Capstone Project - Battle of the Neighbourhoods Report

IBM - Coursera Professional Data Science Certificate Course

Jothika Sundaram
June 02, 2020

Table of Contents

A. Introduction and Business Problem

The city of Toronto is one of the major metropoles in Canada. With a population of over 2.93 million, it is the most populous city in Canada known for its iconic skyscrapers, bustling city life and dynamic ethnic diversity. For these reasons, Toronto is also an international centre for business and finance, and is a major economic hub in Canada.

These factors also encourage entrepreneurs, small business owners and startup companies to open their business in Toronto. This project aims to act as a "startup company guide" to new entrepreneurs. It will provide an analysis on the various business and local venues located across the city, along with local population demographics such as ethnicity and age groups. This information will then give us an idea about what kinds of business should open up in which area of the city, along with the demographics of the local population that will be targeted.

B.1 Data

This analysis requires the following sets of data:

- In order to determine the local population demographics, I used the city's neighbourhood profiles. This dataset can be obtained from the City of Toronto Open Data Portal [1]. This dataset includes population distribution and demographics such as age and ethnicity groups. This is used to discover the characteristics of the audiences surrounding the types of venues in each neighbourhood.
- The location data of the city is obtained using Foursquare API [2]. With this data, I was able to analyse the geographical features, such as popular venues and companies in each borough. This gave an idea of how businesses are spread out in the city which will aid in finding different locations for potential entrepreneurs to open a venue.
- In order to retrieve the information I wanted from the Foursquare API, I needed to provide specific locations that were to be explored. I used a dataset of the different postal code areas in Toronto, along with their respective boroughs and neighbourhoods. This was obtained by scraping this Wikipedia page [3] of postal codes in Toronto.
- To find geographic coordinates of each borough and neighbourhood to feed into the Foursquare API, I used python's geocoding API service provider GeoPy. [4]

B.2 Methodology

The methodology of this project is composed of the following steps:

1. Examine the neighbourhood profiles of Toronto to get an idea of the demographics in each area. These findings are then visualized using choropleth maps to see the distribution of these different groups.
2. Collect and clean the data required to feed our Foursquare API to get the location data of the different venues across the city. These locations are then visualized on a map.
3. Examine the most common venues by category located in each borough, which are then visualized using bar charts.
4. Using the unsupervised, k-means clustering machine learning algorithm, cluster the neighbourhoods in order to partition these areas based on the most common venue type. These clusters are then visualized on a map.
5. Finally, analyse the distribution of venues across the city along with their local population demographics - this will provide potential locations and target audiences for new entrepreneurs based on their business needs.

C. Data Collection and Cleaning

The city's census data I collected was stored into a table. This data was separated by neighbourhood and included population distribution, age group distribution, ethnic origin, household characteristics and so on. First, I focused on the ethnic origin category. I excluded ethnicities that are traditionally found in North American ancestry and focused on foreign ethnic origins and the distribution of these residents across different neighbourhoods. I then examined the different age groups among the population, specifically children aged 0-14, youth aged 15-24, the working population aged 25-54 and seniors aged 65+. The density and distribution of these populations were visualized using python's map rendering library Folium. In the following choropleth maps, the areas in darker colours represent higher populations and black regions indicate no data was available:

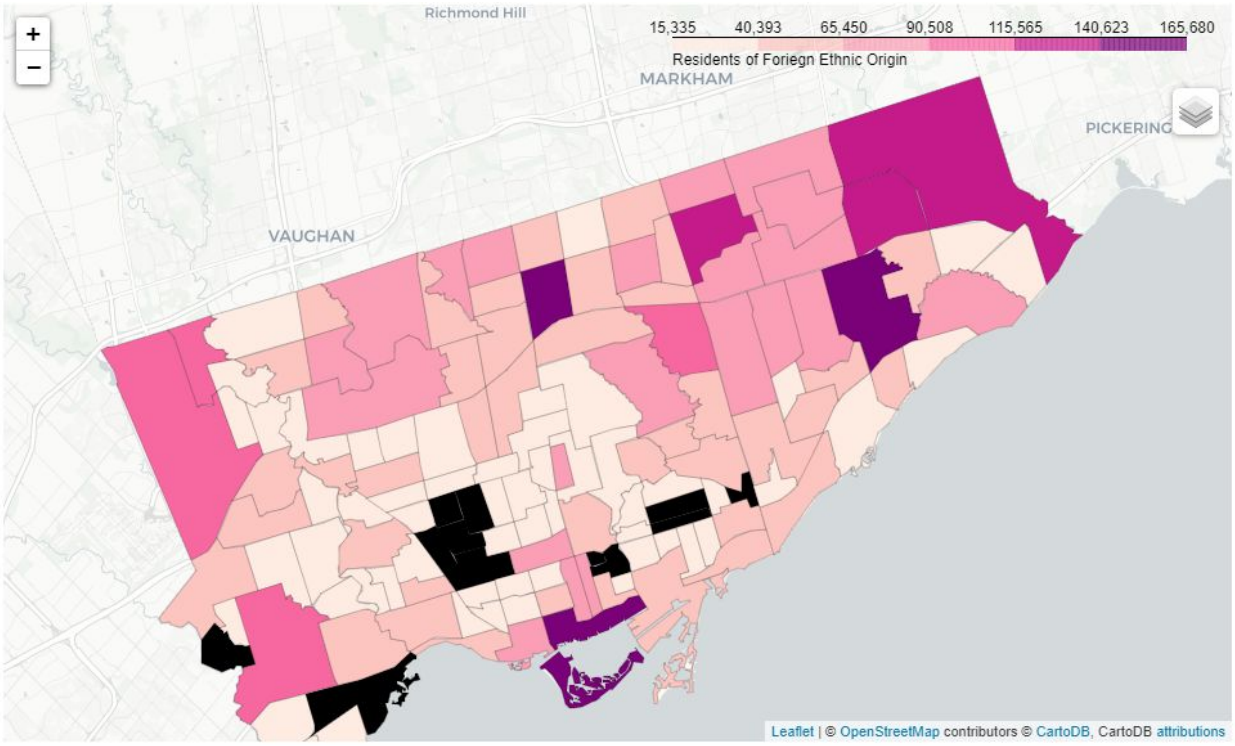

Figure 1. Distribution of residents that are of foreign ethnic origin.

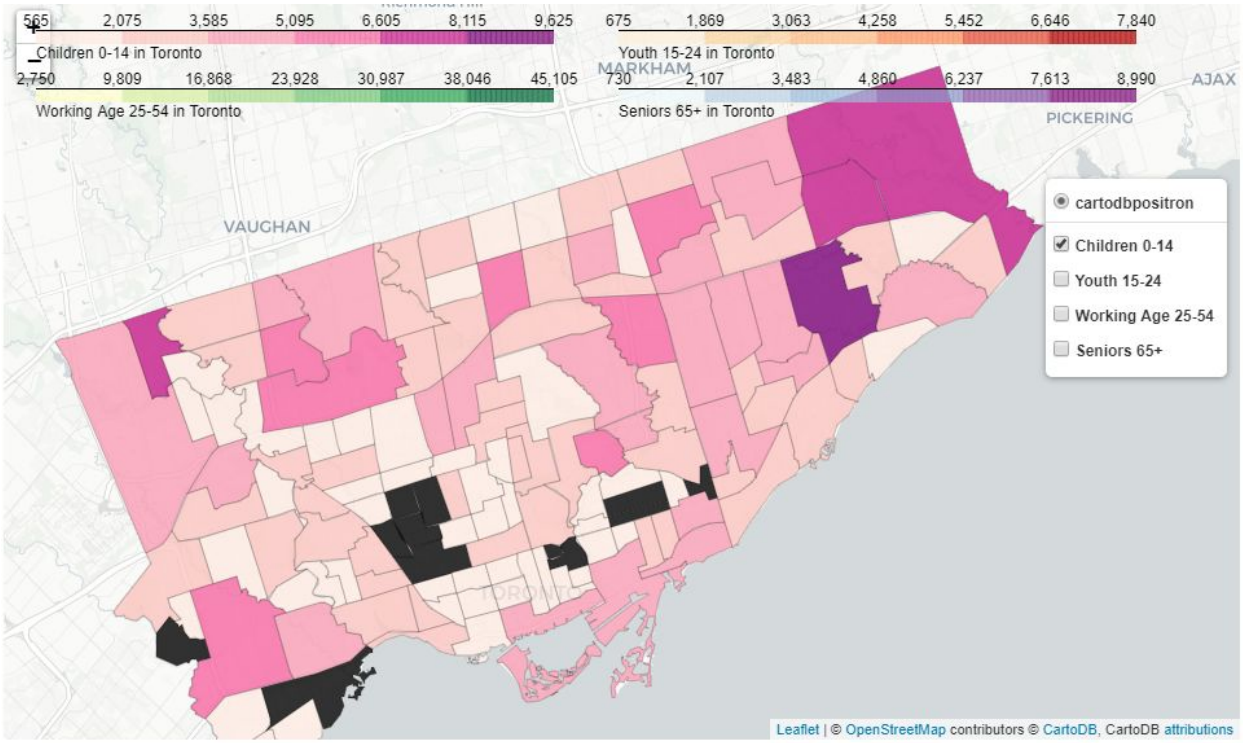

Figure 2. Distribution of children populations.

