Battle of the Neighborhoods
‘Which Apartment Area to Rent?’
Capstone Project

Table of contents:

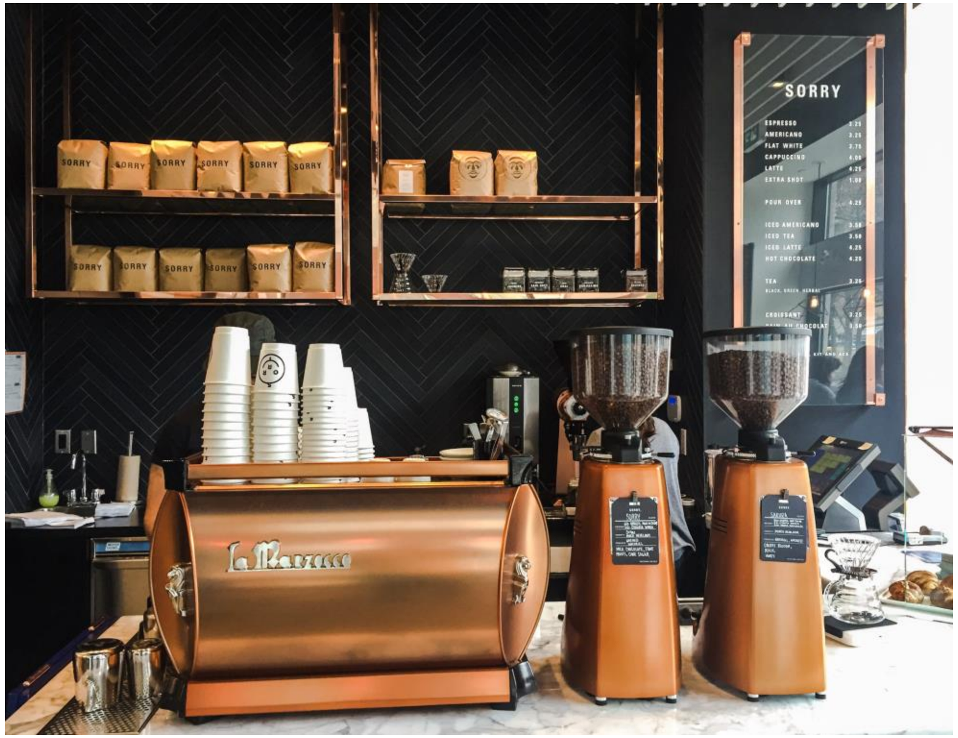
1.Introduction - Identifying business problem and background

Toronto is the provincial capital of Ontario. With population is now estimated at 6,196,731, it is the most popular city in Canada and the fourth most popular city in North America. Its food, culture, diversity, and sights to see, this capital of Ontario is a vivacious place to live.

Downtown Toronto is the main central business district of Toronto, Ontario Canada. Located entirely within the district of Old Toronto, it is approximately 17 square kilometers in area, bounded by Bloor Street to the northeast and Dupont Street to the northwest, Lake Ontario to the south, the Don Valley to the east, and Bathurst Street to the west.

Downtown Toronto is full of great neighborhoods with apartment rentals that have their own unique charm and near all the exciting events and attractions. It is also the location of the municipal government of Toronto and the Government of Ontario and home to three public universities, OCAD University, Ryerson University, and the University of Toronto.

2. Problem Statement



Source from www.google.com

Assuming we are a real estate agent in Downtown Toronto area and one of our clients from Italy came to consult on where is the best neighborhood to rent an apartment in Downtown Toronto area.

So, as a real estate agent, in order to assist our client, we need to perform an analysis in order to identify the best location for our client. Since there are also some request pertaining the area from our clients, we need to include the factors in our analysis.

Factors to be included:

- Areas with Italian restaurants and coffee shops .
- Less populated.

3. Data

3.1 List of data and sources

1. List of Toronto neighborhoods area will be taken from https://www.wikipedia.org/

2. Geographical coordinates of the neighbourhoods with the respective Postal Codes from https://cocl.us/Geospatial_data

3. List of restaurants and shops area will be taken from https://foursquare.com/

3. Data

3.2.1 Descriptions of data - Wikipedia

1. List of Toronto neighborhoods area will be taken from https://www.wikipedia.org/
   i. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. This page will provide the information on Toronto neighborhoods, boroughs and postal codes.

