Where to live when you first move to London? A data analysis approach

by Benoit Fedit

Introduction

As part of the final capstone project for IBM Data Science Professional Certification in Coursera.
I decided to write a post where I will share the different methods used to answer the problem: Where to live when you first move to London?
In this project I will only focus on rent price and nearby venues.

Problem

“Where to live when you first move in London?”

London is a global city, known worldwide as one the great cities to visit, and often regarded as the best place to start a career.

As a result,every year many young professionals decide to move to London on a whim and the only decision to make, is where to live.

However, when people move without planning it carefully, they regret having chosen a specific area for renting.
Thus, they want to move again but finding out somewhere to live can become difficult if we don't have the right information.

In this project I'll attempt to gather different information such as venues by area as well as renting price by area and combined them with the aim of clustering the London's districts based on their venues and renting price.

I decided to keep this project simple as I want to finish this project on time while ensuring to implement all parts learnt throughout the course.
But to make the problem more complex we could've added more variables such as school rate, crime rate, surgeries rating...

Methodology

Data Collection: Web scraping

In order to retrieve data about venues and renting price I will use the Foursquare API and the Findproperly website.
I will need to write some data scraping scripts using python.

Data preprocessing

Using the numpy and pandas' libraries I will clean, transorm and merge the different datasets together.

Data Exploration

Using matplotlib, seaborn and folium libraries I will explore the data and try to extract some interesting insights.

Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
In order to address our initial problem I will apply a clustering algorithm on the extracted data and try to cluster the different areas of London according to their type of venues, rent, number of parks and so on.

Data Collection

Web Scrapping & Data Cleansing

Renting Price from FindProperly

	District	Neighborhood	Areas	weekly_rent_1	weekly_rent_2	weekly_rent_3	weekly_rent_4	sales_1	sales_2	sales_3	sales_4
0	BR1	Bromley	NaN	221	300	348	432	264	366	431	751.0
1	BR2	Bromley	NaN	227	304	391	629	231	367	520	603.0
2	BR3	Bromley	NaN	223	315	376	485	271	364	494	677.0
3	BR4	Bromley	NaN	237	265	409	508	147	353	592	734.0
4	BR5	Bromley	NaN	167	263	318	438	216	308	377	673.0

Map Latitude and Longitude to Postcode

	District	Neighborhood	Areas	weekly_rent_1	weekly_rent_2	weekly_rent_3	weekly_rent_4	sales_1	sales_2	sales_3	sales_4	latitude	longitude
0	BR1	Bromley	NaN	221	300	348	432	264	366	431	751.0	51.416710	0.009042
1	BR2	Bromley	NaN	227	304	391	629	231	367	520	603.0	51.506420	-0.127210
2	BR3	Bromley	NaN	223	315	376	485	271	364	494	677.0	51.415095	-0.035403
3	BR4	Bromley	NaN	237	265	409	508	147	353	592	734.0	51.506420	-0.127210
4	BR5	Bromley	NaN	167	263	318	438	216	308	377	673.0	51.506420	-0.127210

Unpivot the renting and selling columns

	District	Neighborhood	Areas	latitude	longitude	variable	value
0	BR1	Bromley	NaN	51.416710	0.009042	weekly_rent_1	221
1	BR2	Bromley	NaN	51.506420	-0.127210	weekly_rent_1	227
2	BR3	Bromley	NaN	51.415095	-0.035403	weekly_rent_1	223
3	BR4	Bromley	NaN	51.506420	-0.127210	weekly_rent_1	237
4	BR5	Bromley	NaN	51.506420	-0.127210	weekly_rent_1	167

Add new column type Sale or Rent

	District	Neighborhood	Areas	latitude	longitude	bedroom	value	type
0	BR1	Bromley	NaN	51.416710	0.009042	1	221	Rent
1	BR2	Bromley	NaN	51.506420	-0.127210	1	227	Rent
2	BR3	Bromley	NaN	51.415095	-0.035403	1	223	Rent
3	BR4	Bromley	NaN	51.506420	-0.127210	1	237	Rent
4	BR5	Bromley	NaN	51.506420	-0.127210	1	167	Rent

List of areas of London

Scraping Wikipedia

In order to later plot our data onto a map we need to get the borough name for each postcode. To do so we can scrap a Wikipedia page which contains the borough name related to each postcode.

	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

	Location	London borough	Postcode district
0	Abbey Wood	Bexley, Greenwich [7]	SE2
1	Acton	Ealing, Hammersmith and Fulham[8]	W3, W4
2	Addington	Croydon[8]	CR0
3	Addiscombe	Croydon[8]	CR0
4	Albany Park	Bexley	DA5, DA14

Data Cleansing

	Location	London_borough	Postcode
0	Abbey Wood	Bexley, Greenwich	SE2
1	Acton	Ealing, Hammersmith and Fulham	W3, W4
2	Addington	Croydon	CR0
3	Addiscombe	Croydon	CR0
4	Albany Park	Bexley	DA5, DA14

As we can see some location have more than one borough separated with a comma.
So we need to split each separated with a coma and create a new row per borough.

	Location	London_borough	Postcode
0	Abbey Wood	Bexley, Greenwich	SE2
1	Acton	Ealing, Hammersmith and Fulham	W3
2	Acton	Ealing, Hammersmith and Fulham	W4
3	Addington	Croydon	CR0
4	Addiscombe	Croydon	CR0

	Location	London_borough	Postcode
0	Abbey Wood	Bexley, Greenwich	SE2
1	Acton	Ealing, Hammersmith and Fulham	W3
2	Acton	Ealing, Hammersmith and Fulham	W4
3	Addington	Croydon	CR0
4	Addiscombe	Croydon	CR0

	Location	Postcode	London_borough
0	Abbey Wood	SE2	Bexley
1	Abbey Wood	SE2	Greenwich
2	Acton	W3	Ealing
3	Acton	W3	Hammersmith and Fulham
4	Acton	W4	Ealing

