London Venues and Business Opportunities

Svetlana Petrova

April 10, 2020

IBM CERTIFICATION IN DATA SCIENCE

Capstone Project

1. Summary

This report compares London areas in terms of venues and population’s income. The aim of the analysis is to produce insights which would aid potential and existing businesses and contribute to their awareness about the local environment.

The analysis looks at two matters – venues and average income.

The first part of the workings focuses on the venues in various London postcodes. To optimize the data processing, I have chosen to present venues only in district WC1R which is large are covers key areas of Central and West London with a great variety of public places.

The second part compares the average income in the different London districts. The results are presented on a map in order to visualize the areas with highest revenues and presumably better purchasing power of the residents.

2. Introduction

2.1 Background

London, the capital of the United Kingdom, is a big and diverse megapolis which, prior to Brexit, had been the largest city in the EU. With more than 100 nationalities, London is the most international city in the world. Its inexhaustible opportunities make it attractive to tourists, residents and entrepreneurs.

The business side of London is impressive. It is home for many giant companies from all over the world but is also a great environment for small and medium-sized enterprises. The consumer potential is enough for entrepreneurs to try and actually have the chance to succeed.

2.2 Problem

With population of 14 million (2011 census), Greater London’s demand for public venues of any kind is immense. There one can find thousands of offices, sports clubs and entertainment. Therefore, considering the scale, a possible investor looking to open a business in London should carry out a research when choosing a location and if not yet decided, type of venue to open.

If there is no clear idea on the venue to open, information about currently existing venues should be gathered to identify a niche.

Good choice of location very much links to the opportunity of the area considered. That would include situation and infrastructure, but also important demographic factors like local residents’ purchasing power. Information about population income could be a good indicator.

2.3 Interest

Everyone looking to open a business in London could benefit from this report and refer to it when deciding on type of company to open, as well as which London location to go for. Existing businesses could also find this report useful as it is indeed necessary to be aware of the changes and trends in the relevant environment.

3. Data acquisition and cleaning

3.1 Data sources

The analysis is based on data about London districts via a combination of two datasets:

doogal.co.uk – From this website, I retrieved a dataset "London Postcodes" which contains information about the various postcodes in Greater London, including geo coordinates, district and area names and codes district, as well as average income details.

Foresquare – I connect *Foresquare* to extract data with London venues in district WC1R.

3.2 Data Cleaning

"London Postcodes" information on *doogle* has been last updated in February 2020. It is very recent, same for Foresquare data. Linking the csv dataset and the venues extract from *Foresquare* was successful but a few challenges had to be overcome on the way.

In general, obtaining results via API calls to *Foresquare* does not go smooth at every trial and there could be issues iterating through the *json* file. These errors occur randomly so it could be a matter of trying at different times to succeed.

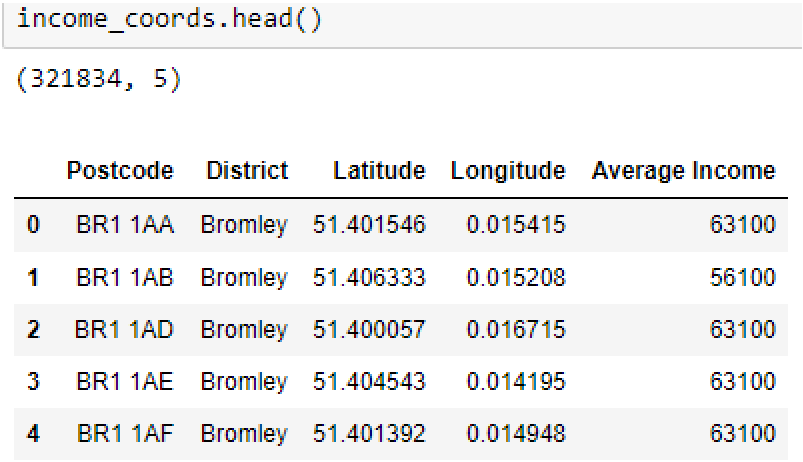
When I tried to relate the venues dataframe with the csv file dataframe, the matching of latitudes was not as productive as when done manually in Excel, although the formatting and values size were the same. The manual *vlookup* match came up with over 70 matching latitudes vs only 14 using *panda.merge*. For the size of data I worked with, this was overcome easily – I extracted manually the postcodes and income for the latitudes of the venues dataframe and prepared two lists which then I appended to the venues.

In the csv file from *doogle* coordinates were provided for each separate postcode, not per district. In the second part where I look at the income in the various London districts, I had to extract from *GeoPy* coordinates for each of the districts. I worked with 329 districts but due to the restrictions for calls per second, I had to split the postcodes list into 22 batches (the optimal number of batches was achieved via trial and error). For this I used a *for loop* with number of iteration equaling the number of batches, and method *time.sleep* to set up a 2-second's pause between iterations.

4. Methodology

4.1 Venues in London District WC1R

I started with reading the csv file with the “London Postcodes” dataset from *doogal*. It contains the details of postcodes, districts and district postcodes, and average income which I then used in both parts of the analysis. This is a large dataset with over 320 thousand of postcodes. It contains 47 columns in total but only a few were needed for my workings, so I extracted them into the below dataframe “income_coords”. It contains all London Postcodes with their coordinates, name of district and average income.