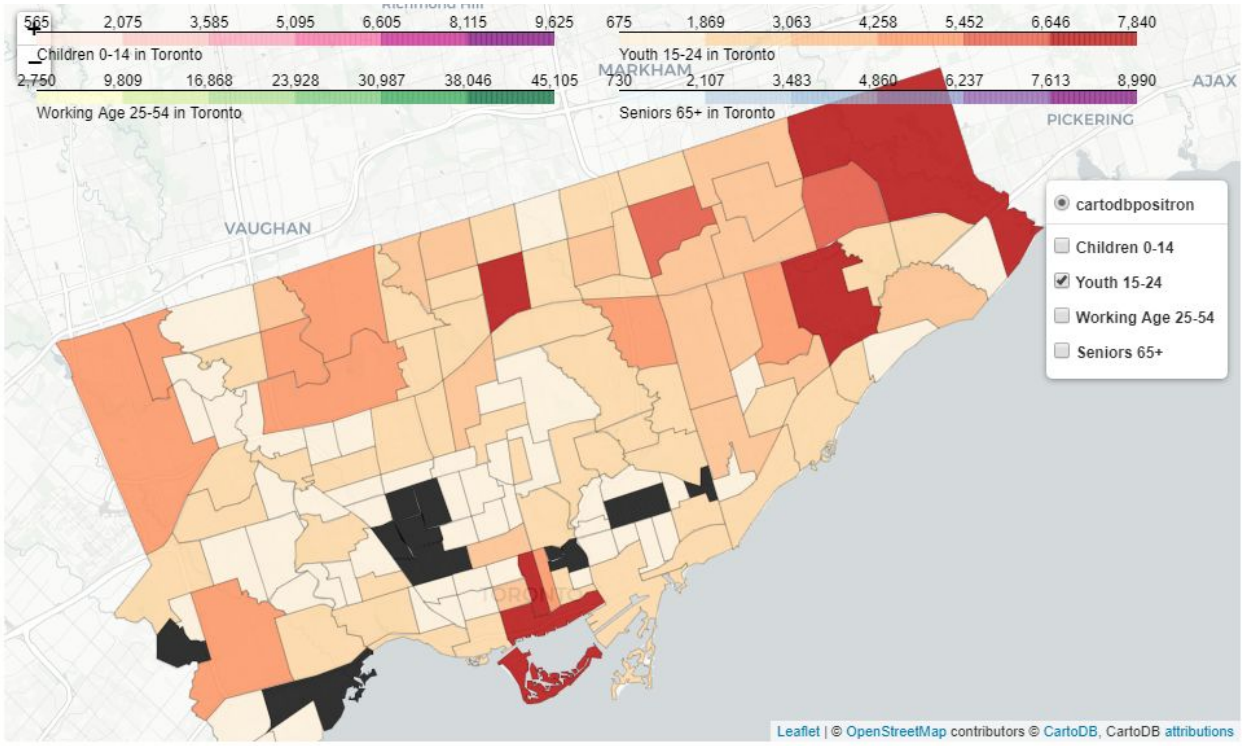

Figure 3. Distribution of youth populations.

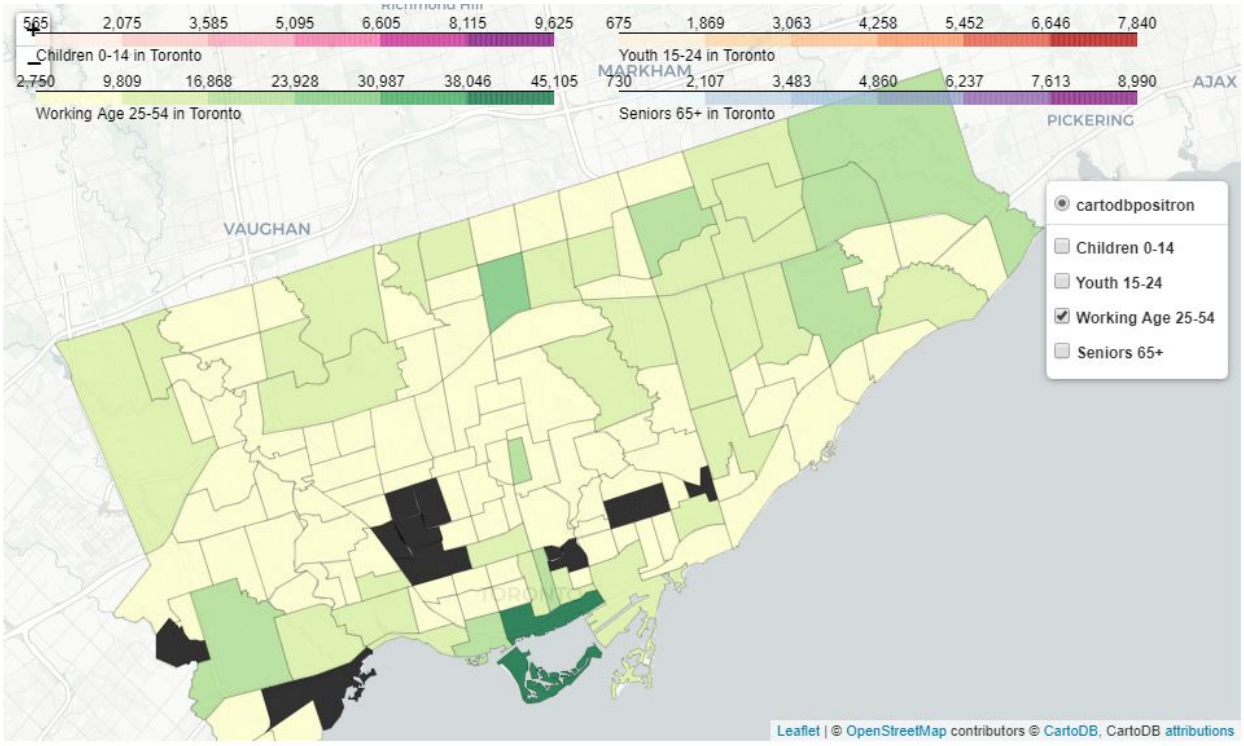

Figure 4. Distribution of adult populations.

