Applied Data Science Capstone

IBM Data Science Professional

Opening a new Bookstore in Toronto, Canada

By: Syed Yusuf

April 2020

INTRODUCTION

There are several ways of growing one's business. One of them is by making your product or service available to a new pool of customers. The most obvious is to open stores in new locations. And in doing so, selecting the best location plays an important role.

In this capstone project for IBM Data Science Professional Certificate. I am creating a hypothetical scenario where a Bookstore wants to expand its franchise in the city of Toronto. Hence the main aim of this project is to help the bookstore decide the best location to start a new outlet of the bookstore.

The location should be selected keeping in mind the potential customers, probably in the neighborhoods where there are more students, for example near hostels and universities. Apart from this to get a good business the location shouldn't have a lot of competition from existing bookstores.

Keeping these things in mind, in this project we would cluster the neighborhoods based on the number of bookstores present and try to identify the best neighborhoods to open a new bookstore.

BUSINESS PROBLEM

The objective of this capstone project is to find the most suitable location for the bookstore to open a new outlet in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if a Bookstore wants to open a new outlet, where should they consider opening it?

TARGET AUDIENCE

The bookstore owner who wants to find a new location to open an outlet for his bookstore.

DATA

To solve this problem, we will need below data:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue data related to Bookstores. This will help us to cluster the neighborhoods and find the clusters that are more suitable to open a Bookstore.

EXTRACTING THE DATA

- Scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

Sources of data and methods to extract them

This Wikipedia page https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto contains a list of neighborhoods in Toronto. We will use web scraping techniques to extract the data form the Wikipedia page, with the help of Python Requests and BeautifulSoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Bookstore category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

METHODOLOGY

Firstly, we need to get the list of neighborhoods in the city of Toronto. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto)We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Toronto.

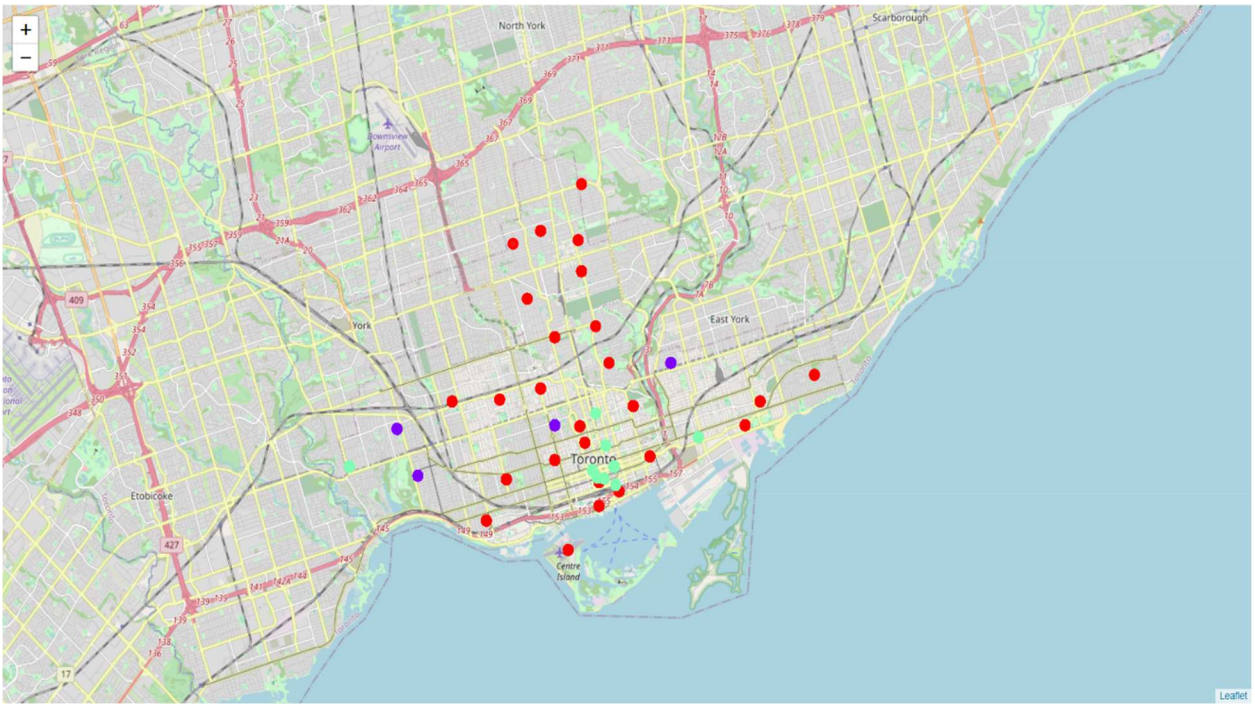
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods 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. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Bookstore” data, we will filter the “Bookstore” 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 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 “Bookstore”. The results will allow us to identify which neighborhoods have higher concentration of Bookstores while which neighborhoods have fewer number of Bookstores. Based on the occurrence of Bookstores in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Bookstores.

