The Battle of Neighborhoods

1.Introduction

1. a) Background

London is the capital and the most populous city of England with an estimated population of around 9 million in 2019. London is an international center of business, finance, arts, and culture which has made it an important destination for immigrants to England. Such a large population of immigrants from around the globe has also made London one of the most multicultural and cosmopolitan cities in the world.

Neighborhoods in London: Greater London is a ceremonial county of England that makes up most of the London region. This region forms the administrative boundaries of London and is organized into 33 local government districts—the 32 London boroughs and the City of London, which is located within the region but is separate from the county. [1]

1. b) Problem Description

London is one of the top destinations for international community in terms of education, jobs and living. Often, its huge size (607 square miles) becomes a problem for the potential property owners in terms of finding an optimal place to live that has great amenities and other types of venues such as gourmet fast food joints, pharmacies, parks, grad schools and so on. Furthermore, the real estate market in a metropolitan city like London is often very competitive and expensive.

1. c) Objective

The aim of this project is to identify and group together similar neighborhoods in London based on the amenities and top venues that exist in a neighborhood. Furthermore, we explore the potential correlation of those neighborhoods with their property rates. This information can be used to establish the association of venues distribution with property prices as well as providing a basis for more accurate property price predictions.

1. d) Target Audience

The findings of this article would be useful for potential property owners in determining a suitable place to live according to their lifestyle and budgeting needs. Moreover, this information can also be useful for people who must relocate in the city of London and are interested to move to similar neighborhoods based on the facilities and venues in their existing neighborhood.

2. Description of Data

The main sources of data used in this project are listed below:

1). Location Dataset: Dataset containing location coordinates of all regions by their postcodes was obtained from the following CSV file: https://www.doogal.co.uk/PostcodeDistrictsCSV.ashx

2). Venue Dataset: Dataset containing top venues for each neighborhood was obtained using Foursquare API: https://foursquare.com/

3). Property Prices Dataset: Data on property prices by postal codes is publicly published by the UK government and was obtained from the following website: https://www.zoopla.co.uk/

