Coursera Capstone

IBM Applied Data Science Capstone Project

Finding the most suitable place to open a restaurant in Toronto

By: Ajay Mathew

July 2020

1.Introduction

1.1 Background

Whenever someone wants to establish a new business, that person would want it to be successful. For the business to be successful, there should be minimum competition for customers in its surrounding area. The aim of this project is to identify places within Toronto that have increased chances of profitability if a restaurant is opened there.

1.2 Problem

Data that might help contribute to finding the solution mainly consists of the nearby venues. The project aims at finding a suitable neighborhood within Toronto to start a new restaurant using the acquired data.

1.3 Interest

The target audience of this project might be a person who is hoping to start a restaurant in Toronto area. For a business like owning a restaurant, less the competition, the better. We would be trying to help locate neighborhoods with the least number of restaurants in Toronto.

2.Data Collection and Cleaning

2.1 Data Collection

The data needed for this project is a combination of data from three different sources. The first data source of the project uses web scraping to retrieve the data from the Wikipedia page : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

We scrape the web page using the pandas.read_html() method. The different columns are:

- **Postal code:** The corresponding postal code of the region
- **Borough:** The name of the borough
- **Neighborhood:** The name of the neighborhood

We would also use the corresponding file to get the coordinate value for each neighborhood from the link: http://cocl.us/Geospatial_data. The data consists of the following columns:

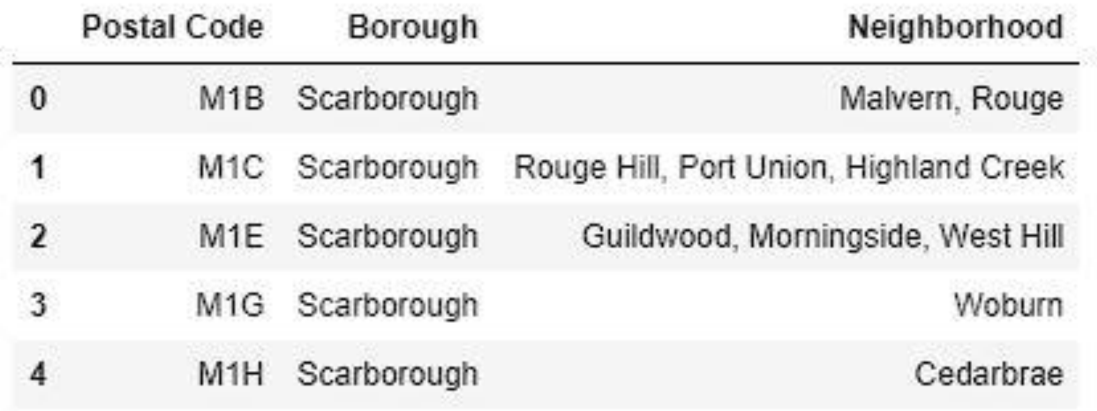
- **Postal Code:** The postal code of a region
- **Latitude:** The latitude of the corresponding region
- **Longitude:** The longitude of the corresponding region

Geocoder library will also be used to retrieve coordinates of needed locations as per demand.

We would be utilizing the Foursquare API to get venue data for the selected neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by more then 125,000 developers.

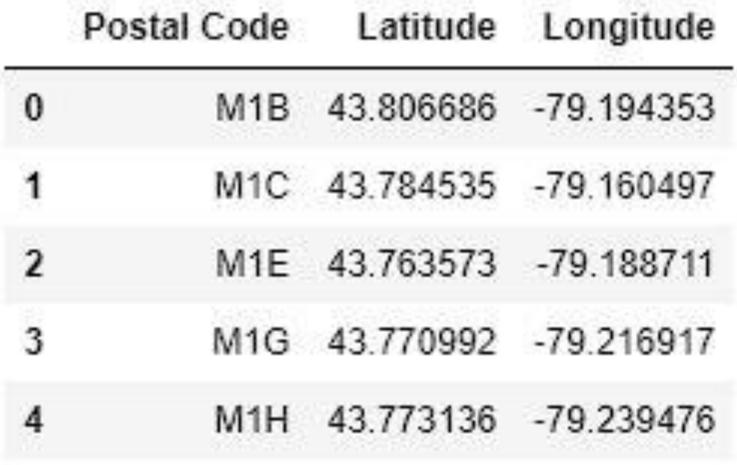
2.2 Data Cleaning

The data extracted from the Wikipedia page is cleaned to remove any rows whose postal code has not yet been assigned a value. We could also group the data according to the Postal CodeThe resulting dataframe after cleaning it is:



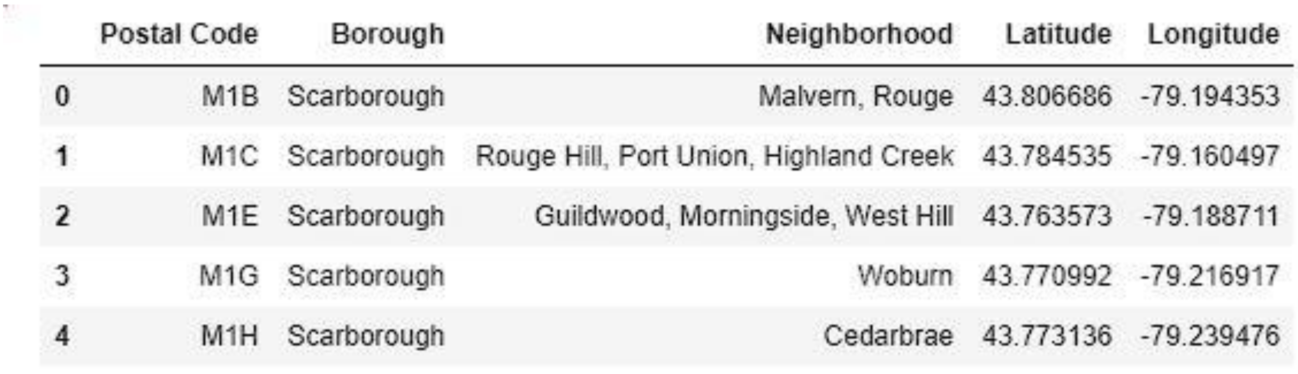
	Postal Code	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

The coordinate dataset for the neighborhoods can be directly used since it does not contain any missing or erroneous values. The corresponding dataframe is :



	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

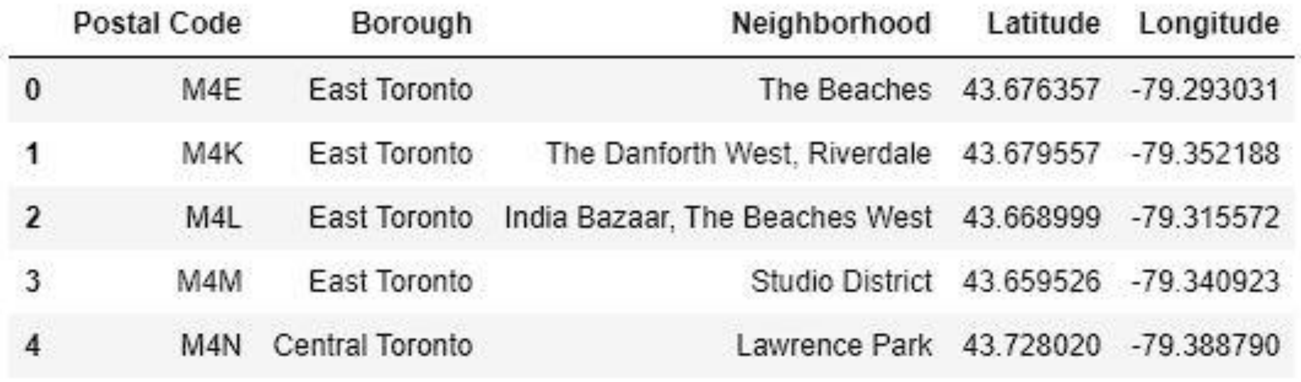
Joining both the above dataframes on the column ‘Postal Code’, we get the corresponding dataframe:



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

This DataFrame would be the basis of all our data analysis.

The results of the Foursquare API calls are in Json files. The resulting Json file from the query would be converted into DataFrames for further use in our program. The corresponding DataFrame is:



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

3. Methodology

Firstly, we need to get the list of neighborhoods in the city of Toronto along with the borough name and the post code. Fortunately, this data in present in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M).).

We extract this from that page using the pandas.read_html() function. However, we need to get the geographical coordinates in the form of latitude and longitude in order to be able use the Foursquare API. This data is available in the csv file (http://cocl.us/Geospatial_data) which can be easily imported into a pandas dataframe. We will be using Geocoder package that will help us convert conventional addresses into geographical data. After gathering the data, we will visualize the neighborhoods in a map using Folium package.

Since, we are interested only in the Toronto city, we will neighborhoods in the Toronto area only.

