Neighbourhoods Recommendation
-A Comparison Among 140 Neighbourhoods of Toronto

1、Introduction

1.1 Background

As the economic center of Canada and one of the largest financial centers in the world,Toronto attracted a large number of investors and immigrants. According to the Census of Population held across Canada in 2016,the population of Toronto was 2731571,which increased by 4.5% in the past 5 years.This number is still climbing. However,as the largest City in Canada,the land square of Toronto is 630.2 square kilometers.Moreover, there are 140 neighbourhoods officially recognized by the City of Toronto and upwards of 240 official and unofficial neighbourhoods within the city's boundaries. The question of which neighbourhood is better for settlement has draw many people's concern.This report will help newcomers to make a decision by comparing and clustering neighbourhoods in Toronto.

1.2 Problem

When concerning a suitable neighbourhood for settlement, people may consider whether house price is affordable ,whether this neighbourhood is safe , whether they are likely to find a job, and what kinds of venues do this neighbourhood have. Therefore, This report will find out a group of neighbourhoods with low crime rate, low unemployment rate and low shelter unaffordable rate. In addition, venues of these neighbourhoods will be explored to provide further reference for us.

1.3 Interest

New immigrants and residents who consider to make a move are interested in the this report. The government of Toronto may also be interested in the comparison result among neighbourhoods.

2、Data Acquisition and Cleaning

2.1 Data resource

There were a lot of classification method of neighbourhoods in Toronto. In this report, I chose the 140 neighbourhoods classification method, because boundaries of these social planning neighbourhoods do not change over time, and 140 neighbourhood classification is also used by government of Toronto. Official data is more reliable. Latitudes and and longitudes of neighbourhoods boundaries, unemployment rate and unaffordable rate data are all derived from 2016 Census Profile on Toronto's city government website. Crime rate data is derived from Toronto police sevice website.

2.2 Data Cleaning

Unemployment rate and neighbourhood classification can be easily get from data resource, but some other data should be dealt first.

2.2.1 Latitudes and Longitudes of neighbourhoods

In order to draw maps and explore venues around neighbourhoods, not only boundaries data but also central latitudes and longitudes are needed. Hence, I first download boundaries latitudes and longitudes data in shp format. Then, I find the center of each neighbourhood using mapwindow software.Last, I convert shp format data into geojson format, so that we can read these latitudes and longitudes files of both boundaries and centers of 140 neighbourhoods .

2.2.2 Unaffordable rate data

In the Census Profile, there is a row showing how many people have to spend 30% or more of income on shelter costs for each neighbourhood. I divide this data by the number of Owner and tenant households with household total income greater than zero,and acquire he percentage of people who has to pay more than 30% income on shelters in each neighbourhood. This result is used to represent the unaffordable rate of each neighbourhood. Compared with houseprice, one advantage of unaffordable rate is that it avoids the influence of some other factors like income levels among neighourhoods.

2.2.3 Crime rate data

The crime rate data downloaded is the average crime rate between 2014-2018 per 100000 people(calculated using 2016 Census population).However, the value of 17th neighbourhood(Mimico) is missing. I used average crime rate to replace the missing value.

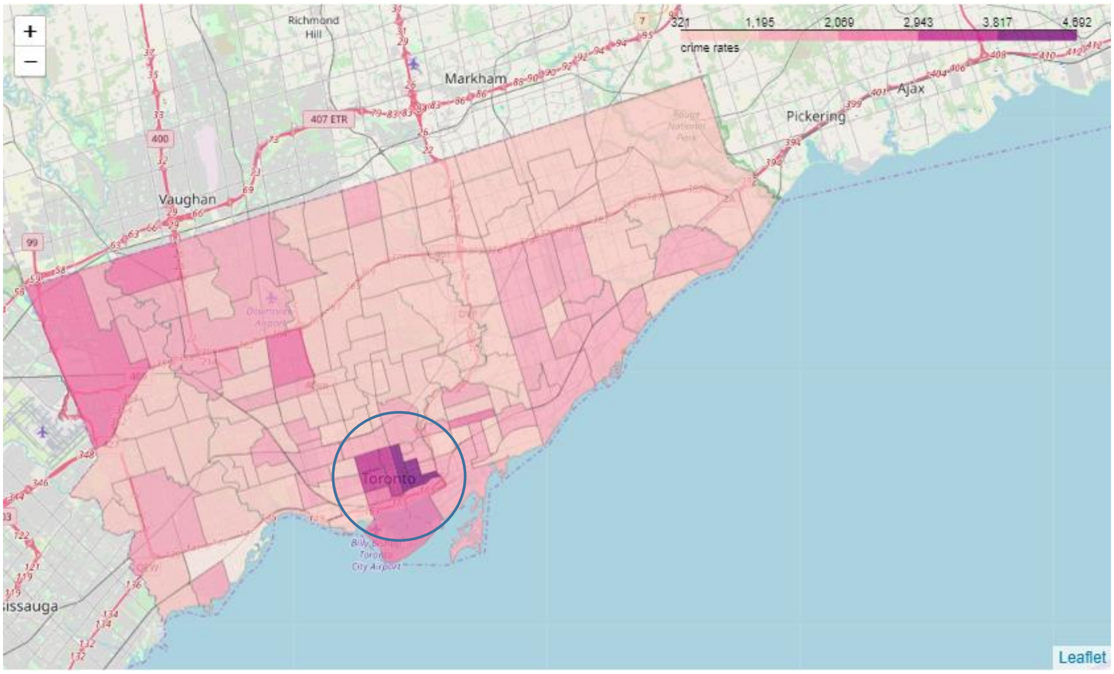
Moreover, to make data more comparable, I multiplied unemployment rate and unaffordable rate by 1000. Thus, data we collected are all based on the incidence probability per 10000 people.

