**SCRUTINIZING ACCOMMODATION IN TORONTO WITH VIRTOUS NEIGHBOURHOOD**

**PRACHI SINGLA**

**DATE: 16-04-2019**

1. **INTRODUCTION**
   1. **Background**

Toronto is Canada’s largest city and is North America’s fifth most populous municipality – with a population of 2.8 million people. Toronto’s motto is “Diversity Our Strength”. It is known as one of the world’s most multi-cultural cities, Toronto prides itself on its wide range of cultures, languages, food and arts. Almost half of its population are immigrants.

Since there is lot of population is immigrants only therefore, they would require accommodation first apart from all other works after coming in Toronto. Therefore it’s advantageous for immigrants if we predict and analyse the places to live with different specifications like nearby restaurants, parks, clubs, shopping malls, gyms etc. with all these amenities. This analysis would help those immigrants, who are searching accommodation with good neighbourhood. Apart from immigrants it would be helpful for businessmen or investors in real estate properties for predicting any property price based on environment around that. Many more insights could be extracted from this analysis which we will discuss after analysing data.

* 1. **Problem: SCRUTINIZING ACCOMMODATION IN TORONTO WITH VIRTOUS NEIGHBOURHOOD**

As we all know, after relocating to some other place every person's first need is to find an accommodation according to his requirements. Since Toronto is world’s most multicultural city where almost half of population is comprised of immigrants from different cities, suburbs, countries. So here this Data Analysis will find solution to problem of finding an accommodation in Toronto city, where some basic specifications/requirements around neighbourhood of Toronto like restaurants/, parks, clubs shopping malls, gyms could be fulfilled.

**1.3 Interest**

Immigrants will be very much interested in this analysis for sure because at new place where we don’t know anything, anyplace, any person we generally prefer online sources to check out our problem. So if any immigrant would find this analysis he/she will surely follow this to find a virtuous place to live.

Apart from immigrants, citizens who already lived there could also check out recent analysis for shifting purpose to any new place in Toronto. Because sometimes, even after living at a particular place, we need to search & analyse places online to survey in a better way.

Real estate businessmen as well as investors could be helped through this analysis in analysing property’s environment to decide property prices.

City Management could also do some improvement or any survey based on this data analysis in future.

Venues density in particular area could decide measure of pollution, noise or crime also around any particular area of Toronto.

1. **Data Acquisition and Cleaning**
   1. **Data sources**

I have extracted data of neighbourhoods of Canada from here <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>. Then after reading this data into pandas dataframe, I have inserted geographical coordinates of the neighbourhoods, using the Geocoder package, here is a link to a csv file that has the geographical coordinates of each postal code: [http://cocl.us/Geospatial\_data](https://cocl.us/Geospatial_data).

Then I used forsquare API to get all venues in neighbourhoods of Toronto i.e. https://api.foursquare.com/v2/venues/explore?

After extraction of all required data, by using useful analytical techniques, I will analyse and give insights which will be helpful to many in different prospectives.

* 1. **Data cleaning**

Data (neighbourhood data) downloaded or scraped from source as mentioned above and transform all neighbourhood information into pandas dataframe after importing all important libraries needed for project.

It’s time to explore & clean data.

**Firstly**, explore dataset and extract shape of dataset along with all columns names and their number which comes out to be as follows:

|  |  |
| --- | --- |
| Dataset shape | (288,3) |
| Postal codes | 103 |
| Boroughs | 11 |
| Neighbourhoods | 209 |
| Rows with Not assigned Borough | 77 |
| Rows with Not assigned Neighbourhood | 78 |
| Rows with Not assigned Neighbourhood and Borough | 77 |

But the problem is this that it includes not assigned values, which is not helpful at all and we need to clean data (either removed or replaced these not assigned values).

So **Secondly**, Eliminate null values (not assigned values) from Boroughs & replacing null values (not assigned values) with Borough names in neighbourhood column. Then get new dataset shape with all column names with nos. & types which comes out to be as follows:

|  |  |
| --- | --- |
| Dataset shape | (211,3) |
| Postal codes | 103 |
| Boroughs | 11 |
| Neighbourhoods | 209 |

Now data has been explored & cleaned accordingly for further processing.

Data frame before cleaning



Data frame after cleaning



* 1. **Feature selection**

### The data frame has 103 Postal codes but it has 211 rows, therefore, each Postal code can present more than one neighbourhood (209 in total).Therefore, the dataframe should be grouped by the Postal code, ending with a dataframe with 103 rows.So thirdly, we have to group the data frame according to postcodes as it contains multiple neighbourhoods which comes out to be new dataset shape of (103,3).

