**Estimation of business type in Toronto city locations**

**Phani Kiran Paluri**

**Jun 12, 2019**

1. **Introduction**

**1.1 Background**

In recent era many technologies are using for finding problems and solutions in real world. Data Science is one of the technologies using for understanding problems and solving them through data analysis and visualization techniques. we found that in the history investors or retailers opened businesses in wrong location faced losses quickly. In this project finding the ideal location for businessperson to open his shop based on the business type in Toronto city, Canada using data science techniques. Foursquare is the company who provides data sets especially related to attractions and businesses around us. Foursquare API data is used to gather some information about the most common venues in each neighbourhood which will help us to group them into a bunch of clusters. Based on the cluster results the more common venue in location decides the type of business that can be established by retailer in that location. For example, a neighbourhood with common venue more constructions or furniture store would be more suitable for a warehouse compared to a neighbourhood with more coffee shops or parks which suggests it’s a residential area.

**1.2 Problem**

Investors or businesspersons are in dilemma in opening new businesses in known or unknown locations. They are not sure whether their business type will be successful in the identified location.

**1.3 Interest**

Investors and businesspersons who are interested in opening new businesses in the Toronto city area.

1. **Data acquisition and cleaning**

**2.1 Data sources**

Data used in this project is related to Toronto city in Canada. The locations from the Toronto can be found in the following link

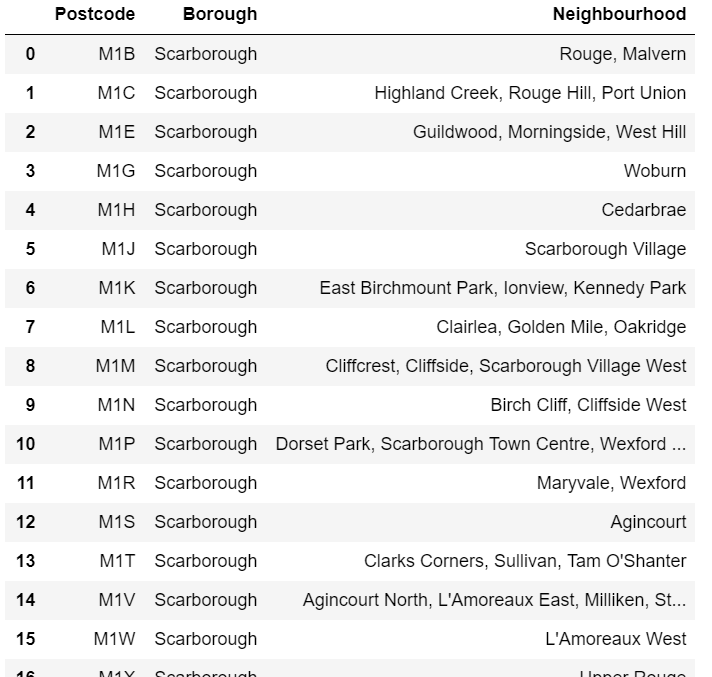
https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

**2.2 Data cleaning**

Data exists in the Wikipedia is not straight forward. Scrape the data from the given link using soup libraries and stored the required data to dataframe using python. The captured data have features of postcode, borough and neighbourhood. Drop the rows which borough is not assigned any value. Merged the neighbourhoods if the rows have same postcode. If the neighbourhood is empty for any post code then kept the same value as borough.

