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

Objective

Cluster 140 Toronto neighbourhoods to different categories that can be used as a guide by potential residents to choose their best and avoid their least favorable neighbourhoods based on their priorities included in the list of characteristics we analyze here.

Neighbourhood grouping criteria (Features)

We use 15 main decision criteria to investigate in our analysis. These are the major characteristics that potential residents consider when choosing their neighbourhood to live in.

They can be divided into two groups: positive and negative indicators. Residents are generally looking for neighbourhoods with:

higher/more
(Positive indicators)

- Number of schools
- Income
- Number of educated residents
- Green space or tree cover
- Walkability
- Public transit
- Shops and stores
- Food and drink places
- Recreation centres & entertainment

lower/less
(Negative indicators)

- Home prices
- Crime rate
- Average rent
- Unemployment rate
- Commuting duration
- Population density

Data acquisition and cleaning

Data is gathered mainly from two resources:

- City of Toronto Open Data Catalogue
- Foursquare API

- Neighbourhoods, their coordinates, and information regarding population, green space, education, crime rate, unemployment rate, housing, public transport, income, walk score, and commute duration are gathered from different catalogue tables in the City of Toronto Open Data Catalogue.

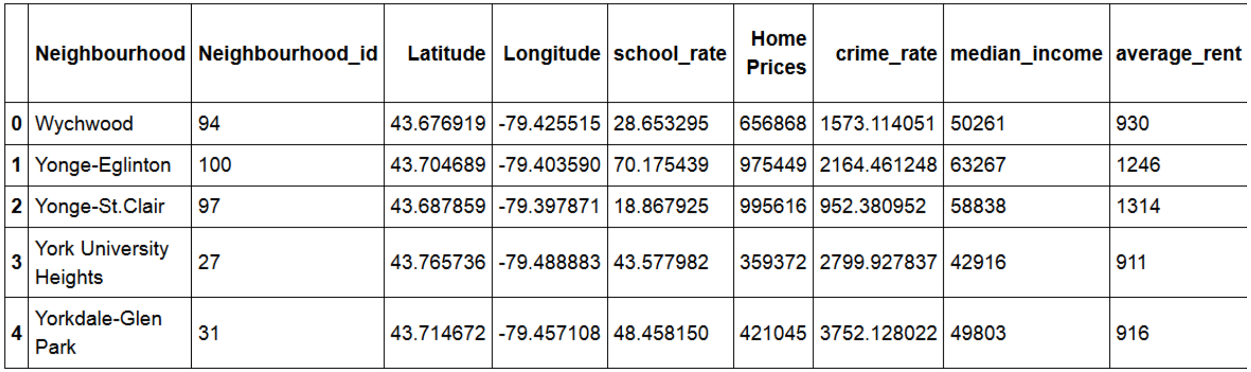
- Foursquare API is used to query venues within 500 meters of neighbourhoods which are divided into three categories of food and drink, shops and stores, and recreation.

- The data tables are parsed and cleaned, invalid values are removed or replaced, and all features are merged into one feature table. Some features are standardized in the form of scores having considered a fixed population.

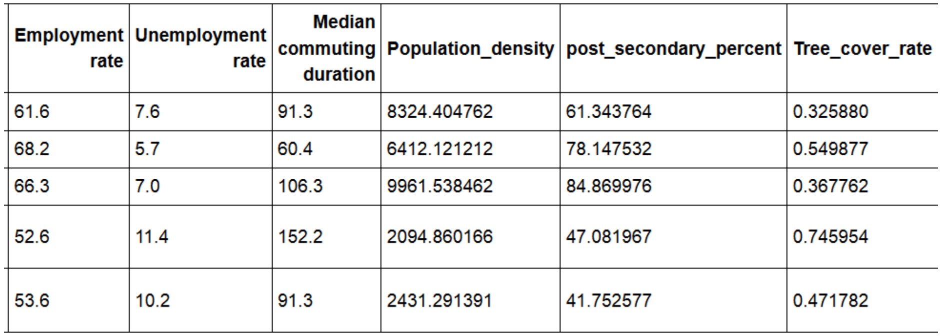
- The final feature tables has 16 features (one of which is dropped in the exploratory data analysis) for each neighbourhood

Data acquisition and cleaning

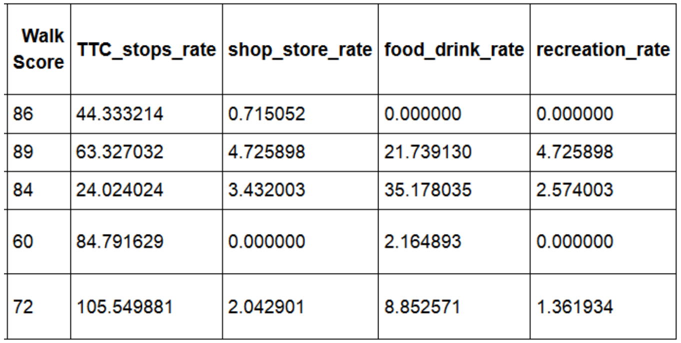
The first few rows of the feature table (for 5 neighbourhoods only)



	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



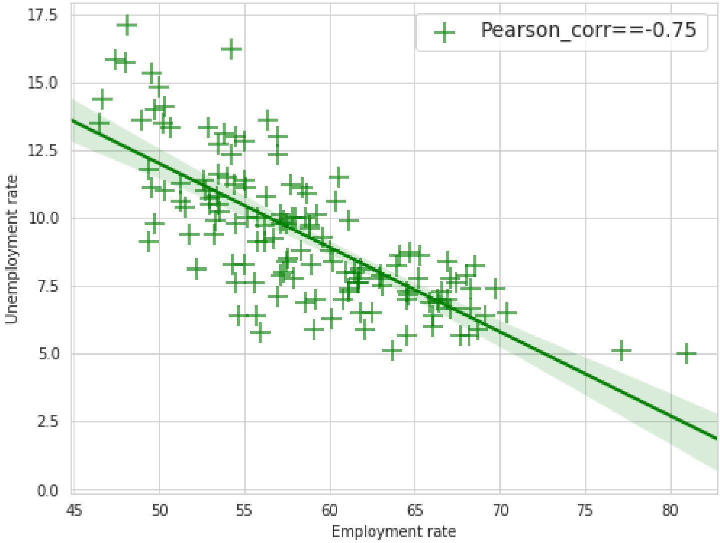
Employment rate	Unemployment rate	Median commuting duration	Population_density	post_secondary_percent	Tree_cover_rate
61.6	7.6	91.3	8324.404762	61.343764	0.325880
68.2	5.7	60.4	6412.121212	78.147532	0.549877
66.3	7.0	106.3	9961.538462	84.869976	0.367762
52.6	11.4	152.2	2094.860166	47.081967	0.745954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

