IBM Data Science Capstone Project

Title: Analysing London Boroughs

Author: Jake Smith

1. Introduction

1.1. Background

London is a large, diverse city in the United Kingdom. Its scope in terms of geography can be defined in different ways, though typically it is divided into 32 London boroughs along with the City of London district. This area is often referred to as Greater London and covers an area of 1572 square kilometres.

London has a large resident population of approximately 9 million inhabitants. In addition to this, a vast number of tourists visit the city each year; over 20 million international visitors were received in 2018. Tourism is a key industry for the economy of the city, and this is emphasised by the fact that tourists who visited London in 2011 were estimated to spend £9.4 billion [1].

1.2. Description of business problem

The resident population of London, in combination with tourists who visit each year, offer lucrative opportunities for the hospitality industry. The business problem I will address in this project is related to this; can specific boroughs within the city be identified as suitable locations to set up a new craft beer bar?

For someone wishing to set up a new hospitality business venture such as this, a key concern would be the frequency that various types of business can be found across different neighbourhoods. When setting up a new bar, an individual would most likely want to know that the location that they identify is in an area in which hospitality businesses are popular and therefore can thrive. Another important factor is knowing where the most popular tourist areas are. This can be assessed in terms of the total tourist traffic they experience per year; a reasonable assumption to make is that areas more popular with tourists are likely to generate more income for hospitality businesses. Investigating both factors can help to identify areas in London that have the right target market.

To address this business problem, I will create a data frame that combines borough names, latitude and longitude coordinates for each borough scraped from Wikipedia, the 10 most commonly found venues in each borough taken from the Foursquare API, tourist trip data sourced from data websites, and finally a geo json file that marks out the separate London boroughs. The boroughs are to be clustered using the unsupervised machine learning technique of K-Means clustering according to the most commonly found venues. This data frame will be used to create an interactive map with Folium that visualises both venue data and tourist trip data for each borough. It will take the form of a choropleth map, which is an ideal way to

visualise the different levels of tourist trips across the boroughs, with coloured labels indicating how the boroughs are clustered based on the frequency with which different venues are found. The labels will also provide descriptive information including cluster descriptions and the popularity of each borough with tourists based on categories.

1.3. Interest in the problem

This business problem is likely to be of interest to anyone who intends to enter the hospitality industry in London, specifically bars, pubs, and other types of drinking establishment. Similarly, those who already have a hospitality business but are considering expanding into more locations would find this information useful.

This information would also be interesting for tourists who want to gain a good idea of the different areas in London before they visit, including popularity of each borough alongside the typical venues they are likely to encounter in different places.

2. Data

2.1. Data description and sources

- The names, latitude and longitude co-ordinates of each London Borough and the City of London district (which is not officially classed as a borough even though it is in central London) will be sourced by web scraping from Wikipedia using BeautifulSoup [2].

- The Foursquare API will be used to obtain the most common types of venue found within the different London boroughs [3].

- Tourism trips to each London borough (thousands per year) for the year 2007 is to be sourced from the UK government data website [4].

- Geo json data that marks out the boundaries for each London borough will be used to create a choropleth map; this is to be sourced from London Data Store website [5].

3. Methodology

3.1. Web scraping using Beautiful Soup

The first step was to web scrape data from Wikipedia about the London boroughs [2] using the Beautiful Soup Python library. The required information was stored in two tables on the Wikipedia page including the required information of borough name and geographical co-ordinates. Several other columns of data were also scraped. The first table contained information on 32 of the London boroughs. A second table contained the same information but for the 'City of London' district only.

The information was scraped from the two tables and stored into two separate Pandas data frames. Unrequired columns were dropped, with the City of London data frame then appended

to the first data frame named 'london_boroughs'. The data frame now contained the names and co-ordinates of the 32 boroughs and City of London district.

Data cleaning was required to remove extraneous text and characters from the name column, such as '[note 1]'. Data wrangling of the 'Co-ordinates' column was also required as it contained the co-ordinates of each borough in two different formats (minute-second and decimal). Splitting of this column was required first into these two formats, before then splitting the decimal column into latitude and longitude. Cleaning was required to remove extraneous text such as 'N', 'E' and 'W', as well as adding a minus symbol at the beginning of any longitude values preceded by a 'W'.

Finally, the data type needed to be changed to float for the latitude and longitude columns. After scraping, cleaning, and wrangling, the 'london_boroughs' data frame was in the following format:

Figure 1 - Pandas data frame with borough names and co-ordinates

	Borough	Latitude	Longitude
0	Barking and Dagenham	51.5607	0.1557
1	Barnet	51.6252	-0.1517
2	Bexley	51.4549	0.1505
3	Brent	51.5588	-0.2817
4	Bromley	51.4039	0.0198

3.2. Joining Tourist Trips data

Tourist trip data was sourced from a UK government data site [4] in the form of a csv file containing trips to each borough in the form of 'Overseas Trips', 'Domestic Stay Trips' and 'Day Trips', all in the format of thousands. This was stored in my github and then read into a Pandas data frame. Wrangling was required to remove commas, change the type to integer, and create a 'TotalTrips' column to be used for the actual analysis by summing the three types of trips. The three individual trip types were then dropped.

A new data frame named 'london' was created by joining the trips data frame with the data frame containing co-ordinates, which looked like:

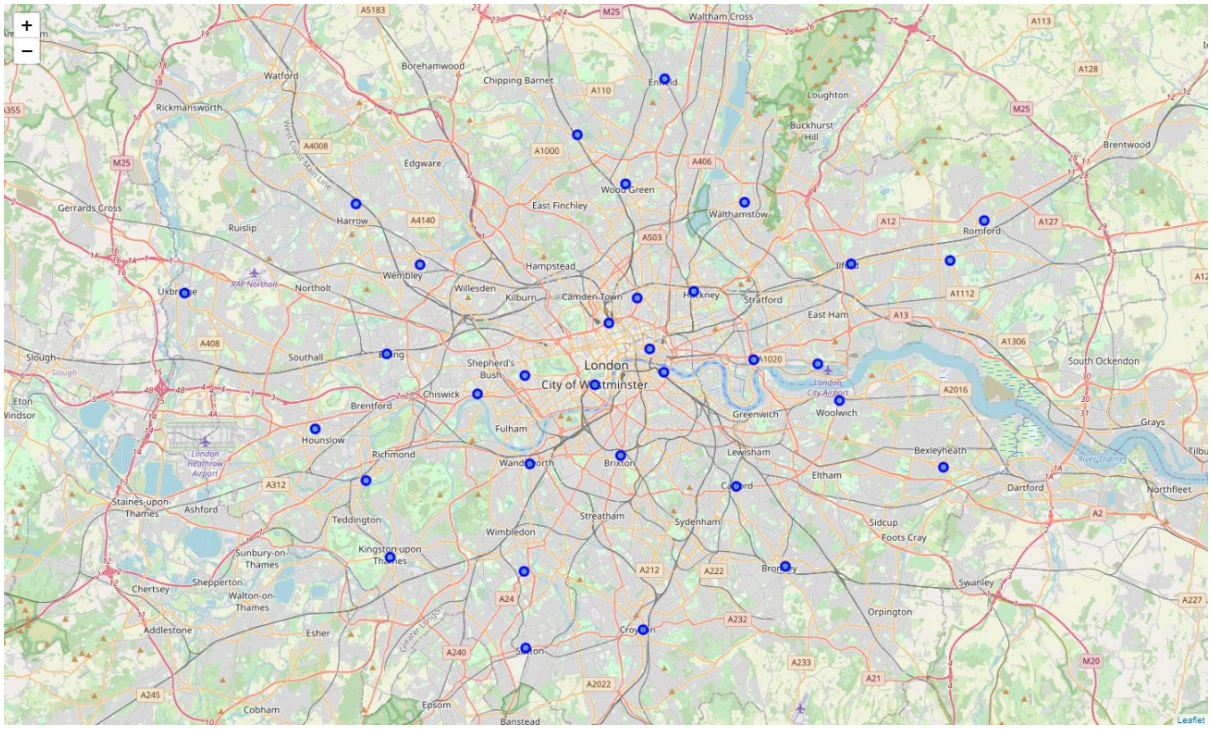
Figure 2 - data frame with TotalTrips data added

	Borough	Latitude	Longitude	TotalTrips
0	Barking and Dagenham	51.5607	0.1557	1140
1	Barnet	51.6252	-0.1517	7169
2	Bexley	51.4549	0.1505	2490
3	Brent	51.5588	-0.2817	3007
4	Bromley	51.4039	0.0198	4364

3.3. Visualising the boroughs on a Folium map of London

To sense check that the information in the data frame was correct a map was created with markers overlaid identifying each of the London boroughs:

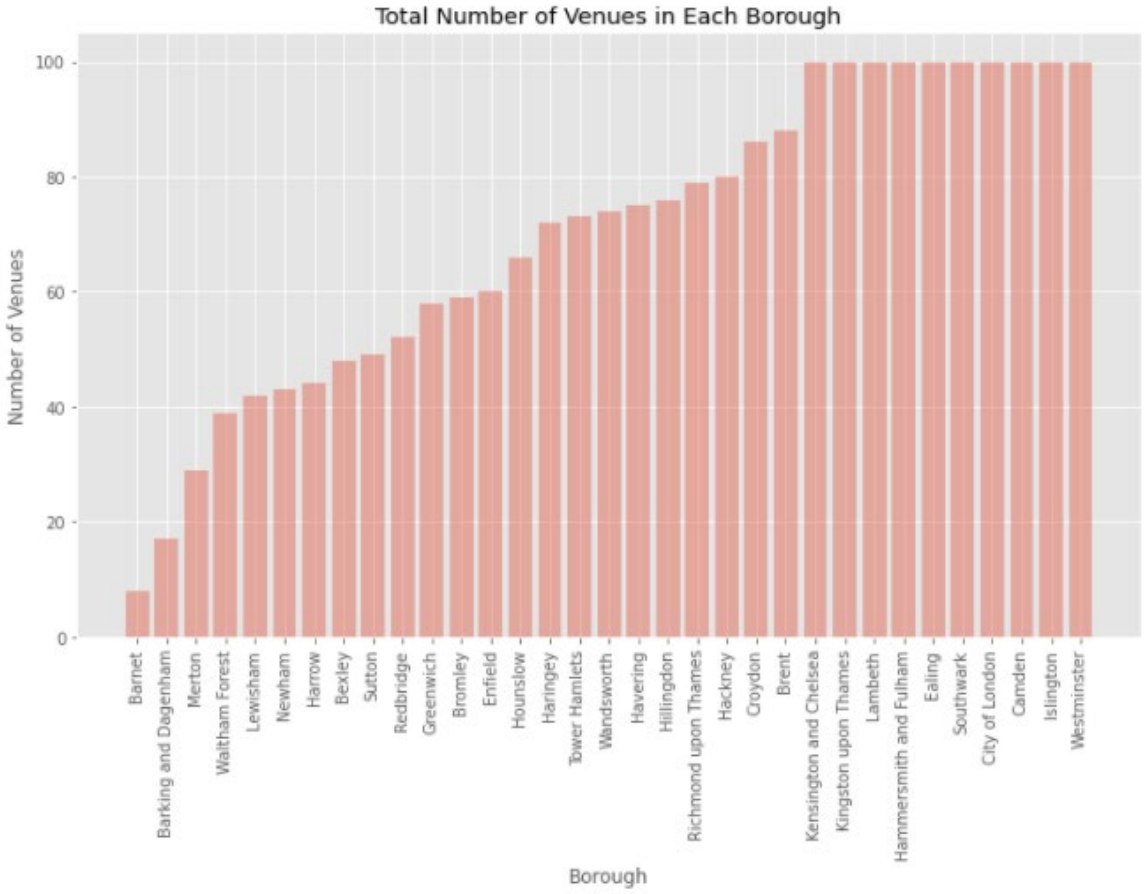
Figure 3 - Folium map with markers identifying London boroughs



3.4. Sourcing venue information for each borough using the Foursquare API

The next step was to collect more information about each borough in the form of most commonly venues found at each, with the API Foursquare was used for this purpose. A function was created that accessed information on which venues could be found within 1000 metres of a location, with a limit of 100 results for each. This function was run using the latitude and longitude co-ordinates from the 'london' data frame, with the results stored in a new data frame. To better understand the venue information, a count of how many venues were returned for each borough was run and visualised in a bar chart:

Figure 4 - number of venues returned per borough



This indicated that some boroughs reached the upper limit of 100 venues returned, with the remaining tending to return between 40 to 80 venues.

