Comparison of Toronto Neighbourhoods as Rental Locations

Manan Bhati

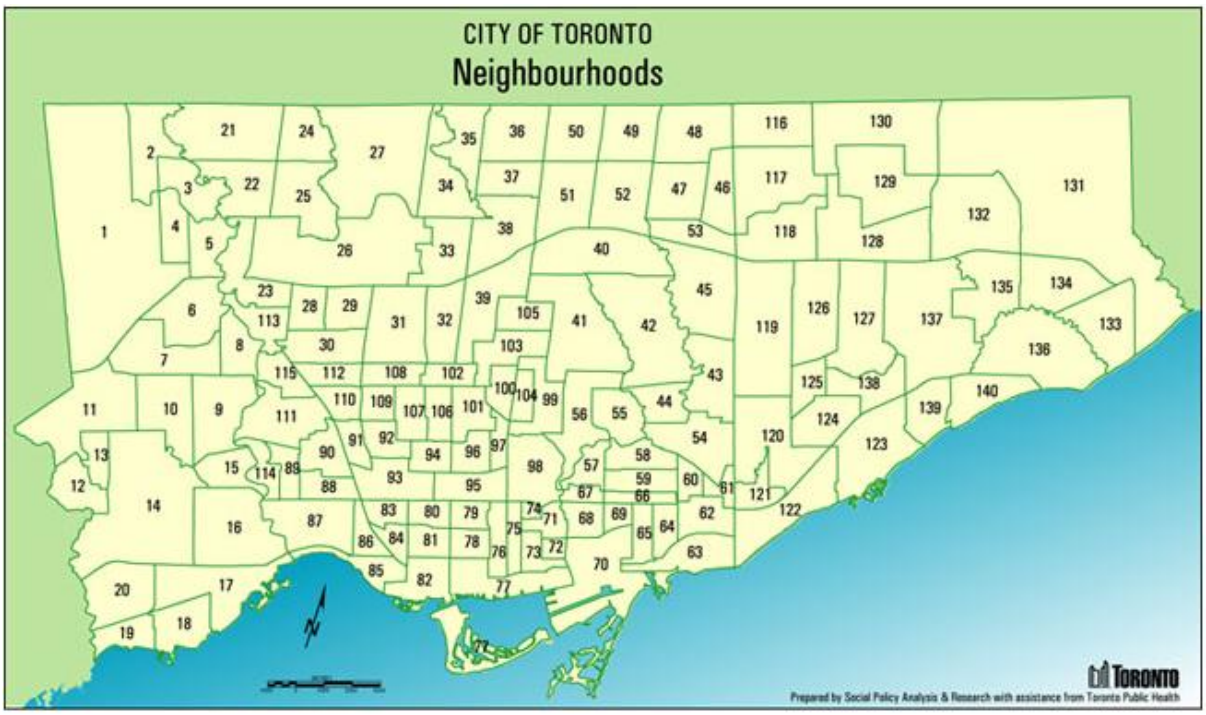
March 14, 2020

1. Introduction

1.1. Background

Toronto is the capital of the Ontario province in Canada and is the most populous city in Canada with a population of over three million residents. It is one of the world's most well-known cosmopolitan cities and is a global hotspot of commerce, finance and arts. Toronto's diverse population continues to boom with a steady influx of immigrants from all over the world. The Greater Toronto Area (GTA), which consists of the central city Toronto and its 25 surrounding suburbs, is the top choice for landing for newly arrived immigrants in Canada.

The cityscape of Toronto consists of 140 officially recognized neighbourhoods, each with its own distinct cultural, demographic and socio-economic profile. For a person or family recently arrived in Toronto, their choice of a neighbourhood to rent a residence in would depend upon several factors, including but not limited to – the distance to their work location, rent affordability, quality of life considerations like availability of good schools, recreational facilities, green spaces, etc., and ethnic and cultural affinities. A profile of each Toronto neighbourhood is available at the City of Toronto's website here.



