

Analysing London for Opening a Pub

Kaan Corbacioglu

June 02, 2021

1. Introduction

1.1 Background

London is one of the most multicultural cities in the world. It is a great tourist attraction and busy at all seasons of the year. The city offers a wide range of restaurants and cuisines. There are currently 8.9m people living in the city. Pubs are a big part of the English culture. The objective of this project is to recommend the perfect spot in London to open a pub. Since there is a pub in every neighbourhood of London, I will be using the Foursquare API along with data scrapped from the web to identify the most suitable location to open a pub in London. I will be using the clustering technique to group neighbourhoods according to the occurrence of pubs in that area. Then I will use some plots to determine the desired location for opening this pub. The analysis in this project is aimed to target business owners who are interested in opening a pub in London.

1.2 Problem

There are over 3500 pubs just in London. This project aims to find the best location for opening a pub using the foursquare API. The prediction will give a good idea where the pub density is less hence more chance of business.

1.3 Interest

The target audience of this project are business owners who are looking to open a pub or expand in London. The results of this project might also be in interest to investors.

2. Data acquisition and cleaning

2.1 Data Source

In order to determine the venues in London in this project web scraping method was used. A Wikipedia page that includes all the London Boroughs/Neighbourhoods and postcodes was utilized.

Source: https://en.m.wikipedia.org/wiki/List_of_areas_of_London

Location	London borough	Post town	Postcode district	Dial code	OS grid ref
Abbey Wood	Bexley, Greenwich ^[7]	LONDON	SE2	020	TQ465785
Acton	Ealing, Hammersmith and Fulham ^[8]	LONDON	W3, W4	020	TQ205805
Addington	Croydon ^[8]	CROYDON	CR0	020	TQ375645
Addiscombe	Croydon ^[8]	CROYDON	CR0	020	TQ345665
Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728
Aldborough Hatch	Redbridge ^[9]	ILFORD	IG2	020	TQ455895
Aldgate	City ^[10]	LONDON	EC3	020	TQ334813

Figure 1: Data Source

2.2 Data Cleansing

Data on this source was scraped using “Beautiful Soup” library and put in to a dataframe. Following the scraping, geopy library was used to gather latitude and longitude data which was merged with the dataframe. Information such as “Dial code” or “OS grid ref” was not in the scope of this project. Hence they were dropped.

There were several problems with the data on this source. First problem was that the location information it contained ranged an exceptionally large are of London inclusive of Greater London. So, the data was filtered to Towns that only contained London area in the table.

Second problem was that some Locations contained the format “(also known as xxx)”. This extra phase resulted in “geopy” to return null values for latitude and longitude information. The work around for this was to drop the phrase from these locations and re-run geopy.

Third problem was that some location names were too generic. For example “Aldwych” is also a name of a place in Derby UK. Hence geopy was returning the wrong latitude and longitude values. These locations were picked up by defining a simple for loop which looked for latitude and longitude values that were outstanding from the rest. The quick fix was to add “London” after the ending of the location name. This ensured geopy to pick up the correct latitude and longitude values.

The fourth problem was that locations such as “Southend” were still being mis identified by geopy even after adding “London” to the naming. The reason for this was that Southend Airport is considered one of London’s airport even tough it is in Southend which is East of the country. So, adding “London” after Southend did not make a difference since Southend London Airport is in Southend. The way to fix this was to add “Lewisham” to the end of

“Southend London” to ensure that the correct location was picked. Lewisham is the name of the borough that Southend London is a part of.

After cleaning the data the final location information was merged with the existing data that was scrapped from the URL. The final dataframe looked like:

	Locations	London Boroughs	Towns	Post Codes	Latitudes	Longitudes
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	51.487621	0.114050
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	51.508140	-0.273261
2	Aldgate	City[10]	LONDON	EC3	51.514248	-0.075719
3	Aldwych	Westminster[10]	LONDON	WC2	51.512625	-0.118568
4	Anerley	Bromley[11]	LONDON	SE20	51.407599	-0.061939

Figure 2: Final version of dataframe

2.3 Feature Selection

As mentioned in section 2.2 the data was filtered further to towns that only contained “London”. This is approximately equivalent to Zone 1-4 in London.