### Now dataset shape is (103,3) because neigborhoods are grouped according to each postcode. Now each postcode contains multiple neighbourhoods separated by commas.

### Data grouped according to postcodes



### It is still showing redundancy in features and need to be more filtered for clear picture. So Fourth, we will add postal Geospatial Coordinates(latitude, longitude) to each postcode & then add it to above extracted dataset of three columns i.e. postcodes, Borough, neighbourhood.

### Dataset with geospatial coordinates



### As we know, we are solving problem for Toronto city only not for Canada. So, we need to eliminate all boroughs except for Toronto only. Fifth, we will select only Toronto in Boroughs with their postcode, neighbourhoods, latitude and longitude and eliminate all others.

### After discarding redundant features, I inspected data frame which results in total 38 nos. of Toronto locations with their neighbourhoods.These Toronto locations are just different angle locations of Toronto like east, west, south, north or central.

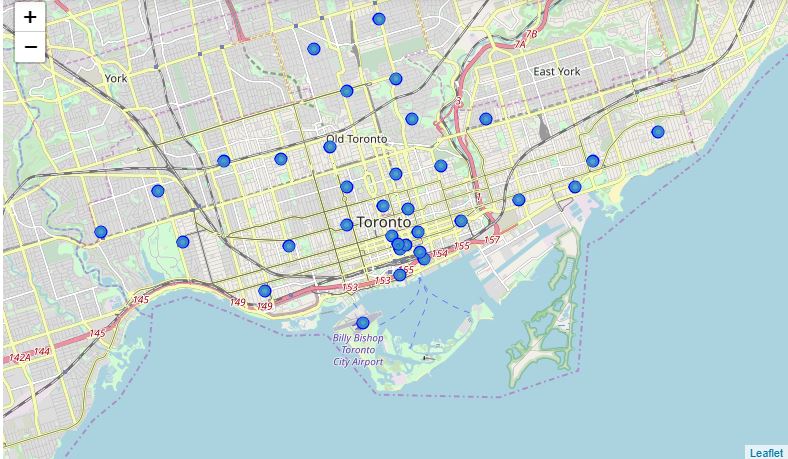
### Data frame of neighbourhoods with Geospatial Coordinates of Toronto only.



### Exploratory Data Analysis

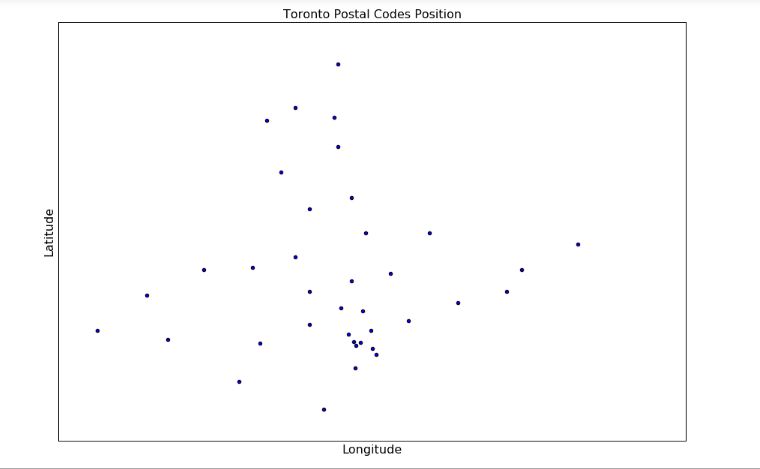
### Calculation of target variable

### First of all we have to create a map of Toronto with Geospatial Coordinates so that in further processing we can get to know nearby places to every particular location (neighbourhood) on this map. Using Geolocator, we got geographical coordinates of Toronto & then by using folium maps library of python, we have created a map of Toronto (with geospatial coordinates). So the target variable would be five categories of venues as specified above near neighbourhoods of Toronto and are mapped on Toronto map as shown below.



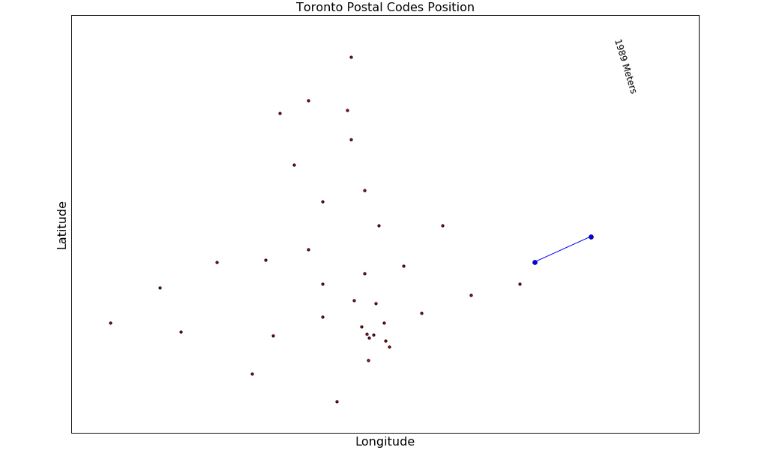
### Scatter plot of Toronto Postcode positions

### Scatter plot helps in visualizing two numeric variables. It helps in identifying the relationship of the data with each variable (i.e. latitude & longitude) i.e. correlation or trend patterns. It also helps in detecting outliers in the plot. There is no outlier in this plot.



