Finding the best location to open a coffee shop

Marios Djakouris

02/08/2021

1. Introduction

1.1 Background

The London Borough of Bromley is the southeasternmost borough of Greater London with an estimated population of around 332000. This borough can be a great opportunity for any coffee shop owner who seeks to capitalise on London's rich market. Nevertheless, choosing a suitable location may be challenging as Bromley is the largest borough of Greater London by total area. Bromley is made up of 8 Postcode districts, with each one having a Post town which can be considered the centre of activity and commerce for each district.

Coffee shops are retailers who sell small quantities directly to consumers at a higher price, therefore a suitable location is the one which offers great accessibility and generates lots of foot traffic, meaning that neighbouring businesses can significantly impact profitably in a positive or negative way. Coffee shops can be considered Monopolistic Competition, they have low barriers of entry with lots of coffee shops selling similar products, thus selecting a location with a low number of competitors is vital.

1.2 Problem

Those who seek to open a coffee shop need to take great consideration on choosing a suitable location, but the main problem they face is the Information gap which exists. Essentially, they usually don't have available all the information required in a presentable format in order to be capable of undertaking a fully-informed decision. Taking into consideration the aforementioned, the main objective is to create maps and charts which will bring more clarity on how location can generate more foot traffic, while also exploring the impact which neighbouring businesses can have on profitability.

1.3 Interest

The ones who will mainly be interested in the best location of coffee shops are the coffee shop owners who wish to enter or expand in the Borough of Bromley market. Furthermore, those who sell similar products or wish to do so may also find this useful. This project has the potential in helping various types of establishments in deciding whether they should add coffee service in their establishment. This is due to the fact that there are no high barriers in serving coffee. Finally, anyone who wish to live or be located near a location with lots of foot traffic and good accessibility to basic services, including consumers, will probably find interest in this project.

2. Data

2.1 Data sources

I used Wikipedia to get information about each postcode district [1], which included Post town, Coverage and Local authority area(s). Moreover, I created and imported a csv file which includes the population of each post code and the coordinates of post towns. Google

Map was used to get the Latitude and Longitude for each post town [2] and postcodearea.co.uk provided the population of each postcode district [3]. Finally, with Forsquare API i got information about all the venues within the proximity of each Post town [4]. Specifically, I got the venue name, latitude, longitude, category and the post town it belongs to.

2.2 Data processing

The database included data about the Local Authority Area which is relating to the administration centre of each postcode district, while the Coverage column provided more detail on what areas each post code covered. The Post Towns are the main towns of each Postcode district and the ones which will most likely generate the highest foot traffic, convenience and profitability for a coffee shop location, so in this analysis we will use them as the centre and explore venues in close proximity. Thus, the Local Authority Area and Coverage are redundant and I removed them from my dataset. Moreover, BR2 has 2 post towns and was therefore shown twice in the database. I decided that for BR2 post code district I should use the large town of Bromley as the Post town, instead of the village of Keston, again due to foot traffic, convenience and profitability concerns. Therefore, I removed from the database the row with the Keston Post town. Next, I grouped my data by Post town, due to being the main focus of this analysis, while joining in a single line post codes with mutual post towns. While working on this project, the website wrongly changed the Post town of West Wickham into London and the BR4 post code into BR4, SE6, CR0. I replaced the wrong information with the correct ones and even though the website was again update with the correct town and post code, I kept the replace code just in case the website change to its previous version, as its not affecting the data. Also, the csv file I created and imported enabled me to have a dataframe which includes each Post town's coordinates and the total population of all Post Code Districts each Post town represents. This format allowed me to merge the population and the coordinates with my main dataset, doing so by creating an inner join on Postcode district.

