Analysis of Toronto Neighbourhoods using a Machine Learning Algorithm

A Report by Jessica Uwoghiren

1. Introduction

1.1. Background

When I began this project, I came across a news article which read *"Canada to welcome 1.2 Million immigrants by 2023".* [1] This made me excited for the millions of people looking for a pathway to Canada since I recently relocated here. A 2020 US news ranking showed Canada as the 2nd best country in the world, so it is not a surprise that every year, thousands of people choose to migrate to Canada.[2] Several factors such as citizenship, quality of life and entrepreneurship were the most significant criteria for the rankings. Asides from having a strong and stable economy, cultural diversity and many opportunities for growth, Canada has offered many immigrants a new home through its programs such as Federal Skilled Workers/Trades Program (Express Entry) or Provincial Nominees Program. In 2019, Canada opened its borders to 341,000 people across the world with India and China being the countries with the most immigrants. 45% of these immigrants chose to settle in Ontario province with over 35% of them settling in the City of Toronto. [3]

1.2. Problem Statement

Toronto covers a land mass of 630km^2 with a population density of 9,410 per km^2 (This is quite high considering average population density for Canada is 4 people per km^2).[4] The City also has 6 districts with a total 140 neighbourhoods. As a new immigrant, a vital question to answer is "What neighbourhood do I settle in?". The aim of this project is to group Toronto neighborhoods in order of desirability using Machine Learning and Data Visualization techniques.

1.3. Basis

There are several factors consider when settling down in any location. For this project, I performed my analysis using on the following criteria:

- Total number of Essential Venues in each neighbourhood
- Primary and Secondary Benchmarks: Primary benchmarks considered were Unemployment rate, Crime rate and COVID-19 rates. The Secondary benchmark was housing price for a one-bedroom apartment in each neighbourhood.

2. Data

2.1. Data Description

For this project, most of the datasets used were obtained from the City of Toronto Open Data Portal. This portal was launched in fall of 2009 to meet the demand for open data and it contains over 370 datasets which are refreshed periodically. Other datasets were scraped from the web as highlighted below:

a) *Neighbourhood Boundaries Map for City of Toronto (GeoJSON)*: This GeoJSON file contains standard geospatial data and geographic features (i.e. coordinates and boundary shapes) for each neighbourhood in Toronto. This was critical for map visualizations. [5]

b) *COVID-19 Data Report for all Toronto neighbourhoods:* This dataset was imported from the City of Toronto website and it contains total number of cases per neighbourhood. [6]

c) *Crime rates for all Toronto neighbourhoods:* This dataset depicts the total number of crimes (Theft, Assault, Homicide etc.) committed in each neighbourhood from 2014 to 2019. [7]

d) *2016 City of Toronto Neighbourhood Profiles/Census dataset:* This dataset is based on data collected by Statistics Canada during the last census campaign in 2016 and was used to extract information on unemployment rate per neighbourhood. [8]

e) *Housing rental prices scraped from websites:* This dataset was the most difficult to come by due to the number of neighbourhoods under consideration. Most of the rental information available were for boroughs or only popular neighbourhoods. However, a rental website, *Zumper*, was used to obtain data on average rental prices per neighbourhood. [9]

f) *Top 100 Venues in Toronto neighbourhoods:* This dataset was obtained from Foursquare site using the Foursquare API.

2.2. Data Cleaning

The two Python modules used for Data Wrangling and Manipulation were *Pandas* and *Geopandas* libraries. The *Geopandas* module was used for spatial coordinates and shape files while the *Pandas* module was used to handle all other neighbourhood datasets. There were three main datasets that were used on this project and in this section will discuss how the datasets were pre-processed prior to carrying out the analysis on them.

a) **Neighbourhood Venues:** A request for all venues and categories in the 140 Toronto neighbourhoods was sent to the Foursquare website using the Foursquare API and my user credentials. The limit was set at 100 Venues. The data from the venues were obtained by extracting the relevant parts of the json file and converting it to a dataframe. This was repeated for all 140 neighbourhoods using a created function – (getNearbyVenues). A total number of 2117 different venues were returned for 138 neighbourhoods. The two neighbourhoods with no Venues were *St. Andrew-Windfields and Willowridge-Martingrove-Richview*

	Neighbourhood Name	African Restaurant	Airport Service	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Astrolog
0	Agincourt North	0	0	0	0	0	0	0	0	0	0	0	0	
1	Agincourt South-Malvern West	0	0	0	0	0	0	0	0	0	0	0	1	
2	Alderwood	0	0	0	0	0	0	0	0	0	0	0	0	
3	Annex	0	0	0	0	0	0	0	0	0	0	0	0	
4	Banbury-Don Mills	0	0	0	0	0	0	0	0	0	0	0	0	

Fig 1.0: Toronto neighbourhoods and Venue Categories based on result from Foursquare API

b) **Primary and Secondary Benchmarks:** The primary benchmarks are Unemployment, Crime and COVID-19 rates while the Housing price was considered as the secondary benchmark. The individual datasets were in either csv or excel formats and were converted into a Pandas dataframe. It is also important to note that for all the datasets in this analysis, the Neighbourhood ID was the primary key (index value) to prevent any duplications or errors during merge operations.