1.2. Problem

An analysis that a person new to Toronto would find helpful in deciding where to take up a rental apartment in the city might include data like average monthly rent per neighbourhood, the proximity of the apartment to the person's job location, availability of public transport for commuting to work, reasonably close supermarkets, schools, entertainment and recreational venues, the number of available rental properties in the neighbourhoods of interest, and so on. This project aims to present the user with Toronto neighbourhoods with residential apartments/ condominiums, clustered on the basis of their distance from the user's work location, and the average monthly rent of a two bedroom apartment in these neighbourhoods.

1.3. Target Audience

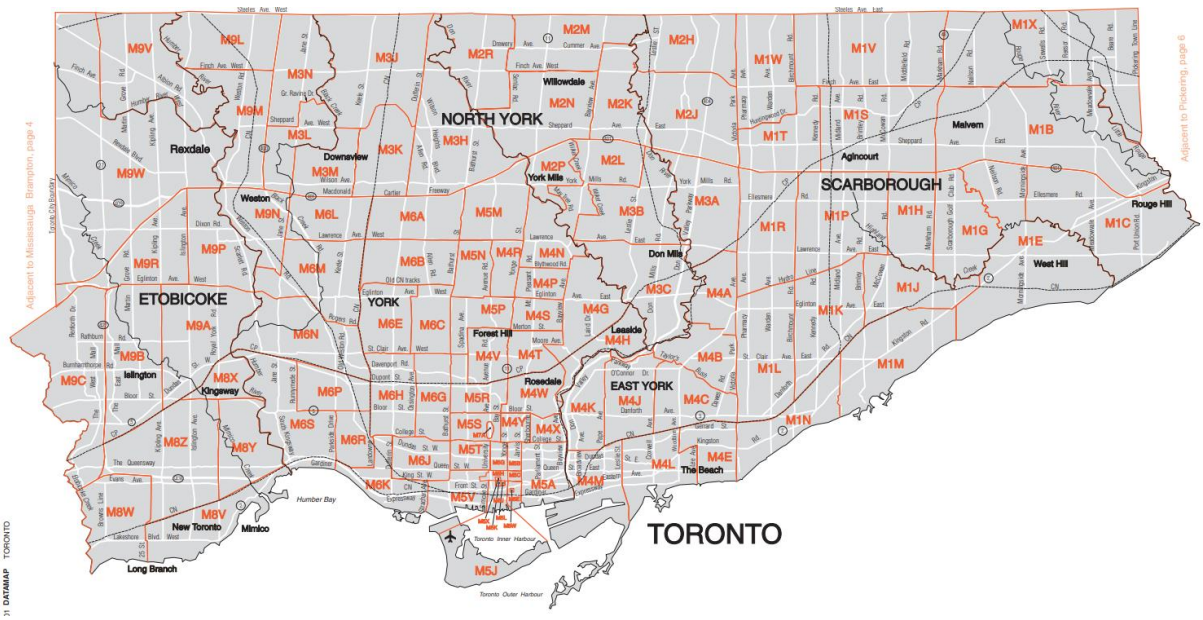
The analysis intends to be of use to people who have recently moved to Toronto for work, know their business/ job location and intend to look for suitable rental accommodation within a certain distance from their work location. The user will be able to find Toronto locations clustered on the basis of distance from work and average monthly rent.

2. Data Sources and Preparation

2.1. Sources

This problem needs to be addressed by using the Foursquare location database. An examination of the Foursquare documentation for their venues API reveals that venues near a location can be searched by name or by a category search, among other parameters. The details included in the results of an 'explore' query venue names and location details include street address, cross-street, city, state, country, postal code, latitude and longitude, and distance from search location. The results do not include neighbourhood data, so the project will henceforth use postal codes as a proxy for Toronto neighbourhoods.

Postal codes in Canada are a sequence of six alphanumeric characters, like M5V 1E5, where the first three letters denote the Forward Sortation Area (FSA) and the next three characters represent location data at a higher granularity like a postal office, street block, etc. All FSAs corresponding to the census sub-division of Toronto start with the letter M.



It's important to note that Toronto FSAs and official neighbourhoods are not the same and do not map one-on-one. Most city statistics available in the public domain are for Toronto's neighbourhoods while the location details available on Foursquare as well as OpenStreetMap (OSM) databases are aligned with postal codes. This presented a major data mapping challenge for this project.

