Battle of the Neighbourhoods

Samuel Angmor Mensah

May 27, 2019

1 Introduction

1.1 Background

A multinational coffee house company recently got the opportunity to expand their business to United Kingdom. Specifically, the city council of Bristol, UK just gave them the permission to establish a coffee shop in Bristol. However, as the UK is a large patronizer of coffee beverages, there exist quite a large number of coffee shops in several neighborhoods in this city. Nonetheless, they intend to take this rare opportunity after several efforts have been made to penetrate into the system. To make the business profitable, they hope for a neighborhood which is at least 4km away from the city center, the closer the better, and have a low coffee shop presence. Selecting from the numerous number of neighborhoods taking into account the presence of coffee shops, and the number of these coffee shops, brings about a challenge. Manually checking neighborhoods is time consuming and requires human intervention, and in some cases not possible. Therefore, it is advantageous for the multinational company to directly identify neighborhoods close to the city center, which have less presence of coffee shops within the city, to help them select the optimal neighborhood to start a coffee shop business.

1.2 Problem

Data that we might require include the neighborhoods around Bristol, as well as coffee shop or cafe venues within the neighborhoods. We center around the city center because these areas will have a large population and have work places where people will like to take a break out for coffee. This project aims to identify neighbourhoods based on the presence of coffee shops, to select the optimal neighborhood to establish a coffee shop in Bristol, United Kingdom.

1.3 Interest

Our findings will be interesting to the stakeholders of the multinational coffee house company. Others who will be interested will be other coffee house companies who have no presence in the UK.

2 Data Acquisition and cleaning

2.1 Data sources

We need information of neighborhoods in Bristol, we scrape neighbourhood information for Bristol in the link below

- Bristol : `https://en.wikipedia.org/wiki/BS_postcode_area`

We convert the addresses scraped from the link to geographical coordinates using **geopy**. Neighborhood information as well as their geographical coordinates can be used together to extract venues.

We use FourSquare `https://developer.foursquare.com/` to extract venues found in neighbourhoods, including the venue category, latitude and longitude and other information.

2.2 Data cleaning

Neighbourhood data from wikipedia pages for Bristol is scraped into dataframes using a python package called BeautifulSoup. We perform preliminary cleaning by removing empty columns in the dataframe. The dataframe have attributes such as 'Postcode district', 'Post town', 'Coverage', 'Local authority area'. Here, the 'Coverage' attribute represent the column for neighborhoods. For easy reference, we will denote 'Coverage' as 'Neighborhood', 'Postcode district' as 'PostCode' and 'Post town' as 'PostTown'.

To stay in the authority area of the city council of Bristol, we perform filtering on the dataframe. We consider only neighborhoods where the 'local authority area' is Bristol.

Next, we realize that the Neighborhood column are aggregated according to the PostCode. We unlist the neighborhoods so that each neighborhood instance is represented by a row. We used an unlistify function published on Github in a pandas issue with url `https://github.com/pandas-dev/pandas/issues/10511` to do this. This presents duplicates within the dataframe but we remove all duplicates.

To be able to get venues around a particular neighborhood, we need the geographical coordinates of the neighborhoods. The data collected so far has no geographical coordinates for the neighborhoods. However we can exploit *geopy*, a python library to convert address to geographical coordinates. We therefore create a column named 'Address' to gather neighborhood names as well as the country UK. This will help geopy to focus on converting only addresses with the neighborhood name located in only UK. We realized that this is more effective because neighborhood names, especially those in the UK exist around the world. We can now employ geopy to convert addresses to latitudes and longitudes and place into respective columns of the dataframe. At this point, the number of neighborhoods we have is 62.

Another important data that we store into memory is the coordinates for the center of Bristol. We want neighborhoods around the city center, so we convert Bristol city center to geographical coordinates and select only neighborhoods around the city center. Our criteria is that the distance between the city center and the neighborhood should be at most $4km$. In the geopy python package there is a distance function which can help to find the distance between any two geographical coordinates. We use this to filter the neighborhoods and select those within 4km from the city center. We gather distances between the city center and the neighborhood into a column in the dataframe. This helps in the analysis when considering distance as a factor to choose a neighborhood. The number of neighborhoods is reduced to 29 after filtering.

2.3 Feature selection

After data cleaning and neighborhood selection, we have 29 neighborhoods which lie within a 4km radius from the city center of Bristol. Also the number of features is 8, namely, 'PostCode', 'PostTown', 'Neighborhood', 'Local authority area', 'Address', 'Longitude', 'Latitude' and 'Distance from city center'. We denote 'Distance from city center' simply as 'Distance-from-center'.

