Market analysis of the districts of London

Introduction

London is one of the biggest, most diverse and expensive cities in Europe consisting out of 32 boroughs including 12 inner and 20 outer boroughs as well as the City of London that stands out as the economical and historical center of the city.

In this analysis we aim to give a market overview of the different parts of the city that can be read as a guideline for entrepreneurs who want to start a business, e.g. open a pub, supermarket or a restaurant and try to find the right location for their project.

For this purpose we cluster the 32 boroughs into 4 different and unique groups that differ among other aspects by the size of population, mean salary and venues present in the area. Thereby, each group of boroughs offers a different kind of market opportunity. The clustering algorithm itself is used to reduce the complexity of the original problem.

Data description

The data we use is obtained from three different sources. First, a table taken from kaggle (*https://www.kaggle.com/justinas/housing-in-london*) containg information about the mean salary, population size, life satisfication, number of jobs and area size of the 32 boroughs from 1991 to 2019. These data mainly help us to group the districts into clusters which have their own descriptive profile.

Second, we use geographical data provided by Wikipedia to obtain the locations of the boroughs in longitude and latitude (*www.wikipedia.org/wiki/List_of_London_boroughs*). This information we finally employ to obtain a list of the nearby venues in the districts using the Foursquare API. Here, we set the limit of returned venues to 100 per borough. Moreover we choose a radius of 1500m for the search, expecting that the obtained section stands representative for the whole borough.

We merge the data into a single table containing all information. Here we restrict our analysis to the year 2018 for which most of the information is available while still being current.

During the data processing phase we noticed that some data for life satisfaction and the population size were missing. We replaced these entries by the mean value of the available observations.

Methodology

In this section we describe how the analysis is conducted and which methods were used. After giving a brief graphical overview of the data, i.e. visualizing the number and category of venues as well as the location of the boroughs, we employ the k-means algorithm to group

the 32 boroughs into clusters that are characterized by different market situations offering distinct business opportunities.

The k-means method is a non-supervised learning algorithm meaning that it works with unlabeled data. In short, the algorithm clusters a given set of data points into k groups.

In a first step k cluster centers are chosen within the algorithm. This can be done randomly or according to an educative guess. In a second step each data point is assigned to its nearest cluster center and finally new cluster centers are selected with coordinates given by the mean values of all data points in the cluster. The algorithm is repeated until convergence is obtained.

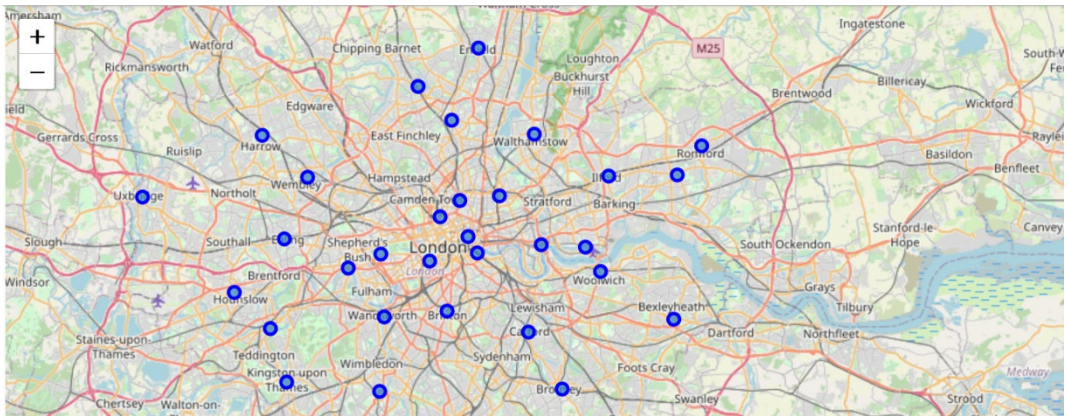
As the best choice for k is not known from the beginning we use the elbow method to obtain a good choice for k. The idea of the elbow method is to plot the mean distance of all data points to its cluster centroid as a function of k. Generically, the cumulated mean distance falls with increasing value of k giving us no information about the best choice for the cluster size. However, empirically, it turns out that a good choice of k is given at the point where the curvature of the cumulated mean distance changes sign or the curve of the cumulated mean distance features a kink.

