Finding the best location for start-up bar Coffee bar

Hanyang hu

May 6th, 2020

TABLE OF CONTENTS

1.0 INTRODUCTION

1.1 BACKGROUND

London is the capital of England and the most densely populated city in Britain. It is a fast-growing city, it has plenty of talent and it is a prime location for securing funding and investments. This makes it an ideal location for many start-up businesses from the food industry especially considering the London restaurant and bar industry has grown exponentially over the course of the last decade and is growing quite considerably. Every street is filled with a variety of restaurants servicing different cuisines more than ever before. However, this also means competition among restaurants has also peaked as there are many substitutes available for consumers. Location, is always a primary factor is determining whether a bar or restaurant will be successful. Many owners dream of opening at a location such as Piccadilly, Mayfair and Soho in Central London, however are they really the best locations? This study attempts to break the problem down and look at the different areas within London and help the client break into the market through securing a good location for opening their new business.

1.2 PROBLEM

In this assignment, we will take a look at the best locations to open their first store in London. The client is running a variety of bars in other parts of the world and would like to open their first coffee bar in London. London is a densely populated city with high competition among consumer goods. Thus, a good location is imperative for the businesses success and we will explore together these potential areas.

1.3 INTEREST

Our client in this case is a business seeking to open their first coffee bar in London. They have restaurants in other parts of the world but would like to use this opportunity to expand into the English market. They have significant funds to back up their start up so money will be less of a limitation in this case.

2.0 DATA ACQUISITION AND CLEANING

2.1 DATA SOURCES

A full list of Post codes with their corresponding latitude and longitude coordinates was found at: https://freemaptools.com. This is referred to as Dataset1.

Full list of Postcodes with corresponding Districts as reference was found at: https://www.milesfaster.co.uk/london-postcodes-list.htm. This is referred to as Dataset2.

In this assignment we will also be using data from Foursquare, a social networking service that provides geographical data. For more information on this software and their services, please refer to their website at: https://foursquare.com.

2.2 Data Cleaning

Firstly, the two data set namely Dataset1 and Dataset2 containing postal code information was transformed into a panda's data frame. This allowed for easy manipulation and organisation of the relational datasets into the formations we require. Individually, Dataset1 was split the data, re-combined using Python pandas and the columns titles were renamed to make for easy identification. We then drop the unneeded columns resulting in a data frame with only two columns 'Postcode' and 'District'. The second step is to merge this data frame with Dataset2 to find the Postcode with the corresponding District and the latitude/longitude coordinates in one table. By resetting the indexes and dropping a few unneeded columns, we create the table outlined in Figure 1.

Figure 1 – Table created from combining Dataset1 and Dataset2.

	Postcode	District	latitude	longitude
0	E1W	Wapping	51.50775	-0.05739
1	E2	Bethnal Green, Shoreditch	51.52939	-0.06080
2	E3	Bow, Bromley-by-Bow	51.52789	-0.02482
3	E4	Chingford, Highams Park	51.62196	-0.00339
4	E5	Clapton	51.55893	-0.05233
...	...	...	...	...
114	W10	Ladbroke Grove, North Kensington	51.52103	-0.21397
115	W11	Notting Hill, Holland Park	51.51189	-0.20424
116	W12	Shepherds Bush	51.50777	-0.22890
117	W13	West Ealing	51.51270	-0.31951
118	W14	West Kensington	51.49488	-0.20923