Since we considered only neighborhoods with Bristol as local authority area, the column 'local authority area' is no more important after the filtering process. Therefore, we drop 'local authority area' attribute from the dataframe. The new features include 'Address', 'Latitude' and 'Longitude'. A summary of features dropped or added is presented in Table 1.

City	**Kept features**	**Dropped features**	**Reason for dropping features**	**New features**
Bristol	PostCode, PostTown, Neighborhood	Local authority area	Same as post town	Longitudes, Latitudes, Address, Distance-from-center

Table 1: Feature selection during data cleaning of neighborhood data.

2.4 FourSquare Data

We exploit Foursquare to extract venues for the candidate neighborhoods. We can extract venues from Foursquare by making calls to the Foursquare API. We do this by constructing a url which sends a request to the API to explore geographical locations. Some of the parameters needed

include client ID of Foursquare, client secret, version number of the API, limit of results generated, geographical coordinates of the place to explore in longitude and latitude, and radius around the geographical position.

The url is in the form:

```
url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}
  &ll={},{}&v={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET,
      latitude, longitude, VERSION, radius, LIMIT)
```

This url makes requests to the FourSquare API to explore venues for the candidate neighbourhoods. The data extracted has a lot of information but we extract the relevant ones for our task. This includes the name of venue, category, address, longitude and latitude, distance to center of neighborhood, neighborhood name and postal code. We are interested in venues in the Coffee Shop category, so we extract those venues for further analysis.

3 Methodology

3.1 Calculation of distance-from-center feature

The 'distance-from-center' was not a feature from the original neighborhood data. However, this feature is important because the goal of our task is to find the closest neighborhood to the center. We therefore calculate the distance between each neighborhood data and the city center. The geopy library already introduced have a distance function that can help solve this task. The distance function outputs a distance measure in either kilometers or miles.

Formally, for any neighborhood coordinate (x_1, y_1) and city center (c_1, c_2). The distance in kilometers can be computed as

$$distance = \text{geopy.distance.distance}((x_1, x_2), (c_1, c_2)).\text{km}$$

We use this to select neighborhoods which have a *distance* less than $4km$ to satisfy the conditions as discussed with stakeholders.

3.2 Exploratory Analysis

We analyse both neighborhood data and foursquare venue data. We show visualizations to see what the data tell us. After applying the distance function from geopy and filtering we have 29 neighborhoods. After exploring venues around these neighborhoods using Foursquare we have a total of 2143 venues. However after further filtering out only those with category Coffee Shop we have a total of 145 venues. The map in Figure 1 shows the neighborhoods for Bristol along with Coffee Shop venues.

From the map in Figure 1 we realize that coffee shops are located close to the city center and spreads generally from the north to the south.

3.2.1 Descriptive Statistics

Figure 1 presents a general idea of locations of coffee shops in Bristol. But is difficult to tell which coffee shop belongs to which neighborhood, or the number of coffee shops found in a neighborhood. Pandas have a function to aggregate coffee shop venues to each neighborhood. We can use a bar graph as shown in Figure 2 to compare the proportion of coffee shops in each neighborhood.

Figure 2 shows that Henleaze have the highest number of coffee shops. We also observe that 14 neighborhoods satisfy the conditions of potential neighborhoods. Thus, neighborhoods with at most 5 coffee shops within a 2000m radius from the center of the neighborhood. We summarize the statistics of the dataset in Table 2

From Table 2 the minimum of coffee shops in a neighborhood is 0, supporting the bar graph. We also realize that the neighborhoods within 4km of the city center have an average of 5 coffee shops. This suggests that finding a neighborhood with at most 5 coffee shops will be ideal for a coffee shop business. Beyond this, the neighborhood can be considered as having a high density of coffee shops. The table also give other statistics such as the quartiles (25%,50%,75%), and the standard deviation (std). The table provides a fine detail of what is represented in the bar graph.

The bar graph however lacks the ability to visualize the position of the neighborhoods. Our goal is to find neighborhoods close to the city center so a visualization on a map will be helpful. A bubble map gives a better visualization.

