WHERE TO SET UP A COFFEE SHOP IN TORONTO

Dung, Nguyen Trung

February 2021

1 Introduction

Coffee is a brewed drink prepared from roasted coffee beans. Coffee is not only a drink, but also a way to relax, to meet people or to work in a friendly manner. People can drink coffee at anytime of the day, especially in the morning when they start a new day, coffee brings a fresh energy.

In Canada, coffee is consumed by adults (ages 18-79) more than any other beverage, even tap water. A coffee study reported that among the number of cups consumed each day by coffee drinkers, men consume an average of 3 cups of coffee, while women drink just 2.4 cups. According to the Coffee Association of Canada, 2/3 of Canadians enjoy at least one cup a day[1].

Toronto is the most crowed city and it's the capital city of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America[2]. With that population, coffee consumption rate is approximately 1.8 million cups of coffee everyday in Toronto. Such high consumption rate indicates a good chance for success for coffee industry, in this case opening a new Coffee Shop.

As many other Food and Beverage shops, the location of a Coffee Shop is an important factor to the business success. The objective of this capstone project is to analyze and select the best locations in the city of Toronto to open a new Coffee Shop. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Toronto, if an investor is looking to open a new Coffee Shop, where it could be located?

This project is useful to investors looking to open a new Coffee Shop in Toronto. This will help the investor to stay away from intensive competition area to ensure a business success.

This project is timely as the market is still very large and there is still a room to open the Coffee Shop in Toronto. A delay for this business, especially when the market is saturated could lead to a failure.

2 Input data

The input data is collected from various sources.

- The population of Toronto: Wikipedia.
- Latitude and Longitude of neighborhoods of Toronto: Coursera. This will be used to plot the map of Toronto and their neighborhoods.
- Data of venues related to Coffee Shop or Café: Four Squares. This data will be used to find the related data to the given problem.

3 Methodology

The data of neighborhoods of Toronto is given on Wikipedia[3]. The data is extracted using web scrapping technique with Beautiful Soup library. As the collected data contains three columns named "postal code", "neighborhood" and "borough". All postal codes with no information in the neighborhood column are dropped. The borough with no information will be assigned as the same name as neighborhood. The next step is to collect only the information related to Toronto by applying a filter in data frame.

[1] https://coffeebi.com/

[2] https://en.wikipedia.org/wiki/Toronto

[3] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

As the information of latitude and longitude from geocoder is unreliable. The information of neighborhood of Toronto provided in Coursera is used instead. The given data is merged with the data of neighborhood collected in the first step above. The merge is executed using the postal code given in both data frame. The result is a data frame containing important information including the name of the neighborhood, their latitude and longitude.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. In order to access to the data of the venues, the Foursquare Developer Account and secret key is used. The main code of Jupyter Notebook will call an API to Foursquare passing in the coordinates of the neighborhoods. Data received from Foursquare is under the format of JSON, the data is extracted with key information of venue name, venue category, venue latitude and longitude. The venues are then grouped by neighborhood and taking the mean of the frequency of occurrence of each venue category. As we are only interested in Coffee, the two key words "Coffee Shop" and "Café" are used as filtering constraints applying to data frame. The frequency of occurrence of these two columns is combined in one column named "Coffee Shop".

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for "Coffee Shop". This classification with a set of existing Coffee Shop on the map allows investors to select where the Coffee Shop should locate.

4 Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Coffee Shop":

Cluster 0: Neighborhoods with low number of Coffee Shop. The mean frequency of occurrence is 0.035.

Cluster 1: Neighborhoods with moderate number of Coffee Shop. The mean frequency of occurrence is 0.142.

Cluster 2: Neighborhoods with high concentration of Coffee Shop. The mean frequency of occurrence is 0.241.

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.