### Scatter plot of Toronto Postcode positions (with closest postcode pair & distance specified)

### After creating above scatter plot, we need to identify closest point for each postcode using Geopy distance library. Then after getting closest distance points with their latitude & longitude positions, we have made a scatter plot of Toronto Postcode positions again with closest distance shown on that plot i.e. 1989 meters. And also we have inserted a distant column to the Dataframe and is used as the radius cover for each postcode. This is showing every postcode’s distance from shortest distanair of postcodes.



### 3.4 Creation of Toronto Map (with different radius for each postal code)

### Using Geolocator, we got geographical coordinates of Toronto & then by using folium maps library of python, we have created a map of Toronto (with geospatial coordinates) with highlighted venues along with all information embedded in that point of map.

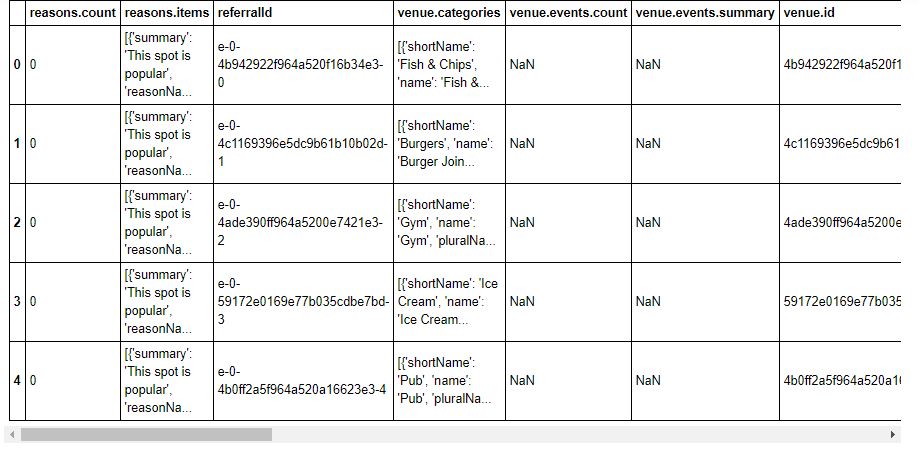
### The map is plotted using different radius for each postal code. Now not only overlapping was avoided but more area of the city will be covered. Consequently, more venues will be retrieved with different areas covered on map.

### C:\Users\nipun\AppData\Local\Microsoft\Windows\INetCache\Content.Word\toronto map with venues.jpg

### 3.5 Get all venues in neighbourhoods of Toronto

### In order to get venues in the perimeter of each postal code, it is necessary to get the geographical coordinates (latitude and longitude) of each one of those and add them to the data frame. So we have used foursquare id and foursquare secret from foursquare developer account.

### To explore the data returned by the Foursquare API, a maximum of 100 venues from the first postcode are requested in a radius of 500 meters.