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 retuned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking mean 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 “Restaurant “ data, we will filter the “Restaurant “ as venue for the neighborhoods.

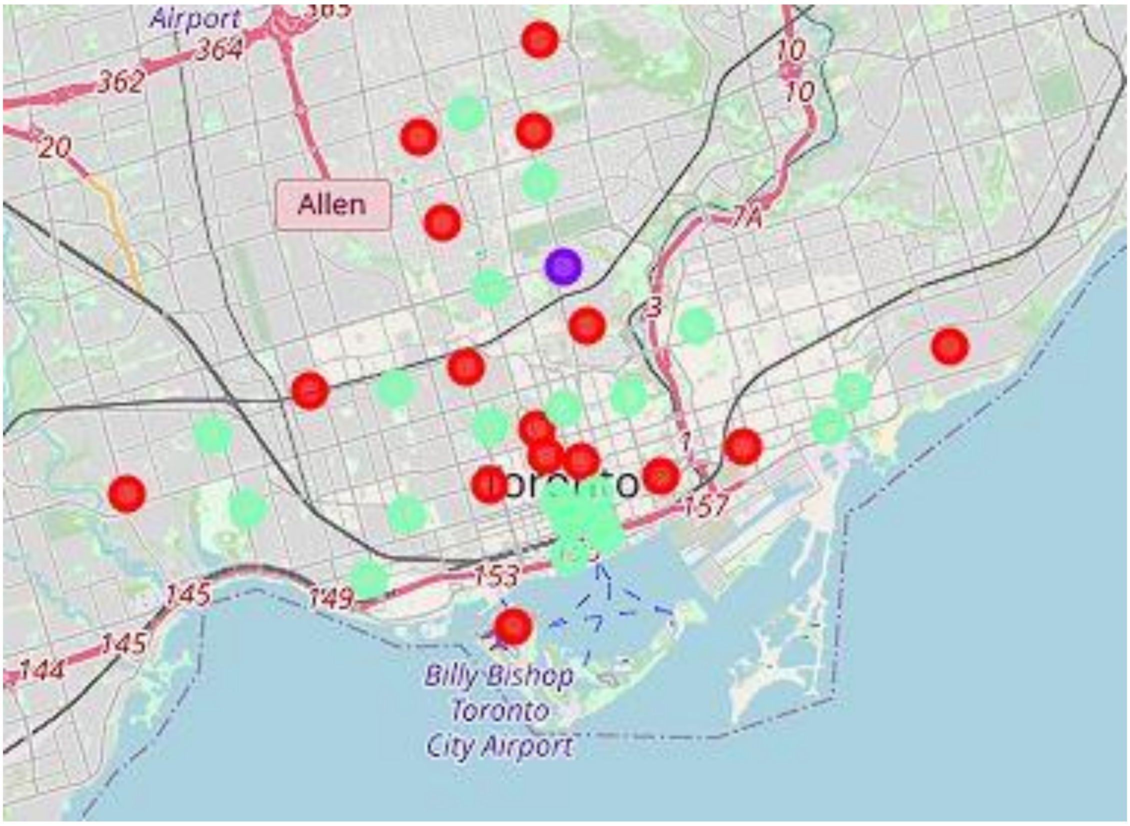
Lastly. We will perform clustering on the data by using the k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 cluster based on their frequency of occurrence of “Restaurant “. The results will allow us to identify which neighborhoods have higher concentration of restaurants while which neighborhoods have fewer number of restaurants. Based on the occurance of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open a new restaurant.

4. 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 “Restaurant “.

- Cluster 0: Neighborhoods with low to no existence of restaurants.
- Cluster 1: Neighborhood with high concentration of restaurants.
- Cluster 2: Neighborhoods with moderate concentration of restaurants.

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in violet color, and cluster 2 in mint green color.



5. Discussion

As observations noted from the map in the Results section, most of the restaurants are present in cluster 1 and 2, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low numbers to no number of restaurants. This presents a great opportunity and high potential areas to open new restaurants as this is very little to no competition from existing restaurants. Meanwhile, restaurants in cluster 1 are likely to suffer from intense competition due to oversupply and high concentration of restaurants. From another perspective, the results show that the oversupply of restaurants mostly happened near a park. The results also show areas of the alternating behaviour of concentration of restaurants. Areas having restaurants are usually nearby to areas with no restaurants. So as we can see from these results, any neighborhood specified in the cluster 0 are good areas to open a restaurant.

6.Conclustion

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. 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 restaurant. The findings of this project will help capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decision to open a new restaurant.