Impact of Toronto neighbourhood on its housing rentals

Babita Khanna

May 04, 2021

1. Introduction

1.1. Background

Toronto is the capital city of the Canadian province of Ontario and the largest city in Canada. With a recorded population of 6,341,935 in the Toronto region as reported by Toronto City Hall. It is the most populous city in Canada and the fourth most populous city in North America.

Toronto covers an area of 630 square kilometres (243 sq. mi), with a maximum north–south distance of 21 kilometres (13 mi). The strength and vitality of the many neighbourhoods that make up Toronto, Ontario, Canada has earned the city its unofficial nickname of "the city of neighbourhoods". There are over 140 neighbourhoods officially recognized by the City of Toronto and upwards of 240 official and unofficial neighbourhoods within city limits.

1.2. Business Problem

Given its population census and variety in neighbourhood, there is a huge demand of exploring and scouting locations in Toronto by various stakeholders for numerous opportunities.

For an investor to launch their business, open a restaurant or a coffee shop, there requires intense research and analysis to search for places with least competition and untapped opportunities for a specific product demand such as cuisine type. Also, for people looking to reside in this popular city, there is a requirement to investigate these areas of specific neighbourhood types with for instance, parks & playgrounds for children, least real estate values or with specific criteria such as least dense population etc. as per their social preference.

Considering this requirement at hand to be resolved, the idea is to investigate how the rental housing rate data and foursquare APIs can be used to answer these questions proactively with the power of data science analytics resolving the demand called out by the said target audience.

2. Data

2.1. Data sources

To resolve the business problem described, following data sources are extracted and explored:

a. *Canadian Postal codes* is an excellent wiki site to retrieve Toronto Neighbourhood details
b. Comma Separated Value file with *Geospatial Coordinates of all Toronto Neighbourhood* to tag the neighbourhood data with its location - latitude and longitude information
c. Public information on Toronto renting index is not readily available. Hence, using the *Housing Market Information Portal* available on Canada Mortgage and Housing Corporation, rental property rates is gathered for each kind of accommodation for the past few years. Computing the average and relevant information, the data sources of Canadian postal codes as well as the geospatial coordinates are combined to represent the average housing range on the recorded properties
d. *Foursquare API* is used to extract all the nearby venues for all combined neighbourhood data.

The following methodology section is a walkthrough of how the data described is investigated, used to create visualization maps, information charts & tables and finally the rental data information on the map is plotted with the price range by clustering based on nearby venue density.

3. Methodology

3.1. Retrieving Data and Pre-Processing

Postal code data of the complete Toronto city and its neighbourhood is extracted from the online wiki site using Data Wrangling via Beautiful Soup feature which reads the HTML data of the website. The table tag is extracted from the raw data retrieved and was processed. The resulting information is stored in a pandas DataFrame with Postal Code, Borough and Neighbourhood information.

To find corresponding location coordinates for all neighbourhood information collected, Geopy library is used to make API requests to Nominatim to retrieve respective geospatial coordinates of latitudes and longitudes. The coordinates retrieved is then merged with the existing pandas DataFrame to the corresponding neighbourhood data.

After much exploration, Housing Market Information Portal found on the Canada Mortgage and Housing Corporation website offers latest housing market data for Canada provinces highly vital for this analysis project. Average rent on the primary rental market by bedroom type is available in various formats. The data is retrieved by using the pandas read html property and the information is stored in another

pandas DataFrame. Using various pre-processing methods, the data is cleaned, converted into required datatypes, irrelevant data columns are removed and data with any null values is deleted. The final information is then merged with the existing pandas DataFrame based on the neighbourhood data which is saved as final DataFrame as shown in the Table 1 below.

	Neighborhood	Borough	Latitude	Longitude	Studio	1BHK	2BHK	3BHK+
0	Ontario Provincial Government	Queen's Park	43.6623	-79.3895	933	1464	1701	2018
1	Garden District	Downtown Toronto	43.6572	-79.3789	1037	1246	1655	2137
2	West Deane Park	Etobicoke	43.6509	-79.5547	1195	1243	1356	1482
3	Don Mills South	North York	43.7259	-79.3409	1091	1281	1433	1647
4	Woodbine Heights	East York	43.6953	-79.3184	962	1003	1197	1493

DataFrame Table 1: Combined data retrieved from data sources for analysis

3.2. Data Analysis

To understand the rental data further, the information was plotted on an histogram for each type of accommodation depicting average sales prices for each month in the Toronto city and its neighbourhood.

