IBM Data Science Specialization: The Battle of Neighborhood

Project Report

**Best Place to Stay in Toronto**



Date: 19 Apr 2019

Rajendra Prasad Puhan

Table of Contents

[1. Introduction 3](#_Toc38746502)

[a. Scenario and Background 3](#_Toc38746503)

[b. Problem Statement 3](#_Toc38746504)

[c. Target Audience 3](#_Toc38746505)

[2. Data Collection and Exploration 3](#_Toc38746506)

[a. Data required 3](#_Toc38746507)

[b. Usage of data 3](#_Toc38746508)

[c. Visualizations 4](#_Toc38746509)

[3. Methodology 4](#_Toc38746510)

[a. Business Understanding 4](#_Toc38746511)

[b. Execution 4](#_Toc38746512)

[c. Analysis 13](#_Toc38746513)

[4. Results 15](#_Toc38746514)

[5. Discussion 15](#_Toc38746515)

[6. Conclusion 16](#_Toc38746516)

# Introduction

### Scenario and Background

Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most [multicultural](https://en.wikipedia.org/wiki/Multicultural) and [cosmopolitan](https://en.wikipedia.org/wiki/Cosmopolitanism) cities in the world. The number of immigrants has been increased for last couple of decades. This is very difficult to choose a place of stay and the place of work without visiting Canada. This project will help the immigrants as well as local people to find optimum location with all the facilities available nearby which matches with previous place of stay.

### Problem Statement

Finding a rental apartment in a good location in Toronto which has similar venues in Bangalore with reasonable cost. Let’s quantify the necessities.

* The venues should be similar to the place I stay in Bangalore
* The rent should be between CAD 1500 and CAD 2200
* The number of bedrooms and bathroom should be 2
* There should be at least one Indian Restaurant nearby

### Target Audience

This project will help any person or family moving to any major city in the world. As we use Foursquare data and mapping techniques, this can be replicated for similar cases for which data is available. People reviewing this project can be included in this group. This is a good start for this data science project and great learning experience for me as well.

# Data Collection and Exploration

### Data required

* [https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M,](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) The postal codes of Canada are available in the above link. The method web-scraping is used to collect the data from the webpage to csv using Beautifulsoup package in Python.
* [http://cocl.us/Geospatial\_data](https://cocl.us/Geospatial_data) The link to a csv file that has the geographical coordinates of each postal code.
* [https://www.kaggle.com/rmenon1998/bangalore-neighborhoods/data#](https://www.kaggle.com/rmenon1998/bangalore-neighborhoods/data) This has Bangalore neighborhood data.
* <https://finkode.com/ka/bangalore.html> Pin code data
* <https://www.kaggle.com/rajacsp/toronto-apartment-price> Apartments available in Toronto with coordinates and price

### Usage of data

1. Using Foursquare and geopy data find the current location in Bangalore neighborhood and top 10 venues nearby.
2. Plot the data in a map.
3. Map top 10 venues for all Toronto neighborhood and cluster in groups.
4. Find out the coordinates of the rental apartments using geopy data.
5. Find out the location of the available apartments using apartment location and price data.
6. Find out Indian Restaurants nearby using Foursquare with search query = ‘Indian’
7. Analyze the cluster groups that matches with the Bangalore location.
8. Finalize the place of stay by plotting all the data on a map.

### Visualizations

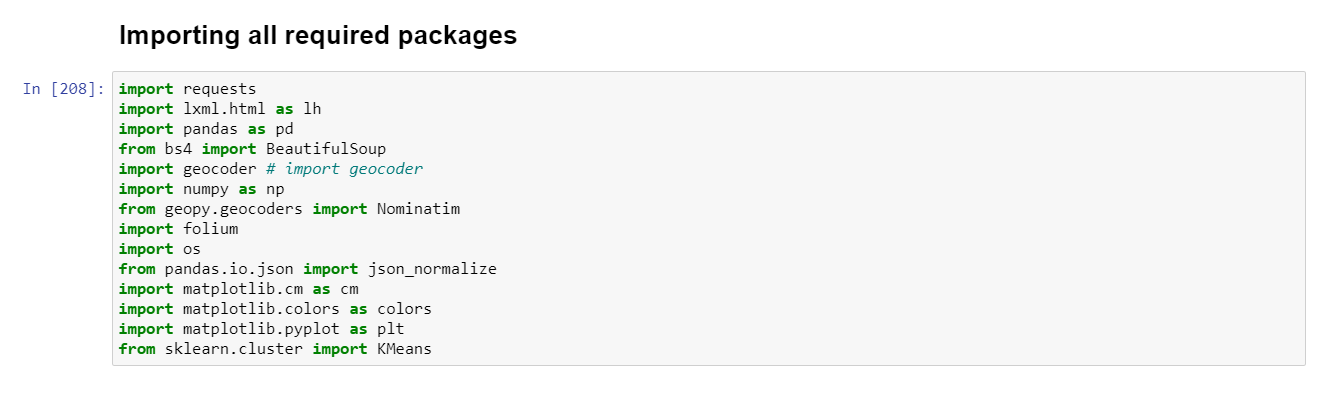
Multiple maps to be created to visualize the locations accurately. The clusters, nearby venues, Indian Restaurants and rental apartments are marked in Toronto map. The packages like Folium, Matplotlib are used to plot data to understand and analyze data.

