IBM – Coursera

Data Science Specialization

Capstone project – Final report

The Battle of Neighbourhoods in Toronto: Where is/is not best to live

Yoones Vaezi

2020

Table of Contents

1. Introduction

1.1. Background

Toronto, Capital of the province of Ontario, is Canada's largest city and a world leader in areas such as business, finance, technology, entertainment and culture. Its large population of immigrants from all over the globe has also made Toronto one of the most multicultural cities in the world. With a population of ~2.8 million as of 2016, Toronto is Canada's most populous city and the fourth most populous city in North America.

According to Global Liveability Index 2019 report published by Economist Intelligence Unit, Toronto ranks 7th (tied with Tokyo, Japan) out of the 140 most livable cities in the world. With thousands of people moving to Toronto from abroad or domestically and pursuing their lives and careers in such a vibrant city, the question becomes: where do I live in Toronto? Which neighbourhoods provide best quality of life in terms of major neighbourhood quality measures? Considering a person's priorities for a neighbourhood, which neighbourhood are less or more likely to satisfy one's expectation of a neighbourhood to live? We will try to answer these questions by data science methods applied on a dataset extracted from multiple data resources and provide some more insights into neighbourhoods' comparisons. We categorize neighbourhoods into multiple categories which share similar features. These categories and features can be a great source of information for someone that has general or even specific priorities in terms of what they would expect from their neighbourhoods.

1.2. Business understanding/Problem description

As one of most liveable metropolitan cities in the world and the centre of many businesses, Toronto is where many newcomers or local people move to and would like to call home. These include people who move from abroad including immigrants or refugees, people who move to Toronto for a new career and business, etc., and also people who move domestically in Canada or even in Greater Toronto Area either in search of better neighbourhood to live in or to live closer to where they work.

A very important decision that these people need to make for themselves and their families is which neighbourhood in Toronto they should select to live. This decision becomes more difficult for many people that move to Toronto for the first time or people who do not know a lot about Toronto's neighbourhoods.

Although Toronto is one of the world's most liveable cities, there are obviously major differences between different neighbourhoods which can make one more appealing than the other to a person considering one's priorities. There are many online resources that can help compare neighbourhoods and facilitate making this decision, however, gathering such information on major decision criteria can be time consuming and there are not many resources out there that bring majority of these criteria into one place. Having a tool that can do this can be very helpful for these people. This is what we have decided to do for

our data science capstone project below: Gather data on major criteria to group neighbourhoods into different categories which can help one decide where to choose as their next neighbourhood of residence.

To do this, we have taken a step backwards and asked the question: What are the main criteria that one generally considers before selecting a neighbourhood for living? Let's say you need to move to Toronto and move to a new neighbourhood within Toronto: What do you generally expect your new neighbourhood have to consider it as a potential future neighbourhood of residence?

The answer we have given to this question is that a person generally would prefer a neighbourhood where people living there or the neighbourhood itself have:

a) the highest number, percentage or amount of:

a-1) schools per 10,000 young residents

a-2) income or salary

a-3) educated residents with post-secondary degrees

a-4) green areas or tree cover (per 10,000 residents)

a-5) walkability to amenities (walk score)

a-6) public transport (bus/street car/etc) stops (per 10,000 residents)

a-7) shops and stores (per 10,000 residents)

a-8) food and drink places (per 10,000 residents)

a-9) recreation centres (gym, sports, entertainment, touristic places, etc) (per 10,000 residents), and

b) the lowest number, percentage or amount of:

b-1) average home prices

b-2) crime rate

b-3) average rent

b-4) unemployment rate

b-5) median journey to work/commuting duration

b-6) population density

We gather data on all these pertinent criteria (features) and analyze them in an effort to be able to group neighbourhoods into different categories that share similar characteristics and find out any neighbourhoods that stand out in terms of having an anomalous number of positive (a) or negative (b) characteristics. This tool can help answer the question of which neighbourhoods are more likely for a person to select as a potential neighbourhood of residence versus the others considering one's priorities.