3. Methodology

In this project the data was scraped from the web and put into a dataframe. After using geopy library and cleansing the data. The locations were mapped using the folium library as can be seen in Figure 3.

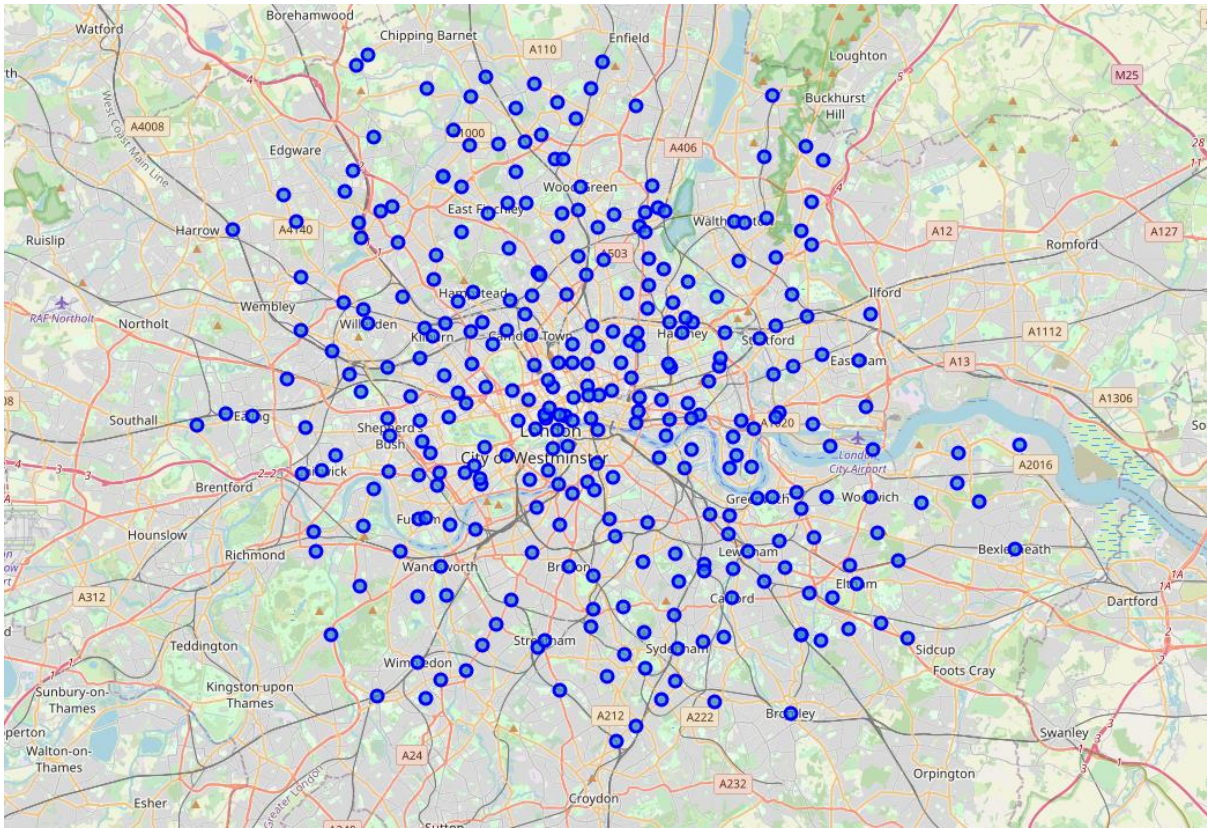


Figure 3: Locations in London

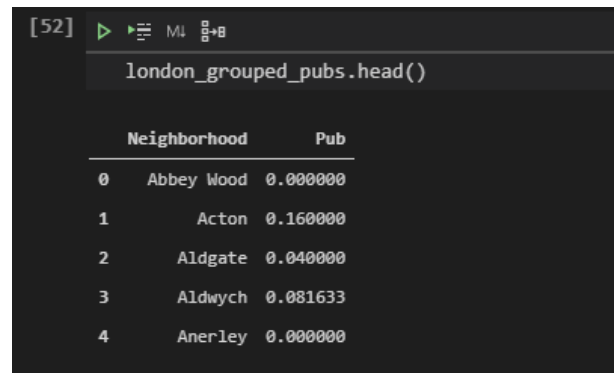
Foursquare API was used to gather venues that are within a radius of 500m within the locations. The results were placed in a new dataframe. In total there was 9748 venues that the API returned and 389 unique categories. To better understand the most popular venues in London, the data was grouped by category (Figure 4).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Coffee Shop	632	632	632	632	632	632
Pub	623	623	623	623	623	623
Café	567	567	567	567	567	567
Hotel	336	336	336	336	336	336
Grocery Store	331	331	331	331	331	331

Figure 4. Top 5 Categories in London

In Figure 4 we can see that Pubs are among the top attraction in London. This data is a indication that opening a pub in London is a feasible thing to do.

Next I used the onehot method to categorical values into numerical values and calculate the frequency of pub appearance in each neighbourhood. This method is used to convert categorical data to numerical data and take the average frequency of the appearance of pubs.



