**The Battle Of Neighbourhoods**

**Author: Alejandro Gutiérrez Sanjurjo** **Date: 23/08/2021**

This report is the final part of capstone project of [IBM Data Science Professional Certificate](https://www.coursera.org/professional-certificates/ibm-data-science) . The main aim of this report is to use all the concept we’ve learned from the previous courses for solving a business problem where we can use the Foursquare location data.

**Table of Content**

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

**1. Business Problem**

The city of Toronto is the largest city in Canada, and one of the largest cities in North America. 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 the country. 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.

With all the options present in the city to taste so many different foods, it is necessary to offer a space where the people can keep the fit to avoid some serious health problems such as cholesterol, diabetes, etc

The objective of this project is to find the best neighbourhood in Toronto to open a Gym using foursquare location data to analyse a little bit the current location of the gyms.

**2. Target Audience**

* Business personnel who wants to invest or open a gym.
* The freelancer who loves to have their own gym as a side business.
* All the people who want to keep healthy

**3. Data Description**

For this project we need the following data:  
***1. Toronto City data (Boroughs, Neighbourhoods)***

* **Data Source**: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
* **Description:** Wikipedia contains all the information we need to explore and cluster the neighborhoods in Toronto. We will be required to web scrape the Wikipedia page and wrangle the data, clean it, and then convert it into a *pandas* dataframe

***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.

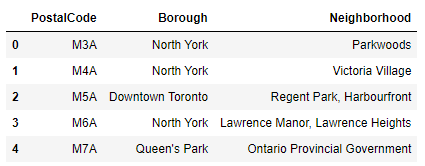


Image 1. Neighbourhood dataset of Toronto

**4. Methodology**

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

* 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.

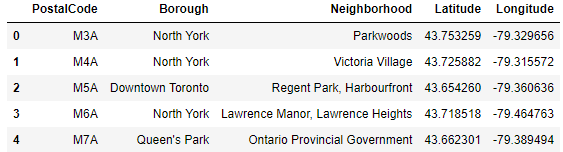
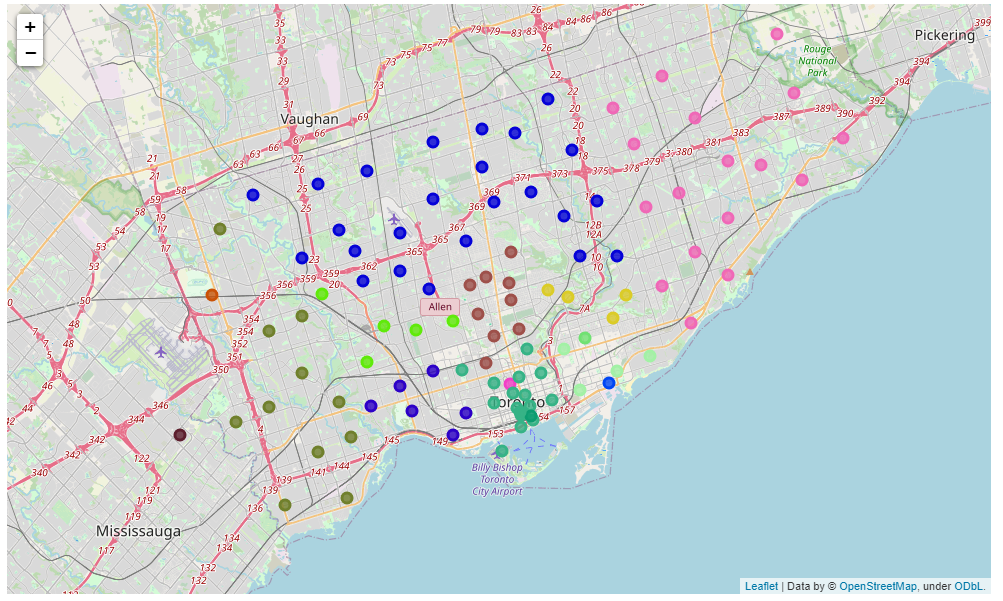


Image 2. Dataset with Geocoder info

Now we will retrieve the venue data present within 1000 meters radius of each neighbourhood using Foursquare API and merge with the above table. With this, we will know what types of venues we can find in Toronto. We have 330 unique venue categories.

Now we need to visualise all neighbourhoods in a map using Folium and colour-coded each. You can find the code on the Jupyter notebook.

The result of the use of Folium package is the following map:



Next, we have to see how many of the venues are of the Gym type. There was a total of 102 gyms in Toronto. Then, we merge 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 gym 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.