3. Methodology

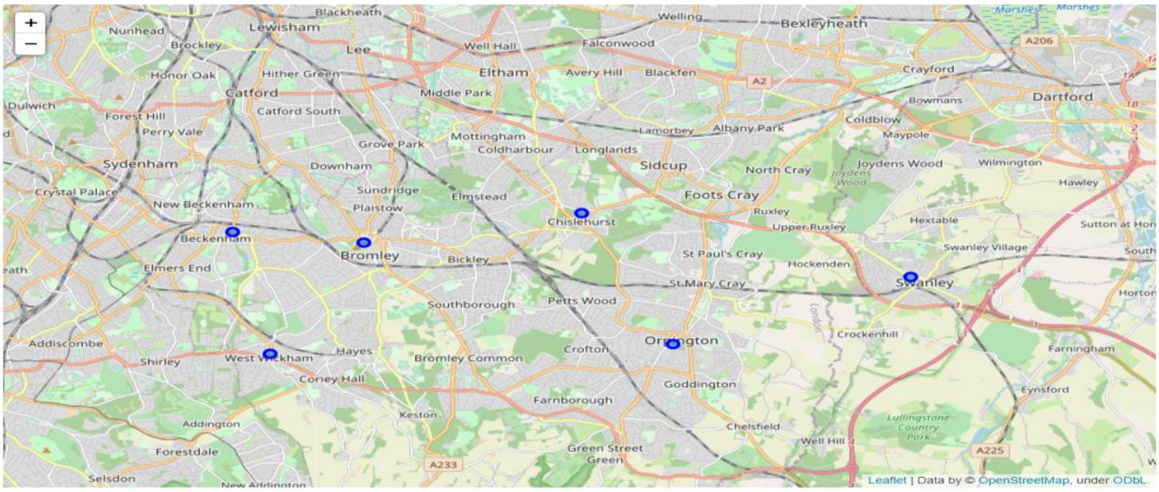
3.1 Foursquare API

As can be seen below, my main dataset includes the Postcode district, Post town, Latitude, Longitude and Population.

Out[14]:

	Postcode district	Post town	Latitude	Longitude	Population
0	BR3	BECKENHAM	51.408780	-0.025260	47,411
1	BR1,BR2	BROMLEY	51.406025	0.013156	100,920
2	BR7	CHISLEHURST	51.413785	0.076756	17,322
3	BR5,BR6	ORPINGTON	51.379588	0.103539	92,084
4	BR8	SWANLEY	51.397170	0.173210	22,053
5	BR4	WEST WICKHAM	51.376826	-0.014540	19,367

Using this dataset and python folium, I was able to visualize them and observe the location of post towns, exploring them on a map and zooming in and out as necessary. This way, we can easily conclude that these post towns are spread out sufficiently throughout the whole Borough, thus using them as the centre point of this analysis is a good way to cover all post codes districts so as to find the best location.



My main dataset and Foursquare API allowed me to get information about the top 100 venues that are within a radius of 500 meters of each Post town. As we can see below, using the main dataset and Foursquare account details, we can send a GET request which gives us a dataframe with the venue's name, latitude, longitude, category and post town. In detail, this was done by creating a Foursquare account with generated Private client ID, Client Secret, Foursquare API version and a default Foursquare API limit value. Please note that there was a limit on the number of venues I could extract due to my account being the free version.

Out[20]:

	Post town	Post town Latitude	Post town Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	BECKENHAM	51.40878	-0.02526	Kelsey House Bar & Kitchen	51.407673	-0.025917	Cocktail Bar
1	BECKENHAM	51.40878	-0.02526	Ristorante Pizzeria Sapore Vero	51.408288	-0.025251	Italian Restaurant
2	BECKENHAM	51.40878	-0.02526	Caffè Nero	51.407498	-0.027477	Coffee Shop
3	BECKENHAM	51.40878	-0.02526	Breeze Yoga	51.410077	-0.023541	Yoga Studio
4	BECKENHAM	51.40878	-0.02526	Tapeo	51.407164	-0.026438	Tapas Restaurant

Looking deeper into the venue categories in each post town, I used one hot encoding which is a method that converts categorical data into binary vectors in order to enable the analysis of categorical variables such as venue categories. Utilising this method, I created a dataframe with the first column being the post town and the rest columns being all the venue categories. An integer value of 1 points out the category of a venue, with all the rest venue categories of that venue having a number of 0 which indicates that a venue doesn't belong to a particular venue category.

