**IBM Coursera Capstone**

**Toronto Neighborhood Analysis**

**IBM Applied Data Science Capstone**

**By**

**Belachew Ayele**

**January 2021**

**Table of Content Page**

1. Introduction ………………………………………………………………………………… 1
2. Business Problem ………………………………………………………………………… 1
3. Target Audience of the project ……………………………………………………. 1
4. Data, source ………………………………………………………………………………… 1
   1. Data from Wiki ……………………………………………………………………… 2
   2. Geo Data………………………………………………………………………………. 2

4.3. Foursquare Data …………………………………………………………………… 5

1. Intergrading Data ………………………………………………………………………… 5
2. Analysis of Data …………………………………………………………………………… 7

6.1. Pre - Clustering ……………………………………………………………………. 7

6.2. Clustering ……………………………………………………………………………. 9

1. Finding/Result ……………………………………………………………………………. 12
2. Conclusion …………………………………………………………………………………. 14
3. Reference ……………………………………………………………………….…………. 15

## Introduction

This is a project to conclude the IBM Data science courses. The project is facilitated with the data from the wiki, and Foursquare API data though a personal account. Python machine learning and use of several steps-by-step process is applied.

The data that involved in this project is focusing on Toronto area neighborhood. It includes many attributes including Toronto Neighborhoods, borough, geographic coordinates and other attributes that at revealed thought Foursquare API data acquire.

In this process, data is download, cleaned, processes, mapped and clustered.

1. **Business Problem**

The business problem is to get data and classify them into cluster using k-means. How data are acquired, processed and clustered, and will answer the type and classes of the foursquare data of Toronto and surrounding area. Data users may lean and use the method and results of this process.

1. **Objective / Target Audience of the project**

With the given set of objectives, this neighborhood has been implemented for the exploration of data using segmenting and clustering techniques applied to the neighborhoods data in Toronto. The objective of Applied Data Science Capstone is given as follows:

1. To learn about clustering and k-means clustering in particular.
2. To showcase this project in the form of the public repository using the GitHub platform.
3. To learn how to use the Foursquare API and clustering to segment and cluster the neighborhoods in Toronto City.
4. To learn how to use different Python package to scrape websites and parse HTML code.
5. To apply the skills acquired so far in this course to segment and cluster neighborhoods in the city of Toronto.

**Target Audience of the project**

This project is the project that can benefit several in the process,

* Any busines that will start work in Toronto and surrounding
* Personnel that are interested in studies and application of python processing such as students, practitioners, researches, professors, entrepreneurs and so on.
* Personal benefit in increasing data sciences knowledge in honing our knowledge.
* General practice in data science application

1. **Data**

Based on the problem statement, data sources are of three sources,

These are:

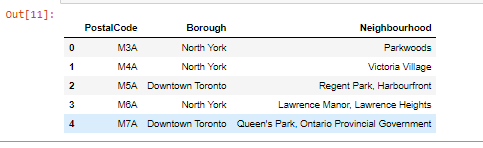
* 1. **Data from wiki**. The following data source of data will be needed to extract/generate to require information [**https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M**](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

This data is on the wiki with Toronto Postal code, Boroughs and Neighborhoods. As this was the data in the wiki, it is not having all the records with complete information. Some of the lines do not have Borough assigned. In such case the recorded gets neglected, however when the borough is assigned and Neighborhood not assigned, the Borough name was given as the Neighborhood. It was captured and cleaned using different of python libraries.

The following is the snip of the extracting script



Here is the partial data that is produced for the above script,



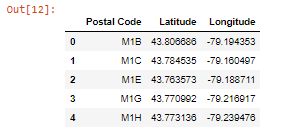
In the above, 103 records were collected

* 1. **Geo Data**: Centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas’ coordinates will be obtained using **Geocoder.** <https://cocl.us/Geospatial_data>

**Postal code coordinates are available in the geocoder site and were grabbed using the python command.**

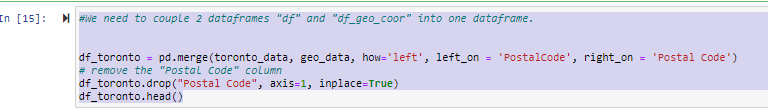


Here are the patial coodinate data of the Postal codes

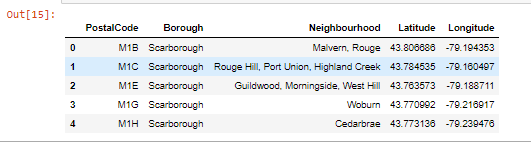


**However, the Postal code, Borough and Neighborhood data should get the coordinates from the geocoder data, by joining the two data sets using the postal code names, and drop the additional postal code.**