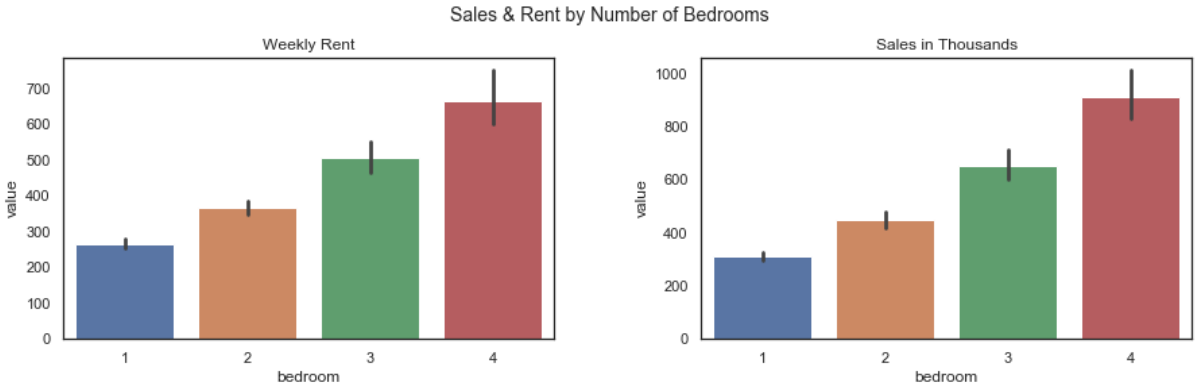
EDA - Exploratory Data Analysis

EDA is one of the most crucial step in data science that allows us to find insights and get a feel for the data. Taking the time to understand the data and being thoughtful about it will help us to better pose a problem or and understand the domain.

By performing EDA thoroughly we can quite often start to draw insight that are essential for the business. We can define what features will be used for our machine learning model and also perform some feature engineering.

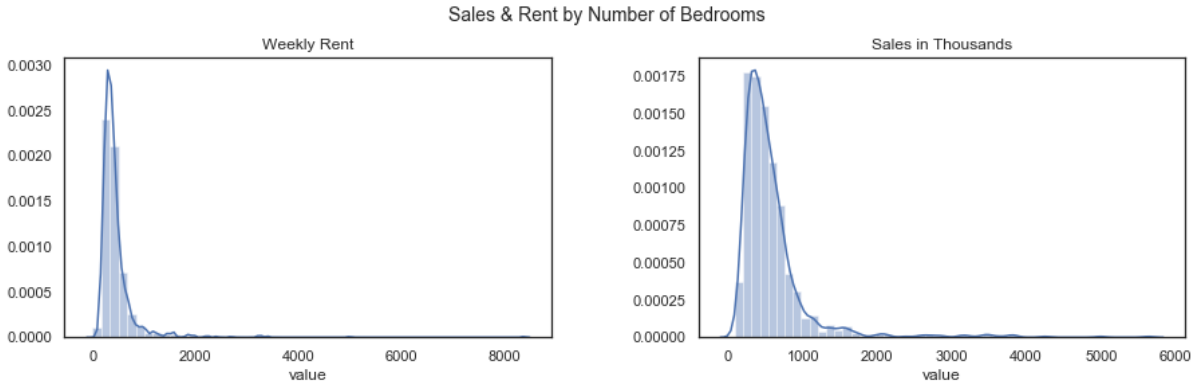
EDA allows us to make initial hypothesis by looking at the data even before running the ML model.
So after running our ML model EDA will help us to validate or to reject our initial assumption about the data.

Renting & Selling Price by Number of Bedrooms



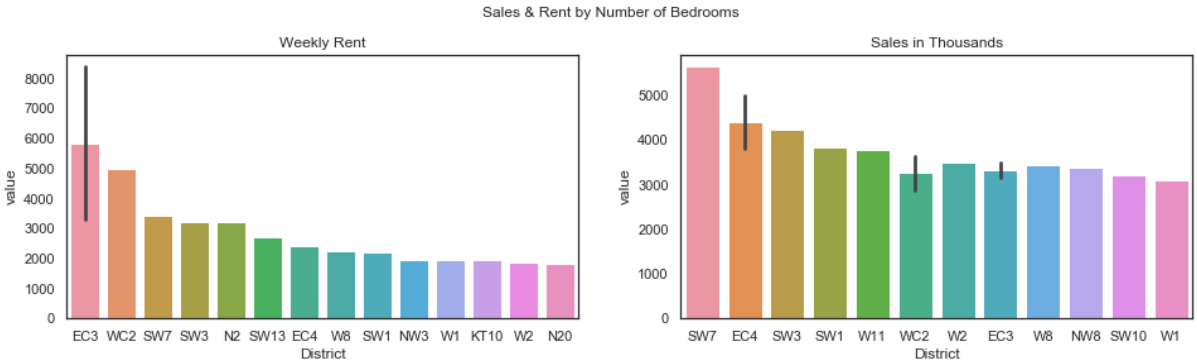
The renting and selling price seem to have a linear growth as the number of bedroom increases.

Renting & Selling Price Distribution



From here we can observe that the distribution of the renting and selling price are both positively skewed.
There seem to be few outliers in the renting price as well as in the selling price.
This is something that we need to pay attention while building a model.
I will use a K-means algorithm in this project and K-means clustering algorithm is actually quite sensitive to outliers, that's because k-means uses the mean and a mean can be greatly influenced by outliers.

Top 15 most expensive districts



This is interesting to observe that some districts are not in the top 15 most expensive area to buy while they are in the top 15 most expensive area to rent.
Those districts could be a good place to invest. I won't go any further into this analysis as this project is about "Where to live" and not "where to invest".

Top 15 most expensive rent not present in top 15 most expensive sales.

	District	Neighborhood	Areas	latitude	longitude	bedroom	value	type
948	N2	Northern	East Finchley	51.58927	-0.16395	4	3228	Rent
769	SW13	Battersea	Barnes	51.47457	-0.24212	3	2694	Rent
971	NW3	North Western\t	Hampstead	51.55506	-0.17348	4	1962	Rent
966	N20	Northern	Totteridge	51.63261	-0.17562	4	1834	Rent

Top 15 most expensive sales not present in top 15 most expensive rent.