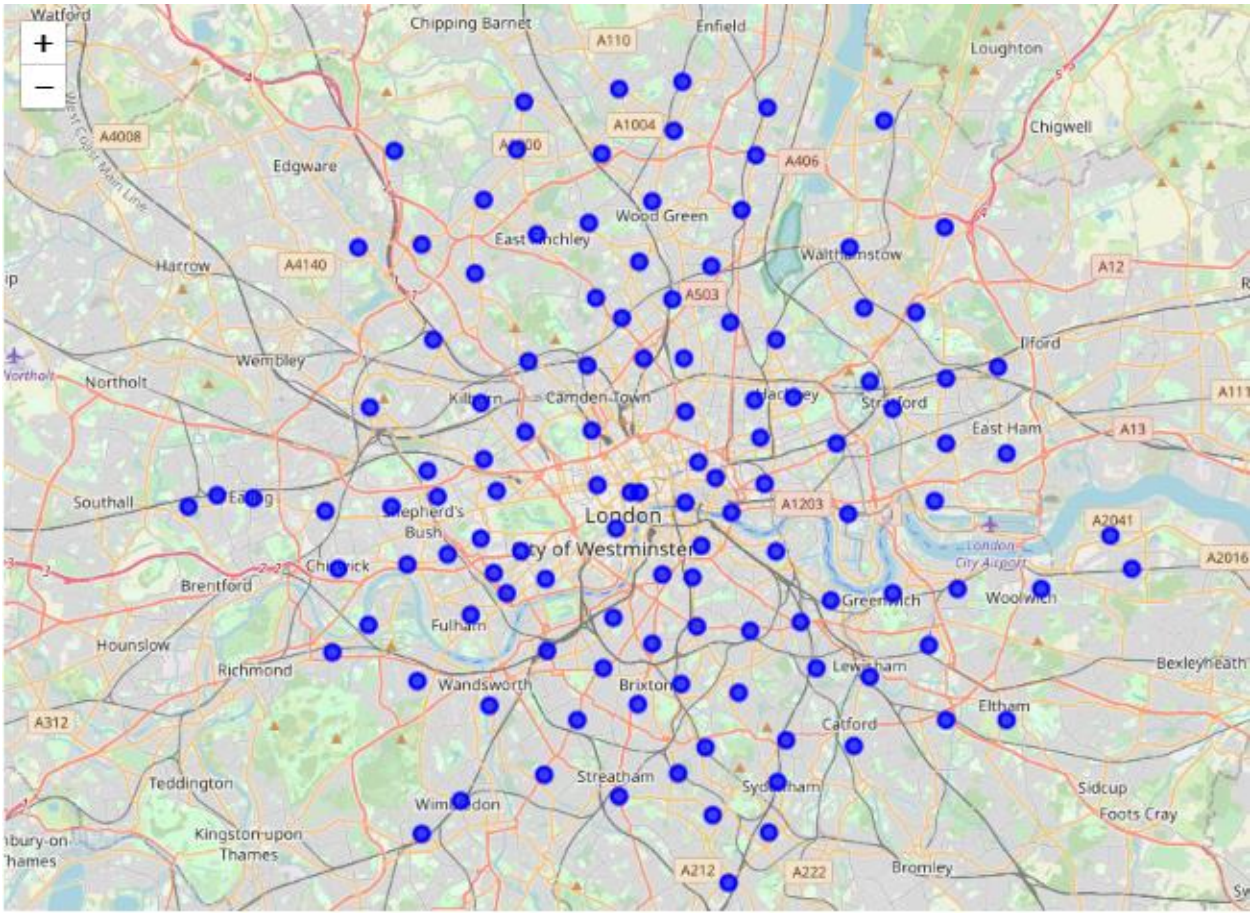
3. Methodology

3. a) Data Acquisition and Wrangling

The Dataset (CSV file) with location coordinates by postcodes for England was parsed and stored in a **pandas data-frame** for further analysis. Data wrangling was performed on the data in the following steps:

1. Data columns used for further analysis are 'Postcode', 'Region', 'Neighborhood', 'Latitude' and 'Longitude'. All irrelevant columns removed.
2. The dataset contains all postcodes of UK. Extracted only the postcodes in the London region that start with N, E, SE, SW and W.
3. Deleted non-Geographic postcodes.
4. Removed rows having missing items or zeros. Verified data types of input values for our model.

The resulting dataset contains 120 unique postcodes along with their neighborhoods and location coordinates. We then use the python **folium** library to visualize geographic details of London and its boroughs. I created a map of London with its postcodes superimposed on top using the latitude and longitude values to get the visual as below:



Next, we use the Foursquare API to explore the venues of all postcodes regions and segment them. We set the LIMIT parameter to 200, which would limit the number of venues returned by the Foursquare API and the radius of 1000 meters to maximize accuracy. Then we merge the resulting venues dataset with the postcodes dataset as shown in the figure below:

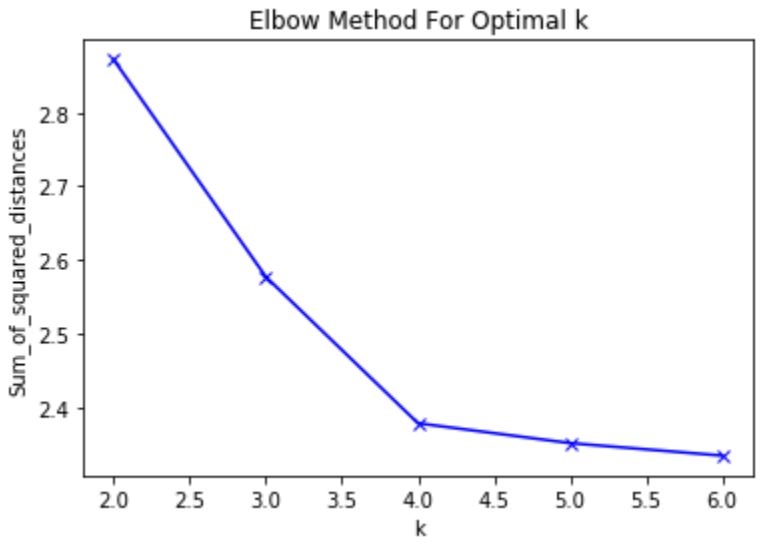
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barnsbury, Canonbury, Kings Cross, Islington, ...	51.5376	-0.0982429	The Life Centre Islington	51.538056	-0.099171	Yoga Studio
1	Barnsbury, Canonbury, Kings Cross, Islington, ...	51.5376	-0.0982429	Steve Hatt	51.538588	-0.099041	Fish Market
2	Barnsbury, Canonbury, Kings Cross, Islington, ...	51.5376	-0.0982429	Pophams	51.536666	-0.096175	Bakery
3	Barnsbury, Canonbury, Kings Cross, Islington, ...	51.5376	-0.0982429	The Bill Murray	51.536226	-0.098750	Pub
4	Barnsbury, Canonbury, Kings Cross, Islington, ...	51.5376	-0.0982429	Saponara	51.536875	-0.096121	Italian Restaurant

