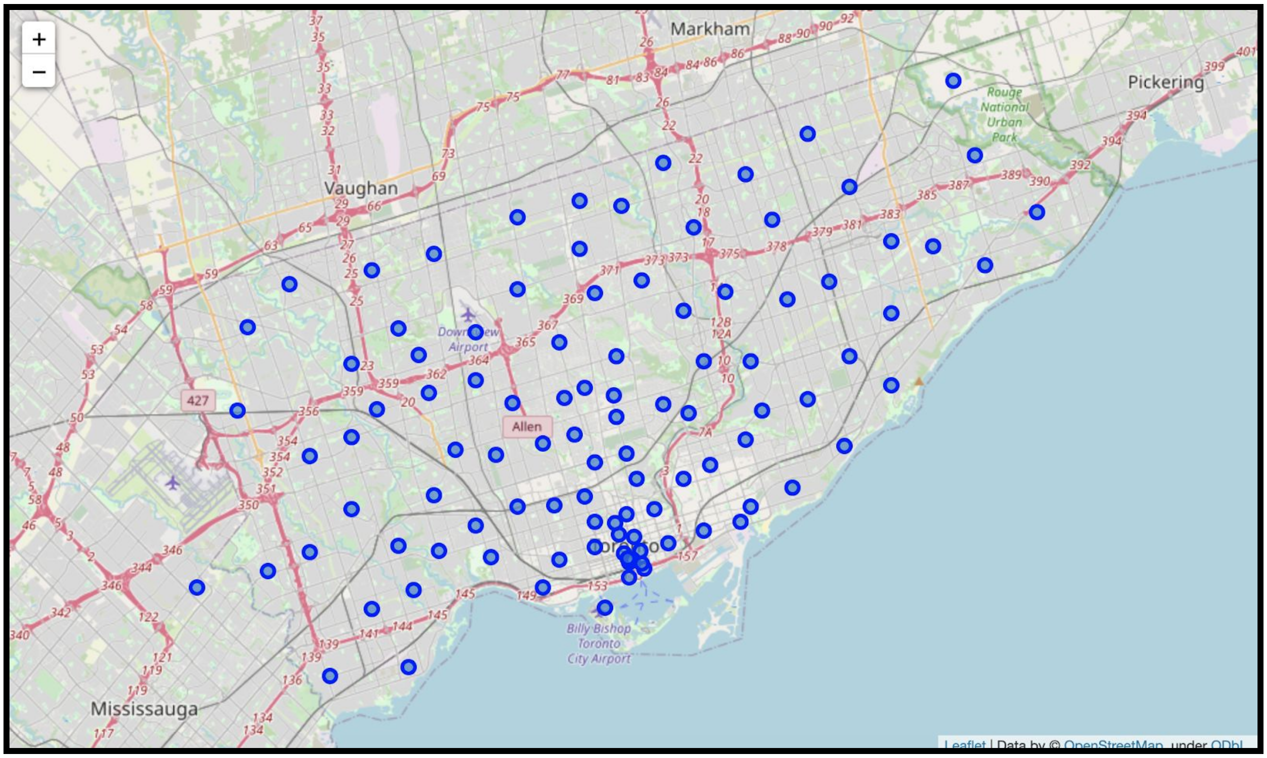
IBM Data Science Professional Certificate offered by Coursera
Capstone Project

Battle of the Neighborhoods

Data Science using Foursquare API

By Eshita Goel


+
−
Markham
Pickering
Rouge National Urban Park
Vaughan
Downsview Airport
427
Allen
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL

Introduction

This report is part of the Capstone Project for IBM's Applied Data Science Professional Certificate offered by Coursera. This is part of the final course in this 9-course series.

We will be using several data visualization techniques, in particular, we will be making use of Foursquare API to retrieve location data for the state of Toronto in Canada and use this data to perform data analysis.

1. Introducing the Project

1.1 Background

Toronto is a big state with numerous neighbourhoods. Each neighbourhood has its share of shops, restaurants, cafes, beaches etc. The different places add to the vibrance of the state and make it a great place to live and travel for others. The place also provides many opportunities for entrepreneurs, especially those that want to start afresh. Being so close to the capital of Canada, it receives its fair share of foot traffic from foreigners. In this project, we aim to analyze the various places in Toronto using location data imported from the Foursquare API.

1.2 The Problem

If a person has to open a new restaurant in a neighbourhood in Toronto then what neighbourhood should he/she choose based on the restaurant type. And if they have a specific restaurant type in mind then what place would be best ensuring a good amount of customer traffic but also keeping in mind the amount of competition. We also cluster the neighbourhoods based on the most popular spots to visit so that we can make it easier for a new business person to choose the right neighbourhood for their restaurant/shop.