	District	Neighborhood	Areas	latitude	longitude	bedroom	value	type
2207	W11	Paddington	Holland Park	51.51244	-0.20639	4	3781	Sale
2088	NW8	North Western\t	St John's Wood	51.53398	-0.17378	4	3391	Sale
2156	SW10	South Western\t	West Brompton	51.48563	-0.18144	4	3224	Sale

	District	Neighborhood	Areas	latitude	longitude	bedroom	value	type
213	SW13	Battersea	Barnes	51.47457	-0.24212	1	658	Rent
491	SW13	Battersea	Barnes	51.47457	-0.24212	2	837	Rent
769	SW13	Battersea	Barnes	51.47457	-0.24212	3	2694	Rent
1047	SW13	Battersea	Barnes	51.47457	-0.24212	4	904	Rent

Merge London Boroughs and London Property price

In order to use the Foursquare API and choropleth map properly I need to add the London_borough from wikipedia into my dataset.
I use an inner join as I want to get rid of the districts for those which we could't find a borough.

	District	Neighborhood	Areas	latitude	longitude	bedroom	value	type	Location	Postcode	London_borough
0	BR1	Bromley	NaN	51.41671	0.009042	1	221	Rent	Bromley	BR1	Bromley
1	BR1	Bromley	NaN	51.41671	0.009042	1	221	Rent	Downham	BR1	Lewisham
2	BR1	Bromley	NaN	51.41671	0.009042	1	221	Rent	Plaistow	BR1	Bromley

	District	Neighborhood	Areas	latitude	longitude	bedroom	value	type	Location	Postcode	London_borough
3	BR1	Bromley	NaN	51.41671	0.009042	1	221	Rent	Sundridge	BR1	Bromley
4	BR1	Bromley	NaN	51.41671	0.009042	1	221	Rent	Widmore (also Widmore Green)	BR1	Bromley

Remove unneeded columns

	Postcode	London_borough	type	bedroom	value	latitude	longitude
0	BR1	Bromley	Rent	1	221	51.41671	0.009042
1	BR1	Lewisham	Rent	1	221	51.41671	0.009042
2	BR1	Bromley	Rent	1	221	51.41671	0.009042
3	BR1	Bromley	Rent	1	221	51.41671	0.009042
4	BR1	Bromley	Rent	1	221	51.41671	0.009042

Remove duplicate rows

As sometime a district belongs to multiple boroughs and vice versa we've generated some duplicate rows.

	Postcode	London_borough	type	bedroom	value	latitude	longitude
0	BR1	Bromley	Rent	1	221	51.41671	0.009042
1	BR1	Lewisham	Rent	1	221	51.41671	0.009042
5	BR1	Bromley	Rent	2	300	51.41671	0.009042
6	BR1	Lewisham	Rent	2	300	51.41671	0.009042
10	BR1	Bromley	Rent	3	348	51.41671	0.009042
11	BR1	Lewisham	Rent	3	348	51.41671	0.009042
15	BR1	Bromley	Rent	4	432	51.41671	0.009042
16	BR1	Lewisham	Rent	4	432	51.41671	0.009042
20	BR1	Bromley	Sale	1	264	51.41671	0.009042
21	BR1	Lewisham	Sale	1	264	51.41671	0.009042
25	BR1	Bromley	Sale	2	366	51.41671	0.009042
26	BR1	Lewisham	Sale	2	366	51.41671	0.009042
30	BR1	Bromley	Sale	3	431	51.41671	0.009042
31	BR1	Lewisham	Sale	3	431	51.41671	0.009042
35	BR1	Bromley	Sale	4	751	51.41671	0.009042

	Postcode	London_borough	type	bedroom	value	latitude	longitude
36	BR1	Lewisham	Sale	4	751	51.41671	0.009042

Duplicate Rows Removed

	Postcode	London_borough	type	bedroom	value	latitude	longitude
0	BR1	Bromley	Rent	1	221	51.41671	0.009042
1	BR1	Lewisham	Rent	1	221	51.41671	0.009042
2	BR1	Bromley	Rent	2	300	51.41671	0.009042
3	BR1	Lewisham	Rent	2	300	51.41671	0.009042
4	BR1	Bromley	Rent	3	348	51.41671	0.009042
5	BR1	Lewisham	Rent	3	348	51.41671	0.009042
6	BR1	Bromley	Rent	4	432	51.41671	0.009042
7	BR1	Lewisham	Rent	4	432	51.41671	0.009042
8	BR1	Bromley	Sale	1	264	51.41671	0.009042
9	BR1	Lewisham	Sale	1	264	51.41671	0.009042

Create Renting dataframe

As already mentioned, my project only focuses on "where to live" and specifically "where to rent", so from here I will only keep the renting price variable.
I will also take the average renting price of each bedroom.