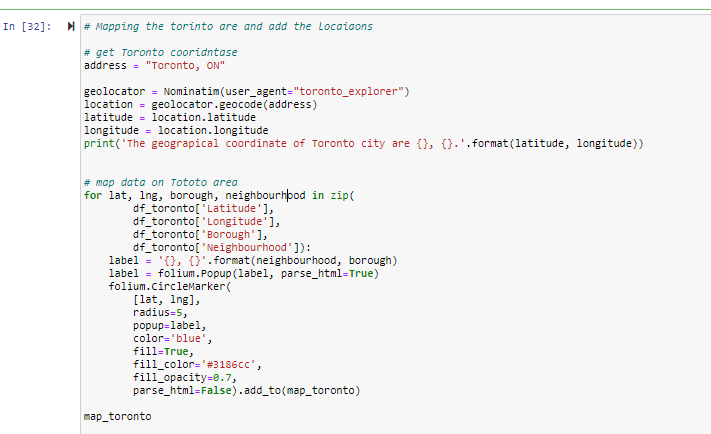
**This is the script used to join the two files and drop the additional postal code,**



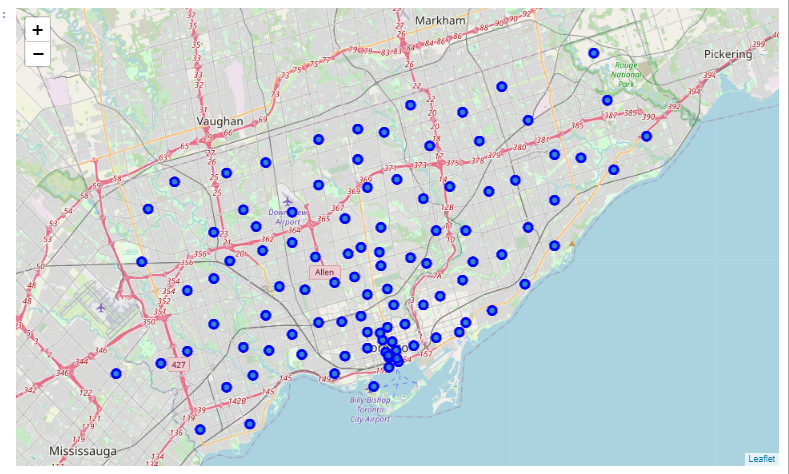
Here are the patial view of the Poctal code, Boough, Negibourhoud and the latitude and longitude coordiates



**As it is called, seeing is believing, above are going to be see which areas they cover on the maps, a folium app command was used to create a map.**

**This is the command script for the mapping of the joined result**

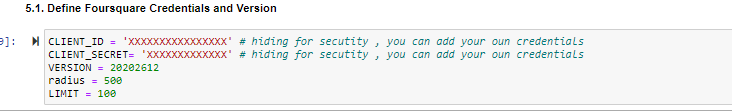
**Map of the joined files show all the record of 103, with postal code, borough, Neighborhood, longitude and latitude.**



* 1. **Foursquare Data**. Number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**

**One of the methods of getting the venue data is getting on the Foursquare option and get data by the type needed. Here in this project, we use the venue to get data surrounding the above mention table records. But to begin with the data extraction through foursquare, one has to get register and get a development permission. Then getting data form the foursquare API is possible.**

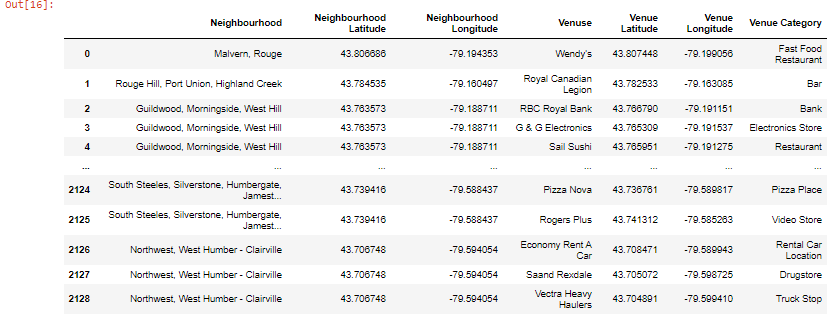
**After registering, the following credential were obtained.**



This is a script of extract data from the Foursquare API



**The foursquare process extracts 2129 data points with the covers that radius of 500 and limit of a 100. After steps of renaming the field names, below is the sample of a table. This table used in the analysis process to determine the next steps**