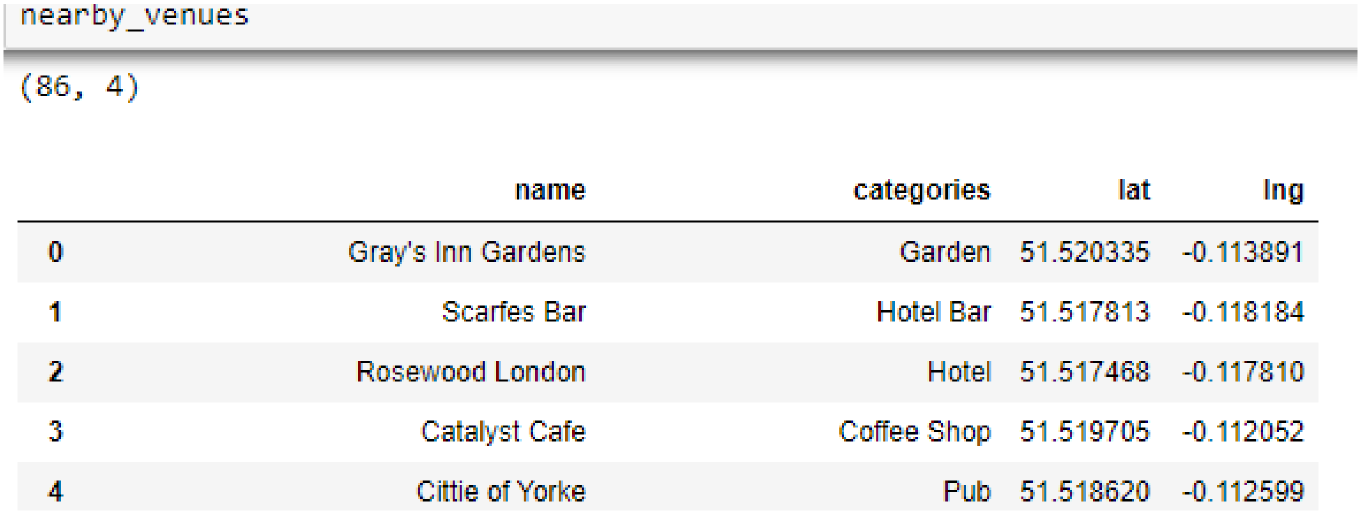


```
income_coords.head()
```

```
(321834, 5)
```

	Postcode	District	Latitude	Longitude	Average Income
0	BR1 1AA	Bromley	51.401546	0.015415	63100
1	BR1 1AB	Bromley	51.406333	0.015208	56100
2	BR1 1AD	Bromley	51.400057	0.016715	63100
3	BR1 1AE	Bromley	51.404543	0.014195	63100
4	BR1 1AF	Bromley	51.401392	0.014948	63100

I used *GeoPy* to obtain the coordinates for London district WC1R. That was my base point for exploring nearby venues. I utilized the Foursquare API to explore the venues and segment them. I set the limit to 100 venues and the radius to 500 meters. I extracted a *json* file with venues in WC1R and prepared the below dataframe “nearby_venues” with venues, categories they fall into and geo coordinates:



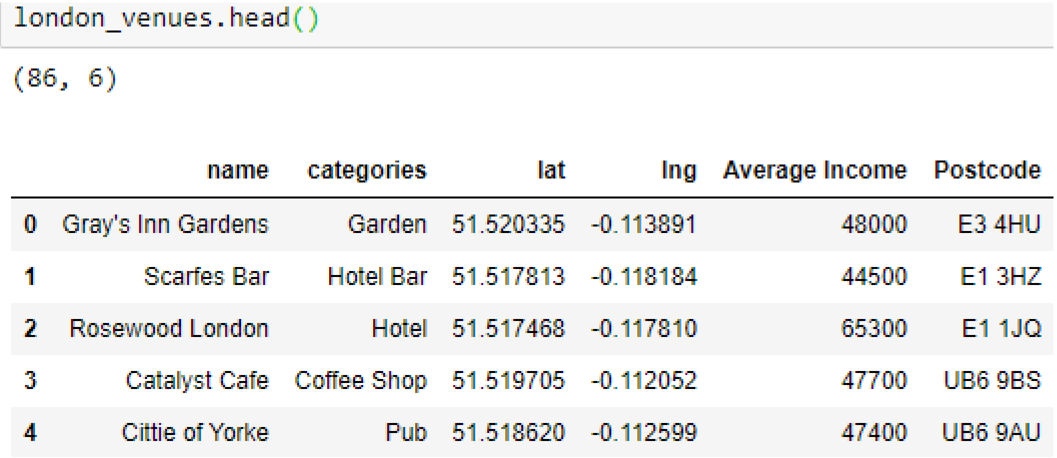
```
nearby_venues
```

```
(86, 4)
```

	name	categories	lat	lng
0	Gray's Inn Gardens	Garden	51.520335	-0.113891
1	Scarfes Bar	Hotel Bar	51.517813	-0.118184
2	Rosewood London	Hotel	51.517468	-0.117810
3	Catalyst Cafe	Coffee Shop	51.519705	-0.112052
4	Cittie of Yorke	Pub	51.518620	-0.112599

As I wanted to visualize on map the venues where popups would show average income too, I prepared two lists with postcodes and income which I added to “nearby_venues”. The values of the lists were taken from the *doogle* csv file, based on the latitude column.

This exercise resulted in the following dataframe “london_venues” with 86 venues samples:



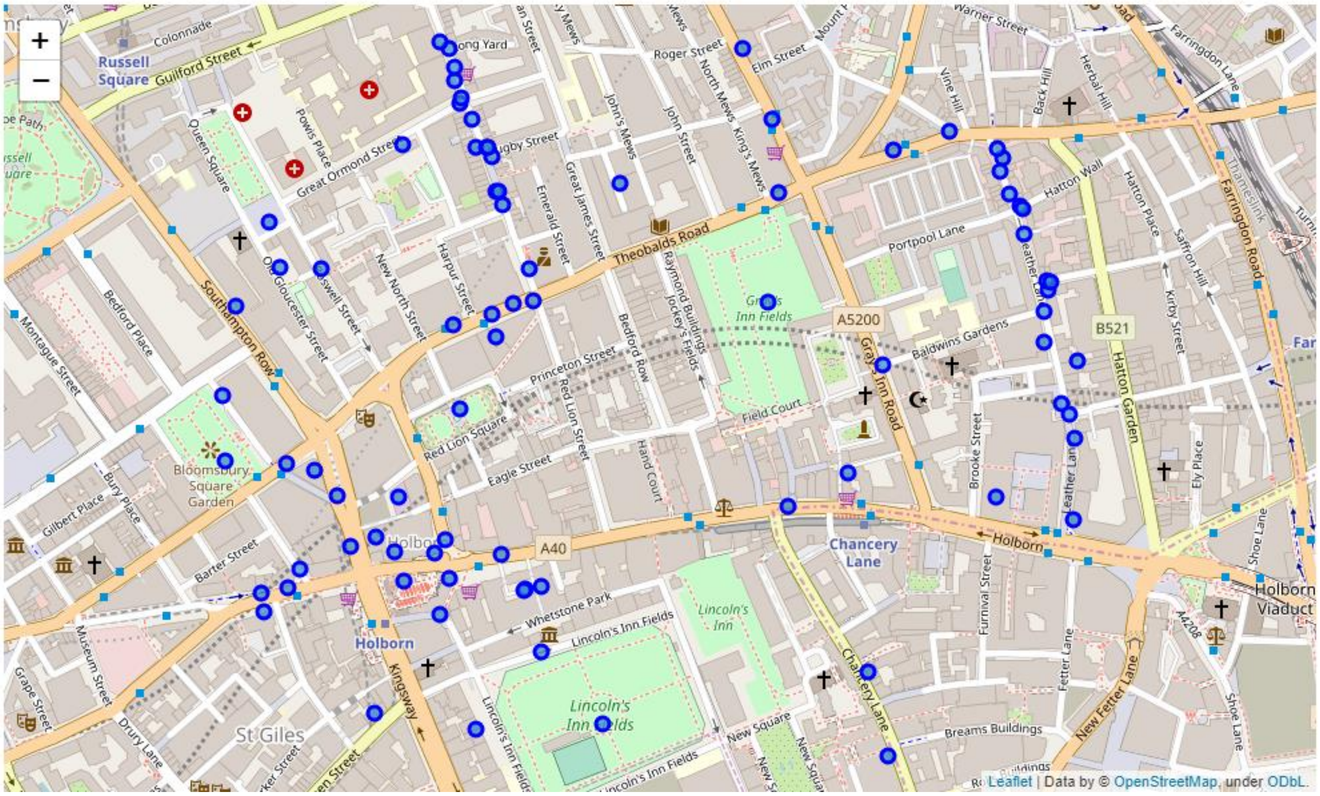
```
london_venues.head()
```

```
(86, 6)
```

	name	categories	lat	lng	Average Income	Postcode
0	Gray's Inn Gardens	Garden	51.520335	-0.113891	48000	E3 4HU
1	Scarfes Bar	Hotel Bar	51.517813	-0.118184	44500	E1 3HZ
2	Rosewood London	Hotel	51.517468	-0.117810	65300	E1 1JQ
3	Catalyst Cafe	Coffee Shop	51.519705	-0.112052	47700	UB6 9BS
4	Cittie of Yorke	Pub	51.518620	-0.112599	47400	UB6 9AU

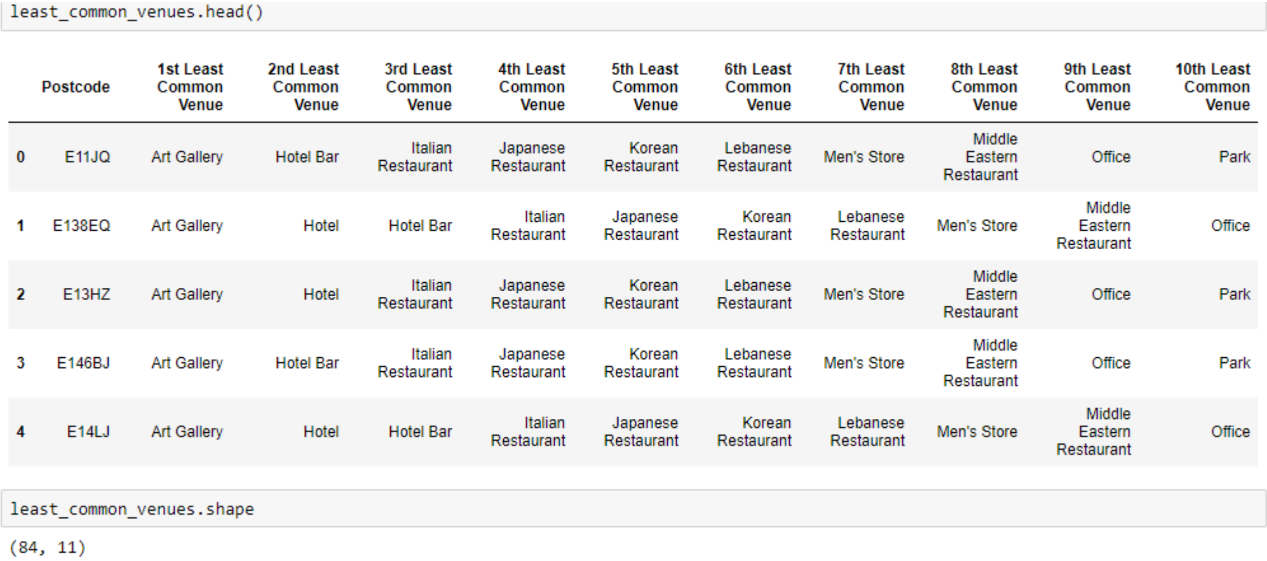
Here are how these venues are situated into Central and West London:

```
map_WestLondon
```



Least Common Venues

Now that I had venues and categories information about my list of postcodes, I could examine these in further detail. My purpose was to spot a possible new business opportunity, therefore I decided to explore the venue categories which were least frequently present. I limited my analysis to 10 results as going further up into the ranking would not be as relevant and could distort the results. I applied *OneHot encoding* onto “london_venues”. With the help of a couple of functions I set up for the purpose, I arrived at a dataframe showing the 10 least common venues for each of the postcodes:



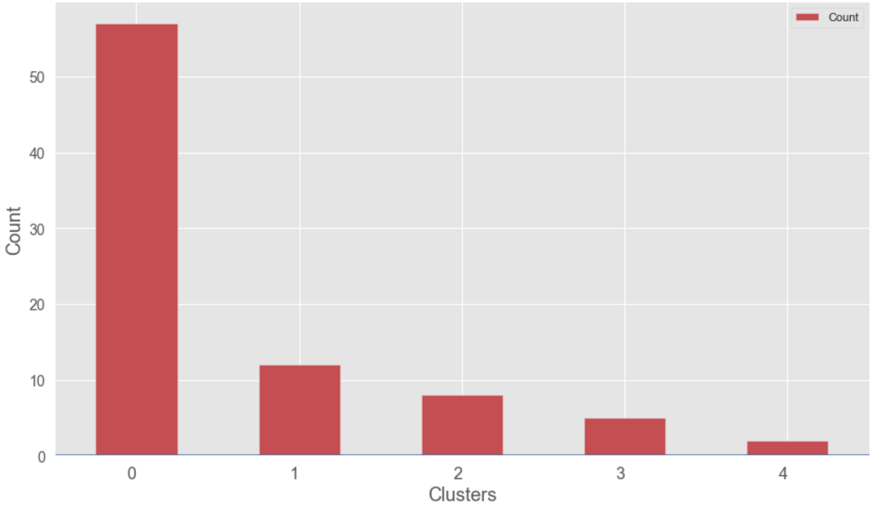
```
least_common_venues.head()
```

	Postcode	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Least Common Venue	6th Least Common Venue	7th Least Common Venue	8th Least Common Venue	9th Least Common Venue	10th Least Common Venue
0	E11JQ	Art Gallery	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office	Park
1	E138EQ	Art Gallery	Hotel	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office
2	E13HZ	Art Gallery	Hotel	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office	Park
3	E146BJ	Art Gallery	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office	Park
4	E14LJ	Art Gallery	Hotel	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office

```
least_common_venues.shape
```

```
(84, 11)
```

In London_venues, there were 86 lines with 84 unique postcodes. When I performed *kmeans* analysis with 5 clusters, it was clear that the result would not be very diverse. Indeed, the majority of the lines fell into the first cluster absorbing 57 out of the 84 postcodes. The below graph shows the large gap between the 5 clusters:

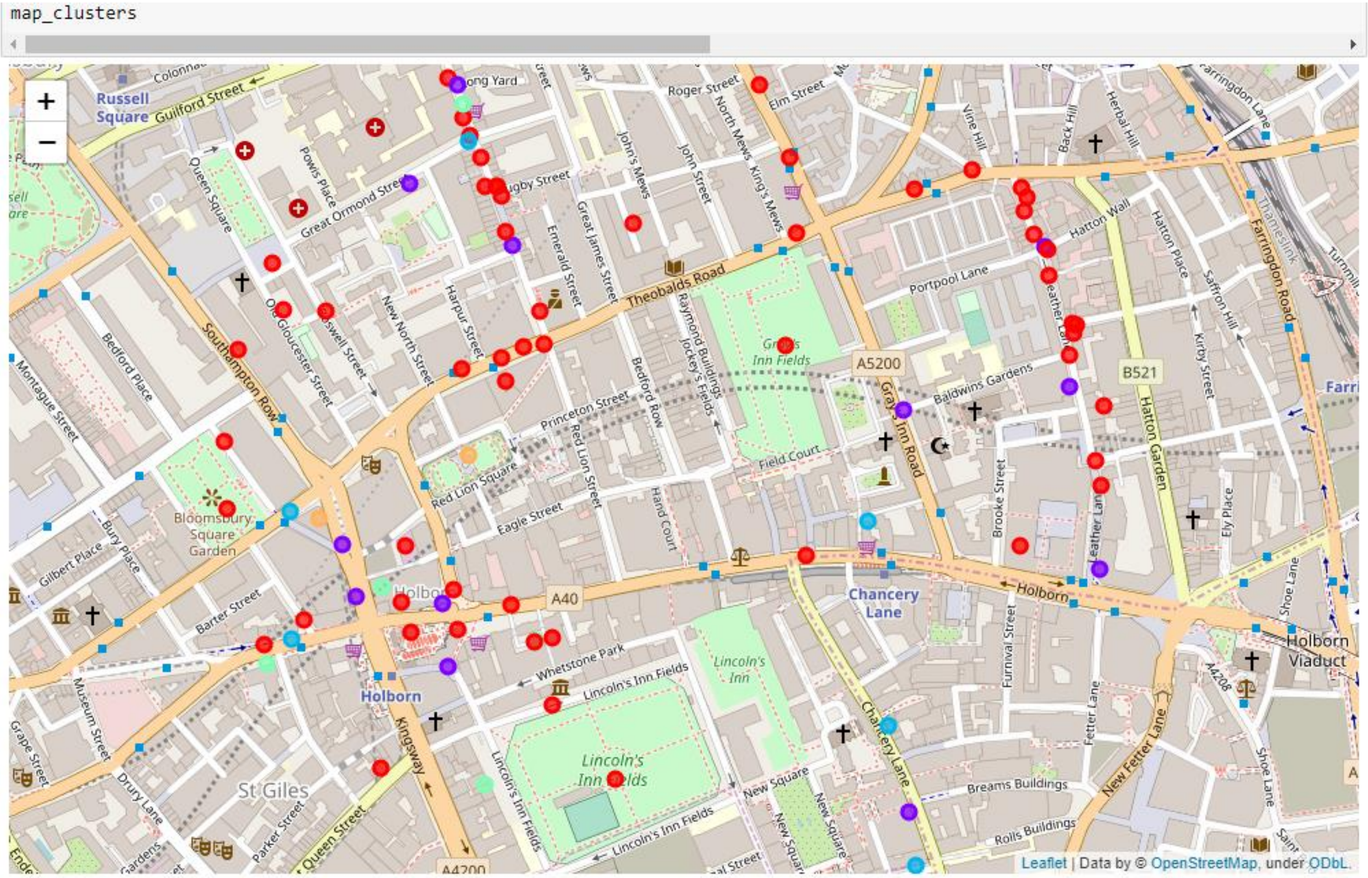


When visualized on the map of London, member of the same cluster could be found in various parts of the West End, and the clusters are not geographically separated from one another. Being an Unsupervised Learning algorithm, *kmeans* does not deal with data labels and data remains anonymous to it. Therefore, any resulting clusters are an attempt of *kmeans* to provide meaningful grouping of the data hence it is crucial that even after removing the labels, the input possesses logic and pattern.

