**Introduction:-**

**Audience/User** - Sports enthusiast students looking for home to rent.

**Business Problem** - An athlete and sports enthusiast student is shifting from his home country to Toronto for higher studies so as result he/she is searching for home to rent near places where sports/fitness areas are located in close proximity. For this reason he has contacted a real estate agent to ease out his work. This project will recommend the areas to the real estate agent where he/she can allocate a home for rent to the student. This project will help the real estate agent in providing better quality of service by allocating proper homes for rent to its customer.

**Data:-**

1. Toronto Postal Code and Neighborhood data will be used from the wiki page - <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
2. Geo Coordinates will be used from the csv file provided - 'Geospatial\_Coordinates.csv'
3. Foursquare Data will be used to analyze different venues in the neighborhoods

**Methodology:-**

**Python Libraries Used:**

1. Pandas and Numpy for Data Analysis
2. Wikipedia for extracting data from Wikipedia Page
3. Geopy to extract Geo Coordinates from an address
4. Sklearn to use K-Means Clustering
5. Matplotlib and Folium for Data Visualization

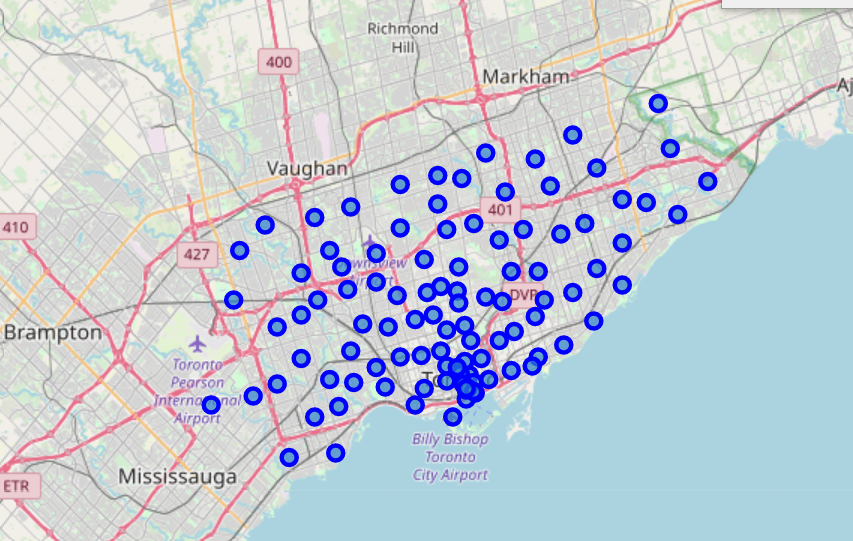
**Data Collection**

1. Toronto Poatal Code and Neighborhood data are collected from the above mentioned wiki page with help of 'wikipedia' library
2. 'Geospatial\_Coordinates.csv' is imported and merged with the Neighborhood data for further processing
3. Foursquare Data is collected with help of API

**Data Wrangling and Exploratory Analysis**

1. Neighborhood data is wrangled by removing unassigned values to Borough and by renaming the columns
2. Geospatial coordinates data is then merged to the Neighborhood data to generate the final dataframe

Toronto Neighborhoods look like below:-



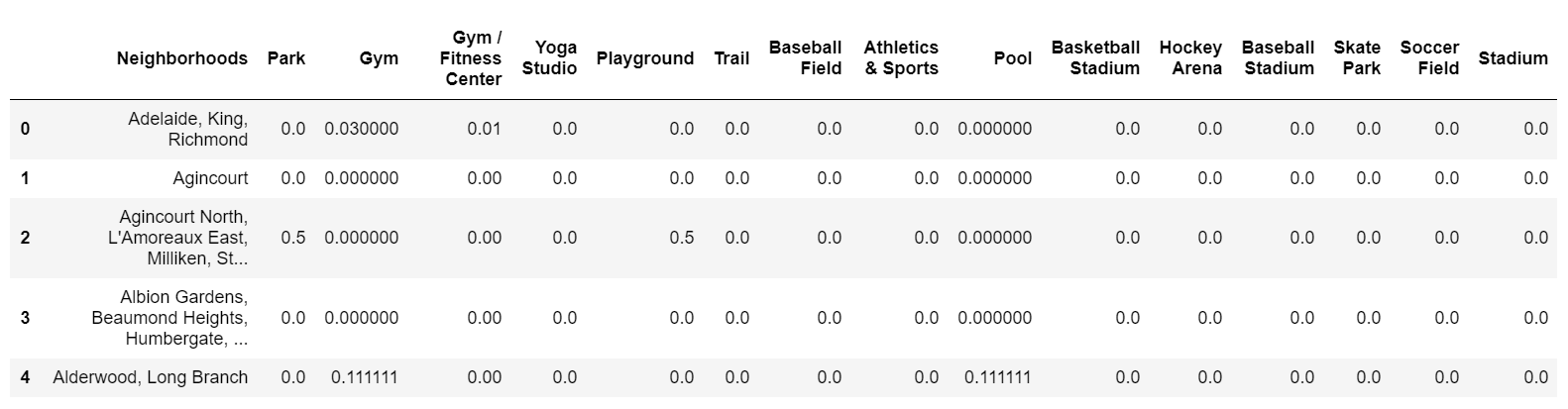
3. A function is written to call the Foursquare API and get top 100 venues in the neighborhoods in Toronto.