### 3.6 Get venues categories

### It is necessary to extract the Category (short Name) of the JSON data. So we have extracted category name clearly.

### venue categories

### After this I have merged this data with neighbourhood data to make synchronization of venues & neighbourhoods

### 

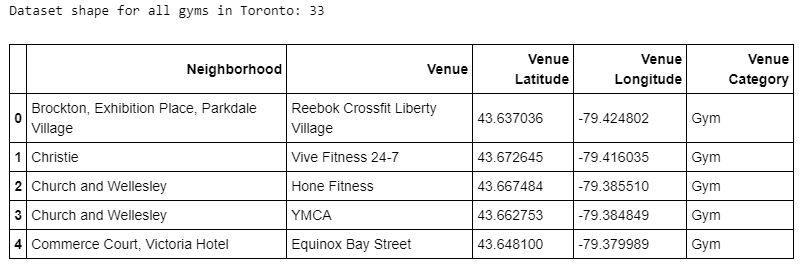
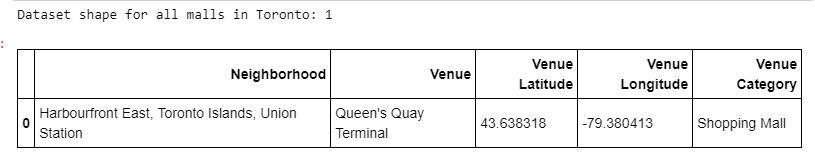
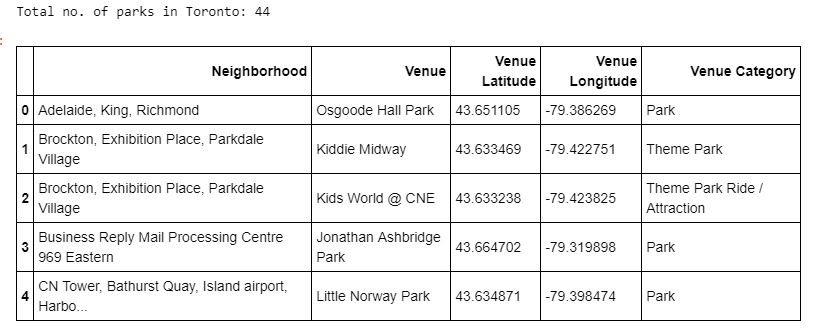


### Above data shows that there are total 1610 venues available in toronto neighbourhoods that include all amenities but we need to find out only those five which we consider for our case problem. So now we have to extract only those from 1610 which we need.

### We also found that, out of 38 unique postcodes of Toronto, 37 no. of postcodes have venues in Toronto. These 37 no. of postcodes in Toronto have total 1610 venues. But we have to find out only those venues as mentioned above out of 1609 i.e. restaurants, parks, clubs, malls, gyms & drop others.

### 3.7 Analyse venues data

### Now we will find total no. of restaurants, parks, clubs, shopping malls, gyms in data frame of Toronto with their respective dataset shape and description according to their neighbourhoods as follows:



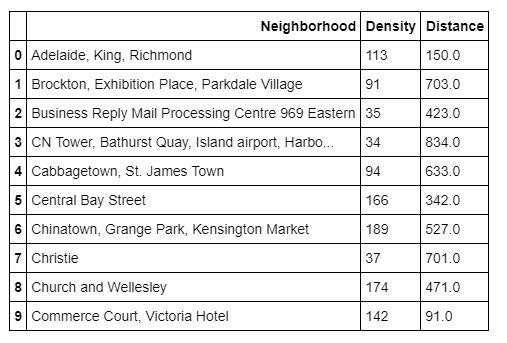
### So total number of all amenities which were specified are found. Now we have to analyse them, visualize them, cluster them to give useful insights.

**3.8 Finding density of all venues**

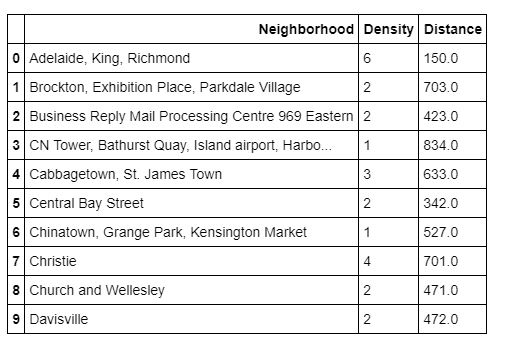
We will find density of all amenities one by one in particular postcodes which constitutes multiple neighbourhoods in a group in each postcode. This will give us number of restaurants, parks, clubs, malls, gyms in that location for further analysis and visualization.

Following are the density and distance table for all amenities:

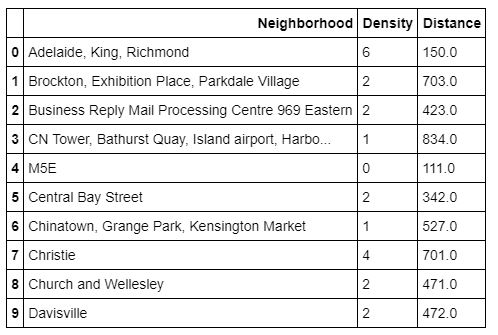
**Restaurant table**



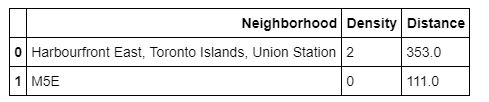
**Park table**



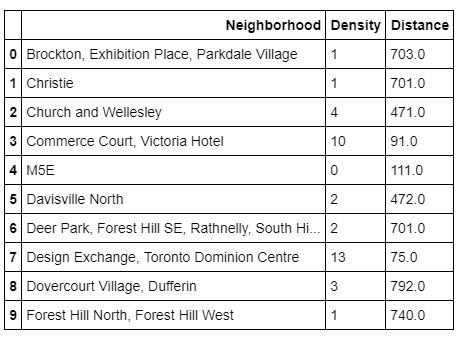
**Club table**



**Shopping Mall table**



**Gym table**



All histograms for these tables of densities will be presented in presentation which automatically says density of venues per neighbourhood group in each postcode..