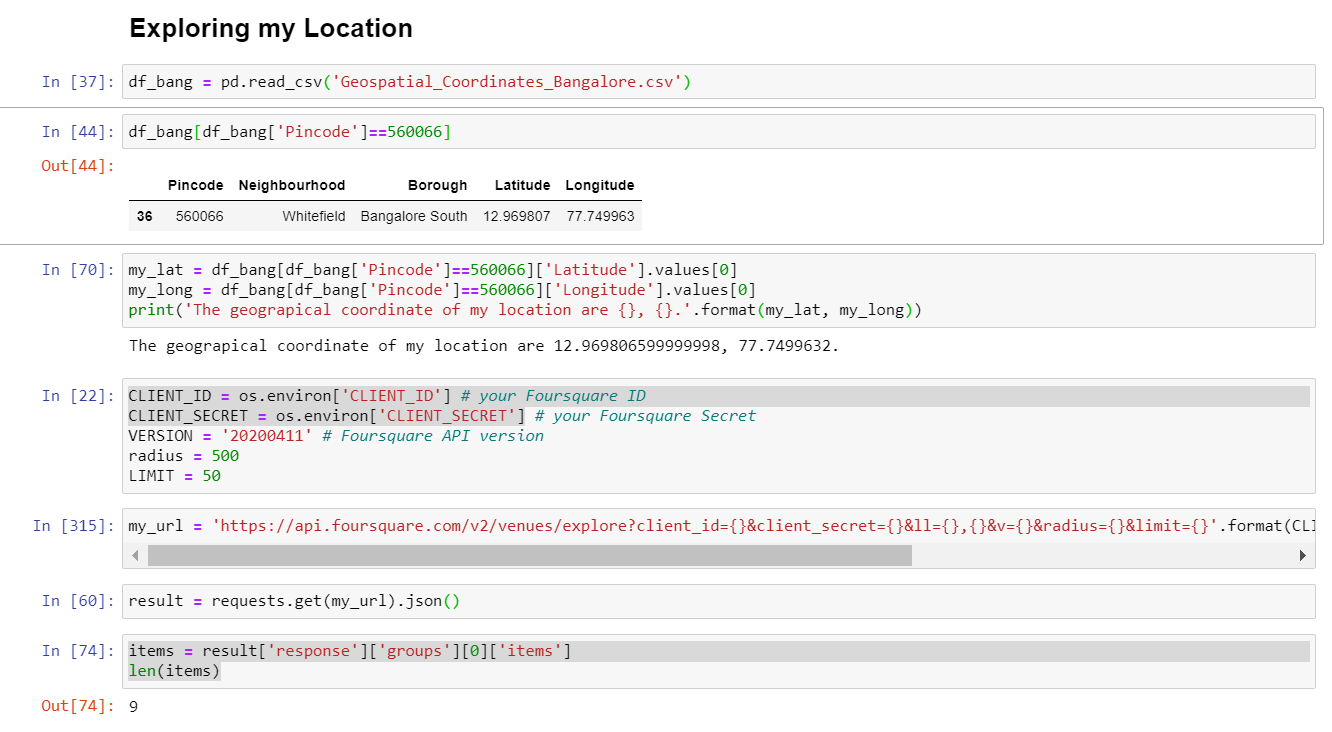
# Methodology

### Business Understanding

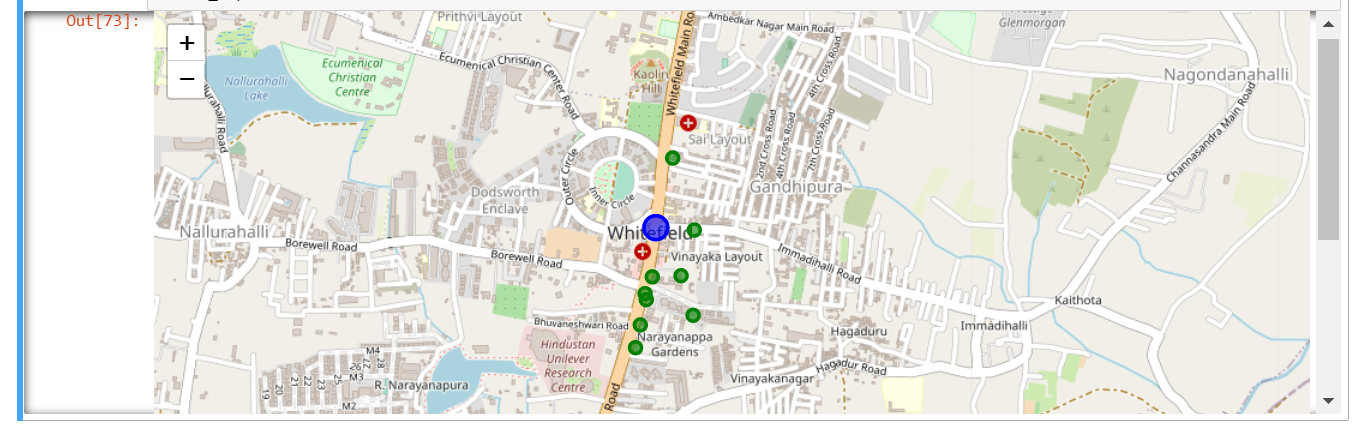
The objective is to find out best location of stay and choose suitable apartment from the list. The apartment should have similar venues nearby as with Bangalore neighborhood. For this we have to analyze the venues nearby my house in Bangalore and venues in Toronto neighborhood.