3. There are several steps we need to perform before we can analyze the data. Those steps are:
   i. Scrape the information from this Wikipedia page.
   ii. Wrangle the data and clean it.
   iii. Read it into a *pandas* Dataframe so that it is in a structured format as the figure below.

	Post Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Pandas Dataframe for data taken from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

3. Data

3.2.2 Descriptions of data – https://cocl.us/Geospatial_data

1. The next data source will provide us with the geographical coordinates of the neighborhoods with the respective Postal Codes

2. The data will be taken from https://cocl.us/Geospatial_data in the form of csv file.

3. Next, the list of geographical coordinates (latitude, longitude) will be merge with the list of Toronto data from Wikipedia to form a pandas Dataframe.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Data from https://cocl.us/Geospatial_data

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Merged data from https://cocl.us/Geospatial_data and https://www.wikipedia.org/

3. Data

3.2.3 Descriptions of data - Foursquare

1. The last data source will be from https://foursquare.com/.

2. We will construct a URL to send a request to the Foursquare API to search for a specific type of venues and to get trending venues around the Downtown Toronto location and construct it in a pandas Dataframe.

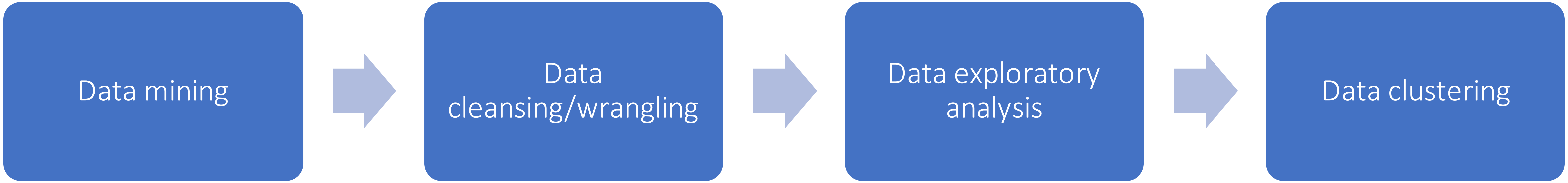
3. Next, we will acquire the information on venue category based on the list of requests given by our client from the Foursquare data.

4. Finally, the data in the Dataframe will be subject to K-Means Clustering.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
1	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

Data from https://foursquare.com/

4. Methodology



4. Methodology

4.1 Data mining

Data on Toronto boroughs and neighborhood are taken from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

The data are mined using pandas.read_html where it returns Read HTML tables into a list of DataFrame objects.

Next, the geographical data from https://cocl.us/Geospatial_data are retrieved in csv format. pd.read_csv is used to read the data and returns the data in Pandas Dataframe.

Lastly, data from https://foursquare.com/ are mined in order to search for a specific type of venues in Downtown Toronto using a URL to send a request to the Foursquare API

4. Methodology

4.2 Data cleansing/wrangling

After the data for Toronto boroughs and neighborhood is collected and converted into Pandas Dataframe, data cleansing is performed. The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood and only cells with an assigned borough will be included in this analysis. Meanwhile, the cells with borough 'Not assigned' is ignored as shown in Figure 4.1.b.

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

a

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

b

Figure 4.1: Data for Toronto boroughs and neighbourhood

4. Methodology

4.2 Data cleansing/wrangling

Next, the data for Toronto geographical coordinates that is collected from https://cocl.us/Geospatial_data as in Figure 4.2.a, are merged with the dataframe data for Toronto boroughs and neighbourhood. The merged dataframe is performed using inner join on the Postal Code column as dipicted in Figure 4.2.b.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

a

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

b

Figure 4.2: Merged dataframe for Toronto neighbourhood and geographical coordinates