1. A Wikipedia page (link) listing the 'M' series FSAs associated with Toronto serves as a useful starting point. An official csv download file for the FSAs is also available on this Stats Canada webpage here. I scraped the Wikipedia page into a Python dataset consisting of 103 Toronto FSAs. Next, a file mapping each of these FSA to their location coordinates (latitude and longitude) is available here.
2. The location coordinates for the user's work location can be retrieved from the Foursquare database using the 'search' query. Ubisoft Toronto, a game development firm, was selected as the location for demonstration of the project. Using the latitude and longitude of the Ubisoft Toronto venue received from Foursquare, an explore query was run to retrieve venues of the category 'residential buildings' within a 12 km radius of the office location. The postal codes of 45 of the 100 venues retrieved were missing. Distance from search location was received for all venues.
3. To retrieve the missing postal codes, the nominatim API from Python's geopy package was used. The venue name, address and cross-street data for the residential buildings with missing postal codes were fed to Nominatim and the results received provided postal codes for a further 43 of the 45 venues. Postal codes were now available for 98 residential buildings.
4. The Canada Mortgage and Housing Corporation (CMHC) maintains the average rent prices for apartments of different configurations by Toronto neighbourhood. The latest relevant data available for Toronto from October 2019 was scraped into a Python data table.

2.2. Data Cleaning and Preparation

At a broad level, the analysis of neighbourhoods requires two data tables to be joined together:

i. Table of residential properties within a 12km radius from the selected work location
ii. Table of average monthly rent prices for a two-bedroom apartment in different parts of Toronto

Table (i) consists of a maximum of 100 results returned by Foursquare location database, with each property's name, address, distance from office, location coordinates and postal code. 45 of the 100 results lacked postal code tagging, so their names and address details were again fed to the Nominatim geocoder service of OpenStreetMap location database to try to plug the gap. With results from Foursquare and OpenStreetMap combined, 98 of the 100 properties had a Toronto FSA code mapped to them. Properties with no postal code mapping were dropped. The 98 remaining properties were tagged with 26 unique FSA codes.

Before feeding data for the 45 properties without postal code tagging to Nominatim, several data cleaning steps were required. While every property had name details, data in address and cross-street fields was often inadequate or missing. The following steps were taken to address this issue:

- Scan the name, address and cross-street fields
- Remove any non-alphanumeric characters from these fields – e.g. '/', '(', '&', etc.

- For each property, the search for an FSA is to be conducted in four passes: name, address, cross-street address, and combined address and cross-street address, in that specific order
- If property name is considered inadequate to return a unique location, e.g. names like '8G' or 'X2', use the street address instead to search for the associated FSA. Thus, '8G' was substituted by the street address '8 Gladstone Avenue' in the search query. The test for adequacy for the name field was that it should not be of a length less than 4 characters
- Where a street address is missing, search by the cross-street for the associated FSA
- If the postal code returned by name is different from that returned by address, retain the former. Similarly, if the postal code returned by address is different from that returned by cross-address, retain the former.

The 43 observations for which a postal code was found were consolidated with the 55 observations for which Foursquare had returned a postal code. All postal codes were trimmed to first 3 characters to yield the underlying FSA code. The 98 observations with postal codes were then joined with the table from Wikipedia mapping each postal code to its location coordinate.

Table (ii), scraped from the CMHC website, consisted of monthly rent information aligned with Toronto's neighbourhoods (not postal codes). This table was cleaned up of extraneous columns and rows and missing values. Thus columns related to bachelor, one-bedroom, three-plus bedroom and 'total' apartments were removed. Rows with missing average monthly rent were removed. A total of 125 observations with average monthly rent were obtained after these steps.

For the required target analysis to be possible, the neighbourhoods in table (ii) had to be tagged to the closest postal code match. The OpenStreetMap location database was used to fetch the postal codes against the neighbourhoods. Before feeding the neighbourhood names into Nominatim API, some cleaning was required. CMHC had in many cases combined adjacent neighbourhoods with a '/' symbol and in other cases, several names were hyphenated. For instance, one observation had the entry 'Woodbine Corridor/Greenwood-Coxwell' as the neighbourhood. Thus, for all neighbourhood entries, names separated by '/' and/or '-'characters were extracted in up to 4 columns and next, from each of these 4 columns, non-alphanumeric characters like '-' and others were removed. This is illustrated in the following image.