3.5. Determining the most common venues found at each borough

One hot encoding was used to determine the frequency with which different types of venues were found at each location. After applying a function to the encoded data, a new data frame with the 10 most common venues found at each borough was returned:

Figure 5 - top 10 most common venues found at each borough

	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
0	Barking and Dagenham	Bus Stop	Construction & Landscaping	Grocery Store	Diner	Discount Store	Chinese Restaurant	Park	Soccer Field	Pool	Supermarket
1	Barnet	Pub	Park	Rental Car Location	Gym	Bus Stop	Fish & Chips Shop	Financial or Legal Service	Film Studio	Filipino Restaurant	Fast Food Restaurant
2	Bexley	Pub	Clothing Store	Coffee Shop	Hotel	Fast Food Restaurant	Supermarket	Italian Restaurant	Pharmacy	American Restaurant	Bowling Alley
3	Brent	Coffee Shop	Hotel	Clothing Store	Bar	Indian Restaurant	Pizza Place	Sporting Goods Shop	Burger Joint	Grocery Store	Italian Restaurant
4	Bromley	Pub	Clothing Store	Coffee Shop	Indian Restaurant	Supermarket	Burger Joint	Electronics Store	Bar	Pizza Place	Gym / Fitness Center

3.6. K-Means clustering analysis

To solve our problem, the boroughs were required to be clustered into different groups based upon similarity in terms of most common venues found at each. The K-Means method of clustering was selected as being suitable for this purpose, it being an unsupervised method of

machine learning. A number of values for K were explored before selecting the optimum value, which in this case was three.

3.7. Visualising and exploring the clusters

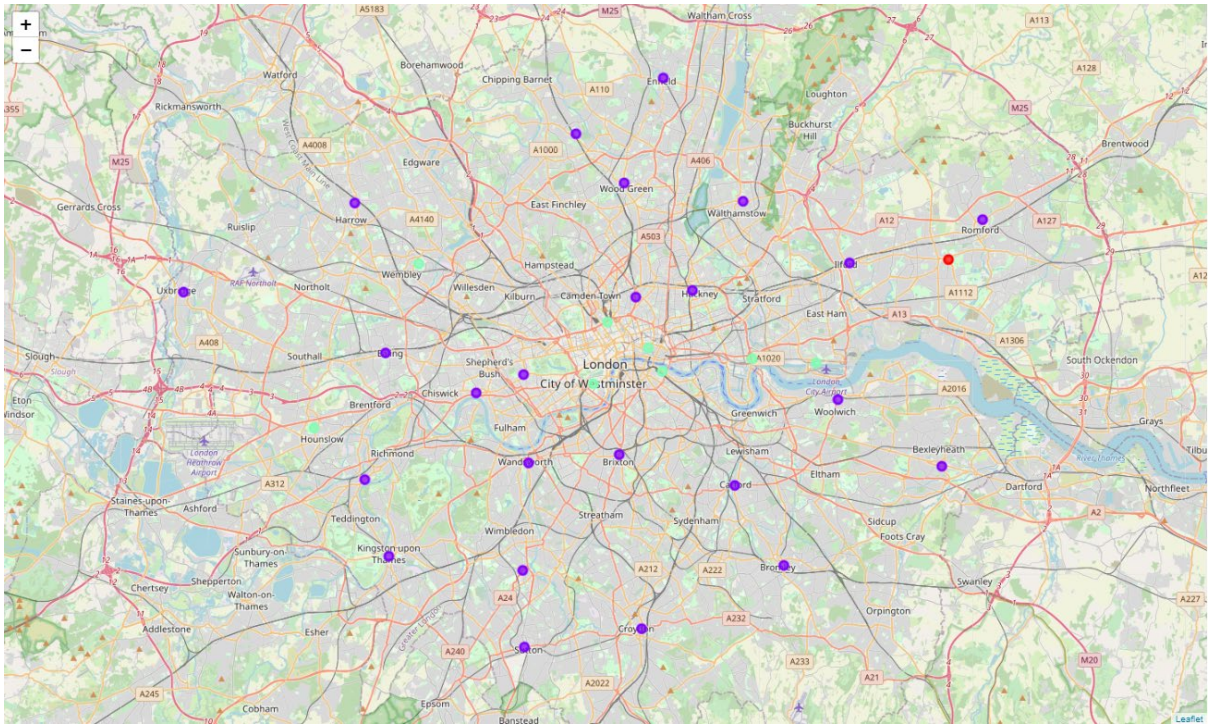
Cluster value labels for each borough, along with most common venue information, were combined with the original borough, co-ordinates and tourist trip data to create a new data frame named 'london_boroughs_clusters':

Figure 6 - data frame with cluster value labels and most common venues added

	Borough	Latitude	Longitude	TotalTrips	Cluster Value	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	Barking and Dagenham	51.5607	0.1557	1140	4	Bus Stop	Construction & Landscaping	Grocery Store	Diner	Discount Store	Chinese Restaurant	Park	Soccer Field	Pool	Supermarket
1	Barnet	51.6252	-0.1517	7169	1	Pub	Park	Rental Car Location	Gym	Bus Stop	Fish & Chips Shop	Financial or Legal Service	Film Studio	Filipino Restaurant	Fast Food Restaurant
2	Bexley	51.4549	0.1505	2490	2	Pub	Clothing Store	Coffee Shop	Hotel	Fast Food Restaurant	Supermarket	Italian Restaurant	Pharmacy	American Restaurant	Bowling Alley
3	Brent	51.5588	-0.2817	3007	3	Coffee Shop	Hotel	Clothing Store	Bar	Indian Restaurant	Pizza Place	Sporting Goods Shop	Burger Joint	Grocery Store	Italian Restaurant
4	Bromley	51.4039	0.0198	4364	2	Pub	Clothing Store	Coffee Shop	Indian Restaurant	Supermarket	Burger Joint	Electronics Store	Bar	Pizza Place	Gym / Fitness Center