The **Unemployment rate dataset** was part of a larger demographics dataset for the City of Toronto. Slicing, transpose and rename methods in Pandas library were used to obtain a clean dataset for further analysis.

For the **Crime rate dataset**, crime rates per 100,000 population were calculated for different crimes committed in 2019 and summed up to one column to give Crime rate. This final crime rate and their respective Neighbourhood IDs were extracted to a new dataframe.
For the **Covid-19 rate dataset**, the Neighbourhood IDs and rate per 100,000 population was extracted from the original dataframe.
The housing dataset was equally read into a pandas dataframe and was utilised in later sections of this project.
All datasets mentioned here were merged on "Neighbourhood ID" column to give one final dataset used to carry out further analysis. See Fig 2.0

	Neighbourhood Name	Neighbourhood ID	Unemployment Rate	Crime Rate	Covid-19 Rate
0	Agincourt North	129	9.80	735.07	542.71
1	Agincourt South-Malvern West	128	9.80	1,384.85	479.86
2	Alderwood	20	6.10	730.05	696.86
3	Annex	95	6.70	1,978.64	782.94
4	Banbury-Don Mills	42	7.20	797.98	314.14

Fig 2.0: Pandas Dataframe showing Toronto neighbourhoods and primary benchmarks

c) Neighbourhood Spatial data: This dataset was used for all map visualizations done in this project. It comprised of geometry for all neighbourhoods and their respective coordinates. The original GeoJSON file described in Section 2.1 was converted to a Geopandas dataframe for use with the Visualization library, Plotly.

3. Methodology

In this section, all Python libraries and techniques used on this Capstone project will be discussed. The libraries and packages used in the Jupyter notebook are listed below:

i) Pandas – For storing and manipulating structured data. Pandas functionality is built on NumPy
ii) Numpy – For multi-dimensional array and matrix data structures.
iii) Geopandas – For storing spatial data coordinates and shape files
iv) Scikit learn – For Machine learning tasks
v) Plotly Visualization Package – For all visualizations (including maps and graphs)
vi) Requests - to send HTTP requests easily
vii) Geopy – To retrieve location coordinates

The main steps for this project can be summarized in the flowchart below:

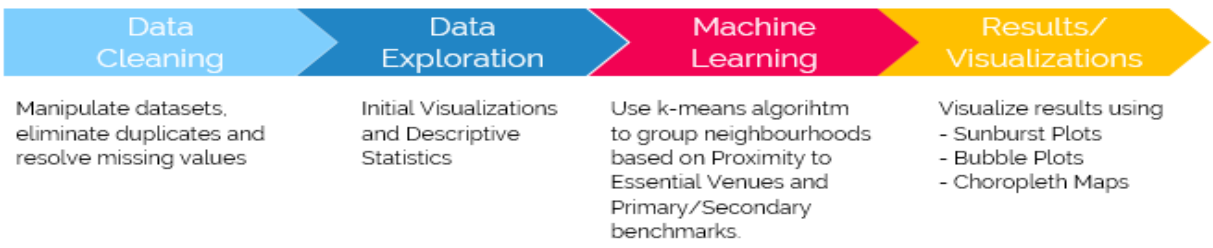

Fig 3.0: Project Flowchart

3.1. Exploratory Data Analysis

3.1.1. Initial Visualization of Toronto Neighbourhoods

The neighbourhoods' spatial coordinates were converted to a *Geopandas* dataframe and imposed on a Street map of Toronto using the Plotly's graphing library. The Scatter Mapbox type under the Plotly express module was used for initial visualization of all Toronto neighbourhoods as shown in Fig 4.0.

