APPLIED DATA SCIENCE

# Exploring London Neighborhoods - Venues

## IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Maria-Lida Kounadi

May 2020

# Introduction

London city is a business hub. According to the Global Power City Index (GPCI)[1], London ranks first on its power to attract people, capital and enterprises from the world. London maintained its position for the 8th consecutive year.

During the daytime, especially in the morning and lunch hours, office areas provide huge opportunities for restaurants and coffee shops. Also, coworking spaces will become a need to provide people with a space to work and at the same time enjoy their coffee and lunch. People who use coworking spaces do not have to own privately leased office and they can use shared spaces for meetings, working environments that can provide them with the desired amenities. Their office can grow dynamically/organically as their teams are growing.

The importance of exploring capital expenditure with its potential return is widely recognised by investors and owners to accurately predict where the next location of their new restaurant/cafe/coworking space would be for an increased ROI (return over investment) and make the investment profitable.

## Business Problem

Using data science methodology and machine learning techniques like clustering, this project aims to help decision making for property developers to open a new cafe/restaurant/co-working space in London based on Foursquare data, to identify the best possible locations.

The problem statement below will be explored as part of this study

### Problem statement:

*“Exploring business idea for a new cafe/ restaurant/ co-working space in London through Foursquare data”*

[1] http://mori-m-foundation.or.jp/english/ius2/gpci2/index.shtml

## Target Audience

This project is particularly useful to property developers and investors, or anyone that would like to open a new cafe/restaurant/co-working space and understand which areas in London would be ideal for this. This investment can always be their side-business as well.

This project can be also useful to any Data Scientist that can use this methodology in order to explore the London neighbourhoods and/or use the Foursquare data.

Figure 1. London Map Art Print by Clair Rossiter

# Data

In this section of the report, the data that will be used to solve the problem and the source of the data is described.

## Description

The following data was used:

1. List of neighborhoods in London from Wikipedia
2. Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
3. GeoPy
4. Land value data to inform the decision making based on the capital investment required.
5. Foursquare data to explore boroughs and its venues, point of interests.
6. Map: https://joshuaboyd1.carto.com/tables/london_boroughs_proper/public

## Data sources

### Neighbourhoods

The wikipedia page: https://en.wikipedia.org/wiki/List_of_areas_of_London contains a list of neighbhours in London with a total of 32 London Boroughs. This study focuses only in the City of London.

### Location

The geographical coordinates were taken using ArcGis and geopy module.

### Land value

The property investment capital of the UK, London has a diverse array of housing and communities across a giant urban area. This site was used to understand the avg price in each borough https://propertydata.co.uk/cities/london

## Foursquare data

Foursquare API was used to get the most common venues of the given Boroughs/Locations of London.

## Maps

For the final visualisation this module was used Choropleth Maps | Python, and data from this location for the map: london boroughs proper

# Methodology

In this section discussion and description of any exploratory data analysis is done, any inferential statistical testing that is performed, if any, and what machine learnings were used.

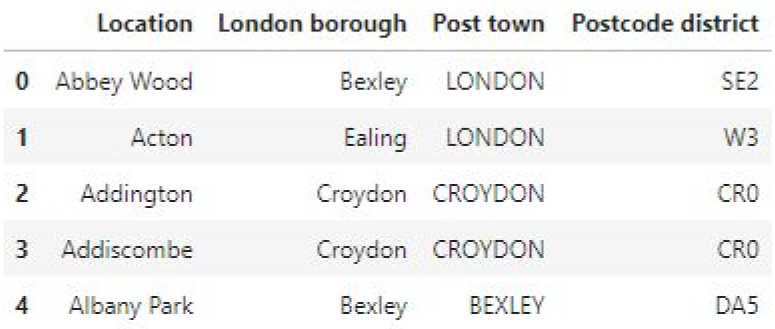
## Data preparation

### Scraping London Neighborhoods from wiki

The wikipedia page was scraped using Beautiful Soup package.

The data was cleaned. The dial code and os gird ref columns were not used so they were dropped. From the london borough column any footnotes and annotations were removed. Few postcodes and boroughs had multiple variables assigned to them so the first one was kept. (this was identified as an assumption). Similarly for the post towns.

The unique values then were checked to understand the different values.



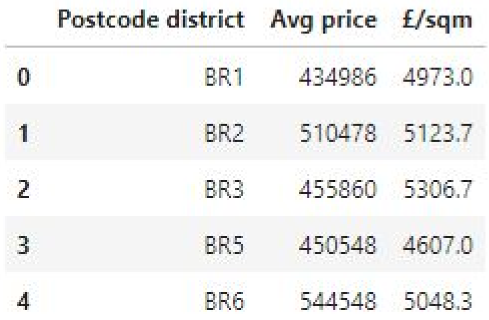
|  | Location | London borough | Post town | Postcode district |
|---|---|---|---|---|
| 0 | Abbey Wood | Bexley | LONDON | SE2 |
| 1 | Acton | Ealing | LONDON | W3 |
| 2 | Addington | Croydon | CROYDON | CR0 |
| 3 | Addiscombe | Croydon | CROYDON | CR0 |
| 4 | Albany Park | Bexley | BEXLEY | DA5 |

### Average Land Price

The Property data site was scraped this time using the pandas read html method.

