## 1.0 Introduction

In this project, the main goal is to analyze the different shops and services that are available in Toronto, Canada, to get an insight of the type of services that each area requires. This will be done using the least common type of shop or service in each region of Toronto using data from Foursquare, Geolocator and Wikipedia.

### 1.1 Background

Most people, especially in urban area like cities depend on different types of shops to buy necessary products. The availability of different types of shops that sell these products is therefore very important for the resident of a location. Some shops/services might be more readily available than others. Finding the least common type of shop can help determine the demand in the location. The demand for the type of shop/service in a locale can be found by analyzing the location data from Foursquare to find the least common types. Using this information, the most profitable shop in each area of Toronto can be determined and a contractor would be able use this information to start a profitable business.

### 1.2 Stakeholders

Some of the stakeholders for this project are discussed below.

- A contractor/Business – They would be interested in finding the most profitable business in a location.
- Residents of the location- They would be interested in improving the accessibility of the particular type of store in their neighborhood.

## 2.0 Data Description

The data required to solve the project is obtained from different source. Before performing any analysis on these data, we must ensure that the data is cleaned and in the suitable form for any analysis technique.

### 2.1 Data Sources

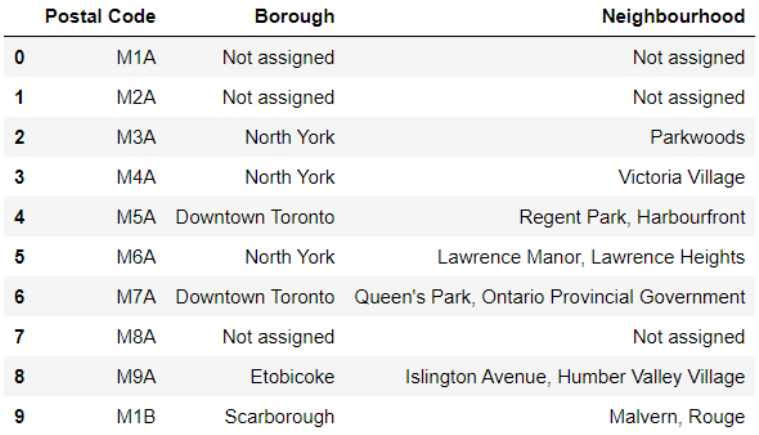
There are four different type of data required for this project. First, it is necessary to get information on the different regions of Toronto. The second type of data would be the type of shops and services in each region. These data types are discussed below.

1. First, the different boroughs in Toronto must be recognized based on the postal code. This information will be obtained from the Wikipedia page ‘List of postal codes of Canada: M’ (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). The Wikipedia page will be read into a pandas data frame and will contain the Postal Code, Borough, and the Neighborhood information.
2. It is also necessary to provide the latitude and longitude location for each postal code. Hence, location data is obtained from the csv file: ‘http://cocl.us/Geospatial_data’.

3. To find the type of store in each area, the boundaries of the neighborhoods are required. This data is obtained from the City of Toronto’s Open Data Catalogue (https://open.toronto.ca/dataset/neighbourhoods/) as a geojson file.
4. Foursquare is used to get location data of the different shops and services present. The data from foursquare will include the name of the shop/service, the type of shop and the latitude and longitude of the shop. This will allow for comparison with the regions of Toronto data and hence divide the shops based on each region.