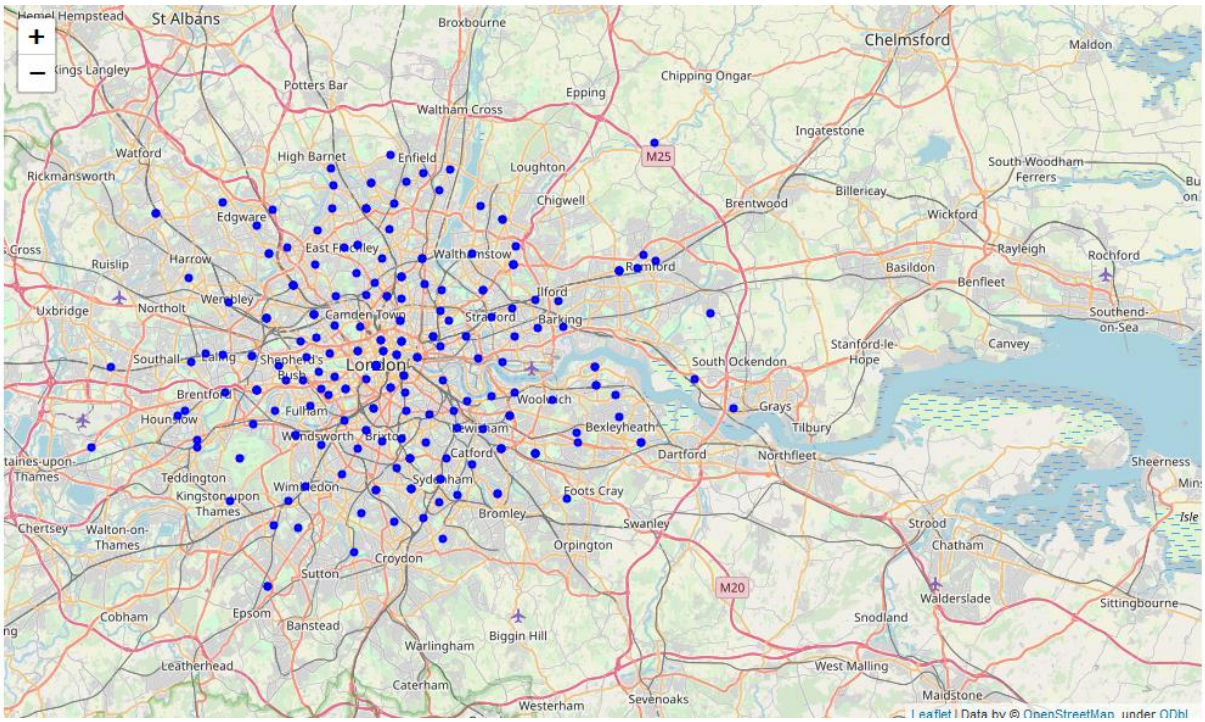
	London_borough	latitude	longitude	avg_rent
0	Bromley	51.416710	0.009042	325.2500
1	Lewisham	51.416710	0.009042	325.2500
2	Bromley	51.506420	-0.127210	335.0625
3	Bromley	51.415095	-0.035403	349.7500
4	Bromley	51.413275	0.087473	344.0000

	London_borough	latitude	longitude	avg_rent
0	Bromley	51.416710	0.009042	325.250000
1	Lewisham	51.416710	0.009042	325.250000
2	Bromley	51.506420	-0.127210	335.062500
3	Bromley	51.415095	-0.035403	349.750000
4	Bromley	51.413275	0.087473	344.000000
5	Croydon	51.384755	-0.051499	323.000000
6	Croydon	51.506420	-0.127210	320.666667
7	Merton	51.402625	-0.143638	327.250000

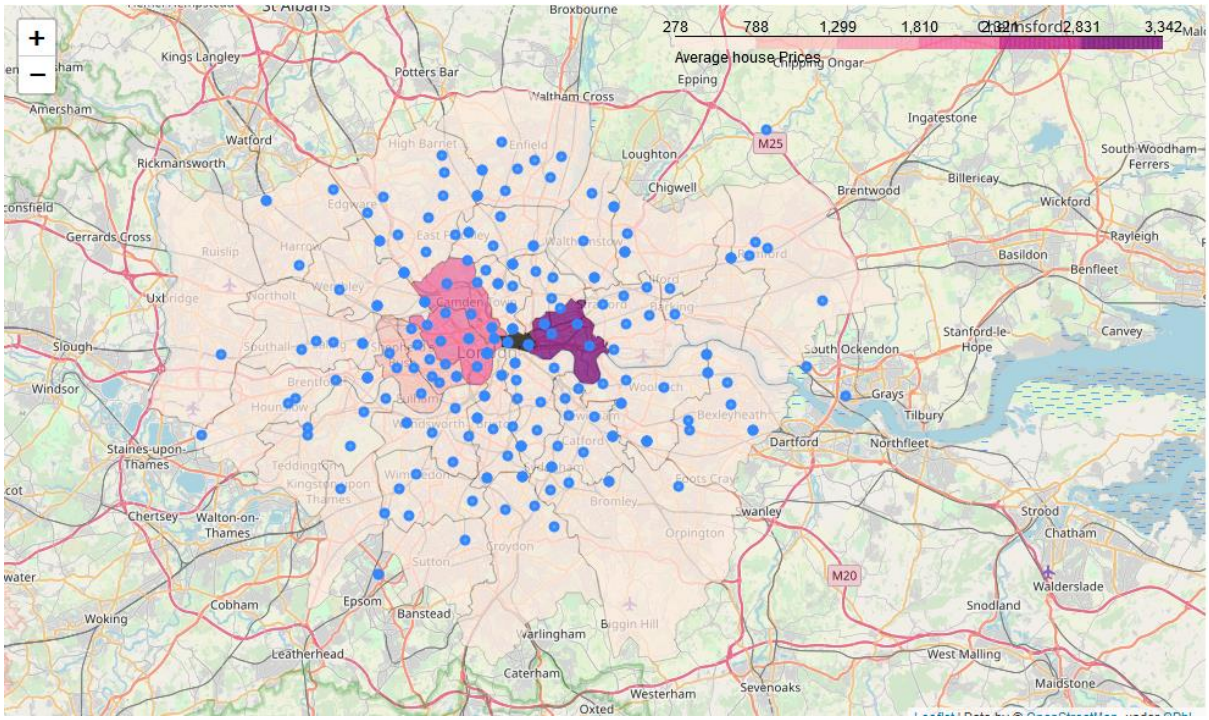
	London_borough	latitude	longitude	avg_rent
8	Croydon	51.396315	-0.106608	308.250000
9	Bexley	51.452068	0.172230	294.875000

Visualise the London districts on Folium map

This map shows all the London's districts that I will try to cluster according to their venues and rents price.



Visualise the London districts with Average Rents by Area



The choropleth map shows all the districts and the boroughs are shaded and coloured according to the average values of their rents.