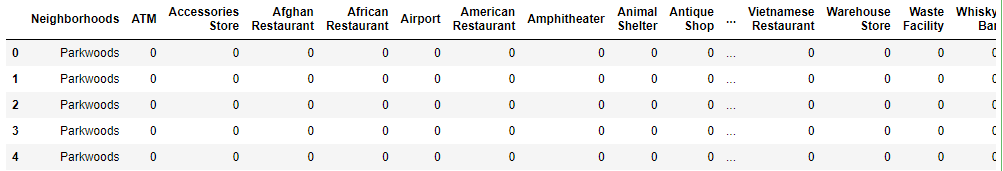


Image 3. One hot encoding using Dummies

Then we grouped those rows by Neighbourhood and by taking the **average** of the frequency of occurrence of each Venue Category. After that, we filter by “Gym” to have only the data we are interest on:



Image 4. Average of the freq of gyms

**K-Means Clustering**

Now we’ll cluster these neighbourhoods based on the frequency of gyms present on the city. 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”**. Here we’re going to use **Elbow method** to get best “k” value. We’ll import ‘*KElbowVisualizer*’ from the *Yellowbrick package.*Then we fit our K-Means model above to the Elbow visualizer

The graph which shows the best K is the following:

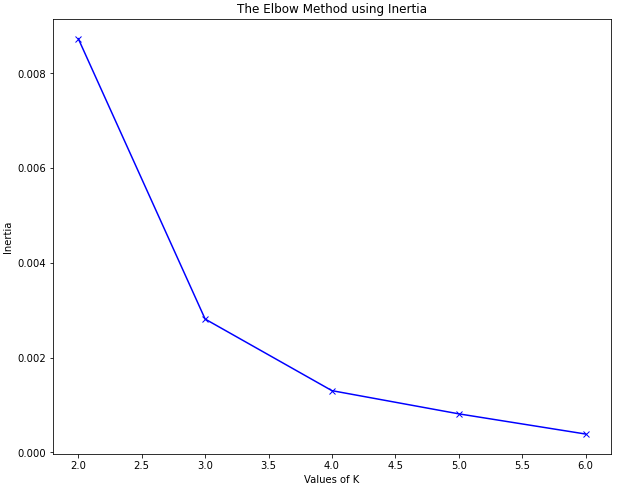


Image 5. Determine the value of "k"

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.

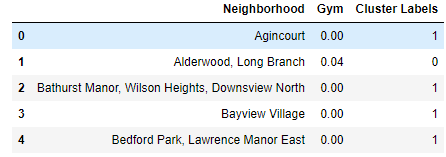


Image 6. Gym labelled by K-means method

**5. Result & Outcomes**

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

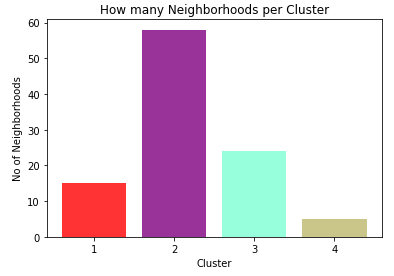


Image 7. Neighbourhood per cluster

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

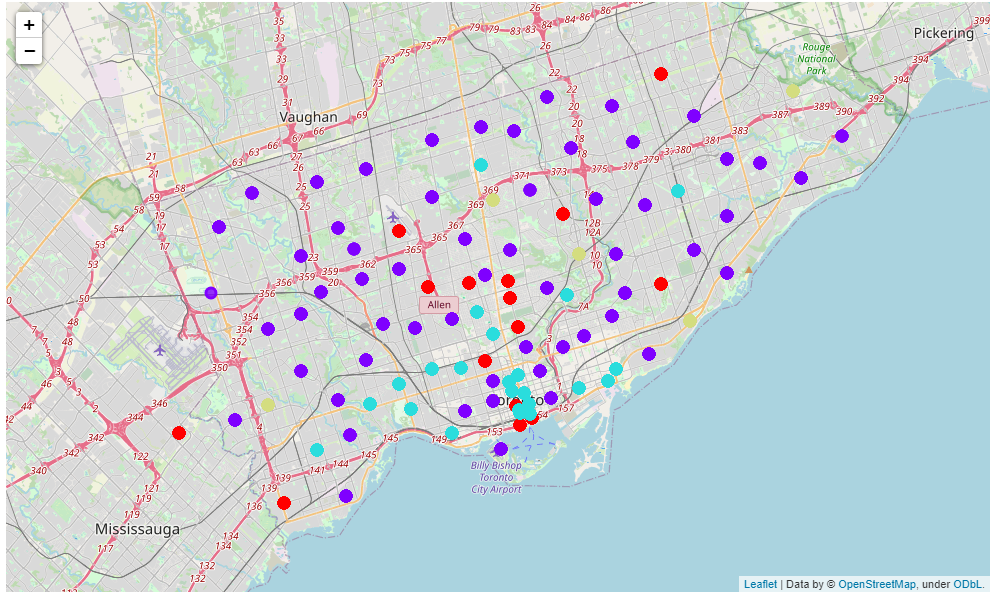


Image 8. Map with the clusters

**6. Conclusion**

In conclusion, to end off this project, we had an opportunity to analyse a little bit a business problem. This is only the beginning, because there are more work to do after obtaining this information, such as analyse more deeply with this information the best place to put a gym or what type or gym are more profitable depends of the neighbourhood. This is only an insight of how we can solve a real world problem.

Thank you very much for the attention.