119 rows × 4 columns

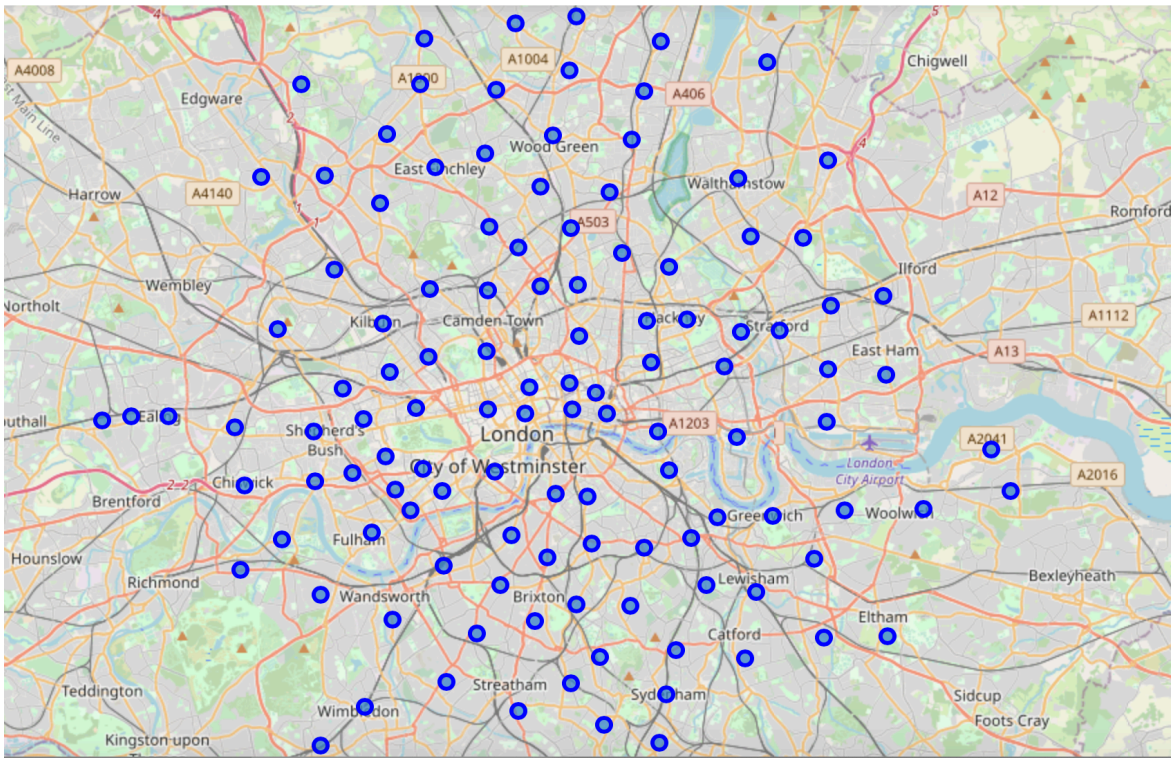
3.0 Methodology

3.1 Exploratory analysis

3.11 The Distribution of London's Districts

Using exploratory analysis, we first take a look at some initial parameters of the dataset. We use the describe function to envision some facts about this set and found the count, mean and standard deviation. We want to be able to dive deeper and look at the distribution of the dataset across London and thus, using the merged data we input the latitude, longitude coordinates along with the names of each district and plotted using Folium on a map of London shown in Figure 2.

Figure 2 – Districts of London plotted using Folium



3.12 Common venues based on District

We explore the most common/popular venues for every district. In order to achieve this, we first use the zip function to create an iterator of tuple of and using three columns namely 'District', 'Latitude', and 'Longitude' of the merged data frame. Using a for loop, we loop through each tuple, appending the name, coordinates and its category to an empty list. Using the group by function we find variance to be significant across the whole dataset based on the number of venues present within each district.

We then use one hot encoding on the venue categories and took an average for the number of occurrences for every district. We append the top most common venues within every

district to a list and convert this list into a data frame. The resulting table can now be used as an input to our machine learning modelling tools.

4.0 K-MEANS CLUSTERING

4.1 FOURSQUARE

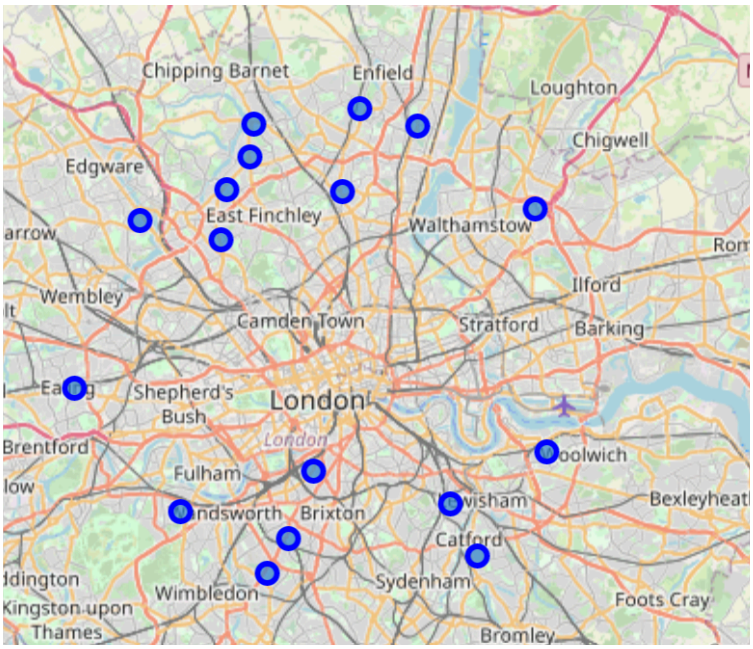
4.1.1 EXPLORE

The objective here using k-means clustering is to find similarities between districts based on the most popular venues around its area. After finding the clusters using the kmeans object from the sklearn library, we attached the corresponding labels it the corresponding data frame. We then look at each cluster individually to try and identify patterns that would help us to identify clusters within London where bars appear more common as a popular attraction. The number of bars in the area is assumed to be directly correlated with the popularity of that type of venue.

We notice that most of the districts in cluster 0 contained coffee roasters/shops as one of the most common venue in that area. This is in itself a likely indication that these have become more popular in the area than other districts. However, this information is not enough and we want to be able analyse further and look into cluster one for more information.

We can also visualise the spread of the cluster using Folium as shown in Figure 3. It appears the districts in Cluster one is quite evenly distributed across London.

