**CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS (WEEK 2)**

By

**JHANSIRANI GUMMA**

**Contents**

[**1. INTRODUCTION** 2](#_Toc54011006)

[**2. DATA SOURCES** 3](#_Toc54011007)

[**3. METHODOLOGY** 5](#_Toc54011008)

[**4. DISCUSSION** 7](#_Toc54011009)

[**5. CONCLUSION** 8](#_Toc54011010)

**1. INTRODUCTION**

The purpose of this Capstone Project is to help people in exploring better facilities around their neighborhood. It will help people making smart and efficient decision on selecting great neighborhood out of numbers of other neighborhoods in Toronto.

Lots of people are migrating to various states of Canada and needed lots of research for good housing prices and reputed schools for their children. This project is for those people who are looking for better neighborhoods. For ease of accessing to Cafe, School, Supermarket, medical shops, grocery shops, mall, theatre, hospital etc.

This Capstone Project aim to create an analysis of features for a people migrating to Scarborough to search a best neighborhood as a comparative analysis between neighborhoods. The features include median housing price and better school according to ratings, crime rates of that particular area, road connectivity, weather conditions, good management for emergency, water resources both fresh and wastewater and excrement conveyed in sewers and recreational facilities.

It will help people to get awareness of the area and neighborhood before moving to a new city, state, country, or place for their work or to start a new fresh life.

**2. DATA SOURCES**

Data Links: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

<https://geocoder.readthedocs.io/index.html>.

<http://beautiful-soup-4.readthedocs.io/en/latest/>

Will use Toronto dataset which we scrapped from Wikipedia on Week 3. Dataset consisting of latitude and longitude, zip codes.

**Foursquare API Data:**

We will need data about different venues in different neighborhoods of that specific borough.  
To gain that information we will use “Foursquare” locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each neighborhood. For each neighborhood, we have chosen the radius to be 100 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood

2. Neighborhood Latitude

3. Neighborhood Longitude

4. Venue

5. Name of the venue e.g. the name of a store or restaurant

6. Venue Latitude

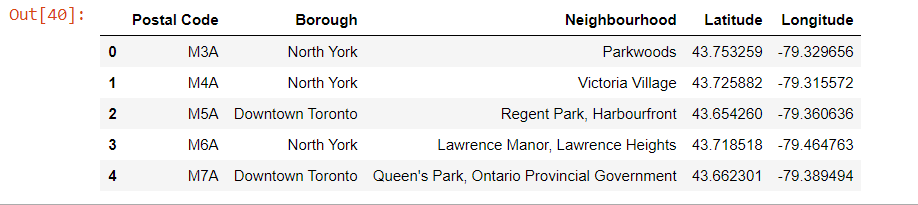
7. Venue Longitude

8. Venue Category

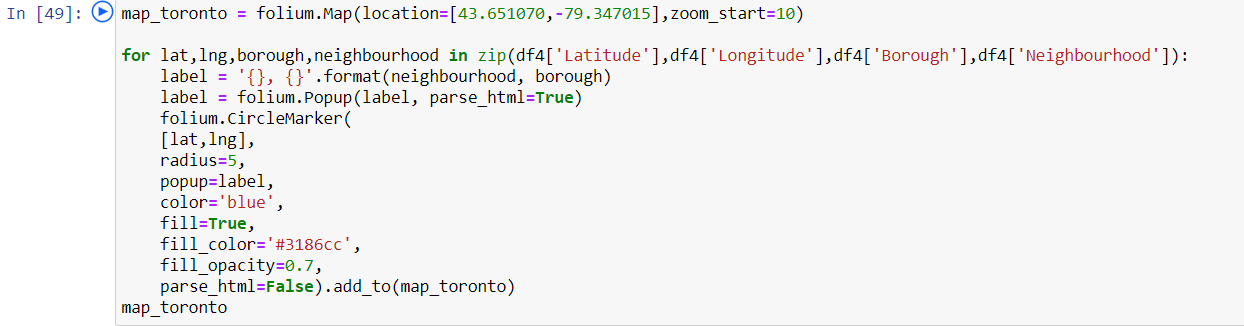
**Data Preprocessing and Cleansing:**

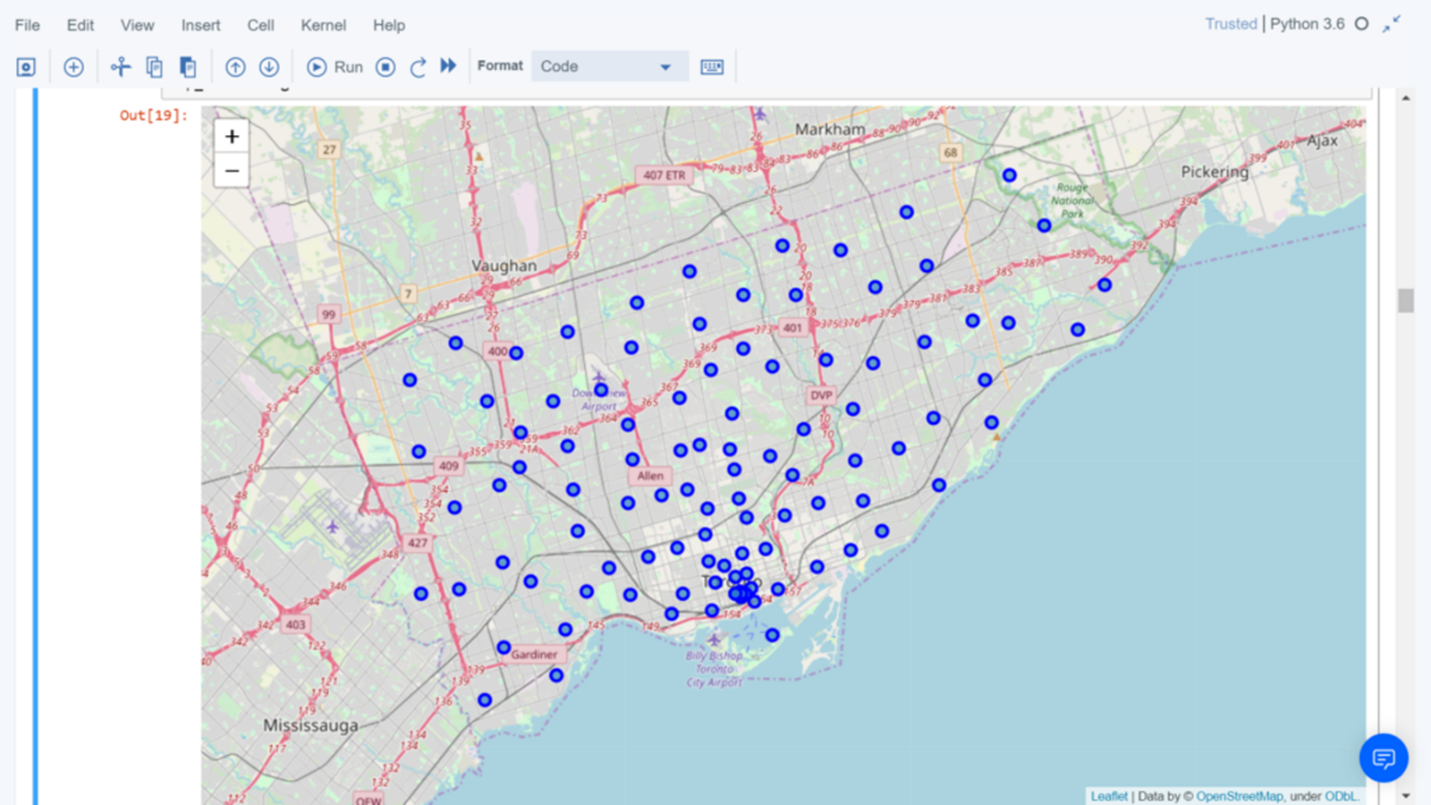
**Data preprocessing** involves the transformation of the Toronto raw dataset into an understandable format.