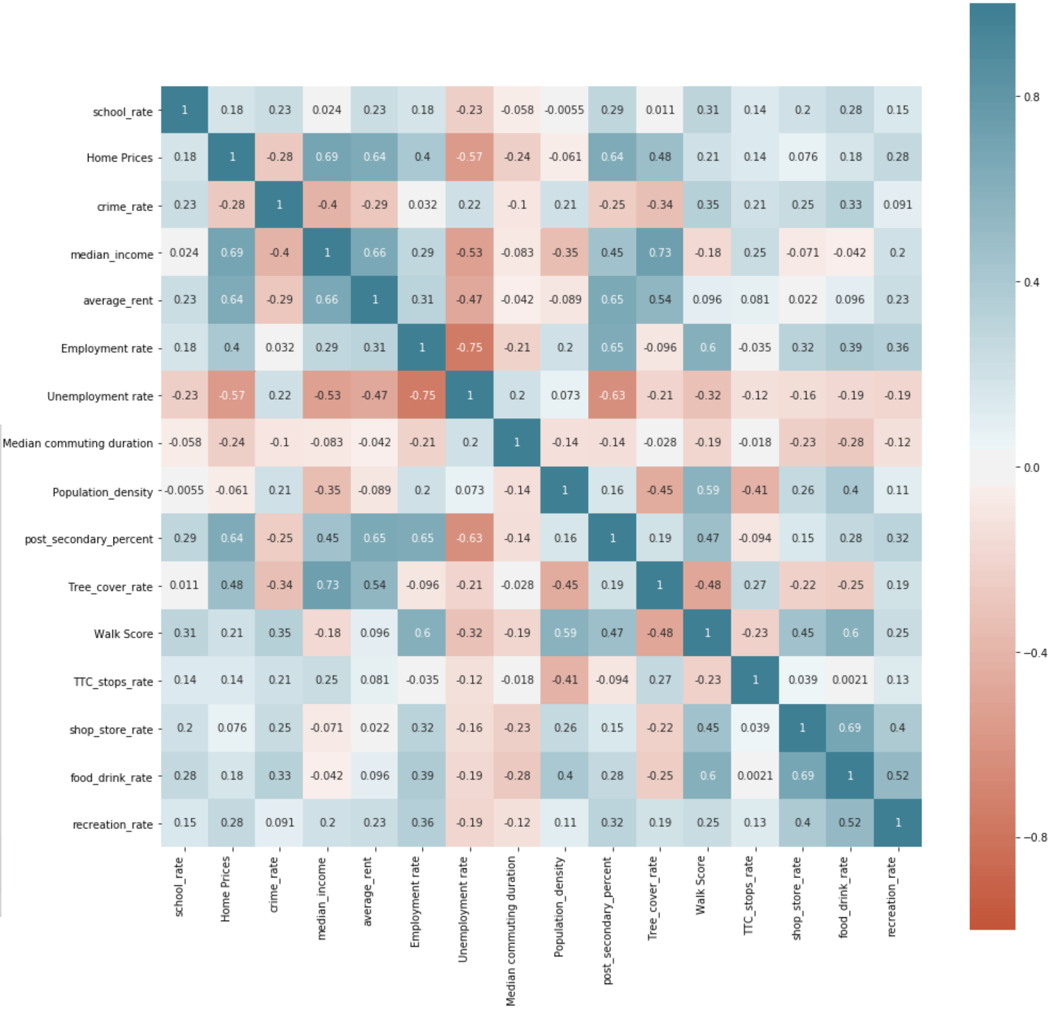


Walk Score	TTC_stops_rate	shop_store_rate	food_drink_rate	recreation_rate
86	44.333214	0.715052	0.000000	0.000000
89	63.327032	4.725898	21.739130	4.725898
84	24.024024	3.432003	35.178035	2.574003
60	84.791629	0.000000	2.164893	0.000000
72	105.549881	2.042901	8.852571	1.361934

Methodology: Exploratory data analysis

- Correlation heatmap and regression plot are used to identify linearly depenedent variables and remove one of them.