```
[52] london_grouped_pubs.head()
```

	Neighborhood	Pub
0	Abbey Wood	0.000000
1	Acton	0.160000
2	Aldgate	0.040000
3	Aldwych	0.081633
4	Anerley	0.000000

Figure 5. London Pubs Frequency

In Figure 5 it can be seen the first 5 elements of the dataframe with the calculated average occurrence of pubs per neighborhood.

3.1 Machine Learning

In order to better understand the data I am working with it was necessary to group them together to come up with some idea of how the areas look like. The next step was to group all the neighbourhoods that had a similar frequency. There are ways of doing this but in this project, I used the k-means clustering method. K-means is a method that runs in a loop and re- calculates the central point of the cluster. The most important part is to define how many clusters we need to use. The best approach for such data in this project is the elbow method.

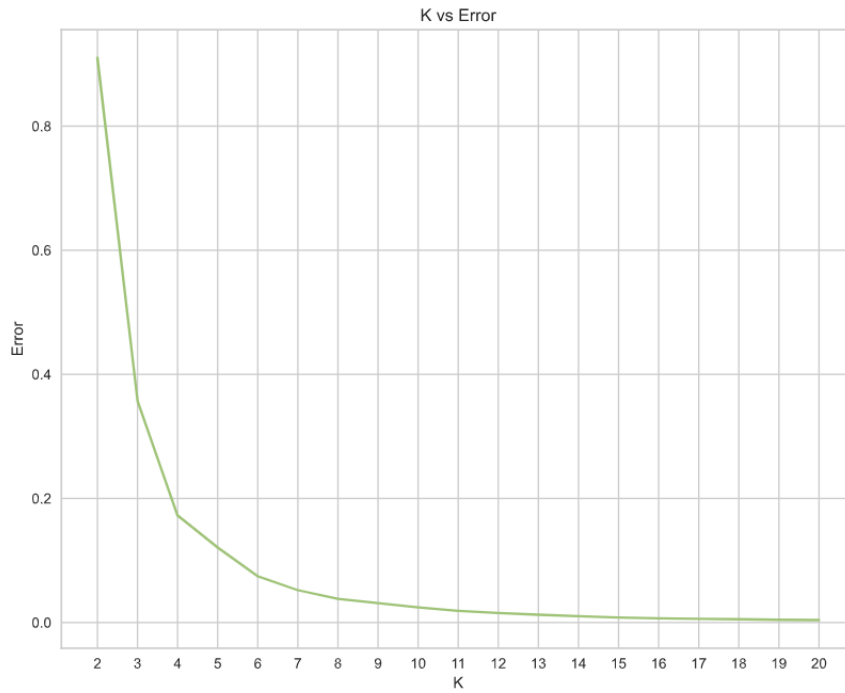


Figure 6. K vs Error Graphic

In this graphic we can tell that the best k value to use will be between 3-6 as this is where the graphic is breaking directions. To better understand the value I used a visualizer to fit the model imported from yellowbrick library. The result is as in Figure 7.