### Execution

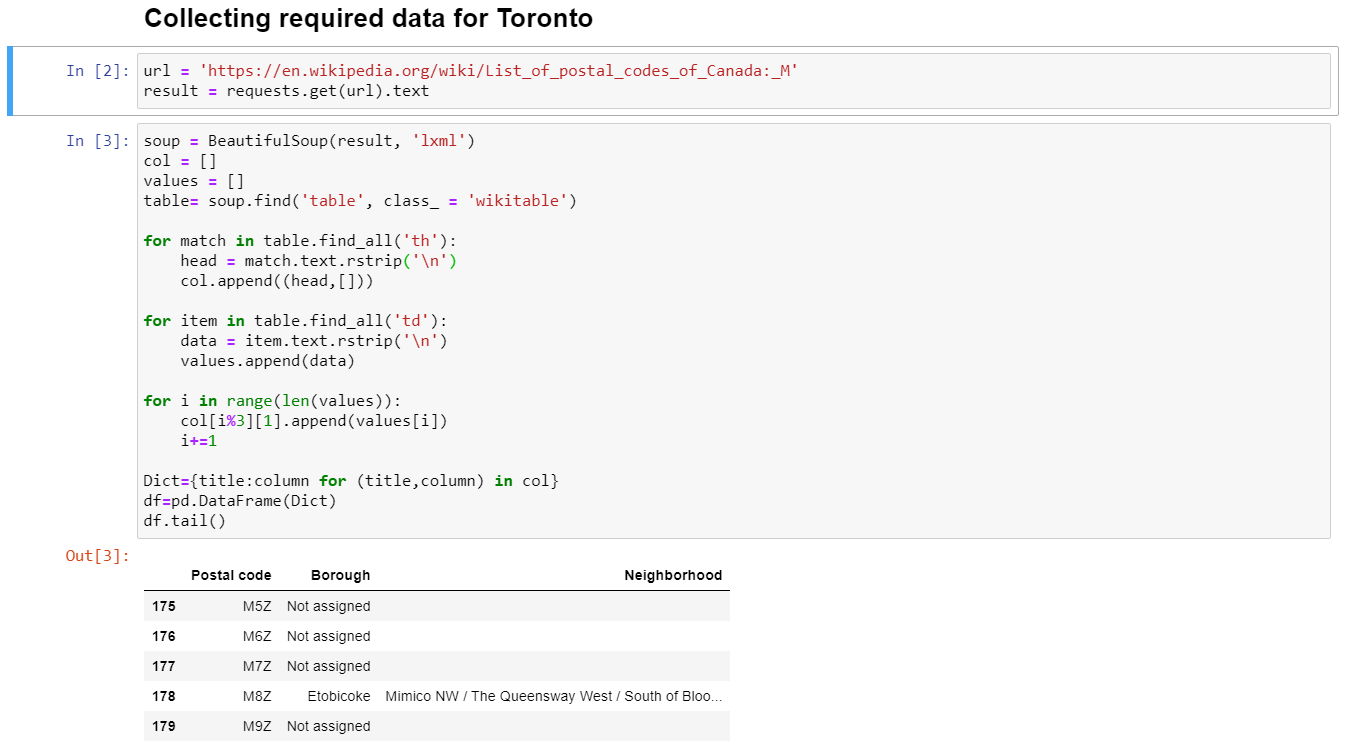




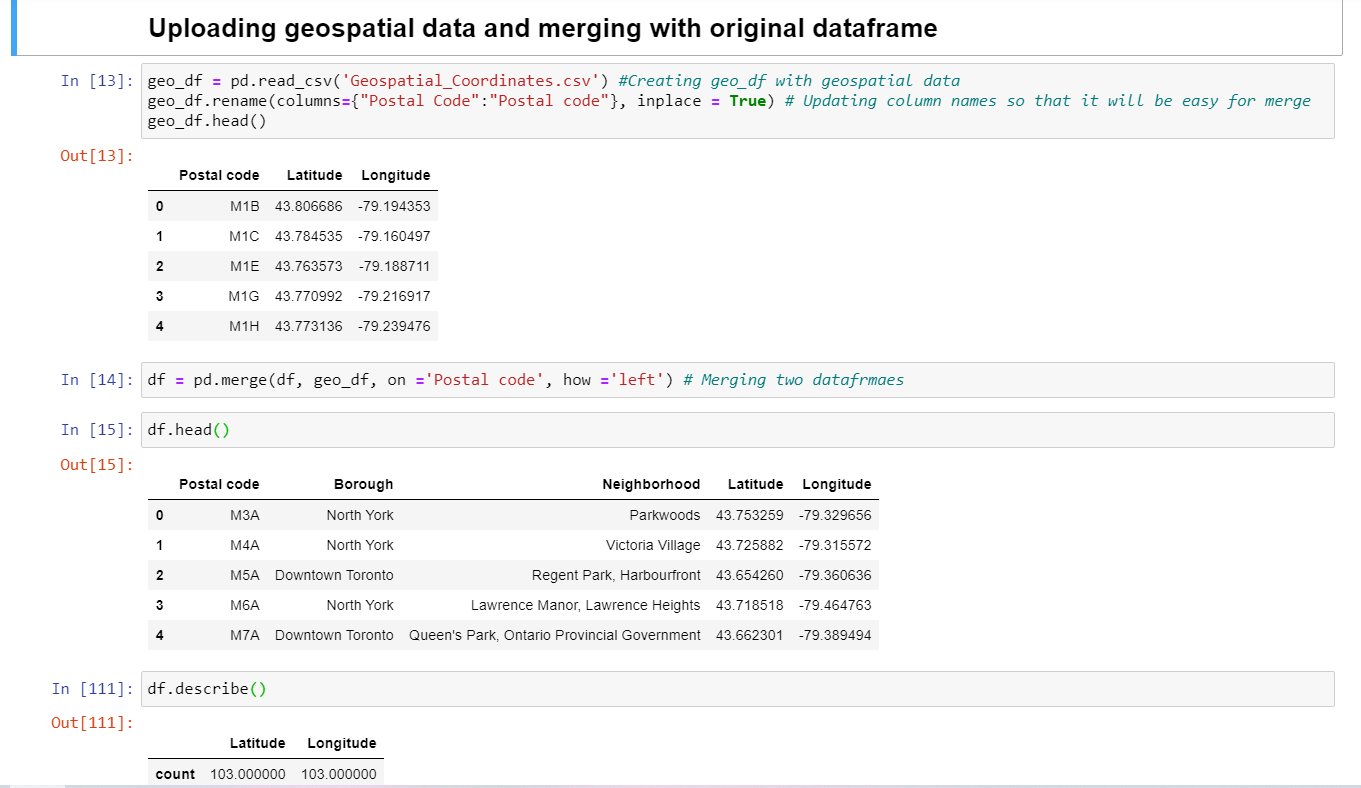
The top 10 venues are found out from Foursquare and plotted on the map with current location.



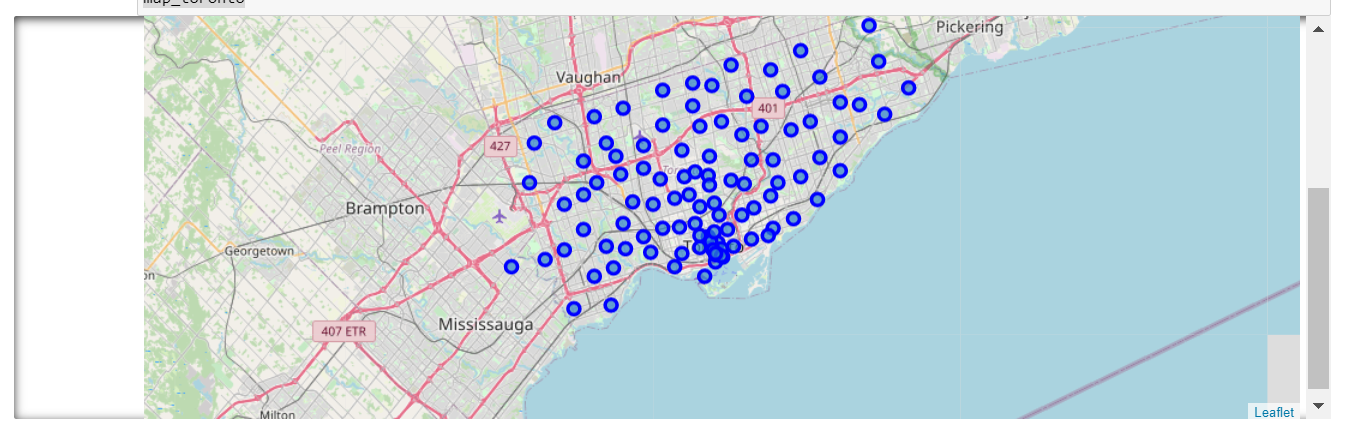
The required neighborhood data is collected from the URL <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,> using beautifulsoup package.



