Clustering London, a data analysis approach

A. Introduction/Business problem

London, the capital of England and the United Kingdom, is a bustling, lively city considered to be one of the world's most important global cities with a diverse range of people and cultures and more than 300 languages spoken in the region. It exerts a considerable impact upon arts, commerce, education, entertainment, fashion and finance.

Administratively, London is divided in 32 boroughs. Together with City of London, these boroughs form 33 local authority districts, which make up Greater London. The region covers 1,572 km2 (607 sq. mi) and had a population of 8,173,920 according to 2019 data.

Due to its significance in the global scene and the variety of opportunities it offers, London still attracts people from all over the world seeking for new endeavors. An established Greek company, which runs its own microbrewery as well as several delicatessen stores, is exploring the idea of setting up a new business in London and needs help in terms of identifying areas with possible opportunities. Data analysis tools could help us providing an answer to a couple of strategic questions regarding this business idea:

- Can we identify boroughs with potential of setting up a microbrewery and/or a delicatessen store?
- How could household income and population affect our decision?

B. Data Description

The following data sources were used to get an insight into our business problem:

1. First dataset consists of London Boroughs with respective median household incomes. This dataset was used primarily to depict the variability of the median household income across London

boroughs, with the use of a choropleth map. We then used it in conjunction with the second dataset to build the final dataset for our analysis. It is compiled from two different sources: CACI & https://www.trustforlondon.org.uk/ and can be found on my GitHub page.

2. Second dataset consists of London boroughs with geographical co-ordinates and population data. As described above, it was used in conjunction with the first dataset to build the final dataset needed for our analysis. Source : https://github.com
3. GeoPy enabled us to get geographical coordinates of London Boroughs in a json file. This served as the primary data source for building a choropleth map. Source: https://githb.com
4. Foursquare API was used to explore London boroughs and retrieve their respective most common venues. Results from these queries were utilized as key feature for clustering London boroughs.

C. Methodology

My aim was to use clustering as a method for segmenting London boroughs in major areas that could give us insights to markets and identify similar features in terms of popular venues. I used Python for my analysis, which provides a large standard library with a variety of tools.

Obtain the Data

First step was to create a dataset consisting of London boroughs and their geographical coordinates, together with their respective population and household income data. For this purpose, I retrieved the datasets described in data section and I created two different dataframes.

Exploratory data analysis

I performed some data wrangling in order to give my dataframes a more appropriate format for further analysis. Removed unnecessary columns and missing values and filtered part of second dataset to retrieve only London related data.

Finally, I merged the two dataframes for the next steps of the analysis.

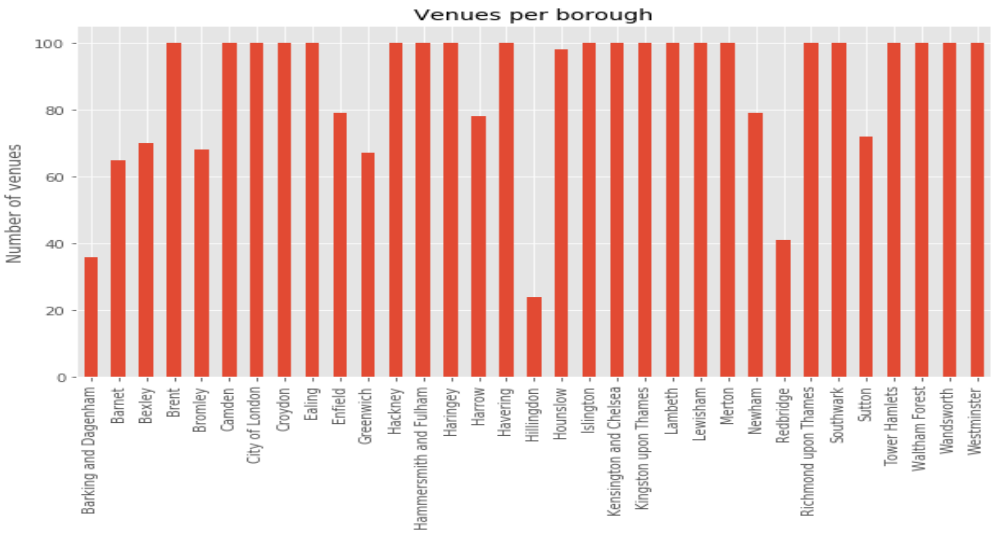
	Borough	Latitude	Longitude	Population	Median income
0	Barking and Dagenham	51.546501	0.124998	185974	21953.0
1	Barnet	51.605499	-0.207715	356064	34163.0
2	Bexley	51.459202	0.136321	231975	28691.0
3	Brent	51.551800	-0.257501	311279	27364.0
4	Bromley	51.391800	0.026393	309225	33659.0
5	Camden	51.534401	-0.143292	220387	36053.0
6	City of London	51.512799	-0.091840	7358	45436.0
7	Croydon	51.368198	-0.096495	363453	27847.0
8	Ealing	51.518002	-0.324967	338429	29918.0
9	Enfield	51.639900	-0.082701	312456	27853.0
10	Greenwich	51.476700	0.051810	254520	26672.0
11	Hackney	51.549599	-0.069847	246136	28043.0
12	Hammersmith and Fulham	51.492100	-0.216469	182563	36527.0
13	Haringey	51.589699	-0.105810	255363	29318.0

