Applied Data Science Capstone

Opening a new Bookstore in Toronto, Canada

By: Riyad Bin Rafiq

June 2019

Introduction

A bookstore is a place where different kinds of textbooks or reference books are sold. It is a very important place for all educated persons. Most of the customers of the bookstore are book lovers of all ages. It is mostly familiar to the students. In a bookstore, there are different shelves which are nicely arranged. From a book shop, we can buy our textbooks, story books, novels, dictionaries, science fiction books etc. In fact, a bookstore always plays an important role for the readers and mostly the students. There are many bookstores in the city of Toronto. Serious considerations are required to open a new bookstore than it seems. So the location of the bookstore is one of the most important decisions that will determine whether the bookstore will be a success or failure.

Business Problem

The objective of this project is to analyse and select the best locations in Toronto city to open a new bookstore. For this project, data science methodology and machine learning techniques like clustering are used. The provided solution gives answer to the question: If an investor wants to open a new bookstore, what would be the best place that you recommend to open?

The project is useful to investors or book lovers who want to open a new bookstore in the Toronto city.

Data

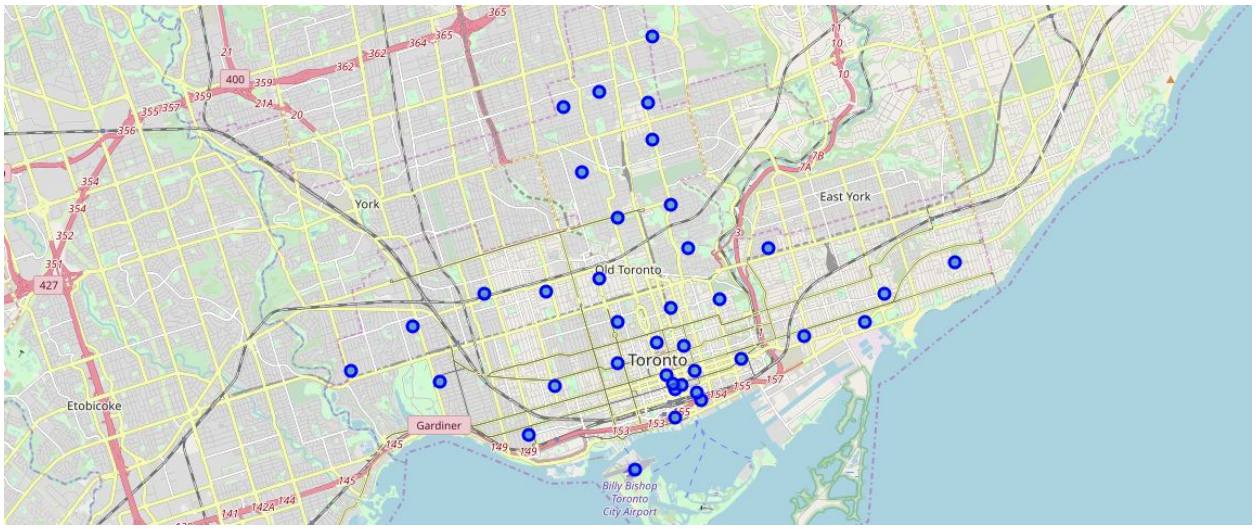
To solve the problem we need the following data:

- List of neighborhoods in Toronto.
- Latitude and longitude coordinates of those neighborhoods. We will use this data to plot the map and also get the venue data.
- Venue data specially bookstores are used to perform clustering.

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, this Wikipedia page contains our required data including postal code, borough and neighborhoods. I will extract the data from the wikipedia page using web scraping techniques. I will also get latitude and longitude of our specified neighborhoods. Then I will use Foursquare API to get the venue data of those neighborhoods. Foursquare API usually provides various categories of the venue data. In this project, I am particularly interested in bookstore data which will help to solve the business problem. Data science methodologies are used in this project such as data scraping, data cleaning, data wrangling, machine learning i.e. K-means clustering, map visualization.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Toronto. Web scraping using Python requests has been done to extract the list of neighborhoods data from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. For this project, the geographical coordinates are needed in the form of latitude and longitude in order to be able to use Foursquare API. After that, the gathered data is stored into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates are correctly plotted in the Toronto city.



