THE BATTLE OF THE NEIGHBOURHOODS

Applied Data Science Capstone Project

Fahamida Haque

Introduction

Coffee is one of the most widely consumed beverages in the world. Its comforting taste and caffein boost are what people have come to rely on. This is especially true in Canada, as Canadians are said to be on of the world's largest coffee consumers. A survey found that almost three quarters of Canadians have consumed coffee in the last day. Like many Canadians, our client is a coffee lover and is looking to open their own gourmet coffee shop. Our client is looking to open his coffee shop in Toronto, as it is Canada's most populous city. Toronto is also an international center of business, culture, and the arts and is one of the most multicultural cities in the world. The ideal location for this new business would be an area with minimal competition and close to the city center.

The goal of this project is to find the ideal location to open a coffee shop in Toronto. We will do this by obtaining a list of neighbourhoods in the greater Toronto region. We will look at venues in each neighbourhood and use data science to find the concentration of coffee shops in each neighbourhood. Lastly, we will use the results from our analysis to recommend the best area to open a new coffee shop.

Data

In this project we will use data from a few sources. We will use a list of Toronto postal codes and neighbourhoods from Wikipedia. We will find the geographical latitudes and longitudes of each neighbourhood using the geocoder package. Finally, the Foursquare API will be used to search for the venues in each neighbourhood.

Methodology

Determining the best location for our coffee shop requires obtaining and cleaning the data. Then we will explore the data and employ machine learning to determine the best neighbourhoods. We will then analyze the data before making our final recommendations.

Data Cleansing

The list of postal codes for the city of Toronto is obtained from Wikipedia. Figure 1 shows the table that is found on Wikipedia. There we see a table where each row contains a postal code, borough name and the neighbourhood's name. To use this data, the table of postal codes is extracted using the Beautiful Soup package and then save the data into a Pandas data frame.

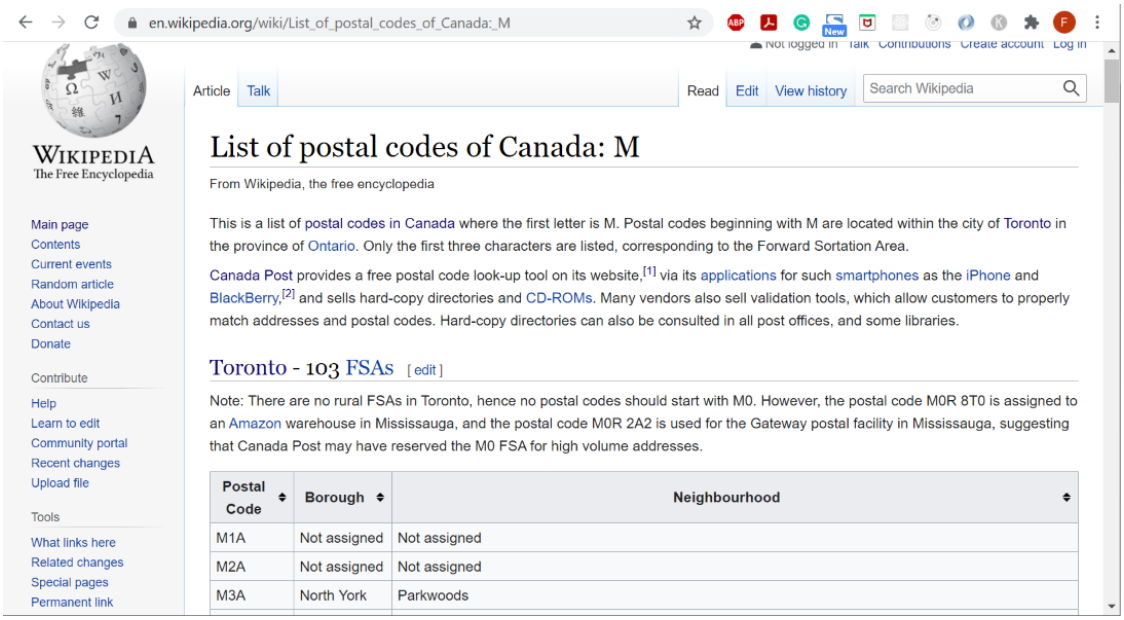

Figure 1: Wikipedia page containing the list of Toronto neighbourhoods with their postal codes

The data cleaning process will look at each row of the table and, discard the rows of data where the borough is not assigned. If the borough is assigned but the neighbourhood is not assigned, the value (the name) of the neighbourhood will be the same as the borough. This results in the final data frame that will be used to explore the neighbourhoods. This can be seen in the figure below.