Extract data from Foursquare

2. Explore Neighbourhoods in London

For every single district of London, I have extracted the top 100 nearest venues within 500 metres. Here is an example of 5 venues that are in Bromley.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bromley	51.50642	-0.12721	Corinthia Hotel	51.506607	-0.124460	Hotel
Bromley	51.50642	-0.12721	Trafalgar Square	51.507987	-0.128048	Plaza
Bromley	51.50642	-0.12721	East Trafalgar Square Fountain	51.508088	-0.127700	Fountain
Bromley	51.50642	-0.12721	Horse Guards Parade	51.504847	-0.126590	Plaza
Bromley	51.50642	-0.12721	ESPA Life at Corinthia	51.506402	-0.125114	Spa

3. Analyse Each Neighbourhood

After performing some transformation I'm able to visualize the most frequent venues for each district.

```
---- Barking and Dagenham----
                        venue  freq
0  Financial or Legal Service  0.25
1                    Pharmacy  0.25
2                  Hookah Bar  0.25
3                Soccer Field  0.25
4            Pedestrian Plaza  0.00


---- Brent----
                 venue  freq
0                Hotel  0.06
1              Theater  0.05
2  Monument / Landmark  0.04
3                Plaza  0.04
4               Garden  0.04


---- Bromley----
                venue  freq
0                 Pub  0.11
1  Italian Restaurant  0.11
2         Supermarket  0.11
3                Park  0.05
4      Ice Cream Shop  0.05
```

…

4. Cluster Neighbourhoods

Run k-means to cluster the neighbourhood into clusters.
First let's merge our two dataframe df_rent (renting price) and london_grouped (venues frequencies)

I will rescale their avg_rent data to have values between 0 and 1. This is usually called feature scaling. One possible formula to achieve this is:

	avg_rent	AVG_Rent_Scaled
0	318.75	0.016909
1	347.00	0.026095
2	457.00	0.061865
3	307.00	0.013088
4	362.75	0.031217

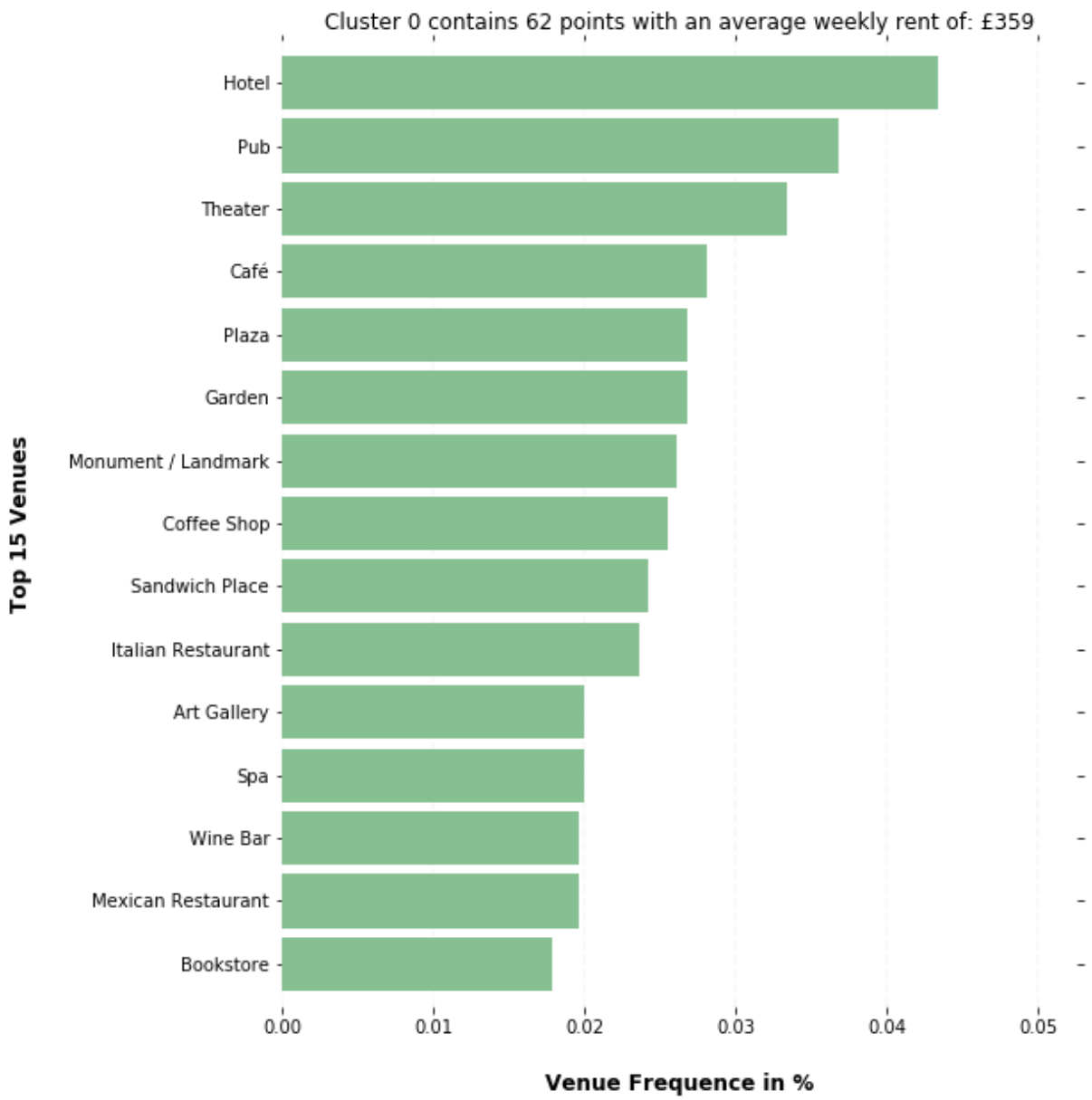
Note that the MinMaxScaler preserves the shape of the original distribution so the K-means algorithm might still be affected by the outliers which will be our maximum value in this case "1"

	Cluster Labels	Neighborhood	London_borough	latitude	longitude	avg_rent	AVG_Rent_Scaled	Venue	Freq
0	0	Barking and Dagenham	Barking and Dagenham	51.572890	0.147528	318.75	0.016909	Accessories Store	0.0
1	0	Brent	Brent	51.506420	-0.127210	347.00	0.026095	Accessories Store	0.0
2	0	Brent	Brent	51.562370	-0.221310	457.00	0.061865	Accessories Store	0.0
3	3	Bromley	Bromley	51.452068	0.172230	307.00	0.013088	Accessories Store	0.0
4	3	Bromley	Bromley	51.426740	-0.055330	362.75	0.031217	Accessories Store	0.0