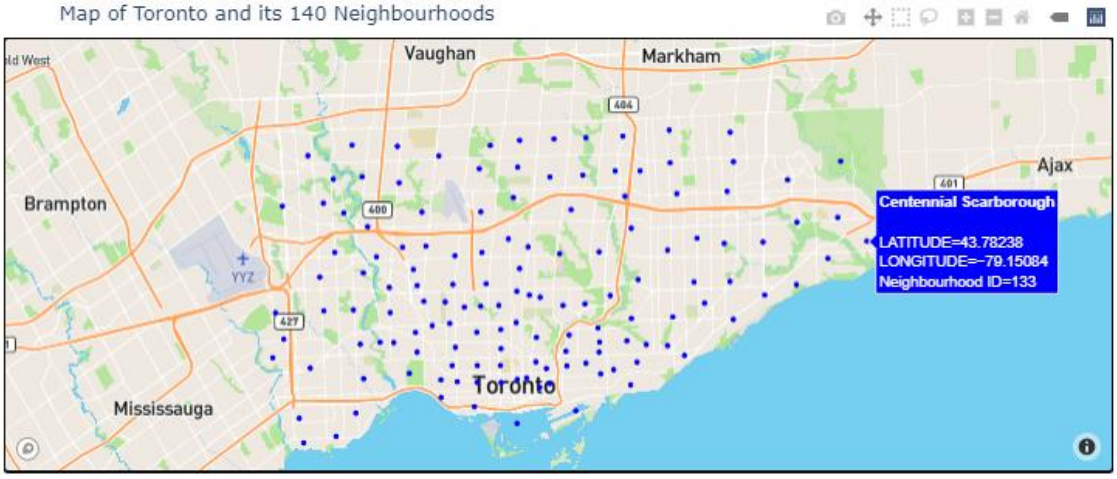

Fig 4.0: Scatter Map of all 140 Toronto neighbourhoods

3.1.2. Determination of Top Venues in Toronto

In this section, the goal was to determine what types of venues were most popular in Toronto. Initial analysis of the dataset from Foursquare website shows that 2118 venues and 291 unique venue categories were obtained.

i) Firstly, One-hot encoding was used to convert venue categories to numerical formats for each neighbourhood.
ii) A new dataframe showing all the neighbourhoods and the total number of each venue type was created. See Fig 1.0 for the resulting dataframe.
iii) Using Transpose and mathematical operation (sum) on the dataframe, all the 291 Venue categories and their total were obtained. The top venues were represented on a bar chart shown in Fig 5.0

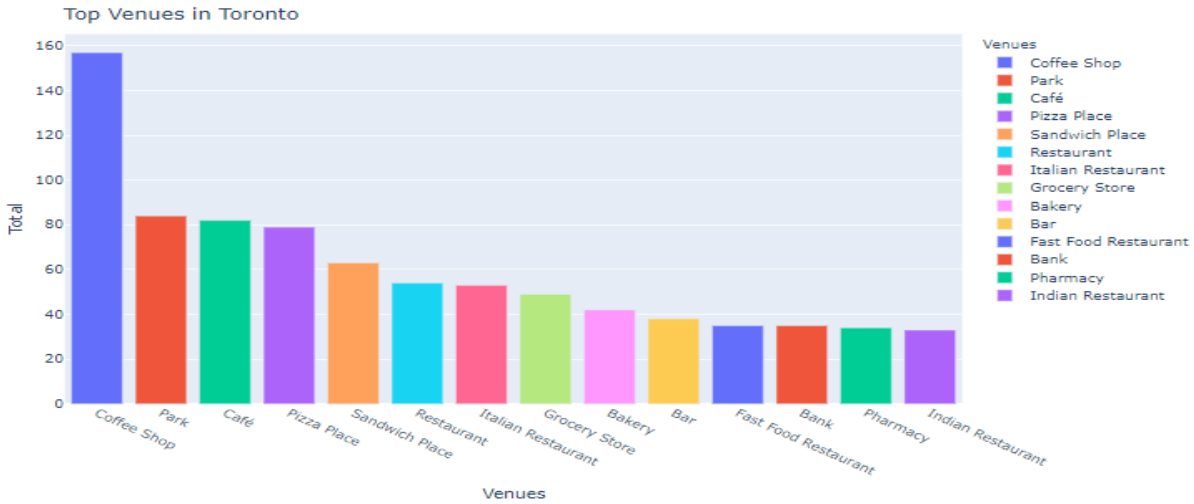

Fig 5.0: Top Venues in City of Toronto

From Figure 5.0, We can see that Coffee shops had the most venues. However, a critical look shows that Restaurants have more venues if we consider the different restaurant sub-categories i.e. (Italian, Fast-food restaurants etc.). After summing up all restaurant sub-categories, there were 900 restaurants in Toronto.

Furthermore, a similar analysis was done for neighbourhoods to determine which neighbourhood had the most venues and the results are shown in Fig 6.0. Church-Yonge and Bay Street Corridor had the most venues. However, the main question is will that necessitate them having a high desirability?

