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

1.Introduction

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

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

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

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

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

3.2.1 Descriptions of data - Wikipedia

List of Toronto neighborhoods area will be taken from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. This page will provide the information on Toronto neighborhoods, boroughs and postal codes. There are several steps we need to perform before we can analyze the data.

Firstly, we need to scrape the information from this Wikipedia page. Next, data cleansing and finally read it into a pandas Dataframe. The next data source will provide us with the geographical coordinates of the neighborhoods with the respective to Postal Codes. The data will be taken from https://cocl.us/Geospatial_data in the form of csv file. Next, the list of geographical coordinates (latitude, longitude) will be merge with the list of Toronto data from Wikipedia to form a pandas Dataframe.

The last data source will be from https://foursquare.com/. 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. Next, we will acquire the information on venue category based on the list of requests given by our client from the Foursquare data. Finally, the data in the Dataframe will be subject to K-Means Clustering.

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 which 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

4.2 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: Postal Code, 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.

Next, the data for Toronto geographical coordinates are merged with the Dataframe that contains data on Toronto boroughs and neighborhood. The merged Dataframe is performed using inner join on the Postal Code column.

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

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

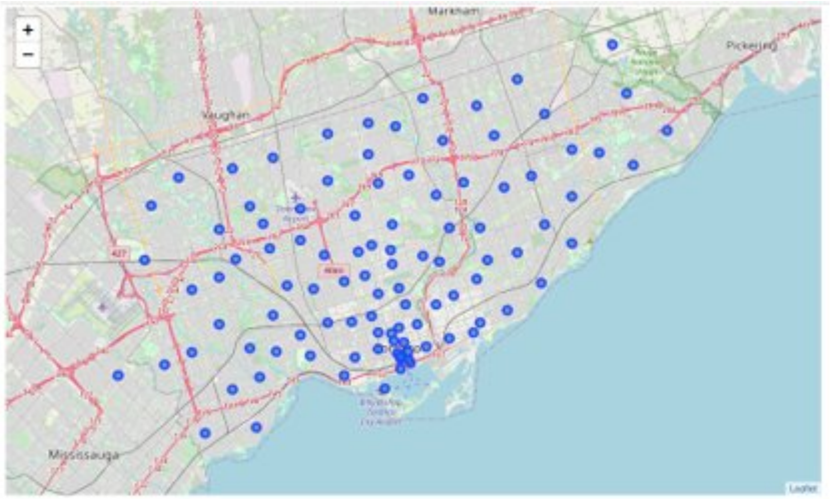


Fig. 1: Map of Toronto

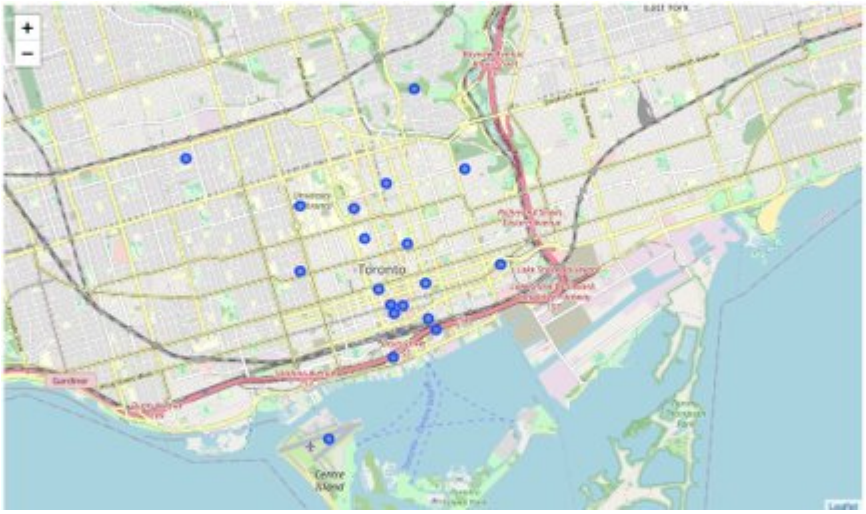
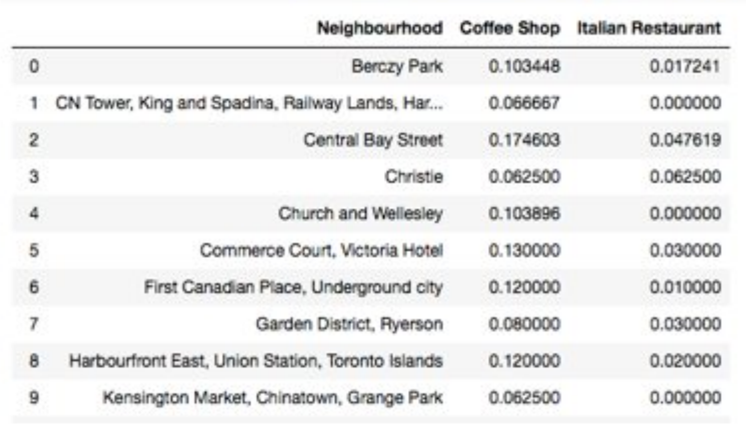