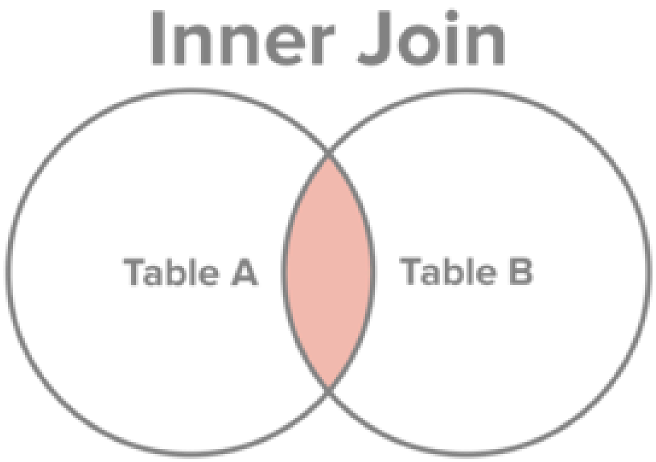
The data was cleaned. The “£” sign and the “,” were removed and the values of the avg price and price/sqft were updated to numeric. Also the price/sft was updated to metric systems using the following equation: sqft = 1/10.764 sqm and or £/sqft = 10.764 £/sqm.

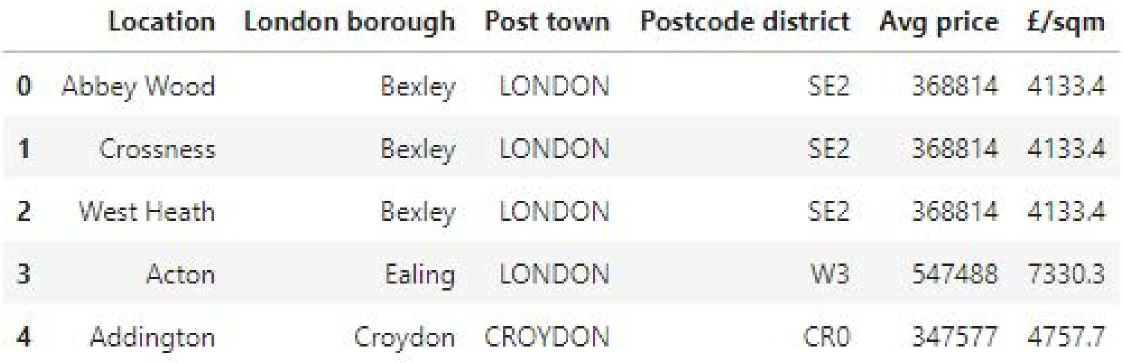
The index was reseted and the Area column name was renamed Postcode district to align with the Neighborhoods data.



| | Postcode district | Avg price | £/sqm |
|---|---|---|---|
| 0 | BR1 | 434986 | 4973.0 |
| 1 | BR2 | 510478 | 5123.7 |
| 2 | BR3 | 455860 | 5306.7 |
| 3 | BR5 | 450548 | 4607.0 |
| 4 | BR6 | 544548 | 5048.3 |

The two tables now are merged with inner join



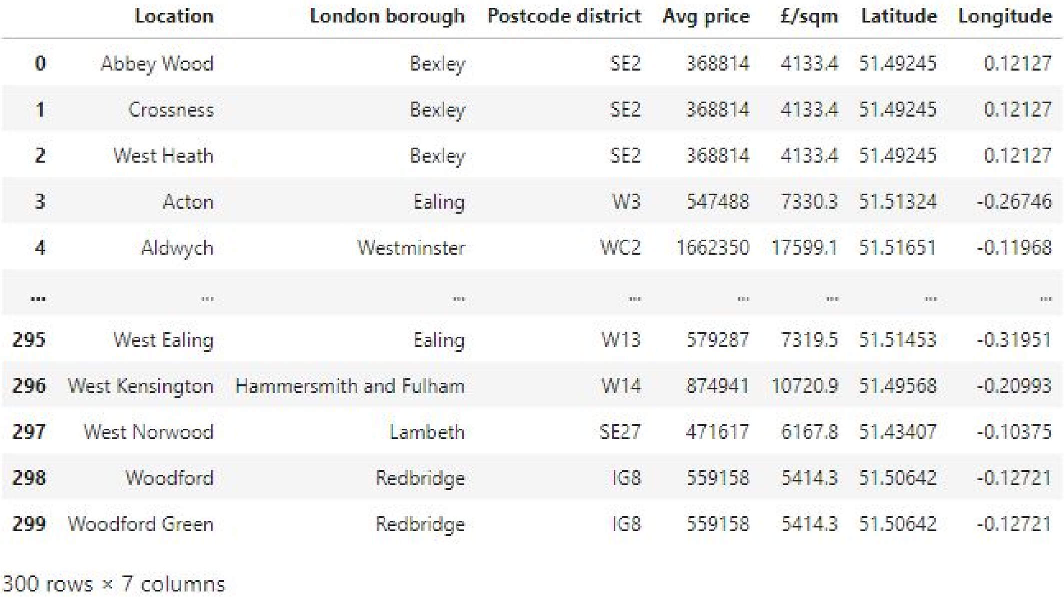


| | Location | London borough | Post town | Postcode district | Avg price | £/sqm |
|---|---|---|---|---|---|---|
| 0 | Abbey Wood | Bexley | LONDON | SE2 | 368814 | 4133.4 |
| 1 | Crossness | Bexley | LONDON | SE2 | 368814 | 4133.4 |
| 2 | West Heath | Bexley | LONDON | SE2 | 368814 | 4133.4 |
| 3 | Acton | Ealing | LONDON | W3 | 547488 | 7330.3 |
| 4 | Addington | Croydon | CROYDON | CR0 | 347577 | 4757.7 |