1.3 Interest

This report will be useful for those who want to start a new business in the state of Toronto. It will also be helpful for those who want to travel to Toronto and want to visit specif locations based on their interest. For example, if a tourist wants to visit multicultural restaurants in Toronto, then what neighbourhoods are best? They should ideally visit those neighbourhoods where multicultural restaurants are popular among the people. The clustering of the neighbourhoods based on the most visited spots allows people to decide where to travel and explore more. Clusters tell us what neighbourhoods are fairly similar to each other so the person can skip travelling to many of the same neighbourhoods.

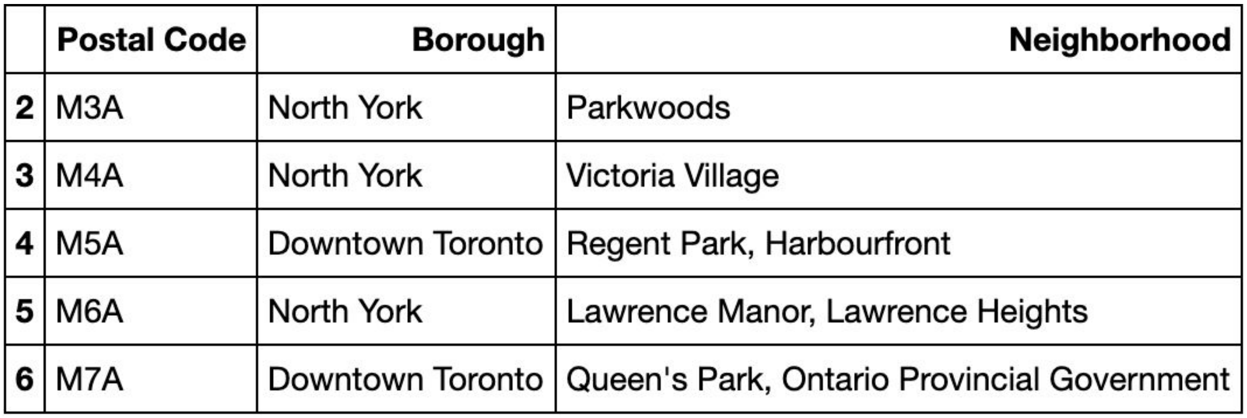
2. Data Acquisition and Cleaning

2.1 Data Available

We make use of a few data sources to get the data required for this project. We get different kinds of Neighborhoods in Toronto along with the Boroughs and Postal codes from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
We get the Latitudes and Longitudes for each postal code through the CSV file provided to us by the Coursera Applied Data Science Capstone Week 3 Module. We get the various location-related data, like the kinds of places in a particular neighbourhood, using Foursquare API. This data will include the type of shops, restaurants, cafes, beaches etc in each neighbourhood.

2.2 Acquiring the Data

We acquire the data about the various neighbourhoods, boroughs and postal codes from the Wikipedia page using Beautiful Soup. We put it into a data frame. The latitudes and longitudes are in a CSV file that can be read using pandas. We will make calls to Foursquare API using our credentials to acquire the location-related data.



	Postal Code	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

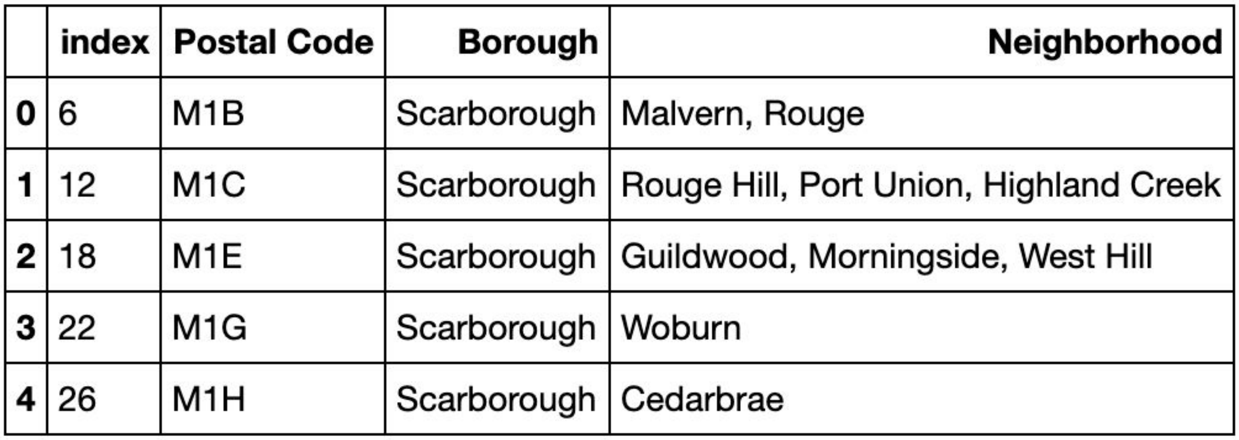
2.3 Cleaning the data

Once we have our data that includes the Postal Codes, Boroughs and Neighbourhoods in Canada, we drop all rows where the borough is unknown. We want to focus on the data that has been assigned a borough.