The geospatial data is downloaded from URL  [http://cocl.us/Geospatial\_data](https://cocl.us/Geospatial_data) and all the co-ordinates are assigned to the postal codes of the neighborhoods of Toronto.



The neighborhood data is mapped in Toronto geographical map.

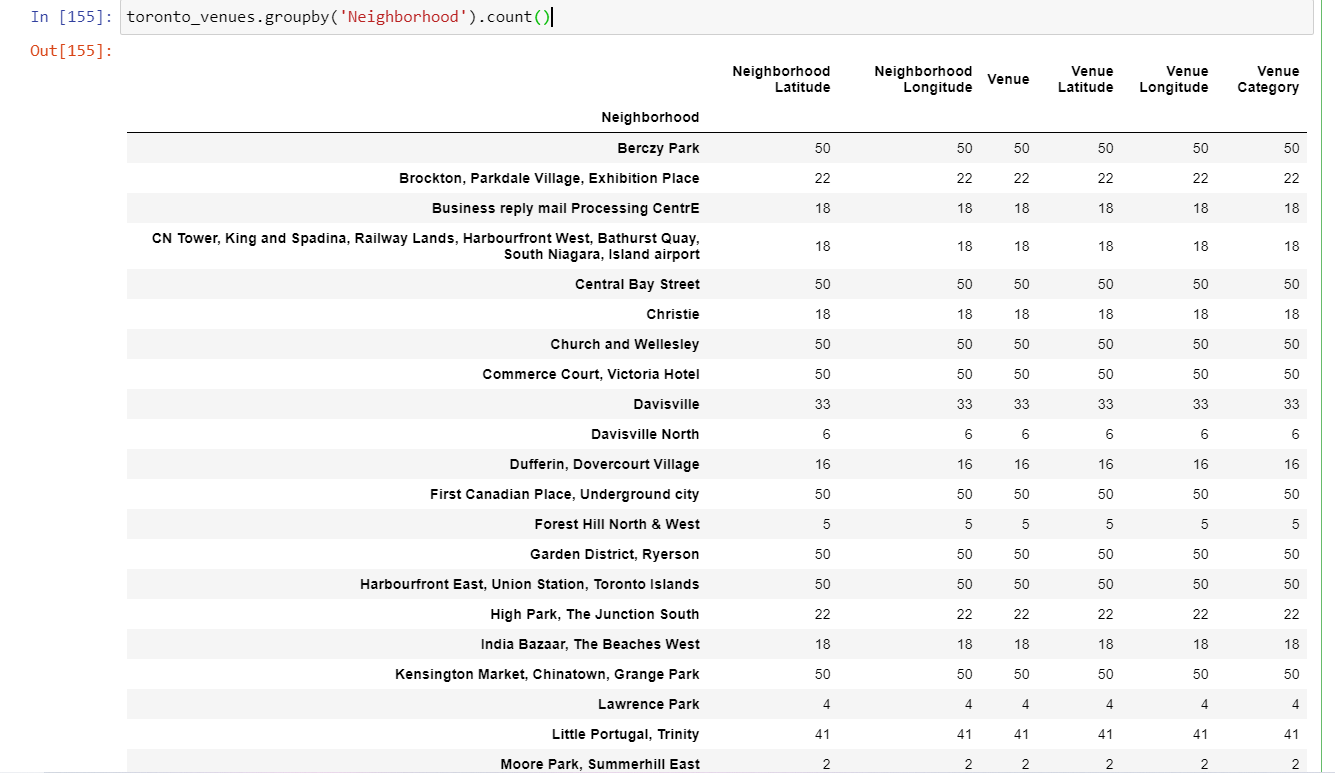


Finding all the nearby venues of the neighborhoods of Toronto using Foursquare credentials.





The venues are counted by grouping with column Neighborhood.



One hot encoding is performed using get\_dummies of Pandas.



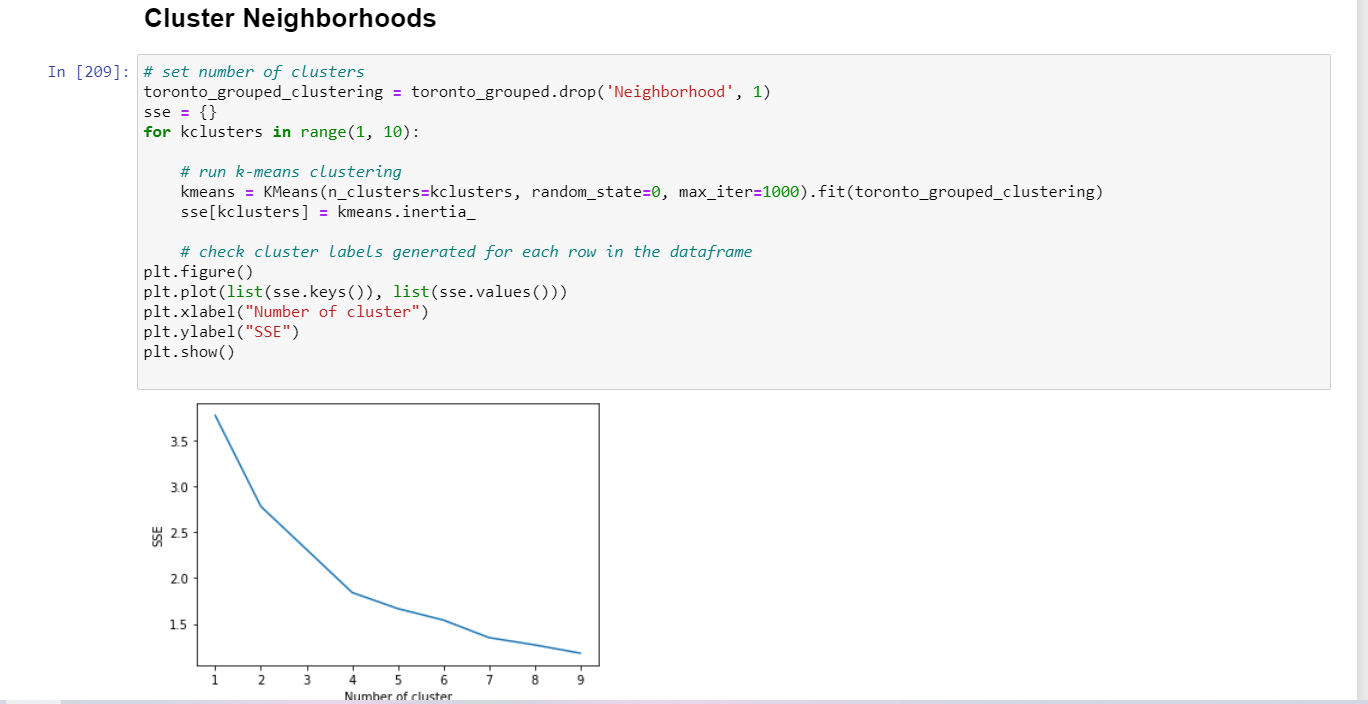
Top 5 venues are shown by neighborhood.