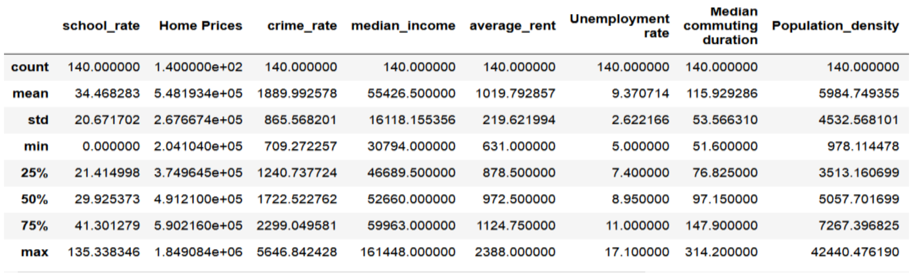




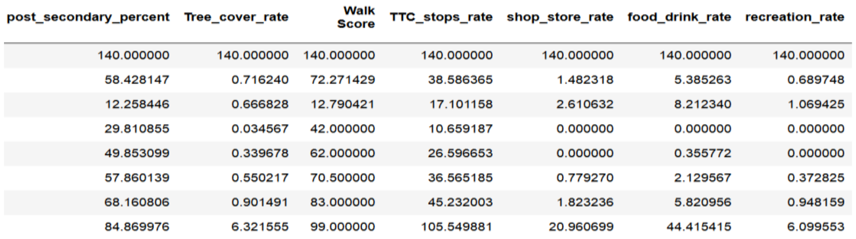
	school_rate	Home Prices	crime_rate	median_income	average_rent	Employment rate	Unemployment rate	Median commuting duration	Population_density	post_secondary_percent	Tree_cover_rate	Walk Score	TTC_stops_rate	shop_store_rate	food_drink_rate	recreation_rate
school_rate	1	0.18	0.23	0.024	0.23	0.18	-0.23	-0.058	-0.0055	0.29	0.011	0.31	0.14	0.2	0.28	0.15
Home Prices	0.18	1	-0.28	0.69	0.64	0.4	-0.57	-0.24	-0.061	0.64	0.48	0.21	0.14	0.076	0.18	0.28
crime_rate	0.23	-0.28	1	-0.4	-0.29	0.032	0.22	-0.1	0.21	-0.25	-0.34	0.35	0.21	0.25	0.33	0.091
median_income	0.024	0.69	-0.4	1	0.66	0.29	-0.53	-0.083	-0.35	0.45	0.73	-0.18	0.25	-0.071	-0.042	0.2
average_rent	0.23	0.64	-0.29	0.66	1	0.31	-0.47	-0.042	-0.089	0.65	0.54	0.096	0.081	0.022	0.096	0.23
Employment rate	0.18	0.4	0.032	0.29	0.31	1	-0.75	-0.21	0.2	0.65	-0.096	0.6	-0.035	0.32	0.39	0.36
Unemployment rate	-0.23	-0.57	0.22	-0.53	-0.47	-0.75	1	0.2	0.073	-0.63	-0.21	-0.32	-0.12	-0.16	-0.19	-0.19
Median commuting duration	-0.058	-0.24	-0.1	-0.083	-0.042	-0.21	0.2	1	-0.14	-0.14	-0.028	-0.19	-0.018	-0.23	-0.28	-0.12
Population_density	-0.0055	-0.061	0.21	-0.35	-0.089	0.2	0.073	-0.14	1	0.16	-0.45	0.59	-0.41	0.26	0.4	0.11
post_secondary_percent	0.29	0.64	-0.25	0.45	0.65	0.65	-0.63	-0.14	0.16	1	0.19	0.47	-0.094	0.15	0.28	0.32
Tree_cover_rate	0.011	0.48	-0.34	0.73	0.54	-0.096	-0.21	-0.028	-0.45	0.19	1	-0.48	0.27	-0.22	-0.25	0.19
Walk Score	0.31	0.21	0.35	-0.18	0.096	0.6	-0.32	-0.19	0.59	0.47	-0.48	1	-0.23	0.45	0.6	0.25
TTC_stops_rate	0.14	0.14	0.21	0.25	0.081	-0.035	-0.12	-0.018	-0.41	-0.094	0.27	-0.23	1	0.039	0.0021	0.13
shop_store_rate	0.2	0.076	0.25	-0.071	0.022	0.32	-0.16	-0.23	0.26	0.15	-0.22	0.45	0.039	1	0.69	0.4
food_drink_rate	0.28	0.18	0.33	-0.042	0.096	0.39	-0.19	-0.28	0.4	0.28	-0.25	0.6	0.0021	0.69	1	0.52
recreation_rate	0.15	0.28	0.091	0.2	0.23	0.36	-0.19	-0.12	0.11	0.32	0.19	0.25	0.13	0.4	0.52	1

Methodology: Exploratory data analysis

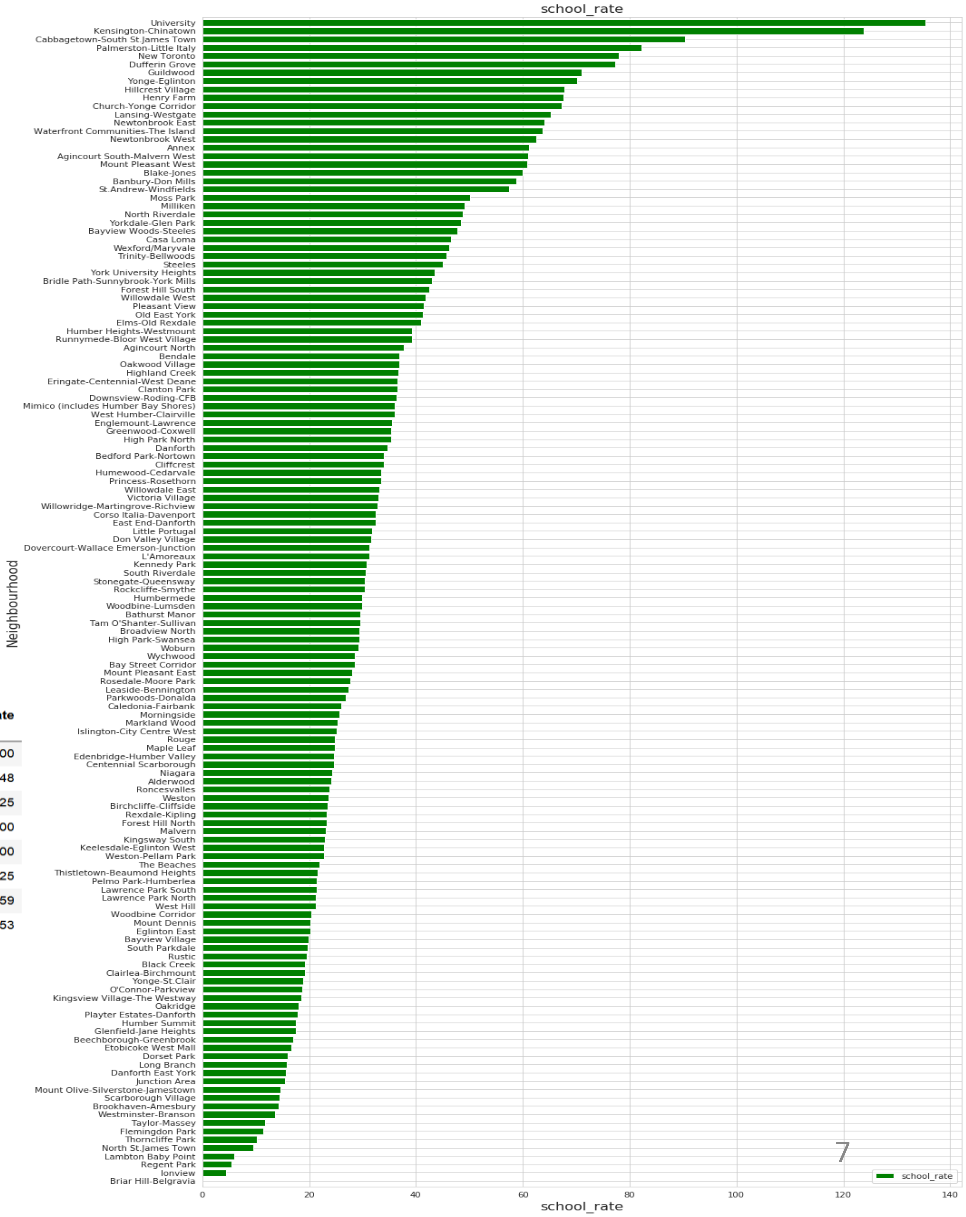
- Bar plots and descriptive statistics are other useful tools to get a better sense of our input data and features.