## Getting coordinates through GeoPy

This study aims to look at the London only post town. For this reason only the Post towns that were named London were kept.This results in 300 unique locations.

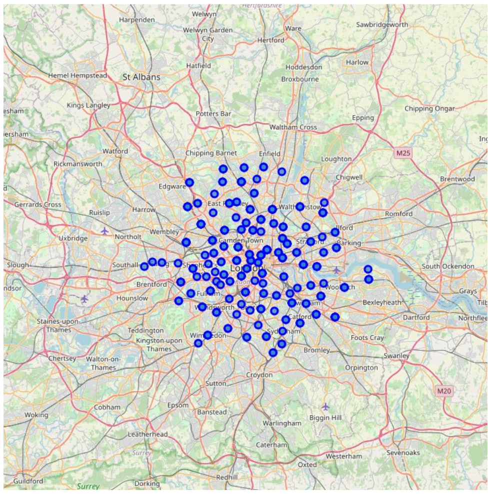
Then the function to get the data from the geocoder using ArcGis was used.  The Latitudes and Longitudes were added on the london_data dataframe based on the postcode district and was named post_london_data dataframe. The ‘Post town ‘



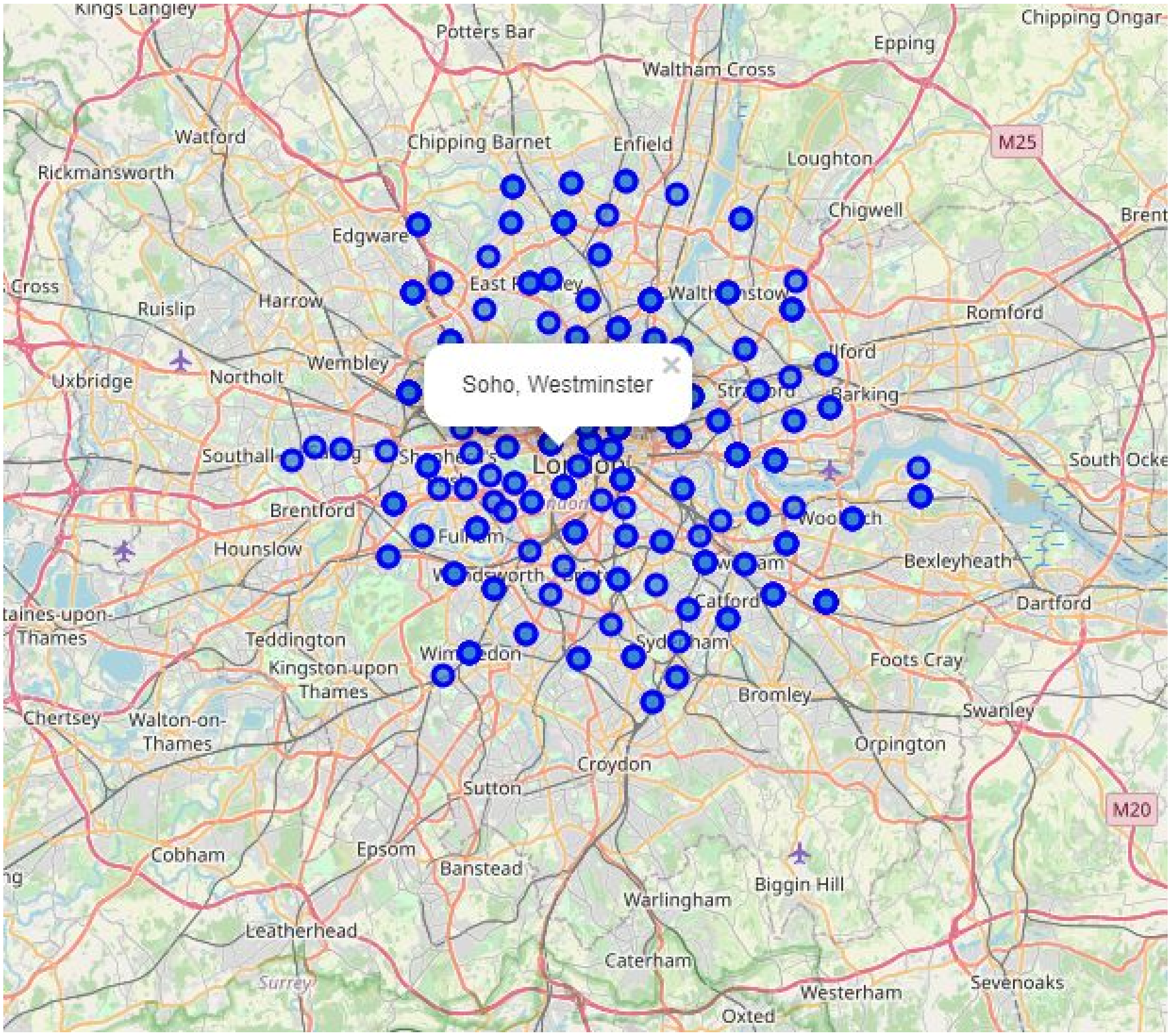
| | Location | London borough | Postcode district | Avg price | £/sqm | Latitude | Longitude |
|---|---|---|---|---|---|---|---|
| **0** | Abbey Wood | Bexley | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 |
| **1** | Crossness | Bexley | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 |
| **2** | West Heath | Bexley | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 |
| **3** | Acton | Ealing | W3 | 547488 | 7330.3 | 51.51324 | -0.26746 |
| **4** | Aldwych | Westminster | WC2 | 1662350 | 17599.1 | 51.51651 | -0.11968 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **295** | West Ealing | Ealing | W13 | 579287 | 7319.5 | 51.51453 | -0.31951 |
| **296** | West Kensington | Hammersmith and Fulham | W14 | 874941 | 10720.9 | 51.49568 | -0.20993 |
| **297** | West Norwood | Lambeth | SE27 | 471617 | 6167.8 | 51.43407 | -0.10375 |
| **298** | Woodford | Redbridge | IG8 | 559158 | 5414.3 | 51.50642 | -0.12721 |
| **299** | Woodford Green | Redbridge | IG8 | 559158 | 5414.3 | 51.50642 | -0.12721 |