A new map was created to visualise how the boroughs had clustered, to see if any initial understanding could be gained based on geography:

Figure 7 - map with coloured markers identifying three clusters



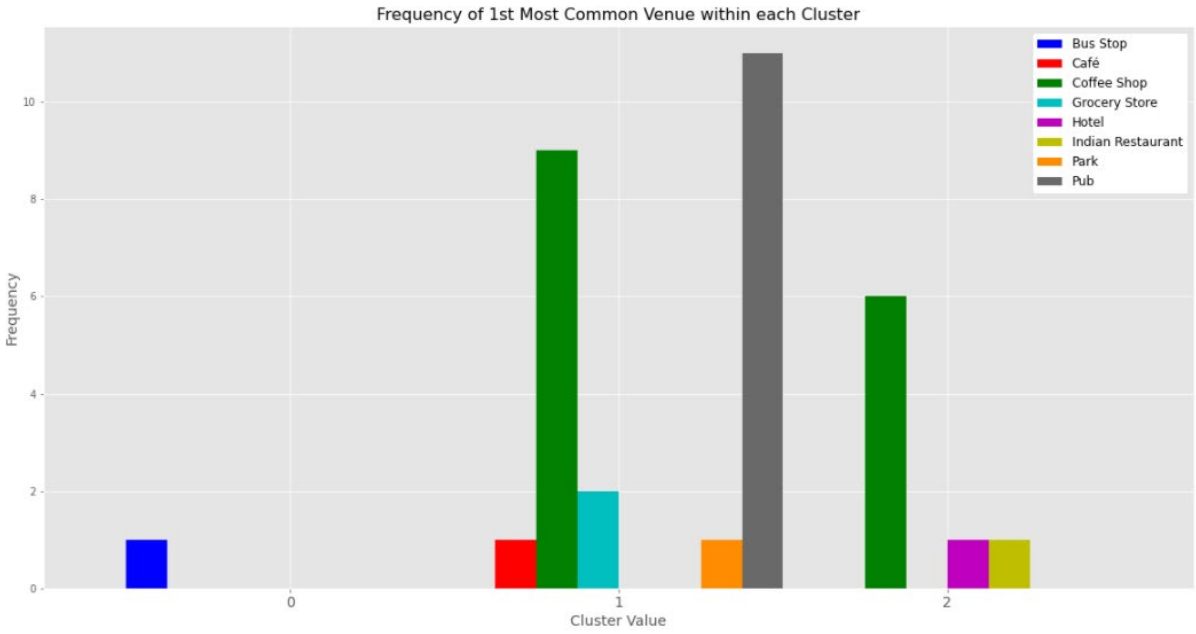
The map indicated that the way the boroughs have been clustered based on most common venues found at each seems to be related to geography of London. Typically:

- Cluster 0 (red marker): Single borough further east
- Cluster 1 (purple markers): Outer located boroughs

- Cluster 2 (light green markers): Centrally located boroughs

Investigating the 1st most common venue found at each borough can help us understand with more clarity how the clusters differ. A bar chart was created based on the frequency with which different 1st most common venues were found for boroughs in each cluster:

Figure 8 - frequency of 1st most common venue found for each cluster



Based on this, it was possible to see how the boroughs have clustered based on just the 1st most common venues found in each, alongside the geographical location, and how that fits together:

- Cluster 0 (red marker) indicated a more atypical location that is further out with it's most common venue being a bus stop
- Cluster 1 (purple markers) indicated an outer borough, with a more diverse range of venues found as first most common, but the majority being pubs followed by coffee shops
- Cluster 2 (light green markers) indicated a centrally located borough with mostly coffee shops

As a final representation of the different clusters and to see if any more details could be distinguished about the clusters, the data frame was filtered to group each separately in turn.

Cluster 0 (red marker):

Figure 9 - single borough within Cluster 0

	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
0	Barking and Dagenham	Bus Stop	Turkish Restaurant	Supermarket	Discount Store	Golf Course	Soccer Field	Grocery Store	Gas Station	Gym / Fitness Center	Park

Cluster 1 (purple markers):

