Location Search for New Coffee Shops

Applied Data Science Capstone Project

Contents

Introduction

An international coffee chain that is specializing in selling coffee and other beverages from different countries around the world is planning to enter the UK market within the following two years. Top management considers this to be a major step in the company's expansion plans that entails considerable risks. United Kingdom hosts main international coffee chains and many more local coffee shops can be found in all major cities. Thus, the company expects to meet strong competition from rival brands and local coffee shops. Moreover, costs for opening and operating a new coffee shop in UK are expected to be very high, especially when it comes in leasing properties in metropolitan areas such as London.

In view of all these, the department that is responsible for planning and implementing the UK expansion plans, has decided that the best course of action is to focus in London, and start with a few coffee shops, three to four, in central London. In order to find the best location and understand and map competition in central London, they assigned to a data analyst to explore coffee shops and venues in areas located in Western [1], Western Central [2], South Eastern [3], and South Western London [4]. They are mostly interested in learning how coffee shops are distributed in these areas, get a segmentation and description for these areas based on venues found there, e.g. are they near theatres, parks or other venue types, and if possible find the average price range and rating for coffee shops in these areas. Hoping to minimize risks and costs, they expected that through this survey they will be able to pinpoint locations where new coffee shops can fit.

Data

The primary data source will be the Foursquare API [5], from which data on venues across different locations in central London will be used. Due to the restrictions that Foursquare imposes on the number of calls and the type of available data, information for approximately 50 to 100 venues per postcode district [6] will be collected in a radius of 500 meters around the centre of each area. Venues will include coffee shops, restaurants, theatres, parks museums and any other venue types available through the Foursquare API. Data will be aggregated to generate statistics per postcode district, such as frequency for different venue categories, top venues according to frequency of occurrence, and

density of coffee shops or similar shops per postcode district. These metrics will be used to identify potential areas where new coffee shops could fit, and for exploring and segmenting areas according to the most frequent venues encountered there.

Some additional metrics that could be used are the average ratings and price range for coffee shops per postcode district. To accomplish this, we need to use the venues details calls in Foursquare API [5] which is a premium call, and a personal account has a limit of 500 premium calls per day. To overcome this obstacle a selection of limited coffee shops, e.g. top 5 per postcode district will be made, or perhaps select coffee shops for only few postcode districts, e.g. after narrowing down the list of interesting areas.

To be able to extract data from Foursquare API [5], coordinates for each postcode district must be available. For this reason a data frame with details and postcodes per area will be scrapped from Wikipedia, see references [1] to [4], and coordinates will be obtained from the Free Map Tools [7].

Methodology

The main objective for the project was to answer questions which will help the team responsible for the UK expansion to decide which locations are perhaps best for opening new coffee shops. Being more specific the expansion team would like first to know in which locations in London coffee shops are less frequent. This would give an indication of where a new coffee shop might fit, and for answering this question the coffee shops frequency was calculated with data from Foursquare. Next, they would like to know how London areas can be segmented, how do the candidate locations differ from other locations and what are their main features in terms of most frequent venues found there. Segmenting London postcode districts with the k-means algorithm was the method chosen for answering these questions. Finally, the expansion team would also like to get some insights about competitors in the candidate locations by exploring ratings and price ranges for coffee shops in London.

The analysis was carried out with Python in Watson Studio [8]. The main libraries used were pandas for scrapping, preparing and analysing data, folium for visualizing maps, and sklearn for k-means clustering. The main steps in the analysis were the following:

1. Scrape data from Wikipedia and Free Map Tools in order to create a data frame with information for London postcode districts, including their coordinates.
2. Using the coordinates from step 1, and the explore and venue details endpoints from Foursquare obtain information about venues in the four areas discussed in the introduction.
3. Generate metrics such as coffee shops and other venues frequency, average ratings and average price range for coffee shops per postcode district.
4. Explore the results with bubble maps to get insights about London locations and perform k-means clustering to segment London postcode districts based on which are the most frequent venue categories in each postcode district.

The Jupyter Notebook is published in GitHub and is accessible through this link.

Results

The main result from the analysis is that in districts around Soho and Chinatown, and in areas such as Kensington, Chelsea and South Kensington, coffee shops are less frequent than in other postcode districts. These areas are shown in figure 1, where a bubble map is shown with the size of the markers being proportional to the frequency of coffee shops for the corresponding district.

Looking into how these areas are segmented, we see that they all belong to the same cluster. This is the biggest cluster by far found when applying the k-means algorithm. The elbow technique was used to find the optimum k, and for k between 3 and 6 there was significant change in the sum of squared distances. Thus, the final selection for k was 6, and the result from k-means was a very big cluster, two smaller clusters, and three clusters with a single district each.

