

ADEBAYO ABDUL-MALIK| IBM APPLIED DATA SCIENCE CAPSTONE | August 25, 2020

**Analyzing neighborhoods Toronto, Canada to determine the best location to open a restaurant**

As a part of the IBM Data Science professional program Capstone Project, we worked on the real datasets to get an experience of what a data scientist goes through in real life. Main objectives of this project were to define a business problem, look for data in the web and, use Foursquare location data to compare different neighborhoods of Toronto to figure out which neighborhood is suitable for starting a new restaurant business. In this project, we will go through all the process 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.

**Introduction**

Toronto, the capital of the province of Ontario, is the most populous Canadian city. Its diversity is reflected in Toronto’s ethnic neighborhoods such as Chinatown, Corso Italia, Greektown, Kensington Market, Koreatown, Little India, Little Italy, Little Jamaica, Little Portugal & Roncesvalles. In this project we will go through step by step process to make a decision on the best location to open a restaurant in Toronto. We analyze the neighborhoods in Toronto to identify the most profitable area since the success of the restaurant depends on the people and ambience. Since we already know that Toronto is the financial capital of Canada, it is a good idea to start the restaurant here, but we just need to make sure it is a profitable idea. If so, where we can place it, so it yields more profit to the owner.

**Target Audience**

**Who will be more interested in this project? What type of clients or a group of people would benefit from this project?**

* Business personnel who wants to invest or open a restaurant in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting the population in Toronto.
* Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.
* Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

**Business Problem**

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. Toronto is the fastest growing city in North America. and is the anchor of an urban agglomeration, known as the Golden Horseshoe in Southern Ontario, located on the northwestern shore of Lake Ontario.

Toronto encompasses a geographical area formerly administered by many separate municipalities. These municipalities have each developed a distinct history and identity over the years, and their names remain in common use among Torontonians. Former municipalities include East York, Etobicoke, Forest Hill, Mimico, North York, Parkdale, Scarborough, Swansea, Weston and York. Throughout the city there exist hundreds of small neighbourhoods and some larger neighborhoods covering a few square kilometers.

Having such vast population and big geographical area, there also exists big competition between businesses. Therefore, it became very challenging for stake holder or new business to decide which area they should start their business to get higher revenue with lowest possible competition.

This project tries to solve the above problem by suggesting to the target Audience the best location to open new restaurants and gaining maximum profits in Toronto.

**Data acquisition and cleaning**

**Data Sources**

* I’ll be using “List of Postal code of Canada: M” (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.
* Then I’m using “[https://cocl.us/Geospatial\_data”](https://cocl.us/Geospatial_data%E2%80%9D) csv file to get all the geographical coordinates of the neighborhoods.
* To get location and other information about various venues in Toronto I’m using Foursquare’s explore API. Using the Foursquare’s explore API (which gives venues recommendations), I’m fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).

From Foursquare API (<https://developer.foursquare.com/docs)>, I retrieved the following for each venue:

* Name: The name of the venue.
* Category: The category type as defined by the API.
* Latitude: The latitude value of the venue.
* Longitude: The longitude value of the venue.

**METHODOLOGY**

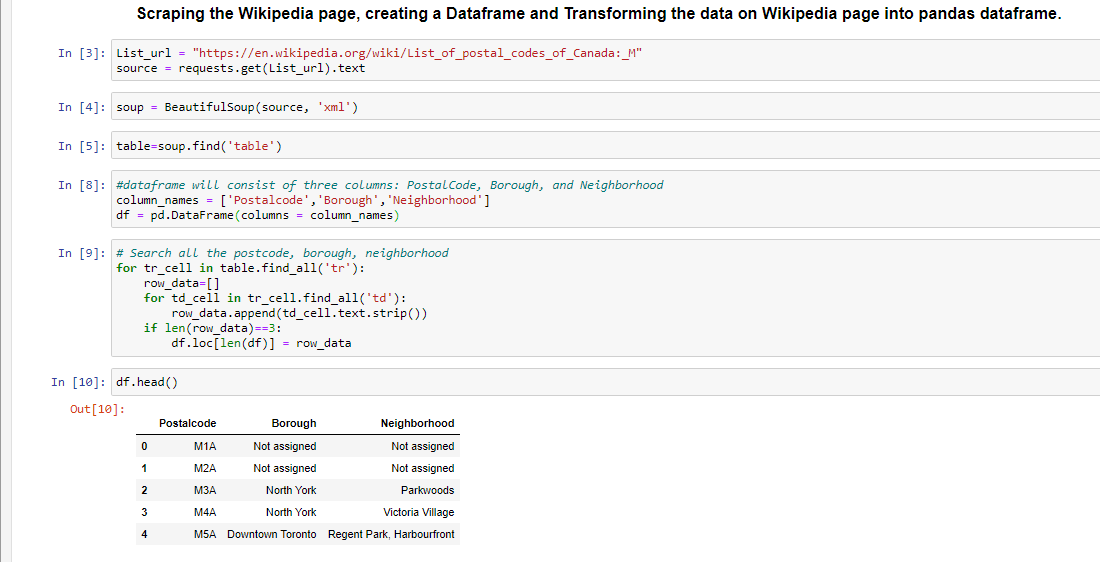
**Scraping the Wikipedia page, creating a dataframe and Transforming the data on Wikipedia page into pandas dataframe.**

Scraped the following Wikipedia page “List of Postal code of Canada: M” in order to obtain the data about the Toronto and its Neighborhoods.

