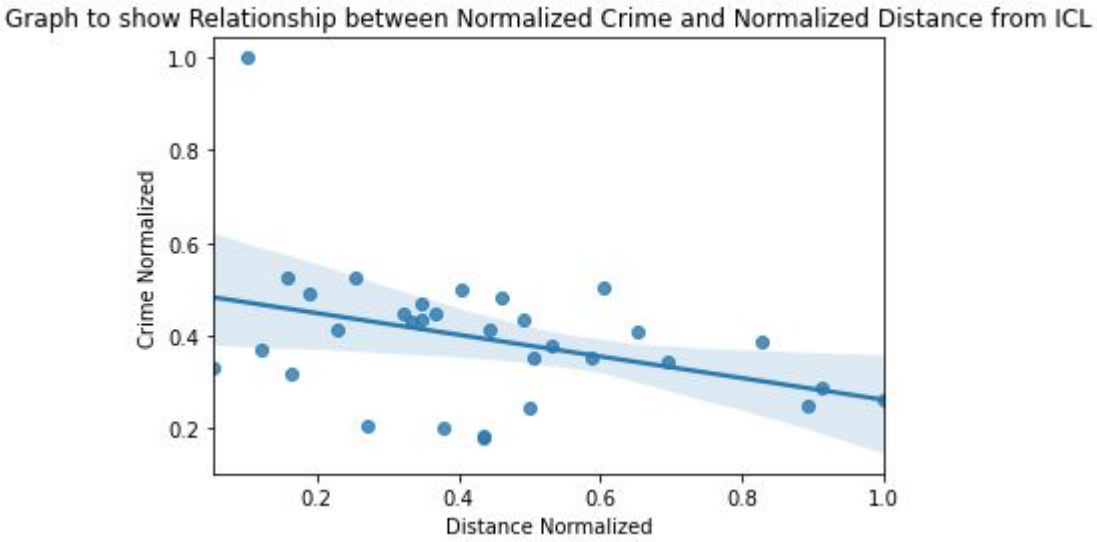

Figure 5: Graph: Normalized Distance vs Normalized Crime

From the graph it can be seen that as the distance increases crime decreases. Although there are some outliers the correlation is relatively strong as a distinct relationship can be observed. This suggests that in order to achieve lower crime the cluster would have to be in further proximity from the university. Although the distance from the university is the most important factor to minimize in this investigation, crime decreases with distance hence the optimum trade off point needs to be found. This require to find a model which finds a cluster that is both in close proximity and low crime.

3.3 Distance from University and Rental

Next the distance from the University and the Rental prices are investigated using the data collected from *mylondon.news* to produce the following graph busing the Seaborn regplot function:

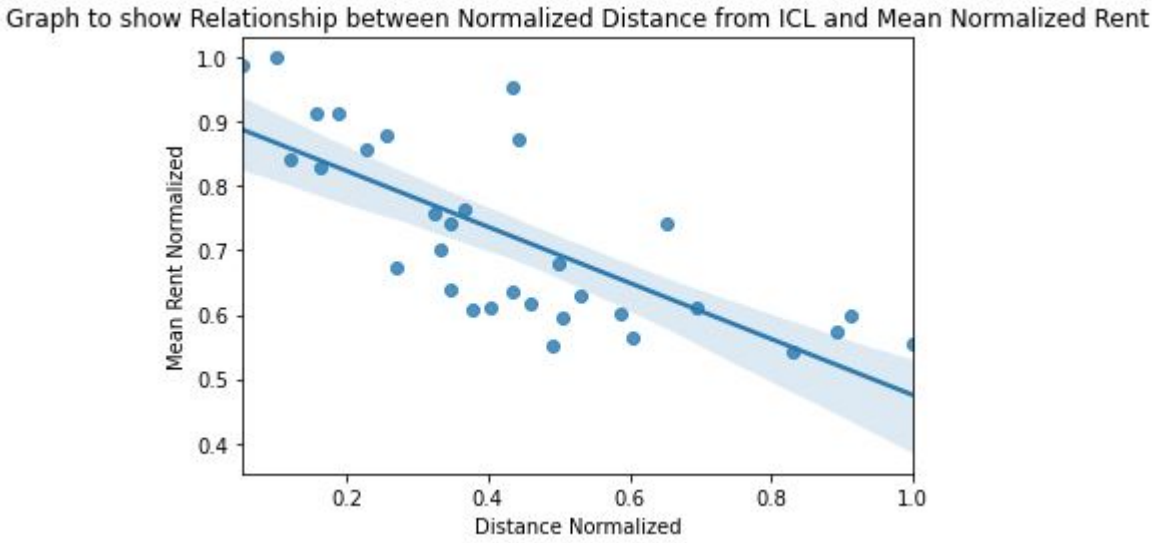

Figure 6: Graph Normalized Distance vs Mean Normalized Distance

It is possible to tell from the graph that as the distance increases the rent significantly decreases. This relationship is much stronger in correlation and has a much steeper gradient. The trade off between distance from the university and the rent is a lot more significant than between distance and crime, however distance from the university still remains an important factor hence having to be still considered. Therefore again a trade off point needs to be found between distance and rent which will be investigated based on the location of the clusters.