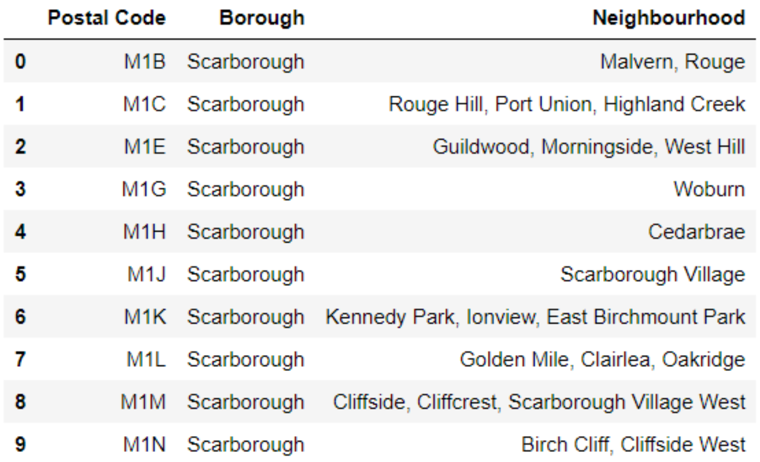
### 2.2 Data Cleaning

In the postal code data from Wikipedia, it was found that some postal codes did not have a corresponding borough. These boroughs and neighborhoods were given a value of ‘Not Applicable’ in the data. Additionally, some postal codes are repeated. This is shown in Figure 1. Hence, the data cleaning process for this data set will need to remove the ‘Not Applicable’ values and merge any postal code that is repeated. After completing this process (Figure 2), we get the data frame of postal codes and neighborhoods in Toronto.



| | 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 |
| 5 | M6A | North York | Lawrence Manor, Lawrence Heights |
| 6 | M7A | Downtown Toronto | Queen's Park, Ontario Provincial Government |
| 7 | M8A | Not assigned | Not assigned |
| 8 | M9A | Etobicoke | Islington Avenue, Humber Valley Village |
| 9 | M1B | Scarborough | Malvern, Rouge |

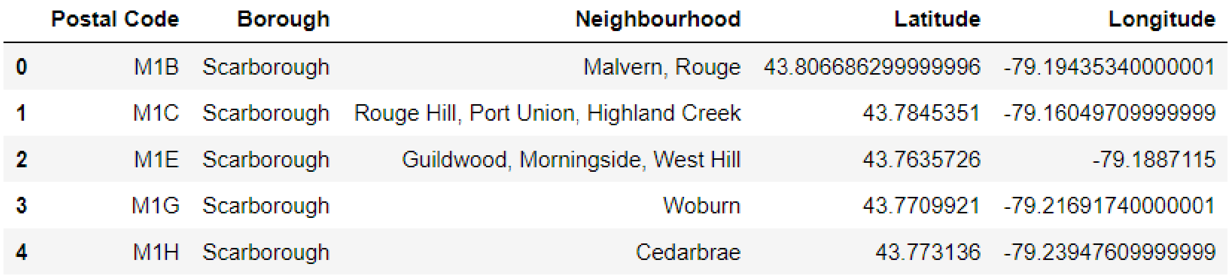
Figure 1: Figure showing the data loaded from Wikipedia.



| | Postal Code | Borough | Neighbourhood |
|---|---|---|---|
| 0 | M1B | Scarborough | Malvern, Rouge |
| 1 | M1C | Scarborough | Rouge Hill, Port Union, Highland Creek |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill |
| 3 | M1G | Scarborough | Woburn |
| 4 | M1H | Scarborough | Cedarbrae |
| 5 | M1J | Scarborough | Scarborough Village |
| 6 | M1K | Scarborough | Kennedy Park, Ionview, East Birchmount Park |
| 7 | M1L | Scarborough | Golden Mile, Clairlea, Oakridge |
| 8 | M1M | Scarborough | Cliffside, Cliffcrest, Scarborough Village West |
| 9 | M1N | Scarborough | Birch Cliff, Cliffside West |

Figure 2: Figure showing the cleaned data.

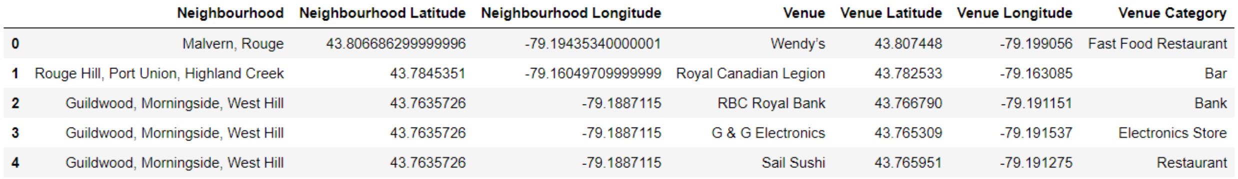
Next, we need to load the latitude and longitude data from the csv file onto the data frame. After this is done, the data frame looks like Figure 3.



