Predicting Neighborhood for Living in Toronto

1. Introduction

1.1 Background

Toronto is the provincial capital of Ontario and the most populous city in Canada. A global city, Toronto is a centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

As the city being giving large opportunities in every field, we will choose Toronto for the project to determine the better Resident Neighborhood to live. If we think of the city residents, they may want to choose the regions where they find all necessities. Now-a-days finding the residence with the categories i.e, Schools, Finance sector of jobs for future, Grocery store and Transportation nearby is getting the real world problem. Everyone wants to consume less time for getting daily routine stuff done. We will find all possible categories available in each Neighborhood of Toronto. All categories plays important role in recent and future routine life.

When we consider all these scenarios, we will create a map and information chart to show the Neighborhood for resident area. This information obtained can be used for further analysis to find residence.

1.2 Problem

Data that might contribute to determine maximum number of categories available in each Neighborhood which will help us to determine the Neighborhood for residence. This project aims to predict the Neighborhood that have all necessities close to find the residence.

2. Data Description

To consider the problem we can list the datas as below:

- Will use the List of Postal Codes of the Canada from Wikipedia link as https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,. The web scraping will be done to convert the web page to structured tabular format. The file will have Postal codes, Borough & Neighborhood. The table will need to be clean to get the Postal codes for Toronto only.

- Using http://cocl.us/Geospatial_data link will get the coordinates of the all Postal codes of Canada. I will clean the data and reduced it to city of Toronto where I

merge the Postal codes and location with above table having Postal codes, Borough and Neighborhood.

- I will use **Forsquare API** to get the most common venues of given Borough of Toronto.

- I used Google Map, ‘Search Nearby’ option to get the center coordinates of the each Borough.

2.1 Data Cleaning and Feature Selection

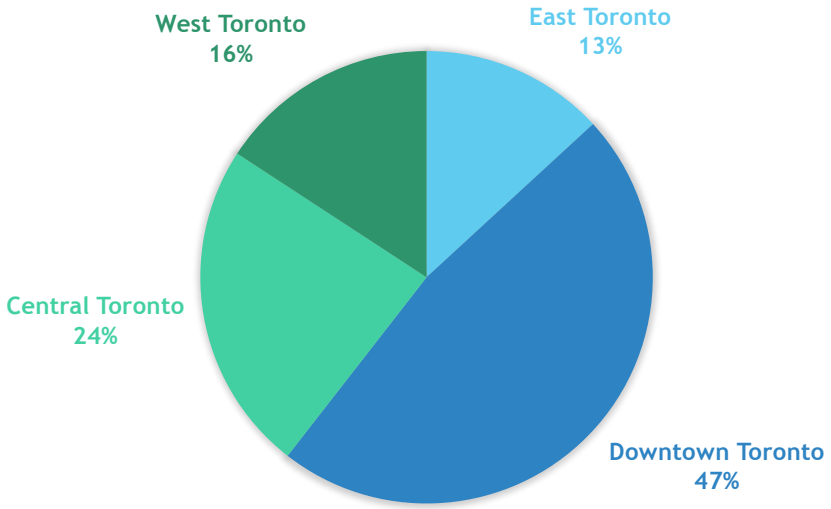
Data downloaded and data web scraped from multiple sources were combined into one table. From the link, the data was web scraped using beautiful soup function and structured in tabular format which is dataframe. The data frame contained the Postal code, Borough and Neighborhood. So initially to clean up the data, removed the data from Borough that didn’t have any name assigned in Bororugh. After we renamed all Neighborhood that didn’t have any name irrespective to Borough. At that point we copied the name of the borough to the name of the neighborhood. The data frame was then filtered by deleting unwanted rows and column. Finally we got the data to start further process.

To get the location of all the neighborhood, the data was read from csv file through pandas method in dataframe. Having postal code same in both dataframe, both the data frame was combined to get the location of each neighborhood and postal code.