Data Analysis - Visualization

I used Foursquare API, to explore London boroughs and retrieve a list of recommended venues within a radius of 2000 meters from each Neighborhood center. There is a set limit in the number of returned venues of 100. A sample of the list of venues retrieved per London borough is provided below:

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barking and Dagenham	51.546501	0.124998	Capital Karts	51.531792	0.118739	Go Kart Track
1	Barking and Dagenham	51.546501	0.124998	Mayesbrook Park	51.549842	0.108544	Park
2	Barking and Dagenham	51.546501	0.124998	Goodmayes Park	51.558503	0.116386	Park
3	Barking and Dagenham	51.546501	0.124998	Co-op Food	51.540093	0.127522	Grocery Store
4	Barking and Dagenham	51.546501	0.124998	wilko	51.541002	0.148898	Furniture / Home Store

We can also see the number of venues retrieved per borough, in the form of a bar chart:



In summary, my query returned *2889* venues, with a number of *270* unique venue categories. I must mention that depending on the time you run the Foursquare API query, results may vary in terms of numbers and category of venues retrieved.

Next step of the analysis was to create a dataframe from the list of venues retrieved by Foursquare API. This dataframe presents all different types of venues found per borough and is the dataframe used for clustering. I applied onehot encoding and grouped rows by

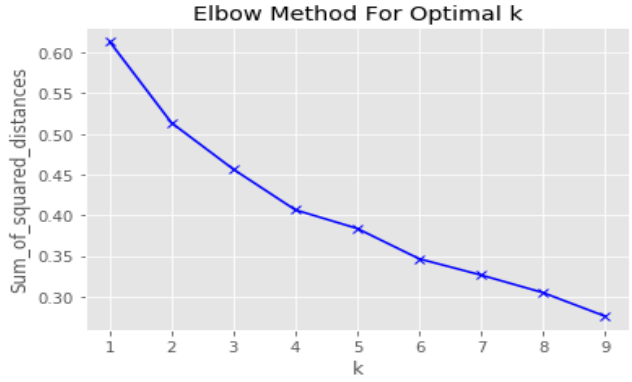
neighborhood and by the mean of the frequency of occurrence of each category. This resulted into the following dataframe (*sample*):