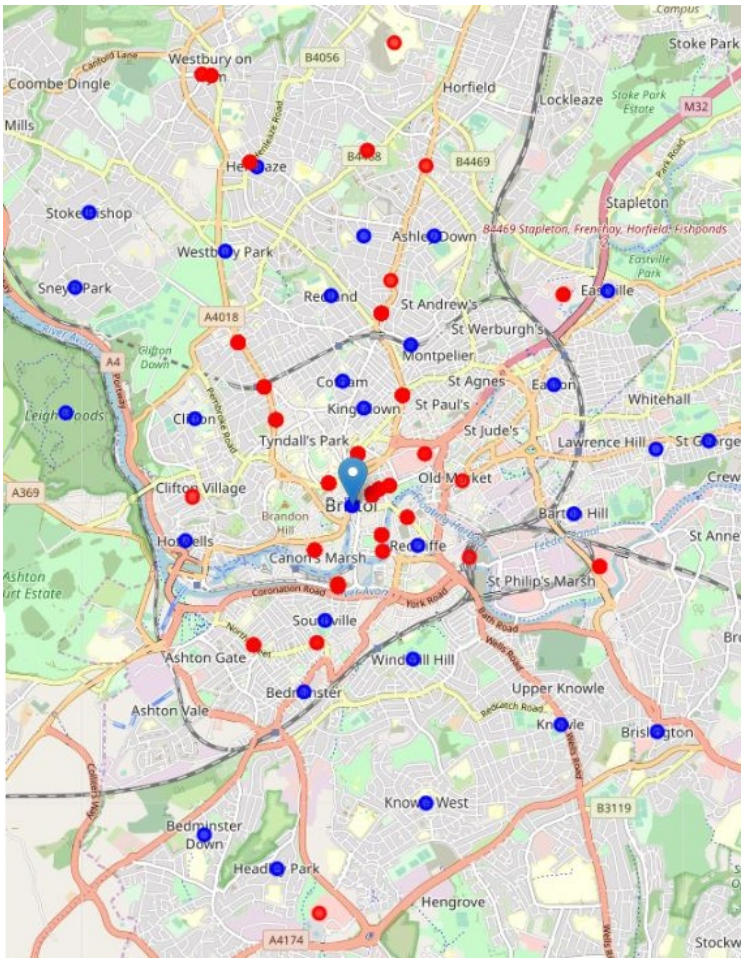

Figure 1: Map of Bristol with neighborhoods(blue) within 4km from the city center, and venues(red) imposed on the map

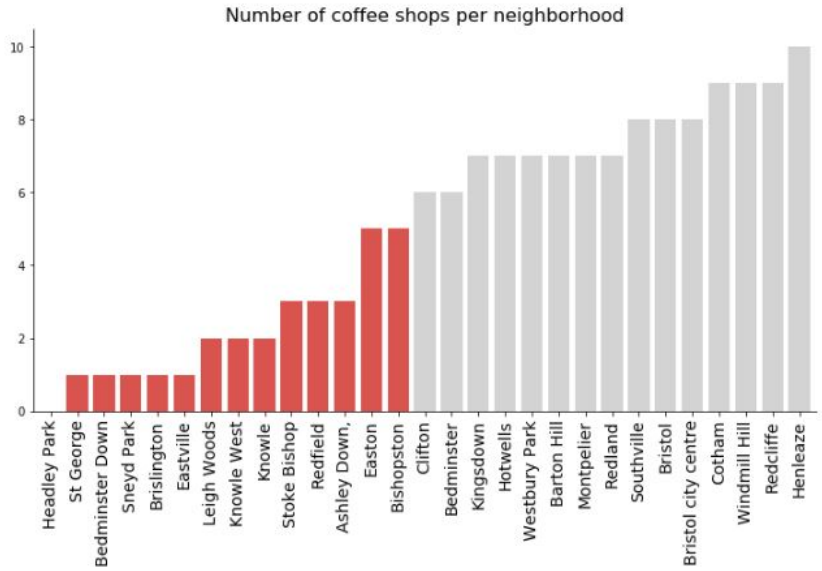

Figure 2: Number of coffee shops per the 29 neighborhoods. Attention is on neighborhoods with less than 5 coffee shops (highlighted by the red bars)

	Venues
count	29
mean	5
std	3.105295
min	0
25%	2
50%	6
75%	7
max	10

Table 2: Statistics of coffee shop venues for the 29 neighborhoods

3.2.2 Bubble Map

We present the bubble map to properly visualize neighborhoods with high or low density of coffee shops. The bubble map scales up the marker representing the neighborhood if the neighborhood has high density. This procedure makes visualization much easier compared to the bar graph. Figure 3 shows a bubble map of neighborhoods, where the radius of the marker is linear to the number of venues in the neighborhood.