300 rows × 7 columns

Also the geographical coordinates of London city was identified as 51.5073219 and -0.1276474.  A map of London was created.



Each circle has the info of the Location and London_borough



## Using Foursquare Location Data

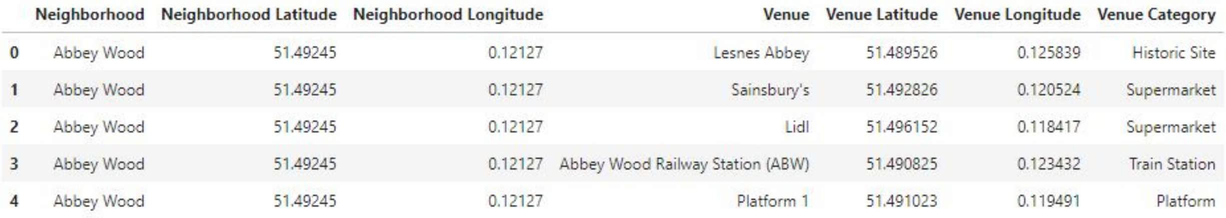
Initiallythe Foursquare credentials were defined along with the version of the API.

A function was defined to explore the venues around a location. A radius of 500 was used and a limit of 100 venues (the free API has a limit)

A url was created as below:

url=https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET,lat,lng,VERSION,radius,LIMIT)

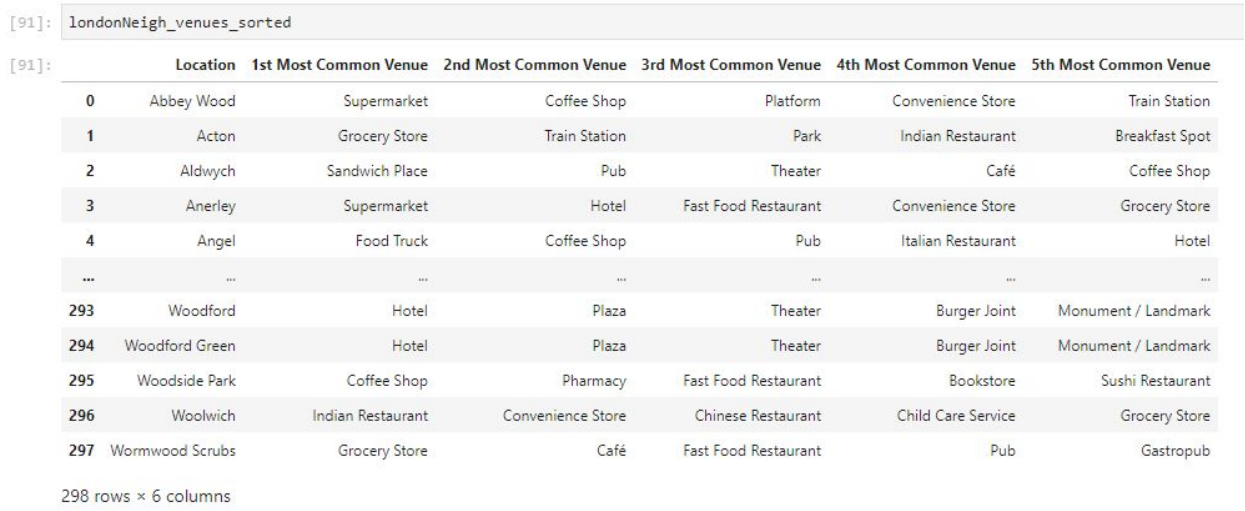
The function presented the 100 venues within the radius of 500 meter for each location from their given latitude and longitude information.



