IBM Data Science Capstone Project

Keivan Mokhtarpour
Coursera

August 17, 2020

1 Introduction

This is a capstone project for IBM Data Science Professional Certificate. In this project, a data-analyzed approach to find the best possible locations for opening up a coffee house in the greater Toronto area (GTA) is considerable. As you may know, many immigrants enter Canada every year using investment and entrepreneurship programs. Most of them end up buying local businesses or simply opening up a franchised coffee house like Tim Hortons or Starbucks in large cities where the turnover of their investment is somehow guaranteed. Therefore, many of them are concerned about the location as a key factor in their annual total sales. GTA has been host to many immigrants over the past few decades and they have enlarged the city in terms of size by occupying new neighbourhoods and territories. The development of the Toronto city has extended its boundaries all the way to the northern regions like York where many immigrants specially Iranians, Koreans and the Chinese have chosen North York as their new home. This is a challenging decision for many of them whether to buy their first business in the north or prefer Downtown Toronto over it. This project is aimed at helping them to narrow down the search area they are looking for opening up a new coffee house that is a $6.2 billion industry in Canada.

2 Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new coffee house in the following boroughs of the greater Toronto area (GTA):

- Central Toronto
- Downtown Toronto
- East Toronto
- East York
- North York
- West Toronto

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 an entrepreneur wants to open a coffee house, where should they consider opening it?

3 Client

The entrepreneur who wants to find the location to open a coffee house.

4 Data

To solve this problem, we will need the following data:

- List of neighborhoods in Toronto, Canada. (Web Scrapping through Wikipedia)
- Latitude and Longitude of these neighborhoods (Geocoder Package)
- Venue data related to coffee houses (Foursquare API). This will help us find neighborhoods that are more suitable to open a coffee shop.

5 Methodology

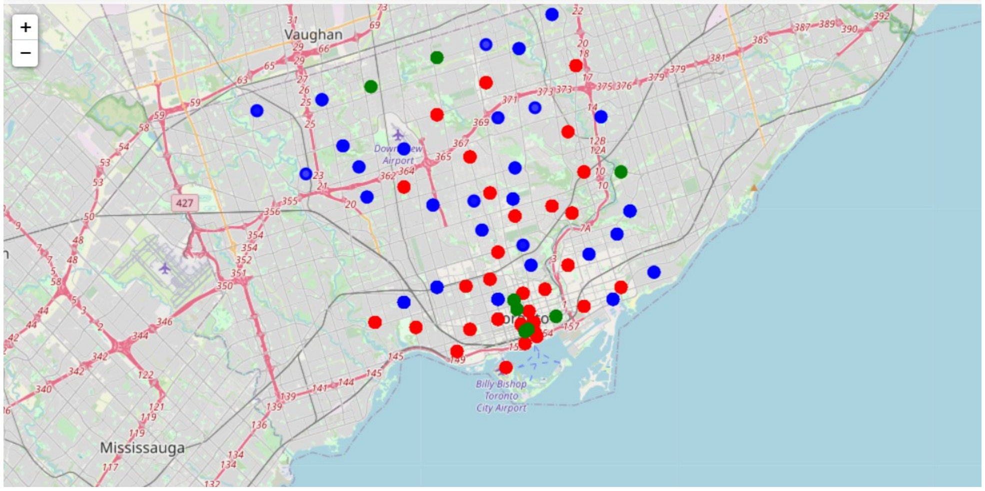
First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia. I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.

However, it is only a list of neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder Package but it was not working so I used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories 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 made a justification to specifically look for 'Coffee Shop'. Lastly, I performed the clustering method 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 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 coffee house.