1. **Predictive Modelling**

We have unlabelled data from starting so we will use unsupervised learning. In this, model works on it’s own with that knowledge that it can find and predict patterns and groupings of unlabelled data. There are four types of models in this machine learning i.e. Dimension reduction, Market Basket analysis, Density estimation and Clustering that can be used to find patterns and trends of dataset.

This is all basically machine learning. We will use Clustering model in which we will use k-means clustering to classify five different venues categories, their locations, their distance, their density in particular location i.e. restaurants, parks, clubs, malls, gyms. After getting all classifications of all venues which we need, we will cluster them and give them categories (cluster names) to identify in future by their category on map. Clustering is basically used for finding structure of data and summarizes it to give useful insights for future.

**4.1 Clustering using k-means**

In this, we will cluster five different categories of venues one by one according to their group of neighbourhood present in each postcode. Clustering is basically used for finding structure of all five different categories of venues and summarizes it in form of different groups to give useful insights for future.

* + 1. **Finding best k using elbow method**

In this we used k-means algorithm that aims to partition no. of restaurants, parks, clubs, malls & gyms into k(value of k is different for every venue) clusters. This k has been extracted using elbow method which is alternatively said as no. of clusters.

After this, there are 3 steps for k-means algorithm which I have followed and then made visualizations

Initialisation – K initial “means” (centroids) are generated at random.

Assignment – **K** clusters are created by associating each venue data with the nearest centroid. Every venue has different no. of clusters and respective centroids. So accordingly every venue has assigned categories in form of low, medium and high.

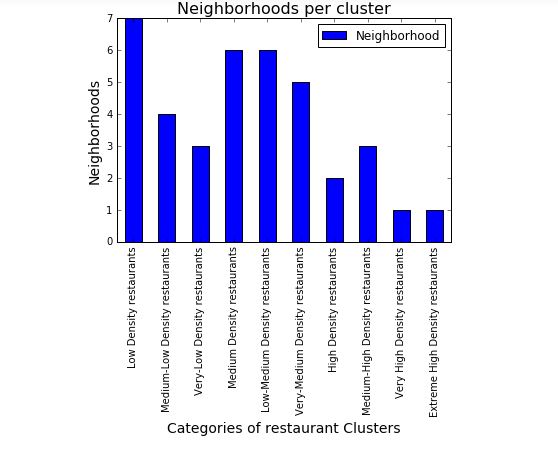
Update – The centroids of the clusters of different venues becomes the new mean

Assignment and Update are repeated iteratively until convergence. The end result is that the sum of squared errors is minimised between points and their respective centroids.

Visualization- Visualize clusters in bar graphs of restaurants, parks, clubs, malls, gyms, that would display no. of neighbourhoods constitutes per cluster and each cluster shows particular no. of centroids.

All visualizations of 5 different venues are as follows:

**Restaurants**



Based on the centroids of each cluster of restaurant formed in k-means, the cluster names are as follows:

Based on the centroids of each cluster of restaurant formed in k-means, the cluster names are as follows:

1. "Low Density restaurants: Centroid equals to 7, Cluster=0

2. 'Medium-Low Density restaurants: Centroid equals to 37, Cluster=NAN

3. 'Very-Low Density restaurants: Centroid equals to 52, Cluster=2

4. 'Medium Density restaurants: Centroid equals to 74, Cluster=NAN

5. 'Low-Medium Density restaurants: Centroid equals to 91, Cluster=4

6. 'Very-Medium Density restaurants': Centroid equals to 113, Cluster=1

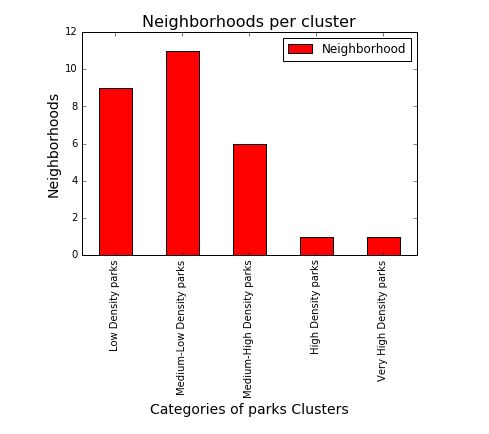
7. 'High Density restaurants: Centroid equals to 142 , Cluster=NAN