Figure 10 - cluster 1 boroughs

	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
1	Barnet	Pub	Park	Bus Stop	Fish & Chips Shop	Gym	Café	Electronics Store	English Restaurant	Escape Room	Exhibit
2	Bexley	Pub	Clothing Store	Fast Food Restaurant	Hotel	Supermarket	Coffee Shop	Pharmacy	American Restaurant	Nightclub	Pizza Place
4	Bromley	Pub	Clothing Store	Coffee Shop	Indian Restaurant	Supermarket	Café	Bar	Pizza Place	Electronics Store	Burger Joint
6	Croydon	Coffee Shop	Pub	Hotel	Platform	Clothing Store	Bookstore	Indian Restaurant	Italian Restaurant	Mediterranean Restaurant	Portuguese Restaurant
7	Ealing	Coffee Shop	Pub	Café	Italian Restaurant	Indian Restaurant	Thai Restaurant	Pizza Place	Park	Bakery	Hotel
8	Enfield	Pub	Coffee Shop	Clothing Store	Grocery Store	Italian Restaurant	Optical Shop	Supermarket	Fish & Chips Shop	Pharmacy	Department Store
9	Greenwich	Grocery Store	Pub	Clothing Store	Coffee Shop	Supermarket	Plaza	Bakery	Pharmacy	Gym / Fitness Center	Fast Food Restaurant
10	Hackney	Pub	Coffee Shop	Brewery	Café	Bakery	Park	Restaurant	Italian Restaurant	Modern European Restaurant	Butcher
11	Hammersmith and Fulham	Pub	Café	Indian Restaurant	Coffee Shop	Park	French Restaurant	Gym / Fitness Center	Japanese Restaurant	Italian Restaurant	Gastropub
12	Haringey	Pub	Café	Turkish Restaurant	Clothing Store	Fast Food Restaurant	Bakery	Grocery Store	Park	Chinese Restaurant	Coffee Shop
13	Harrow	Coffee Shop	Fast Food Restaurant	Indian Restaurant	Sandwich Place	Pharmacy	Park	Clothing Store	Department Store	Gym / Fitness Center	Pub
14	Havering	Coffee Shop	Pub	Clothing Store	Shopping Mall	Park	Fast Food Restaurant	Furniture / Home Store	Supermarket	Café	Grocery Store
15	Hillingdon	Coffee Shop	Pub	Clothing Store	Fast Food Restaurant	Pharmacy	Gym	Italian Restaurant	Grocery Store	Department Store	Park
17	Islington	Pub	Coffee Shop	Café	Park	French Restaurant	Cocktail Bar	Bakery	Mediterranean Restaurant	Gastropub	Pizza Place
18	Kensington and Chelsea	Café	Restaurant	Pub	Italian Restaurant	Juice Bar	Clothing Store	Garden	Coffee Shop	Bakery	Hotel
19	Kingston upon Thames	Coffee Shop	Café	Pub	Clothing Store	Thai Restaurant	Burger Joint	Italian Restaurant	Department Store	Sandwich Place	Bakery
20	Lambeth	Coffee Shop	Pub	Caribbean Restaurant	Pizza Place	Cocktail Bar	Beer Bar	Tapas Restaurant	Market	Yoga Studio	Music Venue
21	Lewisham	Coffee Shop	Park	Supermarket	Grocery Store	Pub	Furniture / Home Store	Italian Restaurant	Platform	Theater	Café
22	Merton	Park	Café	Supermarket	Train Station	Fast Food Restaurant	Pub	Pizza Place	Coffee Shop	Pet Store	Diner
24	Redbridge	Grocery Store	Turkish Restaurant	Clothing Store	Coffee Shop	Supermarket	Pub	Bakery	Sandwich Place	Department Store	Fast Food Restaurant
25	Richmond upon Thames	Pub	Coffee Shop	Italian Restaurant	Grocery Store	Park	Café	Garden	Boat or Ferry	Historic Site	Deli / Bodega
27	Sutton	Coffee Shop	Clothing Store	Café	Pub	Supermarket	Pizza Place	Department Store	Hotel	Italian Restaurant	Bar
29	Waltham Forest	Pub	Café	Gym / Fitness Center	Art Gallery	Coffee Shop	Pizza Place	Brewery	Restaurant	Multiplex	Museum
30	Wandsworth	Pub	Gym / Fitness Center	Grocery Store	Coffee Shop	Café	Park	Clothing Store	Breakfast Spot	Pizza Place	Asian Restaurant

Cluster 2 (light green markers):

Figure 11 - cluster 2 boroughs

	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
3	Brent	Coffee Shop	Hotel	Bar	Clothing Store	Grocery Store	Pub	Pizza Place	Sporting Goods Shop	Indian Restaurant	American Restaurant
5	Camden	Coffee Shop	Café	Hotel	Bookstore	Park	Plaza	Breakfast Spot	Train Station	Science Museum	Burger Joint
16	Hounslow	Indian Restaurant	Coffee Shop	Grocery Store	Hotel	Fast Food Restaurant	Clothing Store	Supermarket	Bus Stop	Bakery	Discount Store
23	Newham	Hotel	Airport Service	Coffee Shop	Sandwich Place	Airport Terminal	Park	Light Rail Station	Bus Stop	Gym / Fitness Center	Chinese Restaurant
26	Southwark	Coffee Shop	Pub	Tapas Restaurant	Cocktail Bar	Hotel	Garden	Bakery	Seafood Restaurant	Scenic Lookout	Brewery
28	Tower Hamlets	Coffee Shop	Park	Hotel	Italian Restaurant	Light Rail Station	Bus Stop	Sandwich Place	English Restaurant	Lounge	Restaurant
31	Westminster	Coffee Shop	Hotel	Theater	Park	Café	Garden	Sushi Restaurant	Sandwich Place	Pub	Sporting Goods Shop
32	City of London	Coffee Shop	Gym / Fitness Center	Hotel	Cocktail Bar	Modern European Restaurant	Steakhouse	Falafel Restaurant	French Restaurant	Theater	Event Space

The above provided the confidence needed to create descriptive labels for each cluster that could be added as a new column ‘Cluster Label’ to the main data frame. Based on the map of cluster locations, along with the above bar chart of Most Common 1st Venue found in each cluster, and finally eyeballing the boroughs found in each cluster when filtered, appropriate labels were assigned to each cluster as follows:

Cluster 0: Atypical, non-hospitality
Cluster 1: Pubs and coffee shops
Cluster 2: Coffee shops & hotels

These labels were added to the data frame in a new column called ‘Cluster Label’:

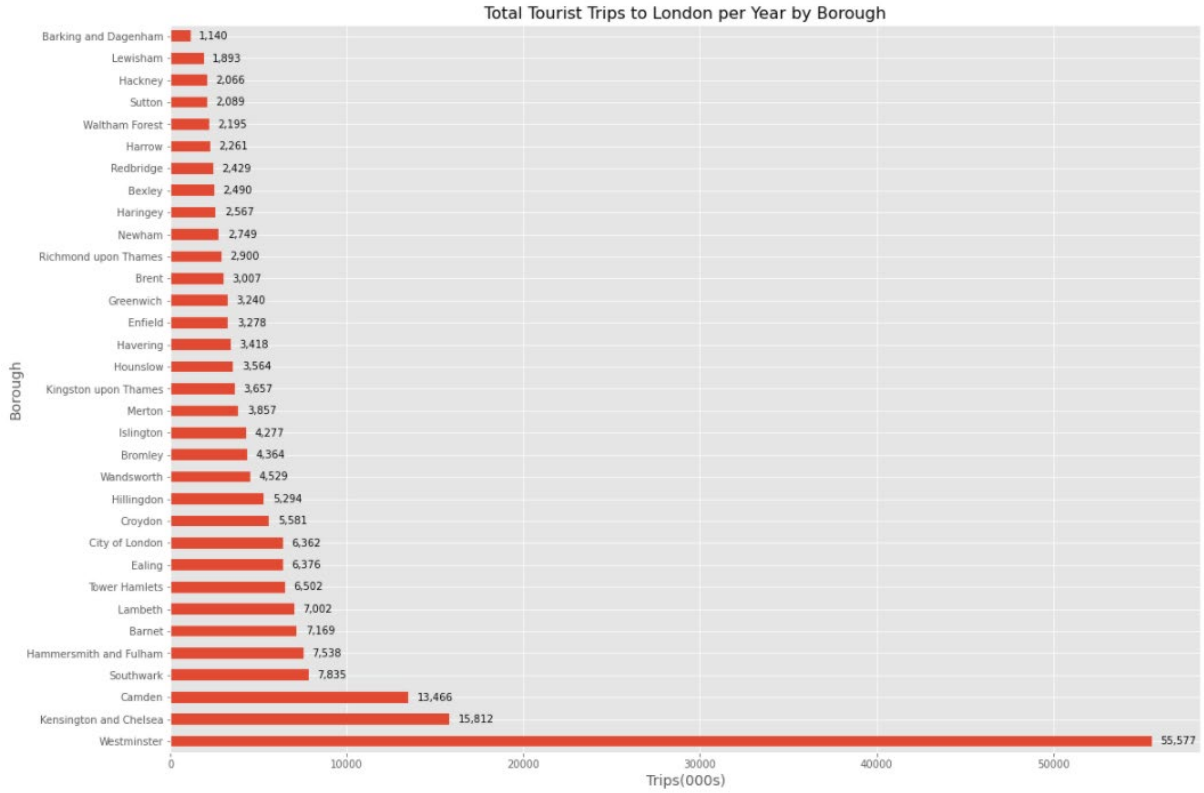
Figure 12 - data frame with Cluster Label added

	Borough	Latitude	Longitude	TotalTrips	Cluster Value	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	Cluster Label
0	Barking and Dagenham	51.5607	0.1557	1140	0	Bus Stop	Turkish Restaurant	Supermarket	Discount Store	Golf Course	Soccer Field	Grocery Store	Gas Station	Gym / Fitness Center	Park	Atypical, non-hospitality
1	Barnet	51.6252	-0.1517	7169	1	Pub	Park	Bus Stop	Fish & Chips Shop	Gym	Café	Electronics Store	English Restaurant	Escape Room	Exhibit	Pubs and coffee shops
2	Bexley	51.4549	0.1505	2490	1	Pub	Clothing Store	Fast Food Restaurant	Hotel	Supermarket	Coffee Shop	Pharmacy	American Restaurant	Nightclub	Pizza Place	Pubs and coffee shops
3	Brent	51.5588	-0.2817	3007	2	Coffee Shop	Hotel	Bar	Clothing Store	Grocery Store	Pub	Pizza Place	Sporting Goods Shop	Indian Restaurant	American Restaurant	Coffee shops & hotels
4	Bromley	51.4039	0.0198	4364	1	Pub	Clothing Store	Coffee Shop	Indian Restaurant	Supermarket	Café	Bar	Pizza Place	Electronics Store	Burger Joint	Pubs and coffee shops

3.8. Categorising tourist trips data

To increase understanding of the tourist trips data, it was first visualised in a bar chart to give an idea of the differences found between boroughs:

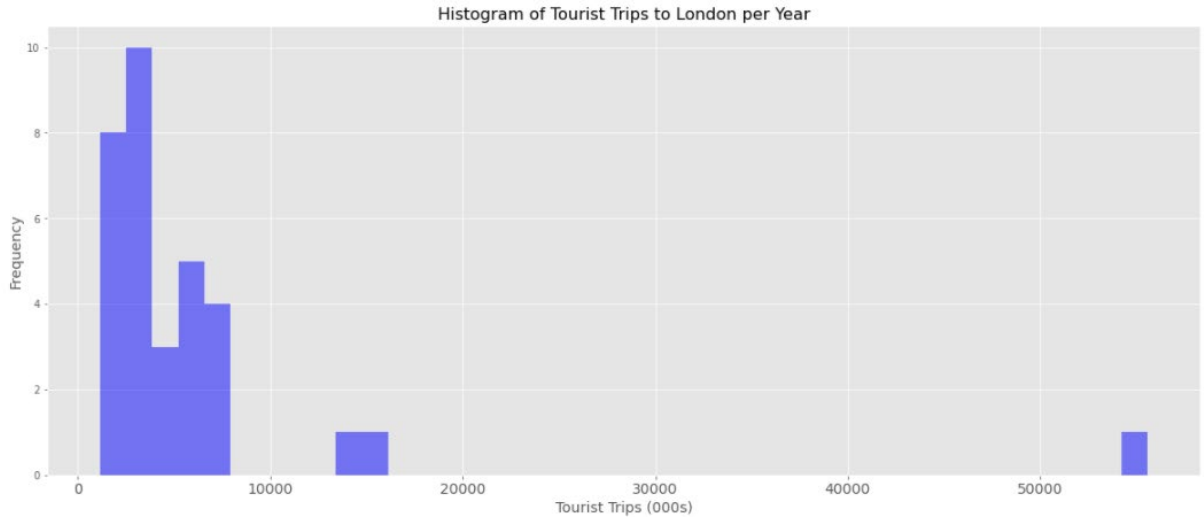
Figure 13 - horizontal bar chart showing number of tourist trips



The chart indicated most boroughs to have between 2,000 and 8,000 total trips per year (in thousands), with a few notable exceptions. Westminster is something of an outlier with over 55,000.

A histogram was also created to help visualise this:

Figure 14 - histogram illustrating number of tourist trips



To help the ease with which the tourist trips data could be worked with, they were categorised based on the following criteria:

- Below 3000: "Low"
- 3000-5000: "Moderate"
- 5000-8000: "High"
- Above 8000: "Very High"

Each borough was assigned a 'Popularity' label based using these criteria, which was added to the data frame. This could then be used to provide additional descriptive information in the label pop out to help a user when selecting a borough on our final map.