4. Methodology

4.3 Data exploratory analysis

After data cleansing, data exploratory analysis is performed on Toronto area by creating a map of Toronto using latitude and longitude values using Folium visualization library. Geopy library is used to get the latitude and longitude values of Toronto as shown in Figure 4.3.

```python
import geopy
from geopy.geocoders import Nominatim

# define the city and get its latitude & longitude
city = 'Toronto'
geolocator = Nominatim(user_agent="foursquare_agent")
location = geolocator.geocode(city)
latitude = location.latitude
longitude = location.longitude

print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
#print(latitude, longitude)
```

```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

Figure 4.3: Codes for the latitude and longitude values of Toronto.

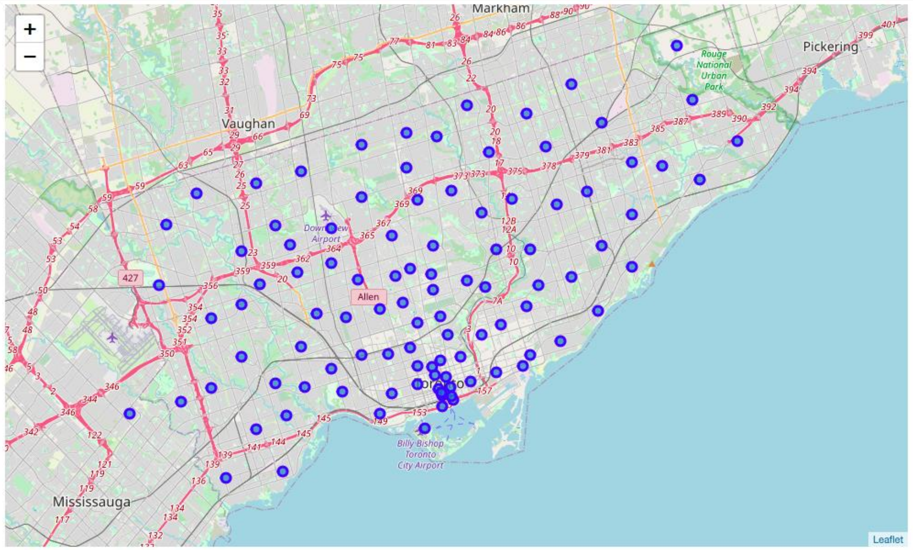


Figure 4.4: The map of Toronto area

4. Methodology

4.3 Data exploratory analysis

Beside generating the Toronto map, there are 10 boroughs and 103 neighbourhood had been identified from the Toronto dataframe collected from wikipedia. Next, from the Toronto dataframe, only the 'Borough' column contains 'Downtown Toronto' are selected and the Downtown Toronto map is generated. Foursquare API tools is used to explore the neighborhoods and venues in the Downtown Toronto.

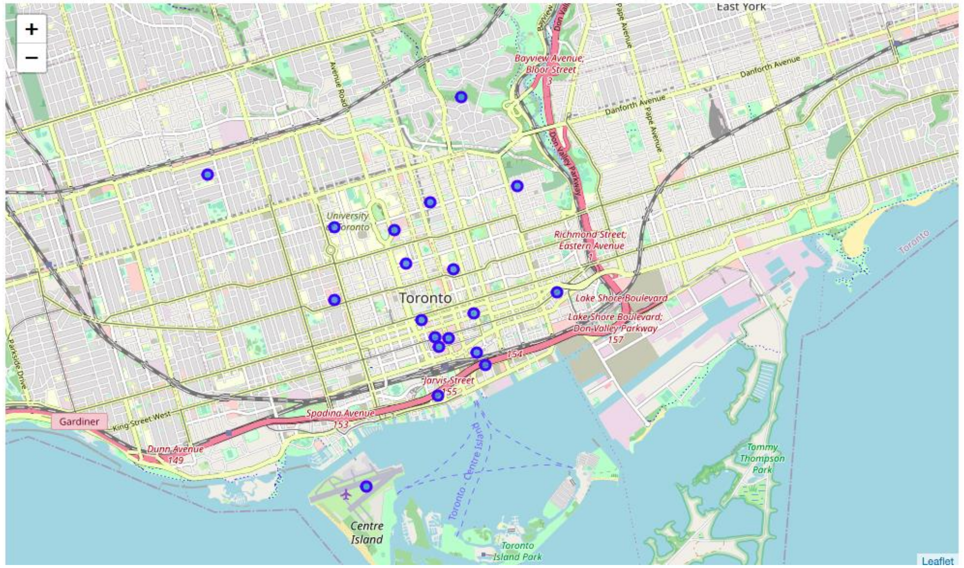


Figure 4.5: The map of Toronto area

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
1	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

Figure 4.6: Dataframe for venues for in each neighbourhood in Downtown Toronto based on Foursuqare data.