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 2: The first five rows of cleaned data frame

Data explorations

To begin the data explorations, the longitude and latitude columns are added to the data frame. Then using the postal codes, and the geocoder API, the data frame is updated to include the geographical latitudes and longitudes of each neighbourhood. This can be seen in in Figure 3.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.75245	-79.32991
1	M4A	North York	Victoria Village	43.73057	-79.31306
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.65512	-79.36264
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.72327	-79.45042
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.66253	-79.39188

Figure 3: The first five rows of the updated data frame with the geographical coordinates

In addition to the neighbourhood coordinates we will also obtain those of the city of Toronto. This will help us to visualize the neighbourhoods on the map. Using the folium library, each neighbourhood is mapped onto the map of the city of Toronto. This can be seen in Figure 4. where the location of the neighbourhoods is marked by the blue circle.

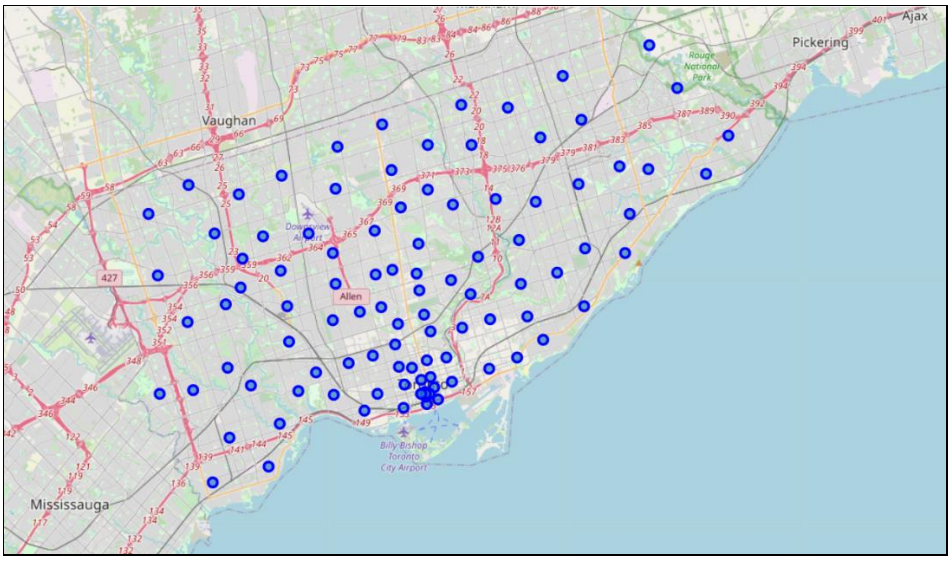

Figure 4: All the neighbourhoods in Toronto that will be used in our analysis

To explore the neighbourhoods, we will use the Foursquare API. Next comes exploring the neighbourhoods. Using the Foursquare API, we will search for 500 venues in each neighbourhood within a 1 km radius from the center of the neighbourhoods. I believe 500 venues per neighbourhood is a good sample size as it will give us a good representation of the businesses that make up the neighbourhood. As many businesses tend to be in the city or neighbourhood center, a 1 km radius will cover a large area.

Machine Learning

Once we have obtained a list of all the venues, we will find all the unique venue categories and find the average mean of each category in every neighbourhood.

Then, we will find the average mean of coffee shops for each neighbourhood. To determine the areas that have the most coffee shops, we will use the k-means clustering method to cluster the neighbourhoods based on their average mean.

Using the elbow point method, we will determine the most appropriate number of clusters. We found that k = 5 clusters would be the most appropriate and this can be seen in figure 5. Finally, we will run the k-means clustering and generate a clustering label for each of the neighbourhood.