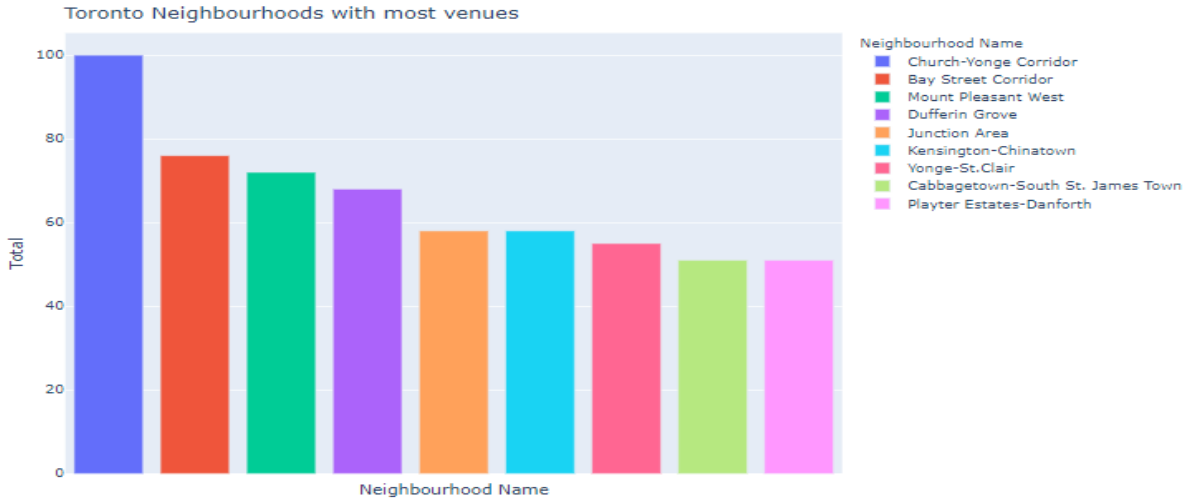

Fig 6.0: Toronto neighbourhoods with most venues

A Sunburst chart was also created as part of this project but that can be viewed in the blogpost and *Jupyter* notebook due to its interactive nature. See the link in the References section [10] [11]

3.1.3. Descriptive statistics for Primary Benchmarks

Initial data analysis was also done on the Unemployment, Crime rate and COVID-19 dataset. The describe function was used to obtain the overall descriptive statistics of the different key indicators. See Fig 7.0 for a summary

	Neighbourhood ID	Unemployment Rate	Crime Rate	Covid-19 Rate
count	140.00	140.00	140.00	140.00
mean	70.50	8.30	1,378.40	913.01
std	40.56	1.90	797.68	584.55
min	1.00	4.50	504.18	236.47
25%	35.75	6.90	869.21	449.91
50%	70.50	8.20	1,232.85	756.24
75%	105.25	9.62	1,532.33	1,129.27
max	140.00	14.60	5,314.57	2,812.36

Fig 7.0: Descriptive Statistics for Primary Benchmarks

Based on the dataframe above, we can see that the average Unemployment rate for City of Toronto is 8.3% for 2019. Average number of crimes committed per 100,000 people is 1378 and 1 in 100 persons has contracted COVID-19 as of October 2020. The percentiles also show some interesting facts such as:

i) 75% of Toronto neighbourhoods have crime rates up to 1532 per 100,000 population while the remainder have crime rates above 1532 per 100,000 population

ii) in 25% of Toronto neighbourhoods, unemployment rate 4.5% -6.9%

3.2. Machine Learning Algorithm

A clustering algorithm called *"k-means"* was used to group the neighbourhoods in order of desirability for new immigrants, k-means is an unsupervised Machine Learning algorithm that groups data into k number of clusters. This method uses a centroid based algorithm to group the neighbourhoods into "k" clusters such that all neighbourhoods with similar characteristics or qualities are in the same cluster. The algorithm works in the following steps:

i) Determine most optimal k (i.e. no of clusters)
ii) Initialize k such that initial means are randomly generated within the data domain
iii) k clusters are created by associating every observation with the nearest mean
iv) The centroid of each of the k clusters becomes the new mean
v) Steps (iii and iv) are repeated until convergence is reached such that all data points belong to a cluster that are significantly distinct from one another

Firstly, k-means clustering was done on the neighbourhoods using the Venue categories (Total number of specific venues) to group each neighbourhood. Another clustering attempt was also done using the primary and secondary benchmarks.

3.2.1. Determining Optimum Number of Clusters

For every clustering attempt, the optimal k was obtained using the "Elbow method". For this method, the dataset is fit with the k-means model for a range of values (1-10). The distortions for each value of *k* is stored and then plotted on a line chart. The point of inflection is a good indication that the model fits best at that point.

The KneeLocator function from the Kneed library can also be used to determine this point of inflection. A snapshot of the code is shown in Fig. 8.0