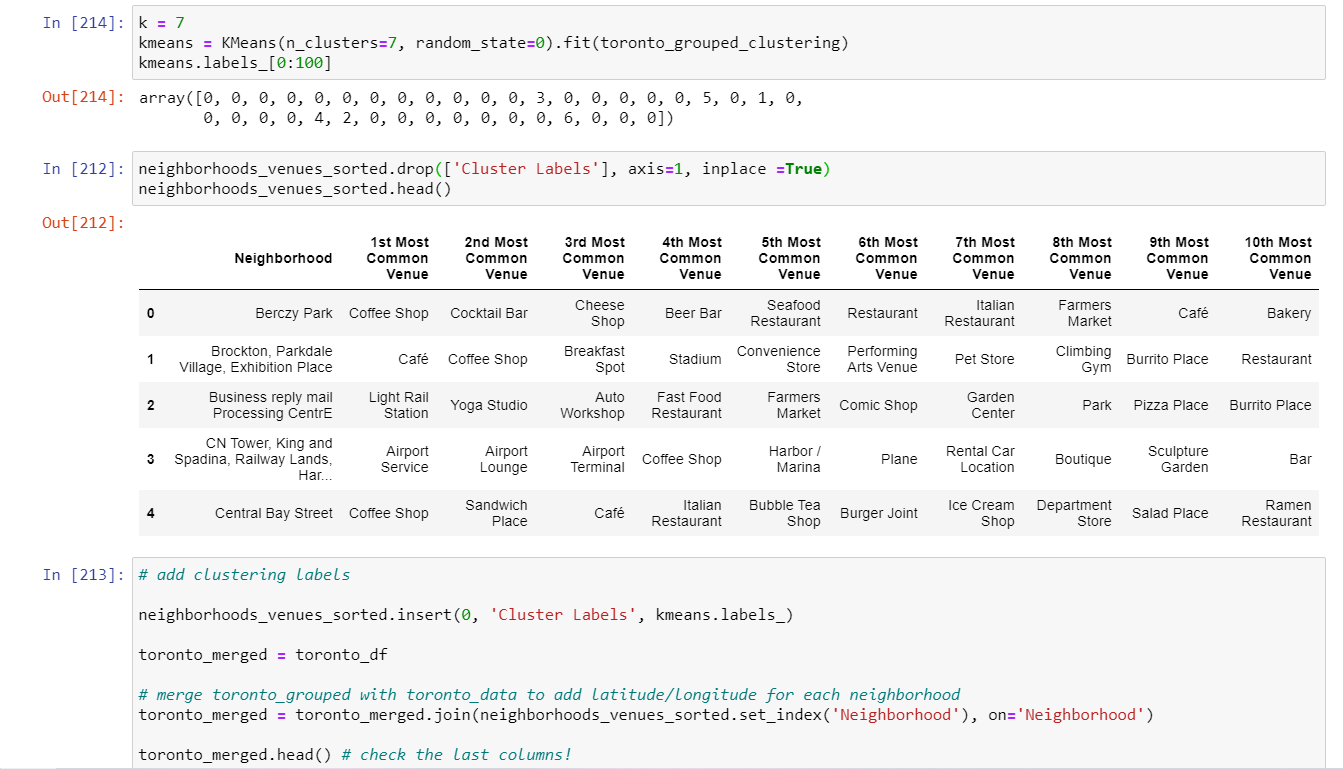
Top 10 venues are assigned to the neighborhoods and columns are arranged in order from 1st to 10th.



The neighborhoods are grouped in to clusters and the number of clusters to be decided by plotting SSE against no of clusters. So k is incremented from 1 to 10 and SSE is calculated for each step. In the graph mentioned below SSE is on Y axis and no of clusters is plotted on X axis. This process is called as Elbow method.



From the above graph we finalized the no of clusters should be 7. With k=7 the model is fitted and each row is assigned with a cluster number from 1 to 7.





The clusters are plotted on a map as shown below.