More than one neighbourhood can exist that has the same postal code, we combine such rows. We make a single row for each postal code and the subsequent neighbourhoods that are associated with that postal code would be put into the same row separated by commas.

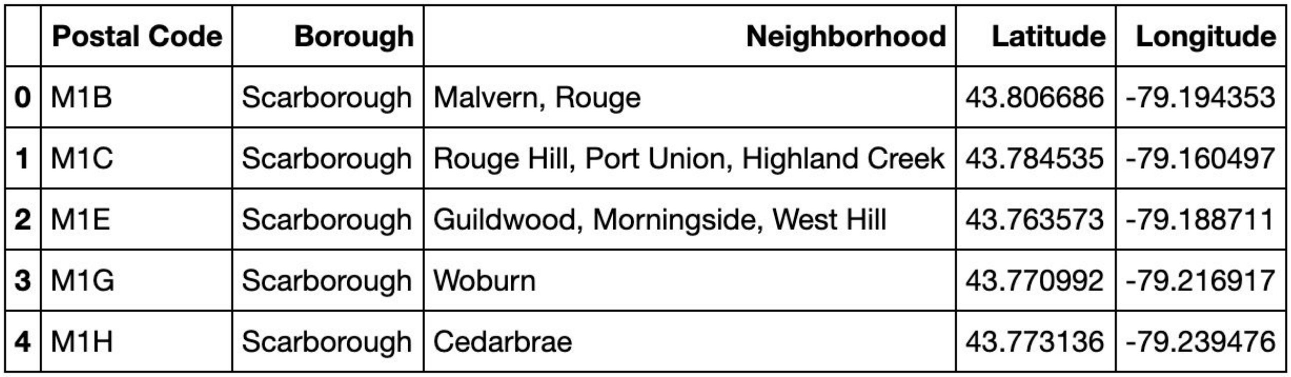
Is a particular neighbourhood is not assigned but that row has an assigned borough, then the neighbourhood is considered to be the same as the borough.

We sort the data by the postal codes.



	index	Postal Code	Borough	Neighborhood
0	6	M1B	Scarborough	Malvern, Rouge
1	12	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	18	M1E	Scarborough	Guildwood, Morningside, West Hill
3	22	M1G	Scarborough	Woburn
4	26	M1H	Scarborough	Cedarbrae

Then we merge the data frame with the stored values of the latitudes and longitudes of each postal code.

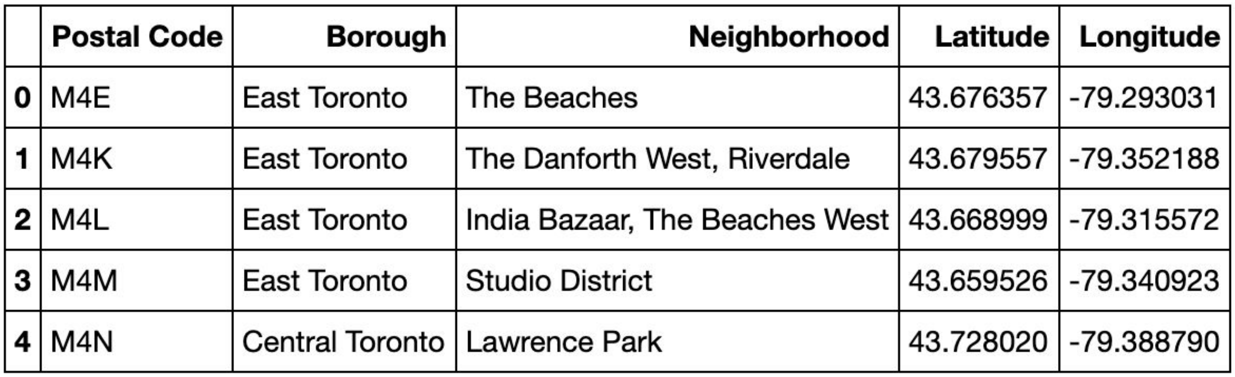


	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917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

Here we get our final data frame.

2.4 Feature Selection

We want to focus on neighbourhoods in Toronto. So we drop all rows that have Boroughs outside of Toronto. We keep the data from Toronto. Finally, our data frame is ready to use.



	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