Capstone Project Report

Finding the best Neighborhoods for Foreign Students in Toronto

Muhammad Abaidullah Anwar

The Capstone Project report submitted in partial fulfilment of

IBM Certification

of

Applied Data Science

February 2021

1. Problem Description

Toronto being the capital city of the Canadian province of Ontario has been recorded as one of the most populous city with a population of 6,254,571 as of 2021 in Canada and the fourth most populous city in North America. The city is surrounding the western end of Lake Ontario is an international center of business, finance, arts, and culture, education, and is recognized as one of the most multicultural and cosmopolitan cities in the world. There are 140 neighbourhoods officially recognized by the City of Toronto and upwards of 240 official and unofficial neighborhoods within the city's boundaries. These are Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown, and many more.

The immigrants during their initial few years and the foreign students always need to find a convenient locality to live which is close to the educational institutions, restaurants of their choices, and shopping facilities. The main objective of this Capstone Project is to identify the neighbourhoods of Toronto for the foreign students from India and Pakistan, particularly, to find a reasonable cost effective residence which is near to their universities, Indian or Pakistani dining facilities, and the shopping malls. There are many cuisines which are difficult to distinguish as Indian or Pakistani and therefore it will mainly serve for both types of foreign students.

We would like answer the question “What is the best convenient neighborhood for a Pakistani or Indian foreign students in Toronto?”.

2. Data Description

The data of Toronto available from Wikipedia page that contain Borough, Neighborhoods, and that has all the information we need to explore and identify the neighborhoods in Toronto. However, this data is not in a format that could be directly used for my Capstone Project. The data must be preprocessed before using it in any project of extracting the useful and required information and knowledge. The data of Toronto on Wikipedia page will be scrapped, wrangled, clean it, and then read it into a pandas dataframe so that it is in a structured format. The additional information of latitudes and longitudes of each neighborhoods will combined with the data from file "CoordinatesToronto.csv" as read from the http://cocl.us/Geospatial_data is used to add the required coordinates of the neighbourhoods.

The Foursquare API will be used to find and filter the neighborhoods for convenient residences for the foreign students from Pakistan and India. The Machine Learning algorithm i.e. Clustering will be used to generate the clusters of interests in this project including universities, Indian and/or Pakistani restaurants, and shopping malls.

3. Foursquare API

We will need data about different venues in different neighbourhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighbourhood. For each neighbourhood, we have chosen the radius to be 500 or 1000 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

- Neighbourhood: Name of the Neighbourhood
- Neighbourhood Latitude: Latitude of the Neighbourhood
- Neighbourhood Longitude: Longitude of the Neighbourhood
- Venue: Name of the Venue
- Venue Latitude: Latitude of Venue
- Venue Longitude: Longitude of Venue
- Venue Category: Category of Venue

Based on all the information collected for both London and Paris, we have sufficient data to build our model. We cluster the neighbourhoods together based on similar venue categories. We then present our observations and findings. Using this data, our stakeholders can take the necessary decision.

4. Methodology

The Python has a long list of useful packages that could be used to explore and analyze the data for multiple purpose. The necessary and required packages and libraries being used in the Capstone Project will be installed and imported. The brief detail of the packages and libraries is given as follows:

- numpy – library to handle data in a vectorized manner
- Beautiful Soup – for pulling data out of HTML and XML files
- pandas as – library for data analsysis
- json – library to handle JSON files
- Nominatim – convert an address into latitude and longitude values
- requests – library to handle requests
- json_normalize – tranform JSON file into a pandas dataframe
- Matplotlib and associated plotting modules
- k-means from clustering stage
- folium – map rendering library

4.1. Data Collection

The data of Toronto using postal codes and of ‘M’ has been scrapped by reading the data from Wikipedia page “https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. This website provides a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario.

We cleaned the data and then read it into a *pandas* dataframe so that it is in a structured format and could be used effectively and easily for the analysis.

4.2. Data Preprocessing

We used and processed the cells that have an assigned borough and ignore cells with a borough that is Not assigned. Also if a cell had a borough but a "Not assigned" neighborhood, then the neighborhood was assigned the same as the borough.

4.3. Feature Selection

Since we needed the only the borough, neighbourhood, postal codes and geolocations (latitude and longitude) therefore end up selecting the columns that we needed.

We will not present the code for any activity but only show the output of the codes. Figure 1 shows that data retrieved from the Wikipedia. Note that "\n" at the end of each information in each cell. We removed the \n" by using *str.replace()* function and the updated data is shown in Figure 2.