4. Methodology

4.4 Data clustering

In order to find the neighbourhood with both coffee shop and Italian restaurant, a machine learning algorithm is applied to a data set of Downtown Toronto. In this project, an unsupervised learning algorithm, K-Means clustering is used. Before running the K-Means on the dataset, a one hot encoding method in applied to the dataset. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
1	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center

Before

	Neighbourhood	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	...	Theme Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Vid Gan Sto
0	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

After

Figure 4.7: Dataframe before one hot encoding and after applying the one hot encoding

4. Methodology

4.4 Data Clustering

After performing the one hot encoding, the Dataframe is grouped by neighborhood and the mean frequency of the occurrence of each venue category is calculated. From this mean of frequency dataset, a new dataframe that only consists of coffee shop and Italian restaurant is created.

Before applying the K-Means clustering algorithm, an optimal number of clusters into which the data may be clustered needs to be determined and the **Elbow Method** is one of the most popular methods used.

	Neighbourhood	Coffee Shop	Italian Restaurant
0	Berczy Park	0.103448	0.017241
1	CN Tower, King and Spadina, Railway Lands, Har...	0.066667	0.000000
2	Central Bay Street	0.174603	0.047619
3	Christie	0.062500	0.062500
4	Church and Wellesley	0.103896	0.000000