In the following map, the five clusters are presented in different colors, where the members of the largest Cluster 0 is in red. Each marker has a popup label with all of the following information about the marker in question:

- Postcode
- 4th Least Common Venue – I have chosen this point of the ranking because the first three were too overlapping across all postcodes. The forth places contained a better variety and thus a more diverse base for orientation.
- Average Income in that location
- Cluster number

When looking at this map, one could focus on the spots with preferred location and know straightaway what the residents' income level is there.

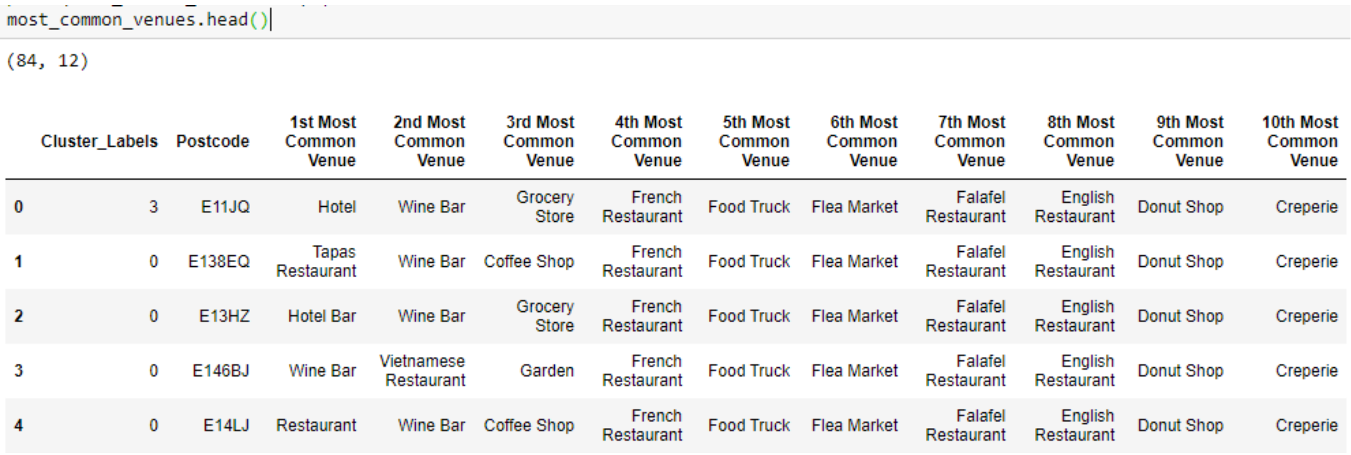


```
map_clusters
```

Leaflet | Data by © OpenStreetMap, under ODbL.

Most Common Venues

Existing businesses interact and should exist in synch as far as possible, therefore it is important to have a more detailed picture of what is around when looking to make a move in certain location. Now that I had examined the least common venues, it was important to find out what were the most common ones. This time I defined a function returning a dataframe showing the top 10 frequent venues. It naturally produced results including cafes, grocery stores, coffee shops and types of restaurants very typical for London like wine bars and pubs.

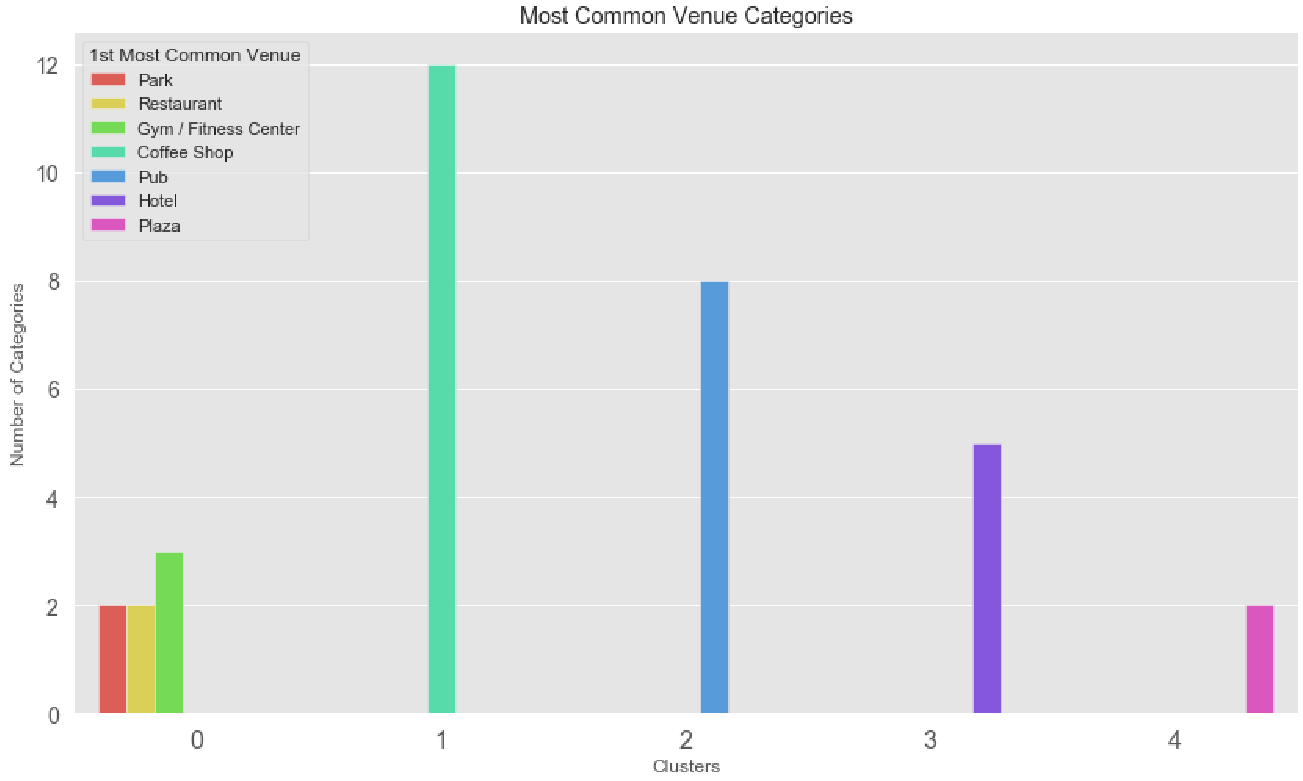


```
most_common_venues.head()
```

```
(84, 12)
```