3、Methodology

3.1Data Visualization

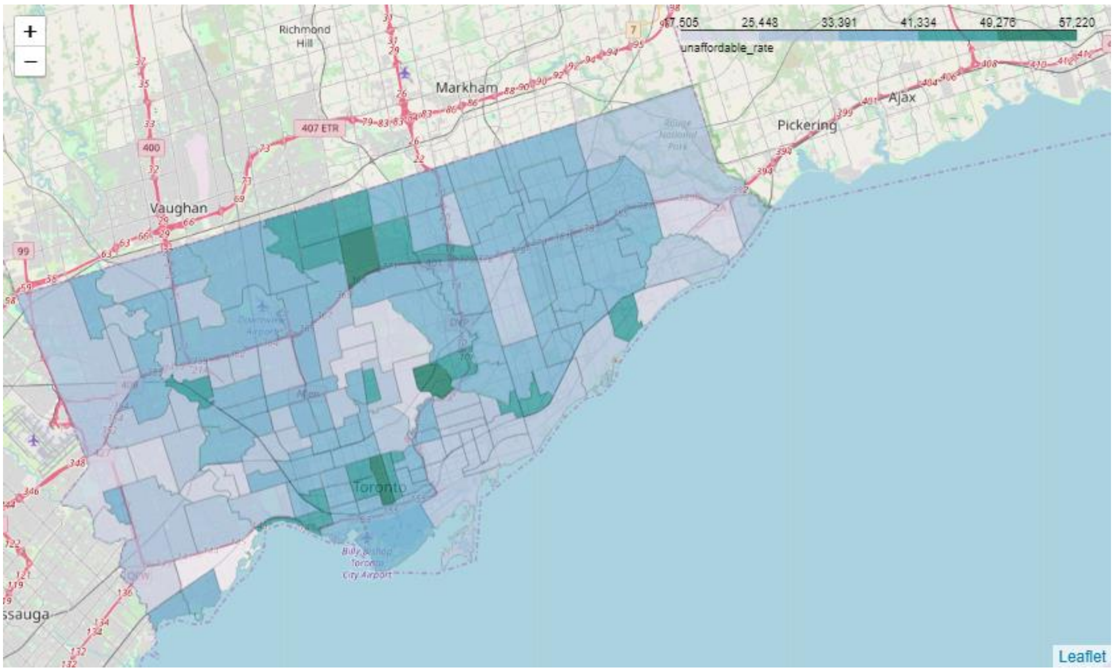
After cleaning the whole data set, I observed the differences of crime rate,unemployment rate and unaffordable rate among neighbourhoods respectively using choropleth maps.