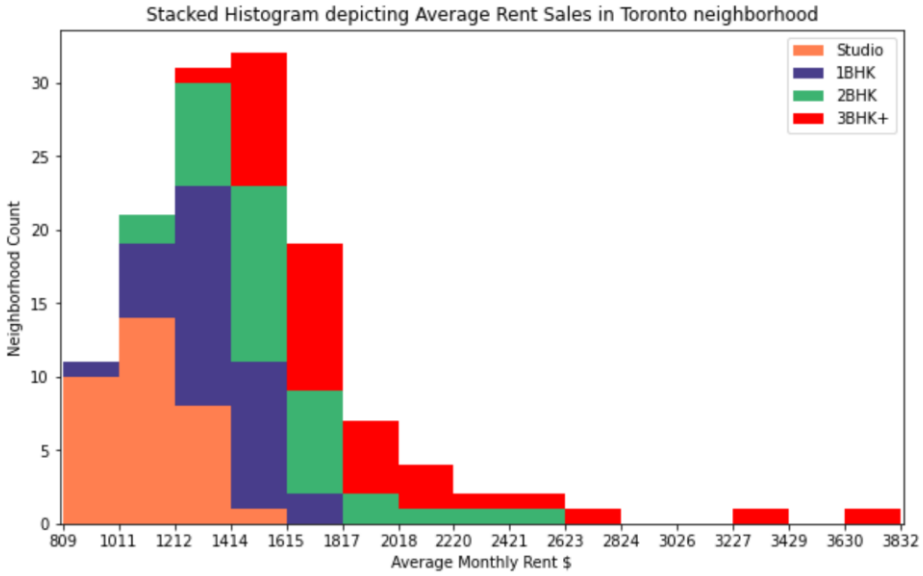

Figure 1: Stacked histogram depicting Average Monthly Rent Sales ($)

As seen in the figure 1 above, most of the rental property rates fall in the lower price range of the graph presenting that most of the housing rental rates must be lower. One of the reasons could be the 2020 pandemic impacting the economy and the Canadian currency rate is running strong.

This is a very good indicator for stakeholders as it impacts their decision-making process. Average rent is then calculated for all the accommodation types and based on the price range, the neighbourhood data is segmented and labelled into four ranges: Low Monthly Rent, Low-Medium Monthly Rent, High-Medium Monthly Rent and High Monthly Rent. These labels are added into a new column variable along with rest of the data.

The best way to visualize the neighbourhoods is by plotting them onto a map. To set a base location, Nominatim is used to retrieve location data for Toronto. The neighbourhood data with its latitude and longitude coordinates is used to plot the folium map. Markers are used to pinpoint the locations are on the map.

