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 12 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 12. The final list of neighbourhoods and their features and their corresponding values.

3. Methodology

We can divide this section into two parts. In the first part we will be looking into the input data and features we have prepared in the previous section and try to get more insight about the data to answer a few questions such as: what is the distribution of each feature among all neighbourhood like? Are there any outliers standing out from the rest of neighbourhoods in each category? How correlated are different features? We obviously need to reduce the number of linearly dependent features which may skew our neighbourhood clustering results.

In the second part, we discuss the clustering technique (K-means) that we use to categorize the neighbourhoods based on the features we ended up keeping and the reasoning behind choosing its parameters.

3.1 Exploratory data analysis

In the first step we plot the feature correlation heat map to identify the highly correlated features. We need to reduce the number of linearly correlated features that can skew the clustering results. We calculate Pearson correlation values and plot them in form of a heat map in Figure 13. The positive and negative correlations are shown in blue and red, respectively.

From this plot, the largest (absolute) correlation is between employment and unemployment rates. The correlation is -0.75, which means they have a negative linear correlation. The larger the employment rate, the lower the unemployment rate. This is actually trivial. One would expect to see this correlation between the two. Therefore, we only keep one of them (unemployment rate) for the rest of our study. The correlation between the two can be seen below in the regression plot of the two features in Figure 14.

This is the only column we are going to remove from the feature space. One may argue that there are other features that also seem to show high correlation and may need to be removed. But these correlations are not trivial and each of them seem to be important on its own and not necessarily controlling the other parameter directly or a reasonable causal relationship cannot strictly be drawn initially. For instance, we see a correlation of 0.73 between median income and tree cover rate. However, in reality, these two features are not dependent on each other. So we keep both.

In Figure 15, we show some quick descriptive statistics of the features, which are created using the *describe* method in *Pandas*. What stands out from this table is that there are neighbourhoods with zero school rate (no schools at all), and zero *shops_store*, *recreation*, and *food_drink* rates. This can really affect the decision of potential residents negatively.

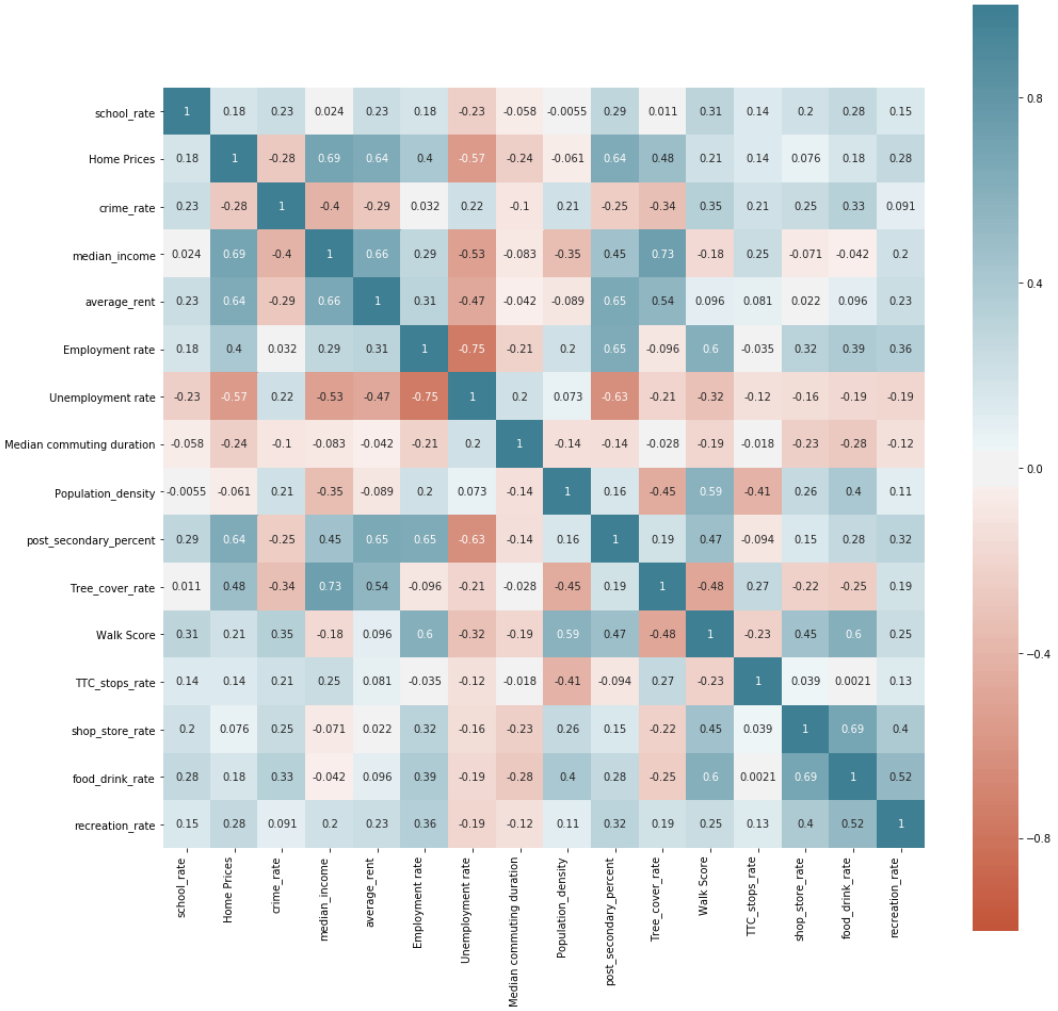

Figure 13. Heat map plot of feature correlation values. Blue and res show high and low correlation respectively.

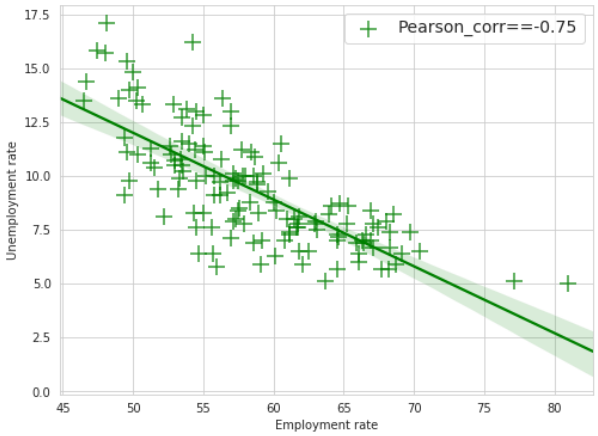

Figure 14. The regression plot showing a clear negative correlation between employment and unemployment rates. The Pearson correlation is calculated to be -0.75, as shown in Figure 13.

	school_rate	Home Prices	crime_rate	median_income	average_rent	Unemployment rate	Median commuting duration	Population_density
count	140.000000	1.400000e+02	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000
mean	34.468283	5.481934e+05	1889.992578	55426.500000	1019.792857	9.370714	115.929286	5984.749355
std	20.671702	2.676674e+05	865.568201	16118.155356	219.621994	2.622166	53.566310	4532.568101
min	0.000000	2.041040e+05	709.272257	30794.000000	631.000000	5.000000	51.600000	978.114478
25%	21.414998	3.749645e+05	1240.737724	46689.500000	878.500000	7.400000	76.825000	3513.160699
50%	29.925373	4.912100e+05	1722.522762	52660.000000	972.500000	8.950000	97.150000	5057.701699
75%	41.301279	5.902160e+05	2299.049581	59963.000000	1124.750000	11.000000	147.900000	7267.396825
max	135.338346	1.849084e+06	5646.842428	161448.000000	2388.000000	17.100000	314.200000	42440.476190

post_secondary_percent	Tree_cover_rate	Walk Score	TTC_stops_rate	shop_store_rate	food_drink_rate	recreation_rate
140.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000
58.428147	0.716240	72.271429	38.586365	1.482318	5.385263	0.689748
12.258446	0.666828	12.790421	17.101158	2.610632	8.212340	1.069425
29.810855	0.034567	42.000000	10.659187	0.000000	0.000000	0.000000
49.853099	0.339678	62.000000	26.596653	0.000000	0.355772	0.000000
57.860139	0.550217	70.500000	36.565185	0.779270	2.129567	0.372825
68.160806	0.901491	83.000000	45.232003	1.823236	5.820956	0.948159
84.869976	6.321555	99.000000	105.549881	20.960699	44.415415	6.099553

Figure 15. Descriptive statistics of the feature data.