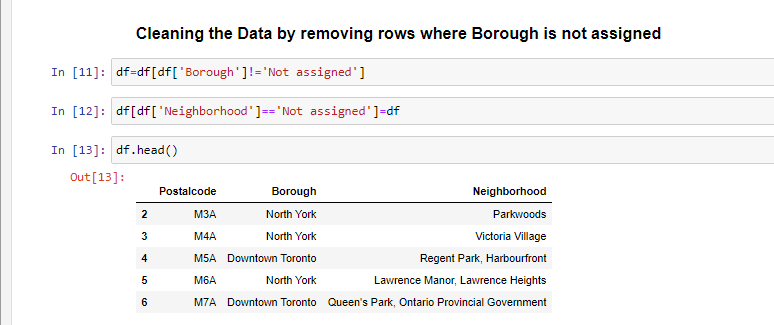
Assumptions made to attain the below Data Frame:

* Data frame will consist of three columns: Postal Code, Borough, and Neighborhood
* Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
* 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 row 11 in the above table.
* If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

I used beautiful soup to scrape the data from Wikipedia

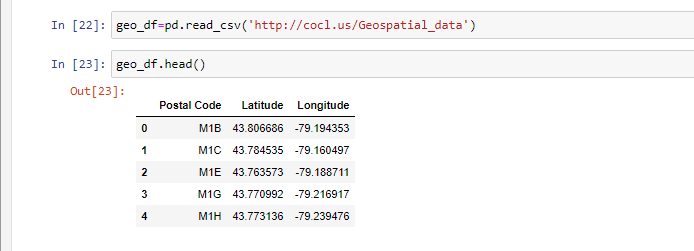


After some cleaning we got the proper dataframe with the Postal code, Borough & Neighborhood information.

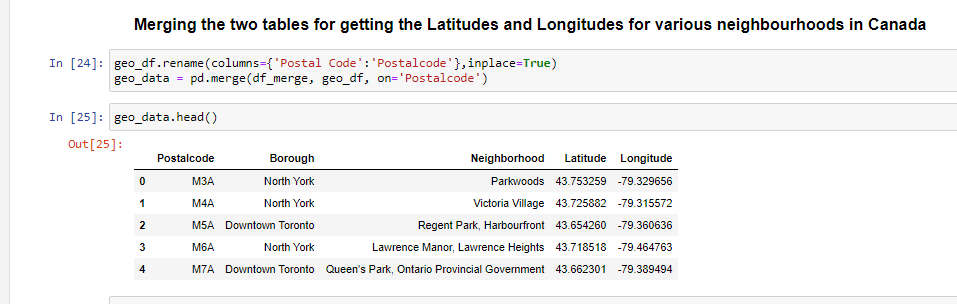


**Adding geographical coordinates to the neighborhoods**

Next important step is adding the geographical coordinates to these neighborhoods. To do so I’m extracting the data present in the Geospatial Data csv file and I’m combining it with the existing neighborhood dataframe by merging them both based on the postal code.

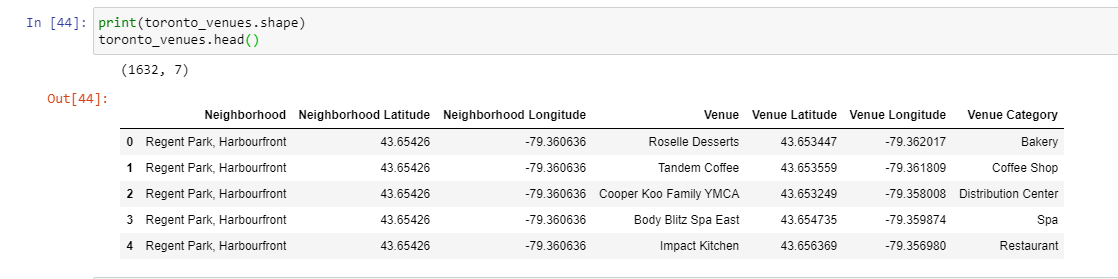


I’m renaming the columns to match the existing dataframe formed from ‘*List of Postal code of Canada: M’ wiki page.* After that I’m merging the two data frames.