Figure1: Choropleth map of crime rate



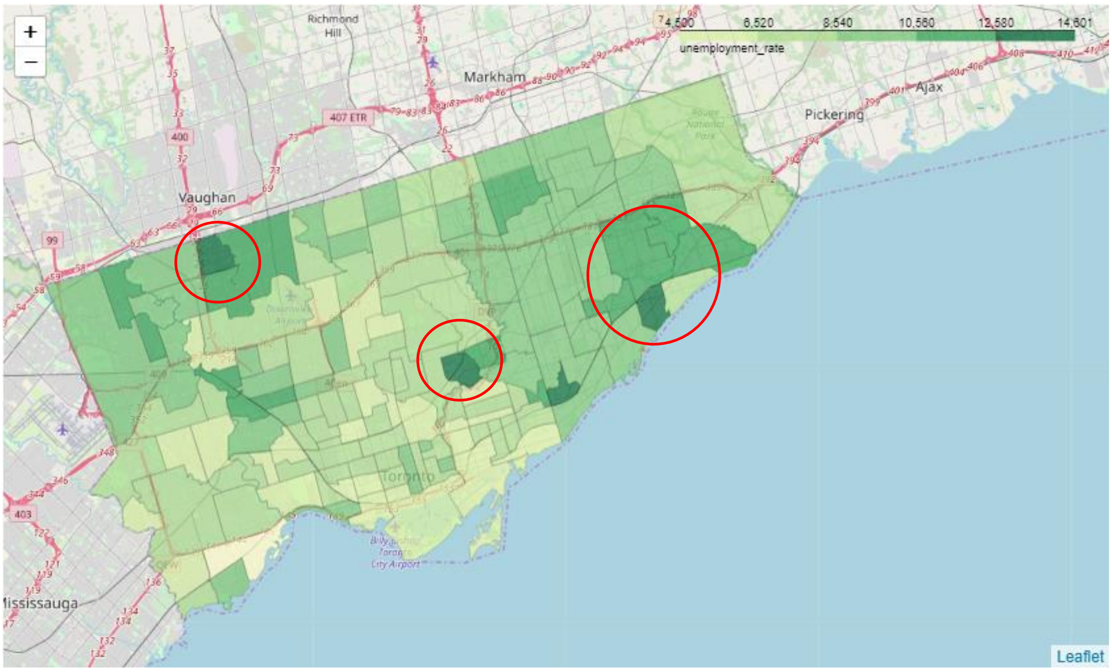
From figure 1, we can find that areas in blue circle have highest crime rate, while the central and eastern Toronto have lower crime rate on average.

Figure2: Choropleth map of unaffordable rate



The figure corresponding to house unaffordable rate indicates that people live in central Toronto are more likely to pay less than 30% income for shelters.The rate is also low in western and eastern Toronto.

Figure3: Choropleth map of unemployment rate



According to figure 3, the unemployment rate is lowest in southwestern Toronto.

Some parts of central and eastern Toronto has relative low unemployment rate as well. Conversely,areas in red circles undertake higher unemployment rate.

All three figures represent that neighbourhoods have different performance on safety, employment rate, and shelters affordability.