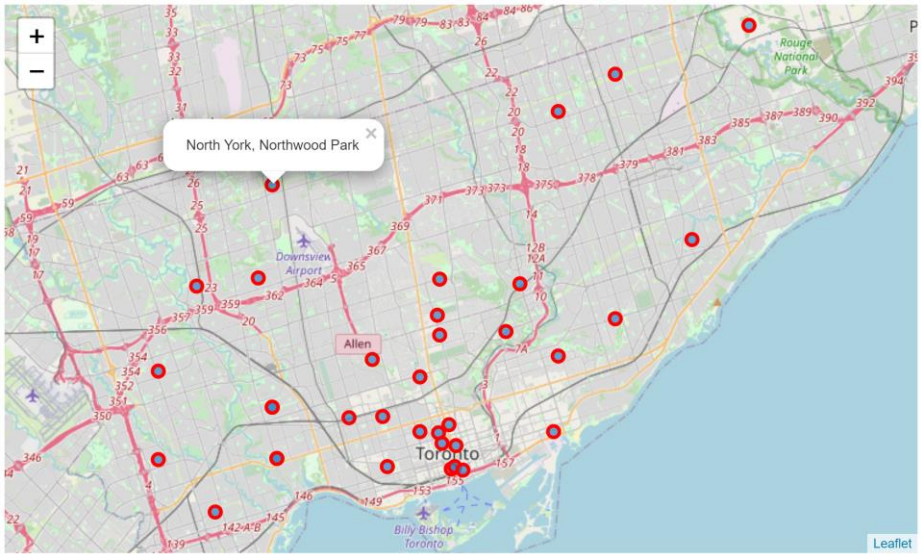

Figure 2: Folium map visualizing geographic details

To further analyse these locations, Foursquare API is utilized to provide location-based experiences with diverse information about venues, users, photos, and check-ins. This also supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. For this project, the API is used to retrieve all nearby venues of each neighbourhood location with the set limit of 100 results and within a radius of 500m. The request retrieved 935 results which is stored in another DataFrame as shown in the table 2 below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ontario Provincial Government	43.662301	-79.389494	Queen's Park	43.663946	-79.392180	Park
1	Ontario Provincial Government	43.662301	-79.389494	Mercatto	43.660391	-79.387664	Italian Restaurant
2	Ontario Provincial Government	43.662301	-79.389494	NEO COFFEE BAR	43.660130	-79.385830	Coffee Shop
3	Ontario Provincial Government	43.662301	-79.389494	Nando's	43.661728	-79.386391	Portuguese Restaurant
4	Ontario Provincial Government	43.662301	-79.389494	T-Swirl Crepe	43.663452	-79.384125	Creperie

DataFrame Table 2: Data retrieved from Foursquare API for nearby venues based on location coordinates

Based on the information retrieved using the Foursquare API, a bar chart is created to visualize the number of venues for each neighbourhood as shown in Figure 3.

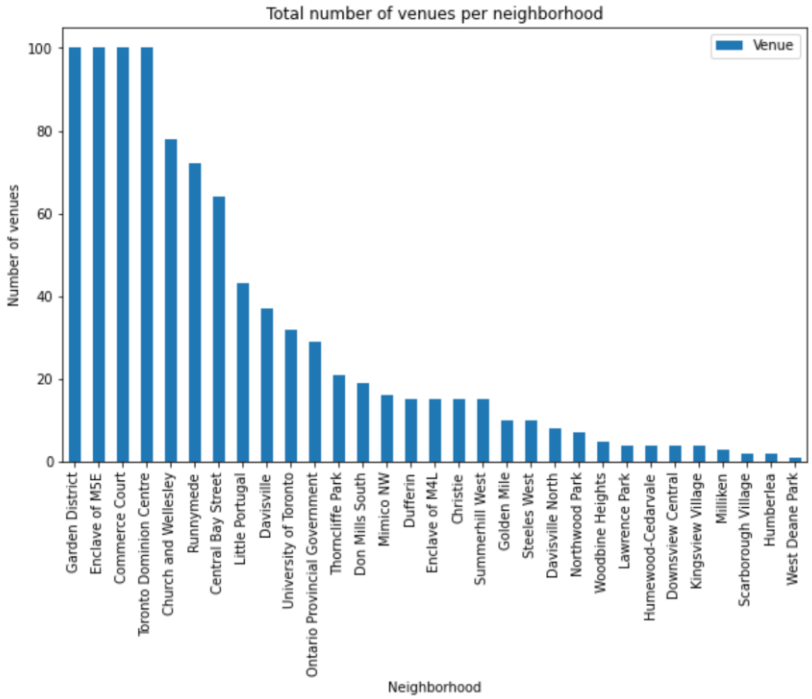

Figure 3: Number of venues per neighbourhood

The bar chart shows that the neighbourhoods Garden District, MSE Enclave, Commerce Court & Toronto Dominion Centre has reached the limit of 100 venues while West Deane Park, Humberlea, Scarborough Village, Mililiken, Kingsview Village and few more others have way less than 10 nearby venues. Based on the geographical coordinates information extracted, the inquiry result may vary. Geocoordinates with higher accuracy may increase the possibility of venue results.

The Venue Category column retrieved from the Foursquare API provides more insight to the neighbourhood areas. There were 197 unique venue categories for the complete neighbourhood. Using One Hot Encoding, venue categories were split into columns for each neighbourhood, normalized and the top 10 venues were sorted into a staging DataFrame.

3.3. Data segmentation and exploration

Based on the data analysed, a lot of venue category types were common amongst the Toronto neighbourhood. With that assumption, clustering the data will be best approach to classify the neighbourhood data into structures for better understanding and manipulation. K-Means clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome dividing the data into k number of clusters.

