**Analyzing Neighborhoods in Toronto**

**Introduction**

Toronto is Capital of Canada. Being Major city attracts lot of people for work, tourism etc. Let’s say if someone got a job in Toronto and want to relocate to the city and he is interested in choosing a neighborhood which has parks, outdoor activities, Gym etc. So we will analyze Neighborhoods of Toronto and try to find the neighborhoods which has above requirements.

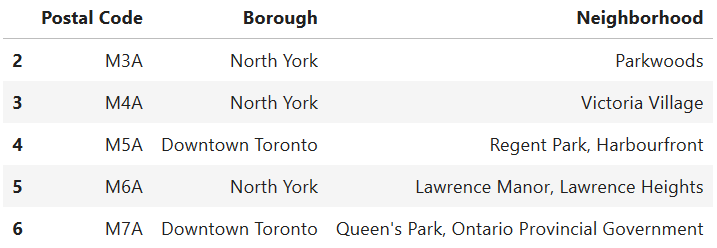
**Data**

For the Toronto neighborhood data, a Wikipedia page exists, *https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M* that has all the information we need to explore and cluster the neighborhoods in Toronto. We will scrape the Wikipedia page into a structured format using Pandas DataFrame.

**Data Wrangling**

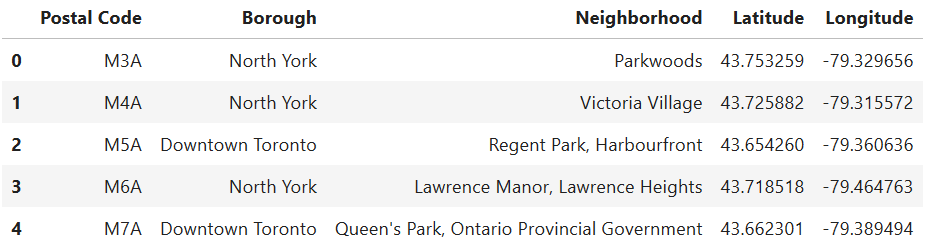
The table from the Wikipedia page consists of Toronto neighborhoods with columns Postal code, Borough and Neighborhoods. The page is scrapped using Pandas *read\_html ()* function and the neighborhood table is loaded into Data Frame.

The DataFrame consists of few cells with no Borough and neighborhood assigned. These cells are deleted and only cells which has neighborhood information is stored. The shape of DataFrame is (103, 3) indicating we have 103 Neighborhoods. The neighborhoods and Borough are grouped by postal code.

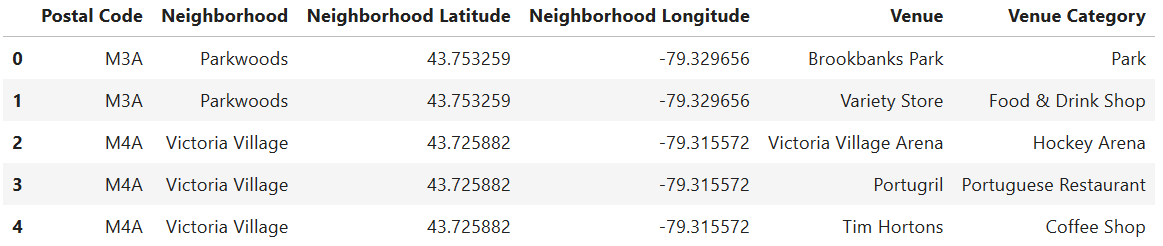


**Geospatial Coordinates**

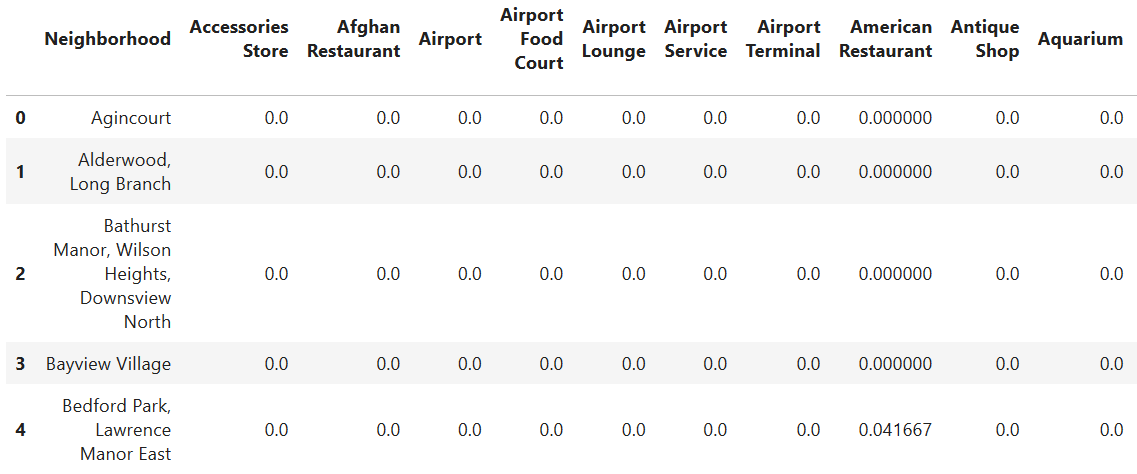
We need the Geospatial coordinates Latitude and Longitude of each postal code in order to extract the venues from each neighborhood. The coordinates for each postal code is available in Geospatial coordinates.csv. The coordinates are extracted using pandas *read\_csv ()* function and are merged with existing pandas DataFrame grouped by postal code. Now that we have all data we proceed to extracting venues.