In order to make better sense of the input data, we also visualize each of the features in the form of horizontal bar plots. This way, we investigate each feature separately across all neighbourhoods and better spot any neighbourhoods that show anomalously high or low value for that feature.

An example bar plot (school rate) is shown in Figure 16. For the purpose of brevity we only show a single plot. Plots for the rest of the 14 features are included in the project's Notebook. In all plots, the variables are first sorted in a descending order. Therefore, all plots show the maximum length of the bars at the top and minimum length at the bottom of the figure.

We can make the following observations from investigating each of feature's bar plots:

1) **School rate**: The *University* and *Kensington-Chinatown* neighbourhoods have anomalously high school rates, while, *Briar Hill-Belgravia* has zero schools.

2) **Home prices**: *Bridle Path-Sunnybrook-York Mills* neighbourhood seem to have anomalously high home prices, and *Flemingdon Park* neighbourhood has the lowest home prices.

3) **Crime rate**: The three neighbourhood *Moss Park*, *Bay Street Corridor*, and *Kensington-Chinatown* show anomalously high crime rates. On the other hand, *Bayview Woods-Steeles* shows the lowest crime rate.

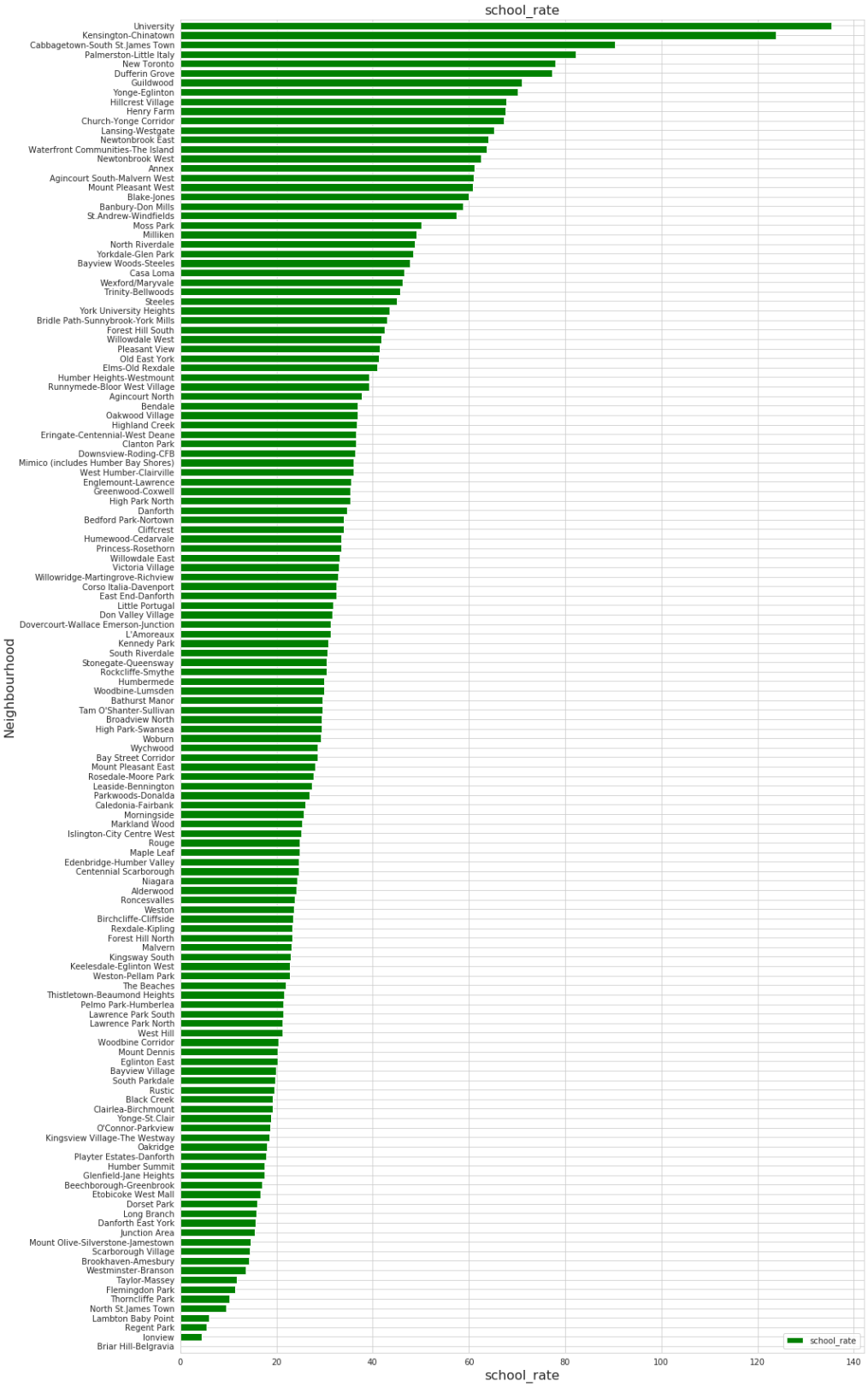

Figure 16. Bar plot showing school rate in each of 140 neighbourhoods in Toronto.

4) **Median income**: The *Bridle Path-Sunnybrook-York Mills* neighbourhoods clearly stands out with its high median income. The Four neighbourhoods *South Parkdale*, *Oakridge*, *North St.James Town*, and *Regent Park*, have the lowest median income.

5) **Average rent**: As in median income, the *Bridle Path-Sunnybrook-York Mills* neighbourhood clearly stands out with its high rents. The neighbourhoods *West Hill*, *Hillcrest Village*, and *Regent Park* have the lowest average rent and are more favorable neighbourhoods for renter, at least in terms of the rental costs only.

6) **Unemployment rate**: The *Oakridge* neighbourhood has the highest unemployment rate, while neighbourhoods *Lambton Baby Point*, *Waterfront Communities-The Island*, and *Niagara* have the lowest unemployment rate.

7) **Commuting duration**: The neighbourhoods *Woburn* and *Rouge* have the longest commuting to work duration, while *Markland Wood* has the lowest.

8) **Population density**: *North St.James Town* neighbourhood has anomalously the highest population density while neighbourhoods *Rouge*, *West Humber-Clairville* and *Bridle Path-Sunnybrook-York Mills* have the lowest population density.

9) **Educated population**: The neighbourhoods *Yonge-St.Clair* and *Waterfront Communities-The Island* have the highest proportion of educated people with postsecondary education, while *Glenfield-Jane Heights* shows the lowest.