To conduct k-means clustering, the number of clusters to divide the data into must be defined first which can be decided based on the elbow method. The elbow method is one of the most popular methods to determine this optimal value of k. The results are plotted on a graph per k value against distortion which is basically the Euclidean distances from the cluster centres of the respective clusters.

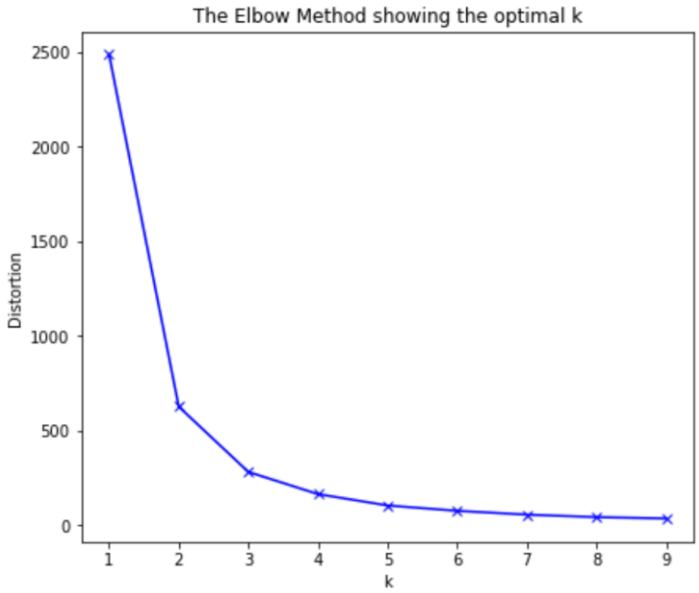

Figure 4: Distortion shown per k value based on elbow method

The Figure 4 shows the graph plotted for the elbow method resulting in k=3 as the optimal value. This k value is used to cluster the venue data after which each neighbourhood data is labelled with a cluster number. This data is then merged into the existing DataFrame with the rental prices.

Neighborhood	Borough	Latitude	Longitude	Average Rent	Cluster Labels	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
Ontario Provincial Government	Queen's Park	43.6623	-79.3895	1529.00	1	Coffee Shop	Sushi Restaurant	Yoga Studio	Smoothie Shop	Café	Sandwich Place	College Auditorium	College Cafeteria	Portuguese Restaurant	Creperie
Garden District	Downtown Toronto	43.6572	-79.3789	1518.75	2	Clothing Store	Coffee Shop	Middle Eastern Restaurant	Café	Italian Restaurant	Japanese Restaurant	Bubble Tea Shop	Cosmetics Shop	Lingerie Store	Movie Theater
West Deane Park	Etobicoke	43.6509	-79.5547	1319.00	1	Filipino Restaurant	Bakery	Yoga Studio	Donut Shop	Fish & Chips Shop	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Ethiopian Restaurant
Don Mills South	North York	43.7259	-79.3409	1363.00	0	Coffee Shop	Gym	Restaurant	Sporting Goods Shop	Discount Store	Bike Shop	Beer Store	Supermarket	Italian Restaurant	Sandwich Place
Woodbine Heights	East York	43.6953	-79.3184	1163.75	1	Curling Ice	Park	Video Store	Athletics & Sports	Skating Rink	Intersection	Beer Store	Yoga Studio	Eastern European Restaurant	Field

DataFrame Table 3: Cluster Labels with top 10 venues for each neighbourhood

Based on most common venue category for each neighbourhood, another bar graph is plotted for each of these clusters to understand the type of venues that are surrounding the area. This is helpful to generalize the neighbourhood area for the investors to scout for location based on the surroundings and brainstorm what opportunities are available to ensure a productive and successful investment.