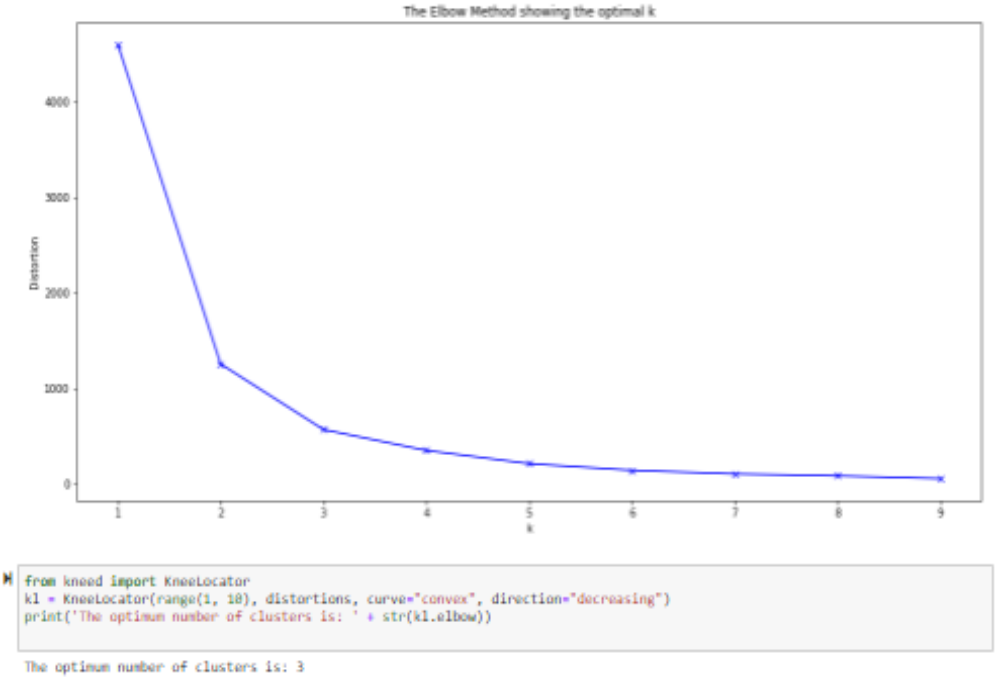

Fig 8.0: Elbow Plot and KneeLocator Function

The optimum k obtained for all clustering attempts was 3 meaning that all 140 neighbourhoods would be grouped into 3 clusters for each clustering attempt.

3.2.2. Clustering Neighbourhoods by Total number of Essential Venues

For this clustering attempt, the aim was to group neighbourhoods based on the number of essential venues that were available in them. These essential venues included places such as Schools, Train stations, Restaurants, Banks, Shopping Malls, Bus Stations etc. For each neighbourhood, the total number of these essential venues were obtained from the larger dataset in Section 3.1.2 and with the *k-means* algorithm, the neighbourhoods were grouped into 3 clusters.

3.2.3. Clustering Neighbourhoods by Primary Benchmarks

In this second clustering attempt, the neighbourhoods were grouped on the three indicators – Unemployment, Crime, and COVID-19 rates. These benchmarks were considered as *primary benchmarks*.

3.2.4. Clustering Neighbourhoods by Secondary Benchmark

Using the outcome of the clustering attempt in Section 3.2.3, the neighbourhoods with lowest Unemployment, Crime and COVID-19 rates were further grouped based on their Housing prices (for one-bedroom apartment). The Housing prices were considered as *secondary benchmark*. The outcome of this final clustering attempt was used to generate the final Neighbourhood Desirability Index.

4. Results and Discussion

Using the steps explained in Section 3, we were able to obtain two distinct maps for Toronto:

- Neighbourhoods Venues Density Map: Provides insight on which neighbourhoods have high, mid, and low number of essential venues
- Toronto Neighbourhood Desirability Map: Shows neighbourhoods based on their desirability with respect to Primary and Secondary benchmarks

4.1.1. Neighbourhoods Venues Density Results and Visualizations

Based on the outcome of the clustering attempt described in section 3.2.2, the 140 Toronto neighbourhoods were put into three distinct clusters. These Clusters have the characteristics described in the Table 1.0.

Table 1.0: Distribution of neighbourhoods based on Total number of Essential Venues

Clusters	Description	No of Neighbourhoods per Cluster	Average Number of Essential Venues	Minimum - Maximum Venues
0	Low Venue Density	92	2	0-4
1	High Venue Density	14	19	14-28
2	Mid Venue Density	34	8	5-13

From the table above, we can see that about 66% of Toronto neighbourhoods have low concentration of “essential” venues within them. The neighbourhoods with high venue density make up only 10% of Toronto neighbourhoods. These neighbourhoods are shown in Table 2.0.

TABLE 2.0: Top neighbourhoods with essential venues

S/No	Neighbourhood Name	Total
1	Mount Pleasant West	28
2	Church-Yonge Corridor	27
3	Yonge-St. Clair	23
4	Bay Street Corridor	23
5	Kensington-Chinatown	20
6	Greenwood-Coxwell	20
7	Playter Estates-Danforth	19
8	Lawrence Park North	17
9	Agincourt South-Malvern West	17
10	Dufferin Grove	17
11	Cabbagetown-South St. James Town	16
12	Yonge-Eglinton	15
13	Corso Italia-Davenport	15
14	South Parkdale	14

The Final Map showing the Neighbourhood Venue Density was developed using Plotly Package and is shown in Fig. 9.0