As features for the k means algorithm we use as numerical values the mean salary, population size, life satisfaction, number of jobs, area size of each borough. To improve on the performance of the k-means algorithm we employ feature scaling provided by the scikit learn library which allows to obtain a feature set with zero mean and standard deviation.

Moreover, we use the Foursquare API to obtain a list of venues in a radius of 1500m around the center of each borough. In order to use this information in the k-means algorithm we employ the one-hot-vector encoding approach.

Analysis

To get acquainted with the problem we show in Fig. 1 a map of London where the 32 boroughs of the city are highlighted with blue markers. One can observe that about one-third of the boroughs are rather central while two-thirds are on the outskirts.



1 Figure 1: Map of London obtained using the Foursquare API where the 32 boroughs oft he city are highlighted with blue markers.

An immediate guess could be that there is a qualitative difference between central and less central boroughs. We will address this question in more detail at the end of this section.

In Tab. 1 all boroughs are listed by name including information about life satisfaction, mean salary, population size, number of jobs, area size and geographical coordinates in longitude and latitude that were used to mark the positions of the boroughs in Fig 1.

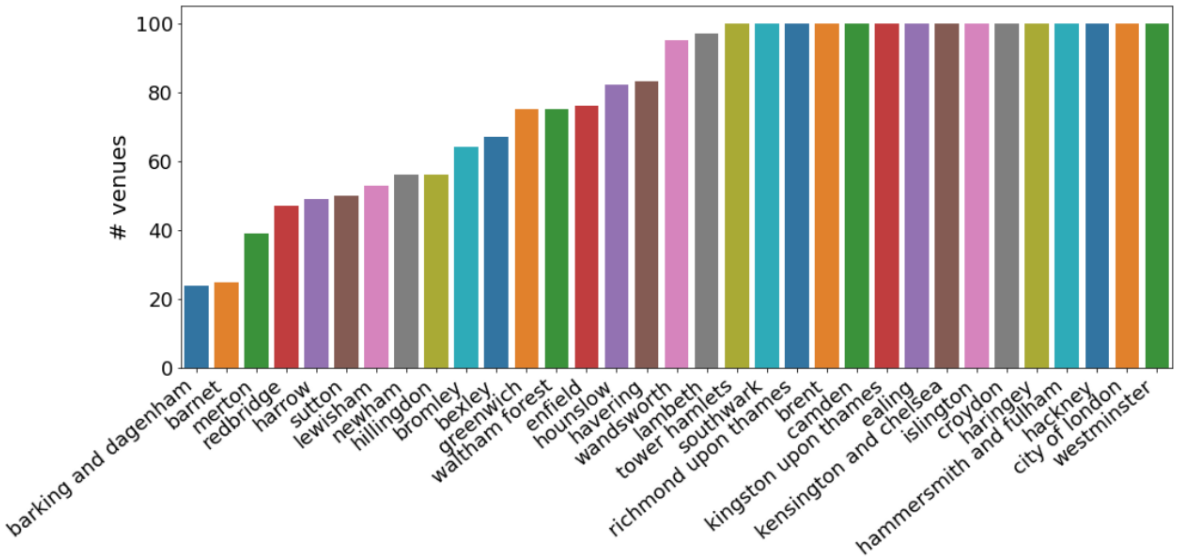
	Borough	life_satisfaction	mean_salary	population_size	number_of_jobs	area_size	Longitude	Latitude
0	city of london	7.583437	90028.00000	8706.0	640000.0	315.0	-0.0922	51.515
1	barking and dagenham	7.520000	32671.00000	211998.0	66000.0	3780.0	0.1557	51.560
2	barnet	7.550000	36776.00000	392140.0	170000.0	8675.0	-0.1517	51.625
3	bexley	7.540000	34496.00000	247258.0	87000.0	6429.0	0.1505	51.454
4	brent	7.710000	35830.00000	330795.0	157000.0	4323.0	-0.2817	51.558
5	bromley	7.570000	35201.00000	331096.0	127000.0	15013.0	0.0198	51.403
6	camden	7.480000	46502.00000	262226.0	403000.0	2179.0	-0.1255	51.529
7	croydon	7.690000	36880.00000	385346.0	149000.0	8650.0	-0.0977	51.371
8	ealing	7.530000	36304.00000	341982.0	148000.0	5554.0	-0.3089	51.513
9	enfield	7.370000	37030.00000	333869.0	133000.0	8220.0	-0.0799	51.653
10	greenwich	7.540000	36288.00000	286186.0	106000.0	5044.0	0.0648	51.489
11	hackney	7.500000	40439.00000	279665.0	152000.0	1905.0	-0.0553	51.545
12	hammersmith and fulham	7.720000	44781.00000	185426.0	155000.0	1715.0	-0.2339	51.492
13	haringey	7.350000	34178.00000	270624.0	91000.0	2960.0	-0.1119	51.600
14	harrow	7.650000	35325.00000	250149.0	91000.0	5046.0	-0.3346	51.589
15	havering	7.680000	34484.00000	257810.0	108000.0	11446.0	0.1837	51.581
16	hillingdon	7.760000	39670.00000	304824.0	209000.0	11570.0	-0.4760	51.544
17	hounslow	7.850000	46102.00000	270782.0	178000.0	5659.0	-0.3680	51.474
18	islington	7.510000	53489.00000	239142.0	255000.0	1486.0	-0.1022	51.541
19	kensington and chelsea	7.230000	42099.00000	156197.0	164000.0	1238.0	-0.1947	51.502
20	kingston upon thames	7.780000	36282.00000	175470.0	96000.0	3726.0	-0.3064	51.408
21	lambeth	7.220000	44814.00000	325917.0	180000.0	2725.0	-0.1163	51.460
22	lewisham	7.490000	33493.00000	303536.0	84000.0	3532.0	-0.0209	51.445
23	merton	7.840000	42062.28125	206186.0	108000.0	3762.0	-0.1958	51.401
24	newham	7.540000	37849.00000	352005.0	139000.0	3857.0	0.0469	51.507
25	redbridge	7.540000	31990.00000	303858.0	92000.0	5644.0	0.0741	51.559
26	richmond upon thames	7.700000	40427.00000	196904.0	111000.0	5876.0	-0.3260	51.447
27	southwark	7.340000	48333.00000	317256.0	329000.0	2991.0	-0.0804	51.503
28	sutton	7.880000	32442.00000	204525.0	78000.0	4385.0	-0.1945	51.361
29	tower hamlets	7.820000	69806.00000	317705.0	333000.0	2158.0	-0.0059	51.509
30	waltham forest	7.460000	32875.00000	276700.0	88000.0	3881.0	-0.0134	51.590
31	wandsworth	7.650000	45317.00000	326474.0	147000.0	3522.0	-0.1910	51.456
32	westminster	7.660000	63792.00000	255324.0	775000.0	2203.0	-0.1372	51.497