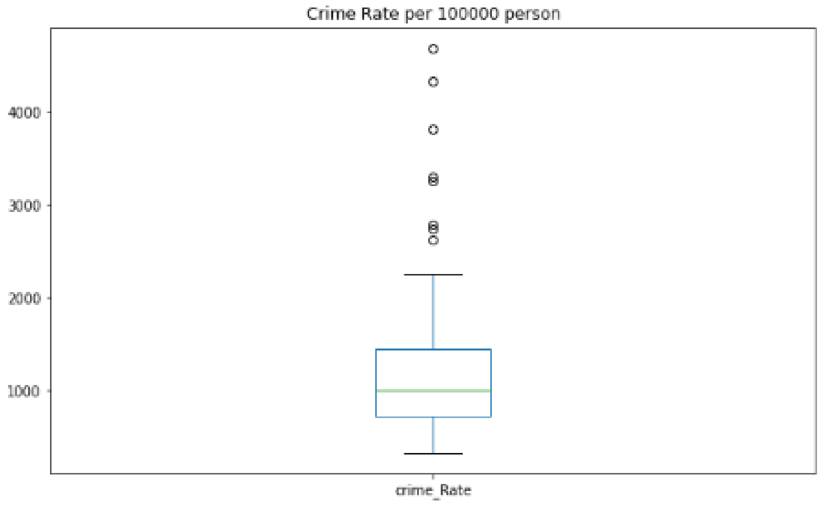
3.2Data Description

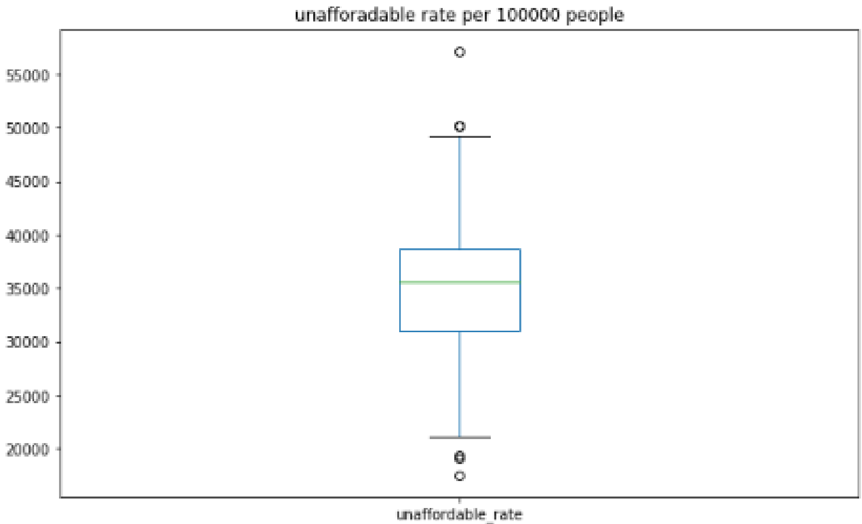
To explore our data in depth, we will take a look of our numeric data with data description.

Table1: Data Description

	Hood_ID	crime_Rate	unaffordable_rate	unemployment_rate
count	140.0000	140.000000	140.000000	140.000000
mean	70.5000	1191.992358	35129.006503	8303.571429
std	40.5586	717.218038	7010.588694	1896.509885
min	1.0000	321.997958	17505.720824	4500.000000
25%	35.7500	726.737742	31087.577963	6900.000000
50%	70.5000	997.096104	35601.983792	8200.000000
75%	105.2500	1440.850918	38718.919969	9625.000000
max	140.0000	4691.365474	57219.610477	14600.000000

Figure4: Box Figure of crime rate, unaffordable rate, and unemployment rate


Crime Rate per 100000 person
4000
3000
2000
1000
crime_Rate


unafforadable rate per 100000 people
55000
50000
45000
40000
35000
30000
25000
20000
unaffordable_rate

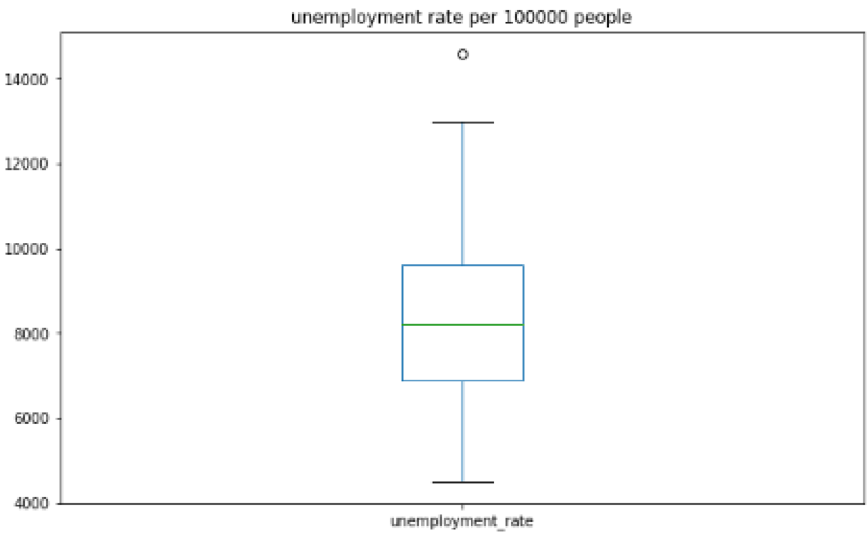
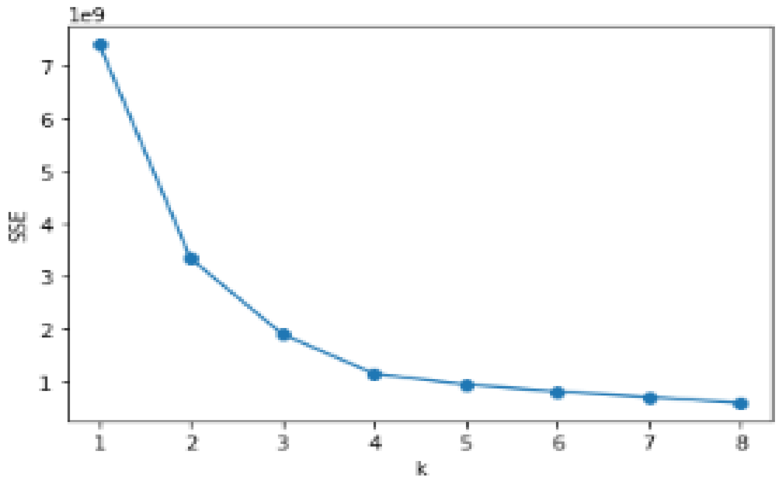

unemployment rate per 100000 people
14000
12000
10000
8000
6000
4000
unemployment_rate

Table and figures above show the distribution of our data.The minimum ,median and maximum value of each factor will help us to distinguish clusters of our data in the data clustering stage.