8. 'Medium-High Density restaurants: Centroid equals to 176, Cluster=NAN

9. 'Very High Density restaurants: Centroid equals to 225, Cluster=3

10. 'Extreme High Density restaurants: Centroid equals to 257, Cluster=NAN

**Parks**



Based on the centroids of each cluster of parks formed in k- means, the cluster names are as follows:

Based on the centroids of each cluster of parks formed in k- means, the cluster names are as follows:

1. 'Low Density parks’: Centroid equals to 1, Cluster=0

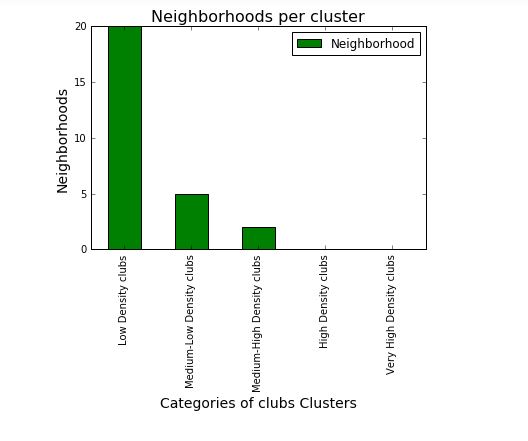
2. 'Medium-Low Density parks': Centroid equals to 2, Cluster=3

3. 'Medium-High Density parks': Centroid equals to 4, Cluster=1

4. 'High Density parks': Centroid equals to 6, Cluster=4

5. 'Very High Density parks': Centroid equals to 8, Cluster=2

**Clubs**



Based on the centroids of each cluster of clubs, the cluster names can be defined as

Based on the centroids of each cluster of parks formed in k- means, the cluster names are as follows:

1. 'Low Density parks': Centroid equals to 1, Cluster=0

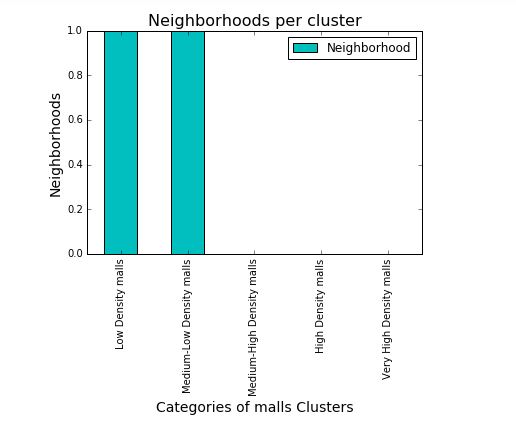
2. 'Medium-Low Density parks': Centroid equals to 2, Cluster=3

3. 'Medium-High Density parks': Centroid equals to 4, Cluster=1

4. 'High Density parks': Centroid equals to 6, Cluster=4

5. 'Very High Density parks': Centroid equals to 8, Cluster=2

**Malls**



Based on the centroids of each cluster of malls formed in k-means, the cluster names are as follows:

1. 'Low Density malls': Centroid equals to 0, Cluster=01

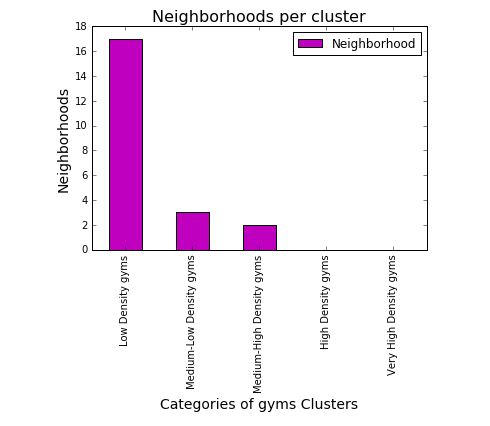
2. 'Medium-low Density malls': Centroid equals to 2, Cluster=0

3. 'Medium-high Density malls': Centroid equals to NAN, Cluster=2

4. 'High Density malls': Centroid equals to NAN, Cluster=3

1. 'Very-High Density malls': Centroid equals to NAN, Cluster=4

**Gyms**



Based on the centroids of each cluster of gyms formed in k-means, the cluster names are as follows:

1. 'Low Density gyms': Centroid equals to 1, Cluster=0

2. 'Medium-low Density gyms': Centroid equals to 5, Cluster=2

3. 'Medium-high Density gyms': Centroid equals to 12, Cluster=1

4. 'High Density gyms': Centroid equals to NAN, Cluster=3

5. 'Very-High Density gyms': Centroid equals to NAN, Cluster=4

**4.2 k-means clustering algorithm problem**

* K-means clustering is very sensitive to scale due to its reliance on Euclidean distance so be sure to normalize data if there are likely to be scaling problems.
* Determining the number of clusters in a data set, a quantity often labelled **k** as in the k-means algorithm, is a frequent problem in data clustering, and is a distinct issue from the process of actually solving the clustering problem.
* If there are some symmetries in our data, some of the labels may be mis-labelled
  1. **Solution to problem**
* Simple technique to find at least approximate value of k is taking mean distance between data points and their cluster centroid.
* The **Elbow method** is a virtuous solution to above k-means algorithm problem and I also have used this method only in my project.