	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



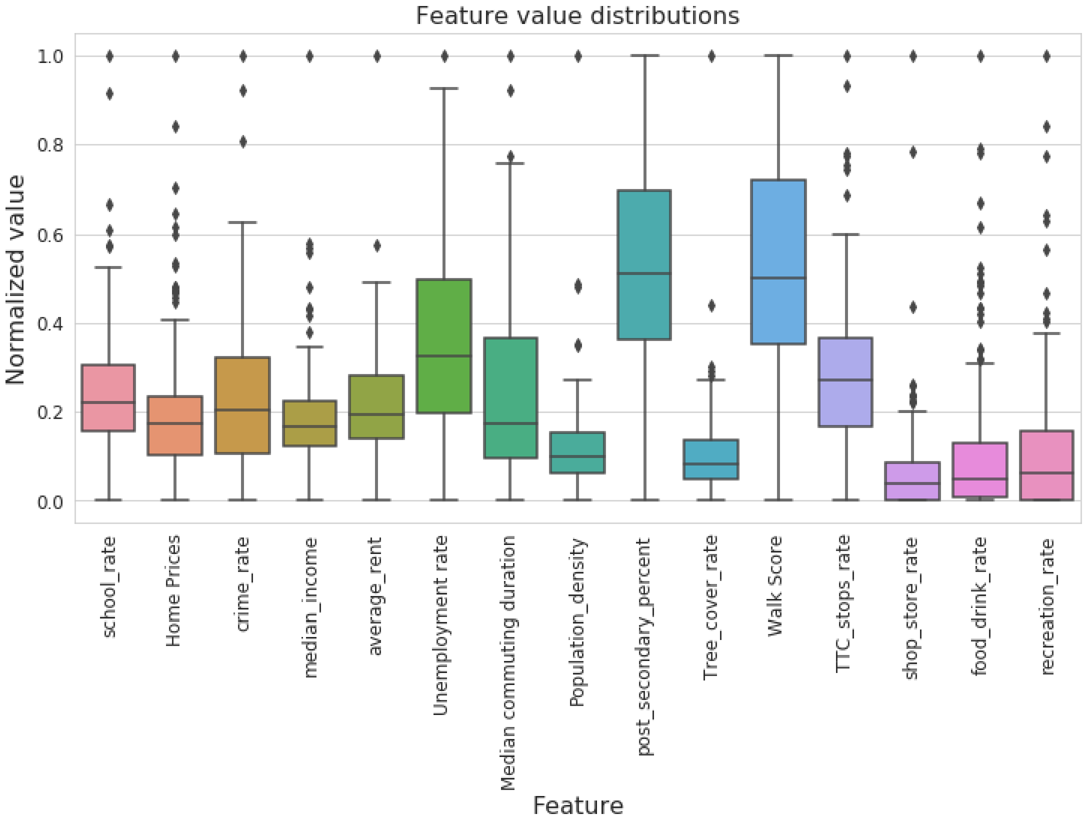
post_secondary_percent	Tree_cover_rate	Walk Score	TTC_stops_rate	shop_store_rate	food_drink_rate	recreation_rate
140.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000
58.428147	0.716240	72.271429	38.586365	1.482318	5.385263	0.689748
12.258446	0.666828	12.790421	17.101158	2.610632	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



Methodology: Exploratory data analysis

- Box plots also helped us identify outlier neighbourhoods in different features

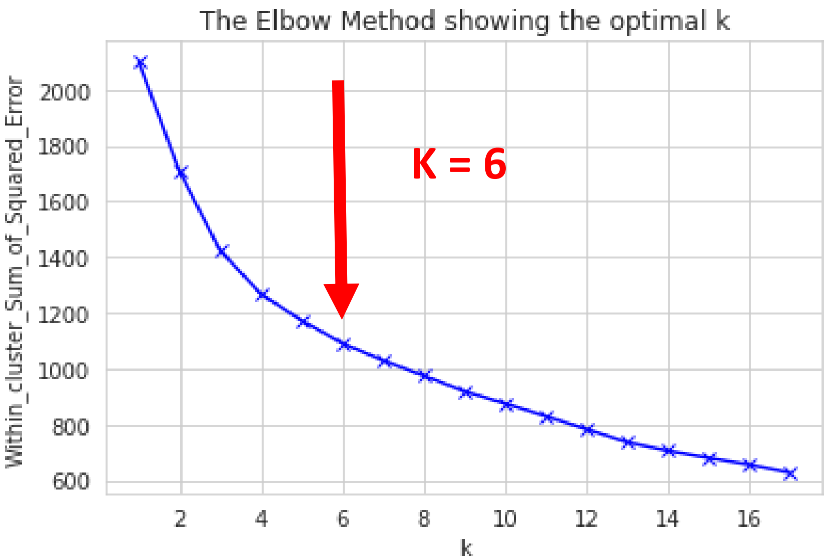
- None of the features seem to have anomalously low values that can be seen as an outlier.
- *Home prices*, *median income,food_and_drink_rate*, and *recreation_rate* show the largest number of large outliers.
- *Walk Score*, *post_secondary_percent*, *Unemployment_rate*, *average_rent* and *crime_rate* show no to very low number of outliers.