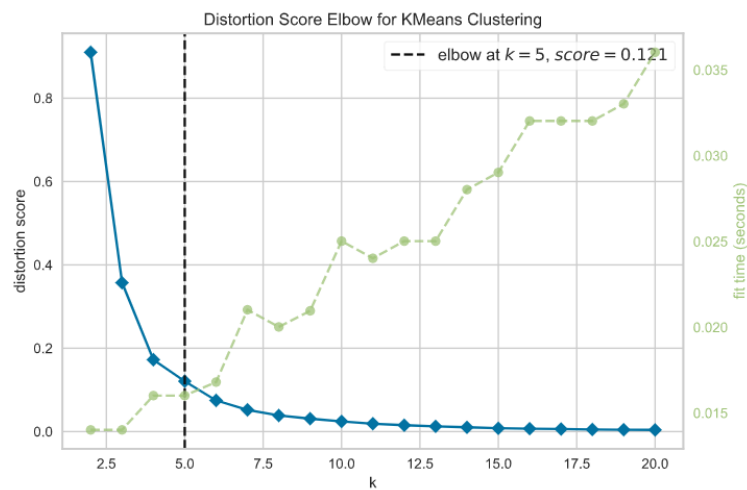


Figure 7. Distortion Score

Figure 7 tells us that the intersection point is at “5” and displays the time it will take to compute using k=5 clusters. So the best cluster number to choose for this data set is given

as 5. This concludes that we will have 5 clusters [0,1,2,3,4] and neighbourhoods with similar frequency of pub occurrence will be grouped in the same cluster.

Next I added a “Cluster Labels” column to the dataframe that contains the frequency of pubs and merged it with the dataframe that contains the venues I got using the Foursquare API.

	Neighborhood	Pub	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	0.00	0	51.487621	0.114050	Co-op Food	51.487490	0.113751	Grocery Store
0	Abbey Wood	0.00	0	51.487621	0.114050	Bostal Gardens	51.486670	0.110462	Playground
0	Abbey Wood	0.00	0	51.487621	0.114050	Abbey Wood Caravan Club	51.485502	0.120014	Campground
1	Acton	0.16	3	51.508140	-0.273261	London Star Hotel	51.509624	-0.272456	Hotel
1	Acton	0.16	3	51.508140	-0.273261	The Aeronaut	51.508376	-0.275216	Pub

Figure 8. Merged Dataframe

Following this step I used the folium library again to map each Neighbourhood with respect to the cluster label it has been assigned. The result is as below in Figure 9.

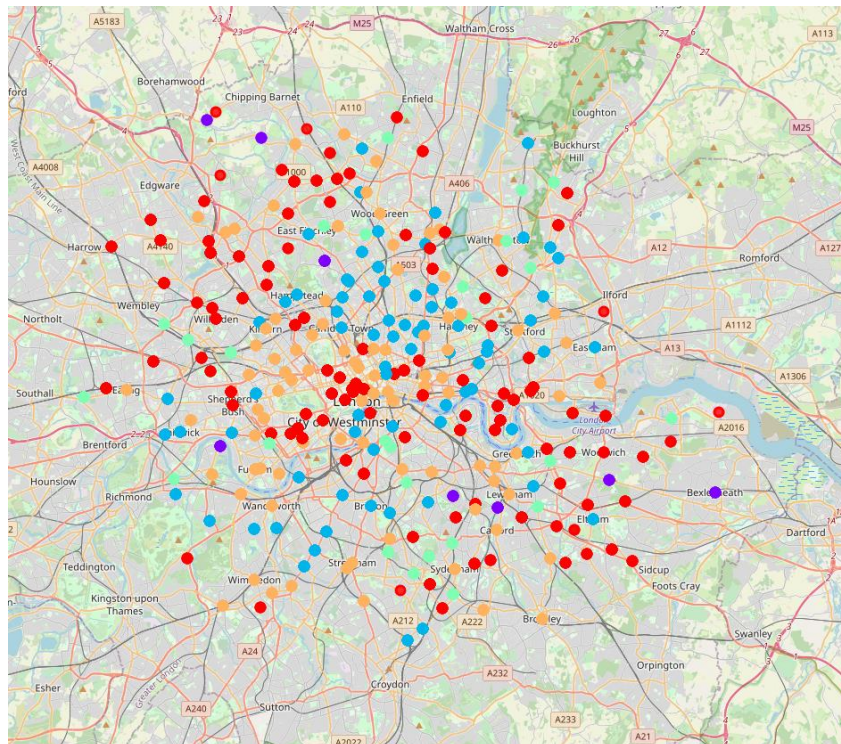


Figure 9. Cluster Map

As it can be seen there are 5 different colours for each cluster on the map. Red, Purple, Blue, Green and Orange represents cluster 1 to 5, respectively.