3.3Data Clustering

We aim at finding out a group of neighbourhoods with better performance on three factors: safety、employment and shelter affordability, which means we need to cluster 140 neighbourhoods according to their common characteristics on performance first. Therefore,I decided to use unsupervised learning K-means algorithm to cluster the neighbourhoods. Before running K-Means to cluster the neighbourhoods, we first use elbow method to find optimum K.

Figure5: Elbow figure



As shown in the figure above, the optimum k would be 4.Therefore，our data is separated into 4 clusters. Samples of each cluster are displayed below：

Cluster0：

	Neighbourhood	Hood_ID	id	crime_Rate	unaffordable_rate	unemployment_rate	Clus
1	Mount Olive-Silverstone-Jamestown	2	Mount Olive-Silverstone-Jamestown (2)	1204.751291	37658.388241	12100	0
5	Kingsview Village-The Westway	6	Kingsview Village-The Westway (6)	930.034130	35112.540193	9900	0
12	Etobicoke West Mall	13	Etobicoke West Mall (13)	620.155039	37008.733624	7900	0
16	Mimico	17	Mimico (17)	1191.992358	39864.864865	6200	0
17	New Toronto	18	New Toronto (18)	1403.631898	39162.790698	8700	0
18	Long Branch	19	Long Branch (19)	1058.655222	35379.812695	7100	0
21	Humbermede	22	Humbermede (22)	1216.068238	37524.752475	10200	0
23	Black Creek	24	Black Creek (24)	1554.004880	36816.939891	12700	0
26	York University Heights	27	York University Heights (27)	1986.538587	40216.855594	10700	0
29	Brookhaven-Amesbury	30	Brookhaven-Amesbury (30)	842.884309	40354.938272	10000	0
31	Englemount-Lawrence	32	Englemount-Lawrence (32)	718.162839	36446.331110	8900	0
33	Bathurst Manor	34	Bathurst Manor (34)	1500.383957	35714.285714	7200	0
37	Lansing-Westgate	38	Lansing-Westgate (38)	891.164202	38584.474886	7200	0

Cluster1：

	Neighbourhood	Hood_ID	id	crime_Rate	unaffordable_rate	unemployment_rate	Clus
34	Westminster-Branson	35	Westminster-Branson (35)	589.706933	46176.470588	9200	1
35	Newtonbrook West	36	Newtonbrook West (36)	1024.924295	46486.790332	8400	1
36	Willowdale West	37	Willowdale West (37)	735.250868	46240.851630	9800	1
43	Flemingdon Park	44	Flemingdon Park (44)	729.596043	44955.300128	10600	1
49	Newtonbrook East	50	Newtonbrook East (50)	1451.014619	48517.298188	8800	1
50	Willowdale East	51	Willowdale East (51)	883.620262	50259.418001	8500	1

Cluster2：

	Neighbourhood	Hood_ID	id	crime_Rate	unaffordable_rate	unemployment_rate	Clus
9	Princess-Rosethorn	10	Princess-Rosethorn (10)	533.355909	19455.252918	6000	2
10	Eringate-Centennial-West Deane	11	Eringate-Centennial-West Deane (11)	616.619360	22684.172137	7400	2
11	Markland Wood	12	Markland Wood (12)	491.788882	24622.531940	6200	2
14	Kingsway South	15	Kingsway South (15)	880.155562	19134.078212	7500	2
15	Stonegate-Queensway	16	Stonegate-Queensway (16)	692.702288	27303.424988	6700	2
19	Alderwood	20	Alderwood (20)	699.137730	21800.433839	6100	2
38	Bedford Park-Nortown	39	Bedford Park-Nortown (39)	911.251981	27008.149010	5500	2
40	Bridle Path-Sunnybrook-York Mills	41	Bridle Path-Sunnybrook-York Mills (41)	886.344751	22839.506173	8000	2
55	Leaside-Bennington	56	Leaside-Bennington (56)	644.676050	22023.346304	6900	2

