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:

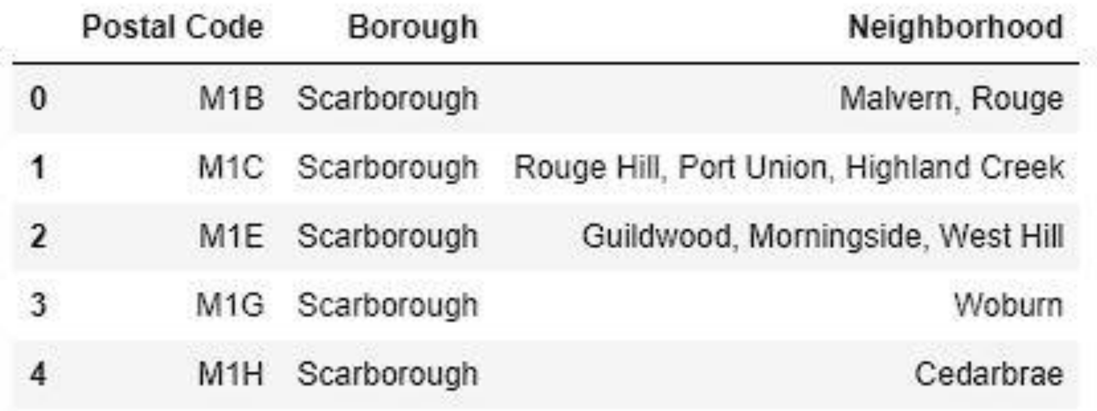
- **Postal Code:** The postal code of a region
- **Latitude:** The latitude of the corresponding region
- **Longitude:** The longitude of the corresponding region

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.

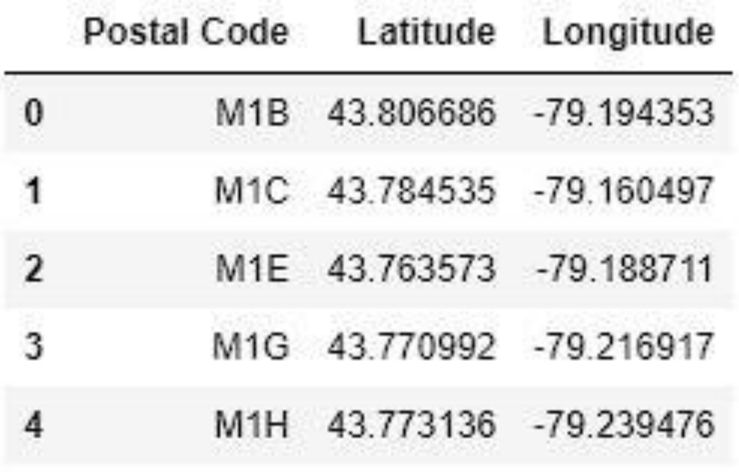
2.2 Data Cleaning

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:



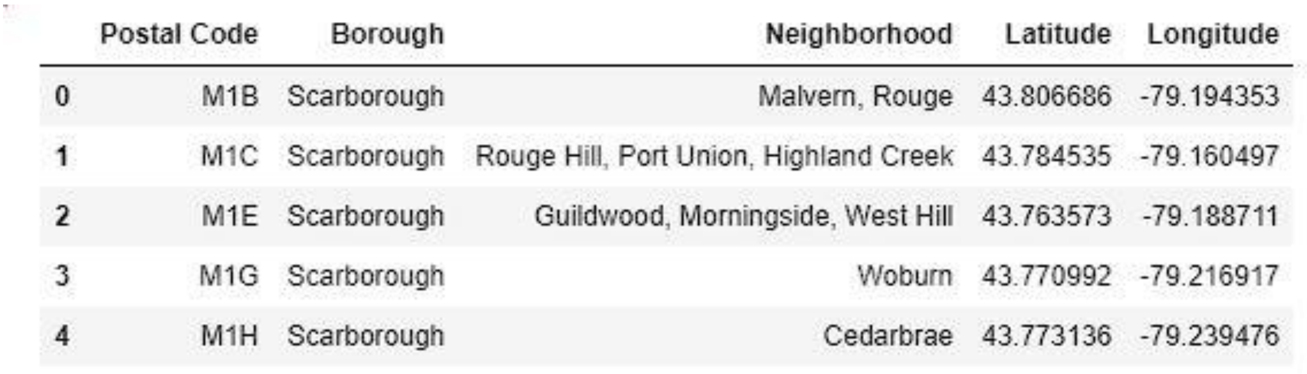
	Postal Code	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

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

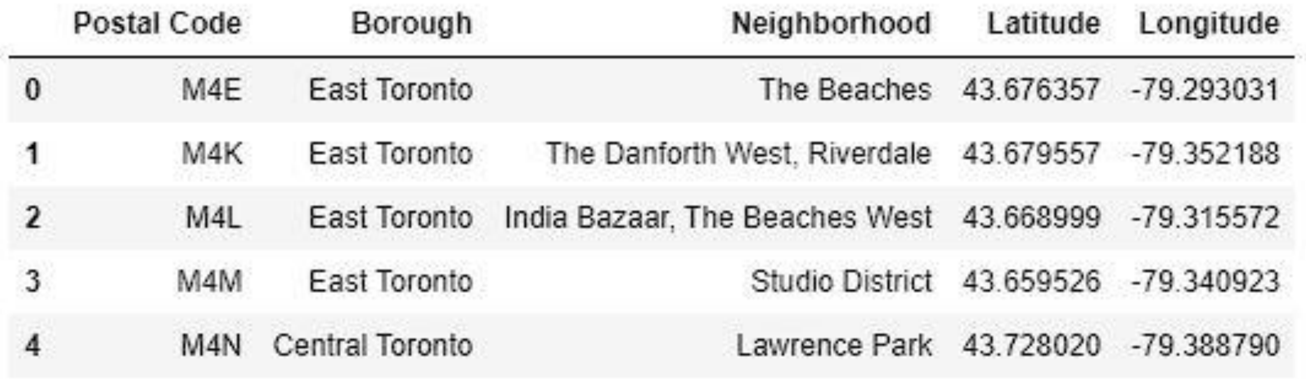
Joining both the above dataframes on the column ‘Postal Code’, we get the corresponding dataframe:



	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