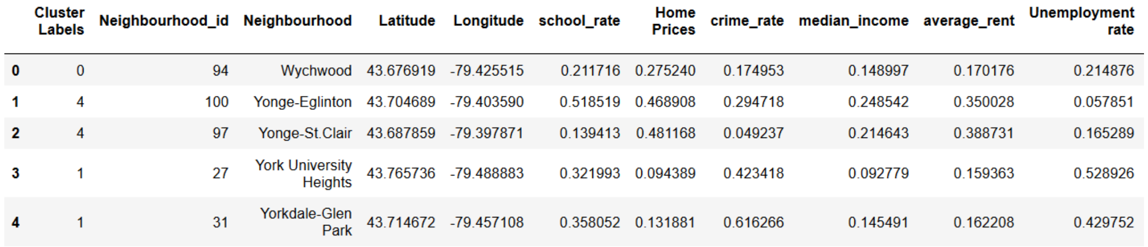


- Majority of neighbourhoods show very low *food_drink_rate*, *recreation_rate* and shop_store*_rate* when compared with their maxima. So are *Tree_cover_rate* and *Population_density*.
- Majority of neighbourhoods seem to have average walk score and proportion of educated people with postsecondary educations

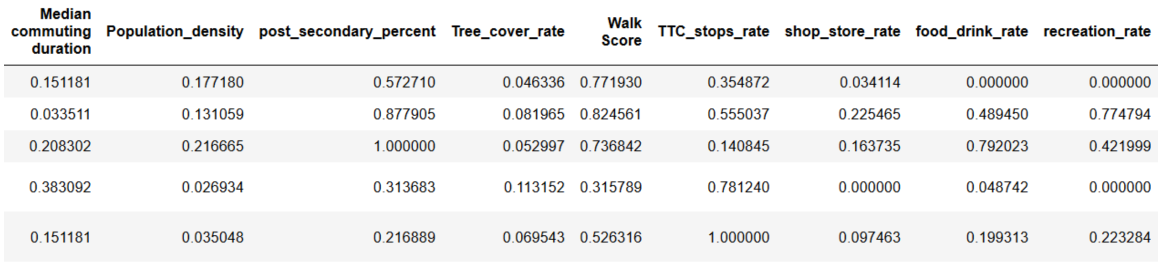
Machine learning (clustering)

- K-means algorithm is used on standardized features for clustering
- An Elbow method is used to find an optimal K-value of 6
- The cluster labels found by K-means are inserted back to the feature data table for further analysis





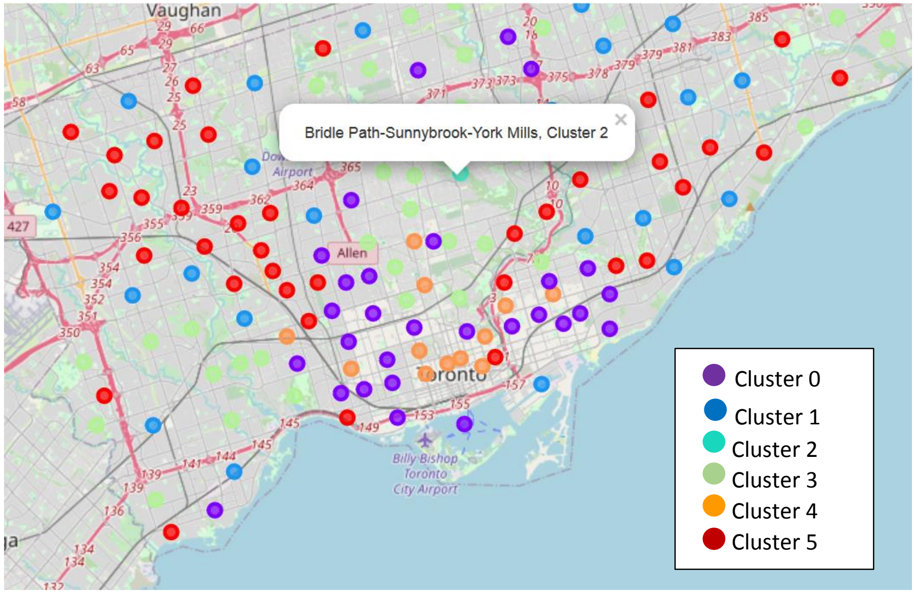
	Cluster Labels	Neighbourhood_id	Neighbourhood	Latitude	Longitude	school_rate	Home Prices	crime_rate	median_income	average_rent	Unemployment rate
0	0	94	Wychwood	43.676919	-79.425515	0.211716	0.275240	0.174953	0.148997	0.170176	0.214876
1	4	100	Yonge-Eglinton	43.704689	-79.403590	0.518519	0.468908	0.294718	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

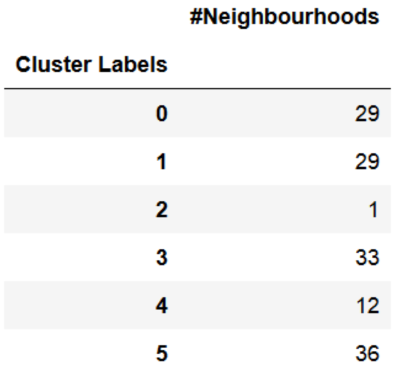


Median commuting duration	Population_density	post_secondary_percent	Tree_cover_rate	Walk Score	TTC_stops_rate	shop_store_rate	food_drink_rate	recreation_rate
0.151181	0.177180	0.572710	0.046336	0.771930	0.354872	0.034114	0.000000	0.000000
0.033511	0.131059	0.877905	0.081965	0.824561	0.555037	0.225465	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

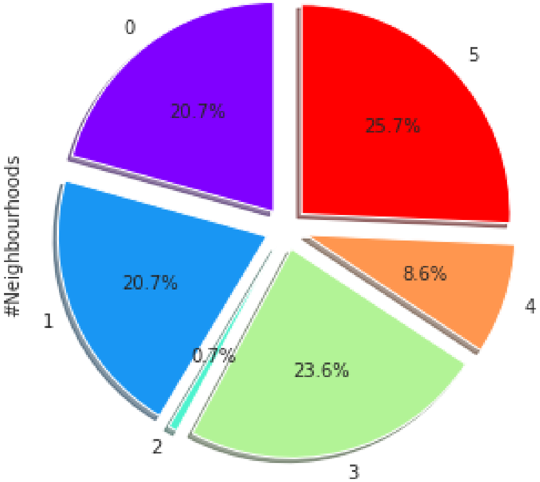
Results

- The 6 neighbourhood clusters are plotted with different colors on a map for visual inspection
- The number and proportion of neighbourhoods in clusters is shows below in form of table and pie chart.
- Cluster 2 includes only 1 neighbourhood highlighted on the map
- Cluster 5 has the highest number of neighbourhoods