Cluster3：

	Neighbourhood	Hood_ID	id	crime_Rate	unaffordable_rate	unemployment_rate	Clus
0	West Humber-Clairville	1	West Humber-Clairville (1)	2743.756959	31840.311587	9600	3
2	Thistletown-Beaumond Heights	3	Thistletown-Beaumond Heights (3)	1214.312199	32519.083969	10400	3
3	Rexdale-Kipling	4	Rexdale-Kipling (4)	1602.825319	30859.375000	10900	3
4	Elms-Old Rexdale	5	Elms-Old Rexdale (5)	1263.778031	33540.372671	10000	3
6	Willowridge-Martingrove-Richview	7	Willowridge-Martingrove-Richview (7)	909.286792	31081.081081	8500	3
7	Humber Heights-Westmount	8	Humber Heights-Westmount (8)	795.342735	32043.530834	7400	3
8	Edenbridge-Humber Valley	9	Edenbridge-Humber Valley (9)	577.814464	30400.000000	6100	3

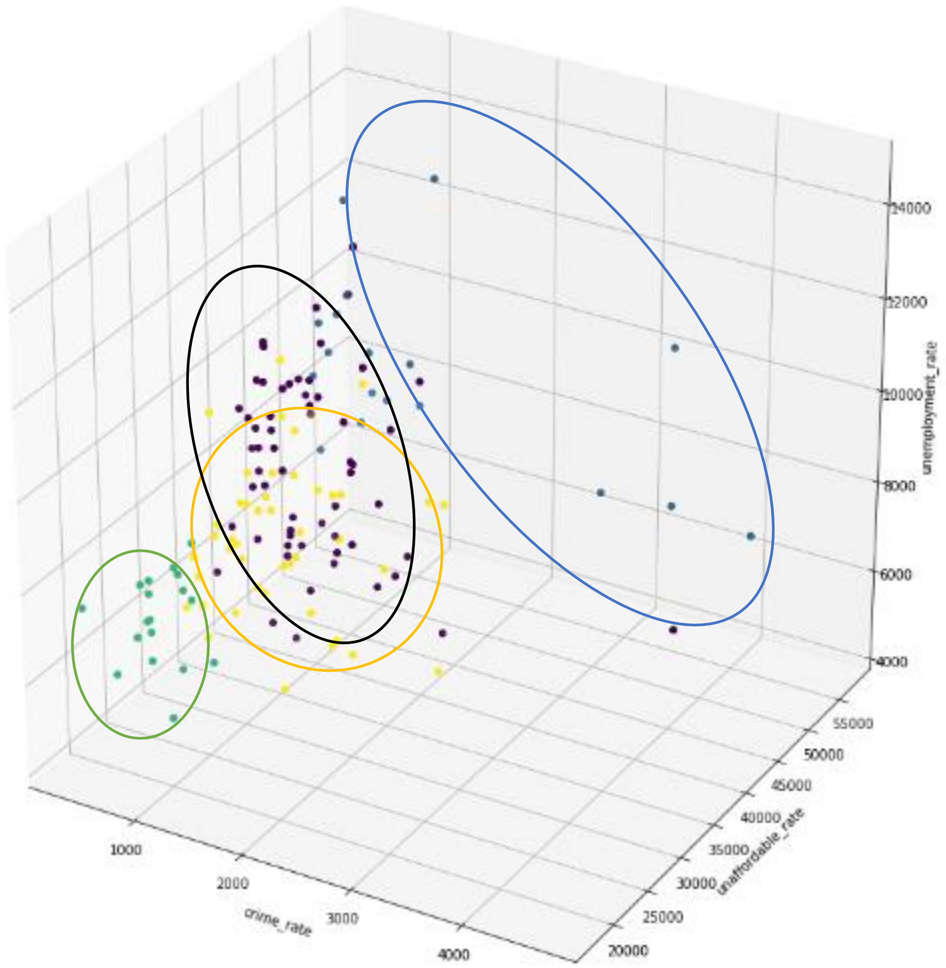
In cluster 0， neighbourhoods have higher on average unaffordable rate， high or low crime rate， and medium to high unemployment rate. The overall performance is not good.

In cluster1, neighbourhoods have very high unaffordable rate， high or low crime rate， and medium to high on average unemployment rate. The overall performance is frustrating.

In cluster2, neighbourhoods have very low unaffordable rate，lower than average crime rate， and lower than average unemployment rate. These neighbourhoods have really good overall performance.

In cluster3, neighbourhoods have medium unaffordable rate，medium crime rate，and medium to high unemployment rate. The overall performance of these neighbourhoods are OK.