**Data cleansing** or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the **data** and then replacing, modifying, or deleting the dirty or coarse **data from Toronto raw dataset. The following table shows the Toronto dataset after cleansing.**

****

**Map of Toronto**

****

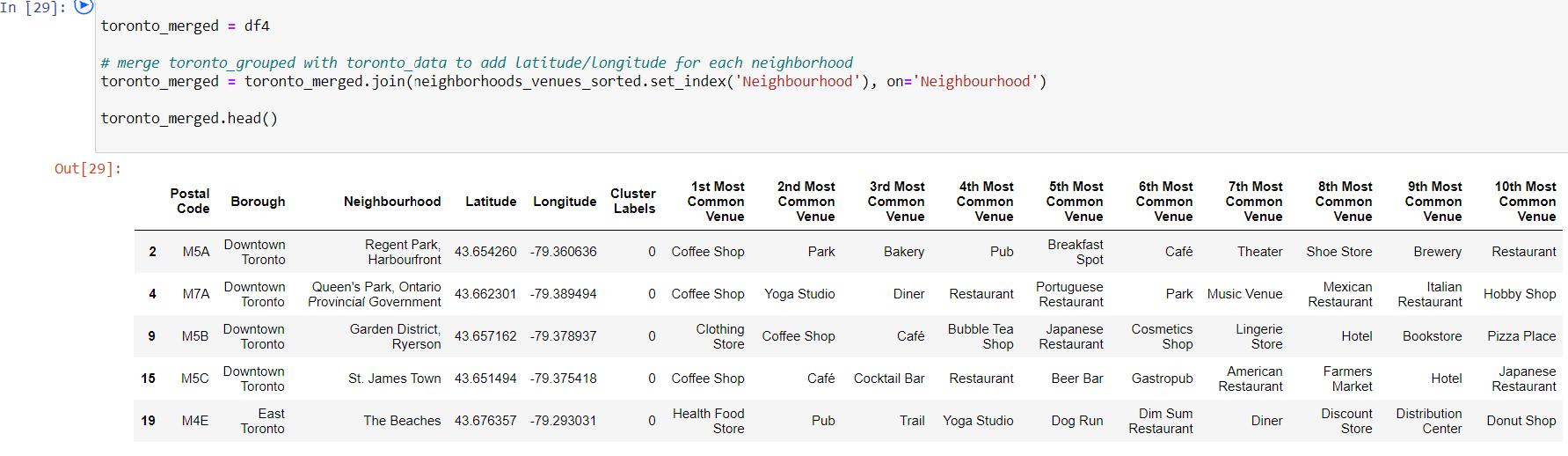


**3. METHODOLOGY**

**Clustering Approach:**

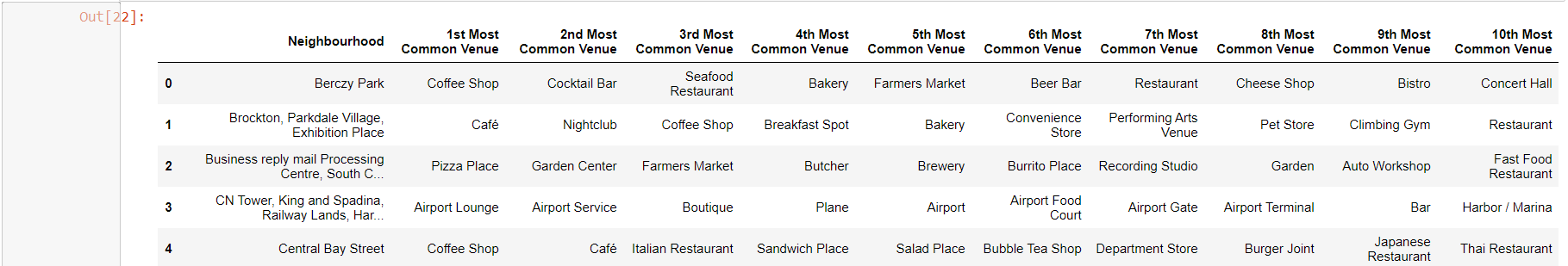
We decided to explore neighborhoods, segment them, and group them into clusters to find good neighborhoods in a big city like Toronto. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