Now we are moving to our aim to identify the different categories venue available in each Neighborhood. To get the different venues we are going to use Foursquare data location. We got all different categories venues in each Neighborhood. As now-a-days people when they are migrating to a place and when they are searching residence, they are actually looking the surrounding. They are looking for necessities things availability like grocery Store, Gyn, restaurant, School and many things. Usually they find the place where maximum venue are near by. So now we will filter the venues that we are looking from all general venues.

3. Exploratory Data Analysis

In this project we will direct our efforts on detecting areas of Toronto Neighborhood and we will limit our analysis to area ~500 meters around Neighborhood. We will try to recognize maximum of the neighborhood that all required venue close by. Below we will see the chart that will show the highest ratio of neighborhood in Toronto Borough.

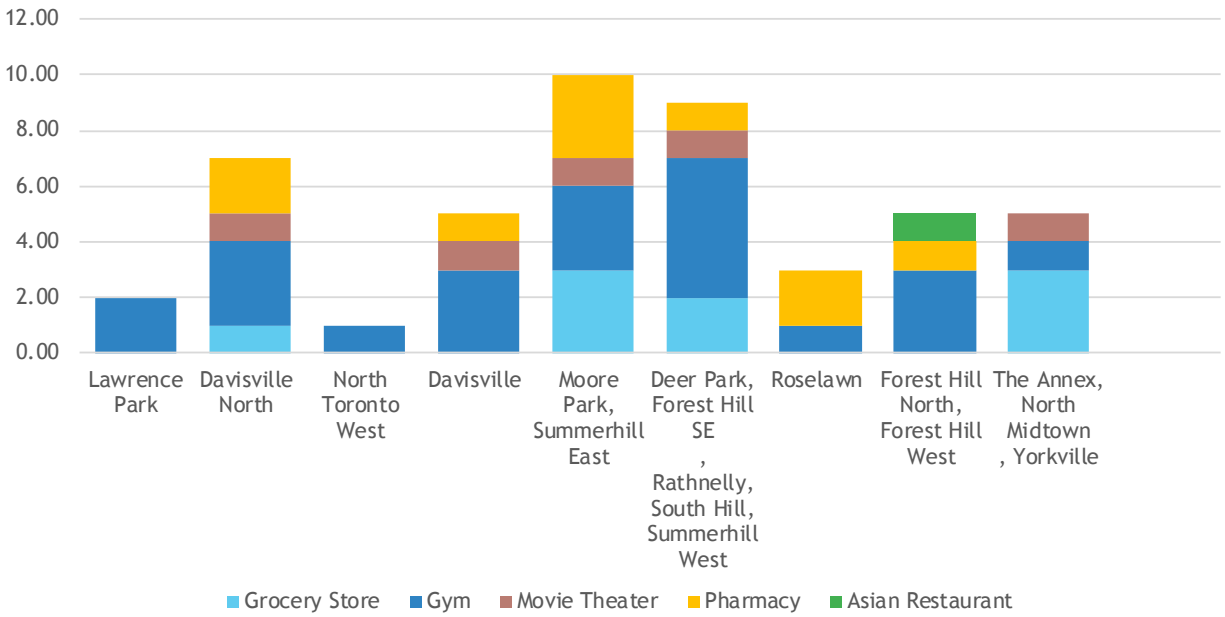


In first step we have collected the required **data: location and type (category) of each Neighborhood in Toronto** . We have also **identified number of venue actually available in Neighborhood** (according to One Hot Encoding technique).