| | Postal Code | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Malvern, Rouge | 43.806686299999996 | -79.19435340000001 |
| 1 | M1C | Scarborough | Rouge Hill, Port Union, Highland Creek | 43.7845351 | -79.16049709999999 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.7635726 | -79.1887115 |
| 3 | M1G | Scarborough | Woburn | 43.7709921 | -79.21691740000001 |
| 4 | M1H | Scarborough | Cedarbrae | 43.773136 | -79.23947609999999 |

Figure 3: Data frame with the latitude and longitude data

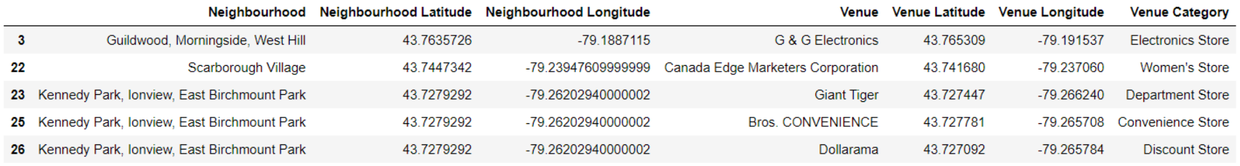
At this point, we need to get nearby stores data from Foursquare. The store name, category and its location information are added to a data frame with the neighborhood information for each entry. This data frame is shown in Figure 4.



| | Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Malvern, Rouge | 43.806686299999996 | -79.19435340000001 | Wendy's | 43.807448 | -79.199056 | Fast Food Restaurant |
| 1 | Rouge Hill, Port Union, Highland Creek | 43.7845351 | -79.16049709999999 | Royal Canadian Legion | 43.782533 | -79.163085 | Bar |
| 2 | Guildwood, Morningside, West Hill | 43.7635726 | -79.1887115 | RBC Royal Bank | 43.766790 | -79.191151 | Bank |
| 3 | Guildwood, Morningside, West Hill | 43.7635726 | -79.1887115 | G & G Electronics | 43.765309 | -79.191537 | Electronics Store |
| 4 | Guildwood, Morningside, West Hill | 43.7635726 | -79.1887115 | Sail Sushi | 43.765951 | -79.191275 | Restaurant |

Figure 4: Data frame with the Foursquare data

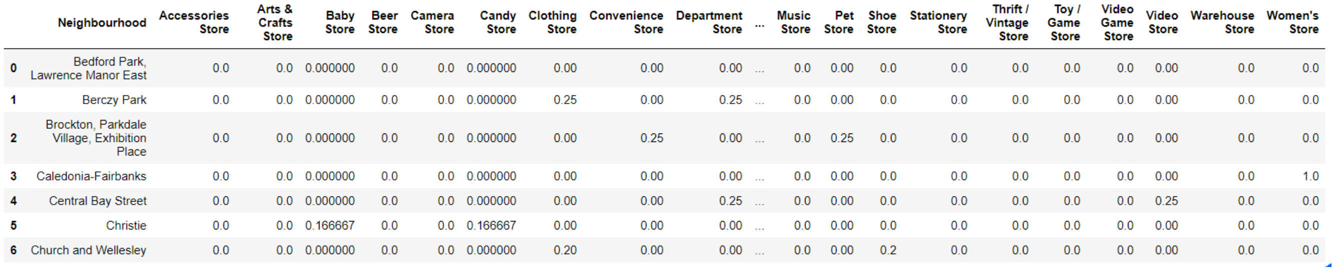
This data frame contains all venue category. For the purposes of this project, we only need the nearby stores for each neighborhood. Therefore, we remove all entries other than those containing the word ‘Store’ in the Venue Category. This data frame is shown in Figure 5.



| | Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 3 | Guildwood, Morningside, West Hill | 43.7635726 | -79.1887115 | G & G Electronics | 43.765309 | -79.191537 | Electronics Store |
| 22 | Scarborough Village | 43.7447342 | -79.23947609999999 | Canada Edge Marketers Corporation | 43.741680 | -79.237060 | Women's Store |
| 23 | Kennedy Park, Ionview, East Birchmount Park | 43.7279292 | -79.26202940000002 | Giant Tiger | 43.727447 | -79.266240 | Department Store |
| 25 | Kennedy Park, Ionview, East Birchmount Park | 43.7279292 | -79.26202940000002 | Bros. CONVENIENCE | 43.727781 | -79.265708 | Convenience Store |
| 26 | Kennedy Park, Ionview, East Birchmount Park | 43.7279292 | -79.26202940000002 | Dollarama | 43.727092 | -79.265784 | Discount Store |

Figure 5: Data frame with only the store data

To make it easier for analysis, this data frame is grouped based of the neighborhood. So, each neighborhood has 1 row and the mean of the different type of store is shown for each row. This is shown in Figure 6.



|  | Neighbourhood | Accessories Store | Arts & Crafts Store | Baby Store | Beer Store | Camera Store | Candy Store | Clothing Store | Convenience Store | Department Store | ... | Music Store | Pet Store | Shoe Store | Stationery Store | Thrift / Vintage Store | Toy / Game Store | Video Game Store | Video Store | Warehouse Store | Women's Store |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Bedford Park, Lawrence Manor East | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.0 | 0.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 1 | Berczy Park | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.25 | 0.00 | 0.25 | ... | 0.0 | 0.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 2 | Brockton, Parkdale Village, Exhibition Place | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.00 | 0.25 | 0.00 | ... | 0.0 | 0.25 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 3 | Caledonia-Fairbanks | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.0 | 0.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.0 | 1.0 |
| 4 | Central Bay Street | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.25 | ... | 0.0 | 0.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.25 | 0.0 | 0.0 |
| 5 | Christie | 0.0 | 0.0 | 0.166667 | 0.0 | 0.0 | 0.166667 | 0.00 | 0.00 | 0.00 | ... | 0.0 | 0.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 6 | Church and Wellesley | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.20 | 0.00 | 0.00 | ... | 0.0 | 0.00 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |

Figure 6: Data frame grouped by Neighborhood.

This data frame gives us all the necessary information and we can move on the data analysis.

### 3.0 Methodology

The main goal of the analysis is to find neighborhoods where similar type of store can be opened. Hence, k-means analysis technique will be used to group together similar areas. First, the neighborhoods in Toronto are visualized in a folium map as shown in Figure 8.