	Borough	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	Warehouse Store	Waterfront	Whisky Bar	Wine Bar	Wine Shop
0	Barking and Dagenham	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.000000	...	0.000000	0.00	0.00	0.00	0.00
1	Barnet	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.000000	...	0.000000	0.00	0.00	0.00	0.00
2	Bexley	0.000000	0.000000	0.029851	0.00	0.00	0.000000	0.00	0.00	0.000000	...	0.014925	0.00	0.00	0.00	0.00
3	Brent	0.000000	0.000000	0.020000	0.00	0.00	0.000000	0.00	0.00	0.010000	...	0.020000	0.00	0.00	0.00	0.00
4	Bromley	0.000000	0.000000	0.014925	0.00	0.00	0.000000	0.00	0.00	0.014925	...	0.000000	0.00	0.00	0.00	0.00
5	Camden	0.000000	0.000000	0.010000	0.00	0.00	0.000000	0.00	0.00	0.000000	...	0.010000	0.00	0.00	0.01	0.00
6	City of London	0.000000	0.000000	0.000000	0.00	0.01	0.010000	0.03	0.00	0.000000	...	0.000000	0.00	0.02	0.00	0.00
7	Croydon	0.000000	0.000000	0.010000	0.00	0.00	0.000000	0.00	0.01	0.010000	...	0.000000	0.00	0.00	0.00	0.00
8	Ealing	0.000000	0.000000	0.000000	0.00	0.00	0.010000	0.00	0.00	0.000000	...	0.000000	0.00	0.00	0.02	0.00
9	Enfield	0.000000	0.000000	0.011765	0.00	0.00	0.000000	0.00	0.00	0.000000	...	0.000000	0.00	0.00	0.00	0.00
10	Greenwich	0.000000	0.014706	0.000000	0.00	0.00	0.000000	0.00	0.00	0.014706	...	0.014706	0.00	0.00	0.00	0.00
11	Hackney	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.000000	...	0.000000	0.00	0.00	0.00	0.02

I also created a second dataframe that displays the 10 most common types of venues for each borough and merged it with the dataframe that contains London boroughs population and income data. A sample of the dataframe below:

Borough	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	Latitude	Longitude	Population	Median income
Barking and Dagenham	Grocery Store	Pub	Bus Stop	Park	Soccer Field	Supermarket	Café	Go Kart Track	Bowling Alley	Sporting Goods Shop	51.546501	0.124998	185974	21953.0
Barnet	Café	Supermarket	Turkish Restaurant	Restaurant	Coffee Shop	Indian Restaurant	Golf Course	Fast Food Restaurant	Park	Pub	51.605499	-0.207715	356064	34163.0
Bexley	Pub	Grocery Store	Supermarket	Coffee Shop	Clothing Store	Fast Food Restaurant	Hotel	Italian Restaurant	Indian Restaurant	Café	51.459202	0.136321	231975	28691.0
Brent	Coffee Shop	Grocery Store	Sandwich Place	Hotel	Clothing Store	Sporting Goods Shop	Indian Restaurant	Warehouse Store	Plaza	Food Court	51.551800	-0.257501	311279	27364.0
Bromley	Pub	Park	Clothing Store	Grocery Store	Pizza Place	Coffee Shop	English Restaurant	Gym / Fitness Center	Sandwich Place	Indian Restaurant	51.391800	0.026393	309225	33659.0

My decision was to use **K-Means** algorithm, one of the most common cluster methods of unsupervised learning, to cluster the boroughs. Elbow method helped me to find optimal **K** value.



Optimal **K** point is usually the elbow of the curve, which might not be so obvious in this diagram. I have chosen a **K** such that the SSE is relatively small but the rate of change of the SSE is relatively high.

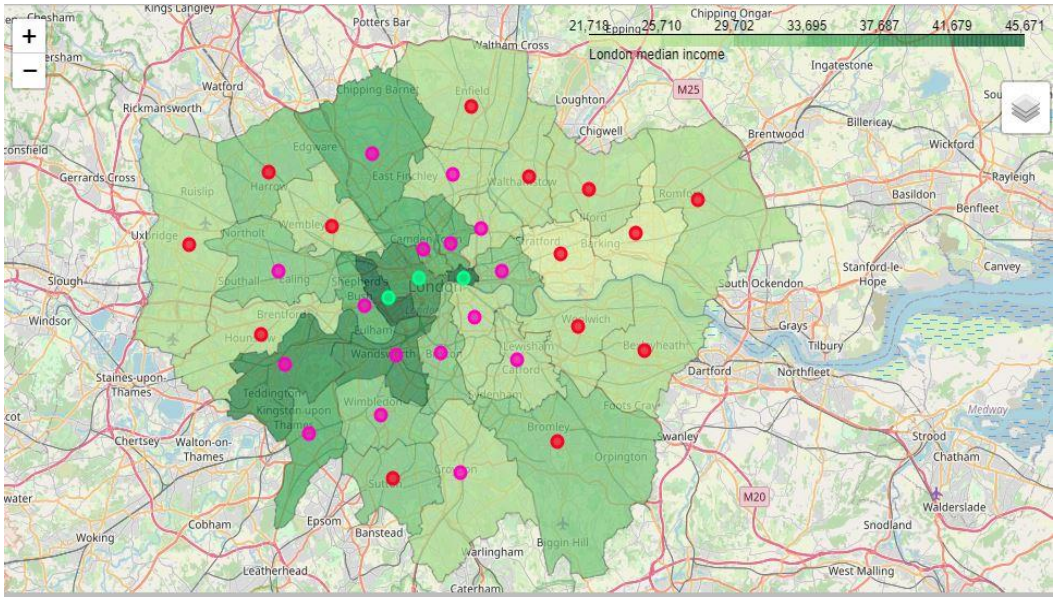
K=3 seems to be the optimal for this case. Increasing **K** to 4 or 5, resulted in added clusters of one or two boroughs, hence with no value added.