5	Commerce Court, Victoria Hotel	0.130000	0.030000
6	First Canadian Place, Underground city	0.120000	0.010000
7	Garden District, Ryerson	0.080000	0.030000
8	Harbourfront East, Union Station, Toronto Islands	0.120000	0.020000
9	Kensington Market, Chinatown, Grange Park	0.062500	0.000000

Figure 4.8: The mean of frequency Dataframe for coffee shop and Italian restaurant category

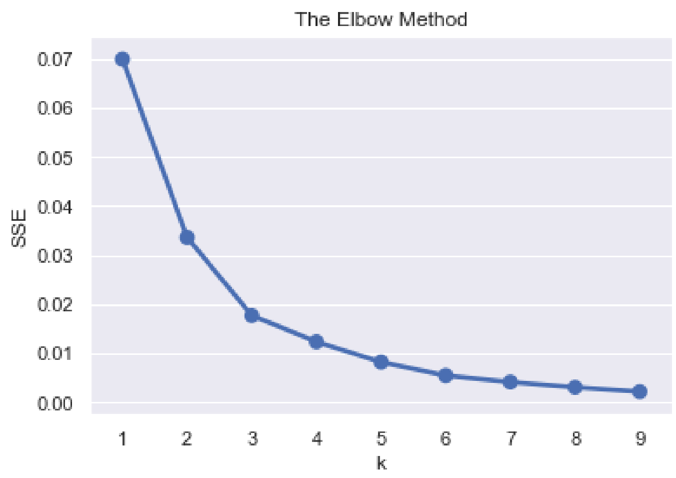
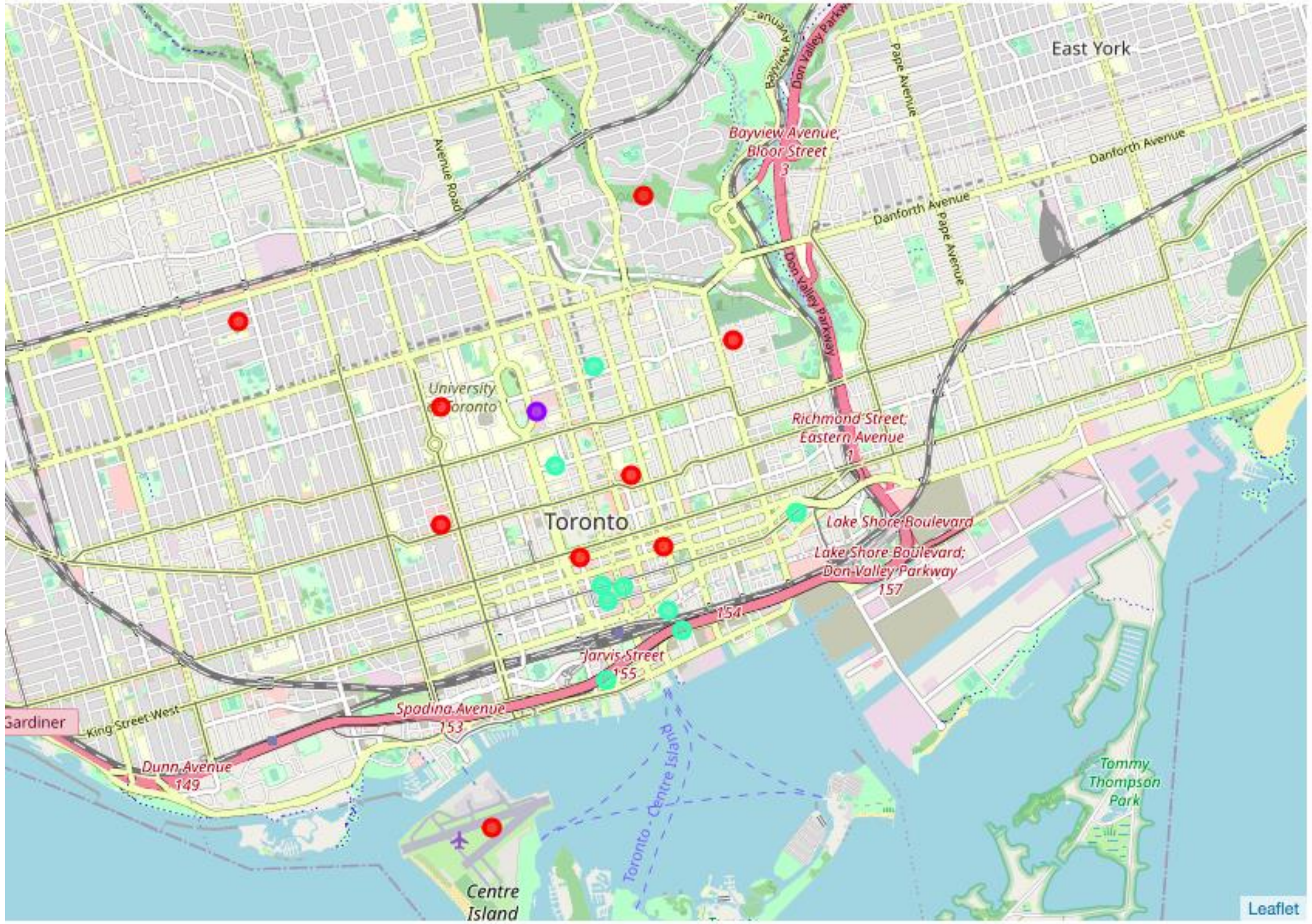


Figure 4.9: Elbow method graph, k vs SSE (Sum of Squared Errors

5. Results



5. Results

From the K-Means clustering, we have a total of 3 clusters around the Downtown Toronto neighborhood.

Red represent the Cluster = 0.
Purple represent the Cluster = 1.
Green represent the Cluster = 2.

Figure 5.1 shows the distribution of all the clusters on the Downtown Toronto area.

Figure 5.1: Downtown Toronto map with the 3-cluster marker output from K-Means algorithm.

5. Results

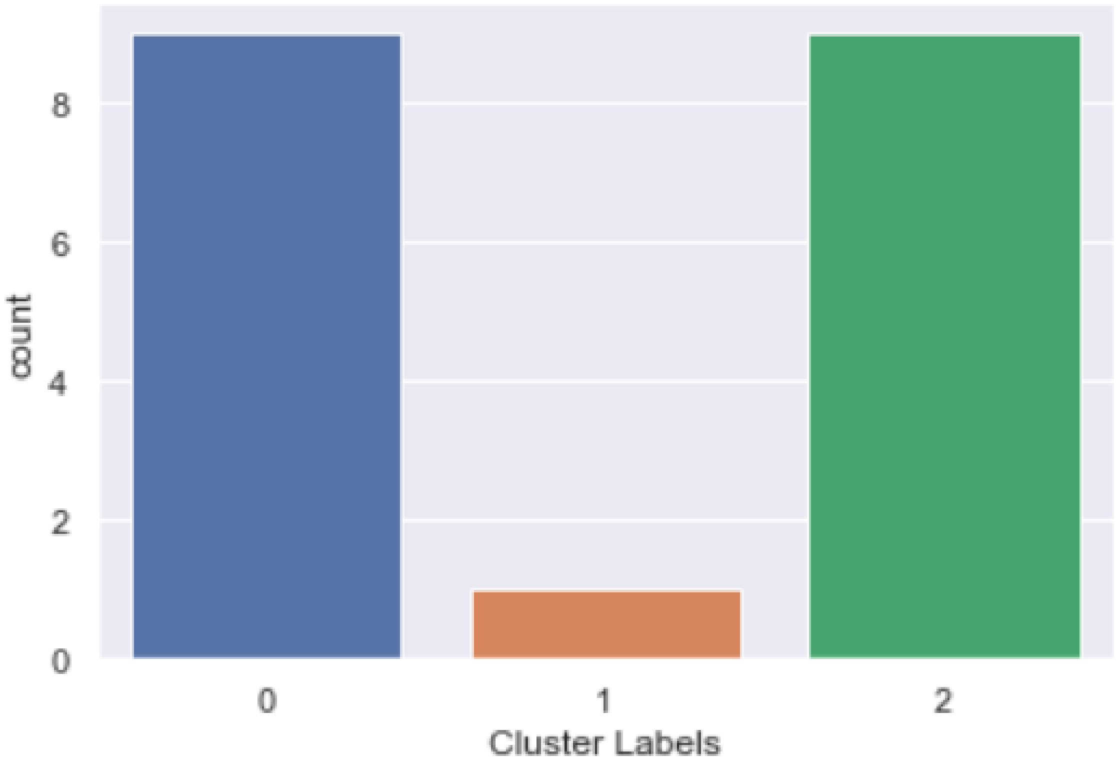


Figure 5.2 depicts the count plot of number of neighborhoods for each cluster. Here we can see that cluster 2 has the lowest number of neighborhood count which equal to 1, as in comparison to other cluster. Meanwhile, Cluster 1 and 2, have the same neighborhood count = 19.

Figure 5.2: Count plot for number of neighborhoods in each cluster.

5. Results

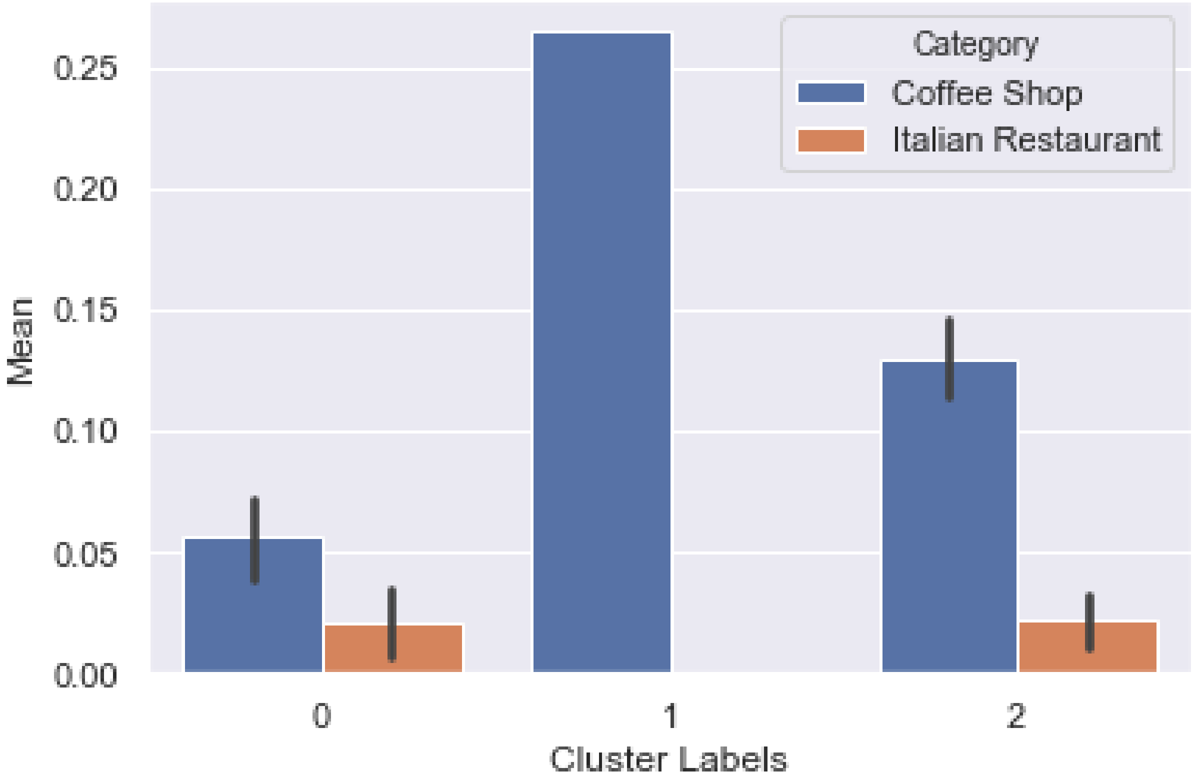


Figure 5.3 displays the mean number of coffee shop and Italian restaurant in each cluster. As shown in the chart, cluster 2 has the highest mean for coffee shop but has 0 Italian restaurant in the cluster.