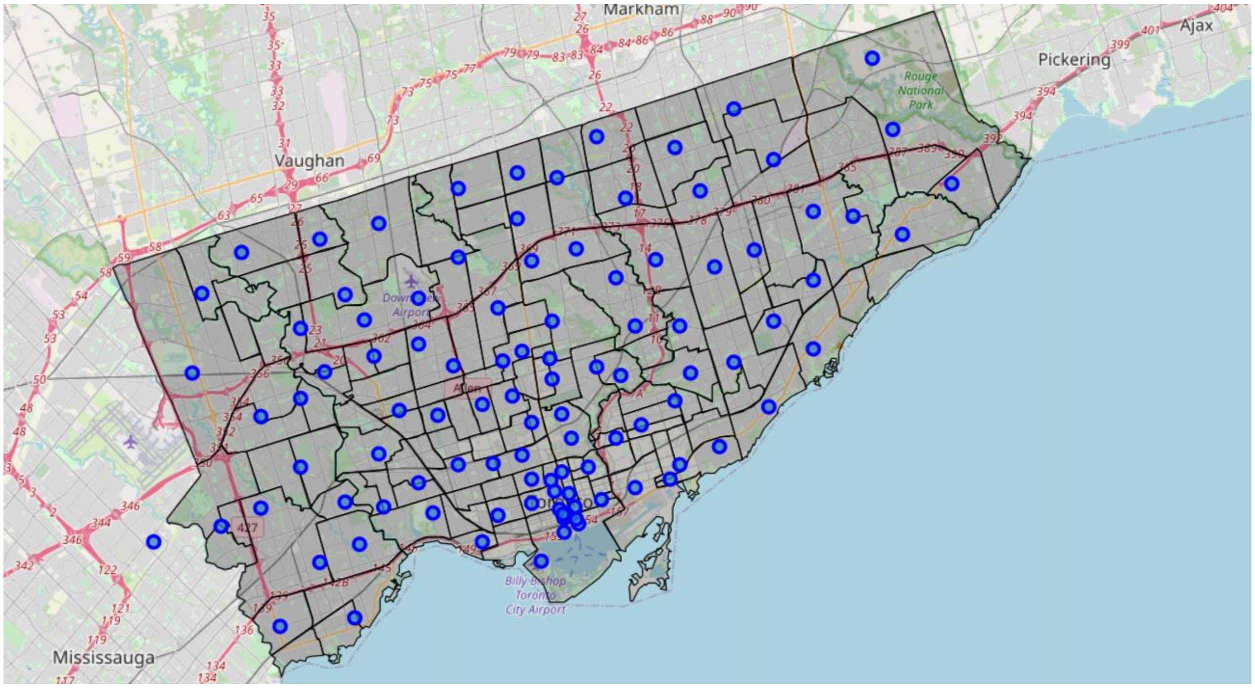


Figure 8: Map of Toronto showing the different neighborhoods.

The K-means clustering is done using a value of 15. This means that the stores are sorted into 15 different categories. The result of this clustering is then displayed is the folium map to get a better understanding of the different neighborhoods. This is shown in Figure 9.

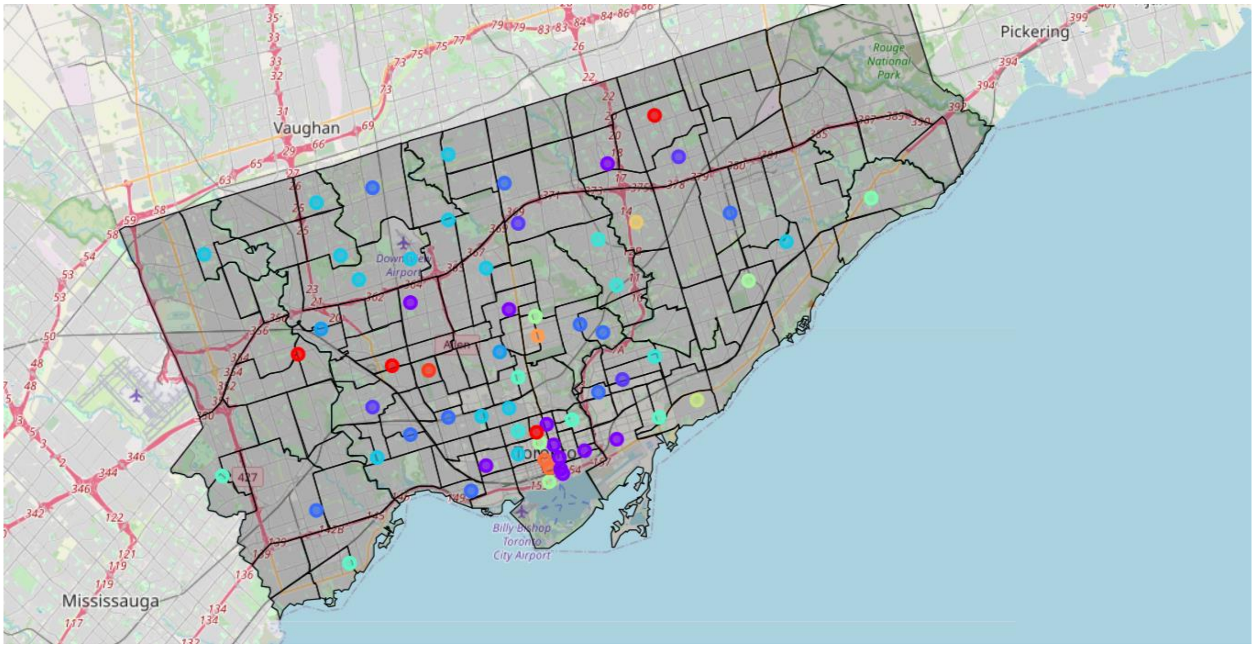
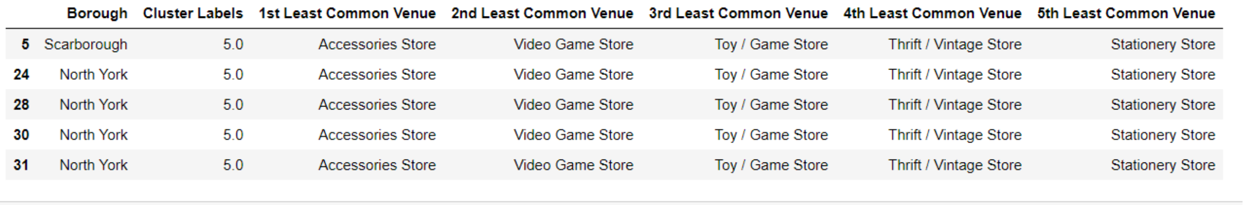