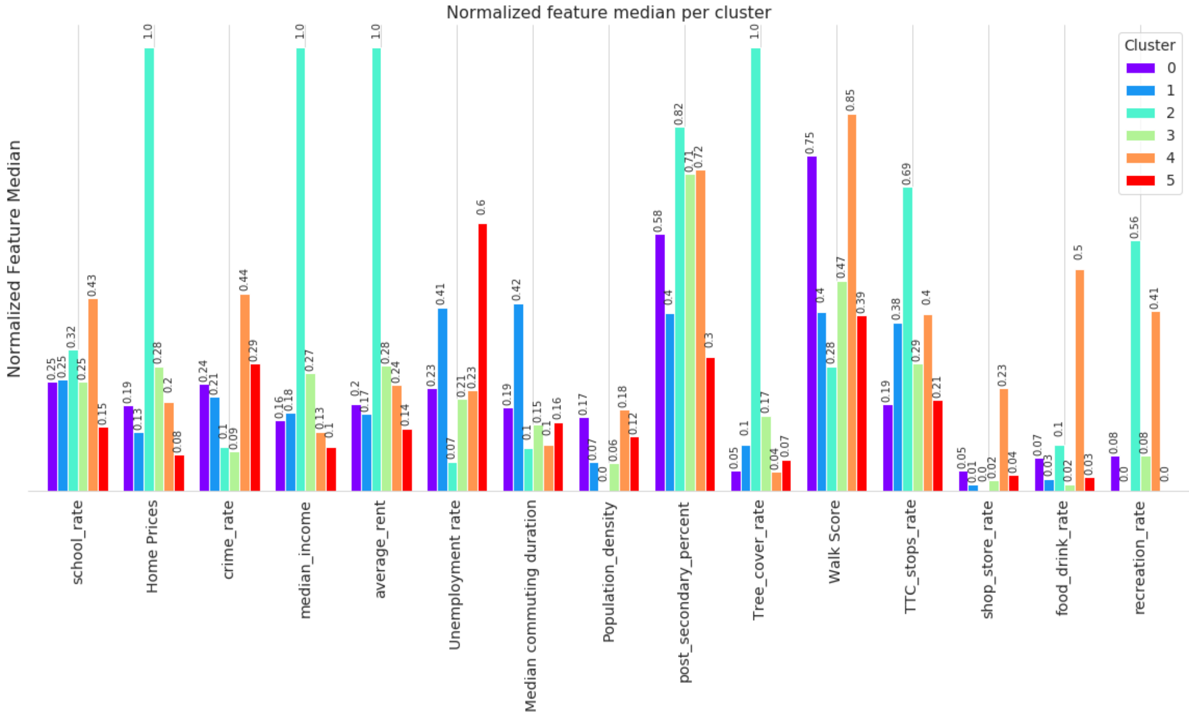


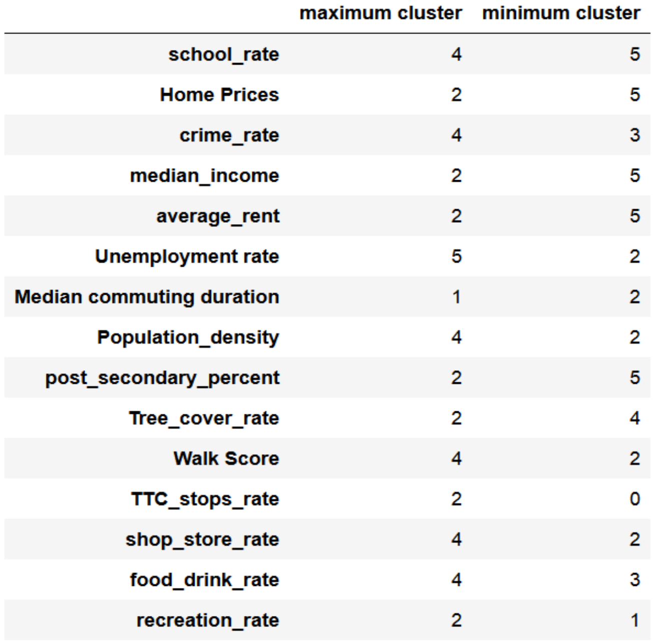


Cluster Labels	#Neighbourhoods
0	29
1	29
2	1
3	33
4	12
5	36



Results





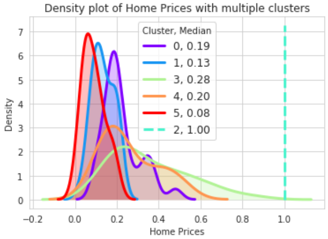
	maximum cluster	minimum cluster
school_rate	4	5
Home Prices	2	5
crime_rate	4	3
median_income	2	5
average_rent	2	5
Unemployment rate	5	2
Median commuting duration	1	2
Population_density	4	2
post_secondary_percent	2	5
Tree_cover_rate	2	4
Walk Score	4	2
TTC_stops_rate	2	0
shop_store_rate	4	2
food_drink_rate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

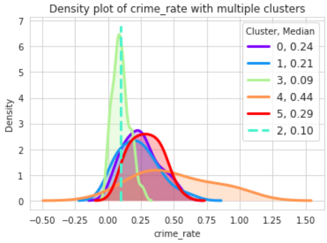
- The median of standardized features calculated across neighbourhoods in different clusters and plotted for each feature as bar plots. A comparison of different clusters based on different features can be made using this plot.
- The clusters that have the highest and lowest median values for different features are listed in the table.

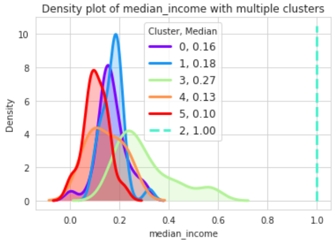
Results

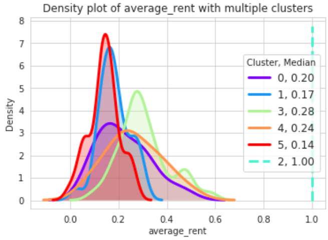
- Density plots used to compare distribution of features across different clusters
- Cluster 2 has only one neighbourhood➔ plotted as a single dashed
- The distributions clearly show differences between clusters.