A final step needed to be taken to deal with the outlier Westminster. With the aim of creating a choropleth map illustrating the number of tourist trips per borough as the solution to our problem, an outlier such as this can skew the visualisation. The likely outcome would be Westminster displayed extremely dark on the map, with nearly all other boroughs displayed as the same very light colour; the map would lack clear differentiation when visualisation of the number of tourist trips between boroughs.

To address this, a new column in the main data frame was created called 'TouristTripsCapped'. This used the same tourist trip figures as found in the 'TouristTrips' column, except for Westminster which was recoded to 20,000. When visualised, the result of this should be to still indicate Westminster as the area with the most trips but allow for clearer differentiation between the other boroughs.

The final data frame to be used for visualisations in the results section therefore consisted of the following:

Figure 15 - final data frame with all information needed to produce choropleth map

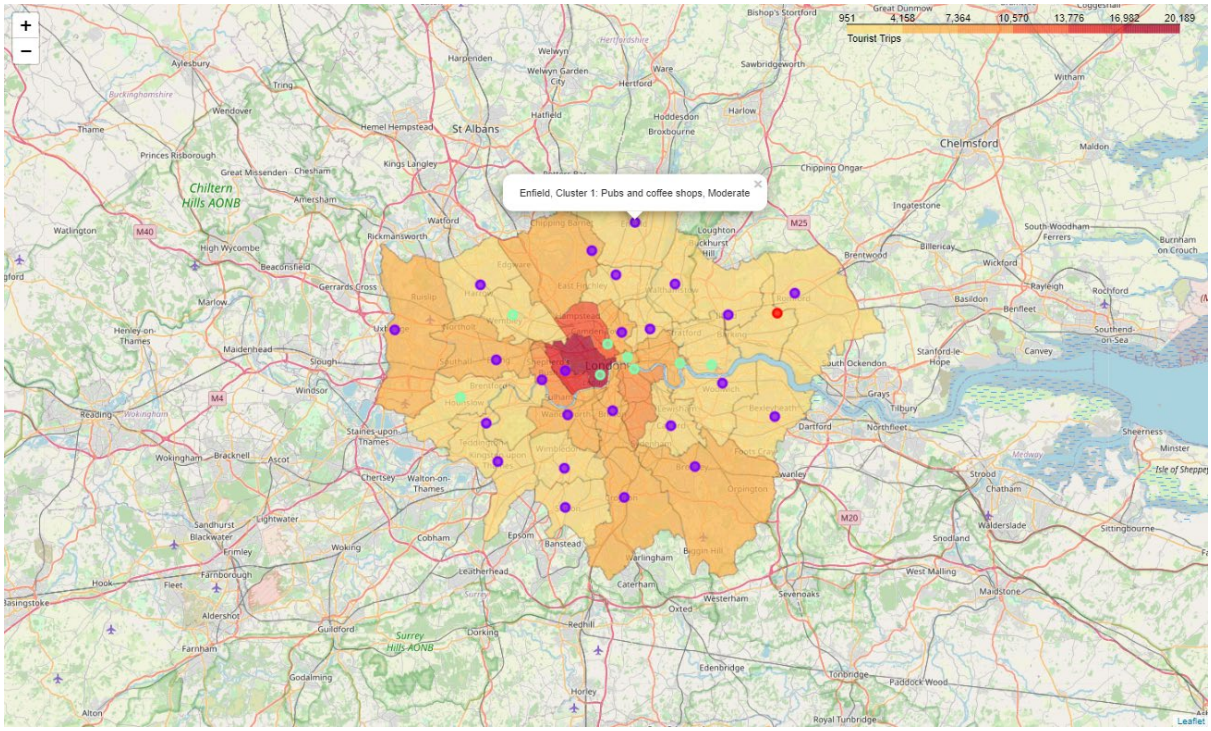
	Borough	Latitude	Longitude	TotalTrips	Cluster Value	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster Label	Popularity	TotalTripsCapped
28	Tower Hamlets	51.5099	-0.0059	6502	2	Coffee Shop	Park	Hotel	Italian Restaurant	Light Rail Station	Bus Stop	Sandwich Place	English Restaurant	Lounge	Restaurant	Coffee shops & hotels	High	6502
29	Waltham Forest	51.5908	-0.0134	2195	1	Pub	Café	Gym / Fitness Center	Art Gallery	Coffee Shop	Pizza Place	Brewery	Restaurant	Multiplex	Museum	Pubs and coffee shops	Very Low	2195
30	Wandsworth	51.4567	-0.1910	4529	1	Pub	Gym / Fitness Center	Grocery Store	Coffee Shop	Café	Park	Clothing Store	Breakfast Spot	Pizza Place	Asian Restaurant	Pubs and coffee shops	Moderate	4529
31	Westminster	51.4973	-0.1372	55577	2	Coffee Shop	Hotel	Theater	Park	Café	Garden	Sushi Restaurant	Sandwich Place	Pub	Sporting Goods Shop	Coffee shops & hotels	Very High	20000
32	City of London	51.5155	-0.0922	6362	2	Coffee Shop	Gym / Fitness Center	Hotel	Cocktail Bar	Modern European Restaurant	Steakhouse	Falafel Restaurant	French Restaurant	Theater	Event Space	Coffee shops & hotels	High	6362

4. Results

4.1. Choropleth map with markers and descriptive labels

Geo json data for London [5] was saved to github and then downloaded and used to create the choropleth map. Markers were added to identify each borough, with the cluster each belongs to indicated by colour. Selecting a marker information provides a pop out with descriptive information including the description of the cluster and popularity label. The shading of each borough relates to the number of tourist trips per year, with deeper shades representing more trips:

Figure 16 - choropleth map with a borough marker selected showing its label information



5. Discussion

The map indicates how the boroughs are clustered by the colour of the markers. A clear difference between more outer located, potentially suburban boroughs (Cluster 1 – purple markers) and more centrally located boroughs (Cluster 2 – light green markers) can be clearly observed. This geographical difference between these two clusters makes sense when considering the most commonly found venues within each. It can be seen that while both have a high number of coffee shops, Cluster 1 certainly has more pubs as the most common venue. On the other hand, The most common venue aside from coffee shops in Cluster 2 boroughs looks to be hotels. This makes sense as hotels are typically located centrally for easy access around a city and to visit tourist destinations.