4. Results

The elbow method returned 5 clusters. To better understand what I am working with, plotted the clusters on three a graphs (Figure 10, 11 & 12). In Figure 10 most of the neighbourhoods are in cluster 1 (red) followed by cluster 5 (orange), cluster 3 (blue), cluster 4 (green) and cluster 2 (purple). Obviously we can see the similar relationship in Figure 12 where cluster 5 consist of the most number of venues.

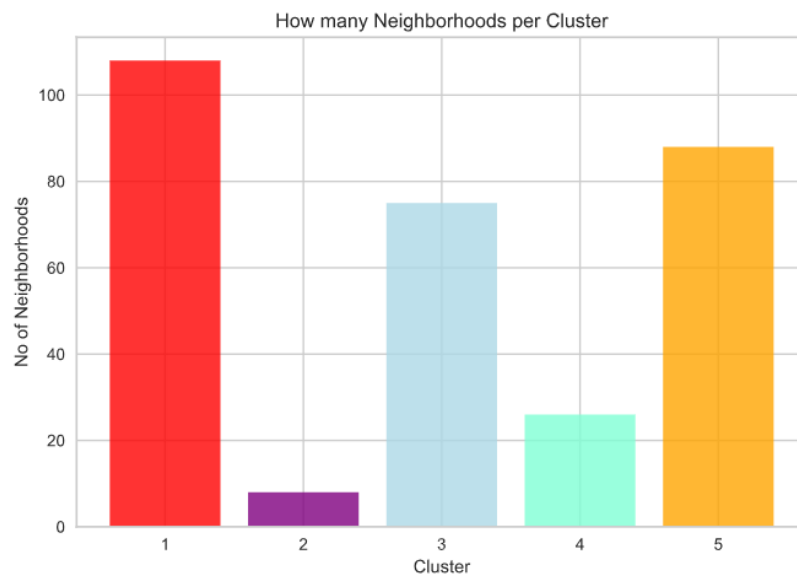


Figure 10. Number of Neighbourhoods per Cluster

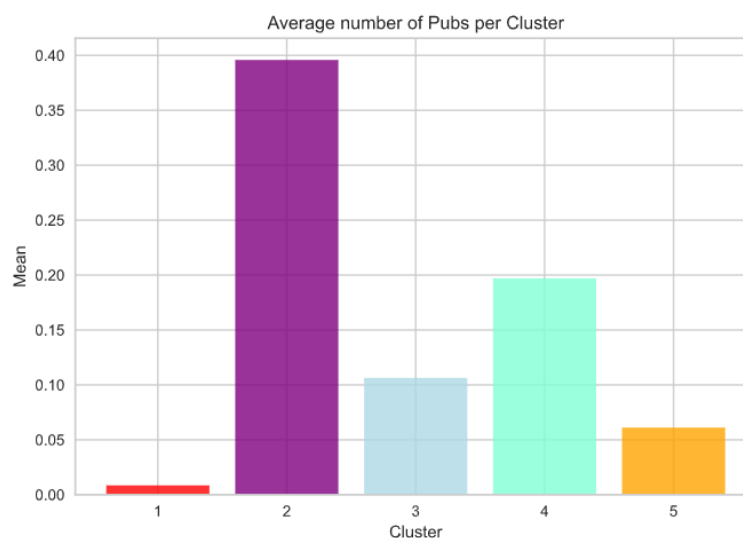


Figure 11. Frequency of Pubs per Cluster

In Figure 11 I plotted the occurrence of pubs per cluster where cluster 2 represents the highest frequency of pubs.