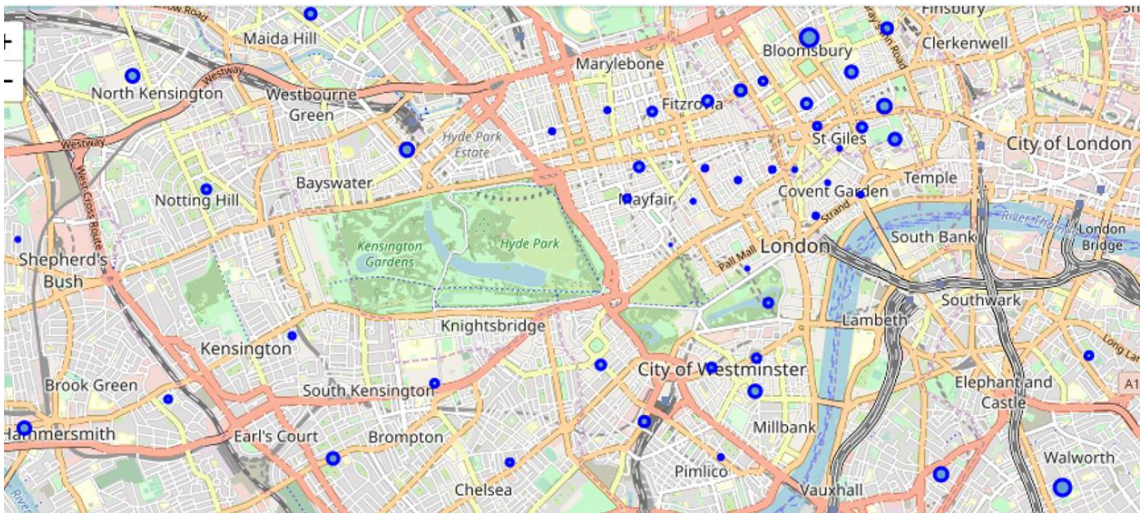

Figure 1 Bubble map for the frequency of coffee shops across different London districts. The size of the marker is proportional to the frequency of coffee shops in the corresponding district.

The biggest cluster had in total 73 districts, which are mainly located in the Western Central part of London, see blue circles in figure 2. These areas are mainly characterised by the fact that they have similar top-10 most common venue categories, such as hotels, theatres, coffee shops and restaurants of different type. The second biggest cluster, see purple circles in figure 2, has 12 districts, which can be characterised as areas of pubs since most of them have as 1st most common venue pubs. Finally, a cluster, see red circles in figure 2, is formed by six areas with common feature the high frequency of grocery stores.

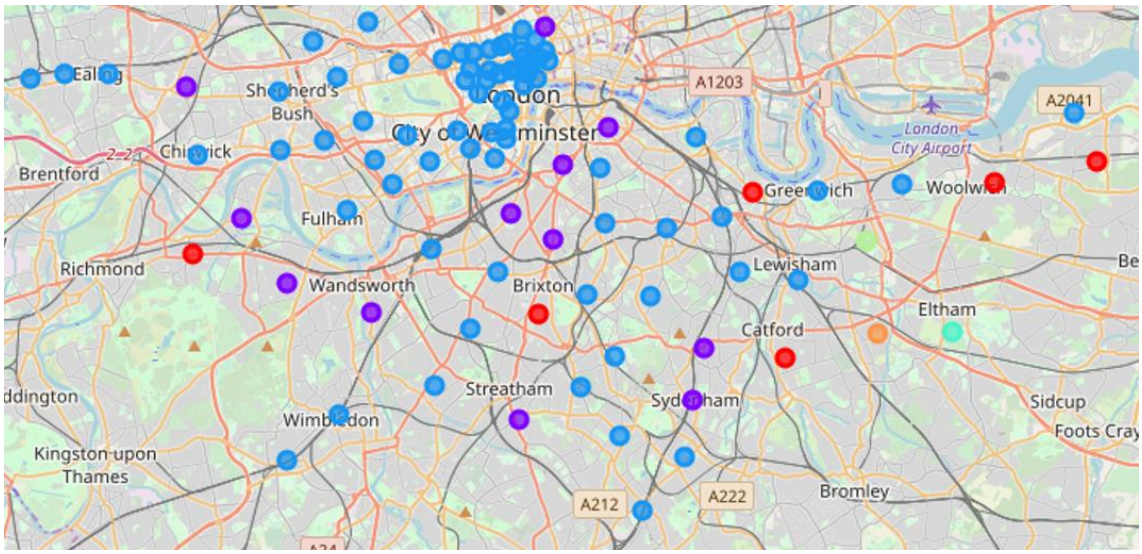

Figure 2 Segments of London districts obtained with the k-means algorithm. The blue circles are districts that belong to the biggest cluster, whereas the purple and red circles correspond to districts that belong to two smaller clusters.

Finally, focusing in the areas with low presence of coffee shops, we see that in terms of prices coffee shops there are characterised by low to medium average price ranges. Furthermore, we see that

coffee shops in these areas are highly rated. This means that the coffee chain will meet two important barriers if a decision is made to enter the London market. First, it won't be easy to compete by offering better prices, and second they will have to compete with other coffee shops which have established a good reputation. Figures 3 and 4 show bubble maps for the average price range and average rating for coffee shops across different London districts near the areas of Soho, Covent Garden and Kensington.