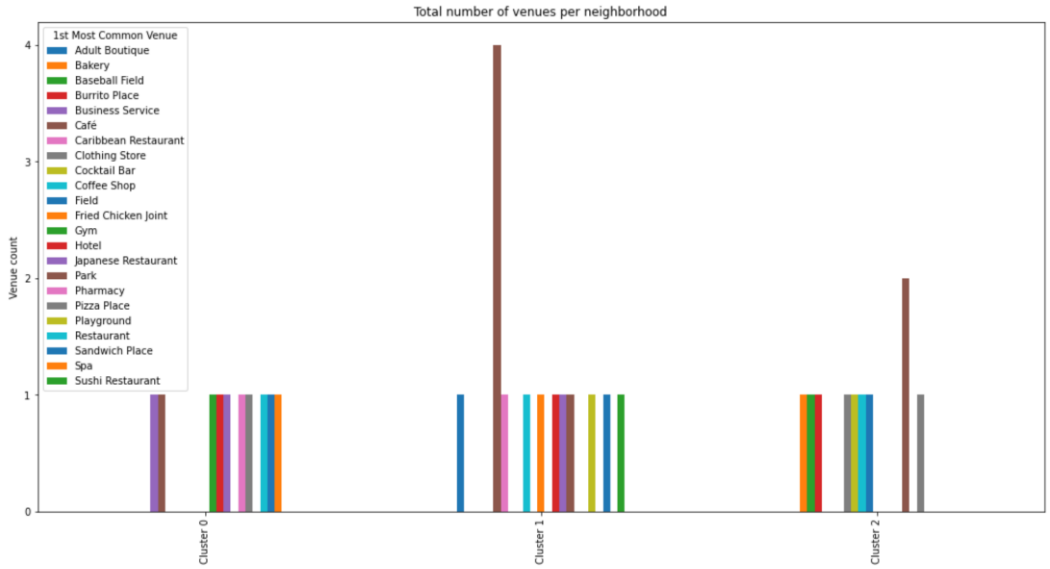

Figure 5: Venue category distribution per neighbourhood cluster

Based on the bar chart information, the neighbourhood data is labelled as below per respective cluster:

- Cluster 0: Dine in Restaurants
- Cluster 1: Café and Gym Venues
- Cluster 2: Parks, Playgrounds & Good Transport Service

Another label was added to the neighbourhood DataFrame with the top 3 venues of the area.

4. Results

The master DataFrame consists of the base rental price data along with new variables with cluster information as shown in the table 4 below. The label variables 'Rent Price Level', 'Cluster Labels', 'Neighbourhood Type' and 'Top 3 Nearby Venues' are the most important outputs which can greatly impact the stakeholder's decision-making.

Neighborhood	Borough	Latitude	Longitude	Studio	1BHK	2BHK	3BHK+	Average Rent	Rent Price Level	Cluster Labels	Neighborhood Type	Top 3 Nearby Venues
Ontario Provincial Government	Queen's Park	43.662301	-79.389494	933.0	1464.0	1701.0	2018.0	1529.00	Low-Medium Monthly Rent	1	Cafe and Gym Venues	Coffee Shop, Sushi Restaurant, Yoga Studio
Garden District	Downtown Toronto	43.657162	-79.378937	1037.0	1246.0	1655.0	2137.0	1518.75	Low-Medium Monthly Rent	2	Parks, Playgrounds & Good Transport Service	Coffee Shop, Clothing Store, Middle Eastern Re...
West Deane Park	Etobicoke	43.650943	-79.554724	1195.0	1243.0	1356.0	1482.0	1319.00	Low Monthly Rent	1	Cafe and Gym Venues	Bakery, Adult Boutique, New American Restaurant
Don Mills South	North York	43.7259	-79.340923	1091.0	1281.0	1433.0	1647.0	1363.00	Low Monthly Rent	0	Dine-in Restaurants	Restaurant, Gym, Coffee Shop
Woodbine Heights	East York	43.695344	-79.318389	962.0	1003.0	1197.0	1493.0	1163.75	Low Monthly Rent	1	Cafe and Gym Venues	Curling Ice, Park, Intersection