I applied **K-Means** clustering on the first dataframe and inserted the results as labels in the 10 most common types of venues dataframe.

Cluster Label	Borough	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	Latitude	Longitude	Population	Median income
1	Barking and Dagenham	Grocery Store	Pub	Bus Stop	Park	Soccer Field	Supermarket	Café	Go Kart Track	Bowling Alley	Sporting Goods Shop	51.546501	0.124998	185974	21953.0
0	Barnet	Café	Supermarket	Turkish Restaurant	Restaurant	Coffee Shop	Indian Restaurant	Golf Course	Fast Food Restaurant	Park	Pub	51.605499	-0.207715	356064	34163.0
1	Bexley	Pub	Grocery Store	Supermarket	Coffee Shop	Clothing Store	Fast Food Restaurant	Hotel	Italian Restaurant	Indian Restaurant	Café	51.459202	0.136321	231975	28691.0
1	Brent	Coffee Shop	Grocery Store	Sandwich Place	Hotel	Clothing Store	Sporting Goods Shop	Indian Restaurant	Warehouse Store	Plaza	Food Court	51.551800	-0.257501	311279	27364.0
1	Bromley	Pub	Park	Clothing Store	Grocery Store	Pizza Place	Coffee Shop	English Restaurant	Gym / Fitness Center	Sandwich Place	Indian Restaurant	51.391800	0.026393	309225	33659.0

Some useful info for London boroughs: the average household income is 31,806 GBP and the average population is 247,695.

I considered presenting, if any, the relationship between the clustered boroughs and the median household income. Hence, I created a London choropleth map with boroughs shaded in proportion to the measurement of the median income. I then used this map to superimpose the clustered boroughs as per the image below:



D. Results

By observing the dataframes of the clustered boroughs, we gain a better insight on common features such as popular types of venues, population and income data. In conjunction with the choropleth map, it provides a helpful view on topography of clustered boroughs and household income.

First cluster includes several London boroughs, with household income higher than the London average (31,806 GBP). Restaurants, cafes and pubs seem to be the most common type of venue found in this cluster.

Cluster Label	Borough	Median income	Population	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	Barnet	34163.0	356064	Café	Supermarket	Turkish Restaurant	Restaurant	Coffee Shop	Indian Restaurant	Golf Course	Fast Food Restaurant	Park	Pub
0	Camden	36053.0	220387	Coffee Shop	Café	Pub	Park	Zoo Exhibit	Vegetarian / Vegan Restaurant	Garden	Italian Restaurant	Market	Greek Restaurant
0	Croydon	27847.0	363453	Pub	Coffee Shop	Hotel	Clothing Store	Supermarket	Park	Bookstore	Indian Restaurant	Mediterranean Restaurant	Italian Restaurant
0	Ealing	29918.0	338429	Pub	Park	Coffee Shop	Café	Hotel	Italian Restaurant	Supermarket	Burger Joint	Indian Restaurant	Persian Restaurant
0	Hackney	28043.0	246136	Cocktail Bar	Pub	Café	Coffee Shop	Bakery	Turkish Restaurant	Pizza Place	Brewery	Park	Restaurant
0	Hammersmith and Fulham	36527.0	182563	Pub	Café	Gastropub	Indian Restaurant	Garden	Japanese Restaurant	Grocery Store	Coffee Shop	Park	Middle Eastern Restaurant
0	Haringey	29318.0	255363	Turkish Restaurant	Café	Coffee Shop	Pub	Park	Mediterranean Restaurant	Bakery	Indian Restaurant	Farmers Market	Lounge
0	Islington	33859.0	206078	Pub	Café	Park	Coffee Shop	Art Gallery	Bakery	Theater	Gastropub	Burger Joint	Cocktail Bar
0	Kingston upon Thames	35805.0	160008	Pub	Coffee Shop	Café	Italian Restaurant	Burger Joint	Thai Restaurant	Indian Restaurant	Supermarket	Park	Gym / Fitness Center