6 Results

Clusters



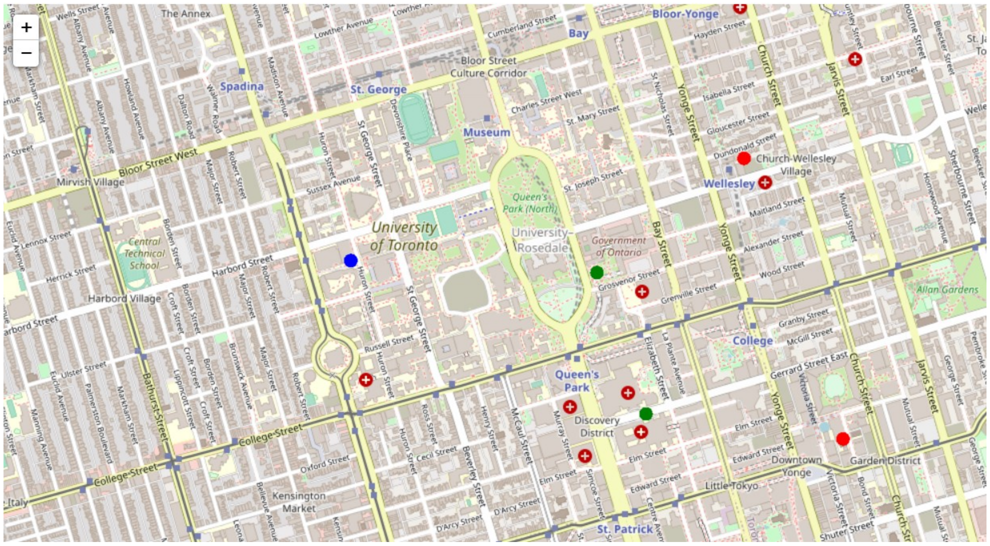
The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many coffee houses are in each neighborhood:

- Cluster 0: Neighborhoods with a high density of coffee houses.
- Cluster 1: Neighborhoods with the least number of coffee houses.
- Cluster 2: Neighborhoods with a limited number of coffee houses.

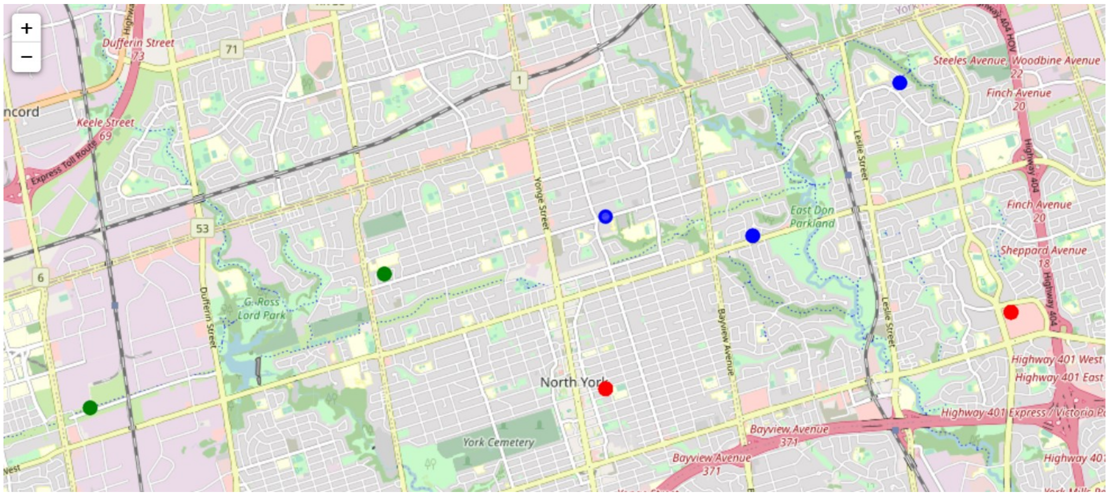
The results are visualized in the above map with Cluster 0 in green, Cluster 1 in blue, Cluster 2 in red.

7 Recommendations

Based on the results section, we recommend the entrepreneur not to open a coffee house in the neighbourhoods denoted with the red circles. As it could be observed, most of the areas of Downtown, Toronto have been over-populated with different coffee houses. The nearest location to Downtown, Toronto which is categorized into the blue cluster with the least number of coffee houses is University of Toronto, Harbord where only one venue named Elchi Chai Shop exists.



On the other hand, there are more locations in North York, Ontario with the least or limited number of coffee houses that are suitable for opening up businesses. Neighbourhoods like Willowdale (Cummer Ave.), Newtonbrook and Bayview village could be good choices in this regard. In total, Northern locations of the Finch Ave. in the North York borough could be considered as potential location for the clients.


+
−
Dufferin Street 73
71
1
Concord
Keele Street 69
Express Toll Route
53
6
Yonge Street
Dufferin Street
G. Ross Lord Park
Steeles Avenue, Woodbine Avenue 22
Finch Avenue 20
Highway 404
Leslie Street
East Don Parkland
Finch Avenue 20
Sheppard Avenue 18
Bayview Avenue
North York
York Cemetery
Highway 401 West
Highway 401 East
Highway 401 Express / Victoria Park
Bayview Avenue 371
Bayview Avenue 371