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

Figure 1: Neighborhoods and Postal Codes

	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: Neighborhoods and Postal Codes

The file "CoordinatesToronto.csv" as read from the http://cocl.us/Geospatial_data is used to add the required coordinates of the neighbourhoods. Figure 3 shows the latitude and longitude of each neighbourhoods.

	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 3: Neighborhoods and Postal Codes

4.4. Visualizing the Neighbourhoods of Toronto for selected Categories

We used the Colleges and University Category ID to show the map of Toronto city with 56 higher educational institutions. Figure 4 shows the map with detail of one of the institutions.

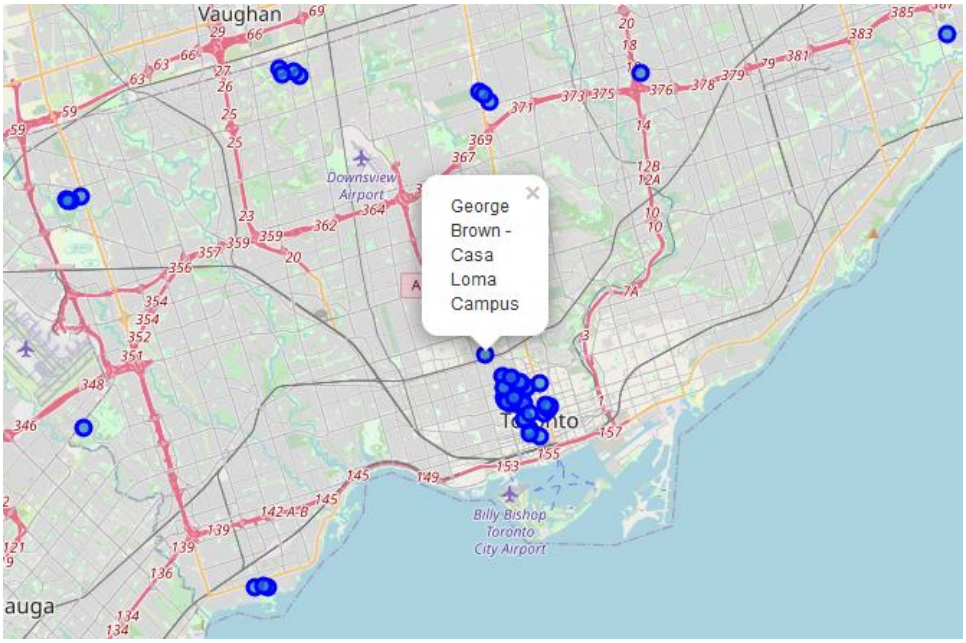

Figure 4: Neighborhoods and Postal Codes

Similarly, we wrote the codes for finding and displaying the Indian restaurants (143), Pakistani restaurants (13), and Halal restaurants (58) in the Toronto city. The maps of these were also generated, which are not shown here, but could be viewed in the notebook on GitHub.

We also explored the region of interest of Toronto (DENC) as follows:

- D - Downtown
- E - East
- N - North
- C – Central

The result is shown in Figure 5 and we found that there are 47 unique categories of interest in DENC.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.654260	-79.360636	George Brown College - SJG Building	43.651888	-79.365574	College Academic Building
1	Regent Park, Harbourfront	43.654260	-79.360636	George Brown College - School of ESL	43.651872	-79.365580	College Academic Building
2	Regent Park, Harbourfront	43.654260	-79.360636	George Brown School Of Design	43.651895	-79.365601	College Technology Building
3	Regent Park, Harbourfront	43.654260	-79.360636	George Brown School of Design	43.651871	-79.365797	College Technology Building
4	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Colaba Junction	43.660940	-79.385635	Indian Restaurant
5	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	The Halal Guys	43.665101	-79.384684	Halal Restaurant
6	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Tandori	43.660377	-79.384680	Indian Restaurant
7	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Dalla Lana School of Public Health	43.659232	-79.393254	College & University

Figure 5: Neighborhoods and Postal Codes

4.5. One Hot Encoding

The main purpose was to find out what are the different kinds of venue categories present in each neighbourhood and then calculate the top 100 or maximum which may be less the 100 common

venues, we use the *One Hot Encoding* to work with our categorical datatype of the venue categories. This was required to convert the categorical data into numeric data. A snapshot of the output is shown in Figure 6.