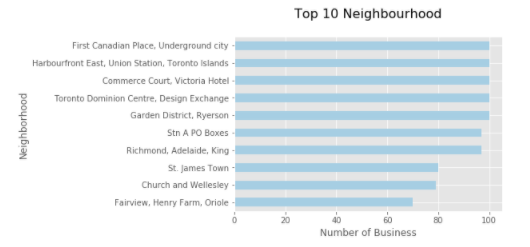


**In data exploration analysis is one step needed to be accomplished. Here down are the top and bottom to neighborhood with the number of businesses in them.**

**Counting the number of businesses (venues) in neighborhood that are in the top ten.**

1. **First Canadian Place, Underground City**
2. **Harbourfront East, Union Station, Toronto Islands**
3. **Commerce Court, Victoria Hotel**
4. **Toronto Dominion Centre**
5. **Garden District, Ryerson**

**Top 10 Neighborhood with number of business**



**Bottom 10 Neighborhood with number of business**



1. **Integrating data**

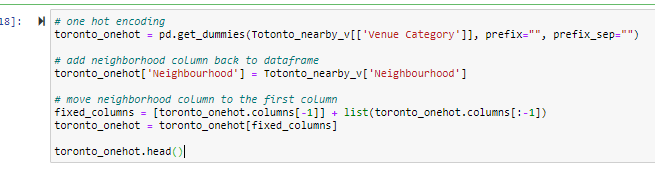
**This is the portion where the data from the wiki and the data acquired through four square API get integrated. Data should be integrated based on common characteristics or feature to join to one file. The coordinate will be shared/transferred from wiki data acquired through the API.**

1. **Analysis of Data**
   1. **Pre-Clustering – this is preparing the data for clustering**

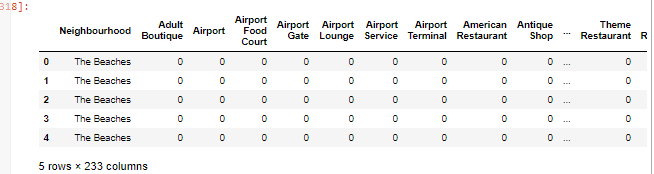
**Data acquired through the Foursquare method should go through a step-by-step process to be analyzed. These steps include**

* **Creating one hot file,**
* **Grouping the data**
* **Get the top five most common venues for teach neighborhood**

**Create onehot file - this is a process that creates a data frame that is testing for each neighbor and assign a zero or 1 based on the exitance of the venue in each Neighborhood with “1” = “exist” and “0” = “does not exist”. It is needed for the clustering of the file/data fame in to cluster categories. The script and the result table are in the snippets below. The one hot table is consisted of 2129 records and 274 features (a feature in here is the number of cases such as Adult Boutique, Airport, Airport food court, etc.)**



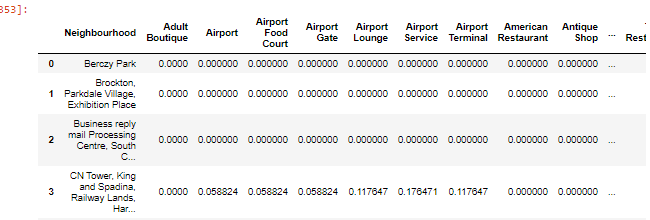
**This is partial one hot table**



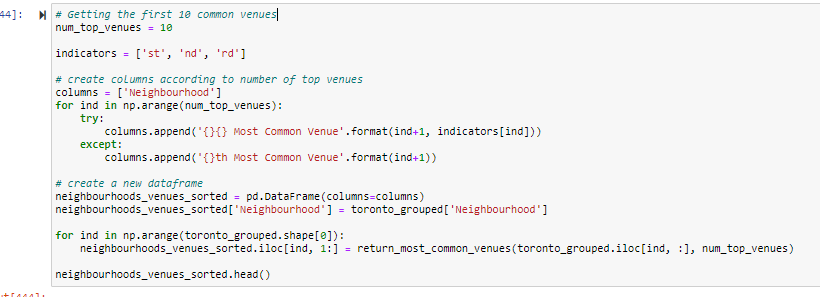
**Grouping the data - the above data which is the product on the one hot was grouped by neighborhood and by the mean of the frequency of occurrence of each category. This produced 96 Neighborhoods (as rows) and 274 columns. The simple scrip and the table as in the snippets below**