Next, Foursquare API will be used to get the top 100 venues that are within a radius of 500 meters. I need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then I need to make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and the venue name, venue category, venue latitude and longitude will be extracted then. With the data, it is possible to check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, I will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, the data is being prepared for use in clustering. Since I am analysing the "Bookstore" data, I will filter the "Bookstore" as venue category for the neighborhoods.

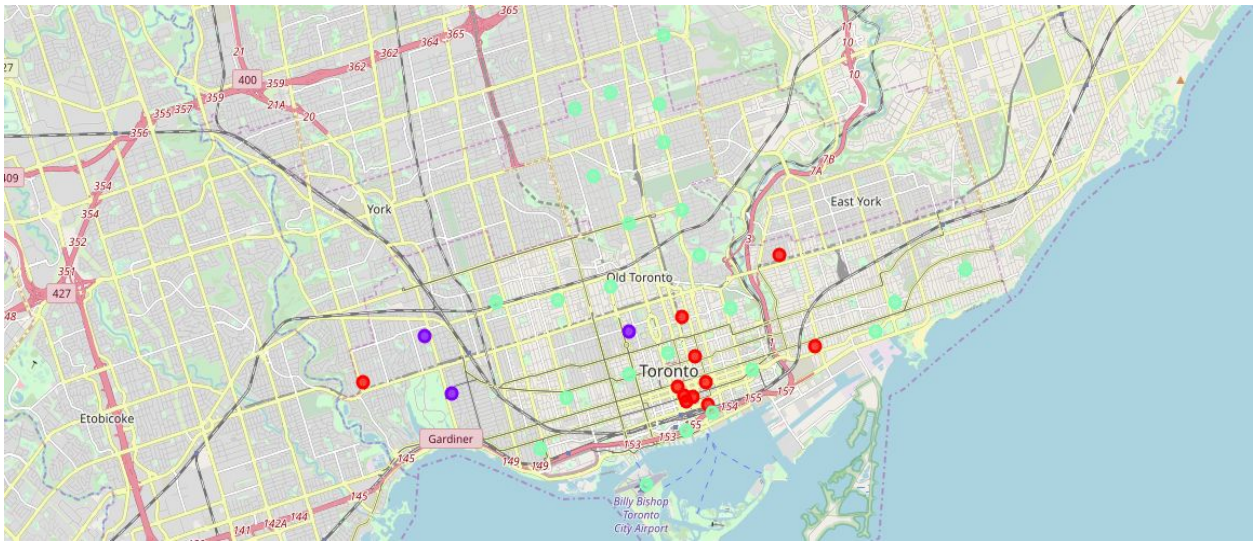
Lastly, I will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. I will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Bookstore". The results will allow to identify which neighbourhoods have higher concentration of bookstores while which neighbourhoods have fewer number of bookstores. Based on the occurrence of bookstores in different neighbourhoods, it will help to answer the question as to which neighbourhoods are most suitable to open new bookstore.

Results

Results show that the neighborhoods are categorized into three clusters using K-means clustering based on the frequency of occurrence for “Bookstore”.

- Cluster 1: Neighborhoods with moderate number of bookstores.
- Cluster 2: Neighborhoods with higher number of bookstores.
- Cluster 3: Neighborhoods with no existence of bookstores.

In map visualization, red color indicates cluster 1, purple and green color specifies cluster 2 and cluster 3 respectively.



Discussion

As shown above by map visualization, neighborhoods of cluster 1 and cluster 2 contain bookstores. On the other hand, cluster 3 contains neighborhoods with no bookstores. So these are the high potential places to open a new bookstore as there will be no competition. If any book lovers or investors want to open a new bookstore in the neighborhoods of cluster 2 which has higher number of bookstores, they will face enough competition. As a result, investors are advised to avoid neighbourhoods in cluster 2 which are high concentrated area of bookstores and suffering from intense competition. We consider only one factor, frequency of occurrence of bookstores in this project. There are also other factors such as population, income of residents, education of residents that could also influence the best location to open a new bookstore. But this information is not available. With this information the research provides a concrete methodology to recommend best places to open bookstores in future.

Conclusion

The project has been approached step by step. At first, a business problem was identified. Then eventually, the processes like data cleaning, data wrangling and sources of the specified data has been detailed to prepare data. Then K-means clustering has been executed to cluster the neighborhoods and recommend the opportunistic neighborhoods to open a new bookstore.