Fig. 2: Map of Downtown Toronto

4.4 Data clustering

In order to find the neighborhood 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.

After performing the one hot encoding, the Dataframe is grouped by the neighborhood value 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 as shown in Fig. 3. 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 as shown in Fig.4.



	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

Fig. 3: Downtown Toronto Dataframe

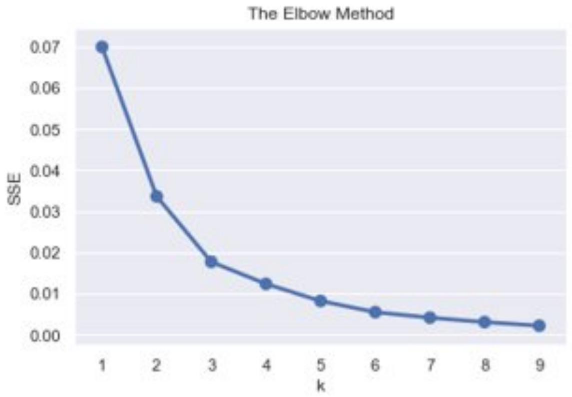


Fig. 4: Elbow method graph, k vs SSE (Sum of Squared Errors

5. Results

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

No.	Cluster number	Color
1.	0	Red
2.	1	Purple
3.	2	Green

Table 1: K-Means cluster maker

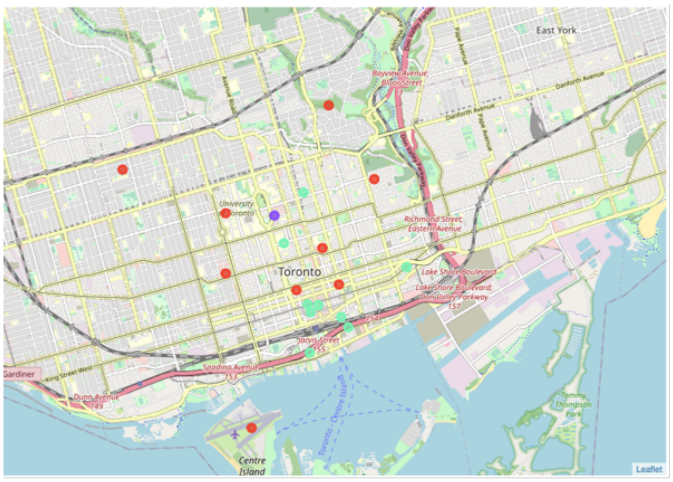


Fig.5: Downtown Toronto map with the 3-cluster marker output from K-Means.

Fig. 6 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 another cluster. Meanwhile, Cluster 1 and 2, have the same neighborhood count = 19.