1 Table 1: Data taken from kaggle about life satisfaction, mean salary, population size, number of jobs, area size and coordinates of the 32 boroughs.

When looking at the data, one directly notices that the City of London is very singular compared to the other boroughs. Even though it is the smallest borough of the city it features the highest mean salary as well as the second highest number of jobs.

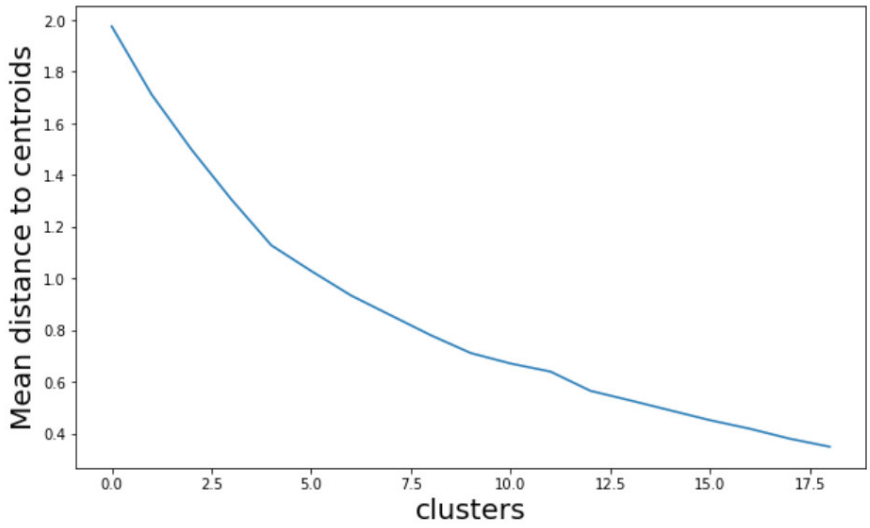
Next we take a look at the venues located in a radius of 1500m around the center of each borough using the Foursquare API. Here we set the limit of maximally returned results to