|  | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Abbey Wood | 51.49245 | 0.12127 | Lesnes Abbey | 51.489526 | 0.125839 | Historic Site |
| 1 | Abbey Wood | 51.49245 | 0.12127 | Sainsbury's | 51.492826 | 0.120524 | Supermarket |
| 2 | Abbey Wood | 51.49245 | 0.12127 | Lidl | 51.496152 | 0.118417 | Supermarket |
| 3 | Abbey Wood | 51.49245 | 0.12127 | Abbey Wood Railway Station (ABW) | 51.490825 | 0.123432 | Train Station |
| 4 | Abbey Wood | 51.49245 | 0.12127 | Platform 1 | 51.491023 | 0.119491 | Platform |

Above is a head table of the list Venues name, category, latitude and longitude information from Foursquare API.

Then, the 5 most frequent venues were identified for each location



```
[91]: londonNeigh_venues_sorted
```

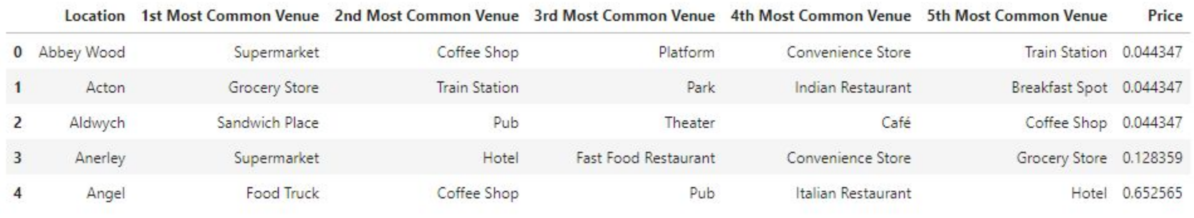
[91]:

|  | Location | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|
| 0 | Abbey Wood | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station |
| 1 | Acton | Grocery Store | Train Station | Park | Indian Restaurant | Breakfast Spot |
| 2 | Aldwych | Sandwich Place | Pub | Theater | Café | Coffee Shop |
| 3 | Anerley | Supermarket | Hotel | Fast Food Restaurant | Convenience Store | Grocery Store |
| 4 | Angel | Food Truck | Coffee Shop | Pub | Italian Restaurant | Hotel |
| ... | ... | ... | ... | ... | ... | ... |
| 293 | Woodford | Hotel | Plaza | Theater | Burger Joint | Monument / Landmark |
| 294 | Woodford Green | Hotel | Plaza | Theater | Burger Joint | Monument / Landmark |
| 295 | Woodside Park | Coffee Shop | Pharmacy | Fast Food Restaurant | Bookstore | Sushi Restaurant |
| 296 | Woolwich | Indian Restaurant | Convenience Store | Chinese Restaurant | Child Care Service | Grocery Store |
| 297 | Wormwood Scrubs | Grocery Store | Café | Fast Food Restaurant | Pub | Gastropub |

298 rows × 6 columns

## Results

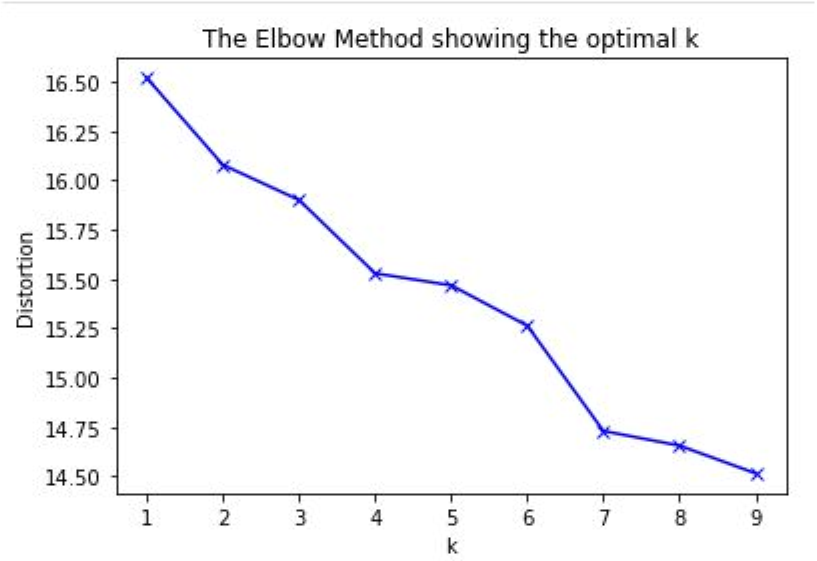
The average normalized  price is added.



| | Location | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | Price |
|---|---|---|---|---|---|---|---|
| 0 | Abbey Wood | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station | 0.044347 |
| 1 | Acton | Grocery Store | Train Station | Park | Indian Restaurant | Breakfast Spot | 0.044347 |
| 2 | Aldwych | Sandwich Place | Pub | Theater | Café | Coffee Shop | 0.044347 |
| 3 | Anerley | Supermarket | Hotel | Fast Food Restaurant | Convenience Store | Grocery Store | 0.128359 |
| 4 | Angel | Food Truck | Coffee Shop | Pub | Italian Restaurant | Hotel | 0.652565 |