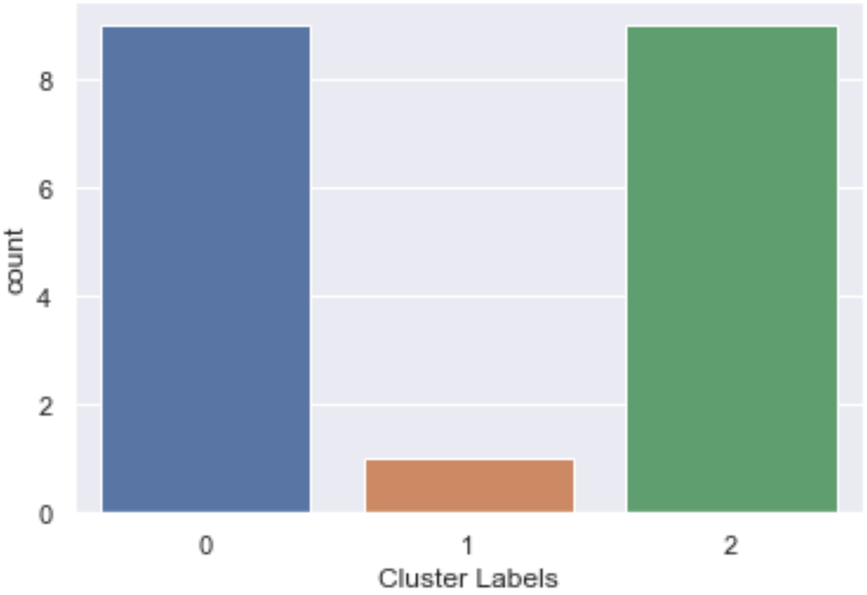


Fig. 6: Count plot for number of neighborhoods in each cluster.

Fig. 7 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.

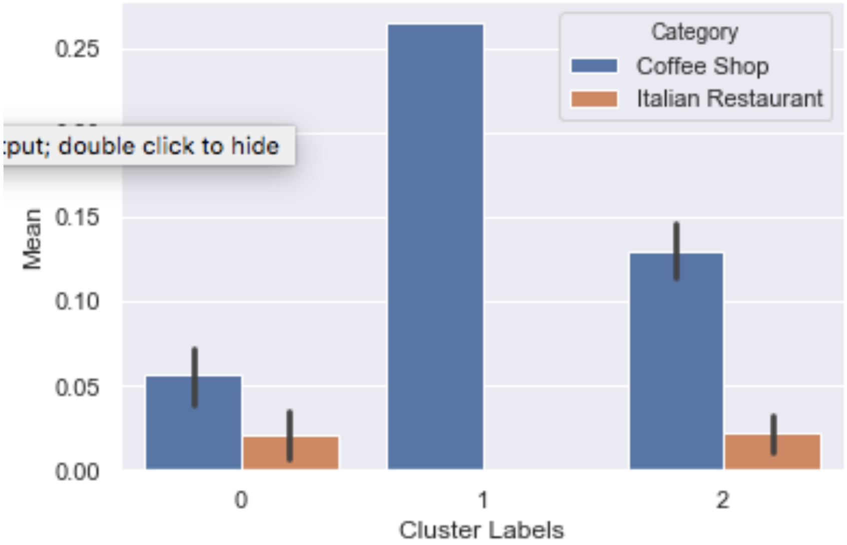
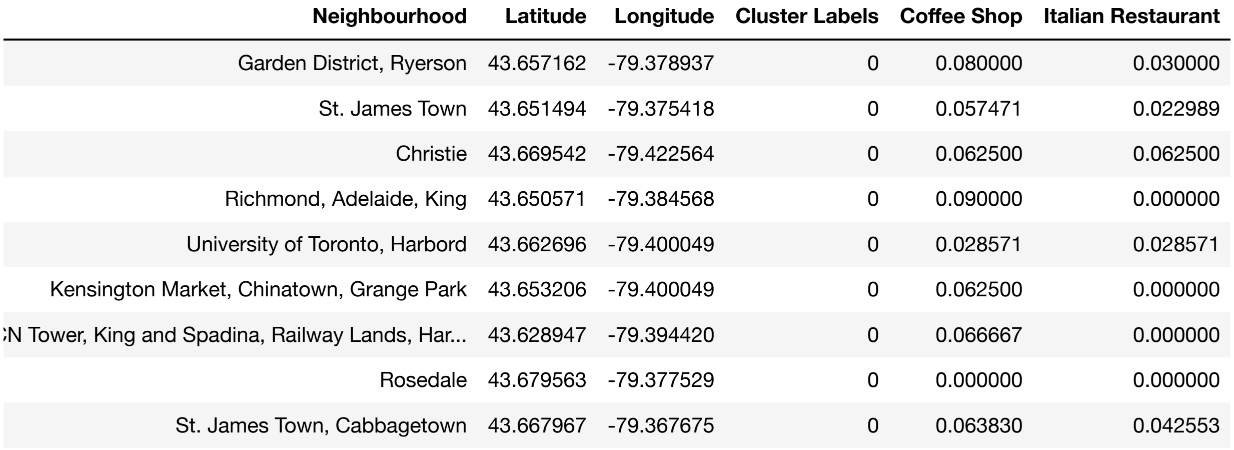


Fig.7: Mean number of coffee shop and Italian restaurant in each cluster

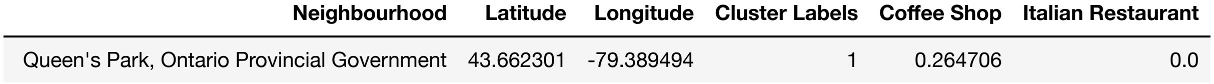
Below are the Dataframe output from the K-Means clustering.