Out[21]:

	Post town	Asian Restaurant	Auto Garage	Bakery	Bar	Bookstore	Bridal Shop	Burger Joint	Café	Chocolate Shop	Clothing Store	Cocktail Bar	Coffee Shop	Deli / Bodega	Department Store	Di
0	BECKENHAM	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	BECKENHAM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	BECKENHAM	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
3	BECKENHAM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	BECKENHAM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

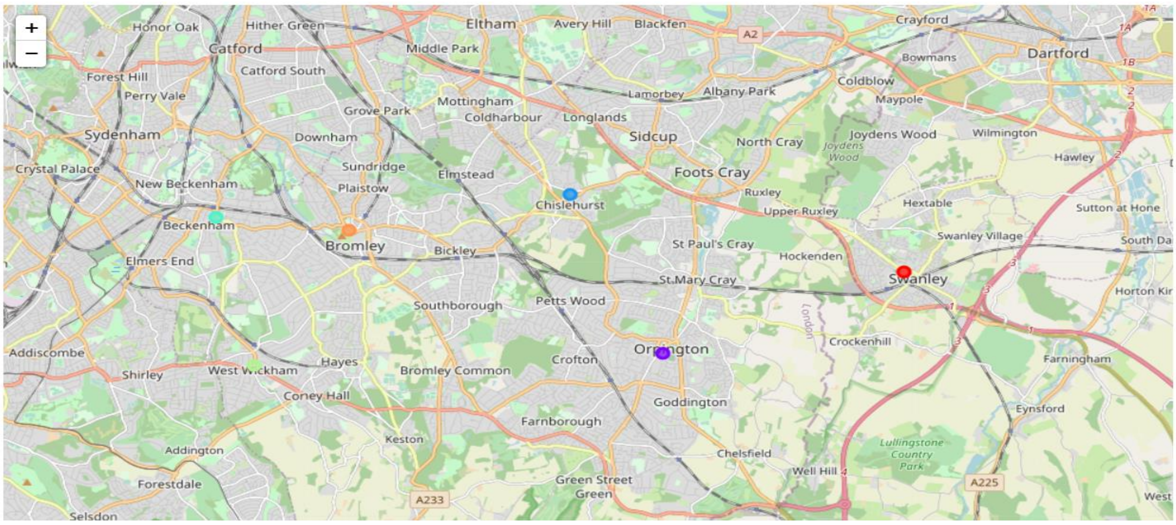
Next, the mean frequency of occurrence of each category was calculated by grouping the rows of this dataframe by the Post town. Therefore, we can see below for instance that 8.8889% of all venues in post town Beckenham belong to the venue category Café.

Out[22]:

	Post town	Asian Restaurant	Auto Garage	Bakery	Bar	Bookstore	Bridal Shop	Burger Joint	Café	Chocolate Shop	Clothing Store	Cocktail Bar	Coffee Shop	B
0	BECKENHAM	0.000000	0.000000	0.000000	0.044444	0.022222	0.000000	0.000000	0.088889	0.000000	0.022222	0.022222	0.088889	0.0
1	BROMLEY	0.022727	0.022727	0.000000	0.045455	0.022727	0.022727	0.045455	0.045455	0.022727	0.113636	0.000000	0.090909	0.0
2	CHISLEHURST	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	0.0
3	ORPINGTON	0.000000	0.000000	0.066667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
4	SWANLEY	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.166667	0.0
5	WEST WICKHAM	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	0.058824	0.0

3.2 K-means Clustering

K- means clustering algorithm is a common type of unsupervised learning and in this case, I used it to cluster the Borough of Bromley into 6 cluster groups, one cluster group for each post town. As I mentioned before, each postcode district has its own Post town which can be considered the centre of activity, commerce and traffic for each district, which is important for a coffee shop location. Also, when we visualised the location of these post towns, we concluded that they spread out sufficiently throughout the whole Borough, thus post towns can act as suitable centroids. Hence, the top 100 venues within a radius of 500 meters of each Post town were being clustered under the same category.