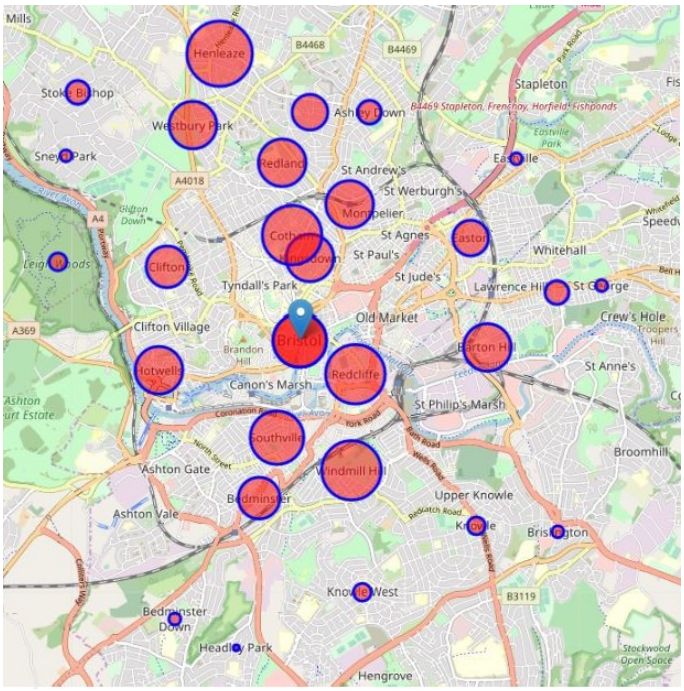

Figure 3: Bubble map show the density of coffee shops in each neighborhood of Bristol

The map shows a high density of coffee shops around the center of Bristol. Well it is expected that most businesses or coffee shops should be around this center because lots of people will be within these neighborhoods, especially tourists, and workers, including travellers who will pass through the busy city center to catch a bus or a train. We also see that there is high density up North and down south. Generally, the density of coffee shops descreases as we move away from the center of Bristol.

Possible neighborhoods we can consider are Sneyd Park, Leigh Woods, Eastville, Bedminster Down, Knowle West, Knowle, Birislington, St. George, East Ville, Red Field, Stoke Bishop, Ashley Down. These neighborhoods have low density but they are further from the city center but within 4km from city center. Leigh Wood is mostly woodland and not ideal for business. Bedminster Down, Sneyd Park and Headley Park is mostly residential and therefore not ideal for business.

Based on this observation and the saturation of coffee shops around the city center, we take the search to neighborhoods which have less than 5 coffee shops to meet the requirements by stakeholders. We exclude Leigh Wood and areas which are mostly residential. Note that our initial search radius was 2000 meters, so this number of coffee shops within the radius is sparse for a coffee shop to thrive. We therefore focus our attention to **Eastville, Brislington, St George,**

Neigborhood	Cluster
Brislington	1
Knowle	1
Knowle West	1
Easton	2
St. George	0
Redfield	0
Eastville	0
Bishopston	2
Ashley Down	2
Stoke Bishop	0

Table 3: K-means clustering of neighborhoods based on all venue features

Knowle West, Knowle, Redfield, Ashley Down, Stoke Bishop, Easton, Bishopston.

We can cluster the neighborhoods selected using k-means clustering to know common characteristics of neighborhoods for batch filtering. This will help narrow it down.

3.3 K-means Clustering

We perform clustering analysis on the neighborhoods. In this analysis we associate each neighborhood with all venues within the 2000meters radius of the neighborhood. We include all venues in this analysis because we aim to find out the presence of coffee shops with respect to other venues. The clustering method we use is k-means.

K-means clustering aims to partition the neighborhoods into k of clusters. Clusters with common features (venues) are grouped in the same cluster. This will make us understand more about the neighborhoods of Bristol.

K-means takes as input a feature matrix, where each row represents a neighborhood and the features represent venues. This implies we have to process the venue data extracted from Foursquare to make it applicable for k-means. To this end, we transform the ''venue' category column into one-hot vectors for each neighborhood. Next, we find the mean of each venue category for each neighborhood. At this point we have the feature matrix which can be used for the k-means function in the python library sklearn.