DataFrame Table 4: Master DataFrame with merged label variables

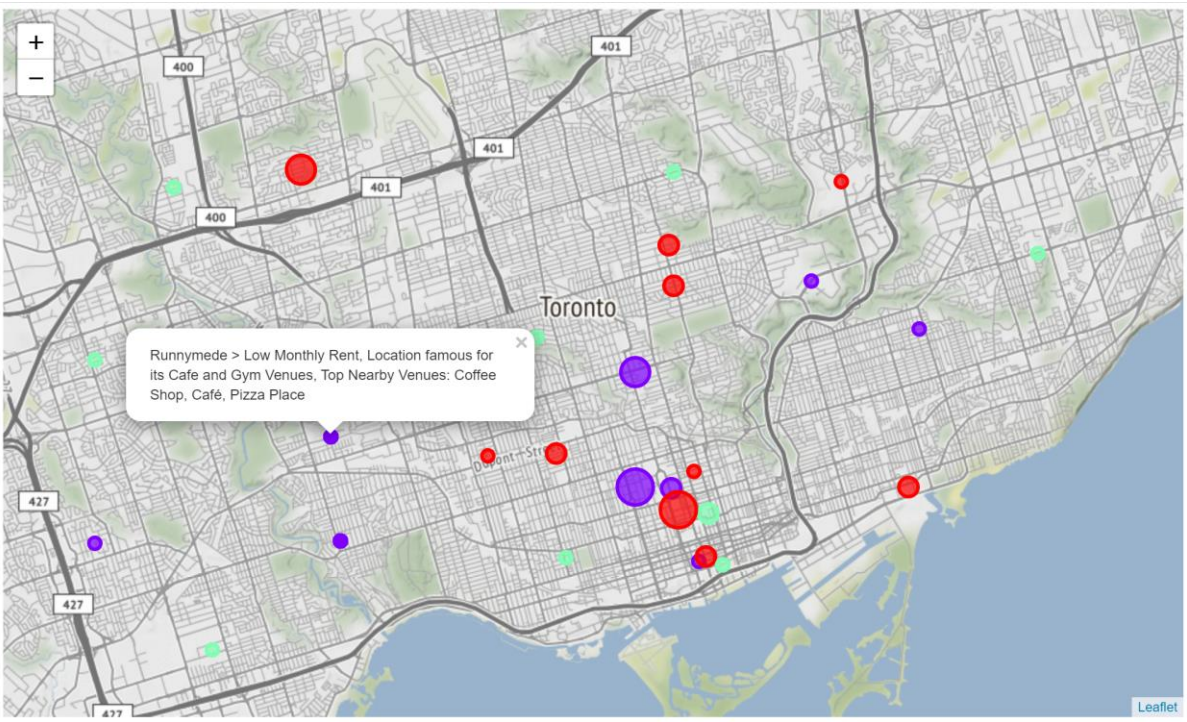

Figure 6: Folium map representing the labels with relevant neighbourhood data

The folium map in figure 6 is the accurate representation of the exploration and analysis of the raw data discussed for this project displaying all the label variables for stakeholders to explore.

5. Discussion

Based on the rental data range, the neighbourhood was also segmented into four tiers: Low Monthly Rent, Low-Medium Monthly Rent, High-Medium Monthly Rent and High Monthly Rent ranges. It was seen most of the neighbourhood fall under the Low Monthly Rent tier followed by Low-Medium Monthly Rent tier.

With the help of the K-means algorithm elbow method, the optimum k value was 3 for the limited data for the clustering analysis. However, this data set can be expanded much more with finer neighbourhood and street details for more detailed and accurate guidance.

There are three main clusters the complete neighbourhood dataset has been segmented into where each are dominated by two-three types of venue categories. The first cluster 0 is known for its Pizza Parlours & Dine-in Restaurants, second cluster 1 for its Cafe and Gym Venues and the third cluster 2 for its Parks, Playgrounds & Good Transport Service. Another information value was added to the result showing us the top three venues in each of the neighbourhood allowing investors and future residents to choose and decide from.

As previously discussed, Toronto is the largest and most populous city in Canada within a narrow area. The total number of measurements and population densities of boroughs in total can vary. As there is such a complexity, very different approaches can be tried in

clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high-quality results for this city.

6. Conclusion

My motive through this project was to answer questions raised by investors and future residents representing their preference of turning to bigger cities to start a new business or set up a new life. Many can achieve way better outcomes through access to such platforms and applications where this information is provided. City or Town hall management and managers can also heavily benefit by using such data analysis reports and platforms.

Future studies can be carried out using various available platforms and packages similarly to my data analysis with the available neighbourhood geographical coordinates and housing market rent data. Where I have presented data visualization of the complete dataset on Toronto folium map, web applications can be built similarly reaching out directly to the targeted investors.

7. References

a. Canadian Postal codes
b. Geospatial Coordinates of all Toronto Neighbourhood
c. Housing Market Information Portal - Canada Mortgage and Housing Corporation
d. Foursquare API