4. Each Neighborhood is analyzed by one hot encoding to produce frequency of occurence of different venues in each neighborhood

5. The data is grouped by Neighborhood and filtered with only below venue categories

* 1. Park
  2. Gym
  3. Gym / Fitness Center
  4. Sporting Goods Shop
  5. Yoga Studio
  6. Playground
  7. Trail
  8. Baseball Field
  9. Athletics & Sports
  10. Pool
  11. Basketball Stadium
  12. Hockey Arena
  13. Baseball Stadium
  14. Skate Park
  15. Soccer Field
  16. Stadium

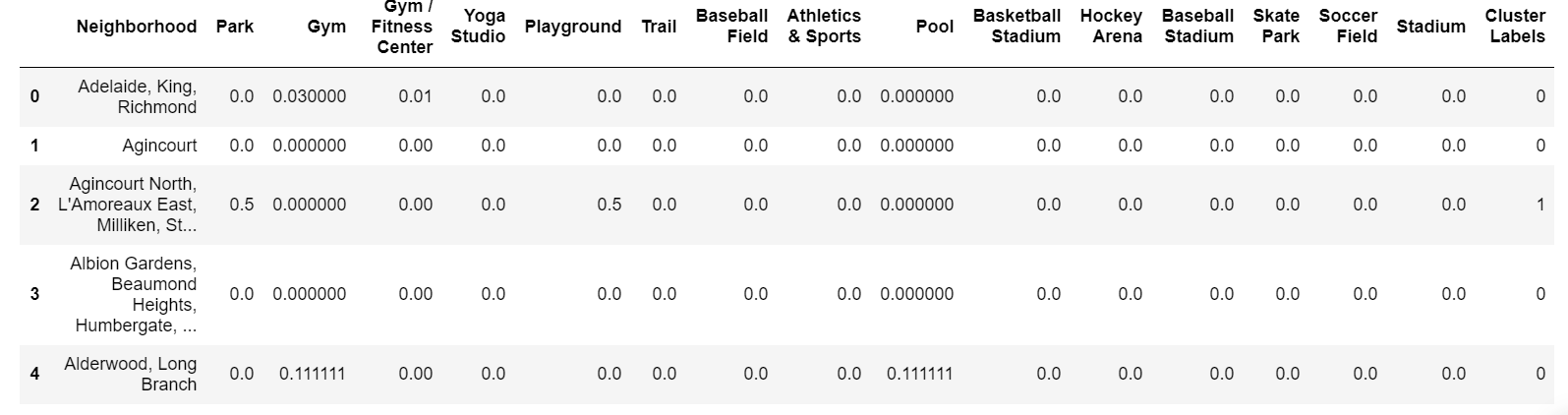
The data looks like below



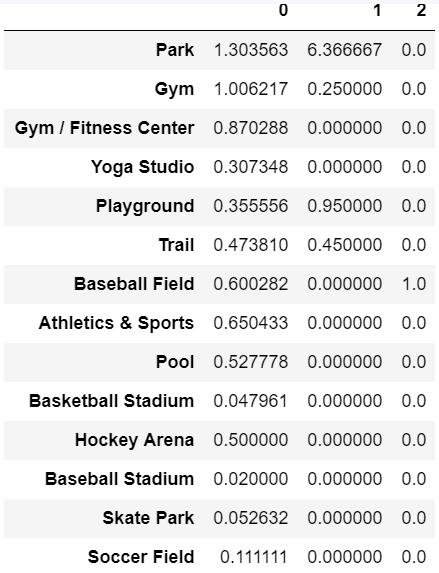
**K-Means Clustering:-**

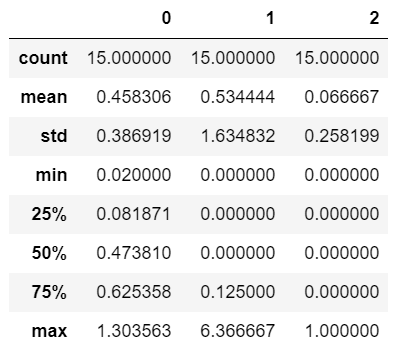
1. A K-Means clustering with 3 clusters have been performed
2. Cluster labels are added to the dataframe