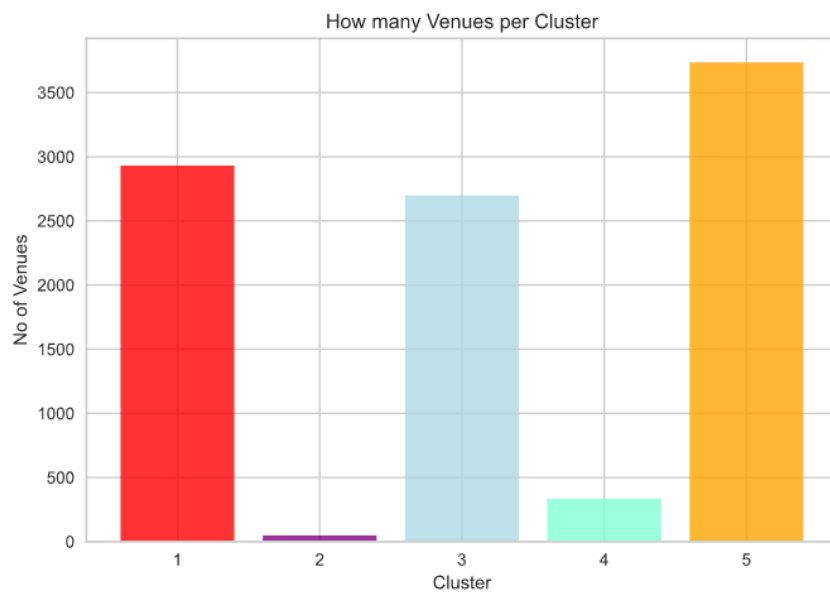


Figure 12. Number of Venues per Cluster

Cluster	Number of Neighbourhoods	Mean value of pubs	Venues returned	Pubs
Cluster 1	108	0.0085	2931	29
Cluster 2	8	0.3958	48	19
Cluster 3	75	0.1064	2698	289
Cluster 4	26	0.1970	335	66
Cluster 5	88	0.0612	3826	242

Table 1

4.1. Analysis of Clusters

When we compare Figure 10,11,12 we can see that most of the venues returned are in cluster 1 (2931) & cluster 5 (3826). But an important fact is that the average number of pubs for cluster 1 & 5 are the lowest 0.0085 & 0.0612.

Also it can be seen in Figure 9 that the map is mostly dominated by cluster 5 & 2 all across London but more dense towards the city centre.

4.1.1 Cluster 1

Cluster 1 is mostly in Northwest, City Centre, Southeast of the city with some out layers in Northeast and Southwest. It can be clearly seen that the density of Cluster 1 is highest in Chelsea towards Westminster. Only 29 venues that were returned were in the category of pub hence the frequency of pubs in this cluster is extremely low. Please see Table 1.

	Neighborhood	London Boroughs	Pub	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	Bexley, Greenwich [7]	0.0	0	51.487621	0.114050	Co-op Food	51.487490	0.113751	Grocery Store
1	Abbey Wood	Bexley, Greenwich [7]	0.0	0	51.487621	0.114050	Frank's Fish Bar	51.487151	0.112231	Fish & Chips Shop
2	Abbey Wood	Bexley, Greenwich [7]	0.0	0	51.487621	0.114050	Abbey Wood Caravan Club	51.485502	0.120014	Campground
3	Anerley	Bromley[11]	0.0	0	51.407599	-0.061939	Aldi	51.408604	-0.059213	Supermarket
4	Anerley	Bromley[11]	0.0	0	51.407599	-0.061939	Co-op Food	51.406243	-0.057830	Grocery Store
5	Anerley	Bromley[11]	0.0	0	51.407599	-0.061939	Tesco Express	51.405853	-0.065041	Grocery Store
6	Anerley	Bromley[11]	0.0	0	51.407599	-0.061939	H&M Kebab House	51.409031	-0.059266	Fast Food Restaurant
7	Anerley	Bromley[11]	0.0	0	51.407599	-0.061939	Costcutter	51.405990	-0.057141	Convenience Store
8	Anerley	Bromley[11]	0.0	0	51.407599	-0.061939	Betts Park	51.408755	-0.067278	Park
9	Arkley	Barnet[12]	0.0	0	51.645583	-0.236258	Arkley Golf Club	51.647774	-0.233413	Golf Course