**This is the partial table result using the group by Neighborhood and get means of occurrents**



**Get the top five most common venues for teach neighborhood- this process extracts the top 10 common venues for each neighborhood and lists them in decreasing order. So, the 1st to 10th common venues table is in the table below**



**This is the partial top 10 common venue table of Neighborhoods**



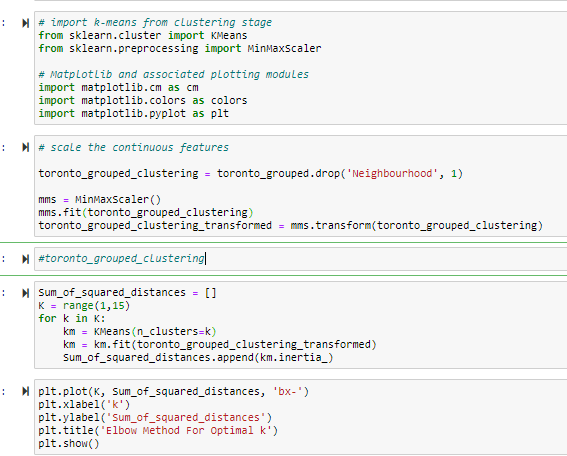
* 1. **Clustering**

**This is classifying the data into clusters.**

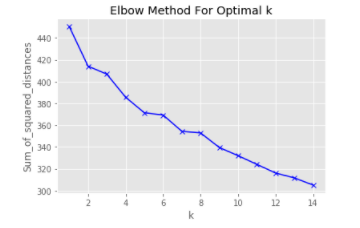
**The ultimate and main goal of this process int to get the clustered output of all the cases into different categories. However, after a thorough process of download, cleaning, join and other critical data processing, now is the time to go through running clustering algorithm or machine learning. To do that, it is usually preceded by determining how many clusters should be assigned? To do this a K-means clustering should be run.**

**Therefore, the following script is run to figure out how many classes were needed to cluster the data file**

**This is the scrip for K-means determination**



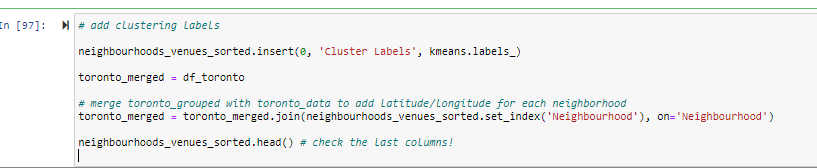
**This is the K-means, number of K for clustering**



**The K- means, roughly was determine where the elbow is curing up a lot (in this case it is hard to determine that but had been taken to be are 7). Note that every run gives a different curve. So, rerunning test again may differ the graph and the result.**

* **Cluster running**

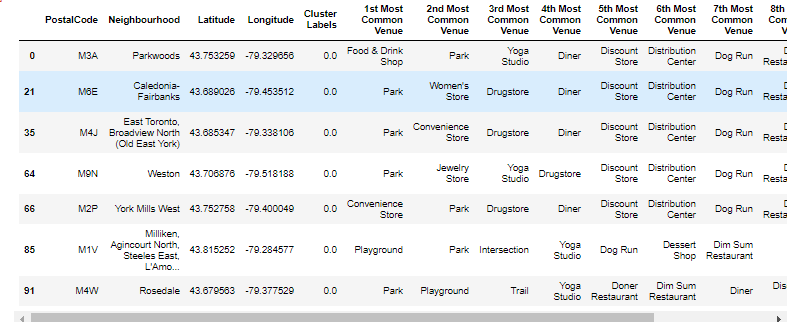
**Based on the number estimated in the k-means, 7 clusters were used in creating the cluster by tunning the script below.**