This is method of interpretation and validation of consistency within [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) designed to help finding the appropriate number of clusters in a dataset.

This method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified.[[1]](https://en.wikipedia.org/wiki/Elbow_method_(clustering)#cite_note-1) Percentage of variance explained is the ratio of the between-group variance to the total

Variance, also known as an [F-test](https://en.wikipedia.org/wiki/F-test). A slight variation of this method plots the curvature of the within group variance.

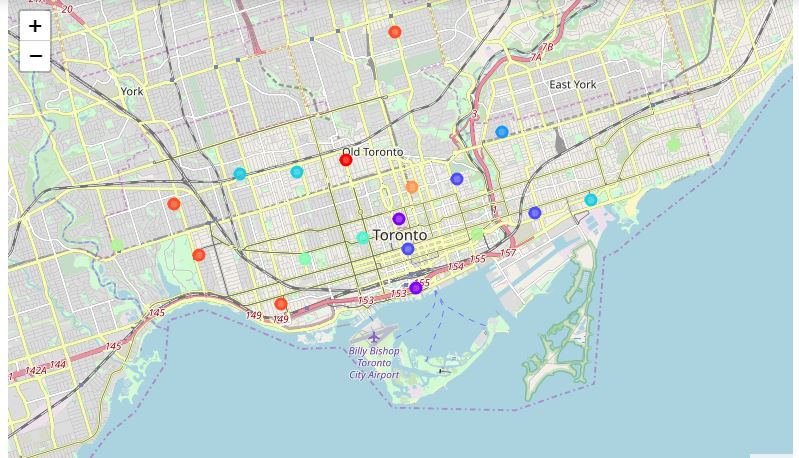
* 1. **Performance of model**
* k-means clustering is rather easy to apply to even large unlabelled data sets
* Clustering analysis is broadly used in many applications such as market research, pattern recognition; data analysis like it did in my analysis of venues in Toronto.
* Clustering of venues would also help marketers, citizens, immigrants and any city management.
* This model can characterize the venues groups based on the different priorities of any individual.
  1. **Conclusion**

In final section, I have created choropleth maps for all five types of venues which we have analysed & contain below information for each borough:

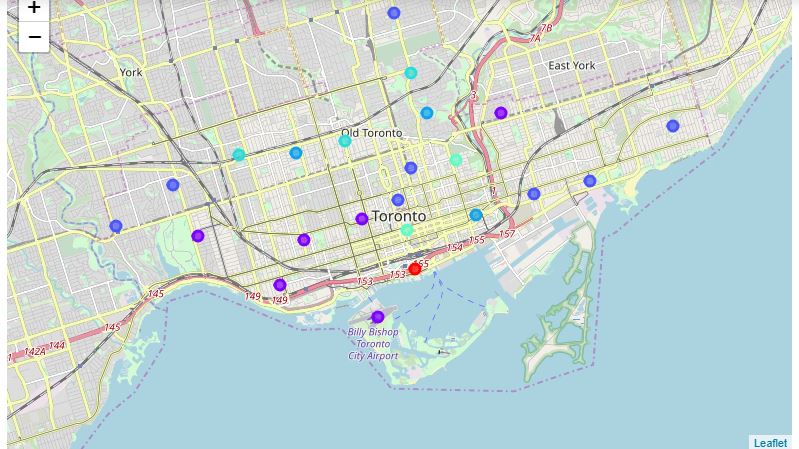
* Postcode
* Borough name
* Neighbourhood name
* Cluster no. which falls into some category which can be seen through bar plots as mentioned above with different colours of clusters based on their densities & category of cluster.