Figure 1 3 Clusters of Coffee Shop in the city of Toronto (Red: cluster 0; Blue: cluster 1; Mint green: cluster 2)

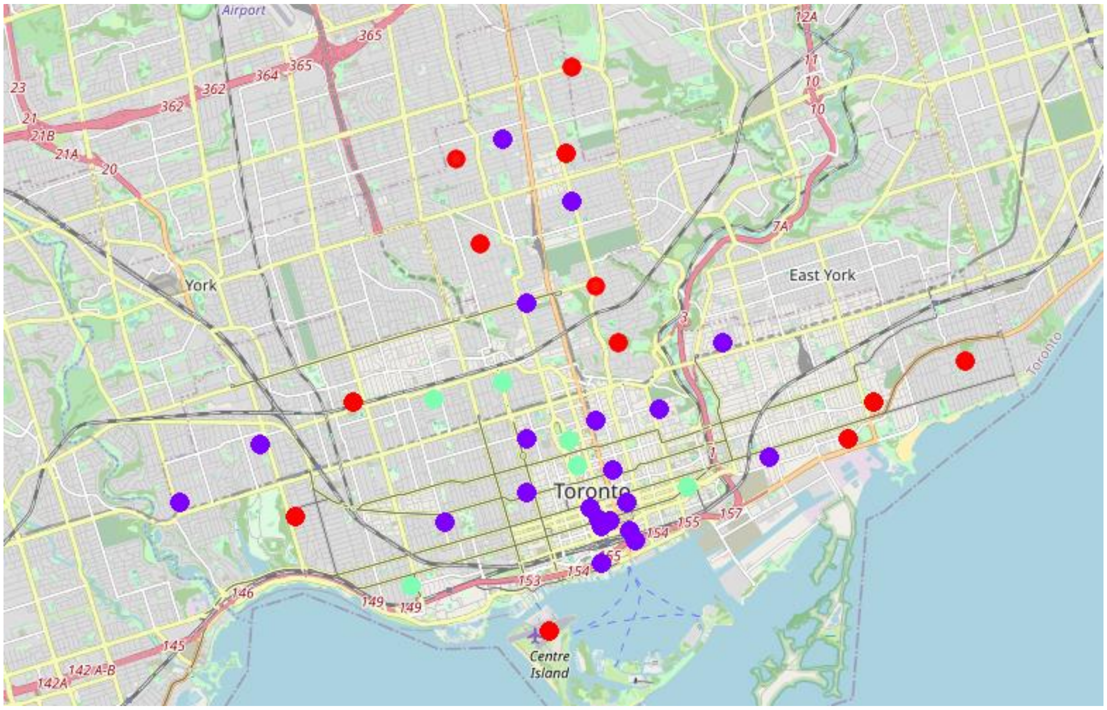


Figure 2 Cluster 0- Low frequency of occurrence