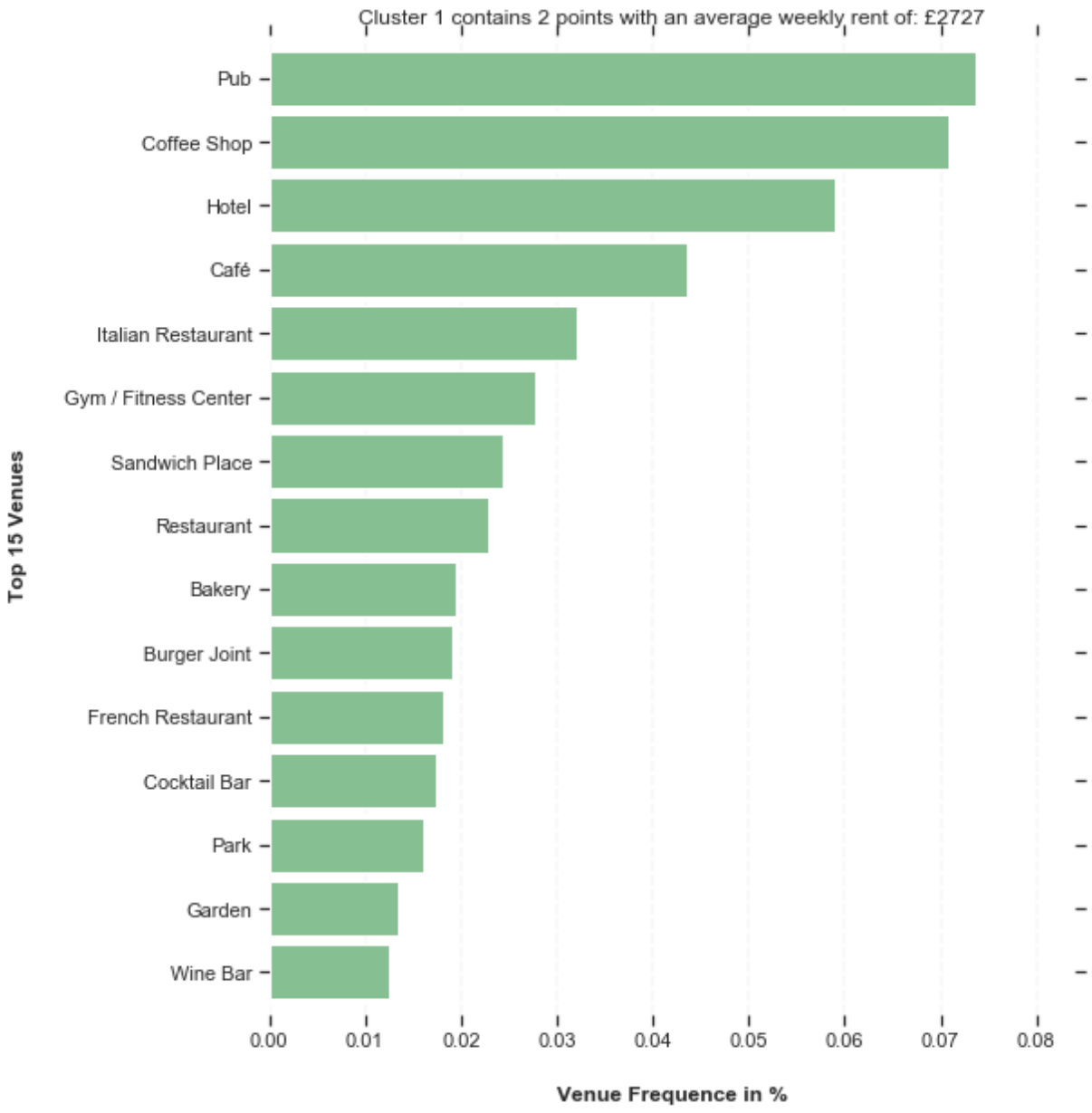
	latitude	longitude
Cluster Labels		
0	62	62
1	2	2
2	17	17
3	96	96

OK what immediately draw our attention is the cluster "1" which has only two values, probably due to the outliers...