Second cluster contains mainly boroughs, with household income closer to or less than the average. Cafes and pubs are common in this cluster too, but and it appears there is a bigger variety in types of venues found in this cluster.

Cluster Label	Borough	Median income	Population	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
1	Barking and Dagenham	21953.0	185974	Grocery Store	Pub	Bus Stop	Park	Soccer Field	Supermarket	Café	Go Kart Track	Bowling Alley	Sporting Goods Shop
1	Bexley	28691.0	231975	Pub	Grocery Store	Supermarket	Coffee Shop	Clothing Store	Fast Food Restaurant	Hotel	Italian Restaurant	Indian Restaurant	Café
1	Brent	27364.0	311279	Coffee Shop	Grocery Store	Sandwich Place	Hotel	Clothing Store	Sporting Goods Shop	Indian Restaurant	Warehouse Store	Plaza	Food Court
1	Bromley	33659.0	309225	Pub	Park	Clothing Store	Grocery Store	Pizza Place	Coffee Shop	English Restaurant	Gym / Fitness Center	Sandwich Place	Indian Restaurant
1	Enfield	27853.0	312456	Pub	Coffee Shop	Supermarket	Pizza Place	Train Station	Indian Restaurant	Grocery Store	Clothing Store	Bookstore	Bar
1	Greenwich	26672.0	254520	Pub	Grocery Store	Park	Coffee Shop	Fast Food Restaurant	Clothing Store	Discount Store	Thai Restaurant	Supermarket	Furniture / Home Store
1	Harrow	31430.0	238913	Coffee Shop	Indian Restaurant	Grocery Store	Sandwich Place	Pub	Park	Gym / Fitness Center	Fast Food Restaurant	Fish & Chips Shop	Bar
1	Havering	29549.0	237245	Coffee Shop	Pub	Supermarket	Fast Food Restaurant	Café	Italian Restaurant	Clothing Store	Shopping Mall	Grocery Store	Park
1	Hillingdon	29222.0	273933	Grocery Store	Pub	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Pharmacy	Pet Store	Rugby Pitch	Park	Sculpture Garden

Third cluster consists of central London boroughs, less populous and with the highest median household income. Hotels, museums and galleries rank highly among most common types of venues, as these boroughs appeal mainly to visitors.

Cluster Label	Borough	Median income	Population	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
2	City of London	45436.0	7358	Hotel	Coffee Shop	Scenic Lookout	Gym / Fitness Center	Grocery Store	Pub	Art Museum	Theater	Falafel Restaurant	Cocktail Bar
2	Kensington and Chelsea	43315.0	158649	Café	Hotel	Garden	French Restaurant	Pub	Italian Restaurant	Science Museum	Gym / Fitness Center	Art Gallery	Ice Cream Shop
2	Westminster	40919.0	219359	Hotel	Cocktail Bar	Clothing Store	Hotel Bar	Art Gallery	French Restaurant	Park	Seafood Restaurant	Coffee Shop	Bookstore

So, how should we address the strategic questions discussed in Introduction section?

Company stakeholders are in favor of setting up a microbrewery in close proximity to pubs, cafes and restaurants, since these venues are a potential target customer. Both microbrewery and delicatessen customers tend to be more affluent and upscale, thus populous areas that maintain a higher household income would be preferred.

Taking into account the above, we could suggest the following:

Boroughs of the third cluster are not so populous and tend to be very expensive in real estate prices. In addition, since the cluster is mostly appealing to visitors and types of venues seem different from the other clusters, I choose to exclude this cluster from our recommendations.

First and second cluster appear to match better the criteria set by the stakeholders. Indicative examples, Lambeth and Wandsworth from the first cluster, populous boroughs with higher income and similar features in terms of common types of venues.