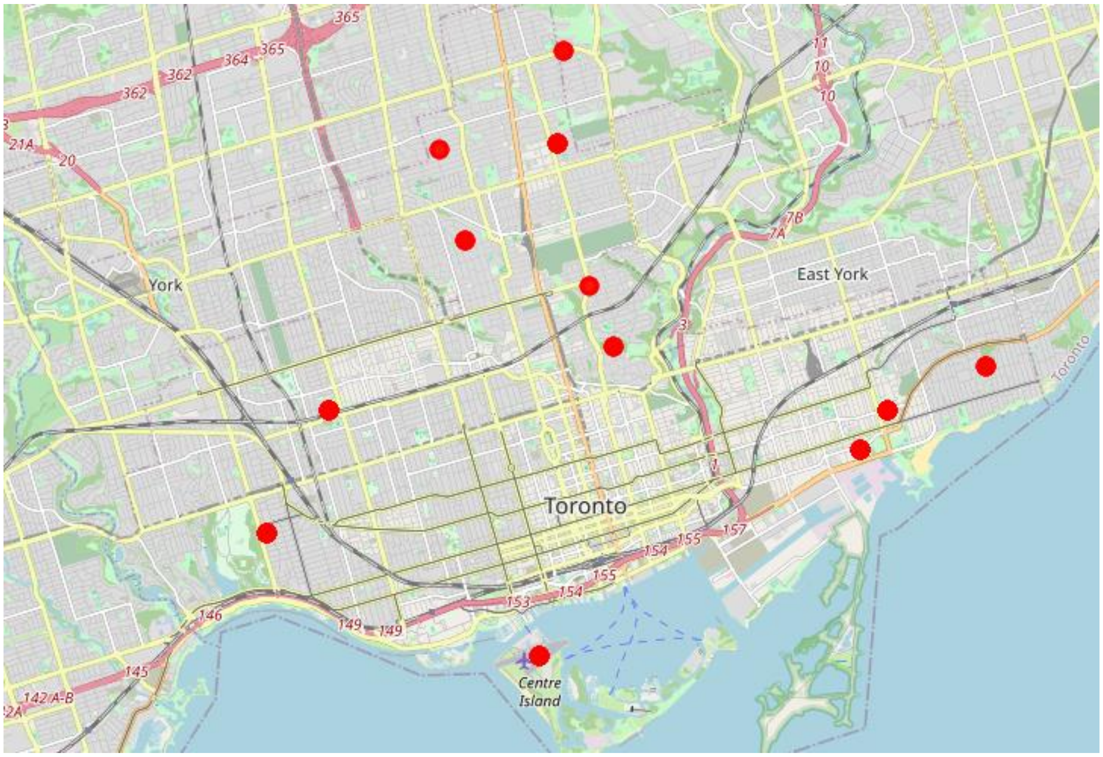


Figure 3 Cluster 1 Moderate frequency of occurrence

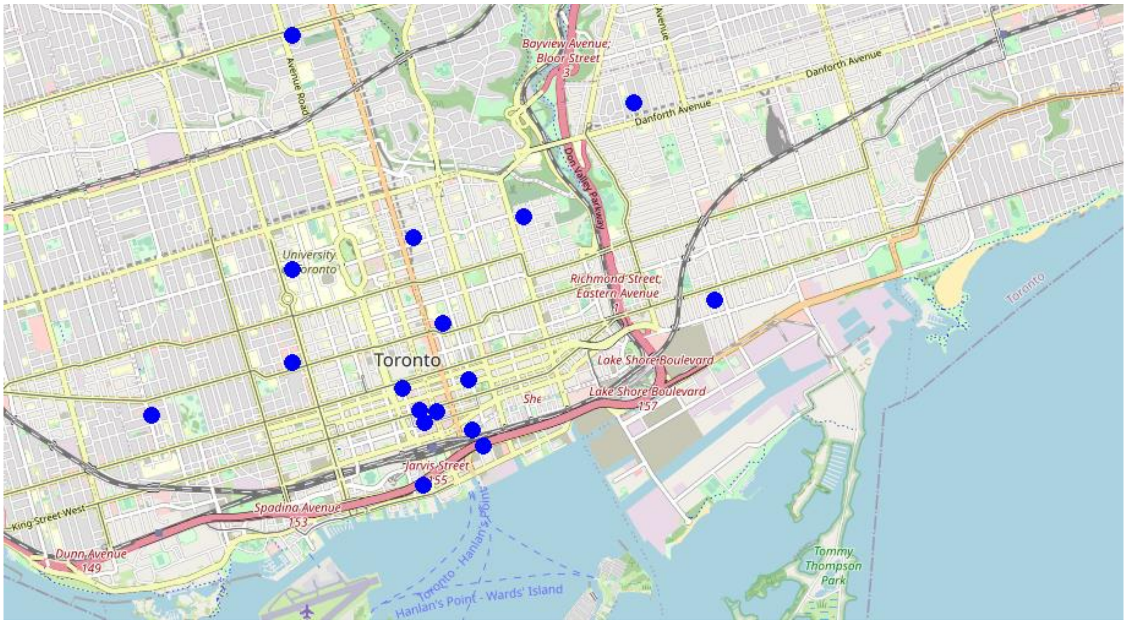
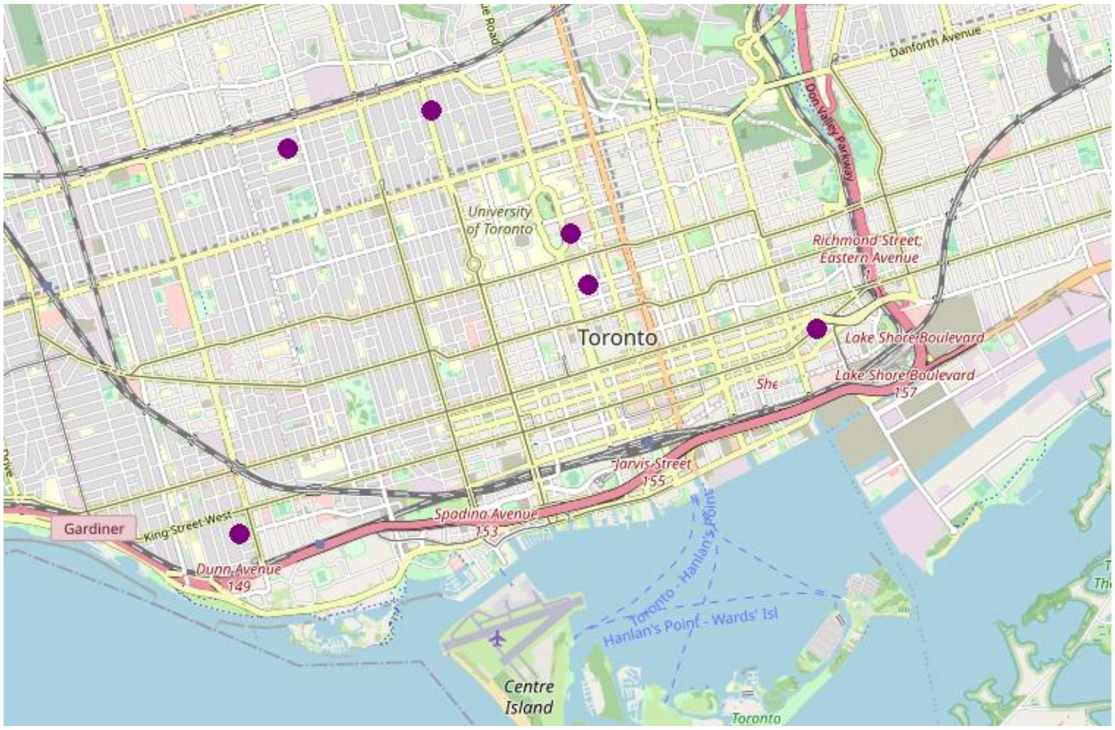