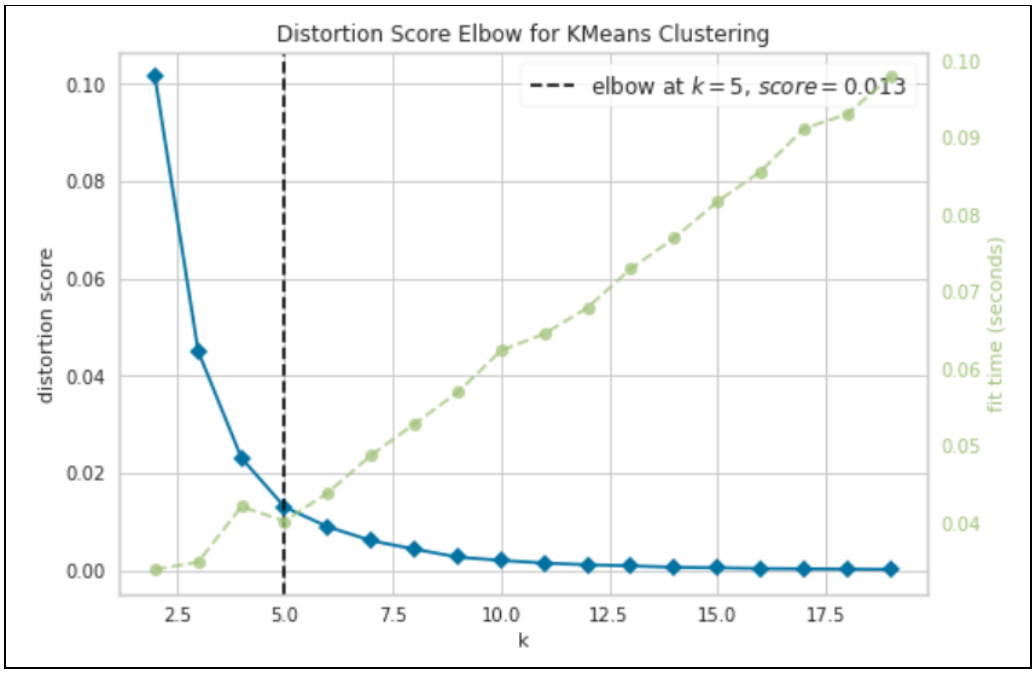

Figure 5: Finding the number of clusters, k using the Elbow Point Method

Data Analysis

To analyze the data clusters, we will plot all the cluster labels from each neighbourhood onto the map of Toronto. This can be seen in figure 6. We will then plot the number of neighbourhoods per cluster and the average number of coffee shops in each cluster. By analysing these two plots we can find clusters with a large density of coffee shops and those that have a low density of coffee shops. Once we have chosen a cluster with a low density of coffee shops, we can then look at all the neighbourhoods present in that cluster and find the neighbourhood that is closest to the city's downtown.

Results

Once, the analysis was complete, we found that there are 386 coffee shops in the city of Toronto. We also found that there are five clusters. Using Folium, we mapped all the neighbourhoods and the cluster they belong to onto the map of Toronto. The figure below depicts this map with all the neighbourhoods and their respective cluster. Neighbourhoods belonging to the first cluster (k = 0) is represented by the red points. Those belonging to the second cluster (k = 1) are represented by the green dots. The third cluster (k = 2) is represented by the blue points. The fourth cluster (k = 3) is represented by the purple points. Lastly, the fifth cluster (k = 4) is represented by the orange points. Figures 7 - 11 represent snippets from the cluster data.