1. Restaurant Map
2. Park Map
3. Club Map
4. Mall Map
5. Gym Map

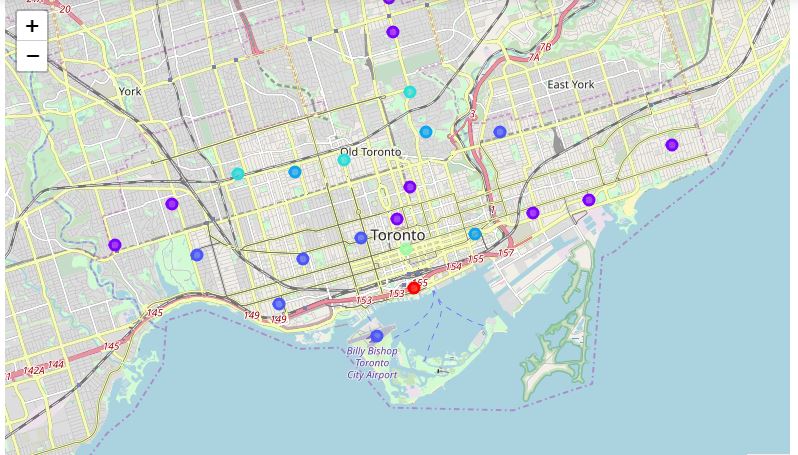
**Restaurant choropleth map**



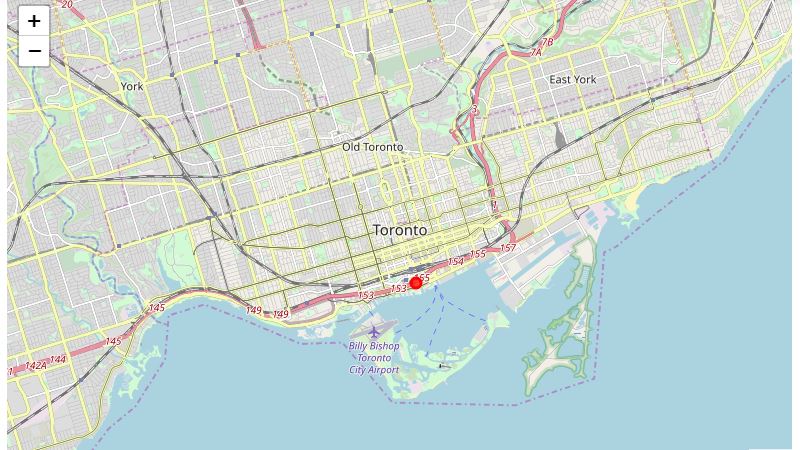
**Park choropleth map**



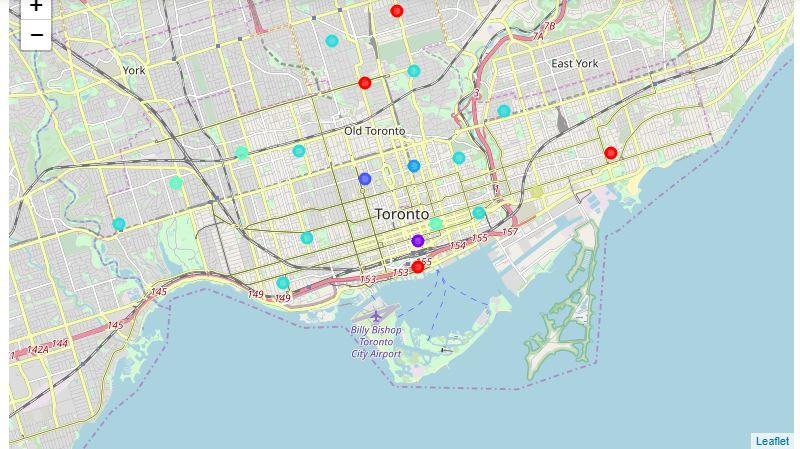
**Club choropleth map**



**Mall choropleth map**



**Gym choropleth map**



### Discussion

As I mentioned before, Toronto is a big city with a high and multicultural population density. The total number of venues present in neighbourhoods of Toronto of the 37 postcodes in total can vary. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this metropolis.

I used the K-means algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to every venue separately. However, only 38 district coordinates were used. For more detailed and accurate guidance, the data set can be expanded and the details of the neighbourhood or street can also be drilled.

I also performed data analysis through this information by adding the coordinates of districts and venues as well as static data on GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and clustering information on the Toronto map. In future studies, web or telephone applications can be carried out to direct investors or Accommodators.

**7. Future Insights**

As we know People are turning to big cities to start a business or for service purpose or for any other work.. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

* + **For investors:** Real estate investors who want to invest in properties could get better idea of property prices by observing this venues data nearby.
  + **For immigrants:** People coming from different areas, suburbs, or different countries could find an accommodation in an easy way by using this data analysis according to their priorities.
  + **Management Purpose**: City managers can manage the city more regularly by using similar data analysis types or platforms.
  + **Citizens prospective:** Citizens who already lived there could also check out this data analysis for shifting/relocating purpose within Toronto.
  + **Social prospective:** Venues density in particular area could decide measure of pollution, noise or crime also.