1.3. Target audience

The project can potentially serve two groups of audience:

a) Future residents: who need to decide which neighbourhood they want to live in, including people who need to move within Toronto because of career related relocation, or unhappiness with their current neighbourhood of residence, etc, or people who are moving to Toronto from another city or abroad to start a career or other reasons, including immigrants, refugees, business owners, etc.

b) City officials: Considering the various decision criteria (features) we are considering for decision making, officials from different and related sectors can investigate how they can improve some of these characteristics in different neighbourhoods to make them more appealing to future residents. For instance, the police can target place with higher rate of crimes, Toronto Transit Commission (TTC) can increase the number of public transit stops in neighbourhoods that need it, the city provide more social housing where housing prices are high or rents are relatively more expensive, etc.

2. Data

City of Toronto website divides Toronto into 140 neighbourhoods (https://open.toronto.ca/dataset/neighbourhoods/).

The features described in the previous section come mainly from two data sources:

1- City of Toronto Open Data Catalogue here

2- Foursquare API here

The features that we are going to use for this analysis are going to come from separate data tables from each of these resources. The data tables queried from each of these resources include:

2-1. City of Toronto Open Data Catalogue:

For this study we mainly use catalogues which are based on Canada's 2011 Census Program because majority of information we need are available.

1-1) List of 140 neighbourhoods, their unique id, latitude and longitude from Toronto Neighbourhoods Catalogue.

The table includes other information which will be dropped as they are not used for our analysis.

The table is cleaned by removing unwanted columns and also parsing it such that each row represents a neighbourhood, with its name, unique id, latitude and longitude coordinates in separate columns. A snapshot of the data is shown below:

	Neighbourhood_id	Neighbourhood	Latitude	Longitude
0	94	Wychwood	43.676919	-79.425515
1	100	Yonge-Eglinton	43.704689	-79.403590
2	97	Yonge-St.Clair	43.687859	-79.397871
3	27	York University Heights	43.765736	-79.488883
4	31	Yorkdale-Glen Park	43.714672	-79.457108

Figure 1. List of 140 Toronto neighbourhoods and their location coordinates.

1-2) Number of schools in each neighbourhood from School Locations - All Types catalogue: This table includes location (latitude/longitude) of all schools in Toronto of all types, their names, address, unique id etc. We calculate the number of schools in each neighbourhood by finding how many of them locates within the shape polygon of each neighbourhood, which are built from the neighbourhoods' GeoJson file on City of Toronto Open Data Catalogues. However, for neighbourhood comparison purposes we need a normalized version of the number of schools in each neighbourhood that are available for a fixed population (here we use 10,000) of young people that are within the age limit of 5 to 19 years old. Here, we call this feature 'school rate'. To calculate school rate, we first find the total population of people with ages between 5 and 19 by summing the population age groups of 5-to-9, 10-to-14, and 15-to-19 from Wellbeing Toronto – Demographics catalogue of population age groups. Multiplying the number of schools in each neighbourhood by 10,000 followed by a division by the total neighbourhood's 5-19 year old population provides school rate. Figure 2 shows a snapshot of the resulting school rates calculated for each neighbourhood.

	Neighbourhood	Neighbourhood_id	school_rate
0	Wychwood	94	28.653295
1	Yonge-Eglinton	100	70.175439
2	Yonge-St.Clair	97	18.867925
3	York University Heights	27	43.577982
4	Yorkdale-Glen Park	31	48.458150

Figure 2. School rate, the number of school in each neighbourhood for unit population of 10,000 people within the age range of 5 to 19 years old.

1-3) Average home prices from Wellbeing Toronto – Housing catalogue: A snapshot of the average housing price table after data manipulation and removing unwanted information looks like Figure 3.