10) **Tree cover**: *Bridle Path-Sunnybrook-York Mills* is the greenest neighbourhood in Toronto while *Church-Yonge Corridor* and *North St.James Town* have the lowest tree cover rates.

11) **Walk score**: There are not any neighbourhoods standing out with anomalous high walk scores compared to others. However, *Bay Street Corridor* has the highest walk score. The neighbourhood *Rouge* on the other hand, has the lowest walkability.

12) **Public transit**: The neighbourhoods *Yorkdale-Glen Park* and *West Humber-Clairville* have the highest public transit rates, while *North St.James Town* is the least favorable in terms of public transit score.

13) **Shops and stores**: The neighbourhoods *Dufferin Grove* and *Junction Area* have the highest shop and store rates. There are several neighbourhoods that show very low to no shops and stores within 500 m of their coordinates.

14) **Food and drink**: The *Playter Estates-Danforth* neighbourhood has the highest food and drink score. There are many neighbourhoods that show very low to no food and drink places within 500 m of their coordinates.

15) **Recreation**: The neighbourhoods *Rouge*, *University*, and *Yonge-Eglinton* have the highest recreation scores. There are many neighbourhoods that show very low to no places for recreation purposes within 500 m of their coordinates.

Box plots are also very useful in visualizing data and identifying ranges and extremes and outliers in the data. To plot all the features in one figure, we first normalize each feature with a min-max technique such that minimum and maximum value of each feature will be 0 and 1, respectively. We then visualize all features in the form of box plots together in Figure 17.

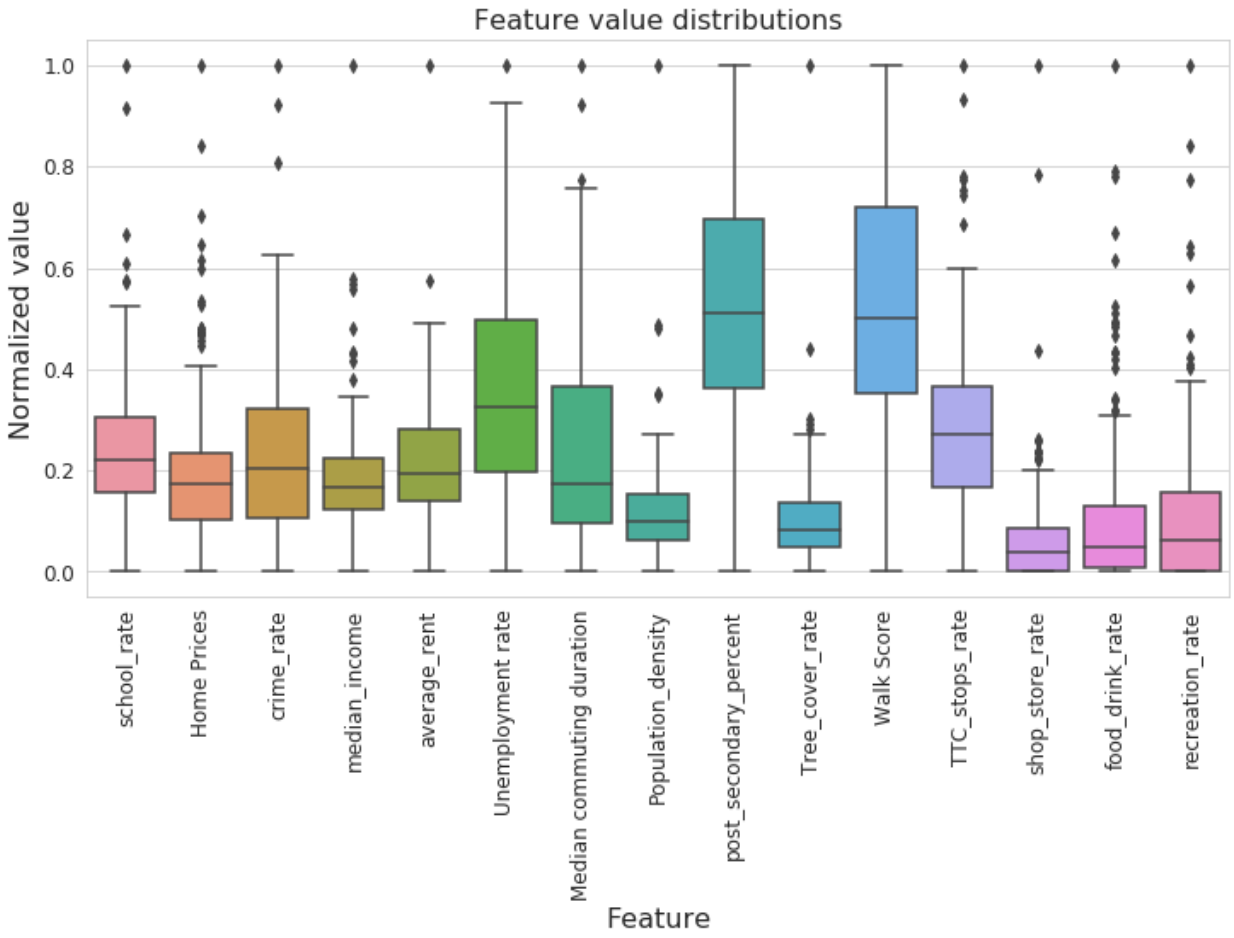

Figure 17. Box plot of all 15 feature after being min-max normalized.

The black segments inside the boxes show the median normalized value for each feature, and the boxes show the interquartile ranges (IQR), the range between 25% (Q1) and 75% (Q3) quartiles. The two caps at the end of the extending lines show the lower and higher extremes of the feature data, which are the smallest and largest data values the lie within [Q1-1.5*IQR, Q3+1.5*IQR], respectively.

Based on Figure 17, none of the features seem to have anomalously low values that can be seen as an outlier. On the other hand, the features *Home prices*, *median income, food_and_drink_rate*, and *recreation_rate* show the largest number of large outliers that lie outside their corresponding high extremes. The features *walk Score*, *post_secondary_percent*, *Unemployment_rate*, *average_rent* and *crime_rate* show no to very low number of outliers.

Majority of neighbourhoods show very low *food_drink_rate*, *recreation_rate* and shop_store_*rate* when compared with their maxima. So are *Tree_cover_rate* and *Population_density*. Majority of neighbourhoods seem to have average walk score and proportion of educated people with postsecondary educations.

3.2. Machine learning (clustering)

The purpose of this study is to categorize the 140 Toronto neighbourhoods into different groups where the neighbourhoods within each group are similar to one another in terms of at least some features, while they are dissimilar to other neighbourhoods in other groups. This means we need to use a machine learning clustering algorithm. Potential residents can check these clusters out and see what sort of features are similar between neighbourhoods in every group resulting from clustering, and see which clusters provide the largest and lowest values for the features that are their most important priorities for choosing a neighbourhood to live in.

There are several clustering algorithms available. However, for this study, we have selected to use K-means technique because it is simple and fast. K-means divides the data into non-overlapping subsets without any cluster internal structures. Examples within a cluster are very similar while examples across different clusters are very different.

K-means is an iterative clustering algorithm in which, first, distance matrix between data points to K (number of clusters) randomly selected points (centroids of clusters) is calculated. Data points are assigned to a cluster depending on their distance to the centroids (initially the randomly chosen points) of the clusters. In the next steps the centroids of each cluster are updated to the mean of the points belonging to them. The distances are calculated again, followed by cluster assignments and cluster centroid determination. This process is repeated until the centroids no longer move.