Figure6: 3D scatters



From figure6, we can see the distribution more clearly. neighbourhoods in cluster 2 (in green circle) have lower unaffordable rate，crime rate，and unemployment rate.

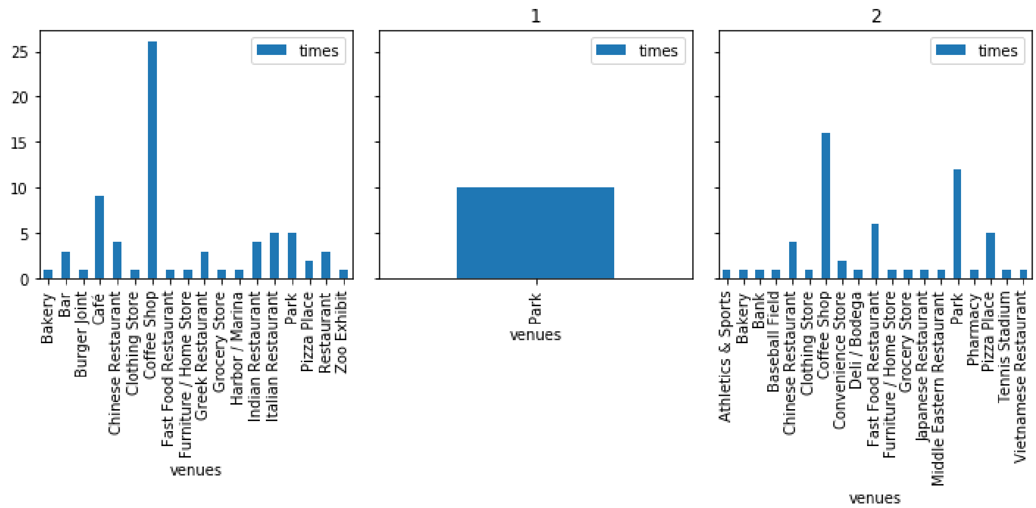
3.4Vennues Exploration

To further explore our neighbourhoods, we need to consider personal preference on venues. I utilized the Foursquare API to explore the neighbouhoods and segment them. I designed the limit as 100 venue and the radius 1000 meters for each neighbourhood from their central latitude and longitude,and acquired 344 unique categories of venues.A sample of Venues data returned from Forsquare API are listed below:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Princess-Rosethorn	43.664928	-79.544787	Princess Margaret Park	43.667835	-79.539934	Playground
1	Princess-Rosethorn	43.664928	-79.544787	Joe Fresh Style	43.660234	-79.546709	Clothing Store
2	Princess-Rosethorn	43.664928	-79.544787	Rosethorn Park	43.659923	-79.541247	Park
3	Princess-Rosethorn	43.664928	-79.544787	Pizza Pizza	43.668780	-79.536730	Pizza Place
4	Princess-Rosethorn	43.664928	-79.544787	VIVID Streetwear	43.662501	-79.533269	Clothing Store

I grouped these venues and list top 10 venue categories for each neighbourhood. As there are some common venue categories in neighbourhoods, I chose K-Means algorithm to cluster them as well.I ran K-Means algorithm with 3 clusters for neighbourhoods. To find proper labels for each cluster, I draw a bar chart displaying the number of 1st Most Common Venue in each cluster .

Figure7: Bar chart for numbers of venues category in each cluster



The proper labels that I found for each cluster according to this bar chart are listed below:

Cluster0: “Cafe & others Venues”

Cluster1: “park Venues”

Cluster2: “Multiple Social Venues”