	Neighborhood	Adult Education Center	Art Gallery	Church	Coffee Shop	College & University	College Academic Building	College Administrative Building	College Arts Building	College Auditorium	College Bookstore	College Cafeteria	College Classroom	College Communications Building	College Engineering Building	College Football Field
0	Berczy Park	0.0	0.0	0.0	0.0	0.00000	0.000000	0.00000	0.0	0.0	0.0	0.0	0.25	0.0	0.0	0.0
1	Brockton, Parkdale Village, Exhibition Place	0.0	0.0	0.0	0.0	0.00000	0.000000	0.00000	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0
2	Central Bay Street	0.0	0.0	0.0	0.0	0.02439	0.195122	0.04878	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0
3	Christie	0.0	0.0	0.0	0.0	0.00000	0.000000	0.00000	0.0	0.0	0.0	0.0	1.00	0.0	0.0	0.0
4	Church and Wellesley	0.0	0.0	0.0	0.0	0.00000	0.050000	0.00000	0.0	0.0	0.0	0.0	0.00	0.1	0.0	0.0

Figure 6: Neighborhoods and Postal Codes

4.6. Top Venues in the Neighbourhoods

We identified, ranked, and label the top ten (10) venues categories in our neighborhood. A snapshot of the output is shown in Figure 6.

	Neighborhood	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	Berczy Park	College Classroom	Indian Restaurant	Student Center	North Indian Restaurant	Community College	General College & University	University	College Math Building	College Library	College Lab
1	Brockton, Parkdale Village, Exhibition Place	Trade School	College Theater	College Lab	College Residence Hall	College Rec Center	College Quad	College Math Building	College Library	College Gym	College Football Field
2	Central Bay Street	College Academic Building	Indian Restaurant	University	Student Center	College Science Building	College Administrative Building	College Lab	Medical School	Government Building	College & University
3	Christie	College Classroom	University	College Rec Center	College Quad	College Math Building	College Library	College Lab	College Gym	College Football Field	College Engineering Building
4	Church and Wellesley	University	Indian Restaurant	General College & University	College Communications Building	Halal Restaurant	High School	College Residence Hall	Trade School	Performing Arts Venue	College Track

Figure 7: Neighborhoods and Postal Codes

4.7. Clustering using KMeans and Visualization

The exciting phase of this Capstone Project was to use the machine learning algorithm KMeans to build five (5) clusters of neighbourhoods. The clusters are shown in the Figure 8.

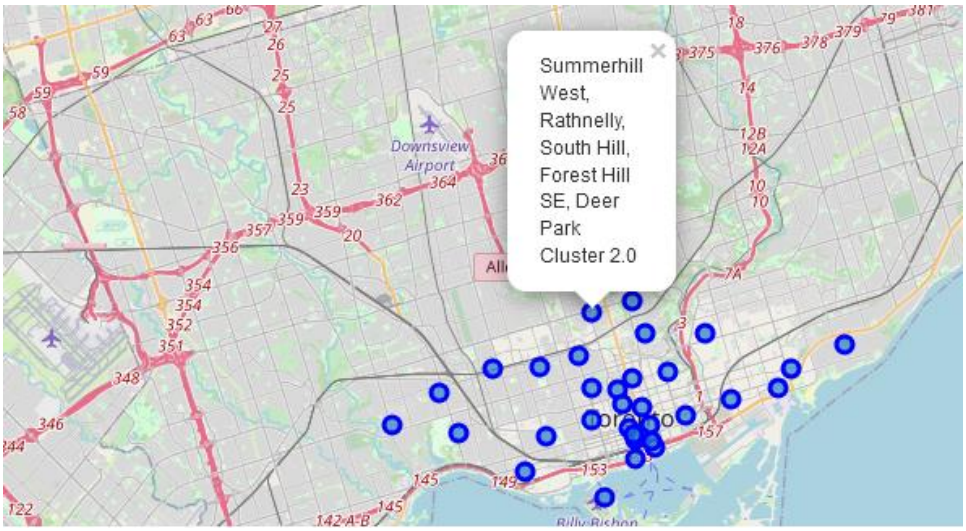

Figure 8: Neighborhoods and Postal Codes

5. Results and Discussion

We generated five (05) clusters and the largest cluster is fourth cluster that includes thirteen (13) neighbourhoods and the second largest cluster is the third cluster which includes seven (07) clusters. The clusters include the information of higher educational institutions and Indian, Pakistani, and Halal restaurants in the Toronto City.

6. Conclusion

We presented the analysis of Colleges, Universities, and different restaurants that has always been useful information for foreign students when come to Toronto for their studies. This information is also useful for the expatriate families living in the city. We explored both the cities based on their postal codes and then extrapolated the common venues present in each of the neighbourhoods finally concluding with clustering similar neighbourhoods together. We could see that each of the neighbourhoods in both the cities have a wide variety of experiences to offer which is unique in its own way.

7. References

1. https://developer.foursquare.com/docs/build-with-foursquare/categories/
2. https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto
3. http://cocl.us/Geospatial_data
4. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M