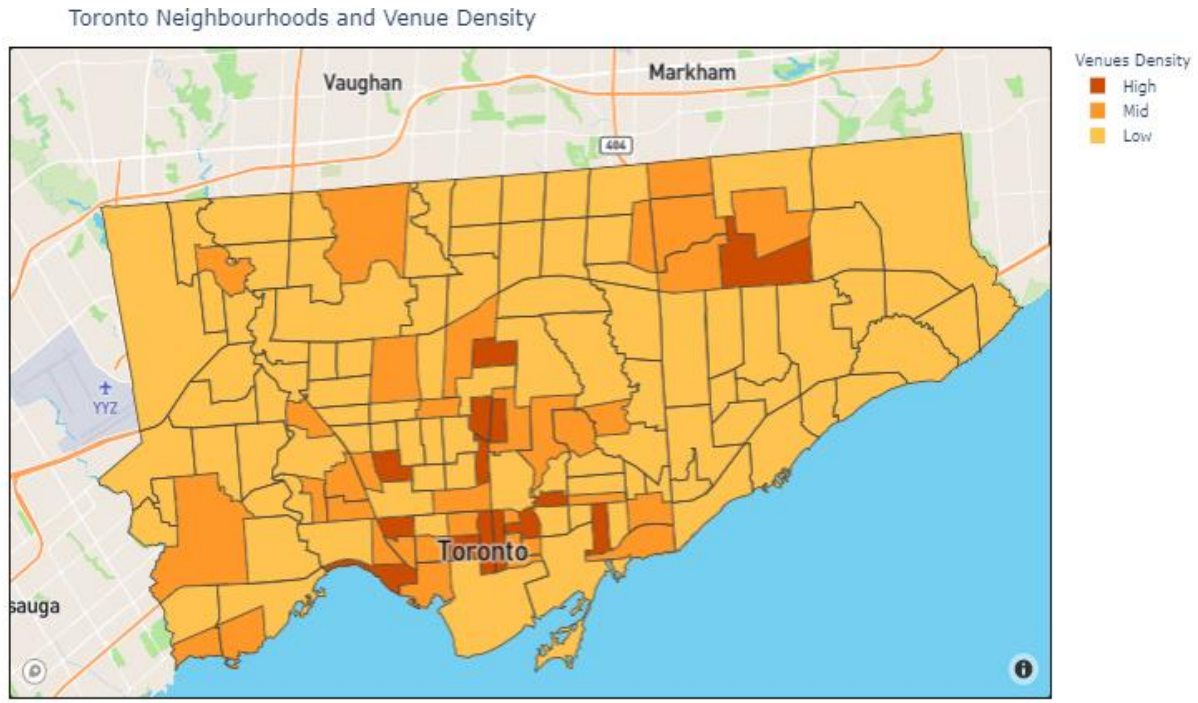

Fig 9.0: Venues Density of Toronto Neighbourhoods

The Plotly Visualizations are very interactive, and this was the reason why they were selected for this project. Fewer lines of code are also required to develop choropleth maps such as the one shown in Fig 10.0. A user is also able to isolate a desired section of the choropleth map to show only the neighbourhoods he/she is interested in i.e. only High Venue Density areas. This can be further explored in the Jupyter notebook platform. [10]

It is also interesting to note that the area of the map with the higher concentration of venues is Downtown Toronto knows as the Entertainment District of Toronto.

4.1.2. Neighbourhoods Desirability Index Results and Visualizations

For the second clustering attempt, the neighbourhoods were first grouped using the *primary benchmarks* – Unemployment, Crime and COVID-19 rates. There were 3 clusters with the results shown in Fig 10.0 and Table 3.0.