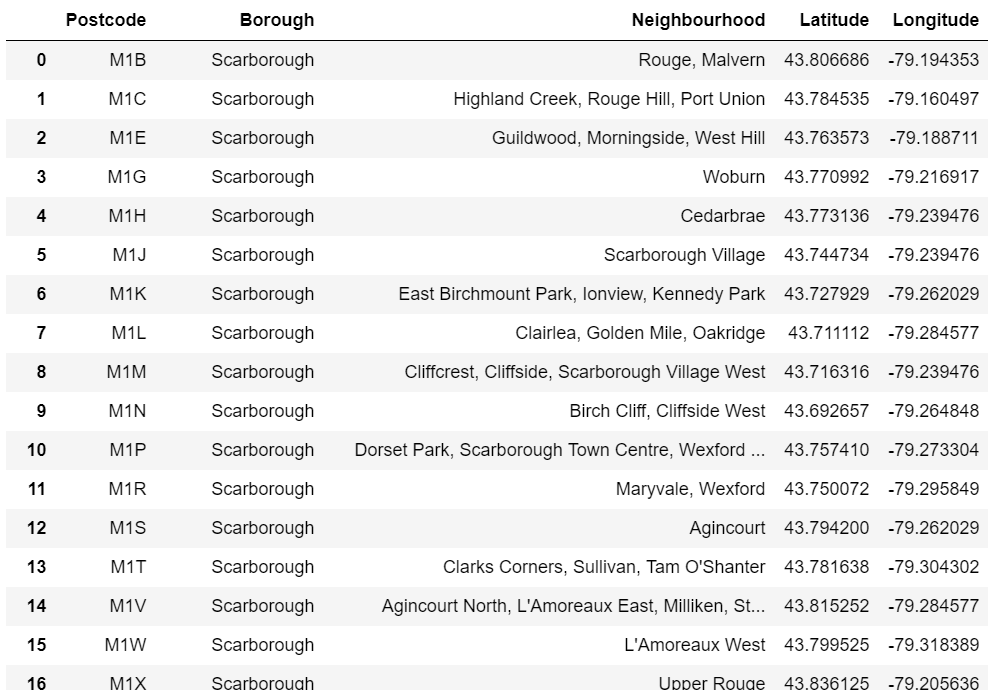
1. **Methodology**

The following is the list of sample postcodes and neighbourhoods obtained after scraping the data in the above link.



Total postcodes used in the above table are 103.

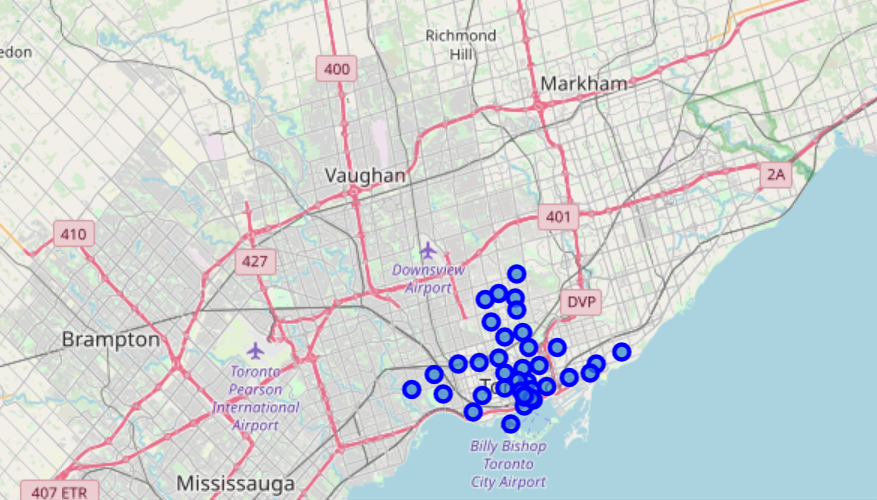
The following is the list of table contains the longitude and latitude obtained using geolocator for the above postcodes.