**Using K-Means Clustering Approach** | Most Common Venue



**Most Common Venues near Neighborhood** | Using Clustering





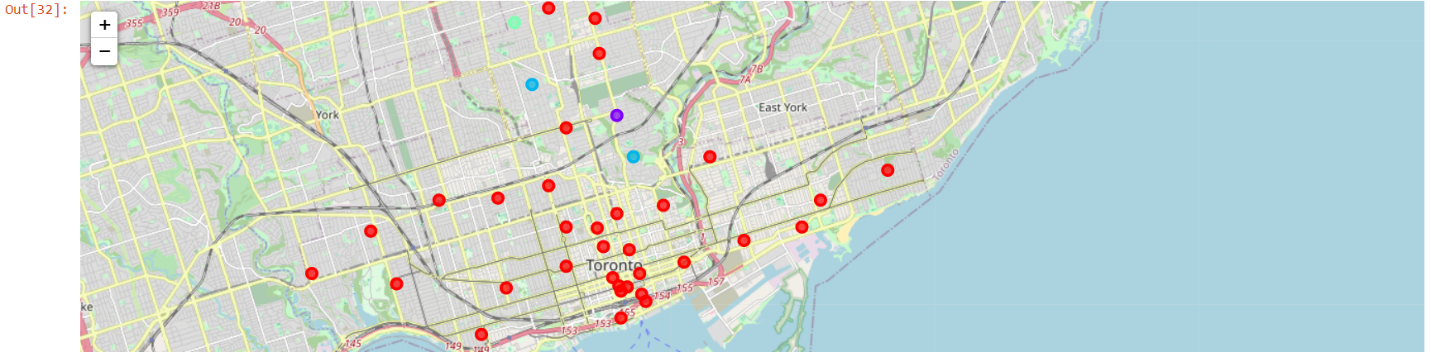
**Workflow:**

Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

would be set to 500.

**4.RESULTS**

**Map of Clusters in Toronto**



**The Location:**

Toronto is a popular destination for new immigrants in Canada to reside. As a result, it is one of the most diverse and multicultural areas in the Greater Toronto Area, being home to various religious groups and places of worship. Although immigration has become a hot topic over the past few years with more governments seeking more restrictions on immigrants and refugees, the general trend of immigration into Canada has been one of on the rise.

**Foursquare API:**

This Capstone project have used Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

**5. DISCUSSION**

**Problem Which Tried to Solve:**

The major purpose of this project is to suggest a better neighborhood in a new city for the person who are shifting there. Social presence in society in terms of likeminded people. Connectivity to the airport, bus stand, city center, markets, and other daily needs things nearby.

* Sorted list of venues near neighborhood around Toronto city.

**6. CONCLUSION**

In this Capstone project, using k-means cluster algorithm I separated the neighborhood into different clusters and different latitude and longitude from dataset, which have very-similar neighborhoods around them. Using the charts above results presented to a particular neighborhood based on most common venues near neighborhood in Toronto.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation.  
This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools.  
The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

**Future Works:**

This Capstone project can be continued for making it more precise in terms to find best neighborhood in Toronto. Best means based on all required things (daily needs or things we need to live a better life) around.

Libraries Which are Used to Develop the Project:

*Pandas: For creating and manipulating data frames.*

*Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.*

*Scikit Learn: For importing k-means clustering.*

*JSON: Library to handle JSON files.*

*XML: To separate data from presentation and XML stores data in plain text format.*

*Geocoder: To retrieve Location Data.*

*Beautiful Soup and Requests: To scrap and library to handle http requests.*

*Matplotlib: Python Plotting Module.*

GitHub Link of Complete Project

https://github.com/jh325/data.git