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 “Bookstore”:

- Cluster 0: Neighborhoods with low number to no existence of Bookstores
- Cluster 1: Neighborhoods with high concentration of Bookstores
- Cluster 2: Neighborhoods with moderate number of Bookstores

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.



DISCUSSION

As observations noted from the map in the Results section, most of the Bookstores are concentrated in the downtown area of Toronto city, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low number to no Bookstore in the neighborhoods. This represents a great opportunity and high potential areas to open new Bookstores as there is very little to no competition from existing bookstores. Meanwhile, Bookstores in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Bookstores. From another perspective, the results also show that the oversupply of Bookstores mostly happened in the downtown area of the city, with the suburb area still have very few Bookstores. Therefore, this project recommends bookstore owners to capitalize on these findings to open new Bookstores in neighborhoods in cluster 0 with little to no competition. Bookstore owners with unique selling propositions to stand out from the competition can also open new Bookstores in neighborhoods in cluster 2 with moderate competition. Lastly, Bookstore owners are advised to avoid neighborhoods in cluster 1 which already have high concentration of Bookstores and suffering from intense competition.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this project, we only consider one factor i.e. frequency of occurrence of Bookstores, there are other factors such as presence of Schools and Universities and the residents of that neighborhood that could influence the location decision of a new Bookstore. However, to the best knowledge of this researcher, such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Bookstore. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

CONCLUSION

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. Bookstore owners regarding the best locations to open a new Bookstore. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new Bookstore if the bookstore is a new one. Else if, the bookstore already has a brand value then it could consider opening its new outlet in the neighborhoods of cluster 2. 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 Bookstore.

REFERENCES

- List of neighborhoods in Toronto. Wikipedia. https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur
- Foursquare Developers Documentation. Foursquare. https://developer.foursquare.com/docs
- Geospatial Coordinates of Toronto Neighborhoods https://cocl.us/Geospatial_data

APPENDIX

Cluster 0

- Berczy Park
- Kensington Market / Chinatown / Grange Park
- Lawrence Park
- Little Portugal / Trinity
- North Toronto West
- Queen's Park / Ontario Provincial Government
- Moore Park / Summerhill East
- Harbourfront East / Union Station / Toronto
- India Bazaar / The Beaches West
- Toronto Dominion Centre / Design Exchange
- Summerhill West / Rathnelly / South Hill
- The Annex / North Midtown / Yorkville
- St. James Town / Cabbagetown
- Regent Park / Harbourfront
- Rosedale
- Davisville
- Forest Hill North & West
- Roselawn
- CN Tower / King and Spadina / Railway Lands
- Central Bay Street
- Dufferin / Dovercourt Village
- Davisville North
- Business reply mail Processing Center
- Christie

Cluster 1

- Parkdale / Roncesvalles
- The Danforth West / Riverdale
- University of Toronto / Harbord
- High Park / The Junction South

Cluster 2

- Stn A PO Boxes
- Studio District
- Commerce Court / Victoria Hotel
- First Canadian Place / Underground city
- Garden District, Ryerson
- Church and Wellesley
- Richmond / Adelaide / King
- St. James Town
- Runnymede / Swansea