Figure 9: Map of Toronto neighborhoods with the result of clustering displayed.

## 4.0 Results and Discussion

From the clustering, we can find the cluster with the greatest number of stores. This was found to be cluster number 5 and the stores in Cluster 5 are shown in Figure 10.



| | Borough | Cluster Labels | 1st Least Common Venue | 2nd Least Common Venue | 3rd Least Common Venue | 4th Least Common Venue | 5th Least Common Venue |
|---|---|---|---|---|---|---|---|
| 5 | Scarborough | 5.0 | Accessories Store | Video Game Store | Toy / Game Store | Thrift / Vintage Store | Stationery Store |
| 24 | North York | 5.0 | Accessories Store | Video Game Store | Toy / Game Store | Thrift / Vintage Store | Stationery Store |
| 28 | North York | 5.0 | Accessories Store | Video Game Store | Toy / Game Store | Thrift / Vintage Store | Stationery Store |
| 30 | North York | 5.0 | Accessories Store | Video Game Store | Toy / Game Store | Thrift / Vintage Store | Stationery Store |
| 31 | North York | 5.0 | Accessories Store | Video Game Store | Toy / Game Store | Thrift / Vintage Store | Stationery Store |

Figure 10: Data frame of the stores in Cluster 5

This cluster contains accessories store, videogame stores, toy stores, thrift stores and stationery stores. From this we can infer that when starting a new store chain in Toronto, it would be most profitable to open stores in cluster 5 such as an accessory store.  With this information, a person starting a new store will be able to decide that starting an accessory store in Toronto is desirable as it has the most potential to become a store chain and hence generate more profit than stores from cluster.

Additionally, any business that specializes in one type of store like grocery store, stationery stores etc. will benefit from this analysis. They will be able to decide which neighborhood their stores would be the most profitable in by comparing Figures 8&9. For example, a company that deals with video game stores will be most profitable in neighborhoods like Scarborough and North York because the number of these stores in these areas are low and hence, they will have a higher demand. Hence using this analysis, we can recommend to the company that these neighborhoods are the target location for opening new stores.

## 5.0 Conclusion

The main goal of the project was analyzing the different types of shops and services available in Toronto to recommend the most profitable type. This was done using Foursquare venue data. Other types of data for the project include location data for Toronto and the different neighborhoods in Toronto. The data sets were processed to make it useful for the analysis. The store data was stored into different categories using K means machine learning technique. The result of this analysis was displayed on a map of Toronto with the different neighborhoods outlined. This will help any future business determine what type of stores would be the most profitable in Toronto. This analysis can also help determine where any existing business should open a new location.

Going forward, more analysis on the number of clusters is required. Increasing the number of clusters would increase the accuracy of the result. But this might also require more effort without much advantage. Also, to get a better result additional factors such as the finances of the store can be considered.

.