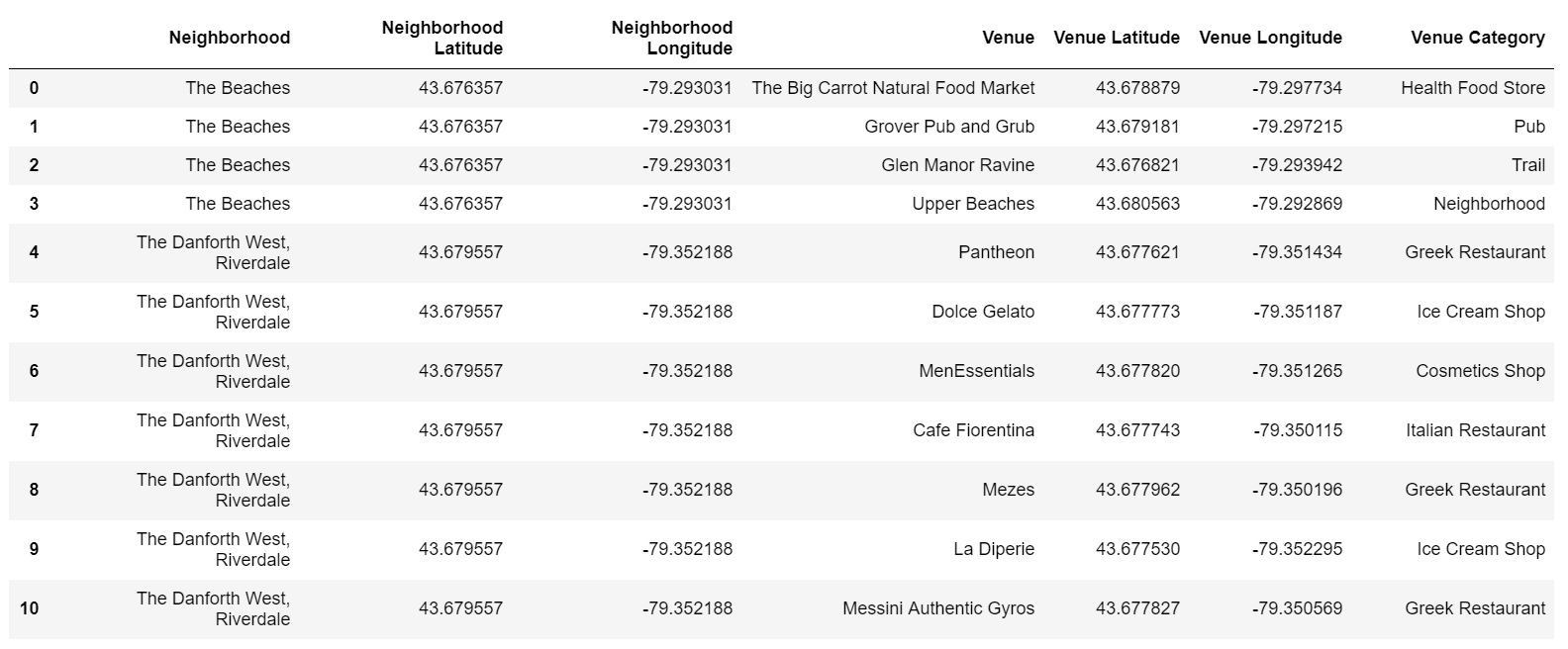
The following is the list of table contains the longitude and latitude of locations within Toronto city.

****

The following is map obtained using folium module by passing above data.



Bypassing the list of neighbourhoods and its latitudes, longitudes to Foursquare API obtained the following the list of rows having venues near to the neighbourhood.



