Contents

Battle of NeighbourHoods

Seema Negi

1 Introduction

With the introduction of industrialization and expansion of technology, relocation has become a very common practice. We are always looking forward to grow our career or looking for a better location to relocate to. The definition of a good location may differ from person to person. For some people, it could be a location with all amenities, while for some other people, it could be a location far from city chaos. There are so many factors that one has to look at while planning to relocate, that it sometimes become very difficult to look at each one of them.

Through this project, we aim to make the planning of relocation easier, by studying and analyzing the neighborhoods of Toronto city and group them into similar clusters on the basis of nearby common venues. Toronto is the provincial capital of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. It is a multi cultural place with diverse population.

2 Target Audience

The results and analysis of the project can be utilised by individuals who are planning to relocate to Toronto. The results from this project can help them in selecting the best location according to their preferred nearby venues.

3 Methodology

3.1 Data Acquisition

For proceeding with the project we need Toronto Neighbourhood data, the geographical coordinates of the neighbourhood and data regarding the most common nearby venues in each neighbourhood of data. The following list shows the various data collection sources for the project:-

1. For the Toronto neighborhood data, there is no structured data present but a Wikipedia page exists `https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M` that has all the information we need to explore and cluster the neighborhoods in Toronto. Scraping the Wikipedia page, wrangling and cleaning the data and then reading it into a pandas dataframe can result in a structured format.

2. To get the geographical coordinates of the neighborhoods , we will be using a csv file `http://cocl.us/Geospatial_data` that has the geographical coordinates of each postal code.

3. We will be using the explore function to get the most common venue categories in each neighborhood of Toronto, and then use this feature to group the neighborhoods into clusters.

3.2 Data Preprocessing

- As discussed in the above section, we need to extract the neighbourhood data of Toronto using Web Scraping. Web Scraping is a method for extracting data from web pages.
  Python's ***urllib*** library can be used to fetch the HTML from the URL that we want to scrape. Once ***urllib.request*** has pulled in the content from the URL, python's ***BeautifulSoup*** library can be used to extract and work with the data within it.

- As it can be seen in Figure 1, some values in column 'Borough' are "Not Assigned", the rows containing these values are removed from the data frame.

Postal Code	Borough	Neighborhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned

Figure 1: Table from Wikipedia Page

- The first five values from resulting data frame is shown in Figure 2

	PostalCode	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

Figure 2: Resulting Data frame values

- The csv containing the geographical coordinates of the neighborhood is read using python's pandas library. The resulting data frame is further merged with the the dataframe containing data of neighbourhood. The figure 3 shows the first five values of the dataframe containing the geographical coordinates and figure 4 shows the first five values of the merged data frame.
- Python's geopy library can be used to get the geographical coordinates(longitude, latitude) of Toronto. These coordinates can be further

	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

Figure 3: Geographical Coordinates of the neighborhood

	PostalCode	Borough	Neighborhood	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

Figure 4: Merged dataframe

used by 'folium' to create the map of Toronto. Figure 5 shows the map created.