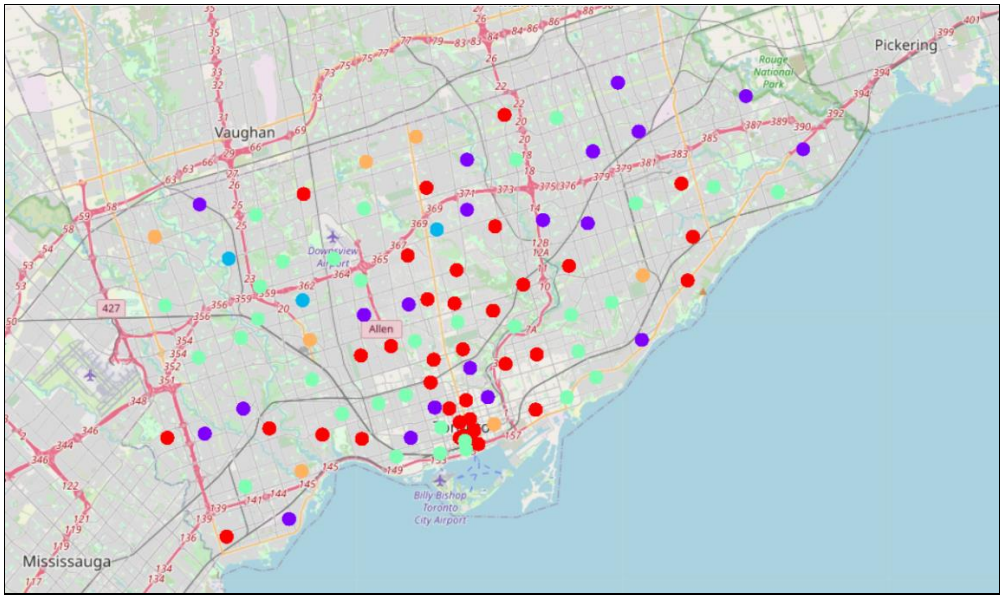

Figure 6: Clusters from the k-means clustering

	Borough	Neighbourhood	Coffee Shop	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Victoria Village	0.1	0	43.73057	-79.31306	Bamboo Garden	43.735968	-79.306236	Chinese Restaurant
1	North York	Victoria Village	0.1	0	43.73057	-79.31306	Pizza Nova	43.725824	-79.312860	Pizza Place
2	North York	Victoria Village	0.1	0	43.73057	-79.31306	The Frig	43.727051	-79.317418	French Restaurant
3	North York	Victoria Village	0.1	0	43.73057	-79.31306	Eglinton Ave E & Sloane Ave/Bermondsey Rd	43.726086	-79.313620	Intersection
4	North York	Victoria Village	0.1	0	43.73057	-79.31306	Wigmore Park	43.731023	-79.310771	Park

Figure 7: Cluster 1 (k = 0)

	Borough	Neighbourhood	Coffee Shop	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Parkwoods	0.0	1	43.75245	-79.32991	Brookbanks Park	43.751976	-79.332140	Park
1	North York	Parkwoods	0.0	1	43.75245	-79.32991	Tim Hortons	43.760668	-79.326368	Café
2	North York	Parkwoods	0.0	1	43.75245	-79.32991	Bruno's valu-mart	43.746143	-79.324630	Grocery Store
3	North York	Parkwoods	0.0	1	43.75245	-79.32991	A&W	43.760643	-79.326865	Fast Food Restaurant
4	North York	Parkwoods	0.0	1	43.75245	-79.32991	Shoppers Drug Mart	43.745315	-79.325800	Pharmacy

Figure 8: Cluster 2 (k = 1)

	Borough	Neighbourhood	Coffee Shop	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	North Park, Maple Leaf Park, Upwood Park	0.25	2	43.71381	-79.48874	Amesbury Sports Complex Arena & Outdoor Pool	43.705935	-79.486880	Athletics & Sports
1	North York	North Park, Maple Leaf Park, Upwood Park	0.25	2	43.71381	-79.48874	Lucky Dragon Chinese Food	43.719463	-79.480755	Chinese Restaurant
2	North York	North Park, Maple Leaf Park, Upwood Park	0.25	2	43.71381	-79.48874	Coffee Time	43.705864	-79.488573	Coffee Shop
3	North York	North Park, Maple Leaf Park, Upwood Park	0.25	2	43.71381	-79.48874	Rustic Bakery	43.715414	-79.490300	Bakery
4	North York	North Park, Maple Leaf Park, Upwood Park	0.25	2	43.71381	-79.48874	Petro-Canada	43.705891	-79.487841	Gas Station