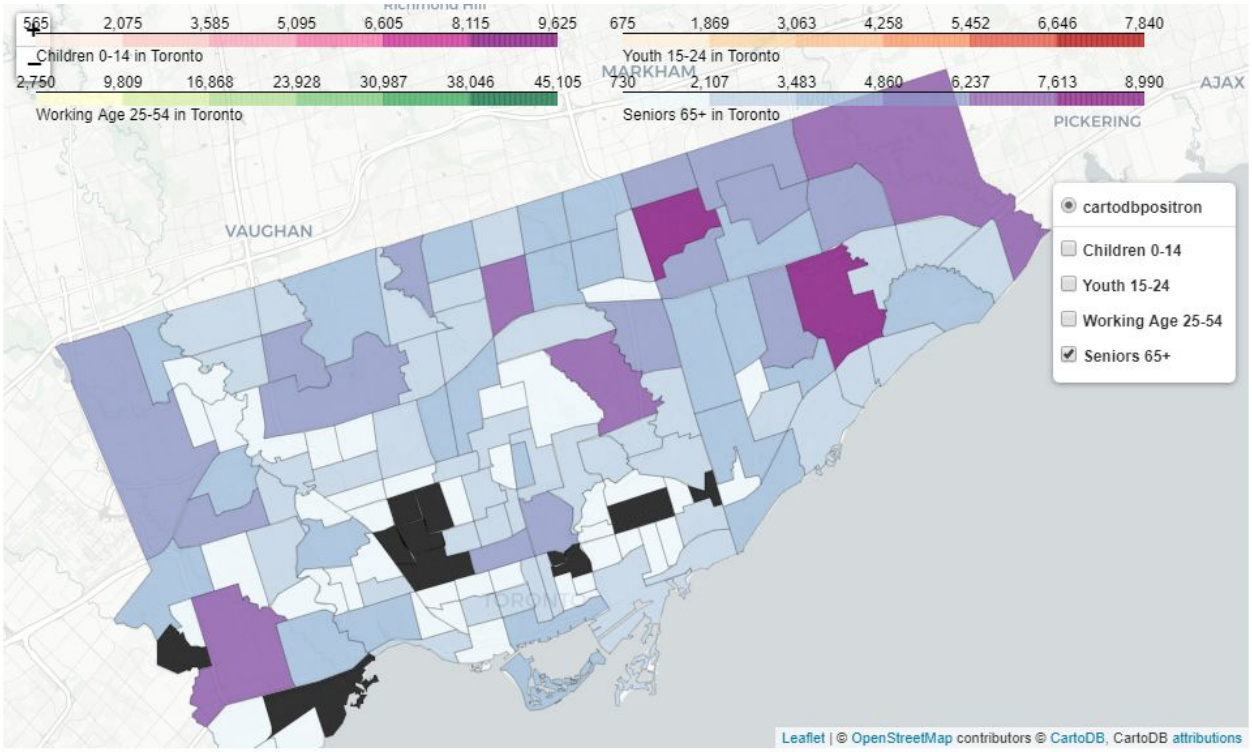

Figure 5. Distribution of senior populations.

Next, I collected and cleaned the data required to use with the Foursquare API. I started with a table of postal codes in Toronto and their respective boroughs and neighbourhoods. Using python's geocoding API, I was able to retrieve the geographic coordinates for each borough and neighbourhood. I then fed the location of each borough into the API to retrieve data of a maximum of 500 venues within a 10 km radius from the borough location. This data was turned into a dataframe for each borough, which were then merged together into one big table containing the venue name, category, venue coordinates, postal code and borough/neighbourhood locations.

This large dataset posed a few problems that I had to fix. Namely, many venues were not assigned to a borough or neighbourhood. If each borough or neighbourhood were examined individually, these venues would not appear in their data. Additionally, if these venues were dropped from the dataset, it may skew the final result of the analysis inaccurately.

To solve this issue, I first separated the dataset into two tables: one containing venues with valid boroughs and neighbourhood an the other containing the invalid data. I then compared the coordinates of the invalid venues to the valid venues to a precision of 2 decimal places in order to match similar locations to a neighbourhood and borough. This worked for most of the invalid data, but I was still left with many venues that were not matched to a location. Instead of dropping these venues, I used reverse geocoding using their venue coordinates and was able to retrieve their borough locations. Unfortunately, this method did not retrieve their valid neighbourhood location, but it was sufficient to fill the missing borough locations in the table. After cleaning this dataset, all venues were marked onto a map as seen below:

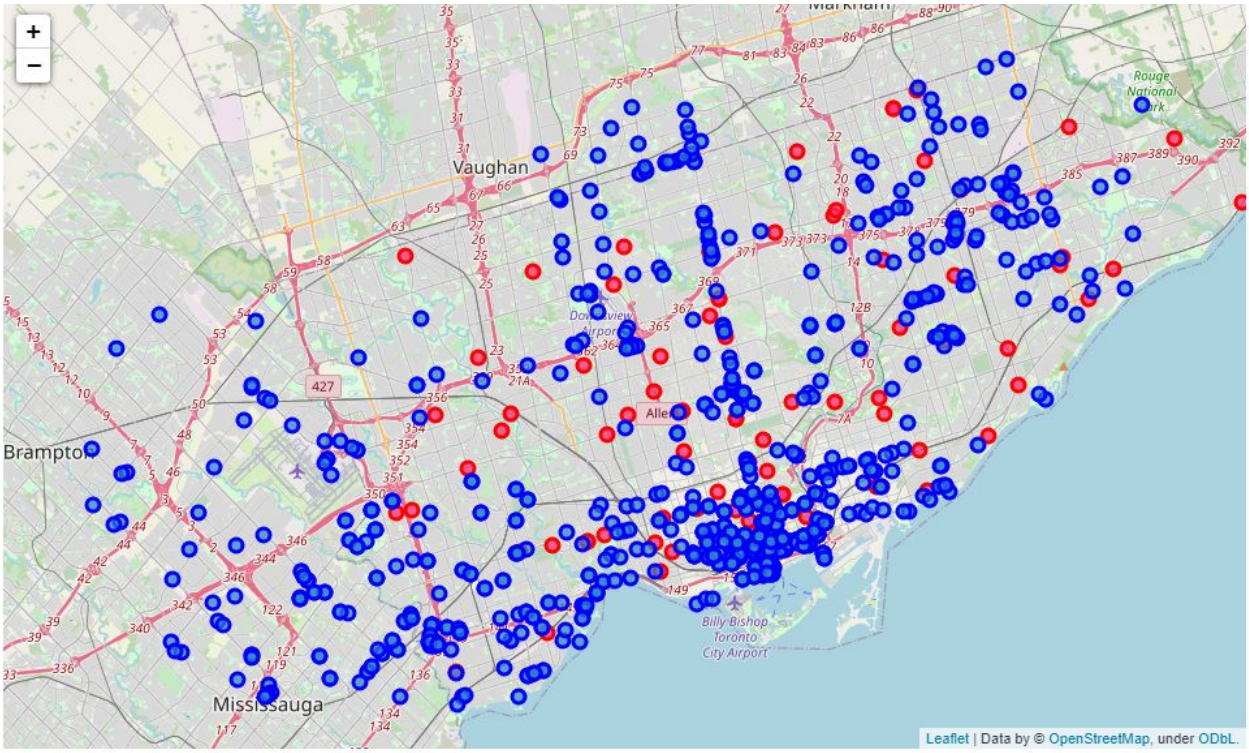

Figure 6. Blue circles indicating venue locations, red circles indicating neighbourhoods that share postal code areas.

D. Analysis

800 venues were returned. I examined the unique category types for these venues and saw that restaurant types were the most common. These restaurants were subcategorized into their respective ethnic category. A few restaurants did not fall under a specific ethnic category, most of them being Steakhouse type venues. These were categorized under "Other". I organized these venues into a table and used a bar chart to visualize the frequency of each restaurant types:

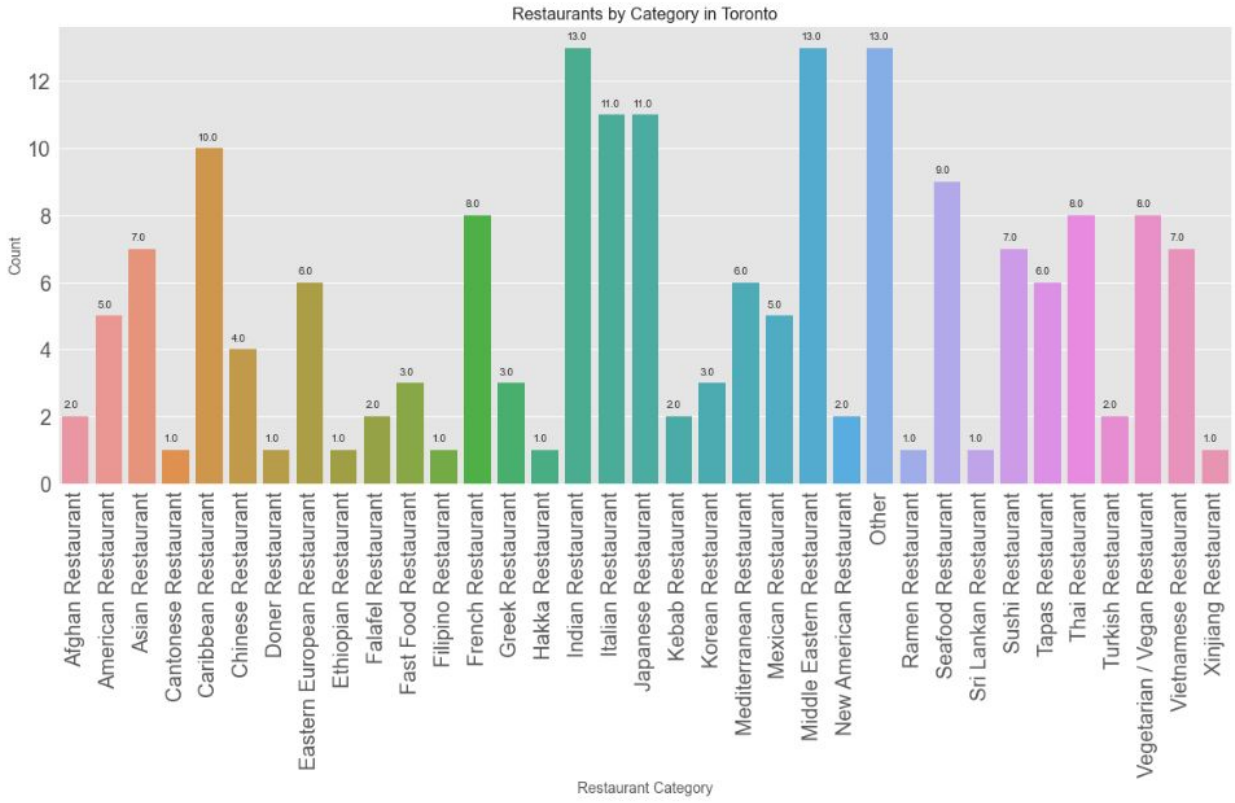

Figure 1. Bar chart showing Restaurants by Category

The results show that, from the data that was retrieved, Indian restaurants, Middle Eastern restaurants and those that fall under the Other category are the most common type of restaurant in Toronto.

I then compared the other venue categories by frequency:

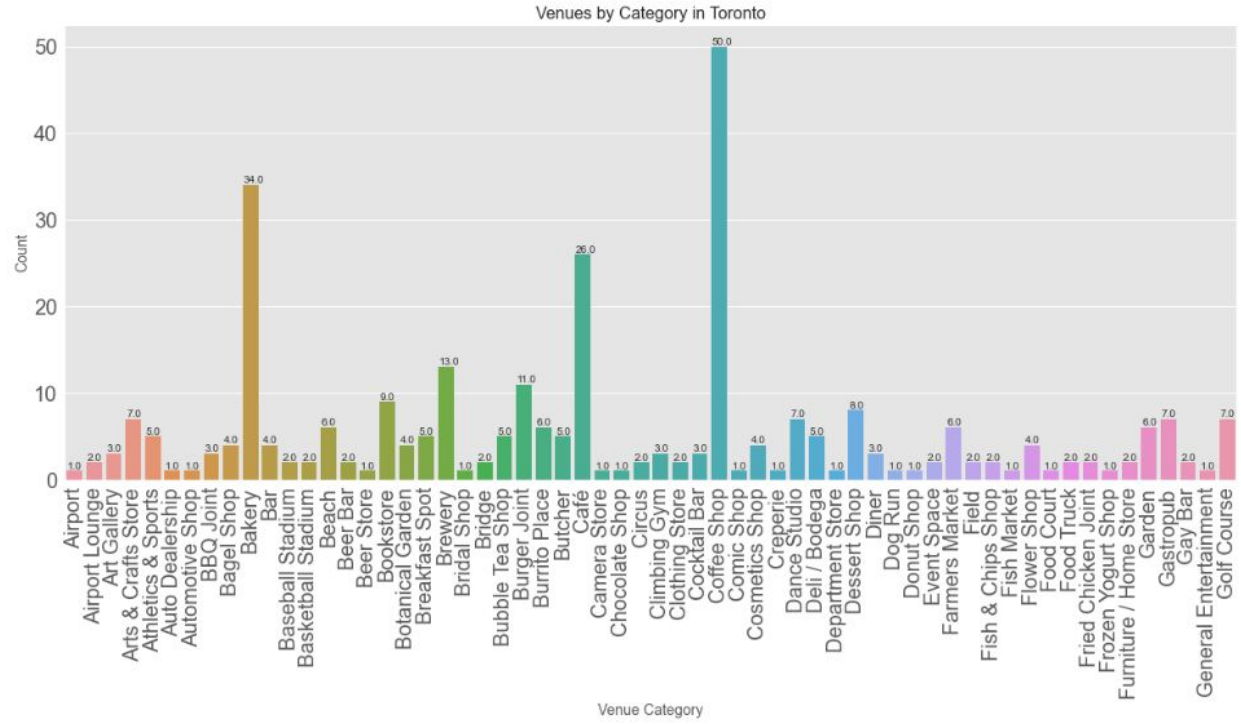

Figure 2.1. Bar chart showing Venues by Category

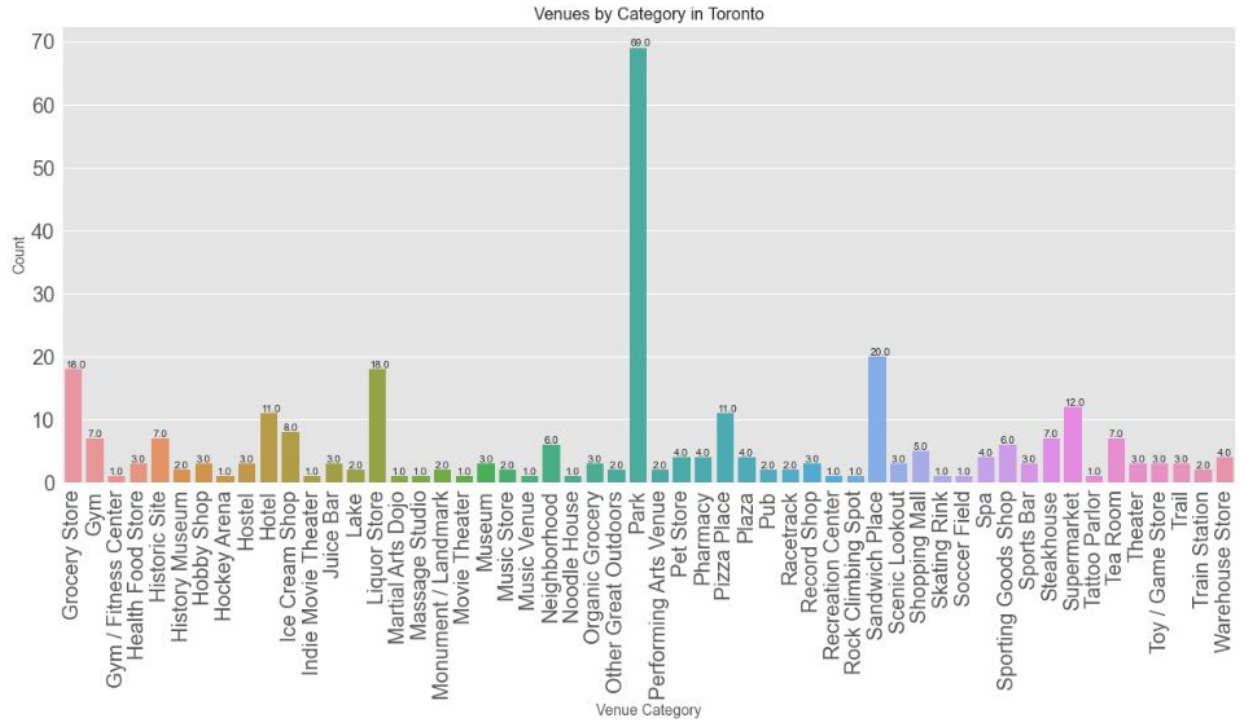

Figure 2.2. Bar chart showing Venues by Category

The results show that after restaurants, park and coffee shop venues are the next most common type of venue.

Next, I prepared the data to cluster the neighbourhoods using k-means clustering. For this section of the analysis, I had to exclude all the venues that were not assigned to a neighbourhood. I transformed the dataset to produce the mean frequency of each venue category occurring in each neighbourhood. I was able to sort the venues for each neighbourhood by most common occurrence:

	Neighbourhood	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	Agincourt	Caribbean Restaurant	Cantonese Restaurant	Gym / Fitness Center	Coffee Shop	Asian Restaurant	Breakfast Spot	Sporting Goods Shop	Bakery	Sri Lankan Restaurant	Yoga Studio
1	Alderwood, Long Branch	Pizza Place	Burger Joint	Park	Café	Seafood Restaurant	Grocery Store	BBQ Joint	Art Gallery	Creperie	Dance Studio
2	Bathurst Manor, Wilson Heights, Downsview North	Park	Bridal Shop	Supermarket	Mediterranean Restaurant	Yoga Studio	Eastern European Restaurant	Falafel Restaurant	Event Space	Ethiopian Restaurant	Diner
3	Bayview Village	Liquor Store	Yoga Studio	Filipino Restaurant	Cosmetics Shop	Creperie	Dance Studio	Department Store	Dessert Shop	Diner	Doner Restaurant
4	Bedford Park, Lawrence Manor East	Bakery	Gourmet Shop	Café	Doner Restaurant	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Eastern European Restaurant	Yoga Studio

Figure 3. Venues by most common occurrence by neighbourhood.

Using the k-means algorithm, I formed 5 clusters of neighbourhoods (labelled 0-4) and each neighbourhood was assigned to a cluster. These clusters are shown on this map:

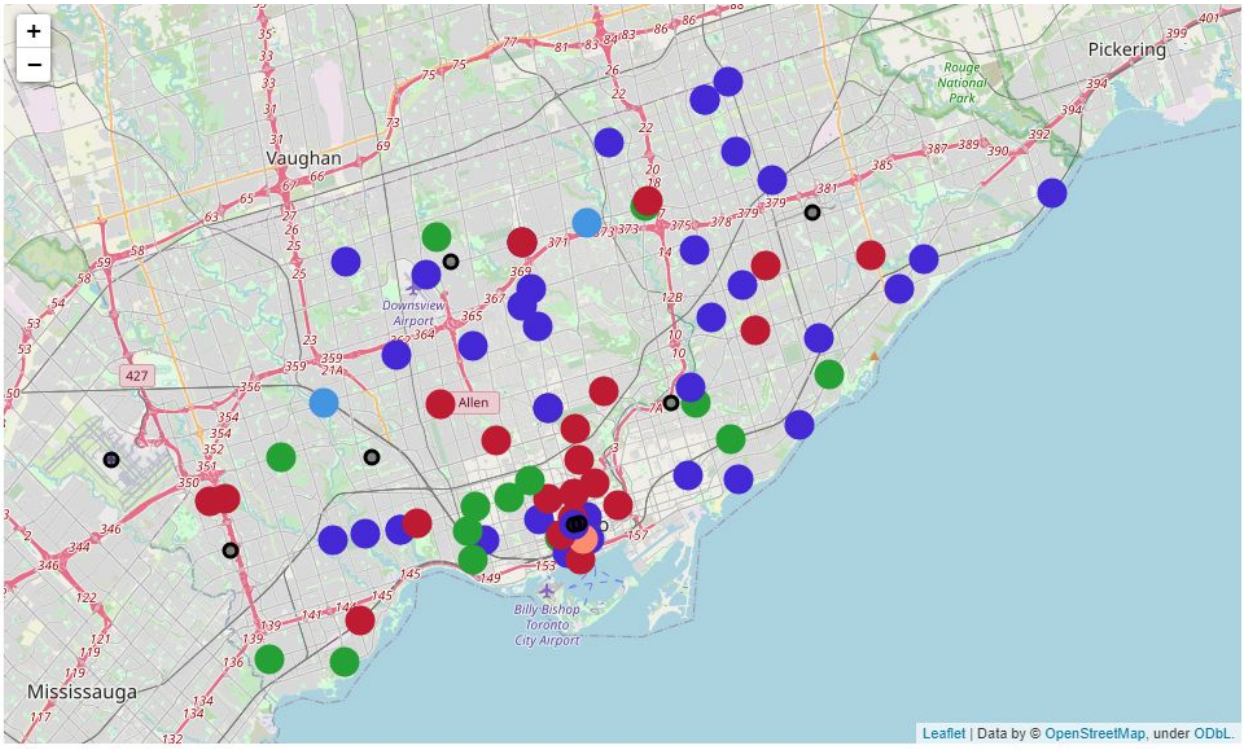

Figure 4. Clusters marked on a map
Red - Cluster 0, Purple - Cluster 1, Orange - Cluster 2, Green - Cluster 3, Blue - Cluster 4

E. Results and Discussion

The defining feature for each cluster is determined by the algorithm, and it was up to me to examine each cluster to find that feature. There were many neighbourhoods assigned to Cluster 0 (Red). In these neighbourhoods, the first two most common venues are dominated by Coffee Shops and Cafés. It can be assumed that this cluster is defined by Coffee Shops and Cafés. However, there were also many ethnic restaurants as common venue types. Looking at the map, it seems that many cafés and ethnic restaurant chains choose to open in the downtown area. As shown on the choropleth maps before (Section C, figure 1), many residents of foreign ethnic

origin resided in the Downtown Toronto area. It makes sense that this area is densely populated with ethnic restaurant chains.

Cluster 1 (Purple) had very diverse common venues, so it was difficult to make out the defining feature. These neighbourhoods are spread out across the city. Again, there were many restaurant and eatery-type venues, and Ethiopian restaurants and Eastern European restaurants were fairly common.

Neighbourhoods in Cluster 3 (Green) have an interesting spread. The most common venues in these areas are Park locations, so they must have a lot of green spaces. These neighbourhoods are mostly situated to the southern side of the city closer to the waterfront and much less in the northern areas near York and North York.

Clusters 2 (Orange) and 4 (Blue) have very few neighbourhoods assigned to them, so there is not enough information to make a thorough analysis from them. Additionally, there does not seem to be any strong relationship between these clusters and the distribution of age groups.

F. Conclusion

This study aimed to analyse the spread of different venues across the city of Toronto in order to gather an idea of what type of business tend to open up in each area of the city. This information will be useful for stakeholders such as business owners and startup companies to decide where they should open a branch that would be ideal for their business.

For example, a restaurant company can see that most of the restaurant chains are centered around Downtown Toronto. These restaurants are all of different ethnic cuisines and are in a highly competitive area, so it is upto that company to decide whether they would like to join the competition, or open up in an area with less competition.

Another example, the neighbourhoods that have lots of parks and green spaces are located. In the locations closer to Downtown Toronto, there is a large population of

youth and young adults (Section C, Figure 3,4). This could be an ideal location to open an outdoor recreational business targeting this age group.

These are only some of the many possibilities that are available to new entrepreneurs and small businesses. The results of this project should be able to guide viewers and stakeholders to find the most profitable locations to open a certain type of business.

G. References

1. City of Toronto Open Data Portal
2. Foursquare API
3. Postal Codes of Toronto
4. GeoPy.