As K-means is a heuristic algorithm, there is no guarantee that it will converge to the global optimum and the result may depend on the initial clusters. It means, this algorithm is guaranteed to converge to a result, but the result may be a local optimum i.e. not necessarily the best possible outcome. To solve this problem, it is common to run the whole process multiple times with different starting conditions. This means with randomized starting centroids, it may give a better outcome. As the algorithm is usually very fast, it wouldn't be any problem to run it multiple times. We use K-means module of Python SKLearn library, which has a default number of 12 runs with random initializations.

To prepare the data for clustering, we first standardize the features by applying a Z-score method, which subtracts the feature mean values from the features and divides the result by standard deviation of the feature values. Data normalization or standardization is required so that all features end up varying in a similar range. Otherwise, extreme features can dominate and skew the modeling.

In K-means technique, we need to first specify the number of clusters, K. The results are very dependent on the number of clusters. One of the very common ways to select K in the K-means method is the Elbow technique. In this technique, K-means algorithm is performed with different values of K. For each K, the inertia, which is the within-cluster sum of squared errors, is calculated. The inertia is plotted against the K values. The rate of the change of inertia on the plot is investigated. The K value associated with the point where this rate of distortion shifts from larger to small values, after which smaller changes in inertia are observed, is selected. We performed the Elbow method and show the results in Figure 18. We find that a K-value of 6 seems like a good choice representing the elbow of the curve.

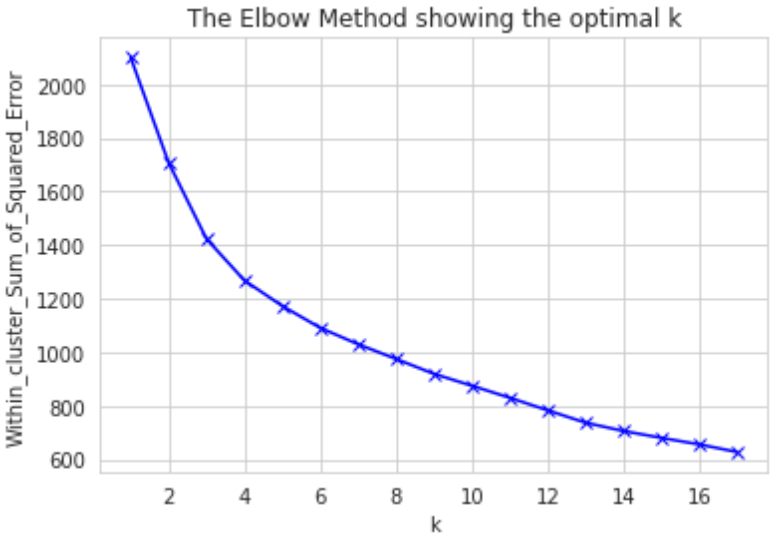

Figure 18. Elbow plot suggesting a K-value of 6 seems like a good choice for the number of clusters to be used in the K-means clustering algorithm.

Using a K-value of 6, we apply K-means algorithm to the standardized feature data table and add the resulting cluster labels to the table, as shown in Figure 19.

	Cluster Labels	Neighbourhood_id	Neighbourhood	Latitude	Longitude	school_rate	Home Prices	crime_rate	median_income	average_rent	Unemployment rate
0	0	94	Wychwood	43.676919	-79.425515	0.211716	0.275240	0.174953	0.148997	0.170176	0.214876
1	4	100	Yonge-Eglinton	43.704689	-79.403590	0.518519	0.468908	0.294718	0.248542	0.350028	0.057851
2	4	97	Yonge-St.Clair	43.687859	-79.397871	0.139413	0.481168	0.049237	0.214643	0.388731	0.165289
3	1	27	York University Heights	43.765736	-79.488883	0.321993	0.094389	0.423418	0.092779	0.159363	0.528926
4	1	31	Yorkdale-Glen Park	43.714672	-79.457108	0.358052	0.131881	0.616266	0.145491	0.162208	0.429752

Median commuting duration	Population_density	post_secondary_percent	Tree_cover_rate	Walk Score	TTC_stops_rate	shop_store_rate	food_drink_rate	recreation_rate
0.151181	0.177180	0.572710	0.046336	0.771930	0.354872	0.034114	0.000000	0.000000
0.033511	0.131059	0.877905	0.081965	0.824561	0.555037	0.225465	0.489450	0.774794
0.208302	0.216665	1.000000	0.052997	0.736842	0.140845	0.163735	0.792023	0.421999
0.383092	0.026934	0.313683	0.113152	0.315789	0.781240	0.000000	0.048742	0.000000
0.151181	0.035048	0.216889	0.069543	0.526316	1.000000	0.097463	0.199313	0.223284

Figure 19. The standardized feature table along with the calculated cluster labels from K-means clustering algorithm, and the neighbourhood names, unique IDs, and coordinates.

4. Results

In this section we visualize and discuss the neighbourhood clusters resulted from K-means algorithm in the previous section.

In Figure 20, we first plot all Toronto neighbourhoods again, but this time each neighbourhood is plotted with a color representing the cluster it belongs to. The popups on the resulting folium map show the cluster number in addition to the neighbourhood name. The legend in the figure show the color associated with each cluster.

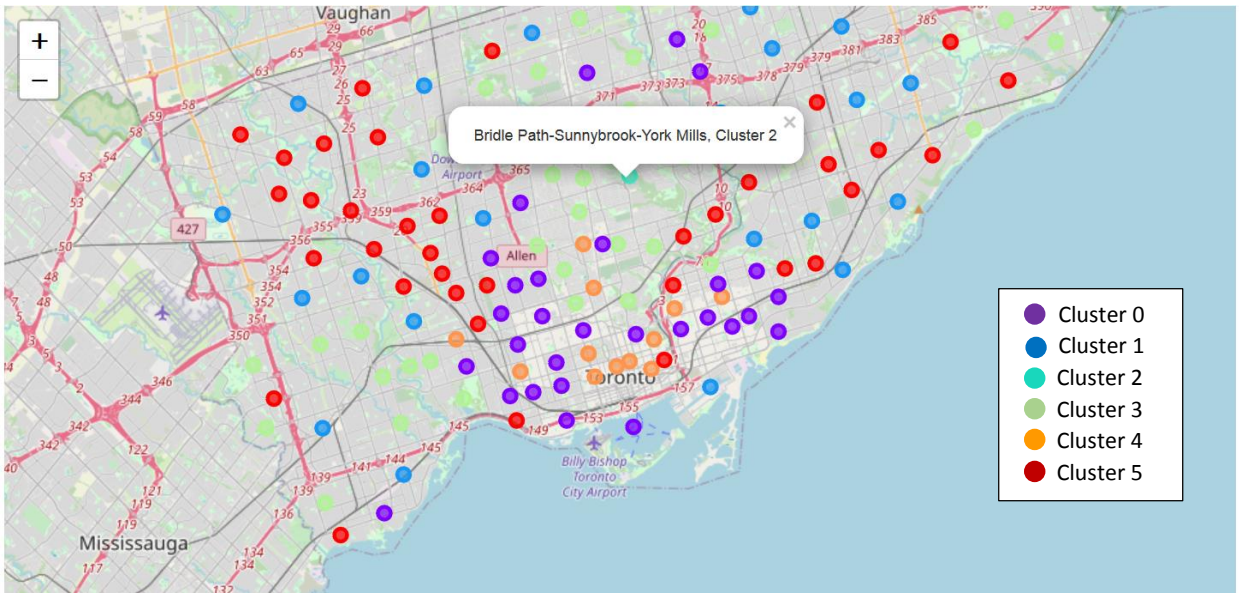

Figure 20. Neighbourhood distribution map where colors represent the clusters they belong to, calculated using K-means technique.