	Cluster_Labels	Postcode	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	3	E11JQ	Hotel	Wine Bar	Grocery Store	French Restaurant	Food Truck	Flea Market	Falafel Restaurant	English Restaurant	Donut Shop	Creperie
1	0	E138EQ	Tapas Restaurant	Wine Bar	Coffee Shop	French Restaurant	Food Truck	Flea Market	Falafel Restaurant	English Restaurant	Donut Shop	Creperie
2	0	E13HZ	Hotel Bar	Wine Bar	Grocery Store	French Restaurant	Food Truck	Flea Market	Falafel Restaurant	English Restaurant	Donut Shop	Creperie
3	0	E146BJ	Wine Bar	Vietnamese Restaurant	Garden	French Restaurant	Food Truck	Flea Market	Falafel Restaurant	English Restaurant	Donut Shop	Creperie
4	0	E14LJ	Restaurant	Wine Bar	Coffee Shop	French Restaurant	Food Truck	Flea Market	Falafel Restaurant	English Restaurant	Donut Shop	Creperie

I performed on the above data a *kmeans* analysis splitting it into 5 clusters. This graph displays the most common venues in each of the clusters.



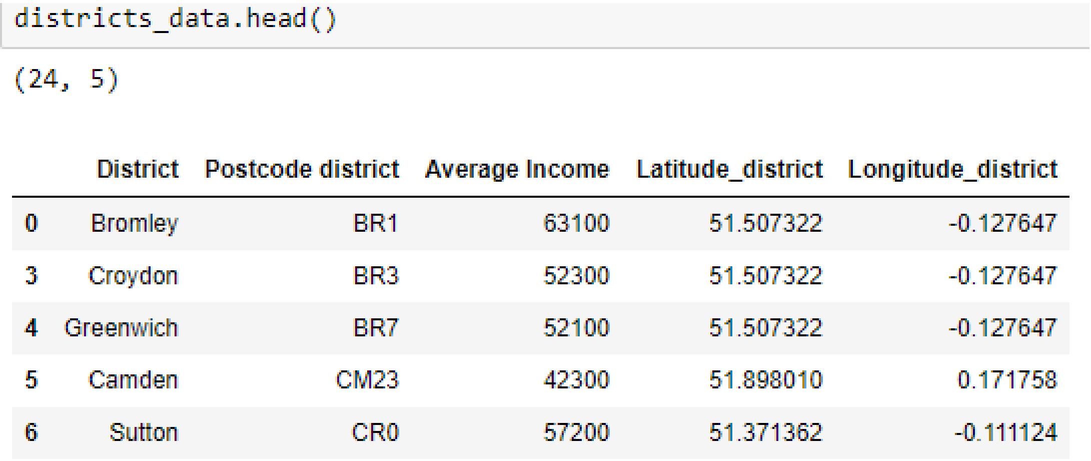
4.2 Average Income in various London Districts

Whilst the first part of my analysis focused on London District WC1R, the second part looks at a broader picture of London, covering 33 different districts.

Every business depends to a great extent on the purchasing power of their potential customers. Therefore, it is important to be aware of the income levels in considered locations. The *doogle* csv dataset contains information about average income for every separate postcode enlisted there. This facilitated my purpose to visualize the income spread amongst various London districts.

To achieve this, I needed the districts with their names, district postcode (which is technically the first part of a postcode), income and coordinates. The first three I took from the "London Postcodes" dataset. The coordinates I obtained from *GeoPy* via API calls. As mentioned earlier, in order to avoid the *GeoPy* restrictions and receiving a "Service Not Available" error, I implemented a *for loop* with 22 iteration, and method *time.sleep* to set up a 2-second's pause between iterations.

I collated the needed information and it tuned out that I had all necessary details for a total of 24 districts, which I put into the following dataframe:

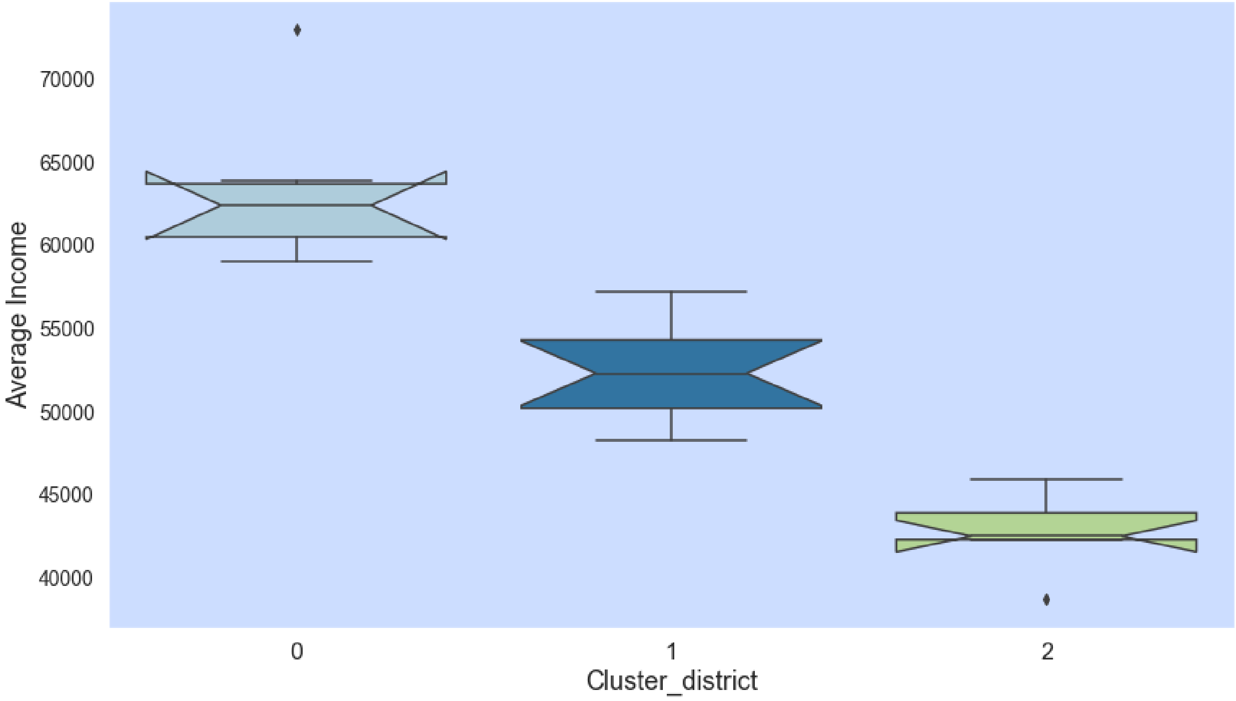


```
districts_data.head()
```

```
(24, 5)
```

	District	Postcode district	Average Income	Latitude_district	Longitude_district
0	Bromley	BR1	63100	51.507322	-0.127647
3	Croydon	BR3	52300	51.507322	-0.127647
4	Greenwich	BR7	52100	51.507322	-0.127647
5	Camden	CM23	42300	51.898010	0.171758
6	Sutton	CR0	57200	51.371362	-0.111124

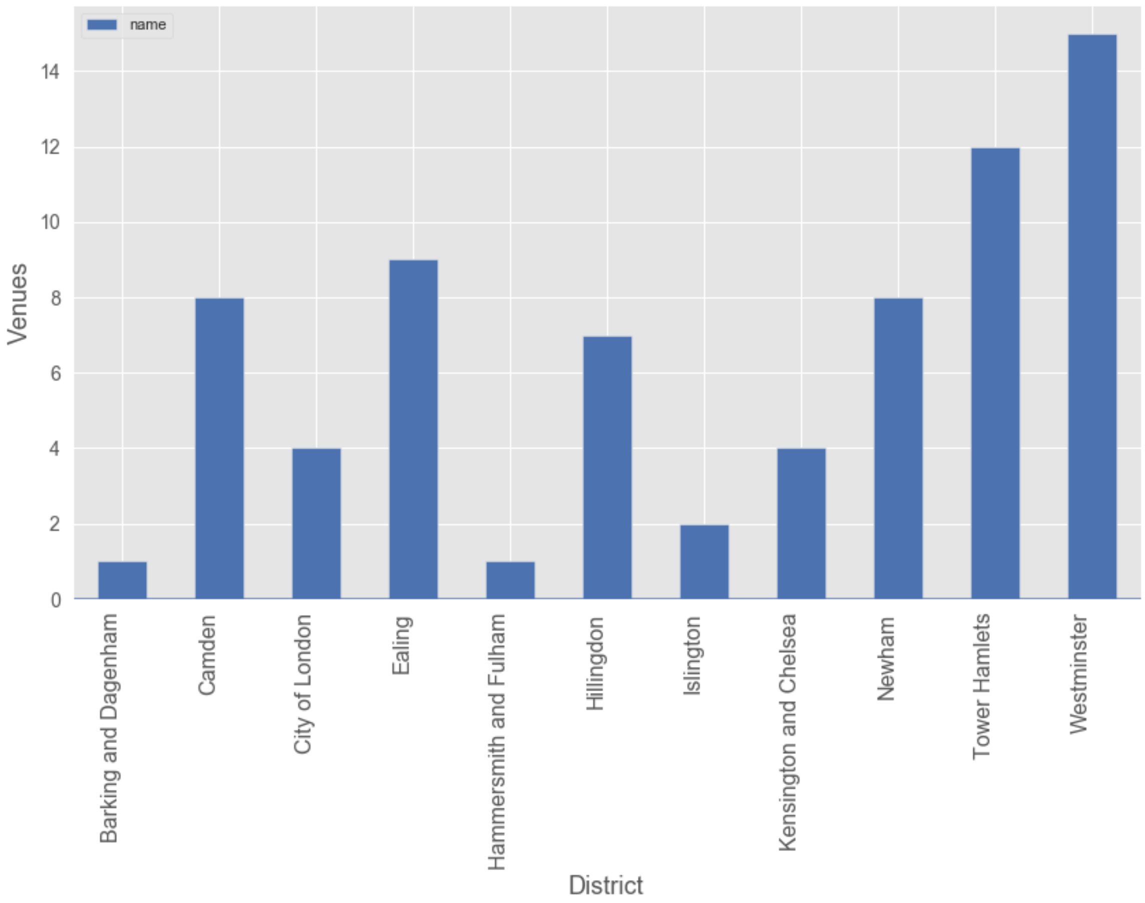
I decided to see what a *kmeans* analysis would show here. This time I grouped my dataframe by Average Income as I was curious to find out if there would be divisible pools. The analysis showed three clear income ranges represented by the three clusters. The below box plot visualizes how they compare:



5. Results

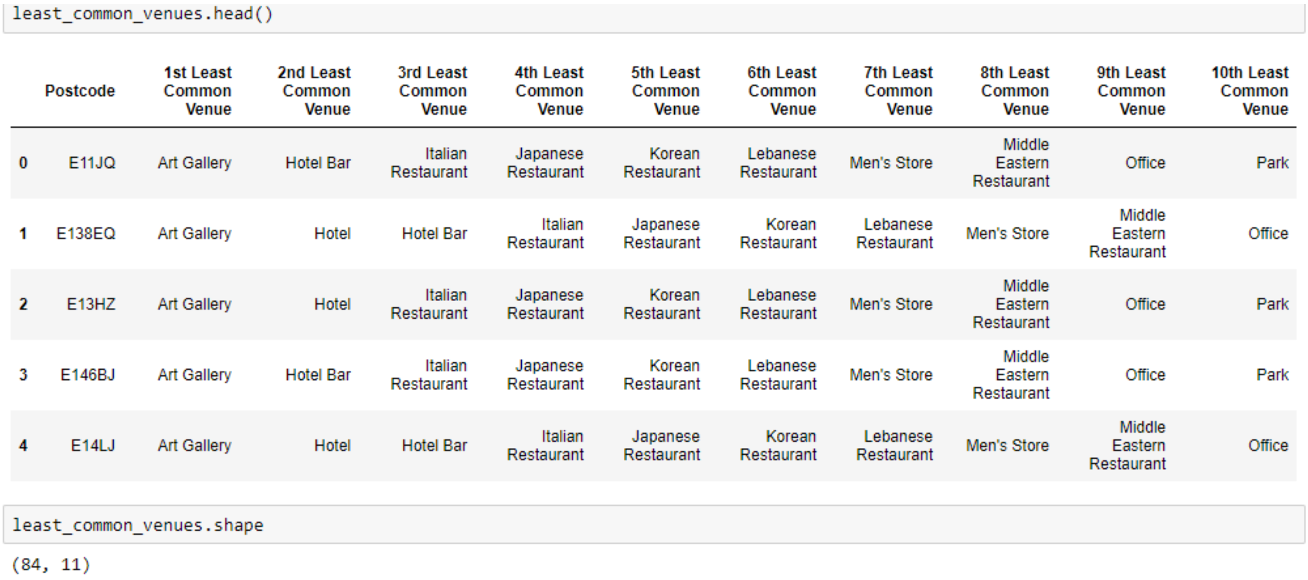
The analysis presented in this report provided some results which could be of good use for individuals and organizations looking to get to know what London offers.

When exploring the venues, it was important to see the saturation of places in the different districts. I added the District name to the venues dataframe and I examined the result. I had a dataframe with 86 samples of venues from 11 different districts. The below graph compares the number of venues in each district:



It is not surprising to see that Westminster is the district with most venues as it includes a significant part of Central London, it is home to the Parliament and Big Ben, Westminster Abbey, South Bank and many others. It attracts millions of tourists and it is justifiable that so many venues are placed there. Tower Hamlets is a neighboring district and the saturation of venues naturally expands into that area.

In the least_common_venues table, one could immediately spot that the least common venue is Art Gallery and this makes sense – art galleries are very specific and unique and in general not supposed to be at every corner. If you move further up into the ranking, there are foreign cuisine restaurants which seem to be rear. Opening a Japanese, Korean or Lebanese restaurant could be a good business option. It is interesting to see that in some areas Italian Restaurants are not very common, taking third place bottom-up. After filtering “least_common_values” for column “3rd Least Common Venue” equaling “Italian restaurant”, it can be seen that there are 52 locations which are possibly lacking Italian cuisine.



```
least_common_venues.head()
```

	Postcode	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	4th Least Common Venue	5th Least Common Venue	6th Least Common Venue	7th Least Common Venue	8th Least Common Venue	9th Least Common Venue	10th Least Common Venue
0	E11JQ	Art Gallery	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office	Park
1	E138EQ	Art Gallery	Hotel	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office
2	E13HZ	Art Gallery	Hotel	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office	Park
3	E146BJ	Art Gallery	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office	Park
4	E14LJ	Art Gallery	Hotel	Hotel Bar	Italian Restaurant	Japanese Restaurant	Korean Restaurant	Lebanese Restaurant	Men's Store	Middle Eastern Restaurant	Office

```
least_common_venues.shape
```

```
(84, 11)
```

On the other hand, the *kmeans* analysis of the most common venues resulted in 5 clusters where the largest one was Cluster 1 with Coffee Shop being a favorite there. There are a lot of pubs and hotels too and this is not at all out of the ordinary. Therefore, the study of both least and most common venues should be considered concurrently.

In terms of Average Income, after the *kmeans* analysis it was easy to recognize the three income ranges in the 3 resulting clusters. The majority of the districts (11 out of 24, or 45%) belonged to the middle range income Cluster 0. This could be read as a sign of a clear middle class in the London population. I should mention that it was encouraging to see that good 25% of the samples were classified into Cluster 1 which represented the highest income range with average revenue reaching up to GBP 73,000 per annum.

6. Discussion

The size and variety of information and datasets relating to London, available online, is immense. If one is considering trading goods or services in London, or is an existing business there, they have indefinite sources of details around all aspects of this megapolis. In particular, my choice of data for this analysis links to what I was trying to achieve, and that was to provide some visibility in regard to the existing competition in London, as well as the capabilities of the future consumers.

As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high-quality results on this occasion. For instance, *kmeans* analysis of the venues would be more insightful if it is applied to a larger area like all London districts, not only WC1R. That would provide an intriguing comparison as London districts very a lot in terms of many elements, including population habits, income and existing venues.

Knowing the specifics of the different London districts could support greatly current and potential investors. It is useful to know which areas are mainly touristic, residential, or purely business districts like the City and Canary Wharf. This largely determines the types of venues which would succeed in a district.

The current report aims to compliment any other research into London in the light of entrepreneurship consideration. On its own it would not suffice as there are many aspects that must be looked into in such a complex process, however it could play a key role in getting to know a future home for one's business.

7. Conclusion

Every year, London plays host to ever more craft, trade and entertainment. It is a collaborative, creative, fast-paced megapolis, with many networks and a lot of support for entrepreneurs. Naturally there is a lot of dynamics in this incredible city, meaning that business owners must stay alert to competition and consumer potential.