3. b) Data Modeling and Analysis

We use One Hot Encoding to group the data on postcodes districts and find out the top twenty venues present in each district.

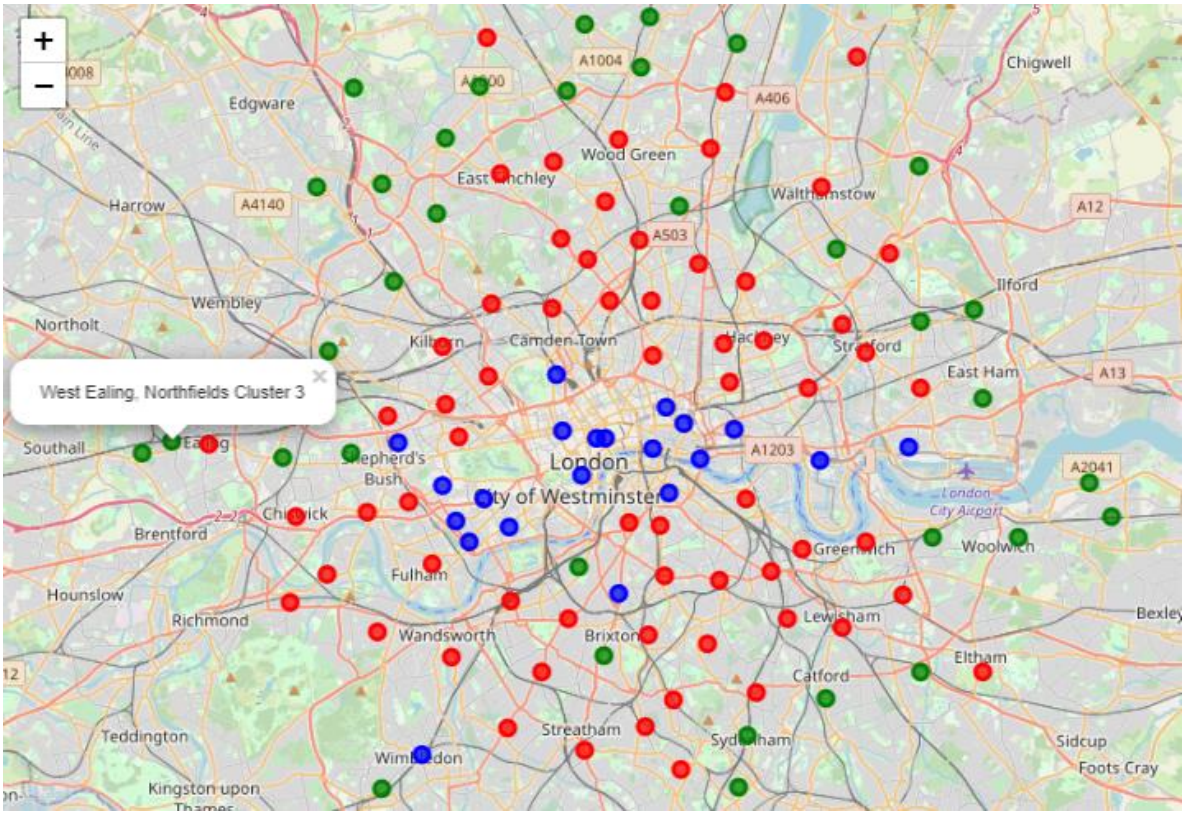
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abbey Wood, West Heath, Crossness, Thamesmead	Supermarket	Grocery Store	Historic Site	Convenience Store	Furniture / Home Store	Train Station	Pub	Platform	Coffee Shop	Exhibit
1	Acton, East Acton, Park Royal, West Acton	Pub	Gym / Fitness Center	Grocery Store	Train Station	Platform	Park	Fast Food Restaurant	Hotel	Café	Coffee Shop
2	Aldgate, Bishopsgate, Whitechapel, Shoreditch,...	Coffee Shop	Hotel	Pub	Indian Restaurant	Café	Bar	Pizza Place	Korean Restaurant	Art Gallery	Bakery
3	Anerley, Crystal Palace, Penge, Beckenham	Pub	Fast Food Restaurant	Grocery Store	Pizza Place	Park	Supermarket	Train Station	Coffee Shop	Tram Station	Track Stadium
4	Balham, Clapham South, Hyde Farm	Coffee Shop	Pub	Grocery Store	Bakery	Pizza Place	Supermarket	Indian Restaurant	Bar	Italian Restaurant	Breakfast Spot