Interestingly, cluster 2 includes only one neighbourhood: *Bridle Path-Sunnybrook-York Mills*, also highlighted in the figure. We already saw in the exploratory data analysis section that this specific neighbourhood has anomalously high or low values in some features, which make it stand out as a separate cluster on its own.

The table in Figure 21 shows how many neighbourhood belong to each cluster. Cluster 5 includes the largest number of neighbourhoods followed by cluster 3 (33 neighbourhoods), cluster 1 and cluster 0 (both 29 neighbourhoods), cluster 4 (12 neighbourhoods), and cluster 2 (only one neighbourhood).

Cluster Labels	#Neighbourhoods
0	29
1	29
2	1
3	33
4	12
5	36

Figure 21. Number of neighbourhoods belonging to each cluster determined by K-means technique.

Figure 22 shows a pie chart of the percentage of neighbourhoods included in each cluster. The cluster number is shown on the outer bound of each wedge. The percentage is inside each wedge. The colors also match the color of clusters on the map in Figure 20.

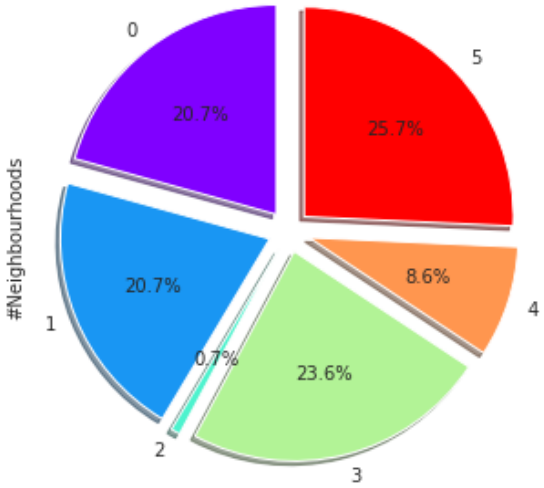

Figure 22. Pie chart of the percentage of neighbourhoods included in each cluster.

To be able to compare different clusters and see what separates one from the other we look into the features of each cluster, their median, and distribution plots.

We first group the normalized feature table shown in Figure 19 by cluster label and take the median of each feature belonging to each cluster. We investigate which clusters have the highest and lowest median value for each feature and show the results in Figure 23.

	maximum cluster	minimum cluster
school_rate	4	5
Home Prices	2	5
crime_rate	4	3
median_income	2	5
average_rent	2	5
Unemployment rate	5	2
Median commuting duration	1	2
Population_density	4	2
post_secondary_percent	2	5
Tree_cover_rate	2	4
Walk Score	4	2
TTC_stops_rate	2	0
shop_store_rate	4	2
food_drink_rate	4	3
recreation_rate	2	1

Figure 23. Table showing which cluster have the highest and lowest median feature values.

Based on Figure 23, on average, the following is found about each feature:

1) **School rate**: Cluster 4 has the highest and cluster 5 has the lowest median values
2) **Home prices**: Cluster 2 has the highest and cluster 5 has the lowest median values
3) **Crime rate**: Cluster 4 has the highest and cluster 3 has the lowest median values
4) **Median income**: Cluster 2 has the highest and cluster 5 has the lowest median values
5) **Average rent**: Cluster 2 has the highest and cluster 5 has the lowest median values
6) **Unemployment rate**: Cluster 5 has the highest and cluster 2 has the lowest median values
7) **Median commuting duration**: Cluster 1 has the highest and cluster 2 has the lowest median values
8) **Population density**: Cluster 4 has the highest and cluster 2 has the lowest median values
9) **Educated population percentage**: Cluster 2 has the highest and cluster 5 has the lowest median values
10) **Tree cover rate**: Cluster 2 has the highest and cluster 4 has the lowest median values
11) **Walk score**: Cluster 4 has the highest and cluster 2 has the lowest median values
12) **Public transit rate**: Cluster 2 has the highest and cluster 0 has the lowest median values
13) **Shop and store availability rate**: Cluster 4 has the highest and cluster 2 has the lowest median values
14) **Food and drink availability rate**: Cluster 4 has the highest and cluster 3 has the lowest median values
15) **Recreation availability rate**: Cluster 2 has the highest and cluster 1 has the lowest median values

In Figure 24, we compare the median of each normalized feature in each cluster against the others with bar plots. Since we are using the min-max normalized features, we are not looking at the actual values of the features. What is important here is that we are looking for relative differences not exact values, which is attained by this plot. Also, using normalized values we can use one single y-axis for all bar plots which makes the comparison much easier. The within-cluster feature median values are plotted as bars, where the height and the number above each show the median value of the normalized feature in that cluster. This plot nicely shows the comparison of the clusters and their feature medians.

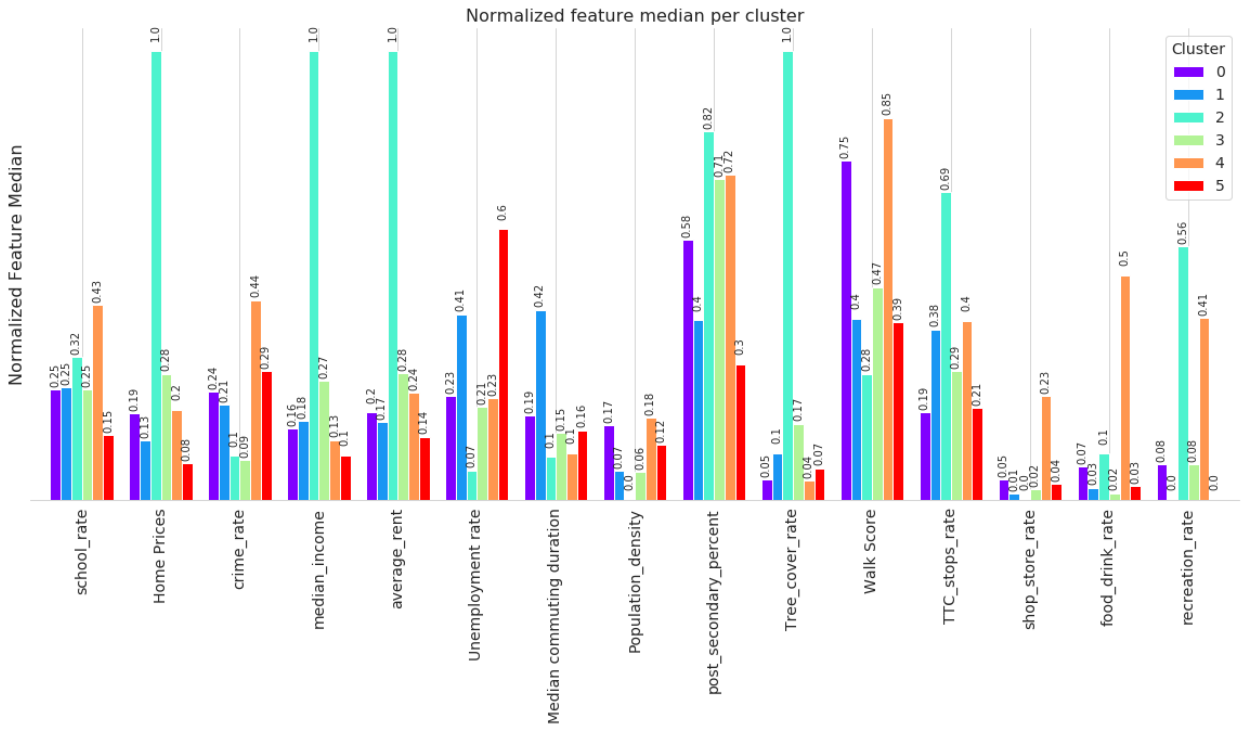

Figure 24. The bar plot of the normalized median feature values in each neighbourhood cluster.