	Neighborhood	Average_Rent	Neigh_1	Neigh_2	Neigh_3	Neigh_4
120	Woburn	1375	Woburn	None	None	None
121	Woodbine Corridor/Greenwood-Coxwell	1427	Woodbine Corridor	Greenwood	Coxwell	None
122	Yonge-Eglinton	2035	Yonge	Eglinton	None	None
123	Yonge-St. Clair	2058	Yonge	St Clair	None	None
124	Yorkdale-Glen Park	1426	Yorkdale	Glen Park	None	None

In the next step, postal codes were retrieved for the neighbourhood names in columns Neigh_1 to Neigh_4. Since neighbourhoods don't map exactly to postal codes, it's possible for an observation with aggregated neighbourhood names against a single monthly rent value to be mapped to several postal codes, as shown below.

	Neighborhood	Average_Rent	Neigh_1	Neigh_2	Neigh_3	Neigh_4	Postal Code1	Postal Code2	Postal Code3
4	Banbury-Don Mills/York Mills	1432	Banbury	Don Mills	York Mills	None	Not found	M2J 5A7	M2P 2E3

A total of 168 mappings of FSA code to average rent were created. 3 duplicates of neighbourhood names but with different average rent values were renamed.

In the next step, the observations were grouped by FSA codes and for each FSA code with multiple average rent values arising from being present across multiple neighbourhoods as defined in the CMHC table, the mean average value was considered as representative.

	Postal Code	Average_rent	Neighborhood
0	L1S	1237.0	Ajax
1	L1V	1237.0	Pickering
2	L3Z	1313.0	Bradford,East Gwillimbury,West Gwillimbury

As can be seen in the preceding image, the neighbourhoods Ajax and Pickering which comprise a single aggregated observation in the CMHC table have been mapped to two FSA codes with same average monthly rent. FSA code L3Z has, however, been mapped to three neighbourhood names extracted from two different observations in the CMHC table. Bradford and West Gwillimbury comprise a single CMHC record with average rent CAD 1,268 and East Gwillimbury is from a separate record with average rent CAD 1,403. The final figure of CAD 1,313 seen in the previous image is the mean of the monthly rent values of the three neighbourhoods associated with FSA L3Z

Table (i) is then left-joined with table (ii) so that each of the 98 residential properties with an FSA code is now linked to the monthly average rent associated with the concerned FSA code. Properties which do not find a corresponding average rent figure in table (ii) are dropped. The merged table is grouped by FSA code and for each FSA code the average distance from the work location in Toronto is calculated based on the individual distances of properties linked with that FSA code.

The result of the preceding data preparation steps is a table which looks as follows. We have 21 unique FSA codes with average rent values available.

	Postal Code	Neigh_Latitude	Neigh_Longitude	Number of Properties	Average_rent	DistanceFromOffice_y
0	M6P	43.661608	-79.464763	2	2005.000000	1202.000000
1	M6K	43.636847	-79.428191	9	1509.428571	3754.444444
2	M5T	43.653206	-79.400049	2	2004.000000	3837.500000

The next step is to run a clustering algorithm on the FSAs based on the average monthly rent value and average distance from office. The rent and distance values are standardized using the ‘StandardScaler’ function in the scikit-learn preprocessing package. Weights can be assigned to these two attributes for apportioning relative importance in determining an FSA’s attractiveness as a rent location.

This completes the data preparation process for this project. After a first run of the k-means clustering algorithm, and visualization using a Folium map, it was found necessary to correct some FSA codes returned for CMHC neighbourhoods from the OpenStreetMap and Foursquare crowd-sourced databases. For instance, a few properties had been tagged to M1J and M2K FSAs which are clearly outside of a 12km radius from the chosen work location. These properties were re-assigned

correct FSAs based on some manual search using Google Maps application and then the clustering algorithm and map visualization were run again.