	Neighbourhood	Neighbourhood_id	Home Prices
0	West Humber-Clairville	1	317508
1	Mount Olive-Silverstone-Jamestown	2	251119
2	Thistletown-Beaumond Heights	3	414216
3	Rexdale-Kipling	4	392271
4	Elms-Old Rexdale	5	233832

Figure 3. List of average home price per Toronto neighbourhood.

1-4) Total number of major crime incidents in each neighbourhood from Wellbeing Toronto – Safety Catalogue: This tables includes number of different types of crimes for each neighbourhood. It also has the total number of major crime incidents which is what we only use for our study. However, number of incidents is usually dependent on population. Therefore, in order to be able to have a better measure of crime for comparison purposes between neighbourhoods, we are going to define the feature 'crime rate', which according to Statistics Canada is defined as 'number of incidents reported to police per 100,000 population'. To obtain this measure, we also query total population per neighbourhood from Wellbeing Toronto - Demographics: NHS Indicators. Figure 4 shows a snapshot of the table that includes total number of crimes, total population, and the resulting crime rate per neighbourhood. Please note that only crime rate column will be kept and used as a feature for later analysis and categorization.

	Neighbourhood	Neighbourhood_id	Total crime number	Total Population	crime_rate
0	West Humber-Clairville	1	1119	34100	3281.524927
1	Mount Olive-Silverstone-Jamestown	2	690	32790	2104.300091
2	Thistletown-Beaumond Heights	3	192	10140	1893.491124
3	Rexdale-Kipling	4	164	10485	1564.139247
4	Elms-Old Rexdale	5	185	9550	1937.172775

Figure 4. Information related to number of crimes and calculated crime rate in each Toronto neighbourhood.

1-5) Median income and average rent per neighbourhood from Wellbeing Toronto - Demographics: NHS Indicators catalogue. From this table we extract and will use the median after-tax household income and average monthly shelter costs for rented dwellings in Canadian Dollars. Figure 5 shows a snapshot of the resulting table after parsing and cleaning the table and extracting the relevant information only.

	Neighbourhood_id	median_income	average_rent
0	1	59703	945
1	2	46986	921
2	3	57522	887
3	4	51194	857
4	5	49425	966

Figure 5. Median income and average rental prices per neighbourhood in Toronto.

1-6) Employment and unemployment rates, median commuting duration and population density from Toronto neighbourhood profile. These features are either extracted directly or calculated (after parsing the table and using other information) from the Toronto Neighbourhood Profiles table which can be found here. Employment and unemployment rates and median commuting (journey to work) duration come from the table directly after parsing and cleaning the table. Note that commuting duration is the median of total amount of time in minutes a person spends in a day for commuting. We also extract information on land area of each neighbourhood in square kilometers from this table. We use it for normalization purposes later on, for instance to calculate population density. We calculate population density by dividing total population of each neighbourhood by the land area of the neighbourhood in square kilometers. Figure 6 shows a snapshot of the information extracted from this analysis. Note that land area will not be used as a feature for neighbourhood comparison as it is not usually a deciding factor for a potential resident in selecting a neighbourhood to live in.

	Neighbourhood_id	Neighbourhood	Land area in square kilometres	Employment rate	Unemployment rate	Median commuting duration	Population_density
0	94	Wychwood	1.68	61.6	7.6	91.3	8324.404762
1	100	Yonge-Eglinton	1.65	68.2	5.7	60.4	6412.121212
2	97	Yonge-St.Clair	1.17	66.3	7.0	106.3	9961.538462
3	27	York University Heights	13.23	52.6	11.4	152.2	2094.860166
4	31	Yorkdale-Glen Park	6.04	53.6	10.2	91.3	2431.291391

Figure 6. Employment and unemployment rates, median commuting to work duration and population density per neighbourhood in Toronto.

1-7) Percentage of educated people from Toronto Education NHS indicator table: The table includes number of people with different levels of education per neighbourhood and also total population of people with age above 15 years old in each neighbourhood. We parse and clean this table and calculate the percentage of educated (with post-secondary degree of diploma) people in each neighbourhood by dividing the number of people with postsecondary certificate, diploma or degree by the total neighbourhood population with an age above 15 years old. Figure 7 shows a snapshot of the resulting table.