In Figure 24, we compared the clusters based on the median of their normalized features. However, instead of a single number, we can also compare the distributions of the normalized features and compare them across different clusters. To do that, we use density distribution plots shown in Figure 25. Continuous density plots are a very useful tool to compare the distribution of different sets from same parameter that have different sizes. Since cluster 2 has only one neighbourhood, it is plotted as a single dashed line at a horizontal-axis value equal to the feature median. The distributions clearly show differences between clusters. The level of separation between clusters is different for different features. For instance, cluster 4 is easily separable from the rest of clusters in the density plots of walk score, shops and store, recreation, and food and drinks, as it seems to be having higher values in these features. This can also be observed in the bar plot of the median feature values above.

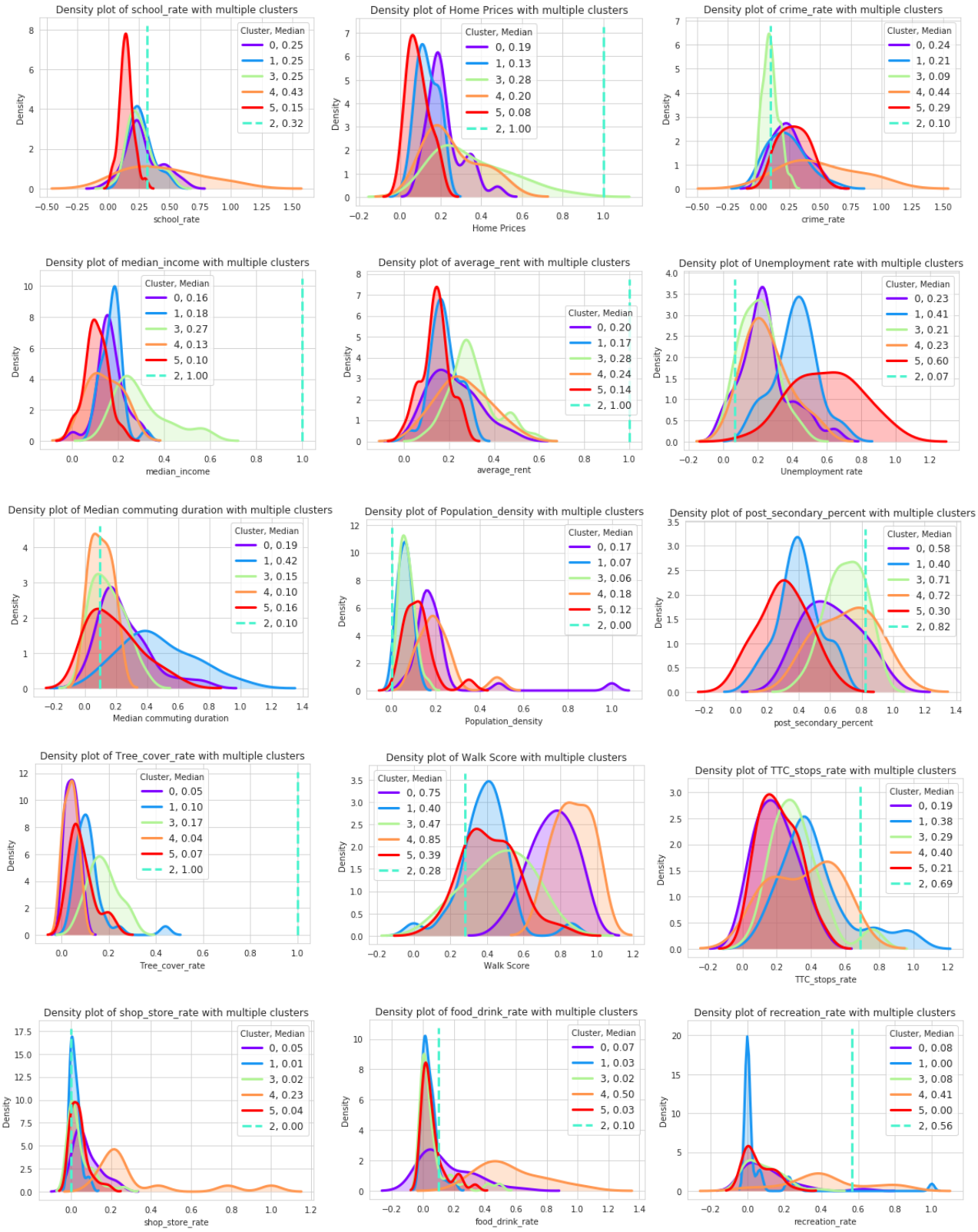

Figure 25. The density distribution plot of different normalized features. Different colors represent different neighbourhood clusters. The feature shown in each plot is mentioned in their horizontal axis label and title.

We also use the definitions of inter-quartile range used in box plots to identify number of outliers in both positive and negative indicators. To do this, we first divide the feature categories into two groups, positive and negative indicators. The positive indicators are those for which high values are favored by potential residents. On the other hand, residents prefer low values for negative indicators. These categories are

a) Positive indicators:

- school_rate
- median_income
- post_secondary_percent
- Tree_cover_rate
- Walk Score
- TTC_stops_rate
- shop_store_rate
- food_drink_rate
- recreation_rate

And,

b) Negative indicators:

- Home Prices
- crime_rate
- Unemployment rate
- Median commuting duration
- Population_density
- average_rent

For each feature, if a neighbourhood has a feature value larger than the extreme value (larger than Q3 by more than 1.5 times the feature's IQR across all neighbourhoods), that neighbourhood is considered an outlier. We then count the number of outliers for both positive and negative indicators and sum them for each neighbourhood. We can then plot the neighbourhoods that have any outliers, both positive and negative, and see which clusters have the highest number of neighbourhoods in both positive and negative indicator groups.

First we plot the neighbourhoods that do not seem to have any outlier features, either positive or negative in Figure 26. For each feature, these neighbourhoods are within the extreme limits of the box plots. There seems to be 90 out of 140 neighbourhoods which meet this condition. In this figure, the colors represent the cluster number each neighbourhood belongs to. There seems to be gaps in the center, central north, and western part of the area, meaning those areas are showing more outliers, which we will investigate in the next plots.