Finally, we can create a k-means object. Since the number of neighborhood have reduced, we choose k as 3. We fit the feature matrix in the k-means object. This labels the neighborhoods based on the features, where neighborhoods with similar features are clustered together. The resulting clustering is seen in the Figure 4 . Interestingly, we find that neighborhoods which are close together have similar features. We also present the cluster of neighborhoods in a Table 3 for easy reference.

3.3.1 Examine the clusters by the most common venues

We note that Cafe is comparable with Coffee Shops. Reason being that cafes normally serve coffee along with other meals. Cafes can be a direct competitor with coffee shops because they might offer more options. If they serve good coffee, one is likely to go there than to go to a coffee shop. Based on this reason we assume cafes as direct competitors and avoid being in the same neighborhood with them if possible.

We examine the 1st most common venue across the cluster of neighborhoods. We find that there is no sign of coffee shops or cafes across the clusters for the 1st most common venue. This is a good sign. This suggests that these neighborhoods are promising for the establishment of a coffee shop. On the other hand, when we analyzed the 2nd most common venue across the clusters, Cluster 2 has an occurrence of Cafe for each of the neighborhoods. This implies that neighborhoods in this cluster have a number of cafes bringing about competition. To this end, we exclude all neighborhoods in Cluster 2.

St. George, Redfield also have Cafes as the 3rd most common venue, with Eastville having a Cafe in the 4th most common. Even though Eastville and St. Geaorge have just 1 occurrence of a coffee shop in each of the neighborhood. They all belong to Cluster 0. Comparing this to Cluster 1, we observe that the coffee shop/cafe do not occur till the 8th most common venue for the neighborhoods in Cluster 1. Based on this reason we exclude neighborhoods in Cluster 0.

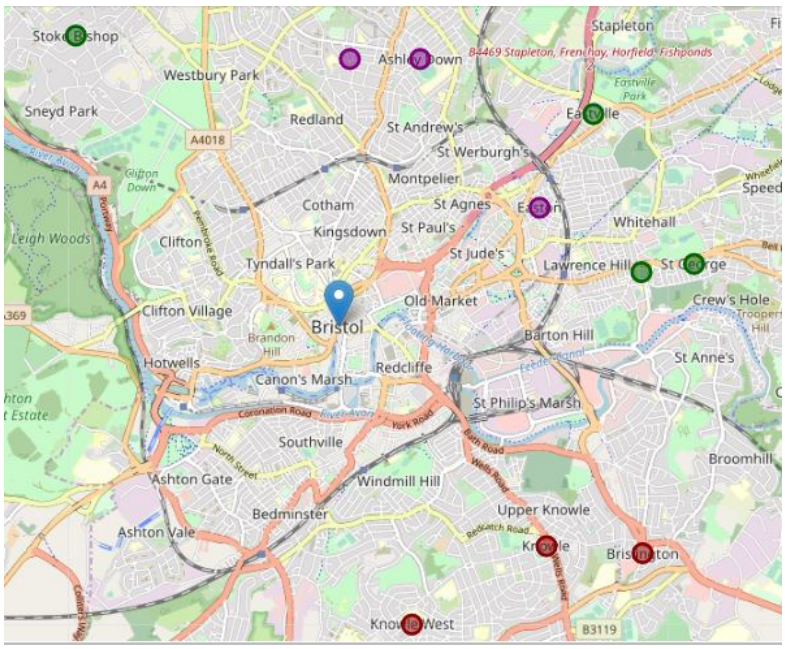

Figure 4: k-means clustering on neighborhoods where $k = 3$

The promising cluster of neighborhoods is cluster 1. Brislington, Knowle and Knowle West are competitive neighborhoods to establish a coffee shop. Reason being that Brislington has no coffee shop but a cafe which is synanymous to a coffee shop. However, Brislington is further away from the city center than Knowle and Knowle West. Knowle West on the other hand is known to be a deprive area, whereby 6 areas out of 8 areas in Knowle West is ranked deprived as noted in *Filwood Ward Profile 2008, p. 9*. As such the best neighborhood to establish a coffee shop is Knowle.

Knowle is about 3km away from the city center and has only 2 coffee shop within a radius of 2000m, and the 8th most common venue from the center of Knowle. More importantly the neighborhood of Knowles is just next to the A37 motorway, creating an excellent opportunity for travellers to stop by and grab a cup of coffee for the road. Knowle has longitude -2.566746 and latitude 51.433783. We show the location of Knowle with respect to the city center in the map in Figure 5.

