**The Battle of Neighbourhoods**

**Author: Manideep Sadineni** **Date: 01/08/2021**

This report is the part of capstone project of Certificate. The main aim of this blog post is to utilise all the concept we’ve learned from this certification for solving a business problem where we can use the Foursquare location data. Let’s see what we are going to solve.

**Table of Content**

1. Business Problem
2. Target Audience
3. Data Description
4. Methodology
5. Results & Outcomes
6. Conclusion

**1. Business Problem**

With a population just short of 3 million people, the city of Toronto is the largest in Canada, and one of the largest in North America (behind only Mexico City, New York and Los Angeles). Toronto is also one of the most multicultural cities in the world, making life in Toronto a wonderful multicultural experience for all. More than 140 languages and dialects are spoken in the city, and almost half the population Toronto were born outside Canada. It is a place where people can try the best of each culture, either while they work or just passing through. Toronto is well known for its great food.

The objective of this project is to find the best neighbourhood in Toronto to open a restaurant using Foursquare location data. In this project we’ll go through the solution for this problem for avoiding or considering low risk criteria and high success rate.

**2. Target Audience**

* Business personnel who want to invest or open a restaurant.
* The freelancer who loves to have their own restaurant as a side business.
* Tourists who want to eat Italian food

**3. Data Description**

For this project we need the following data:  
***1. Toronto City data that contains Borough, Neighbourhoods along with their latitudes and longitudes***

* **Data Source**: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
* **Description:** This Wikipedia page contain all the information we need to explore and cluster the neighbourhoods in Toronto. We will be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a *panda’s* data frame so that it is in a structured format like the Toronto dataset.

***2. Geographical Location data using Geocoder Package***

* **Data Source:** <https://cocl.us/Geospatial_data>
* **Description:** The second source of data provided us with the Geographical coordinates of the neighbourhoods with the respective Postal Codes.

***3. Venue Data using Foursquare API***

* **Data Source:** <https://foursquare.com/developers/apps>
* **Description:**From Foursquare API we can get the name, category, latitude, longitude for each venue.



Neighbourhood dataset of Toronto

**4. Methodology**

After scraping the data from Wikipedia there were Boroughs that were not assigned to any neighbourhood therefore, the following assumptions were made:

* Only process the cells that have an assigned borough. Ignore cells with a borough that is **Not assigned.**
* More than one neighbourhood 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 neighbourhoods: **Harbourfront**and **Regent Park**. These two rows will be combined into one row with the neighbourhoods separated with a comma as shown in **row 11**in the above table.
* If a cell has a borough but a **Not assigned**neighbourhood, then the neighbourhood will be the same as the borough.

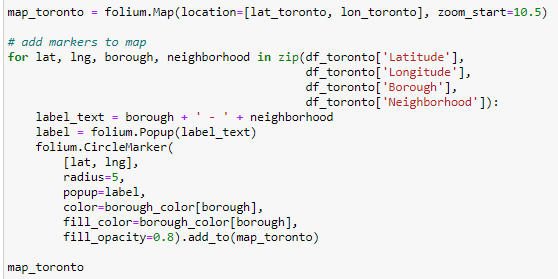
We will merge the two tables together based on Postal Code using the Latitude and Longitude collected from the Geocoder package.



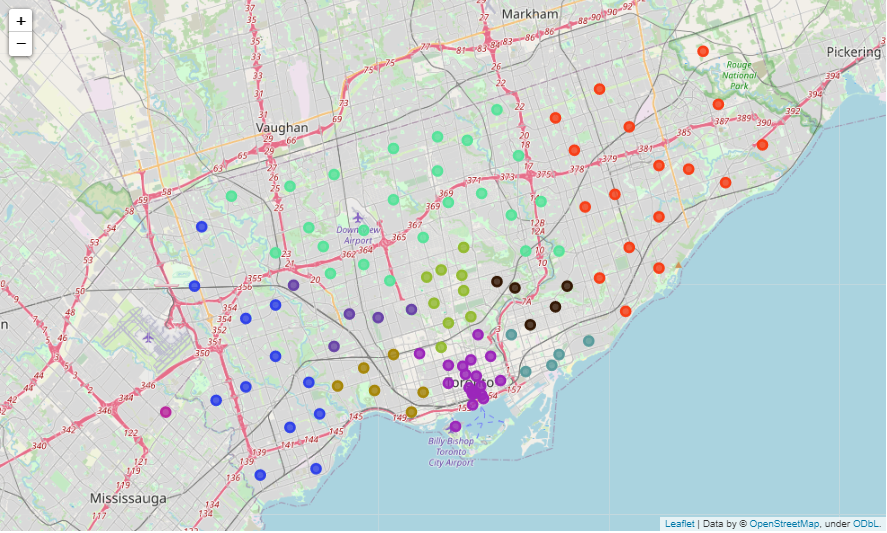
Now we will retrieve the venue data present within 500 meter radius of each neighbourhood using Foursquare API and merge with the above table.



Now we need to visualise all neighbourhoods in a map using Folium and colour-coded each. The below bunch of code needed to do so.