Here is the data look like



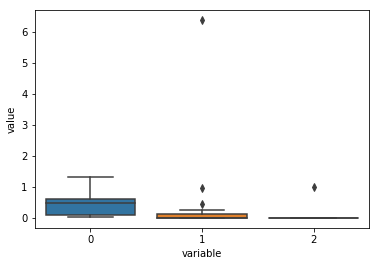
Below are the sum of occurances of each sporting venue in each cluster and statistics of the clusters:-



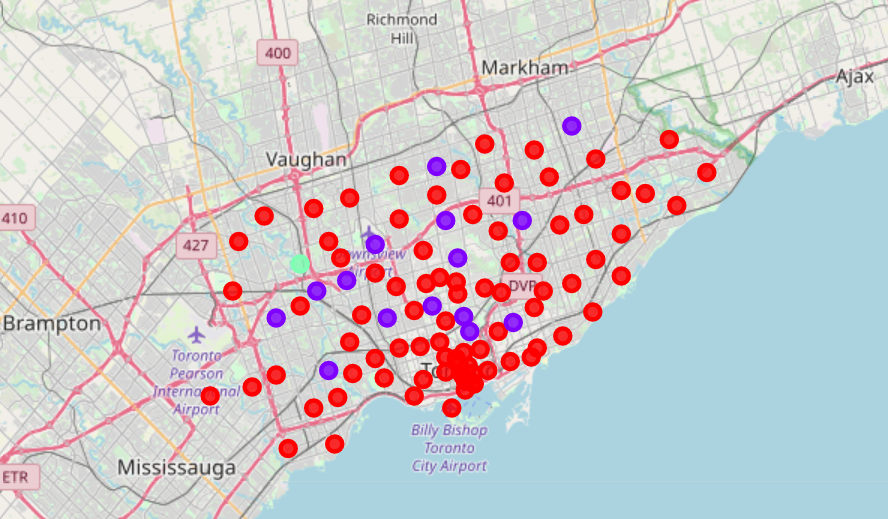


**Data Visualization:-**

1. The below box plot chart shows that wide variation and frequency of Sporting Venues are present in Cluster 0.



2. The Neighborhood clusters in the map of Toronto has been plotted with help of matplotlib and folium.



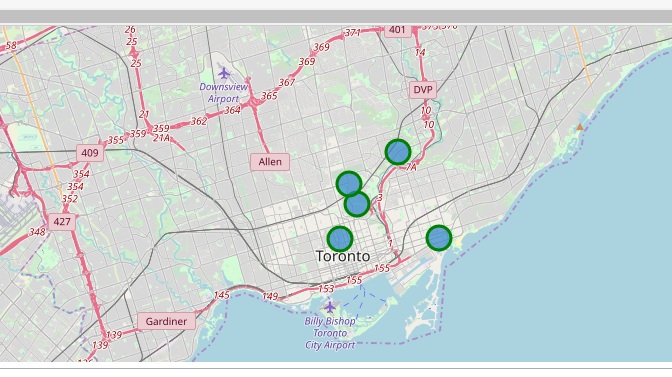
* 1. Red dots belong to Cluster 0,
  2. Purple dots belong to Cluster 1,
  3. Green dots belong to Cluster 2

**Recommendation and conclusion:-**

As per our initial assumption, the real estate agent would like to recommend an area where there are Sporting related venues in close proximity. As per our analysis, Cluster 0 could be recommended. However, there are several neighborhoods in cluster 0. Inorder to recommend top 5 neighborhoods, below steps are performed.

1. The dataframe was sorted based on maximum types and maximum count of venues
2. The top 5 Neighborhoods are recommended to open the store
   1. Business Reply Mail Processing Centre 969 Eastern (Cluster 0)
   2. Rosedale (Cluster 1)
   3. Moore Park, Summerhill East (Cluster 1)
   4. Thorncliffe Park (Cluster 0)
   5. Queen's Park (Cluster 0)

The top 5 recommended neighborhoods look like this:



---------------------------------------Project Report By Rounak Khatua--------------------------------------