Figure 13. Cluster 1

4.1.2 Cluster 2

Cluster 2 is spread out across the city and does not follow any pattern in terms of density. It has the lowest number of neighbourhoods and venues returned. However, it has the highest number of frequency for pubs. Cluster 2 has 19 out of 48 venues which are in the pub category which 0.3958 mean values. In other words, 4 out of 10 venues in cluster 2 is a pub. This is interesting data that shows the areas in cluster 2 there are not many venues but most of them are pubs. There are 21 unique categories in cluster 2.

	Neighborhood	London Boroughs	Pub	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barnet Gate	Barnet	0.333333	1	51.641827	-0.242985	The Gate	51.641994	-0.242582	Pub
1	Barnet Gate	Barnet	0.333333	1	51.641827	-0.242985	Hadley FC	51.642660	-0.243032	Soccer Field
2	Barnet Gate	Barnet	0.333333	1	51.641827	-0.242985	Barnet Gate Wood	51.639062	-0.242836	Forest
3	Bexleyheath (also Bexley New Town)	Bexley[26]	0.400000	1	51.463485	0.148055	The Traveller's Home	51.463449	0.150954	Pub
4	Bexleyheath (also Bexley New Town)	Bexley[26]	0.400000	1	51.463485	0.148055	The Yacht	51.466340	0.145112	Pub
5	Bexleyheath (also Bexley New Town)	Bexley[26]	0.400000	1	51.463485	0.148055	Co-op Food	51.465449	0.146523	Grocery Store
6	Bexleyheath (also Bexley New Town)	Bexley[26]	0.400000	1	51.463485	0.148055	Russell Park	51.463112	0.148876	Playground
7	Bexleyheath (also Bexley New Town)	Bexley[26]	0.400000	1	51.463485	0.148055	One Stop	51.465762	0.146127	Convenience Store
8	Castelnau	Richmond upon Thames[43]	0.428571	1	51.485688	-0.233030	Sips'n'Bites	51.485214	-0.233585	Café
9	Castelnau	Richmond upon Thames[43]	0.428571	1	51.485688	-0.233030	Spoonful	51.485313	-0.233358	French Restaurant

Figure 14. Cluster 2

4.1.3 Cluster 3

In terms of appearance, cluster 3 is very similar to cluster 1 appearing across the entire city but more dense in the central. It has the third highest number of venues and neighbourhoods also the third highest frequency for pubs. The average of pubs in this cluster is 0.1064. The number of venues returned is 2698 where 276 of these are unique categories. In total there is 289 pubs in this cluster and 75 neighbourhoods.

	Neighborhood	London Boroughs	Pub	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	London Star Hotel	51.509624	-0.272456	Hotel
1	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	The Aeronaut	51.508376	-0.275216	Pub
2	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	McBakeme	51.508452	-0.268543	Creperie
3	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	Dragonfly Brewery at George & Dragon	51.507378	-0.271782	Brewery
4	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	Acton Centre	51.506608	-0.266878	Gym / Fitness Center
5	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	Amigo's Peri Peri	51.508396	-0.274561	Fast Food Restaurant
6	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	PureGym	51.508326	-0.277150	Gym / Fitness Center
7	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	Park+Bridge	51.508382	-0.267084	Wine Shop
8	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	The Talbot	51.506527	-0.273585	Pub
9	Acton	Ealing, Hammersmith and Fulham[8]	0.148148	2	51.50814	-0.273261	Lidl	51.507793	-0.270210	Supermarket

Figure 15. Cluster 3

4.1.4 Cluster 4

Cluster 4 has the second highest frequency for pubs (0.1970) and fourth highest venues/neighbourhoods returned. Cluster 4 is more populated in South London but not in Zone 1 which is city centre. A total of 335 venues are in cluster 4 and 66 of these are pubs. There are 94 unique categories in cluster 4.