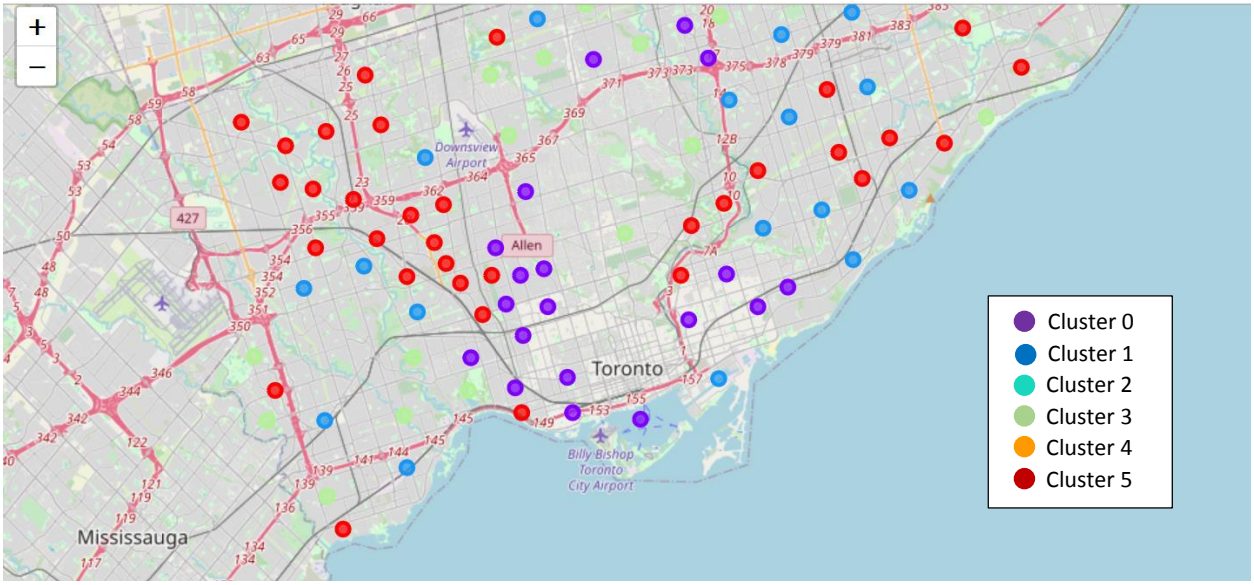

Figure 26. The map distribution of neighbourhoods without any positive or negative feature outliers.

Figure 27 shows the first 10 neighbourhoods with highest number of positive feature outliers. There are two neighbourhoods, *Cabbagetown-South St.James Town* and *Bridle Path-Sunnybrook-York Mills*, which have the largest number of positive indicators, which is 4.

	Cluster Labels	Neighbourhood_id	Neighbourhood	#P_outliers
93	4	71	Cabbagetown-South St.James Town	4.0
90	2	41	Bridle Path-Sunnybrook-York Mills	4.0
136	4	78	Kensington-Chinatown	3.0
108	4	83	Dufferin Grove	3.0
1	4	100	Yonge-Eglinton	3.0
59	4	79	University	3.0
7	3	105	Lawrence Park North	3.0
36	4	67	Playter Estates-Danforth	3.0
133	4	90	Junction Area	3.0
120	0	65	Greenwood-Coxwell	2.0

Figure 27. First 10 neighbourhoods with the highest number of positive feature outliers (column #P_outliers).

Figure 28 shows the first 10 neighbourhoods with lowest number of positive feature outliers. As expected they are all have zero outliers.

	Cluster Labels	Neighbourhood_id	Neighbourhood	#P_outliers
0	0	94	Wychwood	0.0
98	1	120	Clairlea-Birchmount	0.0
95	3	96	Casa Loma	0.0
94	5	109	Caledonia-Fairbank	0.0
92	5	30	Brookhaven-Amesbury	0.0
91	5	57	Broadview North	0.0
89	0	108	Briar Hill-Belgravia	0.0
87	5	24	Black Creek	0.0
86	1	122	Birchcliffe-Cliffside	0.0
85	1	127	Bendale	0.0

Figure 28. First 10 neighbourhoods with the lowest number of positive feature outliers.

Figure 29 show the first 10 neighbourhoods with highest number of negative feature outliers. There are two neighbourhoods, *Kingsway South* and *Bridle Path-Sunnybrook-York Mills,* which have the largest number of negative indicators, which is 2.

	Cluster Labels	Neighbourhood_id	Neighbourhood	#N_outliers
138	3	15	Kingsway South	2.0
90	2	41	Bridle Path-Sunnybrook-York Mills	2.0
39	5	72	Regent Park	1.0
50	3	40	St.Andrew-Windfields	1.0
30	5	121	Oakridge	1.0
118	3	101	Forest Hill South	1.0
117	3	102	Forest Hill North	1.0
38	3	10	Princess-Rosethorn	1.0
43	3	98	Rosedale-Moore Park	1.0
44	1	131	Rouge	1.0

Figure 29. First 10 neighbourhoods with the highest number of positive feature outliers (column #N_outliers).

Figure 30 shows the first 10 neighbourhoods with lowest number of negative feature outliers. As expected they are all have zero outliers. Interestingly, the neighbourhood *Cabbagetown-South St.James Town* which had one of the highest number of outlier positive indicators, has zero outlier negative indicators, which makes it a good neighbourhood, at least in terms of number of outlier features.

	Cluster Labels	Neighbourhood_id	Neighbourhood	#N_outliers
0	0	94	Wychwood	0.0
101	0	92	Corso Italia-Davenport	0.0
100	1	123	Cliffcrest	0.0
99	3	33	Clanton Park	0.0
98	1	120	Clairlea-Birchmount	0.0
96	3	133	Centennial Scarborough	0.0
94	5	109	Caledonia-Fairbank	0.0
93	4	71	Cabbagetown-South St.James Town	0.0
92	5	30	Brookhaven-Amesbury	0.0
91	5	57	Broadview North	0.0

Figure 30. First 10 neighbourhoods with the lowest number of negative feature outliers.

Figure 31 shows the neighbourhoods that have at least one positive indicator outlier. The size of each circle represents the number of outliers for that neighbourhood and the color represents the cluster it belongs to. The popups show these and also the list of outlier features.

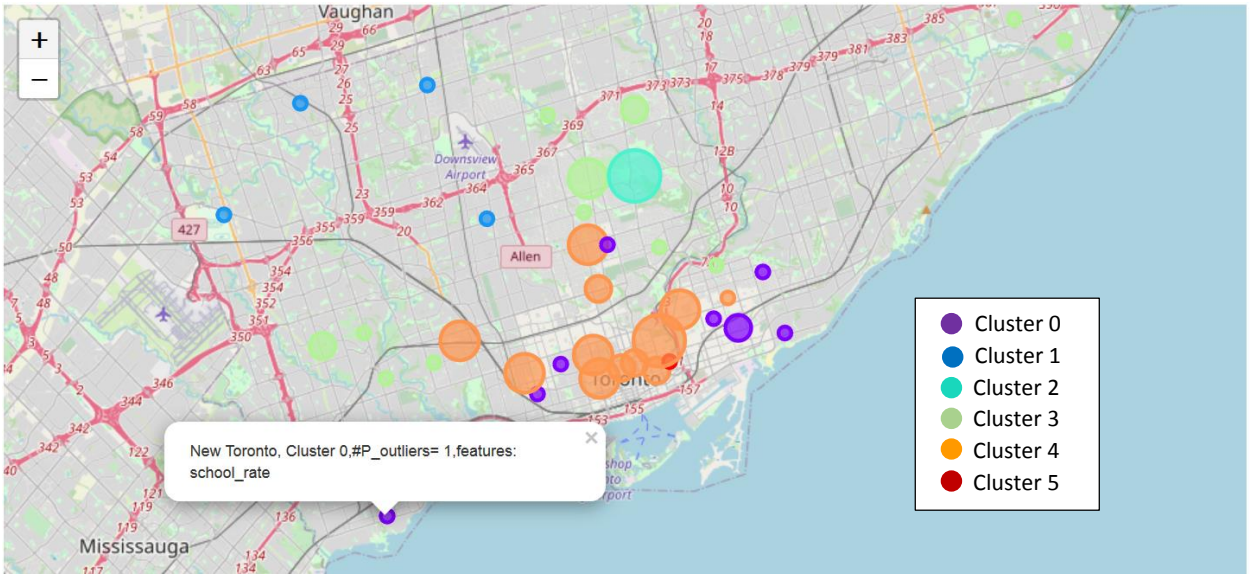

Figure 31. The map distribution plot of neighbourhoods with at least one positive indicator outlier. The size corresponds to the number of outliers (larger size means larger number of outliers) and the color represents the cluster the neighbourhood belongs to.