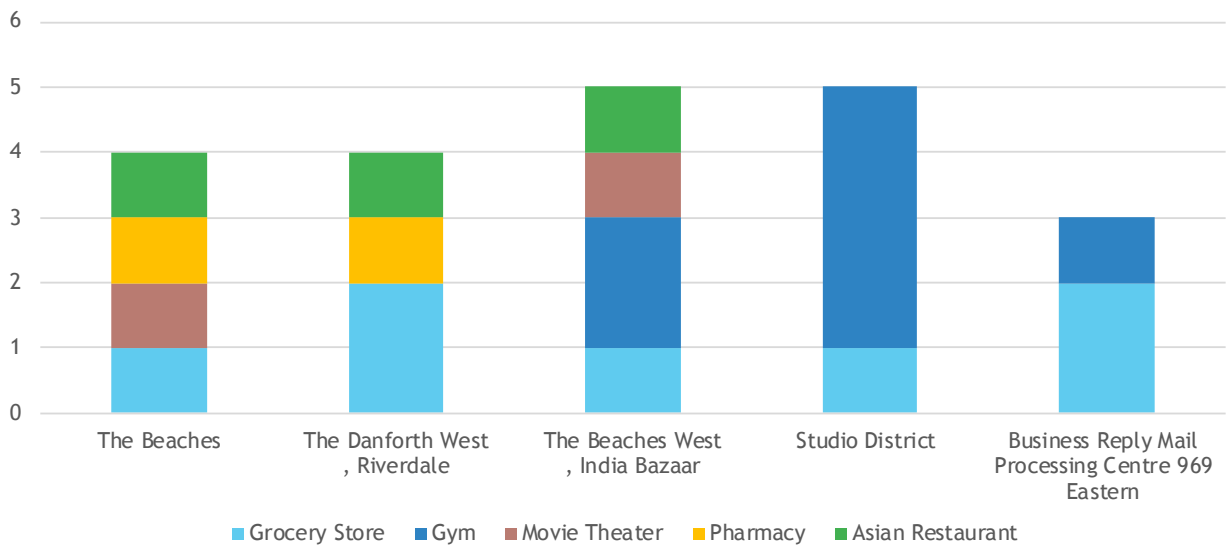
Second step in our analysis will be calculation and exploration of every category in each Neighborhood by finding the weighed average of each category in each Neighborhood.

We have plotted graph for each Borough that shows the proportion of venues available in each Neighborhood.