100 and assume that this section is representative for the whole borough. In Fig.2 the total number of venues in each part of the boroughs is shown. We find that only one-half of the boroughs reach the maximal venue density for the fixed number of returned venues.



2Figure 2: Total number of venues obtained from the Foursquare API fort he 32 boroughs (Limit=100).

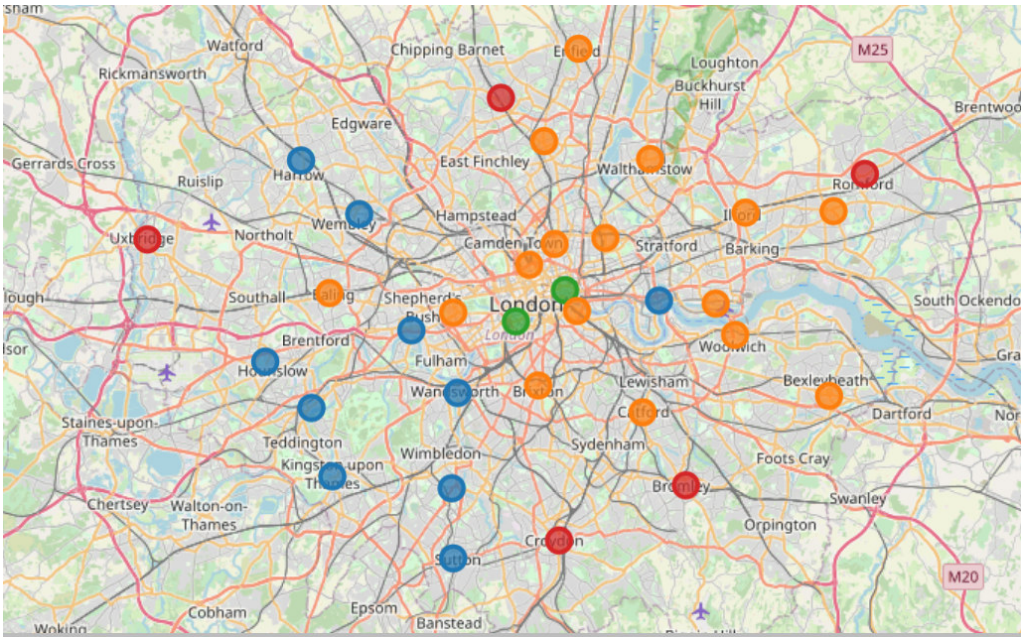
To cluster the boroughs into groups we apply the k-means algorithm where the best value for k can be determined by virtue of the elbow method. In Fig.3 we plot the mean distance of each data point to its centroid after convergence of the k-means algorithm was obtained. A good estimate for the number of clusters is then given by the point at which the curvature changes or the curve features a kink. In Fig.3 we can read off an approximate value of k=4 that we will use in the following.



3Figure 3: Elbow method

The final result of the k-means algorithm is displayed in Fig.4 where again a map of London is shown with clusters indicated by differently colored markers.

There is one central cluster consisting only of the City of London and Westminster (cluster 2, green markers), one in the south-west of the city (cluster 0, blue markers), one that is located around the central borough extending to the northeast part of London (cluster 1 orange markers) as well as another one on the outskirts of the city (cluster 3, red markers).



4 Figure 4: Ma pof London showing the four clusters 0 (blue), 1 (orange), 2 (green) and 3 (red).

While the central cluster (cluster 0) is the smallest one with the lowest population size, it features the highest life satisfaction and mean salary as well as the second-highest number of jobs. In contrast to that mean salary and life satisfaction is rather similar in the other clusters. Cluster 1 is by far the biggest one considering area size, population number as well as the number of jobs. According to our expectations cluster 3, which is located on the outskirts, has the smallest number of jobs and also a comparably small population size while the area size is large.