Due to common venue categories, **K-means** algorithm is used to cluster the postcode regions in London. K-Means algorithm is one of the most common unsupervised clustering techniques and is ideal for this type of dataset. Elbow method is used to determine the optimum value of k. The graph below clearly shows the ideal value of **k = 4**:

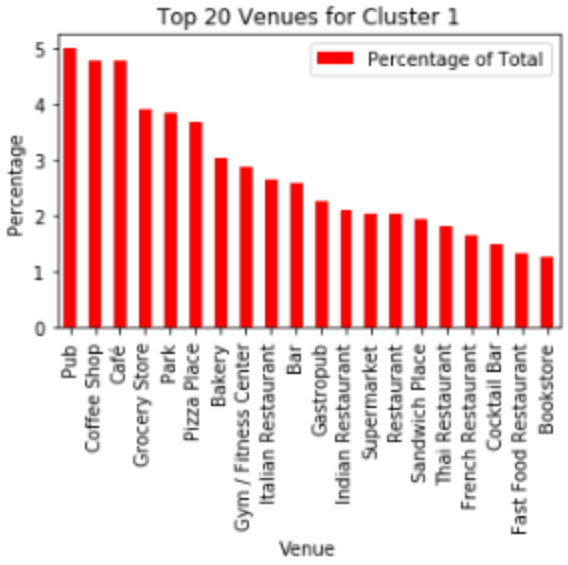


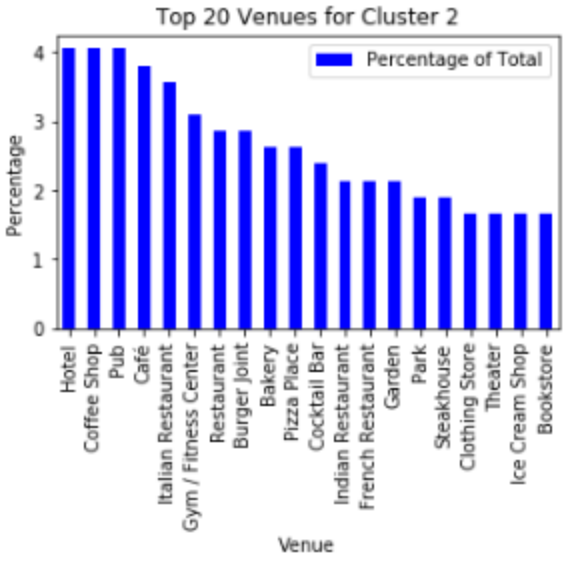
4. Results

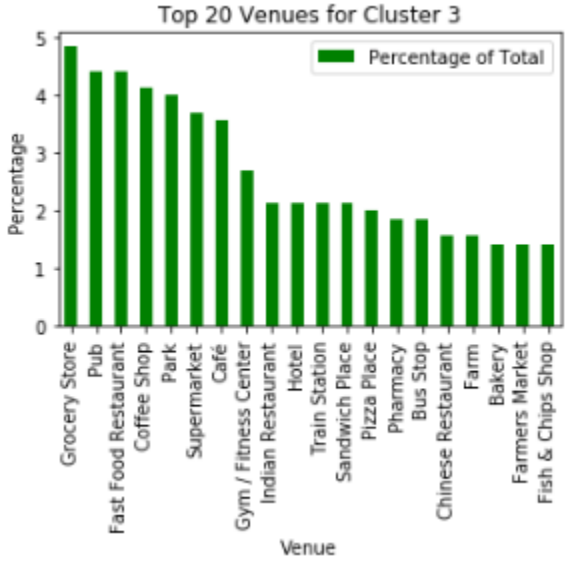
We use the matplotlib and folium packages to visualize the output clusters from the K-Means algorithm on a map of London. The clusters have been color coded to visualize their respective patterns. The map below clearly indicates a circular pattern in the resulting clusters. Moving outwards from central London we observe the clusters are in the following radial pattern: Cluster 2(blue), Cluster 1(red) and Cluster 3(green).



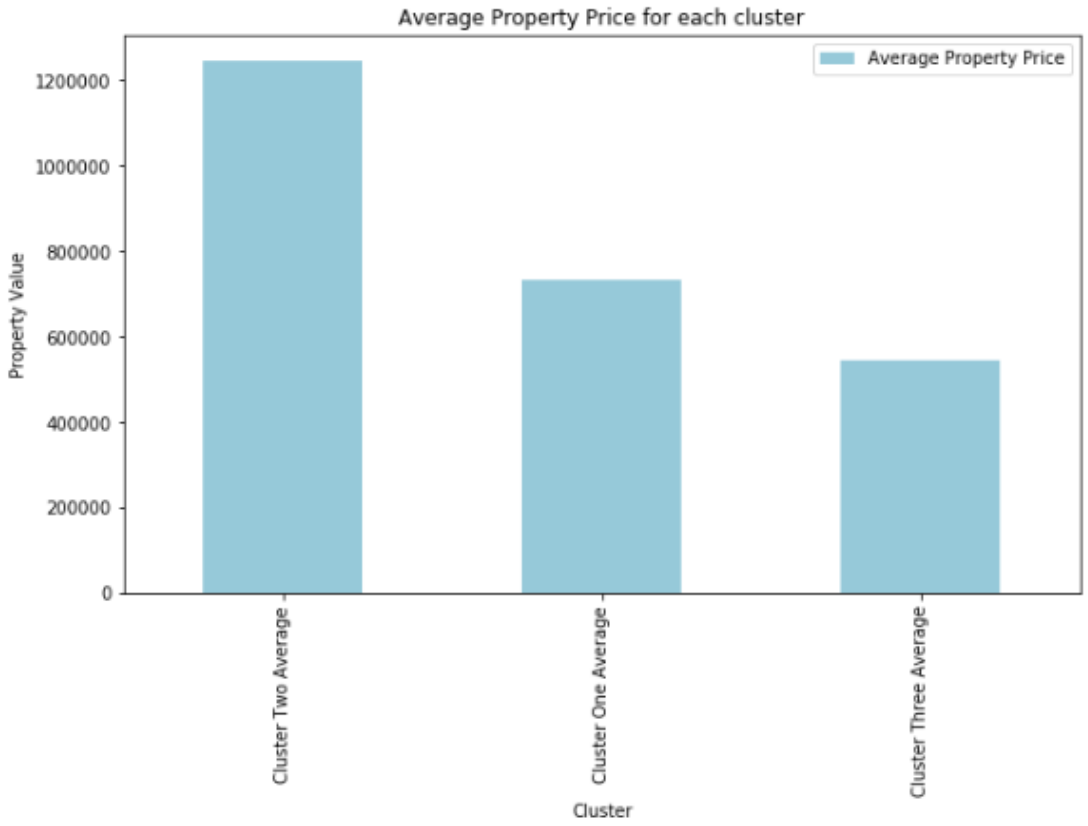
To further clarify our results, we have generated bar charts to visualize the distribution on the top 20 venues in each cluster by their percentage. If we combine both results, we can clearly perceive the difference in the choices of the venues by the residents of respective clusters.



