**Battle of the Neighborhoods**

Finding the best Neighborhood in Toronto using Data Science

This project aims to utilize all Data Science Concepts learned in the IBM Data Science Professional Course. We define a Business Problem, the data that will be utilized and using that data, we are able to analyze it using Machine Learning tools. In this project, we will go through all the processes in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

Table of Contents

1. Introduction
2. Target Audience
3. Data Overview
4. Methodology
5. Discussion
6. Conclusion

3. Data Overview

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources which will provide the list of neighborhoods in Toronto (via Wikipedia), the Geographical location of the neighborhoods (via Geocoder package) and Venue data pertaining to Italian restaurants (via Foursquare). The Venue data will help find which neighborhood is best suitable to open an Italian restaurant.

3.1 Data acquisition:

Source 1: Toronto Neighborhoods via Wikipedia



Figure 1:Wikipedia Page showing List of Neighborhoods in Toronto with respective Postal Codes

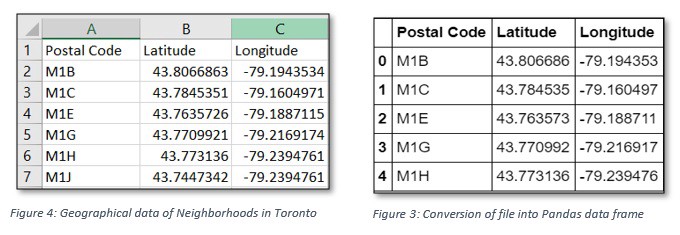
1. <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

The Wikipedia site shown above provided almost all the information about the neighborhoods. It included the postal code, borough and the name of the neighborhoods present in Toronto. Since the data is not in a format that is suitable for analysis, scraping of the data was done from this site (shown in *figure2*).



Figure 2: Data that was scraped from Wikipedia site and put into Pandas data frame

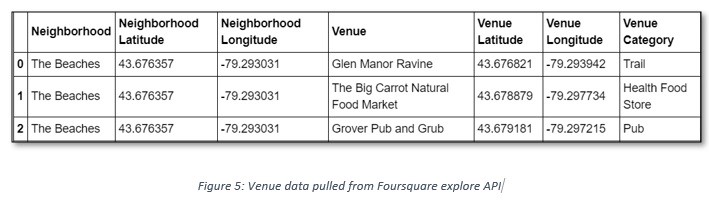
Source 2: Geographical Location data using Geocoder Package



2. <https://cocl.us/Geospatial_data>

The second source of data provided us with the Geographical coordinates of the neighborhoods with the respective Postal Codes. The file was in CSV format, so we had to attach it to a Pandas data frame(shown in figure 3).

Source 3: Venue Data using Foursquare



We performed a bit of data cleansing. It is seen through figure 5 (above) that the neighborhoods are grouped by the name of the neighborhood, so data clustering is made easier later.

4. Methodology

4.1 — Data Cleansing

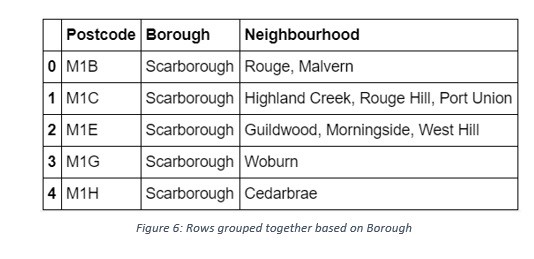
After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were Boroughs that were not assigned to any neighborhood therefore, the following assumptions were made:

1. Only the cells that have an assigned borough will be processed. Borough’s that were not assigned get ignored.

2. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in *Figure2*row 4.

3. If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

After the implementation of the following assumptions, the rows were grouped based on the borough as shown below.



Using the Latitude and Longitude collected from the Geocoder package, we merged the two tables together based on Postal Code.

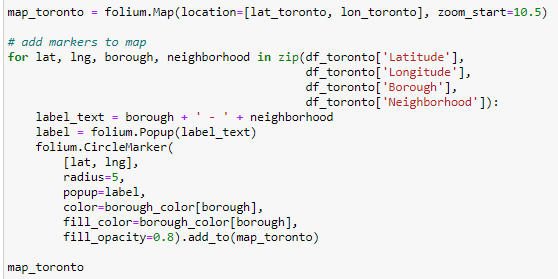


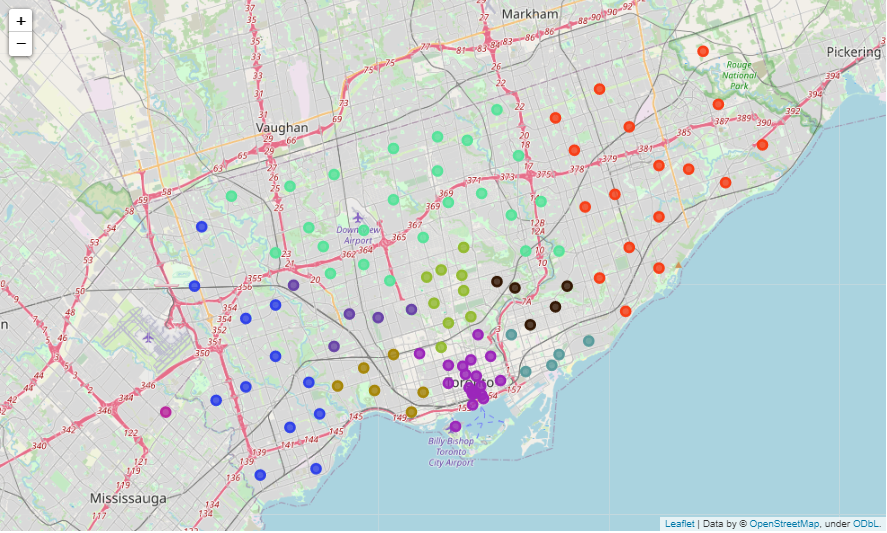
After, the venue data pulled from the Foursquare API was merged with the table above providing us with the local venue within a 500-meter radius shown below.



4.2 — Data Exploration

Now after cleansing the data, the next step was to analyze it. We then created a map using Folium and color-coded each Neighborhood depending on what Borough it was located in.



This snippet of code provided us with the map below:

Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number of Italian Restaurants all over Toronto. There was a total of 45 Italian Restaurants in Toronto. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.