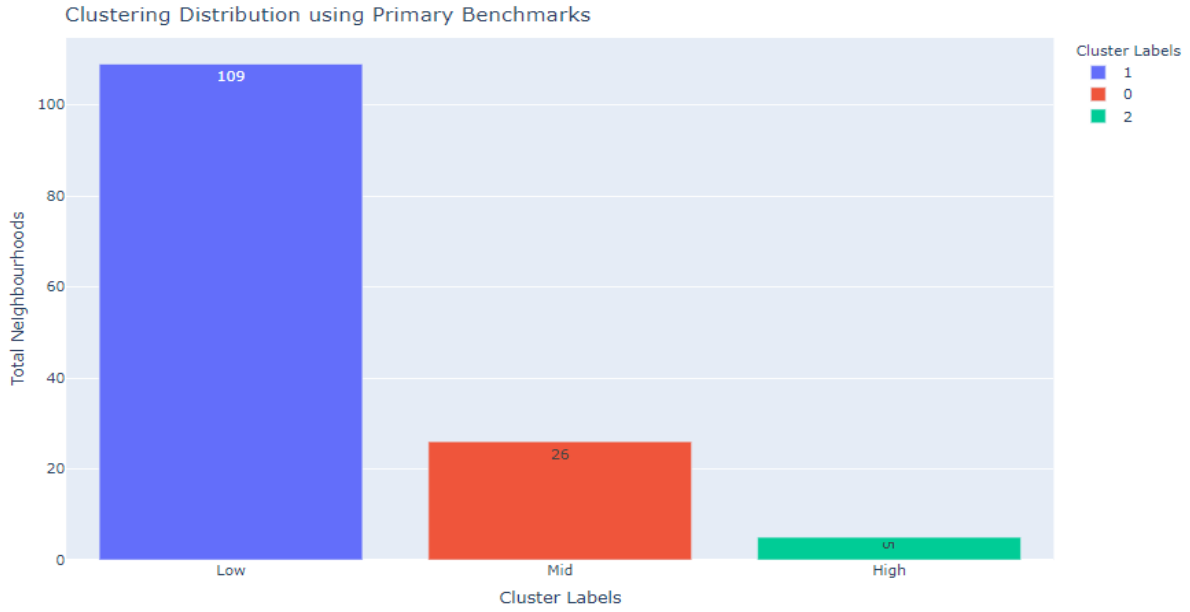

Fig 10.0: Outcome of Clustering using Primary benchmarks

Cluster 1 (Low): Neighbourhoods with lowest Unemployment, Crime and COVID-19 rates

Cluster 0 (Mid): Neighbourhoods with relatively high crime rates* but higher Unemployment and COVID-19 rates

Cluster 2 (High): Neighbourhoods with highest crime rates* but mid Unemployment and COVID-19 rates

* Crime rates were considered the worst of the 3 metrics and this is the reason Cluster 2 was classified as High. A summary of the results can be seen in Table 2.0.

Table 2.0: Result of Clustering using Primary Benchmarks

Cluster Labels	Description	Avg. Unemployment Rate (%)	Avg. Crime Rate (per 100,000 population)	Avg.Covid-19 Rate* (per 100,000 population)
0	Mid	9.75	1648.58	1942.23
1	**Low**	**7.95**	**1169.51**	**669.09**
2	High	8.42	4527.06	878.59

* COVID-19 rates are as of October 22, 2020

Consequently, the 109 neighbourhoods in Cluster 1 (lowest rates) were further grouped using Housing prices as the secondary benchmark. This yielded the result shown in Fig 11.0

Fig 11.0: Outcome of Clustering using Secondary benchmarks

From Fig 12.0, we can see that of the 109 neighbourhoods in Fig 11.0, 7 of them had the lowest housing price with an average of $990 for a one-bedroom apartment. A break down of the average housing prices is shown for each cluster is shown in Table 3.0.

Table 3.0: Housing Prices Categories

Cluster Labels	Housing Prices Description	Average Price (CAD$)
0	Low	990.57
1	Mid	1700.07
2	High	2070.25

Finally, the results from this clustering attempt were used to rank the neighbourhoods into 4 categories. The neighbourhoods that belonged to the *"Mid" and "High"* clusters using primary benchmarks were classified as the Least desirable neighbourhoods while those with Low, Mid and High housing prices were classified as Most Desirable, Desirable and Semi-Desirable respectively. The final distribution of the neighbourhoods is shown below

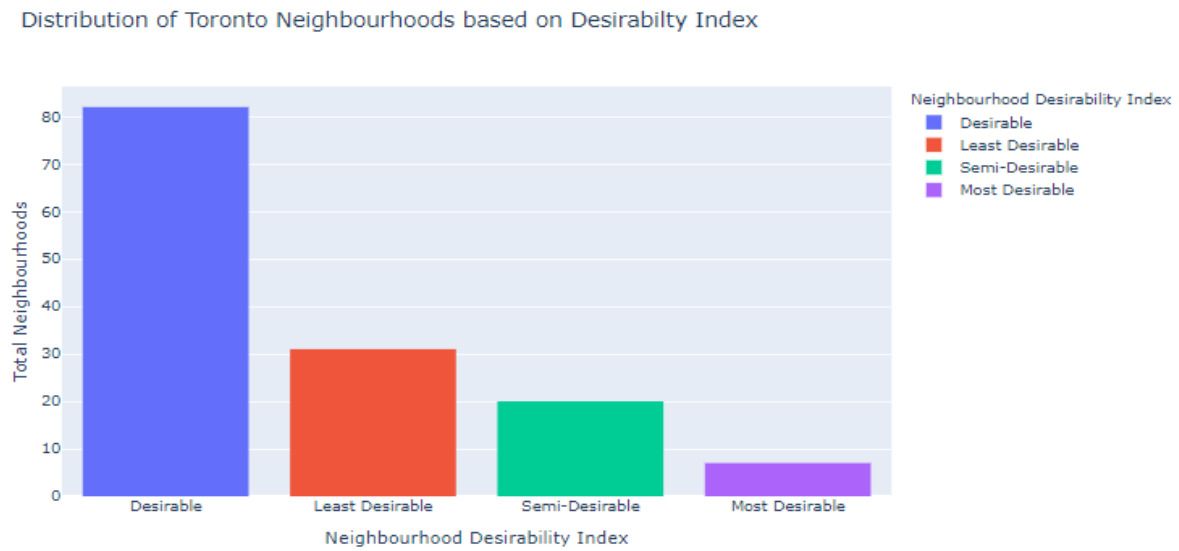

Fig 12.0: Distribution of Toronto neighbourhoods based on Desirability index

From Fig 12.0, we can see that only 7 neighbourhoods fell into the Most Desirable index rank while about 59% of the Toronto neighbourhoods were grouped into the Desirable category based on their medium housing prices and relatively low crime, COVID-19 and Crime rates.