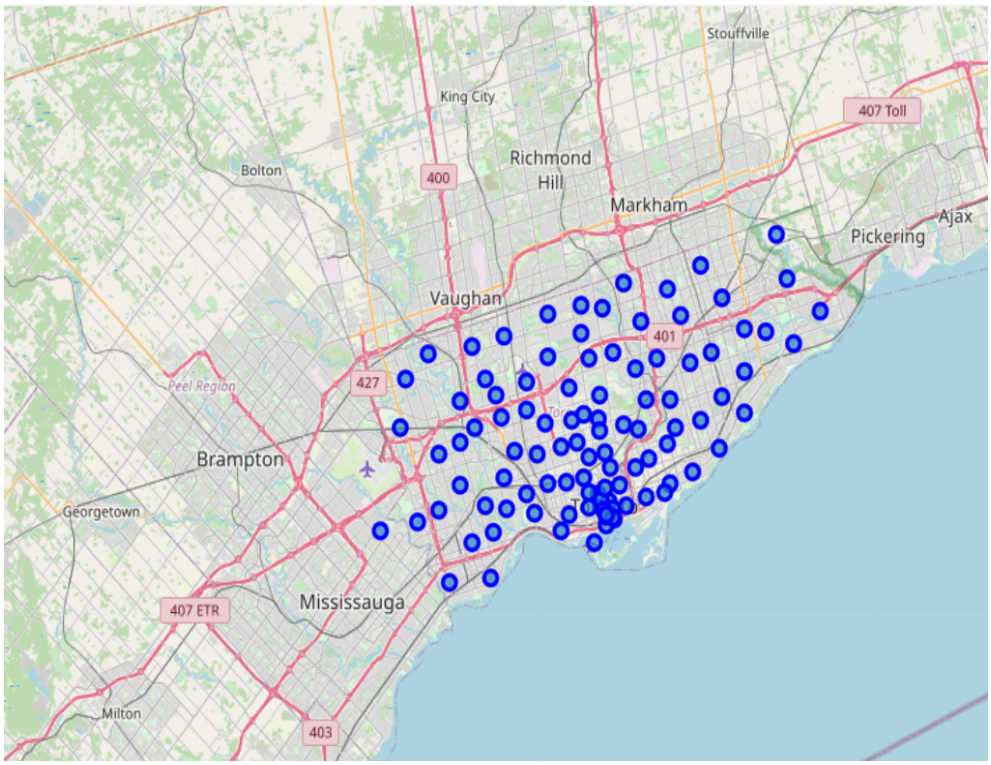

Figure 5: Map of Toronto with neighborhoods superimposed on top

3.3 Feature Extraction

FourSquare API can be used to fetch the venues within a radius of neighbourhood. For our project, we defined the radius of 500. FourSquare is a location data provider. Location data is data describing places and venues, such as their geographical location, their category, working hours, full address, and so on, such that for a given location given in the form of its geographical coordinates (or latitude and longitude values) one is able to determine what types of venues exist within a defined radius from that location.Further information regarding these venues like latitude, longitude, venue category can also be extracted using the API.

3.4 Machine Learning

To analyze the data we performed a technique called One hot encoding in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. Further, we grouped the data by Neighborhood and by taking the average of the frequency of occurrence of each Venue Category. The most common venues extracted act as features for our machine learning algorithm.

3.4.1 K-means Clustering

K-means clustering is an unsupervised machine learning algorithm. It divides the data into k non-overlapping subsets or clusters. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. We can create a k-means clustering model using python's sklearn library. For this project, we selected the value of number of clusters as 5.

4 Results

The Figure 6 shows the different clusters obtained by our algorithm.Neighbourhood within the same cluster are denoted by same colour.It can be observed from the map, the cluster denoted by blue colour contains the highest number of neighbourhoods.

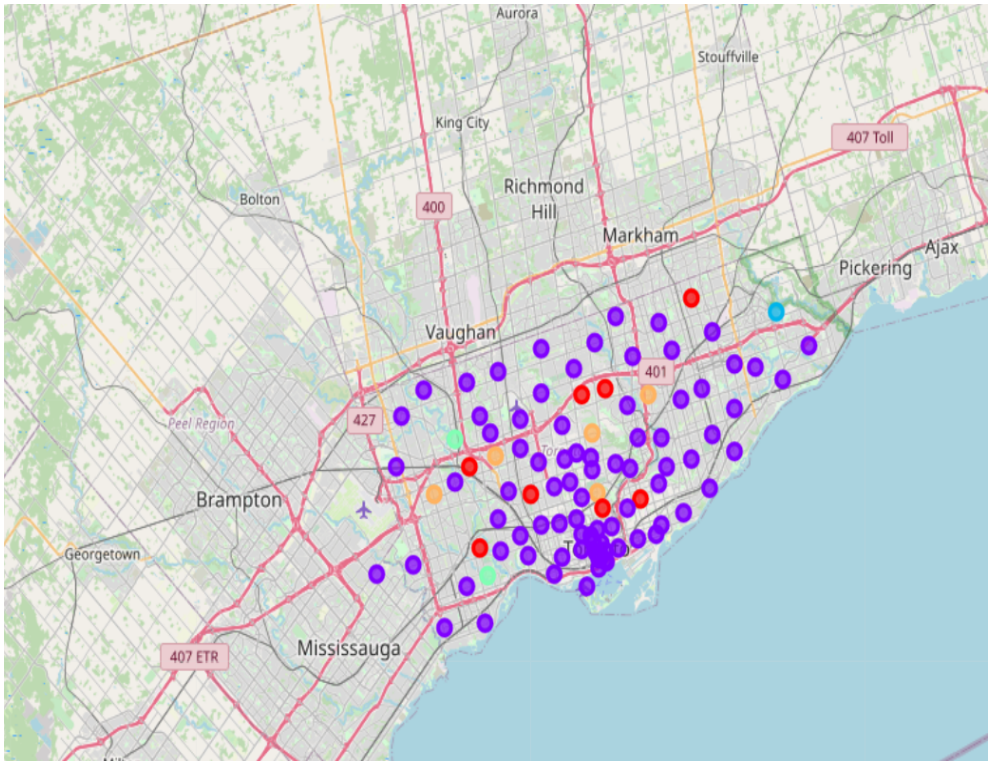

Figure 6: Map showing different clusters

Figure 7 shows the results from the first cluster. It can be seen that the first cluster contains neighbourhoods where most common venues are Park, Pool, Playground, Departmental Store etc

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
21	York	0	Park	Pool	Women's Store	College Stadium	Colombian Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Drugstore
35	East York	0	Park	Convenience Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Deli / Bodega
45	North York	0	Park	Dog Run	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Doner Restaurant	Grocery Store
64	York	0	Park	Dog Run	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Doner Restaurant	Grocery Store
66	North York	0	Park	Convenience Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Deli / Bodega
85	Scarborough	0	Park	Playground	Dog Run	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Doner Restaurant
91	Downtown Toronto	0	Park	Trail	Playground	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center
98	Etobicoke	0	Park	River	Dog Run	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Doner Restaurant