While cluster 1 has mean around 0.05 for coffee shops and 0.02 for Italian restaurant and a higher mean for coffee shop is displayed in cluster 3 at around 0.12 and lower mean for Italian restaurant at 0.02.

Figure 5.3: Mean number of coffee shop and Italian restaurant in each cluster

5. Results

Cluster 1 = Red marker

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Coffee Shop	Italian Restaurant
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	0	0.080000	0.030000
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	0	0.057471	0.022989
6	M6G	Downtown Toronto	Christie	43.669542	-79.422564	0	0.062500	0.062500
7	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	0	0.090000	0.000000
11	M5S	Downtown Toronto	University of Toronto, Harbord	43.662696	-79.400049	0	0.028571	0.028571
12	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	0	0.062500	0.000000
13	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...	43.628947	-79.394420	0	0.066667	0.000000
14	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529	0	0.000000	0.000000
16	M4X	Downtown Toronto	St. James Town, Cabbagetown	43.667967	-79.367675	0	0.063830	0.042553

Cluster 2 = Purple marker

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Coffee Shop	Italian Restaurant
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1	0.264706	0.0

5. Results

Cluster 3 = Green marker

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Coffee Shop	Italian Restaurant
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	2	0.170213	0.000000
4	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	2	0.103448	0.017241
5	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	2	0.174603	0.047619
8	M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	2	0.120000	0.020000
9	M5K	Downtown Toronto	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576	2	0.130000	0.030000
10	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	2	0.130000	0.030000
15	M5W	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	2	0.115789	0.042105
17	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280	2	0.120000	0.010000
18	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	2	0.103896	0.000000

6. Discussion

From the K-means clustering results, the cluster that has both coffee shop and Italian are cluster 1 and cluster 3. Eventhough cluster 3 has a higher mean of coffee shop in comparison to cluster 1, it is also meaning that the area is also highly populated.

Back to the initial request from our client, he requested for :
1. An area with coffee shop and Italian restaurant
2. An area that is less populated.

Hence, from the clustering results, we would propose the cluster number 1 neighborhood particularly Christie neighborhood to our client. This is because Christie has an equal number of mean between the coffee shops and Italian restaurants in the area which is at 0.0625. Also Christie is in cluster 1 which is not as highly populated compared to cluster 3.

7. Conclusion

In conclusion to this project, we have shown how machine learning can be utilized in property management industries. By deploying K-Means clustering algorithm to the geographical data from foursquare and wikipedia, real estate agent can customize the needs of their client easily.