Cluster 1:



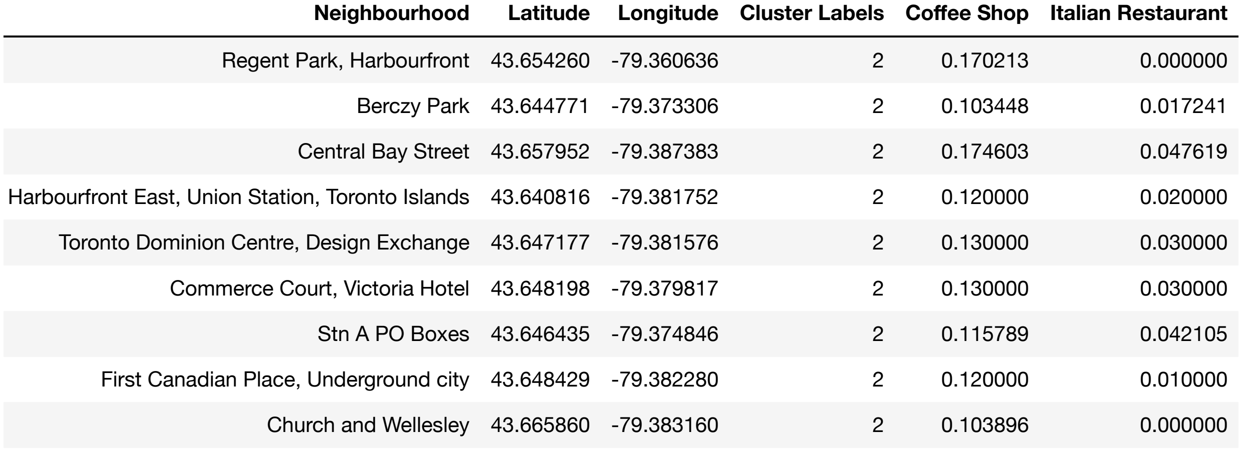
Neighbourhood	Latitude	Longitude	Cluster Labels	Coffee Shop	Italian Restaurant
Garden District, Ryerson	43.657162	-79.378937	0	0.080000	0.030000
St. James Town	43.651494	-79.375418	0	0.057471	0.022989
Christie	43.669542	-79.422564	0	0.062500	0.062500
Richmond, Adelaide, King	43.650571	-79.384568	0	0.090000	0.000000
University of Toronto, Harbord	43.662696	-79.400049	0	0.028571	0.028571
Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	0	0.062500	0.000000
:N Tower, King and Spadina, Railway Lands, Har...	43.628947	-79.394420	0	0.066667	0.000000
Rosedale	43.679563	-79.377529	0	0.000000	0.000000
St. James Town, Cabbagetown	43.667967	-79.367675	0	0.063830	0.042553

Cluster 2



Neighbourhood	Latitude	Longitude	Cluster Labels	Coffee Shop	Italian Restaurant
Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1	0.264706	0.0

Cluster 3:



Neighbourhood	Latitude	Longitude	Cluster Labels	Coffee Shop	Italian Restaurant
Regent Park, Harbourfront	43.654260	-79.360636	2	0.170213	0.000000
Berczy Park	43.644771	-79.373306	2	0.103448	0.017241
Central Bay Street	43.657952	-79.387383	2	0.174603	0.047619
Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	2	0.120000	0.020000
Toronto Dominion Centre, Design Exchange	43.647177	-79.381576	2	0.130000	0.030000
Commerce Court, Victoria Hotel	43.648198	-79.379817	2	0.130000	0.030000
Stn A PO Boxes	43.646435	-79.374846	2	0.115789	0.042105
First Canadian Place, Underground city	43.648429	-79.382280	2	0.120000	0.010000
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. Even though 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.