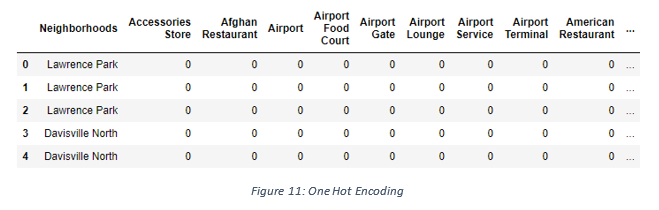
This snippet of code provided us with the map below:



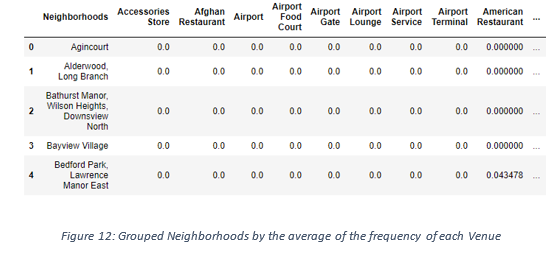
Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analysing the number of Italian Restaurants all over Toronto. There was a total of 45 Italian Restaurants in Toronto. We then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

**Data Pre-processing**

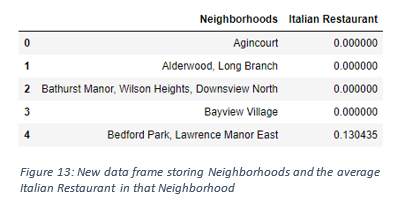
To analyse the respective Italian restaurant, present in that neighbourhood or not, we’ll use **One hot encoding** technique. For each of the neighbourhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.



Then we grouped those rows by Neighbourhood and by taking the **average** of the frequency of occurrence of each Venue Category.

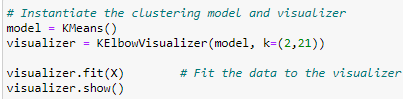


After, we created a new data frame that only stored the Neighbourhood names as well as the mean frequency of Italian Restaurants in that Neighbourhood. This allowed the data to be summarized based on each individual Neighbourhood and made the data much simpler to analyse.



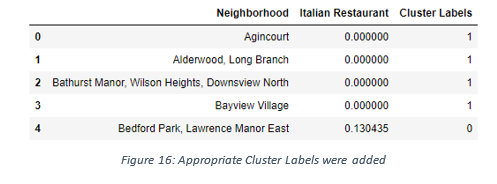
**K-Means Clustering**

Now we’ll cluster these neighbourhoods based on the frequency of Italian restaurants present. To do this we apply k-means clustering algorithm. To avoid the overfitting and underfitting of the model we need an optimum value of **“k”**. There are many techniques like **Elbow method**, **Silhouette score** method to get the best **“k”** value. Here we’re going to use **Elbow method** to get best “k” value. We’ll import ‘*K Elbow Visualizer*’ from the *yellow brick package.*Then we fit our K-Means model above to the Elbow visualizer.



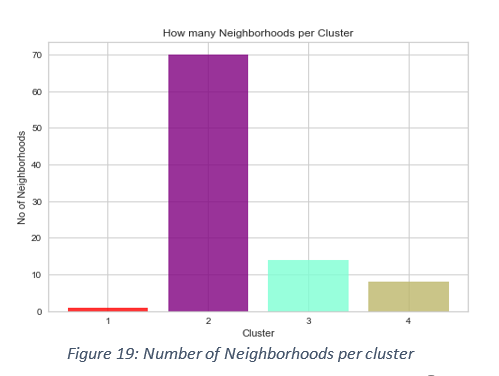
This bunch of code will give this below graph

Here, we can see that the best k value for our dataset is 4. That means we will cluster the dataset into 4 cluster. Each of these clusters was labelled from 0 to 3 as the indexing of labels begins with 0 instead of 1.

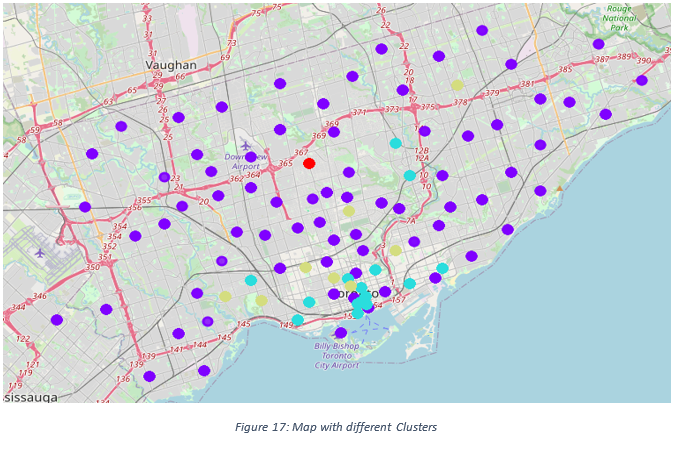


**5. Result & Outcomes**

The below bar chart shows how many neighbourhoods present in each cluster.



The map below shows the different clusters that had a similar mean frequency of Italian restaurants.



**6. Conclusion**

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was like how a genuine data scientist would do.