Figure 9: Cluster 3 (k = 2)

	Borough	Neighbourhood	Coffee Shop	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Lawrence Manor, Lawrence Heights	0.041237	3	43.72327	-79.45042	Marche Istanbul	43.719129	-79.457631	Grocery Store
1	North York	Lawrence Manor, Lawrence Heights	0.041237	3	43.72327	-79.45042	Swiss Chalet	43.717224	-79.456223	Restaurant
2	North York	Lawrence Manor, Lawrence Heights	0.041237	3	43.72327	-79.45042	Red Lobster	43.718952	-79.456245	Seafood Restaurant
3	North York	Lawrence Manor, Lawrence Heights	0.041237	3	43.72327	-79.45042	Dollarama	43.720728	-79.457467	Discount Store
4	North York	Lawrence Manor, Lawrence Heights	0.041237	3	43.72327	-79.45042	Playtime Bowl	43.717427	-79.458148	Bowling Alley

Figure 10: Cluster 4 (k = 3)

	Borough	Neighbourhood	Coffee Shop	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Downtown Toronto	Regent Park, Harbourfront	0.15	4	43.65512	-79.36264	The Grand Hotel & Suites Toronto	43.656449	-79.374110	Hotel
1	Downtown Toronto	Regent Park, Harbourfront	0.15	4	43.65512	-79.36264	The Poet Cafe	43.650637	-79.371276	Café
2	Downtown Toronto	Regent Park, Harbourfront	0.15	4	43.65512	-79.36264	John Fluevog Shoes	43.649896	-79.359436	Shoe Store
3	Downtown Toronto	Regent Park, Harbourfront	0.15	4	43.65512	-79.36264	Izumi	43.649970	-79.360153	Asian Restaurant
4	Downtown Toronto	Regent Park, Harbourfront	0.15	4	43.65512	-79.36264	Bulk Barn	43.649994	-79.370099	Grocery Store

Figure 11: Cluster 5 (k = 4)

To further help us in our analysis, we plotted the number of neighbourhoods in each cluster as well as the average number of coffee shops in each cluster. Figure 12 and Figure 13 show that the first cluster (k = 0) has 38 neighbourhoods with an average of approximately 0.092 coffee shops. The second cluster (k = 1) has 20 neighbourhoods with an average coffee shop density of about 0.018. The third cluster (k = 2) has three neighbourhoods with an average coffee shop density of about 0.222. The fourth cluster (k = 3) has 30 neighbourhoods with an average coffee shop density of about 0.059. Lastly, the fifth cluster (k = 5) has 7 neighbourhoods with an average coffee shop density of about 0.136.