4 Modeling

For this investigation the most appropriate model would be k-means clustering as the question at hand is investigating a task containing unlabeled data which needs to be grouped based on venue density.

4.1 Clustering

The model will be found based on the data table found in Figure 3 using the K-means clustering function found within the scikit learn library. However first the optimum model needs to be found to do this the curve for the error and the cluster amount is plotted for 1 to 9 clusters:

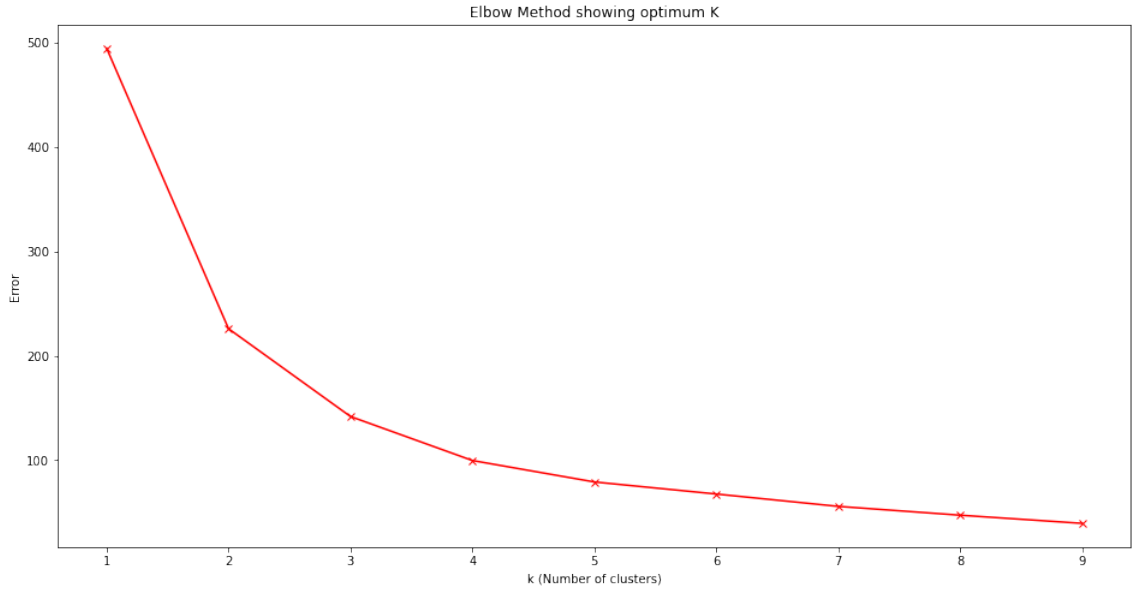

Figure 7: Finding the elbow point for K-means Clustering

From the graph it can be seen that the elbow point is found at $K = 5$ as the improvement diminishes significantly within the model past 5 clusters. Larger amount of clusters will result in the model picking up too much noise leading to high variance.

Since the best model was found to be for $K = 5$ the model is trained using that K value to produce a map of the following clusters:

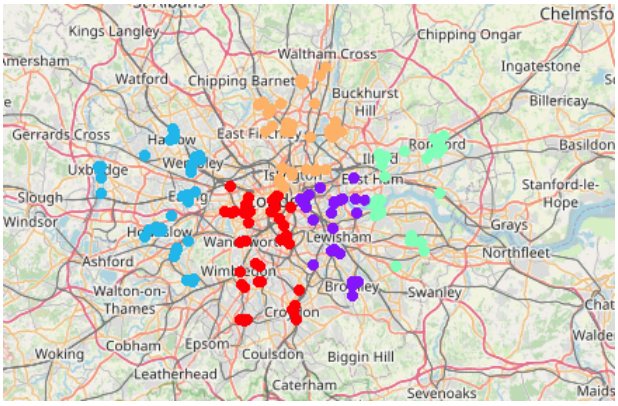

Figure 8: Clustering of Venues in London based on the Investigation Criteria

From the map it is possible to tell that the clusters are formed around the center where the university is located but there are some clusters that are much further away. Since the distance from the university is a such an important factor in deciding the accommodation in this investigation the optimum cluster needs to be found based on the distance alongside the factors of rent and crime.