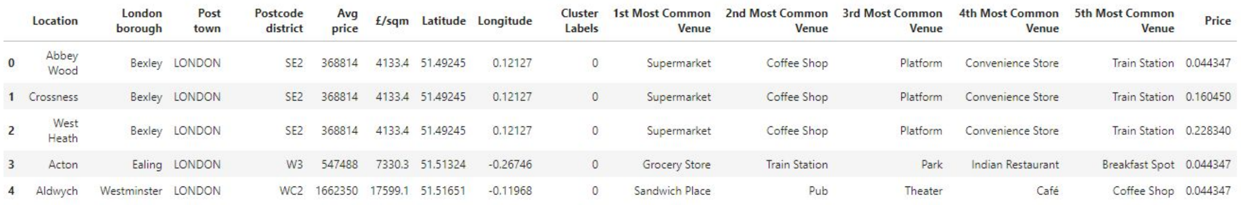
There are some common venue categories in locations and unsupervised machine learning K-means algorithm is used to cluster the locations.

First the dataset was normalised using Standard Scaler method.  Then Elbow method and silhouette score was used to identify the best k.



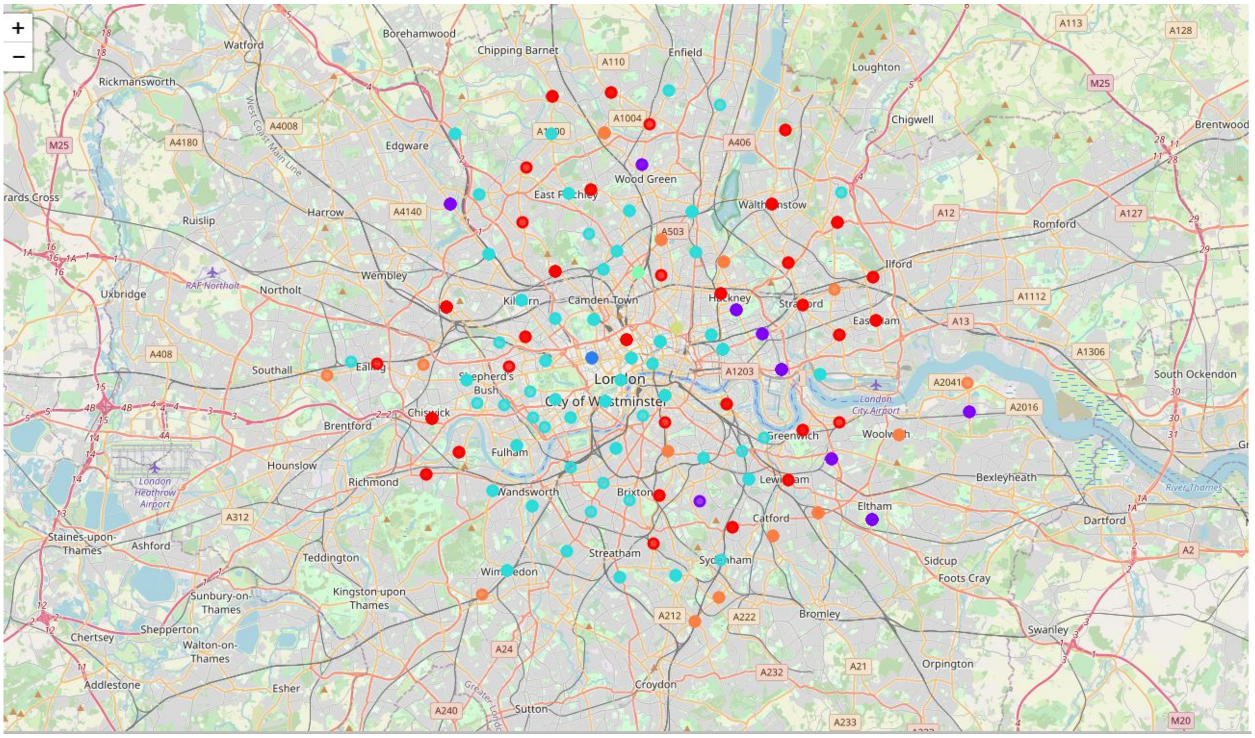
Then the K-Means algorithm was run to cluster the boroughs into 7 clusters because, as shown in the above figures, there is a  7 degree for optimum k of the K-Means.

Then the london_data dataset was merged with the cluster labels for each location.



| | Location | London borough | Post town | Postcode district | Avg price | £/sqm | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | Price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Abbey Wood | Bexley | LONDON | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 | 0 | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station | 0.044347 |
| **1** | Crossness | Bexley | LONDON | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 | 0 | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station | 0.160450 |
| **2** | West Heath | Bexley | LONDON | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 | 0 | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station | 0.228340 |
| **3** | Acton | Ealing | LONDON | W3 | 547488 | 7330.3 | 51.51324 | -0.26746 | 0 | Grocery Store | Train Station | Park | Indian Restaurant | Breakfast Spot | 0.044347 |
| **4** | Aldwych | Westminster | LONDON | WC2 | 1662350 | 17599.1 | 51.51651 | -0.11968 | 0 | Sandwich Place | Pub | Theater | Café | Coffee Shop | 0.044347 |

Also the results were visualised in the map below.