Using a function which included the mean frequency of occurrence of each category and for loop, I created a new dataframe that ranks the top 10 most common venues for each post town. In addition, on this dataframe I inserted the cluster label for each post town, an allocated number from 0 to 5, while also I merged it with the main dataframe, adding the postcode district, latitude, longitude and population for each Post town. Thus, this way we can clearly observe the prevalent venue categories in each cluster, which can give us some indication on a suitable location for a coffee shop. As shown, the new dataframe states for instance that Beckenham's number one most common venue category is coffee shop, which is a negative factor since too much competition should be avoided.

	Postcode district	Post town	Latitude	Longitude	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	
0	BR3	BECKENHAM	51.408780	-0.025260	47,411	3	Coffee Shop	Italian Restaurant	Café	Supermarket	Pharmacy	Platform	Iris
1	BR1,BR2	BROMLEY	51.406025	0.013156	100,920	5	Clothing Store	Coffee Shop	Pub	Bar	Burger Joint	Café	Pa
2	BR7	CHISLEHURST	51.413785	0.076756	17,322	2	Indian Restaurant	Gastropub	Café	Italian Restaurant	Pub	Portuguese Restaurant	Me Re
3	BR5,BR6	ORPINGTON	51.379588	0.103539	92,084	1	Pub	Restaurant	Movie Theater	Hotel	Gym / Fitness Center	Grocery Store	Pa
4	BR8	SWANLEY	51.397170	0.173210	22,053	0	Indian Restaurant	Supermarket	Furniture / Home Store	Coffee Shop	Pizza Place	Fast Food Restaurant	Po Re