Figure 7: Cluster 1

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	North York	1	Portuguese Restaurant	Hockey Arena	Coffee Shop	Intersection	Financial or Legal Service	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run
2	Downtown Toronto	1	Coffee Shop	Pub	Bakery	Park	Theater	Breakfast Spot	Restaurant	Café	Bank	Hotel
3	North York	1	Clothing Store	Miscellaneous Shop	Accessories Store	Boutique	Vietnamese Restaurant	Coffee Shop	Shoe Store	Event Space	Furniture / Home Store	Distribution Center
4	Downtown Toronto	1	Coffee Shop	Sushi Restaurant	Diner	Gym	Discount Store	Sandwich Place	Park	Mexican Restaurant	Italian Restaurant	Hobby Shop
7	North York	1	Gym	Beer Store	Asian Restaurant	Japanese Restaurant	Restaurant	Coffee Shop	Caribbean Restaurant	Café	Dim Sum Restaurant	Discount Store
8	East York	1	Pizza Place	Fast Food Restaurant	Bank	Athletics & Sports	Intersection	Gastropub	Café	Pharmacy	Gym / Fitness Center	Drugstore
9	Downtown Toronto	1	Clothing Store	Coffee Shop	Middle Eastern Restaurant	Bubble Tea Shop	Café	Japanese Restaurant	Cosmetics Shop	Theater	Bakery	Lingerie Store

Figure 8: Cluster 2

Figure 8 shows the results from the second cluster. It can be seen that the second cluster contains neighbourhoods where most common venues are Restauraunts, Coffee shops, cafe, pizza shops and similar venues.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
6	Scarborough	2	Fast Food Restaurant	Doner Restaurant	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Donut Shop	Farm

Figure 9: Cluster 3

Figure 9 shows the results from the third cluster. It can be seen that the third cluster contains only one neighbourhoods.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
57	North York	3	Baseball Field	Food Service	Women's Store	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Dessert Shop
101	Etobicoke	3	Baseball Field	Donut Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Women's Store	Farmers Market

Figure 10: Cluster 4

Figure 10 shows the results from the fourth cluster. It can be seen that the cluster contains two neighbourhoods with almost similar common venues.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	North York	4	Park	Food & Drink Shop	Construction & Landscaping	Electronics Store	Eastern European Restaurant	Drugstore	Donut Shop	Doner Restaurant	Deli / Bodega	Dog Run
49	North York	4	Park	Bakery	Construction & Landscaping	Basketball Court	Trail	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center
61	Central Toronto	4	Park	Swim School	Bus Line	Distribution Center	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Deli / Bodega
77	Etobicoke	4	Park	Bus Line	Pizza Place	Sandwich Place	Discount Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Distribution Center
83	Central Toronto	4	Park	Trail	Tennis Court	Restaurant	Eastern European Restaurant	Drugstore	Donut Shop	Doner Restaurant	Dog Run	Dance Studio

Figure 11: Cluster 5

Figure 11 shows the results from the fifth cluster. It can be seen that the cluster contains neighbourhoods with common venues Park, Bus Line, Dog Run, Distribution Centre etc.

5 Discussion

The results can be used to decide a place for relocation on the basis of common venues present. It can be seen in the results, that cluster 7 and 11 have many venues in common while cluster 9 can be integrated with cluster 8 also. The cluster 2 contains mostly the neighbourhoods which have Restaurants of different types like Portuguese, Fast food, Japanese, Vietnamese, Mexican, Middle east etc, cafes, coffee shops, beer stores and similar venues. If a person wants to relocate to a place which is near to these venues or restaurants, he/she can select neighbourhoods from this cluster while if someone wants to relocate to a place with venues like Parks, Dog Run, Playground, Pools and similar can select a neighbourhood from 1st cluster.

6 Conclusion

In conclusion, the project provides great insights about how k-means clustering can be used to group similar data and solve various business problems. If data is pre-processed and cleaned properly, the machine learning algorithms can provide good results. The project also make use of Python's rich libraries to fetch the information, perform analysis, visualisation and modelling. FourSquare API is used to explore the nearby venues in the neighbourhood of Toronto. Folium is used to create the maps using the geographic coordinates of the location.
The project can be further improved in future by making use of different machine learning algorithm. Currently, the assumptions are made only on the basis of data fetched by FourSquare API. In future, data from different sources can be collected and integrated. Also, more features can be extracted on the basis of underlying business problem.

7 References

1. Wikipedia contributors. (2020, June 28). Toronto. In Wikipedia, The Free Encyclopedia. Retrieved 14:49, July 6, 2020, from `https://en.wikipedia.org/w/index.php?title=Toronto&oldid=964883458`