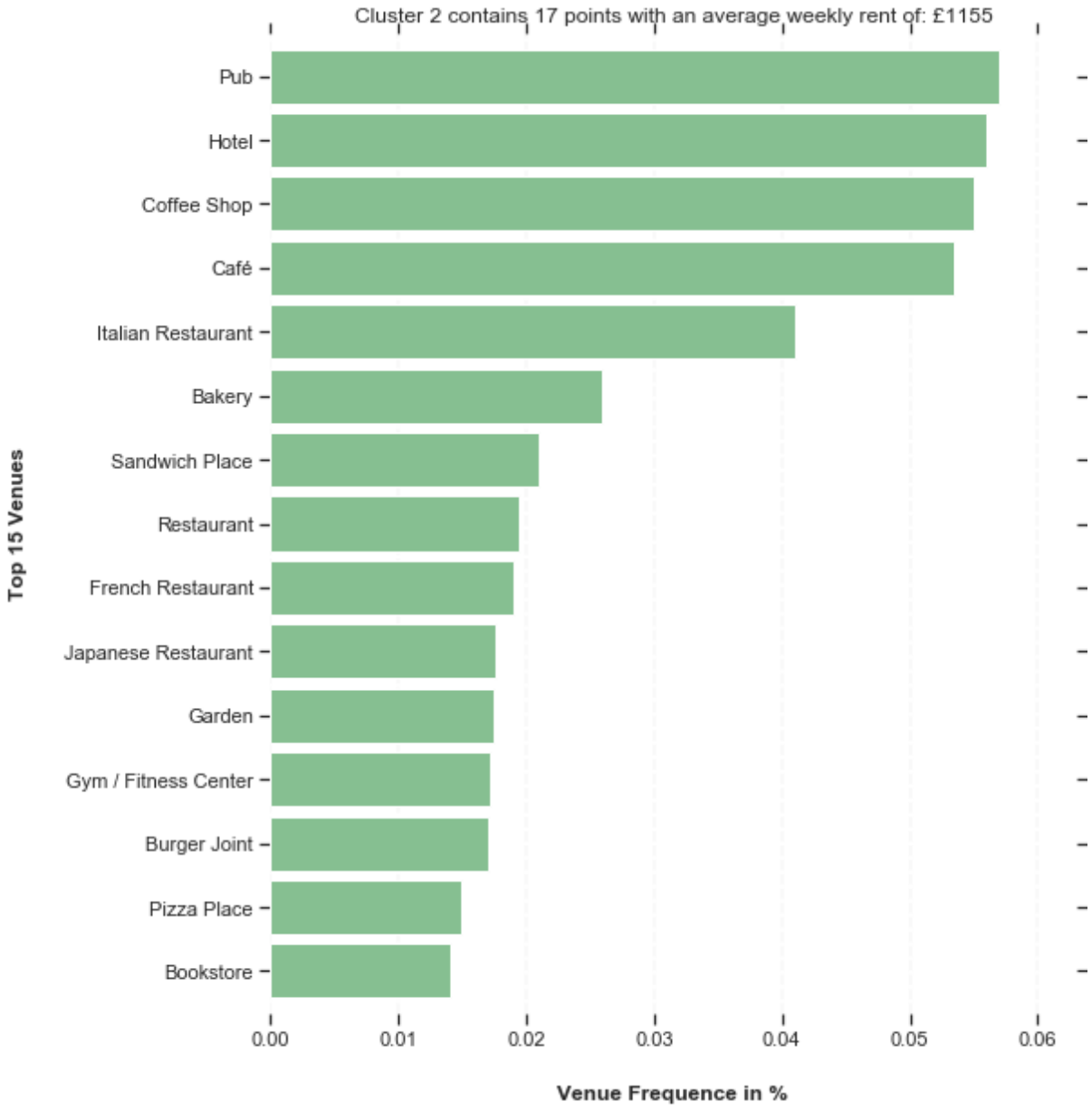
Cluster 1


Cluster 0 contains 62 points with an average weekly rent of: £359
Hotel
Pub
Theater
Café
Plaza
Garden
Monument / Landmark
Coffee Shop
Sandwich Place
Italian Restaurant
Art Gallery
Spa
Wine Bar
Mexican Restaurant
Bookstore
Top 15 Venues
0.00
0.01
0.02
0.03
0.04
0.05
Venue Frequence in %

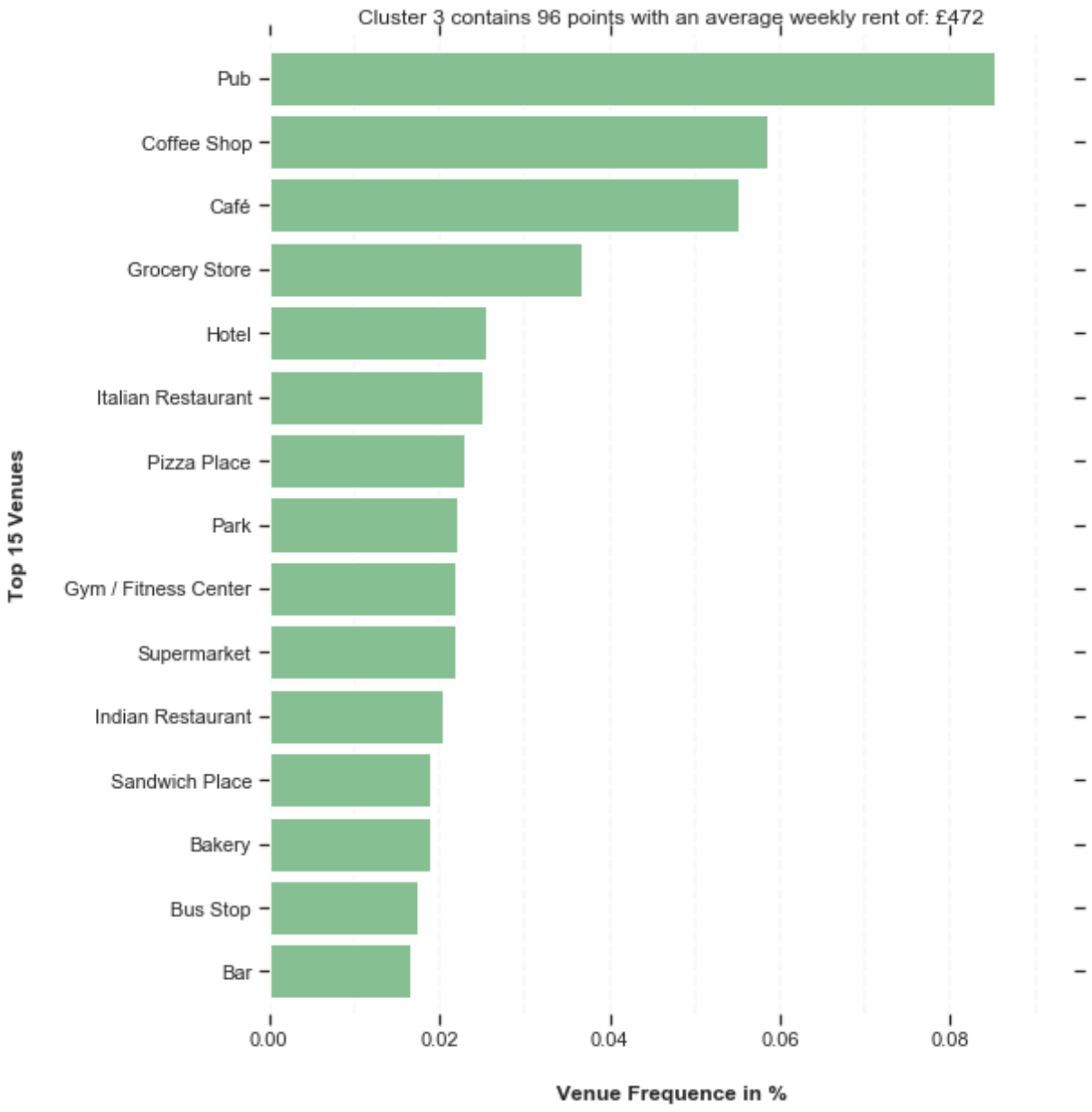
Cluster 2


Cluster 1 contains 2 points with an average weekly rent of: £2727
Pub
Coffee Shop
Hotel
Café
Italian Restaurant
Gym / Fitness Center
Sandwich Place
Restaurant
Bakery
Burger Joint
French Restaurant
Cocktail Bar
Park
Garden
Wine Bar
Top 15 Venues
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Venue Frequence in %