We conclude the analysis. We have suggested the best neighborhood among the neighborhoods in Bristol to establish a coffee shop. All neighborhoods analyzed were within a 4km of the city center of Bristol. We observed that most of the neighborhoods around the city center are already saturated with coffee shops. So we moved our search to the outskirt of the radius. In those areas, either is residential or woodland. Further analysis on the data including external sources hinted us that the best neighborhood is Knowles.

4 Results

The analysis shows that although Bristol has a lot of coffee shops (there is an average of 5 restaurants within a 2000 meters radius of each neighborhood), we find that a couple of neighborhoods not very far from the city center are less densed with coffee shops or cafes. The highest concentration of coffee shops spread from the north down to the south, while we observed that the density of coffee shops reduced as we move further away from the city center. We found out that some neighborhoods with less coffee shop density were either residential areas or forest areas. Excluding those areas in our search, we centered our attention to those which had less than 5 coffee shops within a 4km radius from city center, as that was the basic requirement as discussed with stakeholders. That left us with Easton, Bishopston, Ashley Down, Redfield, Knowle West, Eastville, Knowle, Brislington, St. George and Stoke Bishop.

Directing our attention to these neighborhoods, we perform a k-means clustering to group the neighborhoods based on venues within the neighborhood. This helped us tell the importance of coffee shops in the neighborhood by looking at the most common venue. Noting that cafes are direct competitors with coffee shops, we also examine where cafes are ranked as the most common venues. We wish to avoid neighborhoods with cafes as well to avoid comptetion. This narrowed

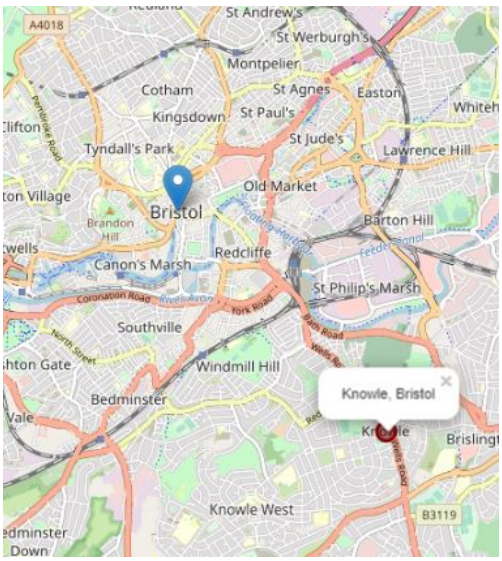

Figure 5: Location of recommended neighborhood (Knowle) relative to Bristol city center

our search to the cluster which contains the neighborhoods Knowle, Knowle West and Brislington.

Using other external information for Knowle, Knowle West and Brislington, we optimally chose the right neighborhood. Specifically, a quick search on Knowle West indicated that the neighborhood is deprived and not econmically stable. Hence a coffee shop business will not survive in such a neighborhood. Also eventhough Brislington have no coffee shops when observing up to the 10th most common venue, they have cafe ranked as the 8th most common venue, which is comparable to a coffee shop. Knowle has a coffee shop as the 8th most common venue. However, Knowle is closer to Brislington and it stretches along the motorway, which will be ideal for travelers to make a stopover to grab a coffee.

5 Discussion

During our analysis, we realized that we had to extend coffee shops to cafes because the presence of cafes also will affect the business of coffee shops. Also care must be taken when filtering the neighborhoods because other unseen parameters could affect business. For this reason, we also look at the environment, whether it is residential, a forest/wood area or a park. We also considered some other external information such as the economic status of the neighborhoods. Even though we suggest Knowle as the optimal neighborhood to start a business, it does not imply other neighborhoods cannot do better. One can also look at other locations within a given neighborhood. Factors that might also be interesting will include proximity to a car park and other social amenities interesting to customers of the coffee shop. The result presented was based on the parameters used and the requirement of the stakeholders.

6 Conclusion

The aim of this project was to identify the optimal neighborhood close to the Bristol city center which has very few coffee shops, with the goal to advise stakeholders on the best neighborhood to establish a coffee shop. We perform extensive analysis using venue data provided by Foursquare as well as neighborhood data scraped from Wikipedia. By calculating certain statics such as the density distribution of coffee shops in neighborhoods, as well as the distance from neighborhoods to the city center we could narrow our search. Further analysis such as clustering using k-means on all venue features of filtered neighborhoods gave us insight on the presence or importance of coffee shops within each neighborhood. Taking into account additional factors like the economic status of the neighborhood and the distance to city, we can strongly advise that Knowles is the best neighborhood to establish a coffee house.