IBM Applied Data Science Capstone

Opening a New Grocery Store in Toronto, Canada

By

M Verma

January 30, 2020

A. Introduction

A.1. Background

Toronto, the capital of the province of Ontario, is the most populous Canadian city with a population of 2,731,571 as of 2016. It is a global city filled with vast opportunity and is home to an array of distinctive and dynamic neighborhoods that reflect the diversity of its population. The city is also known for its vibrant arts and entertainment scene, incredible cultural festivals, delicious food, excellent shopping, thriving business, beautiful parks and beaches, and much more. According to the Youthful Cities Index, 2015, Toronto is considered be the top 10 most appealing cities to live and work. Diversity truly is Toronto's strength, making the city a dynamic, progressive and welcoming place to work and live.

A.2. Problem Description:

Now let me explain the context of this project through a scenario. Say you are a retailer and wanted to open a grocery store in the city of Toronto in Canada. As a retailer, the central location and the large crowd would provide a great distribution channel to market your products and services. Particularly, the location of the grocery store is one of the most important decisions that will determine whether the store will be a success or a failure. Wouldn't be great if you are able to analyze the neighborhoods in Toronto which are most profitable since the success of the store depends on the population and location?

A.3. Objective

The aim of this project is to study and analyze the neighborhoods of Toronto city and group them into similar clusters and, to analyze those clusters to find out those neighborhoods that are more profitable to open a new grocery store. 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, Canada, if a retailer is looking to open a new grocery store, where would you recommend that they open it?

A.4. Target Audience

The information provided by this report would be particularly useful to:

- Retailers and investors looking to open or invest in grocery store in the capital city of Canada i.e. Toronto.
- Freelancer who loves to have their own grocery store as a side business.
- Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

B. Data Description:

To solve the problem, we will need the following data:
a) Toronto City data.

- Data source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Description: The data contains the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Geospace data for each neighborhood in Toronto City.

- Data source: https://cocl.us/Geospatial_data csv file
- Description: The data contains the geographical coordinates i.e. latitudes and longitudes of the neighborhoods.

c) Venue data, particularly data related to grocery store.

- Data source: https://developer.foursquare.com/docs
- Description: Foursquare's explore API gives details such as names, categories and locations (latitude and longitude).

C. Methodology

To start with our analysis, firstly, we need to get the list of neighborhoods in the city of Toronto. The list is available in Wikipedia page, we will do web scraping using Python requests to extract the list of neighborhoods data. However, we need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. In order to do so, we will fetch the coordinate data for all the neighborhoods in in the city of Toronto using the csv file. After gathering the data, we will populate the data into a pandas DataFrame. Next, we will combine both the dataframes i.e. the dataframe with the coordinate data and the dataframe with the list of neighborhoods. Then, we will visualize the neighborhoods in the city of Toronto on a map using Folium package. This will ensure that the geographical coordinate data returned by csv file are correctly plotted in the city of Toronto.

Next, we will utilize Foursquare API to explore the neighborhoods. We will set the LIMIT parameter to 100, and Radius to 1000 meter. This will return the top 100 venues that are within a radius of 1000 meters. To do so, we need to register a Foursquare Developer Account. After registering, we will obtain the Foursquare ID and Foursquare secret key. Then, we will make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. This will give us number of venues returned for each neighbourhood and then we can examine number of unique categories curated from all the returned venues. Then, we will analyze each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. Since we are analyzing the "Grocery Store" data, we will filter the "Grocery Store" as venue category for the neighborhoods.

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 frequently used popular unsupervised machine learning algorithms and is particularly suited to solve the business problem for this project. We will use KMeans algorithm from sklearn library and we will find the number of clusters by using the technique called Elbow Method. The results will allow us to identify which neighborhoods have higher concentration of grocery stores while which neighborhoods have fewer number of grocery stores. Based on the occurrence of grocery stores in different neighborhoods, it will help us to answer the business problem.

D. Results

We have reached at the end of the analysis. In this section we will record all the findings from the above K-means clustering & visualization of the Toronto dataset.

Firstly, let's document the results from the k-means clustering. The clustering shows that neighborhoods are categorized into five clusters based on the frequency of occurrence for "Grocery Store":

- Cluster 0: Neighborhoods with high concentration of grocery stores.
- Cluster 1, 3, and 4: Neighborhoods with moderate number of grocery stores.
- Cluster 2: Neighbourhoods with low number to no existence of grocery stores.

The results of the clustering can also be visualized in the map below with cluster 0 in red color, cluster 1 in purple color, cluster 2 in blue color, cluster 3 in green color, and cluster 4 in yellow color

Secondly, let's record the result from exploratory data analysis:

- From the bar plot in visualization section, we found that out of 11 boroughs only East York, Etobicoke, North York , Scarborough & York boroughs have large amount of Grocery stores. However, Central Toronto, Downtown Toronto, Queen's Park, east Toronto, West Toronto & Mississauga have less number of Grocery stores.
- From the visualization section, it is also being analyzed that it won't be a good decision to open a grocery store in York borough and a neighborhood York Mill West due to large amount of grocery stores. Also, from the clustering section, it is being observed that the above specified borough and neighborhood lies in Cluster 4.

E. Discussion

According to this analysis, most of the grocery stores are concentrated in the East York, Etobicoke, North York, Scarborough & York boroughs of Toronto city, with the highest number in cluster 0 and moderate number in cluster 1, 3, and 4. On the other hand, cluster 2 has very low number to no grocery store in the neighborhoods. This represents a great opportunity and high potential areas

to open a new grocery store as there is very little to no competition from existing stores. Meanwhile, grocery stores in cluster 0 are likely to be suffered from intense competition due to oversupply and large amount of grocery stores. Therefore, this project recommends retailers to open a new grocery store in neighborhoods in cluster 2 with little to no competition.

F. Limitations

However, this project has some limitations. Firstly in this project, we had considered only one factor i.e. frequency of occurrence of grocery stores, but there are other factors which should have been taken into account. Those other factors could be income of residents, market price of land, and percentage of population. Secondly, the clustering is completely based only on data obtained from Foursquare API that came with limitations as to the number of API calls and results returned.

G. Conclusion

In this project, we have gone through the various steps ranging from identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. retailers and investors regarding the best locations to open a new grocery store. The answer to the business question that was raised in the introductory section, “In the city of Toronto, Canada, if a retailer is looking to open a new grocery store, where would you recommend that they open it? “, the answer proposed by this project is: The neighborhoods in cluster 2 are the most preferred locations to open a new grocery store. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new grocery store.