4. Results

4.1 Relevant venue categories in each Post town

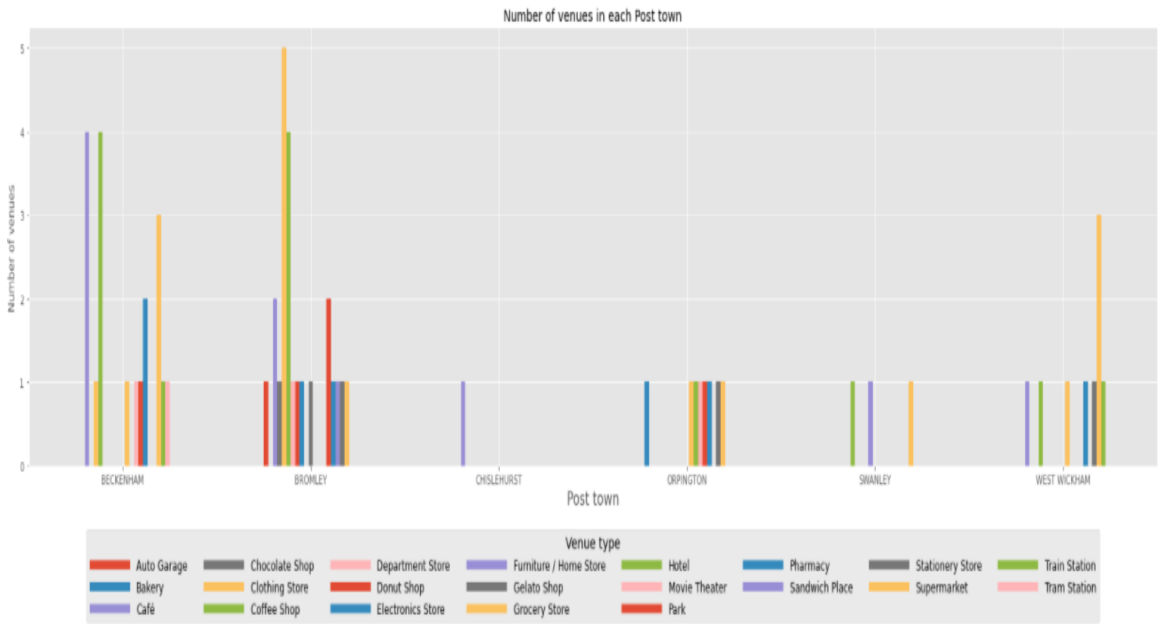
Now we are in a good position to generate relevant and targeted results, with appropriate visualisation which can help us identify the best location to open a coffee shop. Firstly, I got a list with all the unique venue categories included in the data obtained from Foursquare API. My objective was to identify which type of venues will have a negative or positive impact on the profitability of a coffee shop, while also filtering out irrelevant venue categories. Specifically, we need to consider that coffee shop's target customers who will buy coffee and light food between morning and early afternoon. Restaurant, pubs and bars for instance mostly serve and generate traffic during the night, thus is better to remove them from this analysis in order to enhance further the reliability and clarity of the results. On the other hand, venue categories like Clothing Stores, Department Stores, Train Stations and Parks need to be included due to their potential to increase foot traffic and subsequently the profitability of coffee shops. Moreover, venue categories such as Coffee Shop, Cafés, Donut Shops and Sandwich Places are competitors and need to also be included. Obviously, they serve similar or the same products with coffee shops, primarily during the same time of the day, meaning they can detrimentally impact the profitability of coffee shops.

```
array(['Cocktail Bar', 'Italian Restaurant', 'Coffee Shop', 'Yoga Studio',
       'Tapas Restaurant', 'Fish & Chips Shop', 'Grocery Store',
       'Supermarket', 'Café', 'Mediterranean Restaurant', 'Diner',
       'Indian Restaurant', 'Pub', 'Park', 'Kebab Restaurant',
       'Clothing Store', 'Deli / Bodega', 'Portuguese Restaurant',
       'Steakhouse', 'Pharmacy', 'Nightclub', 'Bar', 'Bookstore',
       'Tram Station', 'Pizza Place', 'Fast Food Restaurant',
       'Movie Theater', 'Irish Pub', 'Platform', 'Train Station',
       'Turkish Restaurant', 'Department Store', 'Electronics Store',
       'Gelato Shop', 'Donut Shop', 'Burger Joint', 'Asian Restaurant',
       'Sandwich Place', 'Gym / Fitness Center', 'Stationery Store',
       'Noodle House', 'Chocolate Shop', 'Sushi Restaurant',
       'Shopping Mall', 'Auto Garage', 'Bridal Shop', 'Gastropub', 'Food',
       'Bakery', 'Mexican Restaurant', 'Restaurant', 'Hotel',
       'Furniture / Home Store'], dtype=object)
```

In a similar manner I calculated the mean frequency of occurrence of each category, in this case I calculated the total number of venue category in each post towns, doing so by grouping again by Post town the generated one hot encoding dataframe. This time though, I used the sum function instead of the mean, while also I included only the relevant venue categories I mentioned earlier.

	Post town	Auto Garage	Bakery	Café	Chocolate Shop	Clothing Store	Coffee Shop	Department Store	Donut Shop	Electronics Store	Furniture / Home Store	Gelato Shop	Grocery Store	Hotel	Movie Theater
0	BECKENHAM	0	0	4	0	1	4	0	0	0	0	0	1	0	1
1	BROMLEY	1	0	2	1	5	4	1	1	1	0	1	0	0	0
2	CHISLEHURST	0	0	1	0	0	0	0	0	0	0	0	0	0	0
3	ORPINGTON	0	1	0	0	0	0	0	0	0	0	0	1	1	1
4	SWANLEY	0	0	0	0	0	1	0	0	0	1	0	0	0	0
5	WEST WICKHAM	0	0	1	0	0	1	0	0	0	0	0	1	0	0

Next, in order to be in a position to draw some conclusions from this dataframe, I decided that due to the large number of venue types included, the best method of visualisation in this case was using a bar chart. As you can see below, I had to place the legend below the bar chart, dividing it into 8 columns, adding the “Venue type” title, while setting it into an appropriate size. This was crucial for visualising the large amount of venue types in an organised and clear way. Finally, we can see the number of venues with a positive or negative influence and get some insight into which location is the most appropriate. Bromley for instance has the largest number of clothing stores which is positive, however it’s also having the largest number of coffee shops which is negative since they are direct competitors.