In putting all these onto a map of Toronto, we have the final choropleth map shown in Fig. 13.0.

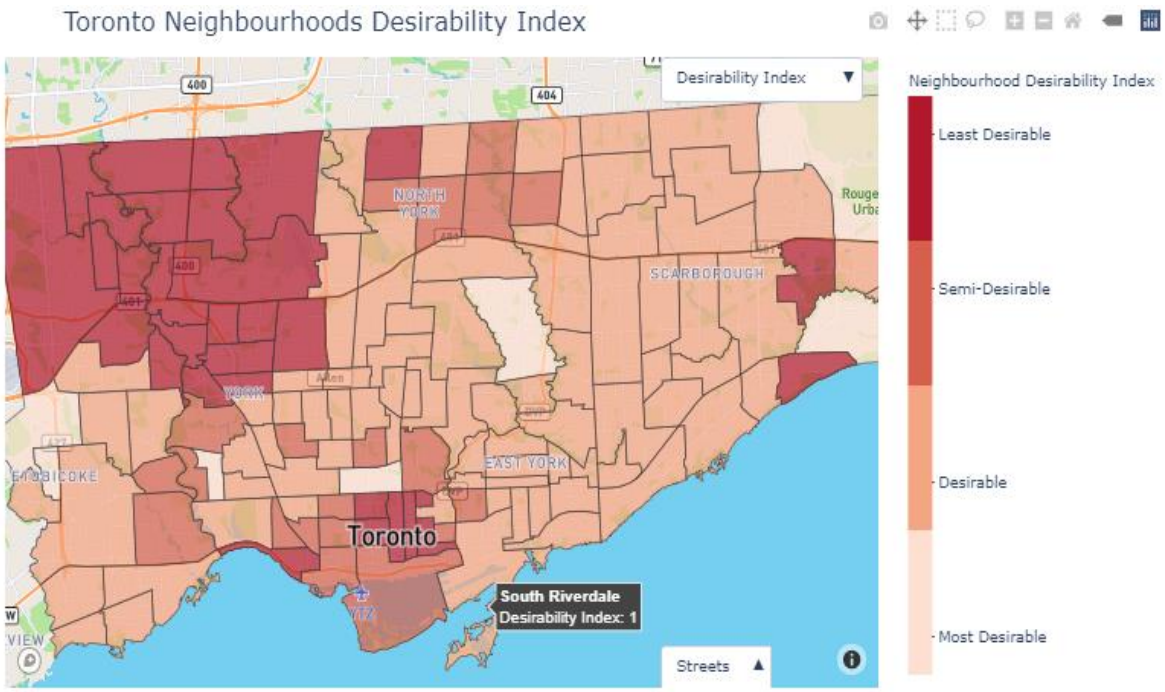

Fig 13.0: Toronto Neighbourhoods Desirability Index Map

5. Conclusion

From the results, we can make the following deductions:

i) Only 10% of Toronto neighbourhoods have high venue density with Mount Pleasant West, Church-Yonge Corridor, Yonge-St. Clair, and Bay Street Corridor taking the lead

ii) **Most Desirable Neighbourhoods:** Consider neighbourhoods in Scarborough area if searching for less pricey apartments. Other neighbourhoods to consider are Banbury Don-Mills and Annex in North York and York districts respectively

iii) **Looking for Entertainment:** Look no further than Downtown Toronto which is also known as the Entertainment District. This area was classified Semi-desirable owing to the higher housing prices. However, if you are looking for fun and have the $$$, it is a great place to settle in

iv) **Presence of Essential Venues:** If you are keen on proximity to essential venues, the neighbourhoods to consider which are also in the Desirable category are Mount Pleasant West, Yonge-St, Clair and Greenwood-Coxwell

v) **Avoid if you Can:** Most neighbourhoods in the North-Western region of Toronto i.e. Etobicoke district were classified as the Least desirable due to the high crime and COVID-19 rates in those neighbourhoods. It is also interesting that this region is home to Jane and Finch which is a “red” neighborhood.

References

Click or Tap on the any of the links below to view the resource

[1] News Article: Canada to welcome over 1.2 million immigrants

[2] US News Best Countries for Immigrants Rankings

[3] News Article: Canada Immigration Statistics

[4] 2016 Canada Population Census

[5] Boundaries of City of Toronto Neighbourhoods

[6] COVID-19: Status of Cases in Toronto

[7] City of Toronto Neighbourhood Crime Rates

[8] City of Toronto Neighbourhood Profiles

[9] Average Housing Rental Prices for City of Toronto

[10] GitHub Repository

[11] Jupyter Notebook Viewer