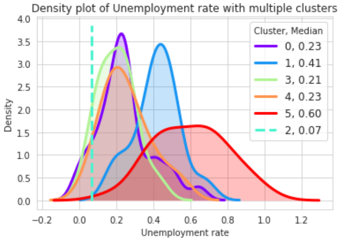
- The level of separation between clusters is different for different features

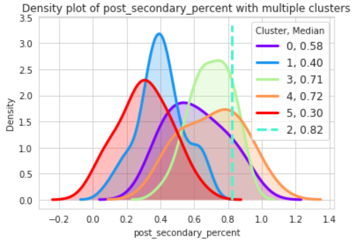


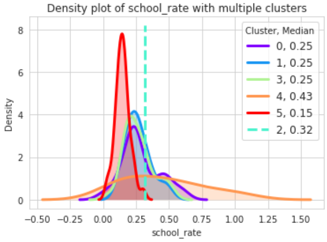


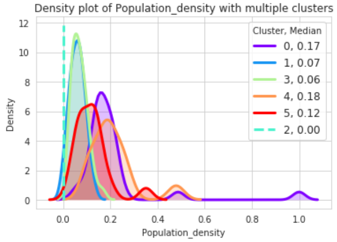


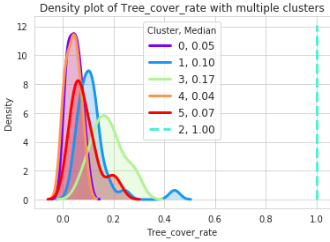


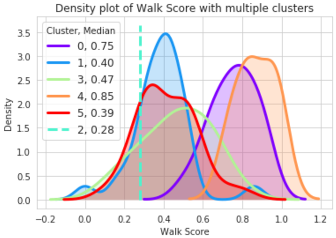


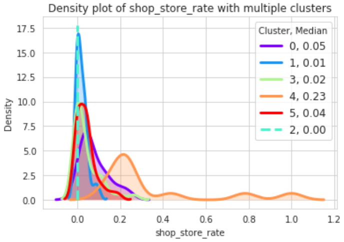


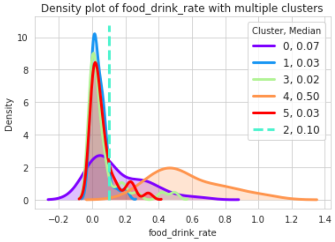


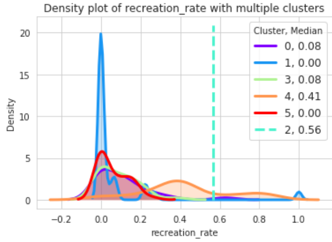


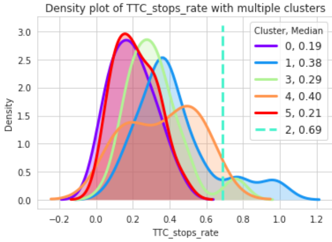


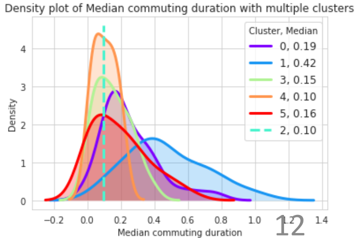








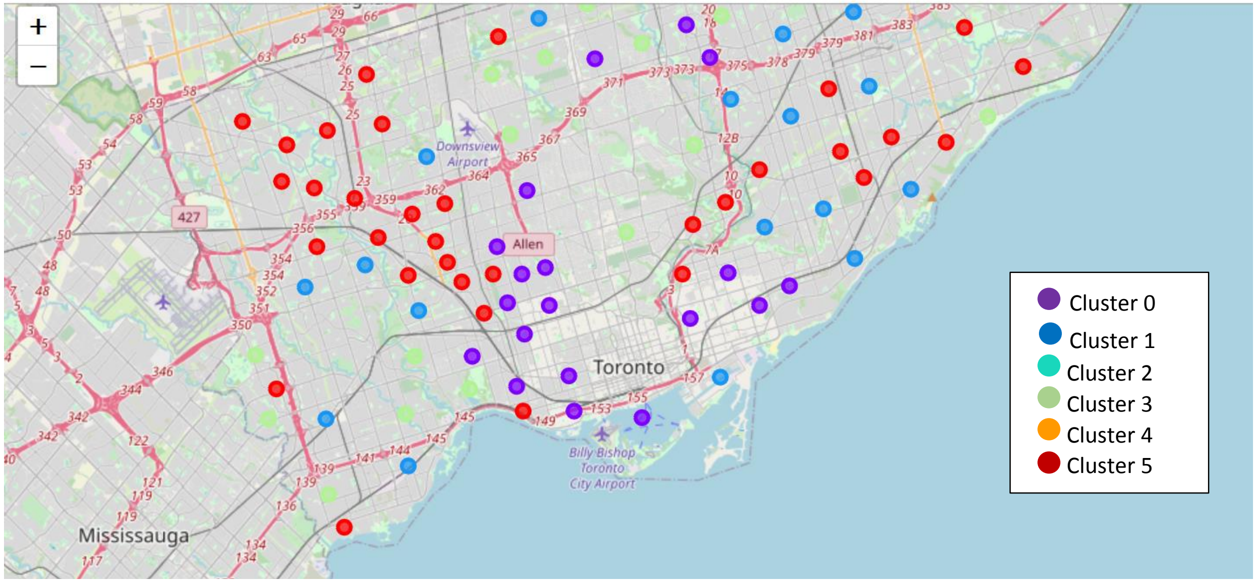




Results

- The number of feature outliers (positive and negative) and their type in each neighbourhood is investigated
- Outlier neighbourhoods are those that fall outside the interquartile range
- 90 out of 140 neighbourhoods show zero number of outlier features.
- Map shows the zero-outlier neighbourhoods and their associated clusters.
- Gaps are observed in the downtown and Western areas.