Red dots: Cluster 0

Purple dots: Cluster 1

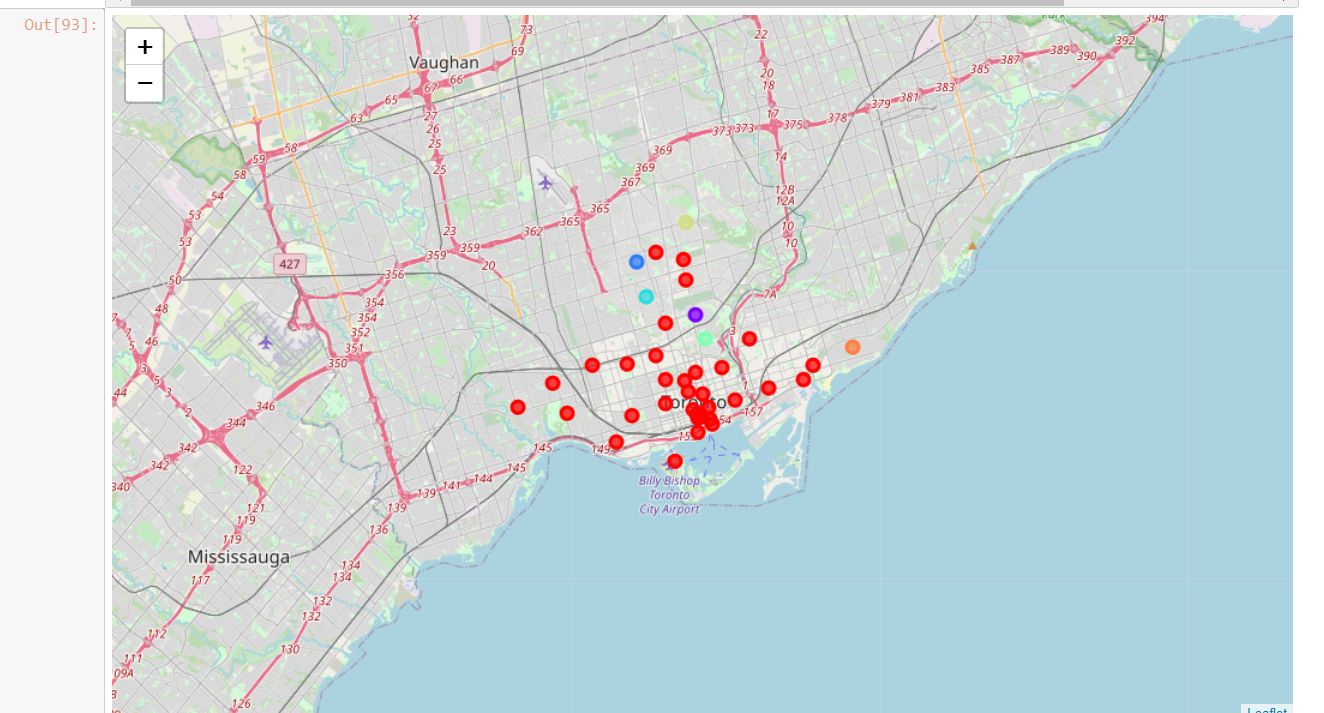
Navy blue dots: Cluster 2

Light blue dots: Cluster 3

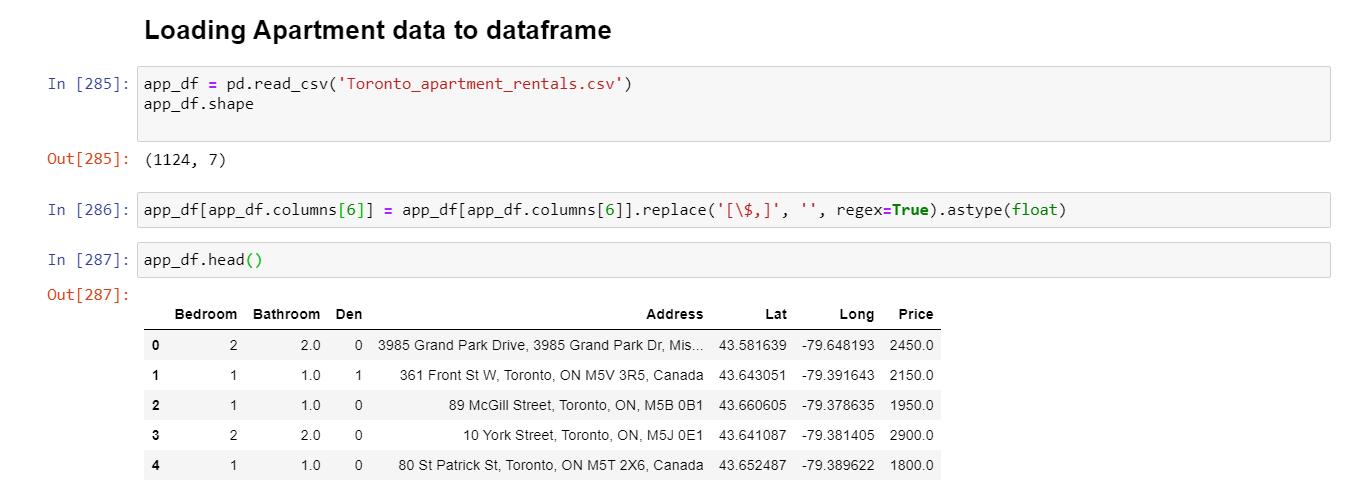
Aqua green dots: Cluster 4

Brown dots: Cluster 5

Orange dots: Cluster 6



Apartment data with co-ordinates and prices are loaded to dataframe.