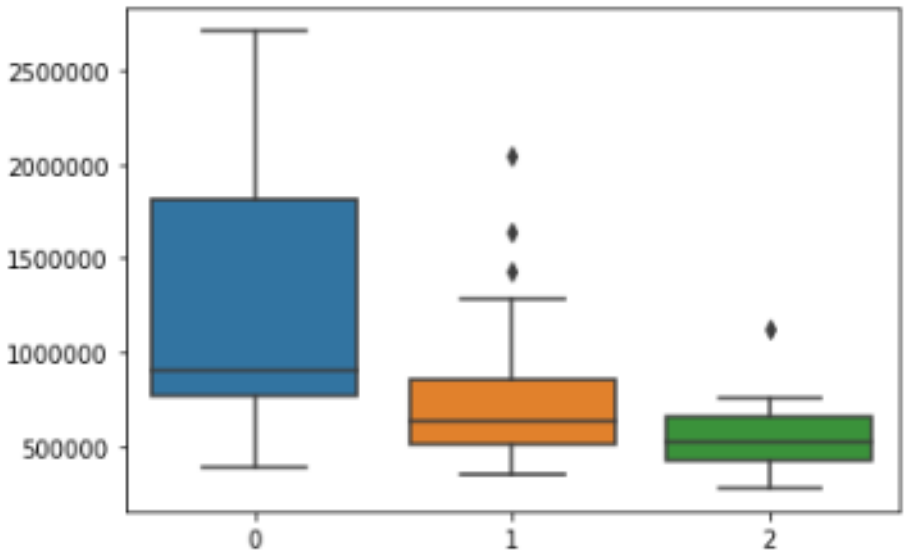




In objective section, one of the aims of this project was also to visualize the correlation of the similar neighborhoods with their property rates. Thus, a bar graph below is used to illustrate the trend of average property prices for each cluster. The property prices increase as we move towards central London.



To further explore the correlation, we compare box plots of each cluster based on their average property value. The figure below shows that, the **range** of property prices also decreases as we move away from central London. The second cluster has the highest mean, median and inter-quartile range followed by first and third cluster.



5. Discussion

London is a metropolitan city with a high population density in a narrow area. There are many factors that can affect the property rates and similarity in neighborhoods. In this study, we have only covered one aspect: venues. As there is such a complexity, very different approaches can be tried in clustering and classification studies.

Furthermore, this technique is only applicable to large-scale cities like London where there are a huge number of neighborhoods. The list of top 20 venues was generated by defining an arbitrary radius of 1000 units. However, the postcodes districts are all different sizes, and some are much smaller than the others. This might result in redundancy in our venues list and at the same time missing other venues. In order to obtain more accurate results, we need to define district specific radius for east postcode region while querying the list of venues.

This data can also be used to establish correlations beyond the scope of this project. For example, this data can be used to determine the best location to open a restaurant or other venue type in London or predict property prices in a neighborhood in London etc.

6. Conclusion

The data suggests that Cluster 2 (mostly lying in the Central London) has highest overall property value followed by Cluster 1, and Cluster 3. These findings show that there is a negative correlation between average property prices and the distance from center e.g. the property rates are highest in the central London. Furthermore, neighborhoods having similar venue distribution also have similar property prices.

7. References

1. contributors, Wikipedia. *Greater London*. 20 April 2020. <https://en.wikipedia.org/w/index.php?title=Greater_London&oldid=951105258>.