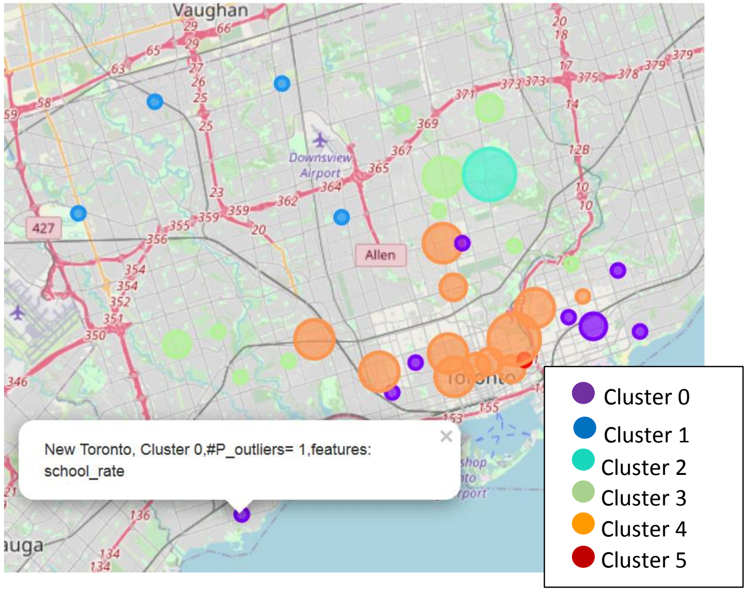
Neighbourhoods with zero outlier features



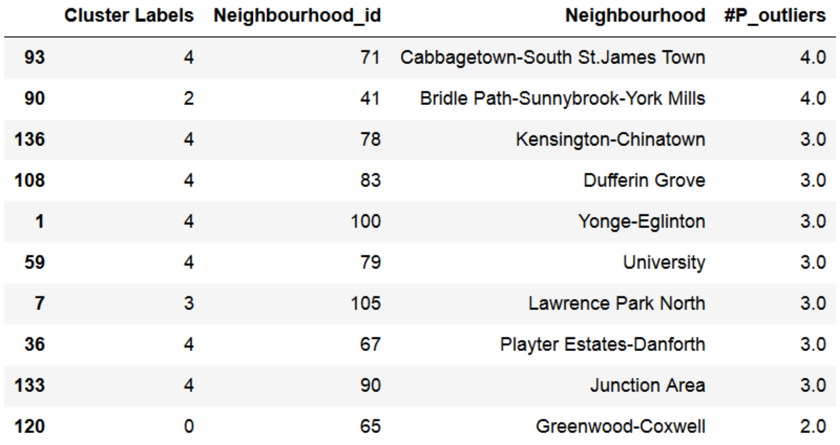
Results

- 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.
- The table shows the first 10 neighbourhoods with highest number of positive feature outliers along with their cluster label.
- Neighbourhoods, *Cabbagetown-South St.James Town* and *Bridle Path-Sunnybrook-York Mills* have the largest number of positive indicators, which is 4.

Neighbourhoods with >0 positive feature outliers



First 10 neighbourhoods with highest number of positive feature outliers

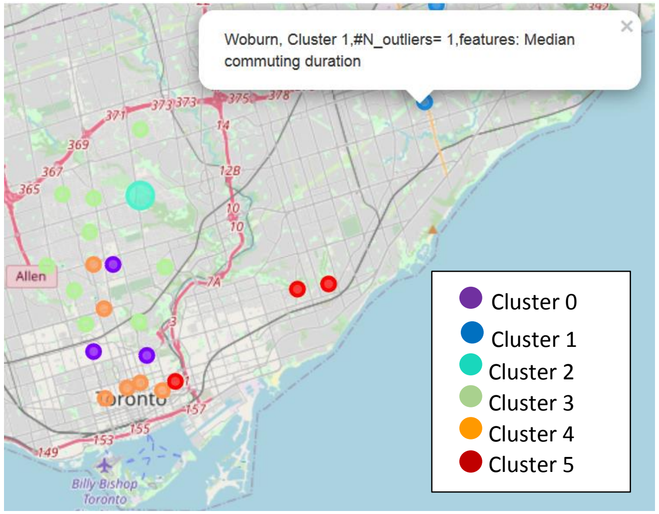


	Cluster Labels	Neighbourhood_id	Neighbourhood	#P_outliers
93	4	71	Cabbagetown-South St.James Town	4.0
90	2	41	Bridle Path-Sunnybrook-York Mills	4.0
136	4	78	Kensington-Chinatown	3.0
108	4	83	Dufferin Grove	3.0
1	4	100	Yonge-Eglinton	3.0
59	4	79	University	3.0
7	3	105	Lawrence Park North	3.0
36	4	67	Playter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

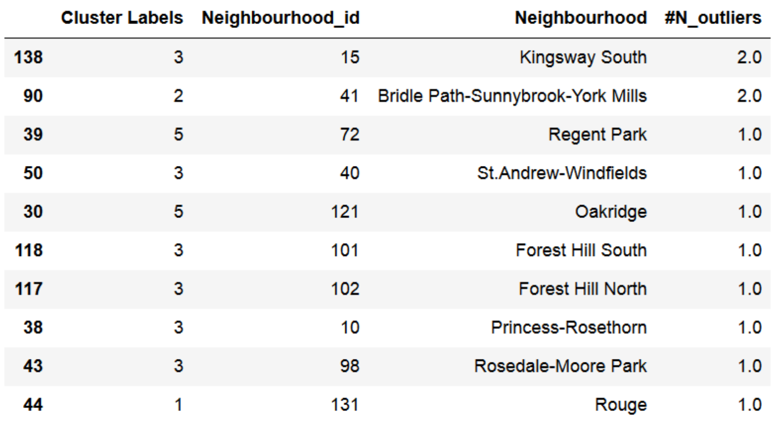
Results

- The map distribution plot of neighbourhoods with at least one negative indicator outlier.
- The table shows the first 10 neighbourhoods with highest number of negative feature outliers along with their cluster label.
- 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.
- 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.

Neighbourhoods with >0 negative feature outliers



First 10 neighbourhoods with highest number of negative feature outliers



	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

Discussions

- 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 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 for cluster 4 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.

Discussions

- Cluster 5 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.

Discussions

- 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.

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*.

Conclusions

- 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.