Coursera IBM Capstone Project: Opening a new Coffeehouse in Toronto

Report Prepared by:
Rishi Mahajan

1. Introduction

1.1 Background

In this Capstone Project, I am creating a hypothetical scenario for an entrepreneur who wants to open 'Coffeehouse' in the Toronto area. The idea behind this project is that there may be many Coffee shops or Coffeehouses in the Toronto area and it might be difficult for an entrepreneur to find the exact location to open a coffee house. So, by using some techniques of data science we will guide him, on which location he should open a shop to gain maximum profit. Coffee house is an establishment that primarily serves coffee(of various types, e.g. espresso, latte, cappuccino). Some coffeehouses may serve cold drinks such as iced coffee and iced tea; in continental Europe, cafés serve alcoholic drinks. A coffeehouse may also serve food such as light snacks,sandwiches, muffins or pastries. From a cultural standpoint, coffeehouses largely serve as centers of social interaction: the coffeehouse provides patrons with a place to congregate, talk, read, write, entertain one another, or pass the time, whether individually or in small groups. A coffeehouse can serve as an informal club for its regular members. By considering all these points about coffeehouse, finding the best location to open such a coffeehouse is one of the most important decisions for the entrepreneur and I am designing this project to guide him find the most suitable location.

1.2 Business Problem

The objective of this project is to find the most suitable location for the entrepreneur to open a coffeehouse in the Toronto area. So, by using some techniques of data science and machine learning methods such as clustering we will guide him, on which location he should open a shop to gain maximum profit. This project aims to provide solutions to answer the business question: If an entrepreneur wants to open a coffeehouse in toronto,where should he consider to open a coffeehouse?

1.3 Target Audience

The entrepreneur who wants to find the best location to open a coffeehouse.

2. Data

To solve this problem we will need data which is associated with the Toronto area. The list of data which will be used is listed below.

1. List of neighborhoods in Toronto, Canada.
2. Latitude and Longitude of these neighborhoods.
3. Venue data related to Coffeehouse. Venue data will help us find the neighborhoods that are most suitable to open a coffeehouse.

3. Extracting the data

1. Scrapping of Toronto neighborhoods via Wikipedia. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
2. Getting Latitude and Longitude data of these neighborhoods via a csv file and merging into the Toronto neighborhood data. (https://cocl.us/Geospatial_data)
3. Using Foursquare API to get venue data related to these neighborhoods.

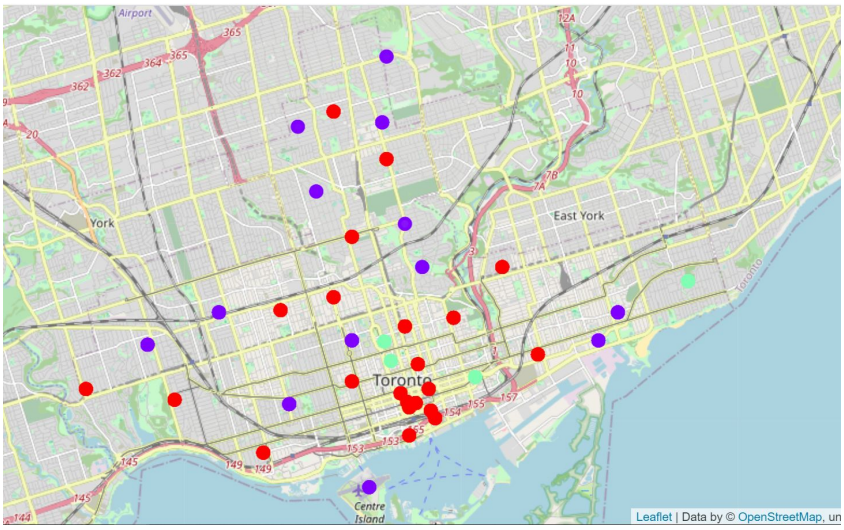
4. Methodology:

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from wikipedia page (“ https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M ”) I did the web scraping by utilizing a beautiful soup python library for pulling data out of HTML and XML files and then used pandas library to frame that data in the dataframe. However, it is only a list of neighborhood names and postal codes. I will need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I used the csv file provided by the IBM team to match the coordinates of Toronto neighborhoods. And after that by using Geocoder library, I found the latitude and longitude of the Toronto area. After gathering all these coordinates, I visualized the map of Toronto using the Folium package to verify whether these are the correct coordinates.

Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have already created a Foursquare developer account in order to obtain an account ID and API key to pull the data. From Foursquare, I am able to pull the venue names, venue categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I noted some points to specifically look for “Coffee Shop”. Firstly I checked if there is any venue category which is related to coffee house. And then I found out Coffee shop venue category is the same as the coffeehouse. After creating dummy variables of the venue category, I counted the number of coffee shops in the Toronto area and its nearly about 27 shops. For doing exact analysis of data, I selected only two columns: Neighbourhood and coffee shop. Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, 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 it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Coffee Shop”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the Coffeehouse.

5. Results:



The results from k-means clustering show that we can categorize Toronto neighborhoods into3 clusters based on how many Coffee Shops are in each neighborhood:

1. Cluster 0: Neighborhoods with a high number of Coffee Shops.
2. Cluster 1: Neighborhoods with little more number of coffee shops than cluster 2
3. Cluster 2: Neighborhoods with very less number of Coffee Shops.

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color and Cluster 2 in light green color. For more detailed information or visualization kindly see the jupyter notebook.

6. Recommendations:

Most Coffee Shops are in Cluster 0 which is around St Andrew, King, Osgoode areas and lowest in Cluster 2 areas which are near to Queens Park, Glen Stewart Park and other areas. Also, there are good opportunities to open near Sackville Street and Glen Stewart Park as the competition seems to be low. Looking at nearby venues, it seems Cluster 2 might be a good location as there are not a lot of Coffee Shops in these areas. There might be good business in the area like Glen Stewart Park, Because tourists might visit that place for some coffee and for some snacks. Also if an entrepreneur wants to open a coffeehouse in the main city then location near 'Queens Park' will be best suggestion to open a coffeehouse because Queens Park and many things are near to this area like hospitals, university of Toronto, some corporate offices. So the people from these locations will definitely visit the coffee house for coffee and for some meetings. Therefore, this project recommends the entrepreneur to open a Coffeehouse in these locations with little to no competition. Nonetheless, if the coffee and snacks taste good, affordable with pleasant ambiance, I am confident that it will have a great following and business anywhere in these suggested locations.

7. Limitation and Future Work

In this project, I only take into consideration one factor: the existence of Coffee shops in each neighborhood. There are many essential factors that can be taken into consideration such as population density, young and corporate people in that area, rent that could influence the decision to open a new shop. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration these factors.

8. Conclusion:

In this capstone project, I have gone through the process of identifying and solving the business problem, finding and specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendations about the exact location to the entrepreneur to open a coffee house. This project will help the relevant entrepreneur to capitalize on the opportunities in high potential locations while avoiding overcrowded areas in their decisions to open a new coffeehouse.

9. References

1. List of neighborhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. CSV File provided by IBM: (https://cocl.us/Geospatial_data)
3. Foursquare Developer Documentation: https://developer.foursquare.com/docs.