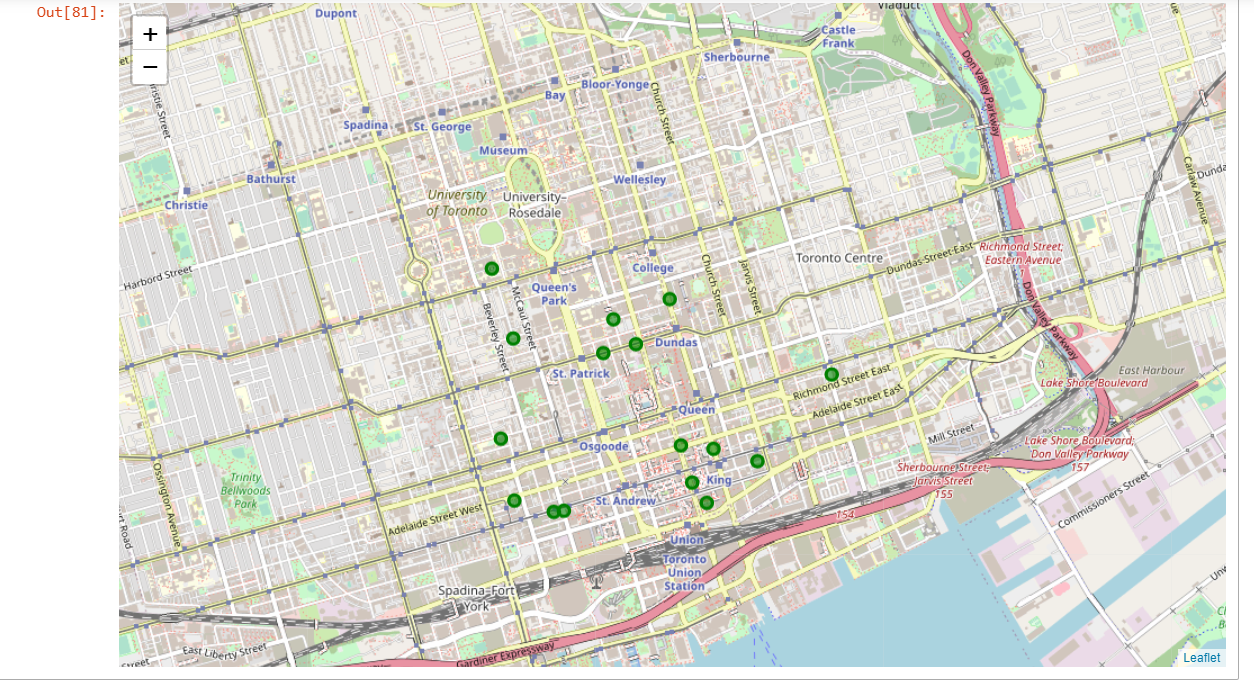


Refining the apartment data with required below conditions and plotting on Toronto geographical map.

* The rent should be between CAD 1500 and CAD 2200
* The number of bedrooms and bathroom should be 2



Finding Indian restaurants in Toronto using Foursquare and search query as ‘Indian'. Plotting the restaurant locations on the map.



Combining all the plots on one map.

Red dots: Cluster 0

Purple dots: Cluster 1

Navy blue dots: Cluster 2

Light blue dots: Cluster 3

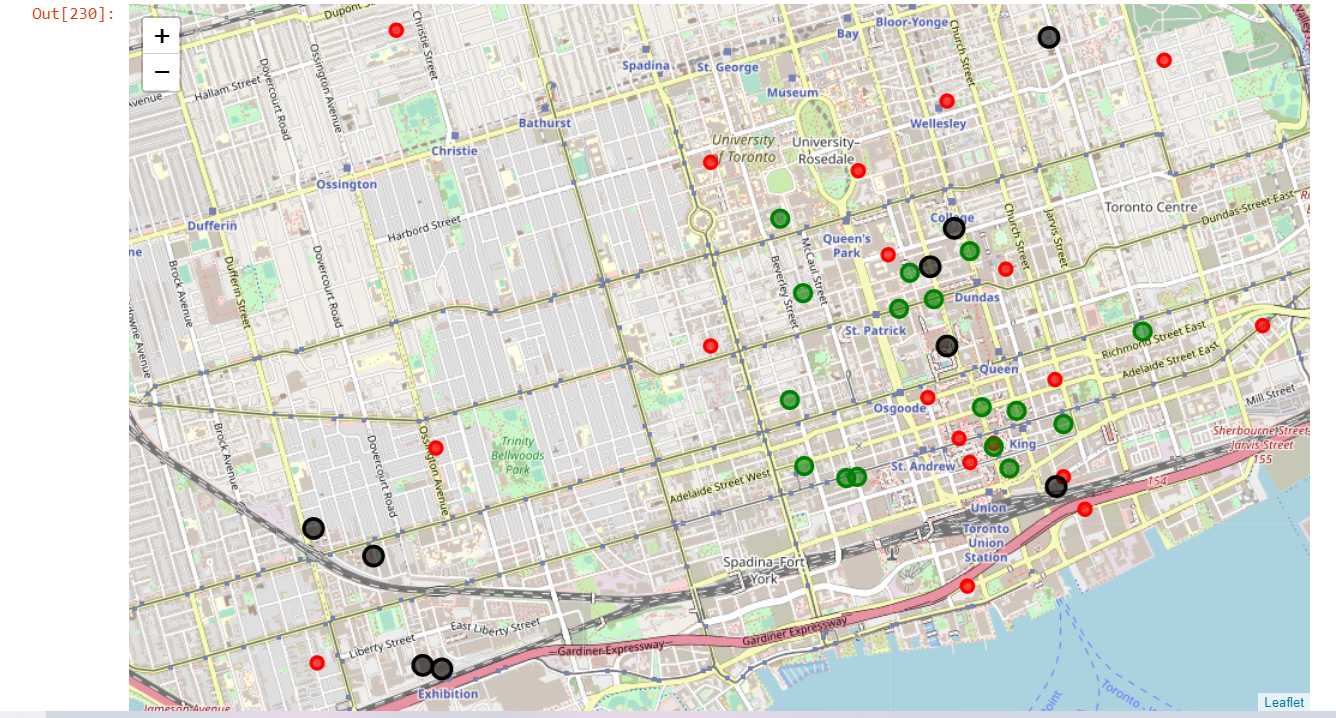
Aqua green dots: Cluster 4

Brown dots: Cluster 5

Orange dots: Cluster 6

Black dots: Apartment locations

Green dots: Indian restaurant locations

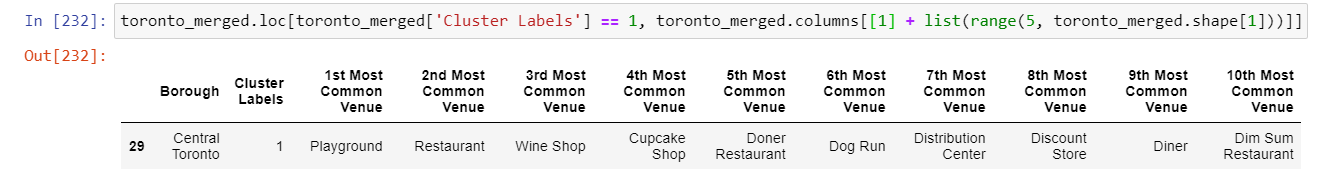


### Analysis

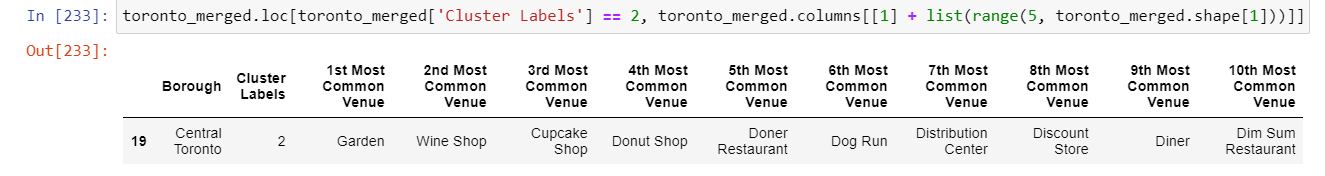
Cluster 0:



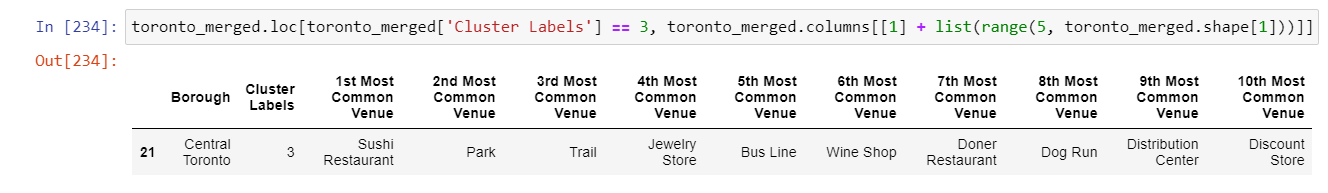
Cluster 1:



Cluster 2:



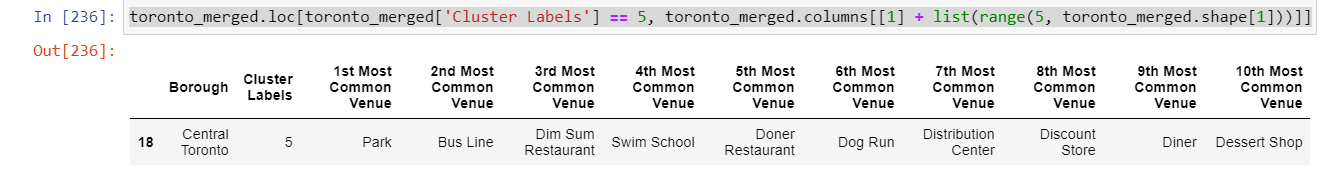
Cluster 3:



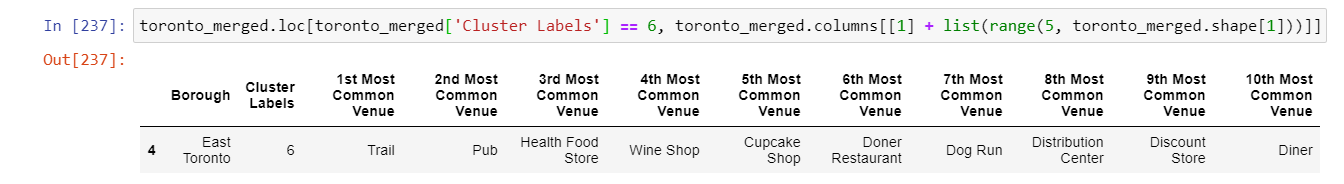
Cluster 4:



Cluster 5:

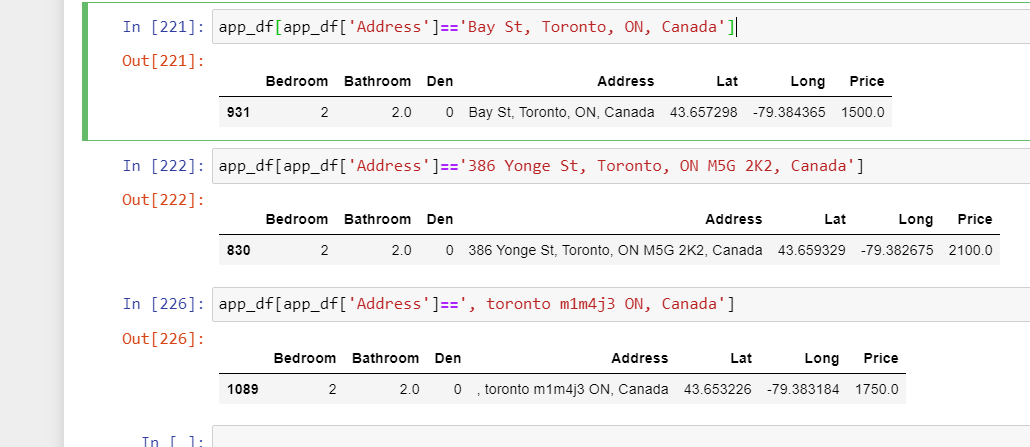


Cluster 6:



As the 7 clusters are not well diversified and Bangalore location venues do not match with cluster 1 to 6. Let us assume that Bangalore current location venues will be similar to cluster 0 in Toronto so that we have to look for apartments in cluster 0. Cluster 0 has covered most of the areas in Toronto so it won’t be much problem to find place to stay.

The apartments available in cluster 0 with Indian restaurants nearby are mentioned below.

1. Bay St, Toronto, ON, Canada
2. 386 Yonge St, Toronto, ON M5G 2K2, Canada
3. , toronto m1m4j3 ON, Canada

As apartment 1 is located in cluster 0 has 4 Indian Restaurants are nearby and the rent is CAD 1500 which is affordable. Hence this can be selected.

# Results

As per the choice of cluster, apartment location and nearby Indian restaurant apartment no 1 is selected. The selected option satisfies all our required criteria.

* The apartment is in a similar cluster as in Bangalore
* The rent is CAD 1500
* There are 2 bedrooms and 2 bathrooms
* There are multiple Indian Restaurant nearby

# Discussion

Below are the assumptions made while working on this project.

* The venues in the neighborhoods in Bangalore and Toronto may not match as per the clustering groups as the culture and lifestyle are different.
* The clusters/groups are created for the neighborhoods of Toronto with available data. These can be improved or enhanced using refining the data.
* The graph used for selection of k in Elbow method is not quite significant.
* Collection of real time data is difficult so sample data from 2018 is used for rental apartment price and location.
* This project can be replicated for analysis of any locations having Foursquare data.
* The quality of the house and furniture can vary as per Price
* Facilities available and quality of the apartments are considered as equivalent

# Conclusion

This project can be very helpful for people migrating to other city or country. This will help choosing best location of stay and finding great and interesting venues nearby. The process of learning and implementing in a real-time project has been a great experience for me. The tools used in the project are very powerful and can help anyone solving real time problems.