	Neighbourhood	Neighbourhood_id	post_secondary_percent
0	Agincourt North	129	47.816806
1	Agincourt South-Malvern West	128	52.137671
2	Alderwood	20	52.497551
3	Annex	95	76.335120
4	Banbury-Don Mills	42	69.306497

Figure 7. Table showing the percentage of educated people in each Toronto neighbourhood having a postsecondary degree, diploma or certificate.

1-8) Amount of tree cover from Wellbeing Toronto – Environment table: This table includes information about green spaces, air pollutants and tree cover in square meters for each neighbourhood. We are only going to use the tree cover information here. However, for neighbourhood comparison purpose, we need to be looking how much green space is available for a fixed amount of population. Therefore, we define a feature names 'Tree cover rate' which measures the amount tree cover in square kilometers per 10,000 people. To calculate this, we use total population of each neighbourhood from previous tables. Figure 8 shows a snapshot of the table that includes the tree cover rate for each neighbourhood.

	Neighbourhood_id	Neighbourhood	Tree_cover_rate
0	94	Wychwood	0.325880
1	100	Yonge-Eglinton	0.549877
2	97	Yonge-St.Clair	0.367762
3	27	York University Heights	0.745954
4	31	Yorkdale-Glen Park	0.471782

Figure 8. Table showing the tree cover rate per Toronto neighbourhood, defined as total amount of tree cover in square kilometers for population of 10000 people.

1-9) Walkability (Walk Score) from Wellbeing Toronto Civics Equity Indicators table: Walk Score measures walkability on a scale from 0 - 100 based on walking routes to destinations such as grocery stores, schools, parks, restaurants, and retail, which is an important deciding criteria to select a neighbourhood as a potential residence. This table includes other information, however, we only extract walk score for our study. Figure 9 shows a snapshot of the Walk Scores for each neighbourhood.

	Neighbourhood_id	Neighbourhood	Walk Score
0	94	Wychwood	86
1	100	Yonge-Eglinton	89
2	97	Yonge-St.Clair	84
3	27	York University Heights	60
4	31	Yorkdale-Glen Park	72

Figure 9. Table showing Walk Score for Toronto neighbourhoods.

1-10) Toronto Transit Commission (TTC) stops from Wellbeing Toronto – Transportation table: Availability of public transit is also an important criteria for a neighbourhood. Many people prefer to use public transport to travel in the city and for commuting to and from work. We use the information on city of Toronto's transportation catalogue to extract the number of TTC stops (includes all bus, streetcar and non-subway stops) for each neighbourhood. This table includes other information such as number of traffic and pedestrian collisions, road kilometers, and road volume which are not critical and not used for our study. Once again, we use a normalized version of the TTC stops for our study because we need to know the amount of public transit available to a specific amount of population to be able to make a fair comparison between neighbourhoods. We again calculate a new feature named 'TTC stops rate' which is the number of TTC stops per 10,000 people (we use neighbourhoods' total populations to calculate this). Figure 10 shows an example of such information.

	Neighbourhood	Neighbourhood_id	TTC_stops_rate
0	Wychwood	94	44.333214
1	Yonge-Eglinton	100	63.327032
2	Yonge-St.Clair	97	24.024024
3	York University Heights	27	84.791629
4	Yorkdale-Glen Park	31	105.549881

Figure 10. Table showing public transport availability (TTC stop score) for Toronto neighbourhoods, calculated as the number of TTC stops per 10,000 residents.