4.2 Evaluating Clusters

Using the table in Figure 2 in combination with the clusters found for the venues the following dataframe is created to show the mean distance from the university, crime and rental cost vary in each cluster:

cluster	Rent	Crime	Distance
0	0.810635	0.430681	0.211838
1	0.733155	0.448975	0.443071
2	0.700841	0.319511	0.492737
3	0.608025	0.325869	0.813448
4	0.702104	0.434828	0.361469

Figure 9: Table showing mean distance from the university, crime and rental cost across clusters

This data is then plotted in a bar chart alongside the average values for each to find the optimum cluster:

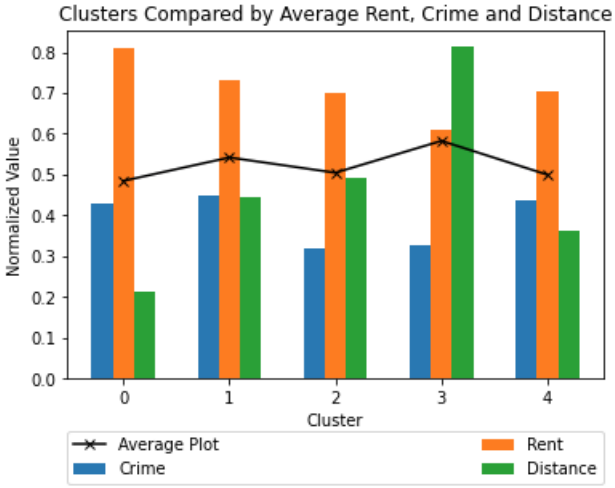

Figure 10: Graph showing the comparison of clusters based on average rent, crime and distance in london

From the graph it can be seen that it depends on which factor is most important to you if distance is important but money isnt an issue the cluster 0 is the best. On the other hand if crime is most important cluster 2 would be the best. And for the lowest rent cluster 3 would be the best. To find the overall best cluster the results will be averaged this can be seen based on the average plot the 0th cluster is the best interms of all criteria however the rent there is the heighest. To improve this investigation and ammend it for different needs the factors would need to be considered with different weightings.

5 Conclusion

Concluding the investigation the results will be discussed as well as the take away from the investigation. At the same time the validity and accuracy of the results will be evaluated based on the methodology used.

5.1 Evaluation

The methodology allowed to yield satisfactory results as from the investigation it was possible to find the best cluster for renting an accommodation, however there are still opportunities in improving the investigation. First of all only one type of model was chosen to find the clusters by using DBSCAN or Hierarchical Clustering the results could have been different maybe leading to higher accuracy in the clustering, The amount of features chosen was quite small and was only based on the preferences of a single person. In the case of another investigation doing some market research through survey can help determine the most important features for aspiring students to find their ideal accommodation. This can be done by personalizing each model for each student by weighting it based on their preferences or finding a model that can encompass the whole student body. The former however would be a better choice as the then the personally tailored model would allow for a more accurate representation of what the student actually wants. This was the case for this investigation as I based the features of factors that are important to me but could have done more background research to determine other possible factors. The data collected was recent and up to date making this investigation relevant as of the current year (2020). The strengths of this investigation are that the features where normalized in order to have equal weighting to minimize the bias towards any one variable and the result of the investigation provides a good estimate for the location to get an accommodation. Overall the investigation was successful however increasing the feature number as well as testing different models could have lead to more deep and accurate investigation,

5.2 Conclusion

Based on the investigation the best cluster is found in south London. This can be shown visually by the following map:

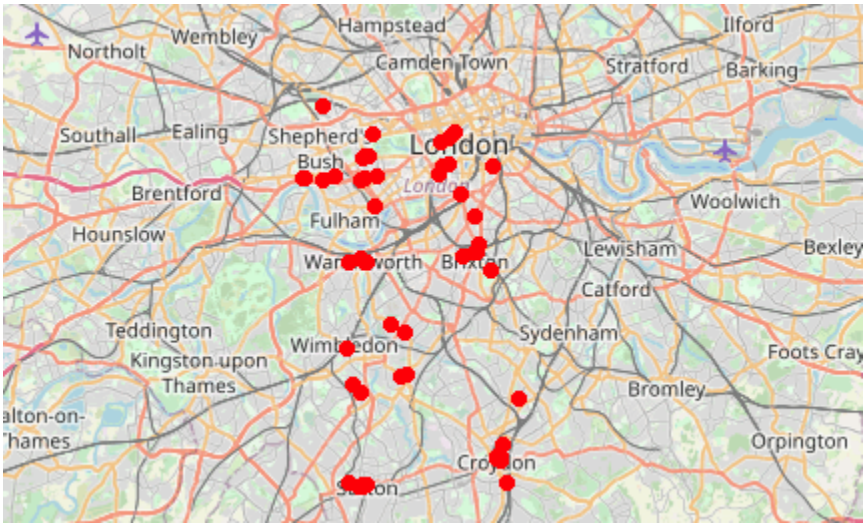

Figure 11: Map of optimum cluster

From this it can be seen that the intracluster distance is actually quite large. However it is obvious that there are some obvious cluster in the center would be the optimal in terms of distance around Brixton and Shepherds Bush. The other smaller clusters are found much more further away. To improve this investigation another kmeans could be run to find the clusters within this cluster as an extension to this investigation. This shows that although the results of this investigation where successful the final clusters had quite large intracluster distance leading to large error.

In conclusion the best location for rental would be shepherd bush or brixton based on distance and crime factors.