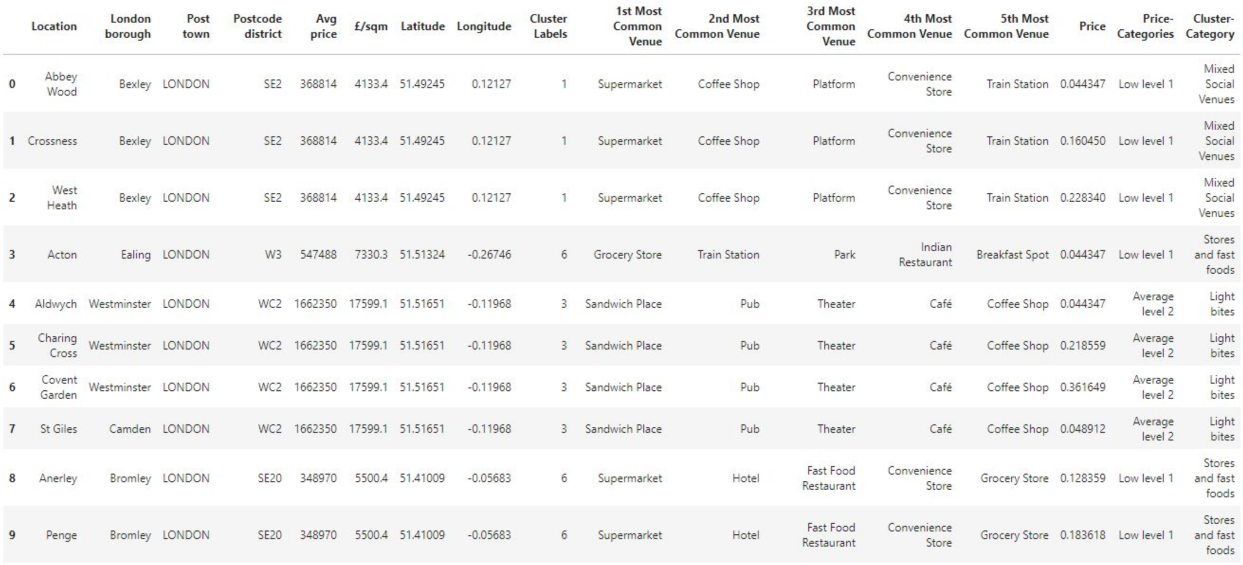


Then bins of average price were identified as below

| Labels | Average Price |
|---|---|
| Low level 1 | <£630k |
| Low level 2 | £630k - £985k |
| Average level | £985k - £1340k |
| Average level 2 | £1340k - £1700k |
| High level 1 | £1700k  - £2400k |
| High level 2 | > £2400k |

Also by examining the clusters the categories below were identified

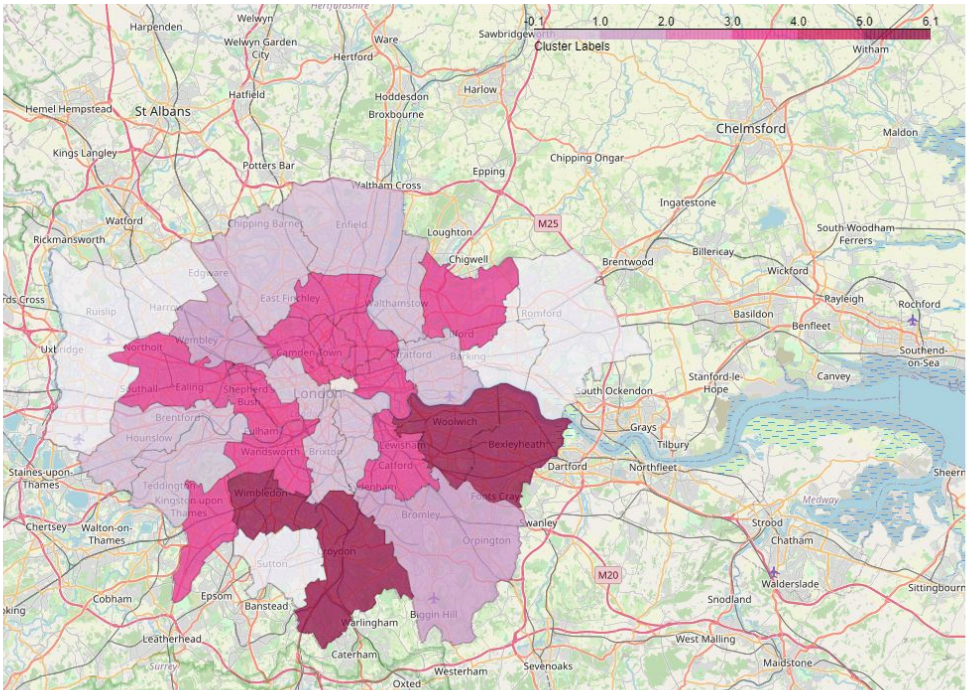
| **Clusters** | **Labels** |
|---|---|
| Cluster 1 | Restaurants |
| Cluster 2 | Mixed Social venues |
| Cluster 3 | Touristic places (cafe - hotel) |
| Cluster 4 | Light bites |
| Cluster 5 | Cafe & Sports events |
| Cluster 6 | All day social venues |
| Cluster 7 | Stores and fast foods |