4.2 Coffee shops per capita in each Post town

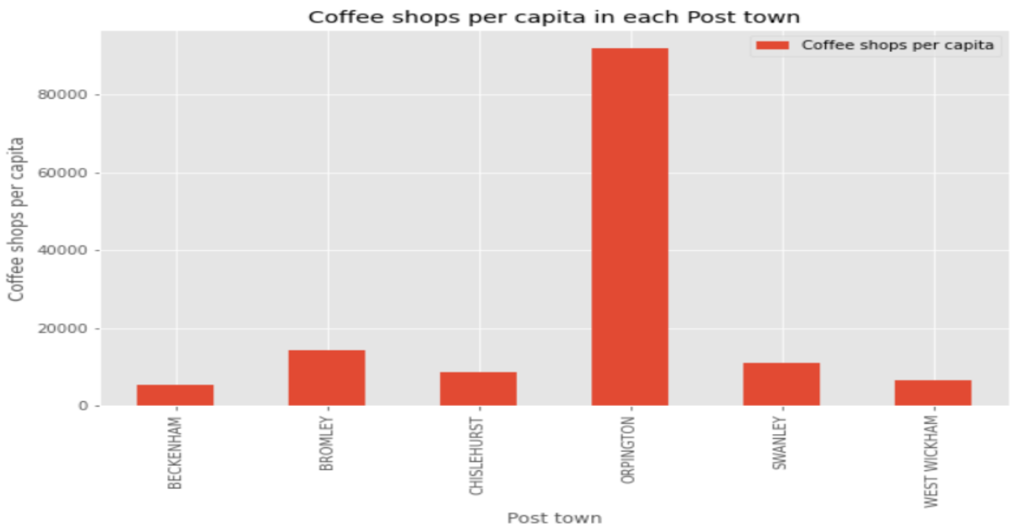
Arguably, the most impactful venue categories on profitability potential are cafes and coffee shops due to being the direct competitors. Therefore, I decided to look deeper into this and created a model which measures the total number of people who will be served by each direct competitor and a new coffee shop. I need to point out that this model is under the assumption that all of the coffee shops and cafés within the same post town will serve an equal number of people, serving only people living in their respective postcode district. Firstly, I filtered the data I extracted from Foursquare API, creating a new dataframe which includes only the venues whose Venue category is Coffee shop or Café.

	Post town	Post town Latitude	Post town Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	BECKENHAM	51.408780	-0.025260	Caffè Nero	51.407498	-0.027477	Coffee Shop
8	BECKENHAM	51.408780	-0.025260	Rendez-Vous Cafe	51.408005	-0.025303	Café
11	BECKENHAM	51.408780	-0.025260	Em and Lou's Kitchen	51.409528	-0.025227	Café
15	BECKENHAM	51.408780	-0.025260	Fee & Brown	51.409188	-0.025156	Coffee Shop
16	BECKENHAM	51.408780	-0.025260	Starbucks	51.406618	-0.028772	Coffee Shop
25	BECKENHAM	51.408780	-0.025260	Love 'A' Slice	51.406668	-0.030324	Café
34	BECKENHAM	51.408780	-0.025260	Costa Coffee	51.406864	-0.028106	Coffee Shop
36	BECKENHAM	51.408780	-0.025260	Village Bistro Cafe	51.407497	-0.027607	Café
50	BROMLEY	51.406025	0.013156	Costa Coffee	51.405653	0.015279	Coffee Shop
65	BROMLEY	51.406025	0.013156	Costa Coffee	51.404458	0.016645	Coffee Shop
68	BROMLEY	51.406025	0.013156	Patisserie Valerie	51.405164	0.016470	Café
69	BROMLEY	51.406025	0.013156	Caffè Nero	51.402653	0.015616	Coffee Shop
71	BROMLEY	51.406025	0.013156	Starbucks	51.405300	0.015538	Coffee Shop

