**The Battle of Neighborhoods**

# Introduction

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario’s northwestern shore. With a recorded population of 2,731,571 in 2016, it is the [most populous city in Canada](https://en.wikipedia.org/wiki/List_of_the_100_largest_municipalities_in_Canada_by_population) and the [fourth most populous city in North America](https://en.wikipedia.org/wiki/List_of_North_American_cities_by_population). Toronto is a financial capital of Canada and is a land of opportunity for entrepreneurs. Knowing that Toronto is a multicultural city and a dream of a ‘foodie’, we would like to use Foursquare location data and regional clustering of venue information to determine what might be the ‘best’ neighborhood in Toronto to open a new Greek restaurant.

The purpose of this whole exercise is for submission of the final capstone project for the "IBM Data Science" course on Coursera as well as to determine what neighborhood of Toronto would be the most optimal to open a new Greek Restaurant. Target audience for this analysis would be individuals/entrepreneurs who are planning to open a Greek Restaurant in Toronto and are looking for the optimal neighbourhood to do it.

# Data

1. List of postal codes of Canada Wiki: [**https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M**](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) for access to neighborhood data of Toronto region.
2. Geographical coordinates of the neighborhoods: [**http://cocl.us/Geospatial\_data**](http://cocl.us/Geospatial_data) for getting the longitude and latitude data for the neighborhoods.
3. Foursquare database: [**https://Foursquare.com**](https://foursquare.com/) to be used in order to explore the desired neighborhood data for various restaurant details and access the JSON files. This data shall be utilized to map the Indian restaurants in various locations.

# Methodology

## Data Collection

First, we extracted the data from the data sources **listed above.** The Wikipedia site (https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M), provided almost all the information about the neighborhoods. It included the postal code; borough and the name of the neighborhoods present in Toronto. However, the data needed to be converted to the appropriate data frame for this analysis. In order to move ahead with analysis, we used Geographical coordinates of the neighborhoods csv file to add latitude and longitude to our analysis. The retrieval of the location, name and category about the various venues in Toronto was collected through the Foursquare explore API.

## Data Preparation

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 that is not assigned are 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.

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

## Analysis

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. 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 Greek Restaurants in Toronto. There was a total of 15 Greek 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. Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called **One hot encoding**. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood. Then we grouped those rows by Neighborhood and by taking the **Average** of the frequency of occurrence of each Venue Category. After, we created a new data frame which only stored the Neighborhood names as well as the mean frequency of Italian Restaurants in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze.

To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Greek Restaurants in that Neighborhood. To do this we used K-Means clustering. To get our optimum K value that was neither overfitting or underfitting the model, we used the Elbow Point Technique. In this technique we ran a test with different number of K values and measured the accuracy and then chose the best K value. The best K value is chosen at the point in which the line has a sharpest turn. In our case we had the Elbow Point at K = 4. That means we will have a total of 4 clusters.

Chart, line chart

Description automatically generated

We integrated a model which would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=4. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster. Neighborhoods that had similar mean frequency of Italian Restaurants were divided into 4 clusters. Each of these clusters were labelled from 0 to 3 as the indexing of labels begin with 0 instead of 1. After, we merged the venue data with the table above creating a new table which would be the basis for analyzing new opportunities for opening a new Greek Restaurant in Toronto. Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label.

# Results

We have a total of 4 clusters (0,1,2,3). Before we analyze them one by one lets check the total amount of neighborhoods in each cluster and the average number of Greek Restaurants in that cluster. We see that Cluster 2 has the least neighborhoods (1) while cluster 1 has the most (91). Cluster 3 has 5 neighborhoods and cluster 4 has only 2.

Chart

Description automatically generated

Then we compared the average Greek Restaurants per cluster. This information is crucial as we can see that even through there is only 1 neighborhood in Cluster 2, it has the highest number of Greek Restaurants while Cluster 1 has the most neighborhoods but has the least average of Greek Restaurants.

Chart, bar chart

Description automatically generated

# Discussion

Most of the Greek Restaurants are in cluster 2 represented by the purple clusters, that would probably mean that opening a Greek Restaurant in The Danforth West and Riverdale (East Toronto) wouldn't be a great idea, since there is huge competition. Even though there is a huge amount of Neighborhoods in cluster 1, there are no Greek Restaurants at all, so it would be one of the optimal neighborhoods to consider when planning to open a new Greek Restaurant. We also see that the North York area (cluster 3) has the second last average of Greek Restaurants. Looking at the nearby venues, Fairview, Henry Farm, and Oriole, would be a great opportunity to open up a new Greek Restaurant. We wouldn't suggest opening a Greek Restaurant in Central Toronto, as there is also a lot of competition. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Italian restaurant in these locations with little to no competition.

# Conclusion

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in way that it was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, to control the content and to break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Toronto, get great measure of data from Wikipedia which we scraped with the Beautifulsoup Web scraping Library. We also visualized utilizing different plots present in seaborn and matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.

Places that have room for improvement or certain drawbacks gives us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Liquor Store and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data-science.