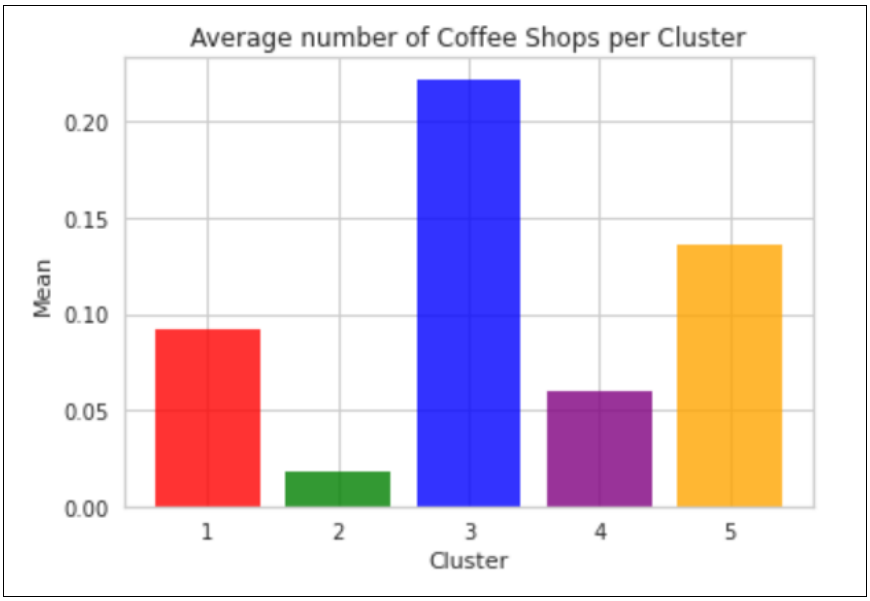

Figure 12: Plot showing the average number of coffee shops in each cluster

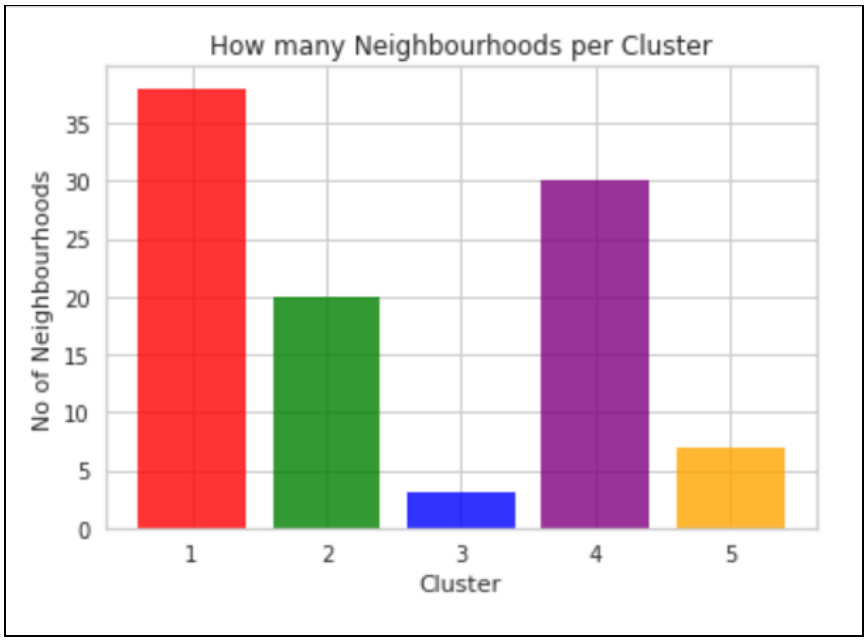

Figure 13: Plot showing the number of neighbourhoods in each cluster.

Thus, the order of cluster in terms of density of coffee shops follow is as follows:

1. Cluster 2 (0.018)
2. Cluster 4 (0.059)
3. Cluster 1 (0.092)
4. Cluster 5 (0.136)
5. Cluster 3 (0.222)

Discussion

From our analysis, we can see that clusters 1, 2 and 4 have the lowest density of coffee shops with the greatest number of neighbourhoods. Clusters 3 and 5 have the highest density of coffee shops with the least number of neighbourhoods. This leads us to believe that the ideal location for our coffee shop will be either in cluster 1, 2 or 4 since a low density of coffee shops indicates that there will be less competition with similar businesses. Since all three clusters contain neighbourhoods in Downtown Toronto, there are quite a few neighbourhoods that could be the ideal location. From cluster 1, we found that locations near Ryerson and central Bay street could have good potential. Similarly, great locations in cluster 2 include Adelaide and King. The locations with great potential in cluster 4 include St James and Rosedale.

There are some drawbacks to our analysis. The main drawback is that all the data used in our analysis comes from the Foursquare API. As we used only one database, our recommendations are based on this data. Using more databases would have resulted in a more complete recommendation. Another drawback comes from the fact that with the data cleaning we have eliminated some neighbourhoods, thus they were excluded from our analysis. Analyzing more neighbourhoods would result in a better recommendation of our final location.

Conclusion

The purpose of this project had been to determine the ideal location to open a new coffee shop in Toronto. According to the client, the ideal location was to be one that was close to the downtown and had minimal competition from similar businesses. To determine the ideal location, we used the postal codes of the city to determine the neighbourhoods and used the Foursquare API to search for venues in each neighbourhood. From this data we determined the average density of coffee shops. Using the k-means clustering we found 5 clusters. By analysing the data, we found that the downtown neighbourhoods in clusters 1, 2 and 4 are the best locations to open the coffee shop.