It is also clear to note the difference in shading of the boroughs based on tourist trips. Very centrally located areas see more trips, which is to be expected when considering the tourist venues found in these locations such as Buckingham Palace and the Houses of Parliament. However, it is also notable that differences can be see between boroughs that are further out from central locations.

So, in answer the business problem of this project: where would be some ideal locations to open a new hospitality venue in London, specifically a craft beer bar?

I would argue that any borough where tourism is 'High' or 'Very High' according to the Popularity labels assigned would be a useful first criterion. Secondly, any boroughs that are also in Cluster 1 are likely to be viable locations due to hospitality venues, specifically pubs, already being commonly found there. This is likely to mean that a new craft beer bar would have a good chance of being successful with a good target audience already in place. A final consideration not fully explored here but nonetheless important is the cost of commercial rentals; it is likely to cheaper to rent a commercial space in Cluster 1 boroughs as they are less centrally located. Cost of premises is clearly an important consideration when opening a business.

A final list of suitable boroughs can therefore be created that is based on the above criteria:

1. Tourist trips: a 'Popularity' label of 'High' or 'Very High', indicating above 5000 tourist trips per year.

2. Location / commonly found venues: any boroughs belonging to Cluster 1, where a borough has a less central location and a majority of pubs in the area.

The main data frame can be filtered based on the above criteria to give us a final list of suitable boroughs:

Figure 17 - final list of suitable boroughs

	Borough	Latitude	Longitude	TotalTrips	Cluster Value	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	Cluster Label	Popularity	TotalTripsCapped
1	Barnet	51.6252	-0.1517	7169	1	Pub	Park	Bus Stop	Fish & Chips Shop	Gym	Café	Electronics Store	English Restaurant	Escape Room	Exhibit	Pubs and coffee shops	High	7169
6	Croydon	51.3714	-0.0977	5581	1	Coffee Shop	Pub	Hotel	Platform	Clothing Store	Bookstore	Indian Restaurant	Italian Restaurant	Mediterranean Restaurant	Portuguese Restaurant	Pubs and coffee shops	High	5581
7	Ealing	51.5130	-0.3089	6376	1	Coffee Shop	Pub	Café	Italian Restaurant	Indian Restaurant	Thai Restaurant	Pizza Place	Park	Bakery	Hotel	Pubs and coffee shops	High	6376
11	Hammersmith and Fulham	51.4927	-0.2339	7538	1	Pub	Café	Indian Restaurant	Coffee Shop	Park	French Restaurant	Gym / Fitness Center	Japanese Restaurant	Italian Restaurant	Gastropub	Pubs and coffee shops	High	7538
15	Hillingdon	51.5441	-0.4760	5294	1	Coffee Shop	Pub	Clothing Store	Fast Food Restaurant	Pharmacy	Gym	Italian Restaurant	Grocery Store	Department Store	Park	Pubs and coffee shops	High	5294
18	Kensington and Chelsea	51.5020	-0.1947	15812	1	Café	Restaurant	Pub	Italian Restaurant	Juice Bar	Clothing Store	Garden	Coffee Shop	Bakery	Hotel	Pubs and coffee shops	Very High	15812
20	Lambeth	51.4607	-0.1163	7002	1	Coffee Shop	Pub	Caribbean Restaurant	Pizza Place	Cocktail Bar	Beer Bar	Tapas Restaurant	Market	Yoga Studio	Music Venue	Pubs and coffee shops	High	7002

6. Conclusion

6.1. Summary

This study has demonstrated the usefulness of several data science techniques. Web scraping using Beautiful Soup is a useful method of efficiently sourcing publicly available data. Foursquare is a valuable API that can access useful location data. Folium is a great map builder with the flexible to combine different types of information into powerful visualisations. The K-Means method of clustering is a machine learning technique that can successfully segment a dataset and provide valuable insights into the similarity of different locations. The approach taken has helped to successfully narrow down multiple locations in a large city to a shortlist of contenders, in this case as potentially ideal locations to open a new hospitality business. Therefore, to a reasonable extent the business problem set out in the Introduction section has been addressed.

The interactive map created in this study is likely to be of use to anybody with an aim of entering the hospitality industry. They would have the ability to study different areas and get an initial impression of the most and least suitable areas to locate their business. Equally someone who already runs a hospitality-based business and desires to expand into other London locations would find value.

Tourists visiting London would be another population who could find this project useful; they can quickly discern the ideal places to visit based on their trip requirements, particularly where time is limited.

Finally, it could also be of interest to those wishing to move to London or somewhere else in the city if already a resident. Knowing the typical types of venues found at each location can save time for someone who has a good idea of what they want to live near.

6.2. Future Research

While this study has provided some useful insights, it is important to recognise where improvements could be made and any future directions for subsequent research. An obvious improvement would be to integrate even more information to help answer the business problem - when deciding where to open a bar there are certainly other relevant factors to

consider. One factor could be the average cost of a commercial letting in each borough. For example, while Kensington & Chelsea is on the shortlist, it being in Cluster 1 and having a 'Very High' tourist label, it is likely to have extremely expensive commercial rental rates, making it potentially less viable as a location to start a hospitality business.

Knowing more about the demographic make-up residents in each borough would also be valuable to know; one might assume that boroughs with a lower average age are likely to generate more business based on an increased orientation towards more socially focused venues such as pubs and bars.

References

[1] Wikipedia – London: https://en.wikipedia.org/wiki/London

[2] Wikipedia – London boroughs: https://en.wikipedia.org/wiki/List_of_London_boroughs

[3] Foursquare API: https://developer.foursquare.com/

[4] UK Government data site - tourism trips to London boroughs: https://data.gov.uk/dataset/ee5038be-d2be-4ab6-a612-70ade60eca12/tourism-trips-borough

[5] London Data Store - geo json data for London boroughs: https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london?resource=9ba8c833-6370-4b11-abdc-314aa020d5e0