Figure 3 – Cluster 0 London distribution



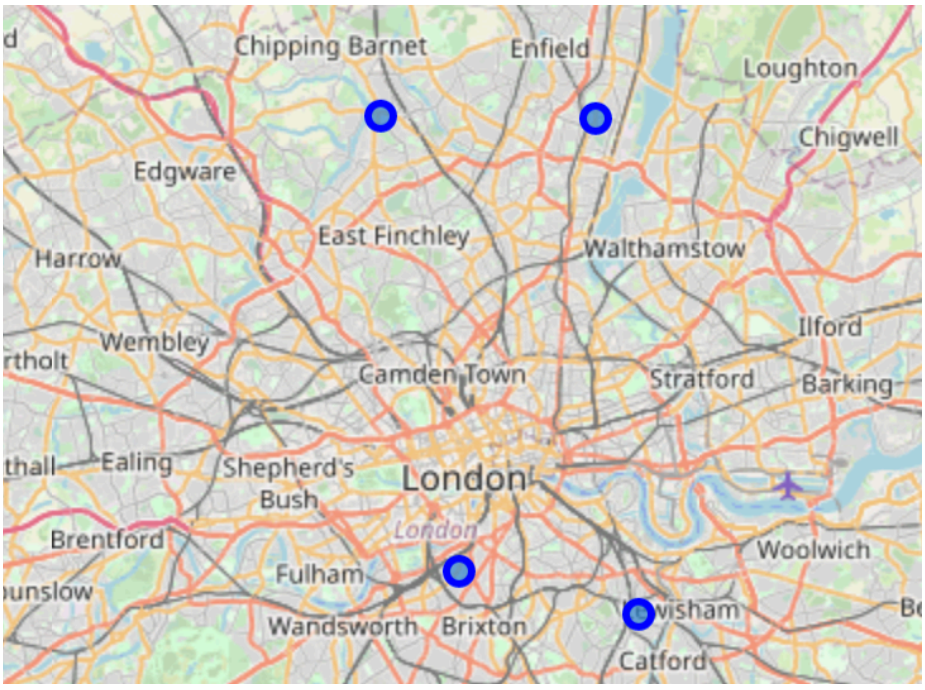
4.1.2 SEARCH

The goal here is to find, quantitatively, the number of coffee shops present within each district of Cluster 0. To do this, we use the search end point along with a query. We then merge and take the segment of the data frame containing only rows from cluster 0 with the district's corresponding latitude and longitude coordinates. Using a similar function used for the explore end point, this information was appended to a list and convert into a data frame for easy visualisation and manipulation. Using the group by method, we find the number of coffee bars within the area. We look deeper into each individual cluster and find that cluster one contains only those where the number of bars is more than 4. Therefore, competition is likely to be fierce in these locations. Cluster 0, contain only locations with less than 3 bars. This maybe an indication that bars are not as popular as a venue in these locations due to a lack of demand or other reasons and thus, not prime locations for the client to start their business.

Finally, we take out cluster 2 from the data frame where there is an above average number of coffee related shops in the location meaning higher than average demand, but not to the extent where there may be high barriers of entry due to the immense competition of already established brands operating in close proximity. Therefore, cluster 2 would logically be the best option for the client's start-up.

5.0 RESULTS

Figure 4 – Optimum locations



As shown in Figure 4, most points are focused on the outskirts of London, even moving into other towns. We have now trimmed our data down to the districts where the study feels would be the most optimum locations. However, many factors other than location goes into play when opening a business such as budget, target customers, marketing, competition and business plan. The client can now define these parameters themselves and with the help of this assignment find the best location. Points closer to the city or central London may be a popular attraction for bars and thus, in theory be a good location, however as this study suggests, the huge competition within those areas due to a larger number of already established bars may cause high barrier to entry.

6.0 DISCUSSION

Through exploring the dataset, we've established a list of locations that are good locations for a start-up coffee bar business. The client can now review these locations and based on their criteria, preferences and defined parameters, conclude on which location they think will best fit all their requirements. However, there are still many limitations to this study such as we have not considered the actual demand for bars in each of the districts. This parameter is difficult to measure and thus, not considered for our study. The other limitation is that, many areas within Central London may not have bars as the most common venue as a result of restaurant and cafes being more prominent. However, this does not mean this isn't significant demand for bars in these areas especially in the centre of the city. This shows why demand is an important parameter to consider. If we can establish a way of determining or estimating the actual demand for a product in an area, we can then more freely and accurately determine which locations are the best for its target markets. The client can even utilise this report for understanding the distribution of locations where bars are more prominent and analyse these locations individually further using other machine learning algorithms. A deeper look into the individual bars, we can identify anomalies and outliers that are skewing the results such as a situation where many chain businesses operating in the same locations are inflating the number of bars in the area when in reality they all belong to one single entity.

7.0 CONCLUSION

In this report, we looked at different datasets, conducted exploratory analysis and using k-means clustering, we were able to identify good locations for a start-up coffee bar business. We interpret the results and their meaning, and applied real world logic to the problem in order to form an argument backed by evidence produced in this study. In conclusion, the results shown in this study gives an indication, however are subjected to a few limitations which can be overcome using further research.