Cluster Label	Borough	Median income	Population	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	Lambeth	30753.0	303183	Pub	Coffee Shop	Park	Grocery Store	Pizza Place	Gym / Fitness Center	Market	Brewery	Bakery	Restaurant
0	Lewisham	26541.0	275988	Pub	Coffee Shop	Grocery Store	Café	Gastropub	Supermarket	Park	Food Truck	Turkish Restaurant	Bar
0	Merton	33003.0	199777	Park	Coffee Shop	Sushi Restaurant	Gym / Fitness Center	Bar	Grocery Store	Pub	Burger Joint	Italian Restaurant	Supermarket
0	Richmond upon Thames	41493.0	186984	Pub	Italian Restaurant	Coffee Shop	Café	Park	Rugby Stadium	Garden	Bakery	Deli / Bodega	Thai Restaurant
0	Southwark	29192.0	288265	Brewery	Café	Italian Restaurant	Coffee Shop	Bar	Pub	Art Gallery	Beer Bar	Pizza Place	Tapas Restaurant
0	Tower Hamlets	30760.0	254073	Coffee Shop	Pub	Café	Hotel	Italian Restaurant	Pizza Place	Cocktail Bar	Beer Bar	Bar	Park
0	Wandsworth	37911.0	306949	Pub	Park	Coffee Shop	Café	French Restaurant	Supermarket	Cocktail Bar	Pizza Place	Breakfast Spot	Bakery

Bromley from the second cluster seems to fit the model as well. Its proximity to the boroughs of Lambeth and Wandsworth, all three situated in South London, could create an area of interest and further investigation for potential opportunities.

Cluster Label	Borough	Median income	Population	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
1	Barking and Dagenham	21953.0	185974	Grocery Store	Pub	Bus Stop	Park	Soccer Field	Supermarket	Café	Go Kart Track	Bowling Alley	Sporting Goods Shop
1	Bexley	28691.0	231975	Pub	Grocery Store	Supermarket	Coffee Shop	Clothing Store	Fast Food Restaurant	Hotel	Italian Restaurant	Indian Restaurant	Café
1	Brent	27364.0	311279	Coffee Shop	Grocery Store	Sandwich Place	Hotel	Clothing Store	Sporting Goods Shop	Indian Restaurant	Warehouse Store	Plaza	Food Court
1	Bromley	33659.0	309225	Pub	Park	Clothing Store	Grocery Store	Pizza Place	Coffee Shop	English Restaurant	Gym / Fitness Center	Sandwich Place	Indian Restaurant
1	Enfield	27853.0	312456	Pub	Coffee Shop	Supermarket	Pizza Place	Train Station	Indian Restaurant	Grocery Store	Clothing Store	Bookstore	Bar
1	Greenwich	26672.0	254520	Pub	Grocery Store	Park	Coffee Shop	Fast Food Restaurant	Clothing Store	Discount Store	Thai Restaurant	Supermarket	Furniture / Home Store
1	Harrow	31430.0	238913	Coffee Shop	Indian Restaurant	Grocery Store	Sandwich Place	Pub	Park	Gym / Fitness Center	Fast Food Restaurant	Fish & Chips Shop	Bar
1	Havering	29549.0	237245	Coffee Shop	Pub	Supermarket	Fast Food Restaurant	Café	Italian Restaurant	Clothing Store	Shopping Mall	Grocery Store	Park
1	Hillingdon	29222.0	273933	Grocery Store	Pub	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Pharmacy	Pet Store	Rugby Pitch	Park	Sculpture Garden

E. Discussion – Conclusion

The purpose of this project was to identify common characteristics in London boroughs in terms of popular venues and help stakeholders in narrowing down the search for optimal location for setting up a microbrewery and/or a delicatessen store. Clustering of London boroughs resulted in creating major areas of interest, which could serve as a starting point for further exploration.

The final decision on the chosen location will also depend on a number of other factors, which have not been part of this analysis. Those would include characteristics such as real estate prices, proximity to relevant markets and other places of interest, potential competitors and social trends of every borough.

We can consider this case study as a good example of the value that data analysis may bring in an organization.

Applying tools and techniques from data science has proven an efficient approach that can be used to gain insights from scattered data. Leveraging data analysis in decision-making process can help identify new business opportunities and narrow down the range of indecisiveness.