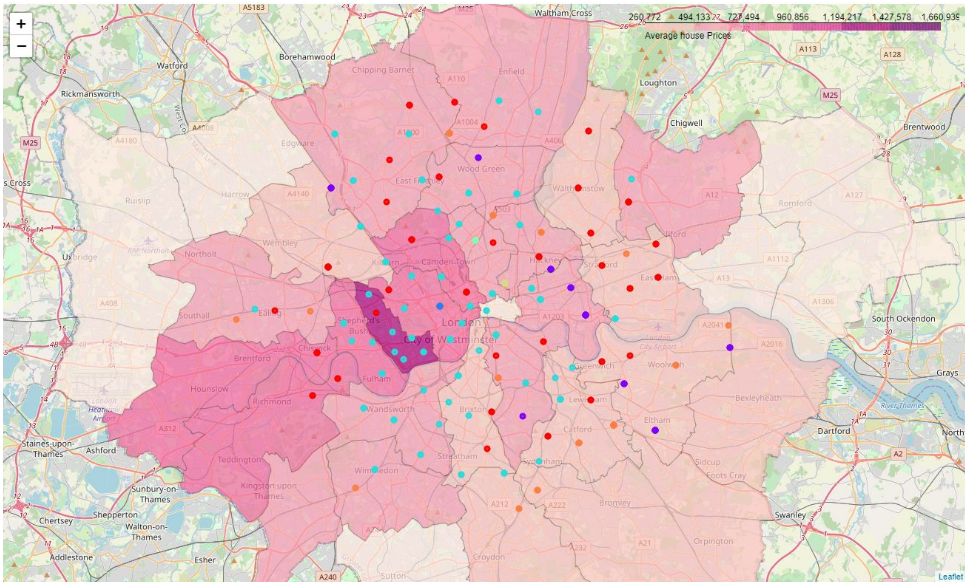


| | Location | London borough | Post town | Postcode district | Avg price | £/sqm | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | Price | Price-Categories | Cluster-Category |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Abbey Wood | Bexley | LONDON | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 | 1 | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station | 0.044347 | Low level 1 | Mixed Social Venues |
| **1** | Crossness | Bexley | LONDON | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 | 1 | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station | 0.160450 | Low level 1 | Mixed Social Venues |
| **2** | West Heath | Bexley | LONDON | SE2 | 368814 | 4133.4 | 51.49245 | 0.12127 | 1 | Supermarket | Coffee Shop | Platform | Convenience Store | Train Station | 0.228340 | Low level 1 | Mixed Social Venues |
| **3** | Acton | Ealing | LONDON | W3 | 547488 | 7330.3 | 51.51324 | -0.26746 | 6 | Grocery Store | Train Station | Park | Indian Restaurant | Breakfast Spot | 0.044347 | Low level 1 | Stores and fast foods |
| **4** | Aldwych | Westminster | LONDON | WC2 | 1662350 | 17599.1 | 51.51651 | -0.11968 | 3 | Sandwich Place | Pub | Theater | Café | Coffee Shop | 0.044347 | Average level 2 | Light bites |
| **5** | Charing Cross | Westminster | LONDON | WC2 | 1662350 | 17599.1 | 51.51651 | -0.11968 | 3 | Sandwich Place | Pub | Theater | Café | Coffee Shop | 0.218559 | Average level 2 | Light bites |
| **6** | Covent Garden | Westminster | LONDON | WC2 | 1662350 | 17599.1 | 51.51651 | -0.11968 | 3 | Sandwich Place | Pub | Theater | Café | Coffee Shop | 0.361649 | Average level 2 | Light bites |
| **7** | St Giles | Camden | LONDON | WC2 | 1662350 | 17599.1 | 51.51651 | -0.11968 | 3 | Sandwich Place | Pub | Theater | Café | Coffee Shop | 0.048912 | Average level 2 | Light bites |
| **8** | Anerley | Bromley | LONDON | SE20 | 348970 | 5500.4 | 51.41009 | -0.05683 | 6 | Supermarket | Hotel | Fast Food Restaurant | Convenience Store | Grocery Store | 0.128359 | Low level 1 | Stores and fast foods |
| **9** | Penge | Bromley | LONDON | SE20 | 348970 | 5500.4 | 51.41009 | -0.05683 | 6 | Supermarket | Hotel | Fast Food Restaurant | Convenience Store | Grocery Store | 0.183618 | Low level 1 | Stores and fast foods |

# Discussion

As closer to the centre a location is, the higher price is required for a property. Also, however boroughs with frequent restaurants are scattered, areas can be identified with “all day venues”. Also as more out of the centre we are located the lower priced stores and fast foods are located.





# Conclusion

The method and results of this study can be used to identify potential areas to open a new restaurant.

The resulting maps can visualise and help the decision making of the potential business that can be open to fit within the context of the neighborhood, taking into account the required initial capital expenditure.