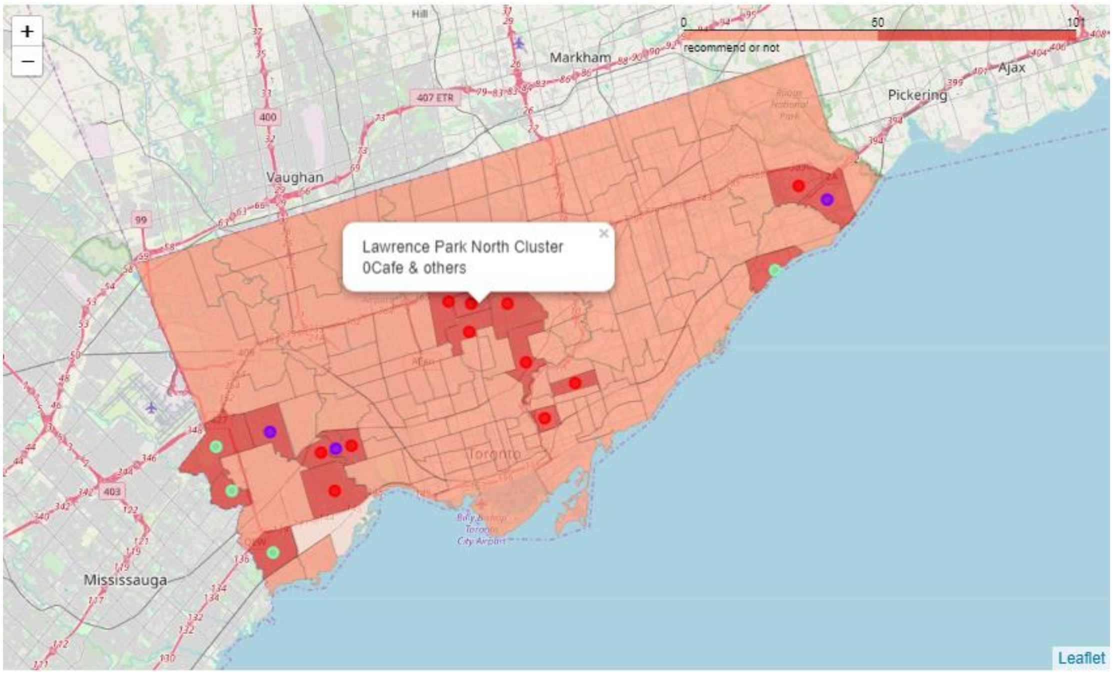
4、Results

In data Clustering section, I already found out the best neighbourhoods cluster according numeric data. Thus, I create a new table giving each neighbourhood in cluster 2 a 100 recommendation score, while other neighbouhoods get a zero recommendation score.

Secondly, I added labels of venues clusters to the new created table.

Lastly, I created a choropleth map which highlight our recommended neighbourhoods with dark red,while other neighbourhoods in light red color. I also add markers on the choropleth map to distinguish venues clusters and text information indicating neighbourhoods name.

Figure8: Final choropleth map-recommendation



The final integrated recommendation lists are shown below:

Hood	Neighbourhood	crime_Rate (per 100000 people)	unaffordable_rate (per 100000 people)	unemployment_rate (per 100000 people)	venues clustername
10	Princess-Rosethorn	533.36	19455.25	6000	park
11	Eringate-Centennial-West Deane	616.62	22684.17	7400	Multiple Social Venues
12	Markland Wood	491.79	24622.53	6200	Multiple Social Venues
15	Kingsway South	880.16	19134.08	7500	Cafe & others
16	Stonegate-Queensway	692.70	27303.42	6700	Cafe & others
20	Alderwood	699.14	21800.43	6100	Multiple Social Venues
39	Bedford Park-Nortown	911.25	27008.15	5500	Cafe & others
41	Bridle Path-Sunnybrook-York Mills	886.34	22839.51	8000	Cafe & others
56	Leaside Bennington	644.68	22023.35	6900	Cafe & others
59	Danforth - East York	628.02	27084.82	6900	Cafe & others
68	North Riverdale	676.66	26152.30	5300	Cafe & others
89	Runnymede-Bloor West Village	935.63	21073.30	5100	Cafe & others
103	Lawrence Park South	518.30	22929.38	7500	Cafe & others
105	Lawrence Park North	516.78	22437.67	6400	Cafe & others
114	Lambton Baby Point	580.09	23200.00	7600	park
133	Centennial Scarborough	359.96	17505.72	7600	park
134	Highland Creek	914.76	24357.24	8500	Cafe & others
140	Guildwood	762.35	24060.15	7900	Multiple Social Venues

5、Discussion

This report only considered 4 factors:crime rate, unaffordable rate,unemployment rate and venues in recommendation. However, there are other factors that may make sense. For example, many people regard transportation and education as important factors for their decision. Therefore, further studies should be made according to a broader range of data.

6、Conclusion

In conclusion, neighbourhoods located in some parts of eastern , central and western Toronto have lower crime rate, unemployment rate and unaffordable rate. New imigrants and residents who decide to move may consider these neighbourhoods.Moreover, park venues in our recommendation lists are mainly located in western Toronto, while Cafes venues are mainly located in central Toronto. People looking for a better environment may take these results as a reference.

7、 References

[1] Wikipedia-List of neighbourhoods in Toronto

[2] Toronto city-government neighbourhood-profiles

[3] Forsquare API

[4] Crime data on Toronto Police website