Cluster 3


Cluster 2 contains 17 points with an average weekly rent of: £1155
Pub
Hotel
Coffee Shop
Café
Italian Restaurant
Bakery
Sandwich Place
Restaurant
French Restaurant
Japanese Restaurant
Garden
Gym / Fitness Center
Burger Joint
Pizza Place
Bookstore
Top 15 Venues
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Venue Frequence in %

Cluster 4



After exploring our 4 clusters we can name them based on their most frequent venues as well as their avg weekly rents.

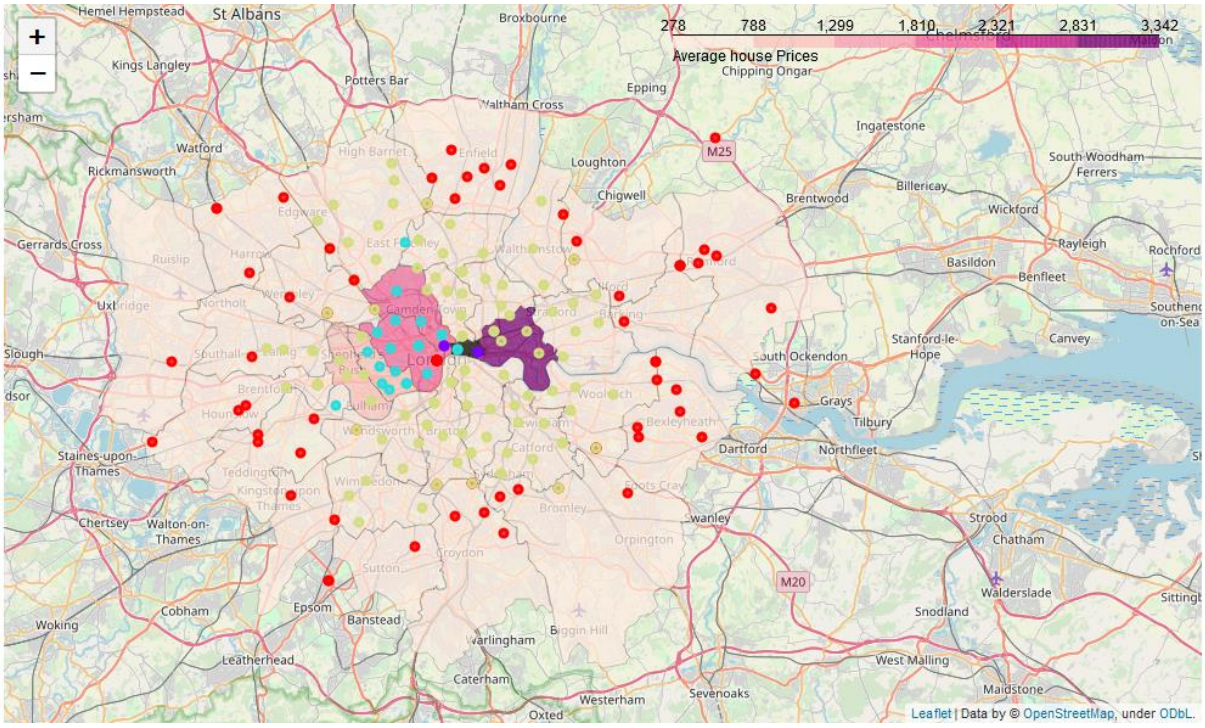
Cluster 0 is the most affordable in term of renting and seems to be a good place for going out as well as enjoying a walk in a park visiting galleries or tasting wines.

Cluster 1 has exorbitant renting price and has only two districts in its cluster. It is indeed caused by the outliers that we have already detected during the EDA part.

Cluster 2 is also quite expensive and has many hotels, pubs, cafes and restaurants the points of this clusters are likely to be located in the city centre.

Cluster 3 is a bit more expensive than cluster 0 but much cheaper than cluster 1 and 2, apart from the
renting price cluster 3 is very similar to cluster 2.

According to their price and top venues I decided to name the 4 clusters as follows:
Cluster 0: Quiet & Cheap
Cluster 1: Exorbitant
Cluster 2: High rent and best districts to go out
Cluster 3: Affordable and good for going out



Result

We came to the result that the most of the districts in central London and close to the centre are similar in term of their venues, what really differentiates them is their renting price. K-means algorithm has been heavily influenced by the variable "avg_rent", we can clearly observe that one cluster contains all the districts of central London, one cluster contains the district of inner London while the last contains the districts of outer London.

Discussion

As mentioned earlier K-mean is sensitive to outliers and this is clearly what has mostly influenced the clusters. Things that can be considered to improve this project:

- Try other models such as K-medoids which is more robust to outlier than K-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances.
- Add other features such as crime rate, school rate...
- Drop the outlier records?
- Cap outliers data

- Transform outliers like transforming the rent price to percentile
- Compare sales vs rents and figure out which district delivers better yields

Conlusion

In this study, I analysed the different districts of London based on the renting price and their venues.
I've implemented the k-means algorithm, some web scrapping and data analysis techniques learnt throughout the IBM Data Science certificate which was mainly taught by Dr Saeed Aghbozorgi and Dr Alex Aklson.

References

- https://en.wikipedia.org/wiki/K-means_clustering
- https://en.wikipedia.org/wiki/Elbow_method_(clustering)
- https://www.coursera.org/professional-certificates/ibm-data-science

Link to the project and data files

https://github.com/f-benoit/IBM-Data-Science/tree/master/London%20CLustering