2.2 Foursquare API

We use Foursquare API to query venues within 500 meters of each neighbourhood coordinates. All venue categories are inspected and we decided to divide venues into three main groups:

2-a) Food and drink: This group include venues that has to do with either food or drinks and that have categories which contain either of the following keywords:

Burger,Restaurant,Breakfast,Coffee,Bakery,Pizza,Buffet,Sandwich,Salad,Poutine,Bagel,Tea,Café,Pub, Chicken, Bar, BBQ, Ice Cream, Diner, Yogurt, Steakhouse, Chips, Brewery, Wings, Beer, Food, Taco, Cheese, Pie, Donut, Noodle House, Snack, Burrito, Pastry

2-b) Recreation: This group includes sport facilities, entertainment and places to visit for recreation and fun. Venue category keywords that are used here include the following, excluding the venues that match the Food and Drink group: Gym, Rink, Yoga, Bowling, Pool, Playground, Trail, Racetrack, Hockey, Rock Climbing, Tennis, Baseball, Soccer, Curling, Basketball, Stadium, Field, Athletics, Zoo, Beach, Museum, Entertainment, Garden, Theater.

2-c) Shops and stores: This group includes venues that are to do with shopping and different types of stores providing services and goods and include venues that have either of the following keywords in their venues category and are not listed in the previous two Food and Drink and Recreation categories: Market, Store, Shop, Supermarket, Tattoo, Nail, Shoe, Grocery.

The total number of venues in each category are counted for each neighbourhood and saved into a table. Once again, what we need for relative neighbourhood comparisons, is a version of the number of venues normalized by total population. So we have defined new features 'food_drink_rate', 'recreation_rate' and 'shop_store_rate', which, for each neighbourhood, are the number of corresponding venues in each category per 10,000 residents. Figure 11 shows an example of the resulting table of these features.

	Neighbourhood	Neighbourhood_id	shop_store_rate	food_drink_rate	recreation_rate
0	Agincourt North	129	1.651255	3.632761	0.000000
1	Agincourt South-Malvern West	128	0.909504	6.366530	0.454752
2	Alderwood	20	0.840336	2.521008	0.000000
3	Annex	95	1.028101	5.140507	0.000000
4	Banbury-Don Mills	42	4.087700	2.972873	0.371609

Figure 11. Table showing food and drink, shop and store, and recreation availability for Toronto neighbourhoods, calculated as the number of venues per each category per 10,000 residents.

2.3 Combined data

The data and the features queried or calculated in the two previous sections are combined into a final table that will be analyzed and used for subsequent study of Toronto neighbourhoods in this project. Figure 11 shows a snapshot of this table including the neighbourhood names and their unique id, latitude and longitudes and theirs corresponding feature (attribute) values.

	Neighbourhood	Neighbourhood_id	Latitude	Longitude	school_rate	Home Prices	crime_rate	median_income	average_rent
0	Wychwood	94	43.676919	-79.425515	28.653295	656868	1573.114051	50261	930
1	Yonge-Eglinton	100	43.704689	-79.403590	70.175439	975449	2164.461248	63267	1246
2	Yonge-St.Clair	97	43.687859	-79.397871	18.867925	995616	952.380952	58838	1314
3	York University Heights	27	43.765736	-79.488883	43.577982	359372	2799.927837	42916	911
4	Yorkdale-Glen Park	31	43.714672	-79.457108	48.458150	421045	3752.128022	49803	916

Employment rate	Unemployment rate	Median commuting duration	Population_density	post_secondary_percent	Tree_cover_rate
61.6	7.6	91.3	8324.404762	61.343764	0.325880
68.2	5.7	60.4	6412.121212	78.147532	0.549877
66.3	7.0	106.3	9961.538462	84.869976	0.367762
52.6	11.4	152.2	2094.860166	47.081967	0.745954
53.6	10.2	91.3	2431.291391	41.752577	0.471782

Walk Score	TTC_stops_rate	shop_store_rate	food_drink_rate	recreation_rate
86	44.333214	0.715052	0.000000	0.000000
89	63.327032	4.725898	21.739130	4.725898
84	24.024024	3.432003	35.178035	2.574003
60	84.791629	0.000000	2.164893	0.000000
72	105.549881	2.042901	8.852571	1.361934

Figure 11. The final list of neighbourhoods and their features and their corresponding values.