Figure 4 Cluster 2 - High frequency of occurrence



5 Discussion

As seen from the map in the Results section, the highest frequency of occurrence classified as cluster 2 of Coffee Shop is in the central (mint green color). The Cluster 1 represent the area where the frequency of occurrence is moderate. The lowest frequency is in area of cluster 0. This is the area where there is low to no competition. This represents a great opportunity and high potential areas to open new Coffee Shop as there is very little to no competition from Coffee Shop.

Coffee Shop in cluster 2 are likely suffering from intense competition due to oversupply. It means that the oversupply of Coffee Shop mostly happened in the central area of the city. In the suburb area and minor road, the presence of Coffee Shop is lesser.

Therefore, this project recommends investors to consider locating the new Coffee Shop in neighborhoods in cluster 0 where there is little to no competition. Investors are advised to avoid neighborhoods in cluster 2 which already have high concentration of Coffee Shop and suffering from intense competition.

6 Conclusion

This project runs thought a typical approach of data science. The business problem is firstly identified, then data required is specified, they will be extracted and cleaned. The machine learning techniques used to classify processed data by clustering the data into 3 clusters based on their similarities. Lastly recommendations are provided based on the findings. To answer the business question on the location of a new Coffee Shop, the recommendation is given for the cluster 0 where the competition is less severe.

7 Reference

Foursquare Developers Documentation. Foursquare. Retrieved from
https://developer.foursquare.com/docs

Coffee Business Intelligence
https://coffeebi.com/

Toronto
https://en.wikipedia.org/wiki/Toronto

List of Postal Code of Toronto
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M