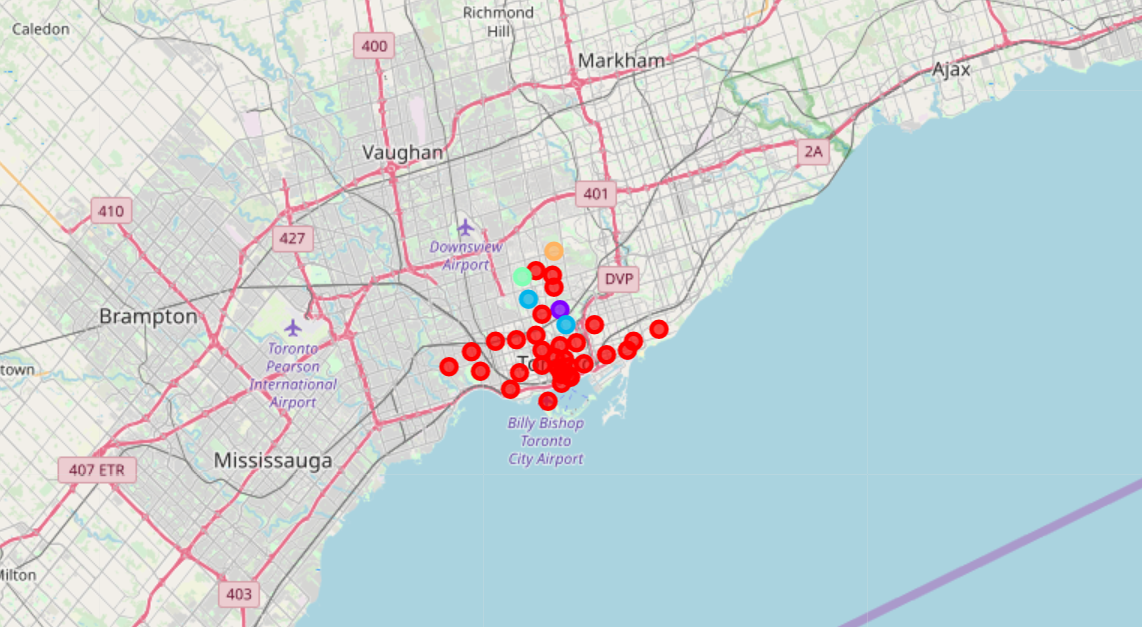
The top 10 common venues printed below for each neighbourhood which is obtained after processing of data from foursquare API.

****

Data divided into clusters using k means cluster algorithm and labelled the clusters which is shown below in the table

****

Categorised clusters show below in the map.

****

1. **Results**

Cluster 0 has the following common venues.

****

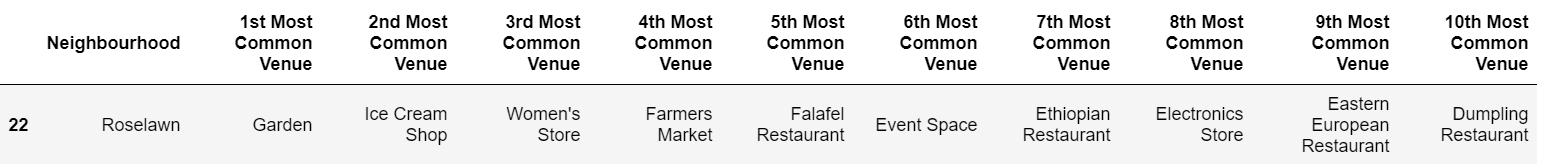
Cluster 1 has the following common venues.

****

Cluster 2 has the following common venues.

****

Cluster 3 has the following common venues.

****

Cluster 4 has the following common venues.

****

1. **Discussion**

I have shown some of the results using folium map. But tables give better idea about the common venues in the clusters.

I used the k means algorithm for segmentation and clustering of data set. The data set has 39 city locations if you have chosen more locations you may get better idea about the businesses surrounding to the postcodes.

I haven’t tried other Machine learning algorithms of clustering and segmentation. We may get better insight into the venues if we try all the machine algorithms of clustering.

**6. Conclusions**

After examining the results, we can conclude that

cluster 0 has more café shops and restaurants. Stores supplying ingredients for food making and coffee making may become successful.

cluster 1,2 sports related business may be successful.

cluster 3 has garden and farmer market. Businesses related to gardening and farming become successful.

Cluster 4 has park and schools. Businesses related to training of games or sports related become successful.

References

1. [Toronto Data](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
2. [Four square API](https://developer.foursquare.com/)
3. [Python libraries](https://docs.python.org/3/library/)