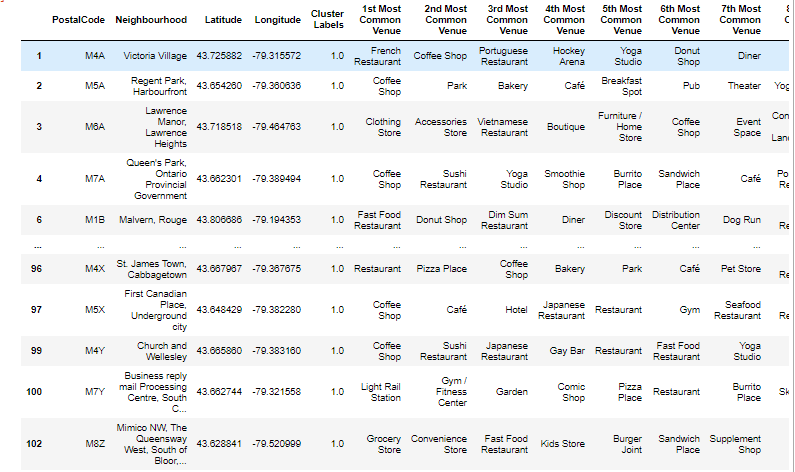


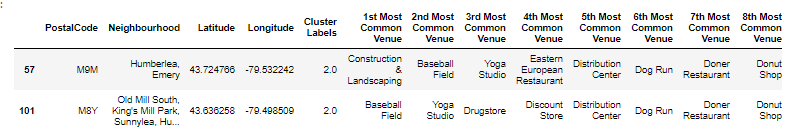
**Running the above, produced the following clusters:**

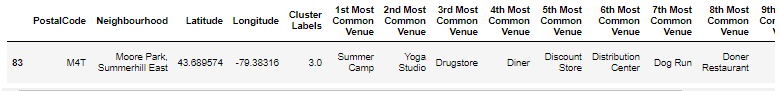
**(Cluster 0. Custer 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5, Cluster 6**

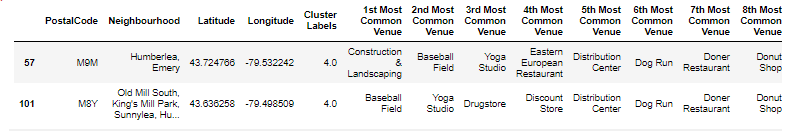
**Cluster 0**



**Cluster 1**

**Cluster 2**

**Cluster 3**

**Cluster 4**

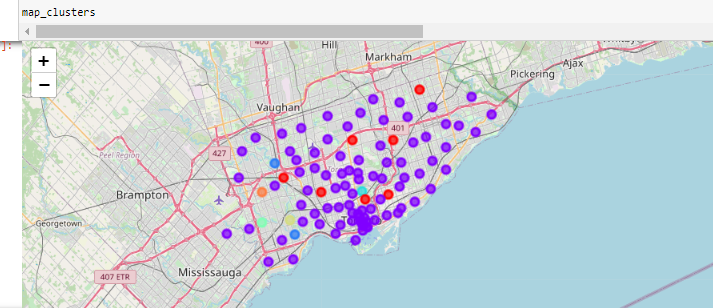
**Cluster 5**

**Cluster 6**

1. **Result and findings**

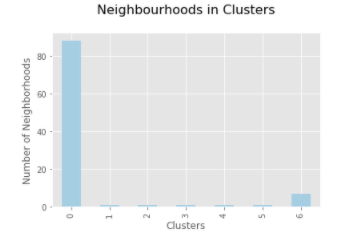
**Cluster classification was run using the data acquired and through Foursquare API and processed for one hot and extracted the top 10 venues in each neighborhood. A test for the K- means value and cluster was performed, followed by clustering the data. A clustering to 7 class as run and the following result was extracted.**

1. **Map of the custard business (folium)**
2. **Major classes in cluster**
3. **Count of Neighbors in each cluster**
4. **Graph of counts**

Cluster map:

|  |  |  |
| --- | --- | --- |
| Cluster | Number of Neigborhoods | Some of the Common Venues in cluster |
| 0 | 88 | Bars, Coffee Shopes, Pizza Places |
| 1 | 1 | Fast Food Restaurant, Yoga Studio |
| 2 | 1 | Martial Arts School, Electronic Store |
| 3 | 1 | Baseball Field, Dog Run |
| 4 | 1 | Filipino Restaurant, Eastern European Restaurant |
| 5 | 1 | Pizza Places |
| 6 | 7 | Playground, Parks, |

Number of Neighborhood in Cluster



1. **Conclusion**

The purpose of this project was to get location data in the Toronto area and classify them into classes/cluster. On the way, it is to learn about the uses of Foursquare API method of downloading data, process and analyze it, and present it to users. This has been demonstrated all along in the above processes. Moreover, anyone who needs to see the classified business in Toronto or who tries to replicate the process to get similar or better results, the foundation is built thought this process. Many can benefit out this process, that incudes but is not limited to researchers, students, teacher, business entrepreneurs, local government official and so forth.

Data was classified into 7 clusters. Although the k-means are not clear because it changes in every k-means run. However, the general principle remains the same.

**Reference**

1. **List of postal codes of Canada: M**

**en.wikipidia.org ---- (visited on December 2020)**

[**https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M**](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

1. **List of postal code from online ---- (visited on December 2020)**

<https://cocl.us/Geospatial_data>

1. **Several IBM data science notes and lectures from the IBM Coursera class**