	Neighborhood	London Boroughs	Pub	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	Mara Interiors & Café	51.477672	0.019368	Furniture / Home Store
1	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	M&S Simply Food	51.476772	0.020189	Grocery Store
2	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	The British Oak	51.476161	0.026150	Pub
3	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	Standard Fish Bar	51.476722	0.021685	Fast Food Restaurant
4	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	The Royal Standard	51.476717	0.018799	Pub
5	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	Moca	51.476833	0.021119	Deli / Bodega
6	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	Sun Ya	51.476918	0.019469	Chinese Restaurant
7	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	Royal Grill Cafe	51.476910	0.020881	Café
8	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	Blackheath Complementary Healthcentre	51.477579	0.019344	Spa
9	Blackheath Royal Standard	Greenwich	0.25	3	51.478392	0.021284	Brother's Bakery	51.477043	0.018729	Bakery

Figure 16. Cluster 4

4.1.5 Cluster 5

Cluster 5 has the highest number of venues returned (3826) and also the second highest number of pubs (242). However, the frequency of pubs is 4th highest slightly above cluster 1. Having the second highest number of neighbourhoods it is spread across the city but denser in the city centre. There are 294 unique categories in the 88 neighbourhoods it contains. Having the highest number of venues yields it to have a very low frequency for pubs (0.0612).

	Neighborhood	London Boroughs	Pub	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aldgate	City[10]	0.04	4	51.514248	-0.075719	Dorsett City London	51.514036	-0.075812	Hotel
1	Aldgate	City[10]	0.04	4	51.514248	-0.075719	The Association	51.513733	-0.079132	Coffee Shop
2	Aldgate	City[10]	0.04	4	51.514248	-0.075719	Hotel Indigo	51.512740	-0.075920	Hotel
3	Aldgate	City[10]	0.04	4	51.514248	-0.075719	Katsu Wrap	51.515883	-0.077849	Food Truck
4	Aldgate	City[10]	0.04	4	51.514248	-0.075719	Discount Suit Company	51.516705	-0.075506	Cocktail Bar
5	Aldgate	City[10]	0.04	4	51.514248	-0.075719	Benk + Bo	51.515731	-0.075875	Bakery
6	Aldgate	City[10]	0.04	4	51.514248	-0.075719	Hotel Motel One London-Tower Hill	51.512635	-0.075513	Hotel
7	Aldgate	City[10]	0.04	4	51.514248	-0.075719	1n1 Fashion Pizza	51.516037	-0.075865	Pizza Place
8	Aldgate	City[10]	0.04	4	51.514248	-0.075719	1Rebel	51.515569	-0.080040	Gym / Fitness Center
9	Aldgate	City[10]	0.04	4	51.514248	-0.075719	Mattarello	51.515518	-0.075434	Italian Restaurant

Figure 17. Cluster 5

5. Discussion

Most of the pubs are in cluster 3 which is shown in blue. Cluster 3 is more towards Hackney area in London. It seems like areas in cluster 3 would not be so good to start a new business for pubs as the competition is already quite high. Even though cluster 1 has a remarkably high number of venues and the highest number of neighbourhoods, the number of pubs is very little. Therefore the locations in cluster 1 would be an ideal for non-competitive environment. With 108 neighbourhoods and only 29 Pubs cluster 1 is the best choice for starting a new business. On the other hand, it wouldn't make sense to open up in cluster 2. The number of venues is 48 which indicates that the area is not popular for people to go dining or drinking. Also, a very high percentage ~40% of the venues are already pubs.

One drawback of the analysis is that it doesn't take into account facts like, the ethnicity of the area. London is a multicultural city and analysing the area separately might change the outcome of the analysis. The recommendation in this project is based on the Foursquare API and density of pubs in areas.

6. Conclusion

In conclusion in this project, central London was swept using the Foursquare API. The areas were put into clusters depending on the occurrence of pubs. The results were visualized using graphics and maps. From a business perspective it was concluded that the best area to open a pub would most likely be in cluster 1.

One of the drawbacks was the depth of the analysis. In future the project can be improved by adding in analysis such as; ethnicity, business locations, pricing etc. However, these were not in the scope or timeline of this project.