**Feature Extraction**

The top 100 venues within radius of 500m for each neighborhood is collected using Foursquare API. A new DataFrame is created by adding each venue and its category.

The shape of the DataFrame is (2224, 6) indicating we have 2224 venues which are found using Foursquare API. There are total 268 unique venue categories. Using one hot coding the Data Frame is converted to columns with each unique venue category. Then the DataFrame is grouped by neighborhood and by taking the mean of frequency of occurrence of each venue category.



A new DataFrame is created to display the top 10 venues for each neighborhood.

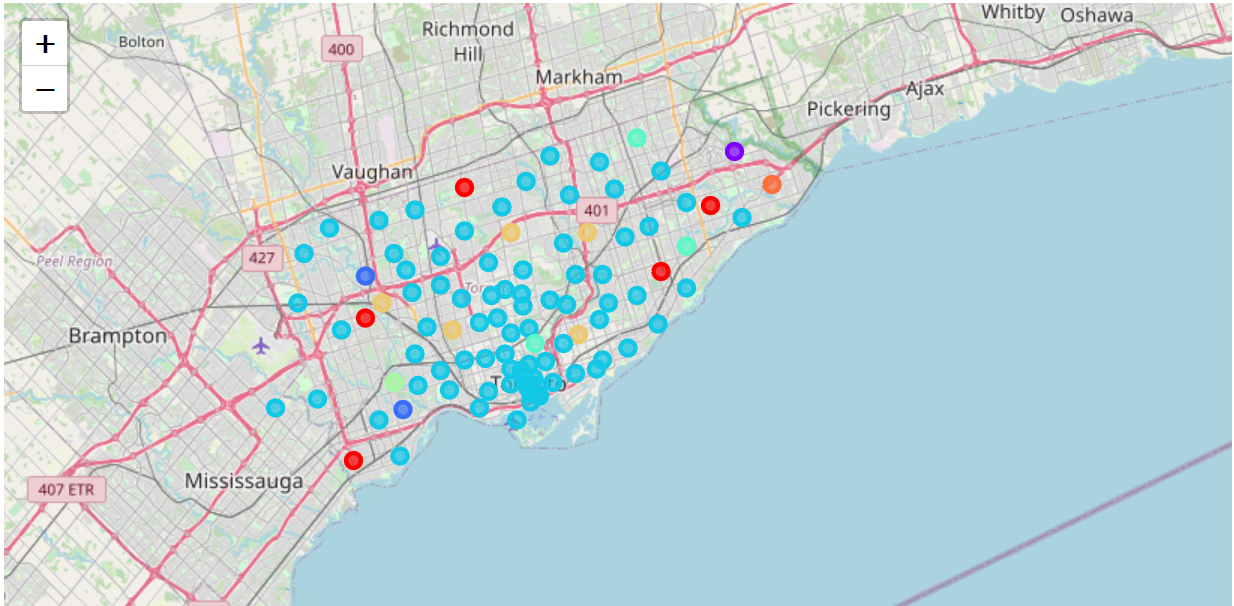


**Model**

We will use K-Means clustering algorithm for clustering our data. K-Means clustering algorithm identifies k number of centroids, and allocates every data point to its nearest cluster, while keeping the centroids as small as possible. We will cluster the neighborhoods into 8 clusters.

**Results**

The clusters gave us idea which neighborhoods are having high density of parks and outdoor activities. The below image shows the Toronto neighborhood is divided into clusters.



|  |  |
| --- | --- |
| **Cluster** | **Popular Venue Category** |
| Cluster -1( Red) | Pizza |
| Cluster-2( Purple) | Restaurant |
| Cluster -3 (Blue) | Sports field |
| Cluster-4 ( Light Blue) | Coffee shop |
| Cluster-5 ( Light Green) | Park |
| Cluster-6 ( Green) | Pool |
| Cluster-7 ( Orange) | Park |

**Conclusion**

The neighborhoods in cluster 5 and 6 have high density of Parks. These neighborhoods are ideal if you choose to relocate to Toronto and likely to live in neighborhoods with lot of free space and parks.