Figure 32 shows the neighbourhoods that have at least one negative indicator outlier. Again, larger size of circle means higher the number of outliers for that neighbourhood and the color represents the cluster it belongs to. Neighbourhood *Bridle Path-Sunnybrook-York Mills* which had one of the highest number of positive feature outliers also has one of the highest negative outliers.

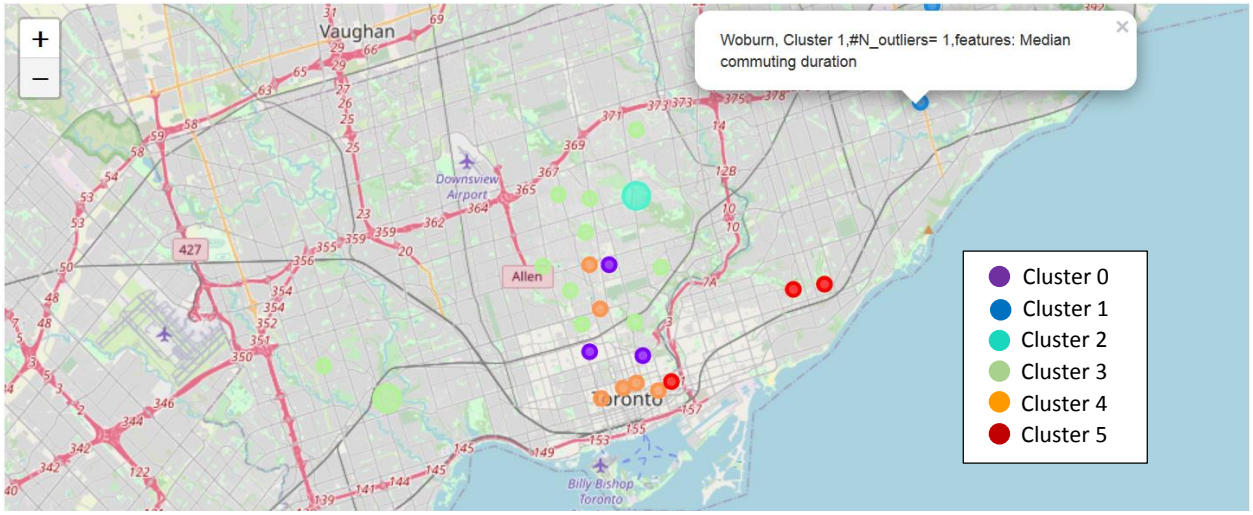

Figure 32. The map distribution plot of neighbourhoods with at least one negative indicator outlier. The size corresponds to the number of outliers (larger size means larger number of outliers) and the color represents the cluster the neighbourhood belongs to.

5. Discussions

As expected, there is no cluster and neighbourhood that can stand out in terms of being either the best or the worst neighbourhood in all features we have used for this study. Choosing to live in a neighbourhood within a cluster really depends on one's priorities in terms of selection criteria. With the tables and plots shown above, and knowing what the main selection criteria are for a person, one can decide which cluster of neighbourhoods fit their needs better.

For instance, for a family that have children within school age range, cluster 4 seems to be the best option as the school rates are the highest, there are plenty of food and drink and shops and store and also recreation places available. Also walkability is the best, with many amenities within walking distance. The cluster 4 also shows decent public transit and one of the lowest commuting durations. The major downsides are however, relatively less green spaces and tree cover and highest crime rates. The neighbourhoods within this cluster are mainly located in the downtown area. The neighbourhood *Cabbagetown-South St.James Town* seem to be a good neighbourhood in this cluster as it shows one of the highest number of positive indicator outliers and zero negative outliers.

Cluster 5 on the other hand, seems to be the least favorable cluster. It shows relatively high crime rate, the highest unemployment rates, the lowest recreation rates, the least proportion of educated population, the lowest amount of income, relatively poor public transit and the lowest school rates. These make this the least favorable cluster for a family.

If one does not mind expensive home prices and rent (which could mean better houses and apartments) and their main priorities are access to recreation centers, food and drink places, lowest amount of commuting duration, lowest unemployment rates, highest proportion of educated people, low crime rates, quiet and less busy neighbourhood, and high school rate, and specifically, the highest amount of

green spaces and tree cover, cluster 2 seems like a very good option. This cluster includes only one neighbourhood: *Bridle Path-Sunnybrook-York Mills*.

Compared to the rest of the clusters, cluster 3 neighbourhoods can be considered overall above average. It has one of the lowest population density, lowest crime rate, the second highest median income, second least unemployment rates, second best tree cover rates, and relatively average commuting duration, school rates, and walk scores.

Overall, clusters 0 and 1 seem to be at the average and below average levels. They can compete with each other on many levels. Depending on one's priorities, either one can be preferred over the other. For instance, cluster 0 has higher home prices and rents, higher crime rates, lower median income, higher population density, less tree cover, and less public transit. However, on the positive side, compared to cluster 1, it provides lower unemployment rate, larger proportion of educated people, much better walkability, and larger number of shops, store, food and drink places, and recreation rates.

6. Conclusions

In this project, we have gathered data on most important criteria that one would consider in choosing a neighbourhood to live in Toronto. We use these features in a K-means clustering algorithm to group the 140 Toronto neighbourhoods into a few non-overlapping categories. The neighbourhoods within each category share similar characteristics and are dissimilar to neighbourhoods across other clusters.

We have used two main data resources, City of Toronto Open Data catalogues and Foursquare API, to query data on 15 different important features, namely, availability of schools, housing prices, rental costs, crime rate, household income, unemployment rate, commuting duration, population density, percentage of educated population, amount of green space and trees, walkability, public transit availability, access to shops and stores, access to food and drink places, and access to recreation.

Having performed detailed exploratory data analysis and applied a K-means clustering algorithm, we have divided the neighbourhoods into 6 main clusters, 0 to 5. Cluster 4 contains mainly neighbourhood in downtown Toronto. Cluster 0 includes neighbourhoods to the east and west of cluster 4. Cluster 3 neighbourhoods are mainly located to the North and Southwest. Clusters 1 and 5 seem to be mainly located on the East and Northwest side of Toronto. Interestingly, cluster 2 only includes one single neighbourhood, *Bridle Path-Sunnybrook-York Mills*.

Our detailed analysis of the features in the neighbourhood clusters has shown that cluster 4 is most likely the preferred cluster for majority of people and cluster 5 is the least favorable. However, choosing one cluster or neighbourhood over the rest completely depends on one's priorities. For instance cluster 4 can be very appealing to someone looking for high walkability and school availability, but can be less favorable for someone looking mainly for low population density and crime rates.

Overall, we believe this work provides a comprehensive analysis of the main decision criteria for choosing neighbourhoods in Toronto to live in and can be considered as a useful guide by potential residents looking to select most suitable neighbourhood depending on their priorities.