Figure 3 Bubble map for the average price range for coffee shops across different London districts. The size of the marker is proportional to average price range for coffee shops in the corresponding district.

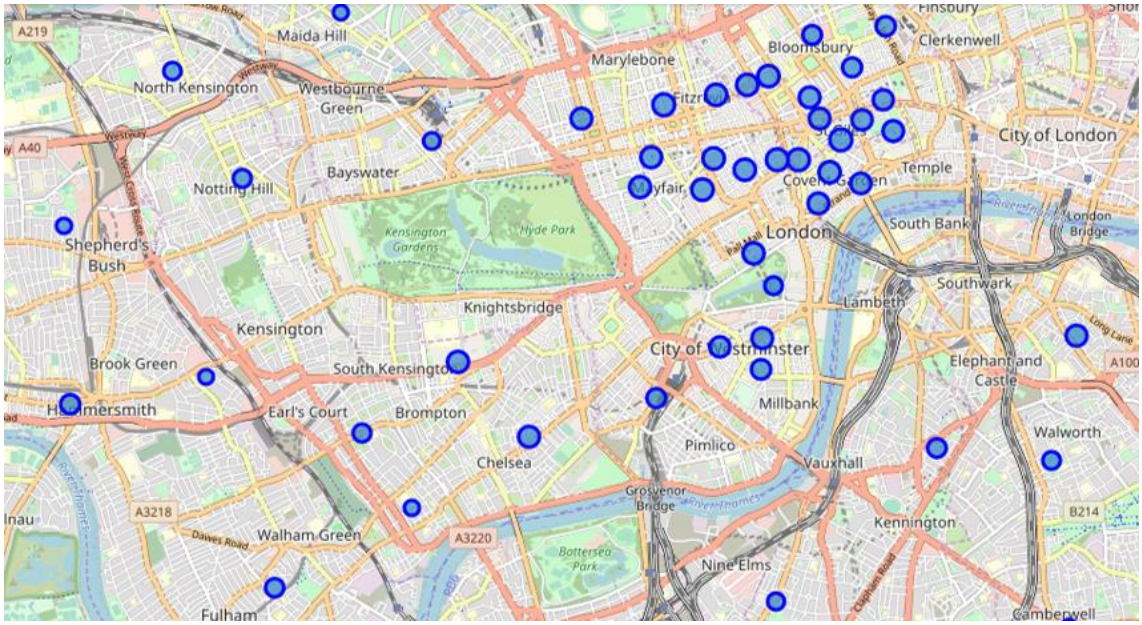

Figure 4 Bubble map for the average rating for coffee shops across different London districts. The size of the marker is proportional to average rating for coffee shops in the corresponding district.

Discussion

Although the analysis has shown that there are interesting areas in London for the company to launch new coffee shops, further investigation is needed. Getting data from other location data providers,

might add more value by revealing more interesting insights about London districts. Moreover, the analysis here focused only on a few aspects of the problem and one must expand the analysis across other dimensions. These can be cost factors such as labour costs, or commercial properties costs, demographics for these areas, i.e. local population age profile, income and spending statistics. In addition to all these, surveys among Londoners might give more insights about their habits and preferences when it come to personal time and leisure and can be useful insights when designing the concept of the new coffee shops and marketing campaign for promoting them.

Conclusion

It is well known that London is a rather competitive environment for a new enterprise across any line of business. For this reason, is important carefully plan all steps from the beginning to the end, i.e. from conceptualization of the new business, design, development and launching. This project aimed at assisting a team of people in making the right decision on opening or not, and where to open new coffee shops in London. Exploiting the Foursquare API, an attempt to find potential areas for new coffee shops was made. The analysis revealed that there are areas where a coffee shop might fit, but the company must pay attention to the fact that existing coffee shops are characterised by low to medium range prices and high ratings. Moreover, the analysis must be extended to other dimensions to capture various cost factors and London demographics and give insights about risks that the company might face when launching the new coffee shops.

References

[1] Wikipedia, “W Postcode Area,” [Online]. Available: https://en.wikipedia.org/wiki/W_postcode_area. [Accessed 3 10 2019].

[2] Wikipedia, “WC Postocode Area,” [Online]. Available: https://en.wikipedia.org/wiki/WC_postcode_area. [Accessed 3 10 2019].

[3] Wikipedia, “SE Postcode Area,” [Online]. Available: https://en.wikipedia.org/wiki/SE_postcode_area. [Accessed 3 10 2019].

[4] Wikipedia, “SW Postcode Area,” [Online]. Available: https://en.wikipedia.org/wiki/SW_postcode_area. [Accessed 3 10 2019].

[5] “Foursquare Developers Portal,” Foursquare, [Online]. Available: https://developer.foursquare.com/.

[6] Wikipedia, “London postal district,” [Online]. Available: https://en.wikipedia.org/wiki/London_postal_district. [Accessed 03 10 2019].

[7] Free Map Tools, [Online]. Available: https://www.freemaptools.com/download-uk-postcode-lat-lng.htm. [Accessed 3 10 2019].

[8] IBM, “IBM Watson Studio,” [Online]. Available: https://www.ibm.com/watson.