Next, I created a dataframe with 2 columns. The first column is the post town and the second one is the Competitors column which includes the total number of cafes and coffee shops within each post town. Using an outer join, I merged this dataframe with the main dataframe which contains the total population in each post code district. Finally, I added a last column which calculates the total coffee shops per capita, doing so with the following calculation: “Population” column / (“Competitors” column+ 1). The +1 is the assumption that we open a new coffee shop in that post town, including it in the total number of coffee shops and Cafes. As shown below, the Coffee shops per capita column shows for instance that we can assume 5267.888889 people in Beckenham are served by each coffee shop and café.

	Post town	Postcode district	Latitude	Longitude	Population	Competitors	Coffee shops per capita
0	BECKENHAM	BR3	51.408780	-0.025260	47411	8	5267.888889
1	BROMLEY	BR1,BR2	51.406025	0.013156	100920	6	14417.142857
2	CHISLEHURST	BR7	51.413785	0.076756	17322	1	8661.000000
3	ORPINGTON	BR5,BR6	51.379588	0.103539	92084	0	92084.000000
4	SWANLEY	BR8	51.397170	0.173210	22053	1	11026.500000
5	WEST WICKHAM	BR4	51.376826	-0.014540	19367	2	6455.666667

In order to visualise the results in a clear manner, I used again a bar chart to display the number of coffee shops per capita within each post town. Now we can easily see for instance that Orpington has a much larger customer potential when taking into consideration only post code district population and number of direct competitors who serve the same Post town.



4. Discussion

The primary concept which influenced my work was the knowledge that Coffee shops are retailers who sell directly to consumers, with the level of accessibility, foot traffic and competition having a direct impact on profitability. Hence, using the models discussed can bring more awareness to neighbouring businesses, pointing to the locations which have the higher number of neighbouring venues with positive influence. In addition, the models point out the locations with the highest direct competition, by bringing into the equation also the total population of each post code district. The London Borough of Bromley is the largest borough of Greater London by total area, as a result each post code district has a Post town which is the centre of activity. In addition, all these Post towns are sufficiently spread, therefore they were suitable centroids for the K- means clustering algorithm which was used to explore each post code district. Nevertheless, this project has several limitations needed to be addressed. Firstly, I was using the free Foursquare API account version, thus I was only able to get a limited number of venues. An updated account is needed in order to incorporate a much larger number of venues in the dataframe and enable the generation of more reliable results. Moreover, these models don't take into consideration the costs associated with opening a new store within a particular post town. When taking into consideration that a location closer to a post town will probably be more expensive, these models may not be suitable for those without substantial initial capital. In this case, a K-means clustering algorithm which includes a larger number of areas within the Borough of Bromley is more suitable. Also, a more detailed analysis on positive impact on profitably could prove to be useful. This is especially true when it comes to assessing how accessible is each post town to the whole population of Greater London instead of being focused within each post code district. Finally, constant feedback is always necessary when these models are applied in the real world to take decisions, making sure they are up to date and continuously improved.

5. Conclusion

The objective of this project was to create models which can aid in the selection of a suitable location to open a coffee shop in The London Borough of Bromley, attempting to close the information gap. We need to think of a coffee shop within a community the same way as a living organism within an ecosystem, in order to thrive the right environment and integration within this system is needed. Investors who seek to make profit in the coffee market need to utilise dynamic up to date models which are able to navigate this fluid, fast-paced environment, especially when they need to choose a specific location. Indeed, this is especially the case in a large and developed economy like the one of Greater London, which can be a source of great opportunity, but also of challenges.

References:

[1] https://en.wikipedia.org/wiki/BR_postcode_area

[2] https://www.google.com/maps

[3] https://www.postcodearea.co.uk/postaltowns/bromley/

[4] https://developer.foursquare.com/