	Cluster Labels	mean_salary	life_satisfaction	population_size	number_of_jobs	area_size
0	0	3.891165e+06	696.337000	224790955.0	133993000.0	3570340.0
1	1	3.431277e+06	640.976875	390912225.0	233019000.0	5125298.0
2	2	7.691000e+06	762.171875	26403000.0	141500000.0	251800.0
3	3	3.154175e+06	660.634000	146107350.0	65211000.0	4656795.0

2 Table 2: The table lists the mean slary, life atisfaction, population size, number of jobs and area size for each cluster.

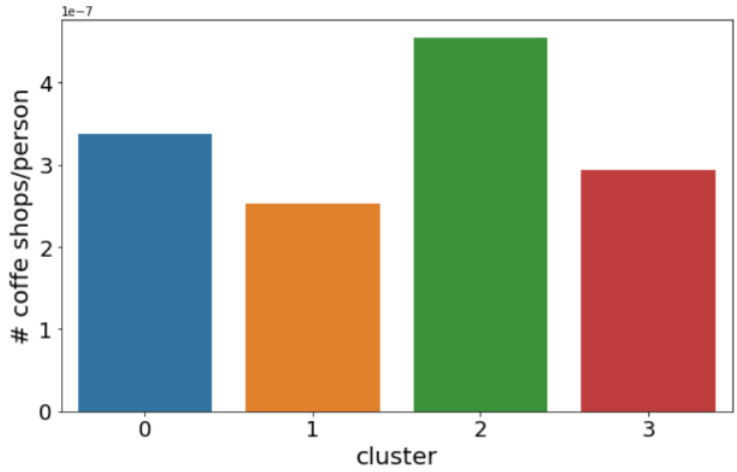
When considering the total number of venues in the four clusters we find that especially coffee shops and pubs, at least in clusters 0,1,3, are very common in most clusters. It stands out that the most common venues in cluster 2 are hotels indicating that the central boroughs are very much focused on tourism. Shops selling goods for daily use as grocery stores are more common in cluster 0,2 and 3 implying that people rather live in these clusters while they more likely to work in the center of the city. This observation is also backed up by the fact that parks are very common in clusters 0,1 and 3.

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	0	Pub	Coffee Shop	Indian Restaurant	Café	Park
1	1	Pub	Coffee Shop	Park	Café	Grocery Store
2	2	Hotel	Coffee Shop	Café	Gym / Fitness Center	Plaza
3	3	Coffee Shop	Pub	Clothing Store	Grocery Store	Park

3Table 3: The table shows the most common venues in each cluster.

Finally, we address the question of which cluster a possible investor should consider as a profitable location to start a business as open a shop, restaurant or another kind of venue. Here one important consideration is that the offered product finds the right market. For example, it would be more advisable to open an expensive jewelry shop or fashion store in a borough belonging to cluster 2, i.e. in the City of London or Westminster where the mean salary is comparable high and clients are more likely able to afford luxury items. Similarly, opening a hotel in cluster 2 seems to be more profitable than in one of the other clusters as most of the tourists want to visit the main attractions in the city. In contrast to that clusters 0 and 1, in which most of the people live, are more suitable for shops selling goods for daily use as grocery stores.

A second important point is market saturation. It is only advisable to offer a product in a certain area if the market in this area is not saturated yet. As an example, we can consider



5 Figure 5: The barplot shows the number of coffe shops per person in each cluster.

coffee shops that are common venues in all clusters. When looking at the number of coffee

shops per person living in each cluster, cf. Fig.5, we find that the market is least saturated in cluster 1 indicating that a borough belonging to this cluster would be a good location for opening a coffee shop.

Discussion and summary

In this work we analyzed the market structure of London using a k-means clustering algorithm that reduces the complexity of the problem.

We found that the 32 boroughs of the city can be grouped into 4 clusters when information about mean salary, population size, life satisfaction, number of jobs, area size and common venues are used. Here, it is important to note that there might be many more features that could have a significant influence on how boroughs are grouped. One example is the total number of tourists visiting a certain area. For sake of simplicity we reduced the analysis to the six features mentioned above.

Even using this reduce set of inputs we obtain a reliable classification that can act as a first guide for possible business decisions.