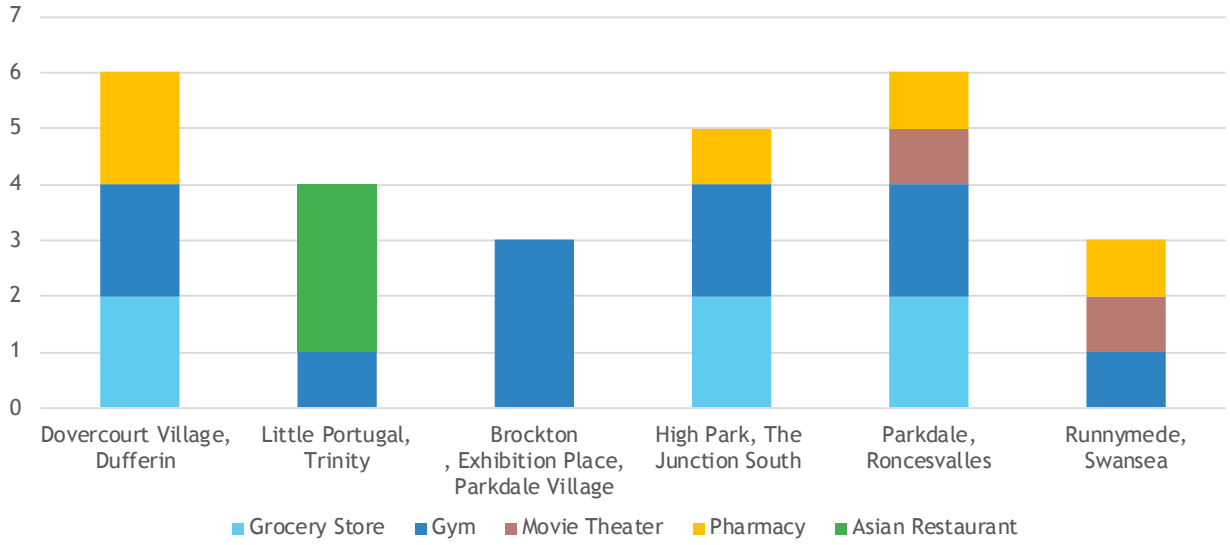
3.1 Central Tororntо



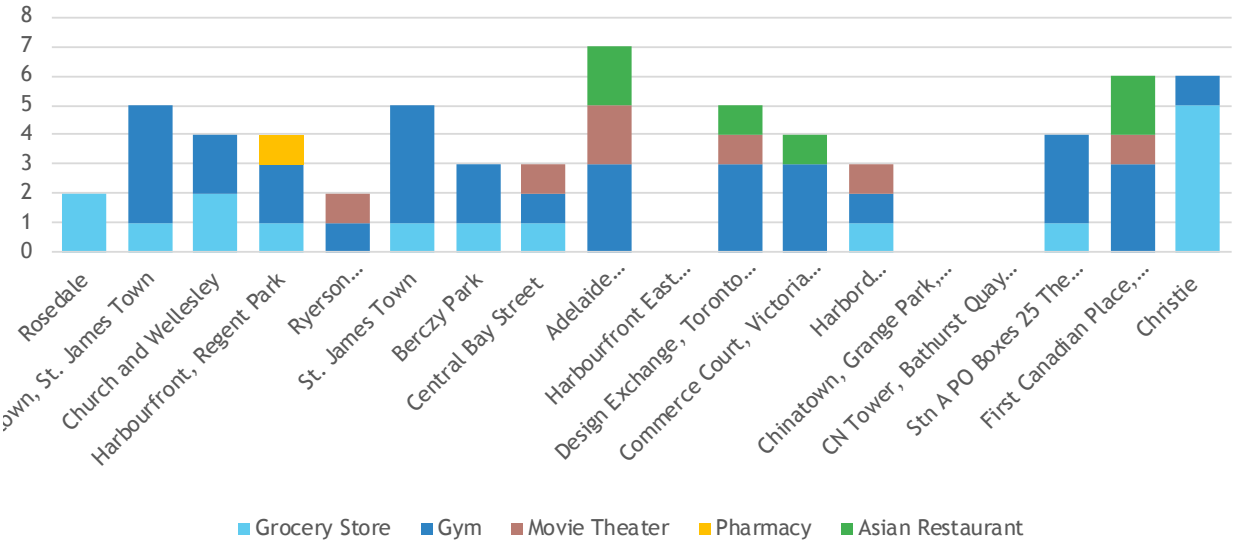
3.2 East Toronto



3.3 West Toronto



3.4 Downtown Toronto



All above graph shows the vertical axis is the number of venues that are available in each Neighborhood where horizontal axis is the list of the Neighborhood. Each neighborhood shows the stacked column with venues available in that area. The venues are listed on bottom line with different color which represent in stacked column which represent the venue available in that neighborhood. These graph will help at last to find optimal location of resident based on analysis on all neighborhood to find the maximum venue available.

In final step we will focus on most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion: we will take into consideration locations with **most common venue in each Neighborhood**. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location. The reason behind using k-means clustering machine learning algorithm is the clustering algorithm provide us with insight into the dataset and lead us to group the data into number of clusters.

4. Results and Discussion

Our analysis shows that although there is a great number of Borough and Neighborhood near Toronto but we focused Neighborhood and Borough in Toronto which offer a combination of popularity among tourists, closeness to city center and strong socio-economic dynamics. The reason behind choosing the location is having the happening in life as location around is happening.

Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is 38 zones containing largest number of potential new locations based on number of and distance to existing venues. This, of course, does not imply that those zones are actually optimal locations for a new resident! Purpose of this analysis was to only provide info on areas in Toronto center where we can get enough information on the availability of daily necessities things. Those criteria would make life more easier and happier. Having those location for resident would be the good achievement. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

5. Conclusion

Purpose of this project was to identify the Neighborhood in Toronto where people can find all necessities things around the residence. By finding the different categories venues in each neighborhood from Foursquare data we have first identified all categories venues that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration.

Final decision on optimal resident location will be made by people/client based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.