**Get location data using Foursquare**

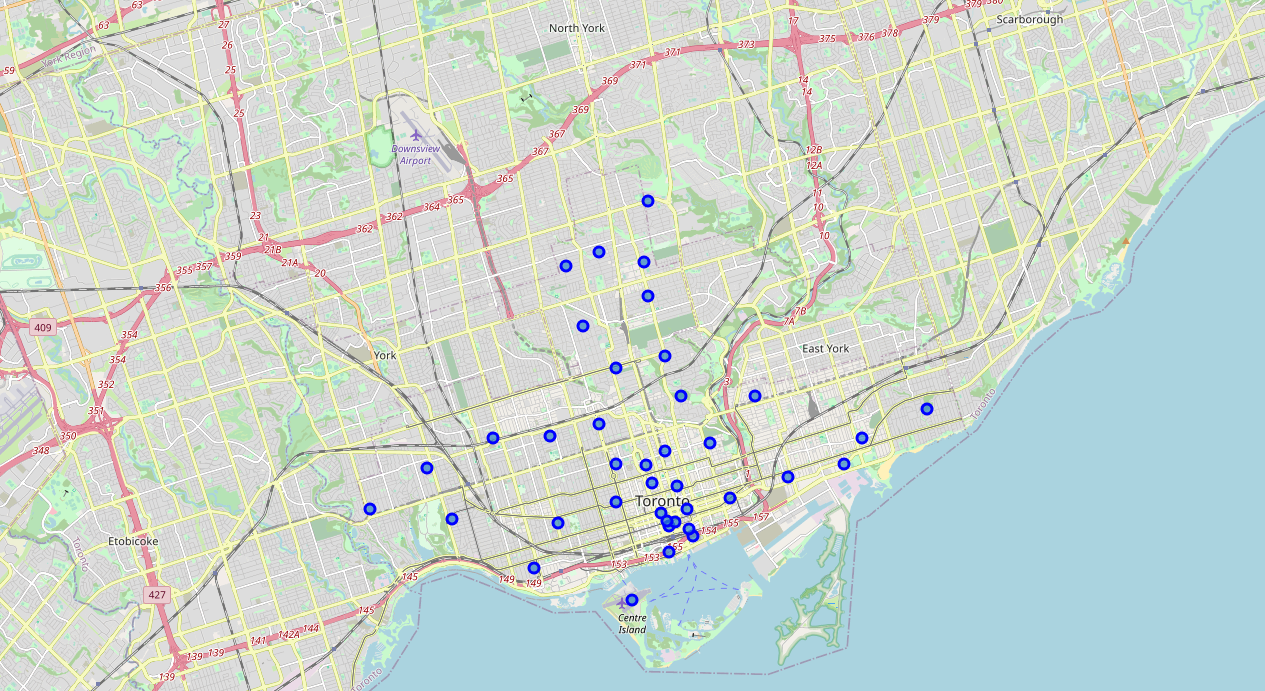
Foursquare API is very useful online application used my many developers & other applications like Uber etc. In this project I have used it to retrieve information about the places present in the neighborhoods of Toronto. The API returns a JSON file and we need to turn that into a data-frame. Here I’ve chosen 100 popular spots for each neighborhood within a radius of 1km.



**Exploratory Data Analysis**

**Folium Library and Leaflet Map**

Folium is a python library; I’m using it to draw an interactive leaflet map using coordinate data.



#### **Clustering**

With an assumption of 5 clusters, use K-Cluster algorithm to come up with 5 different clusters in Toronto with similar set of Venues.Explore each cluster and determine the discriminating venue categories that distinguish each cluster. Identify the clusters & Boroughs/Neighborhoods with Maximum number restaurants and their types.

**Results and Discussion**

We have reached the end of the analysis, in this section we will document all the findings from above clustering & visualization of the dataset. In this project, we started off with the business problem of identifying a good neighborhood to open a new restaurant. To achieve that we looked into all the neighborhoods in Toronto, analyzed the population in each neighborhood & number of restaurants in those neighborhoods to come to conclusion about which neighborhood would be a better spot.

The analysis shows that although there is a great number of restaurants in Toronto, there are pockets of low restaurant density fairly close to city center. We have 4 boroughs and 74 neighborhoods inside geograpical coordinate of 43.653963, -79.387207.

Based on our initial assumption of the cluster with maximum number of restaurants will have the best possibility to have a new restaurant due to the need in the area. Based on the resultant clusters it looks like Cluster 1 and Cluster 5 have higher number of restaurants then rest of the clusters.

It is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

**CONCLUSION**

Finally to conclude this project, We have got a chance to on a business problem like how a real like data scientists would do. We have used many python libraries to fetch the data , to manipulate the contents & to analyze and visualize those datasets. We have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then get good amount of data from Wikipedia which we scraped with help of Wikipedia python library. We also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.

Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Similarly we can use this project to analysis any scenario such as opening a different cuisine restaurant or opening of a new gym and etc. Hopefully, this project helps acts as initial guidance to take more complex real-life challenges using data-science.