**Neighbourhoods Recommendation**

**-A Comparison Among 140 Neighbourhoods of Toronto**

1. **Introduction**
   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. 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.

1. **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.

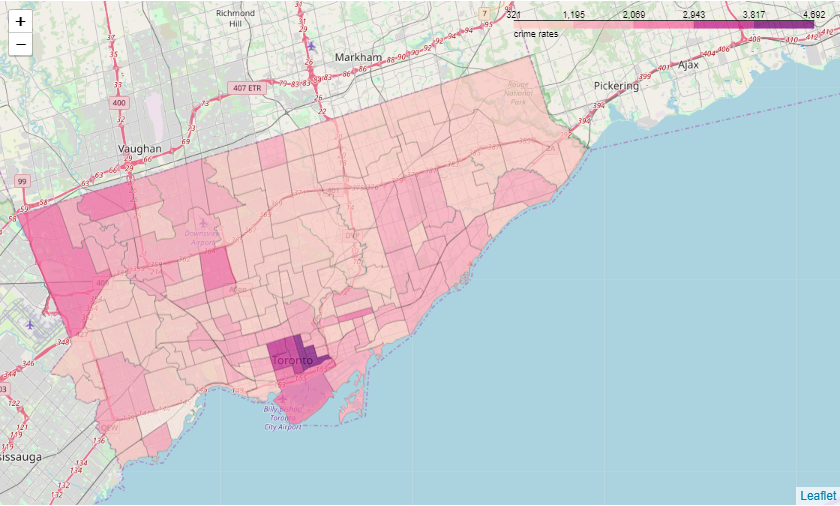
Moreover, to make data more comparable, I multiplied unemployment rate and unaffordable rate by 1000. Thus, data we collected are all based on the incidence probability per 10000 people.

1. **Methodology**

3.1Data Visualization

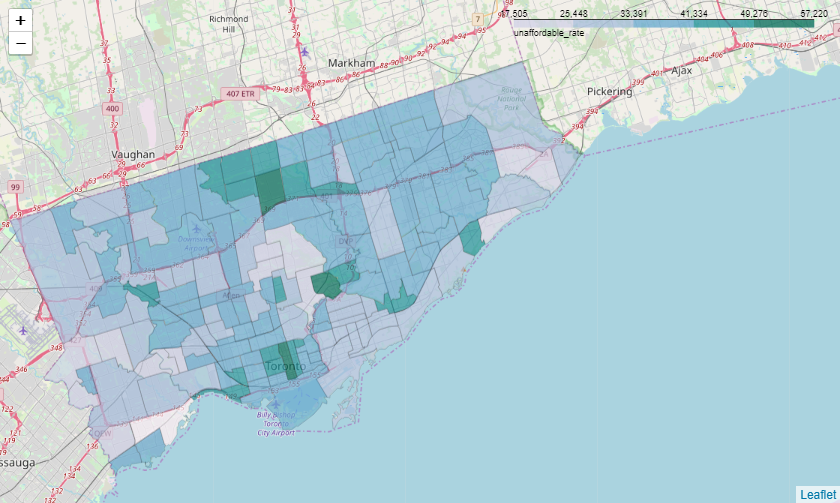
After cleaning the whole data set, I observed the differences of crime rate,unemployment rate and unaffordable rate among neighbourhoods respectively using choropleth maps.

Figure1: Choropleth map of crime rate



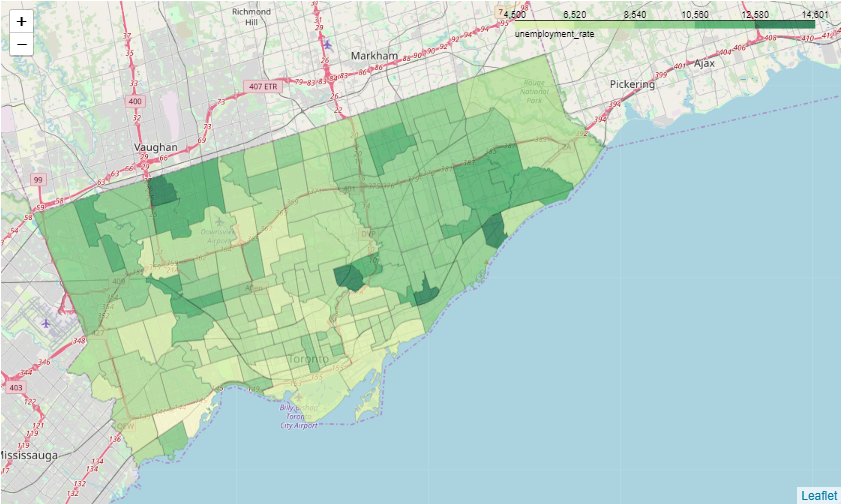
From figure 1, we can find that areas in blue circle have highest crime rate, while the central and eastern Toronto have lower crime rate on average.

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

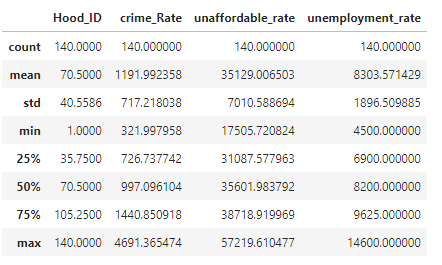


Figure4: Box Figure of crime rate, unaffordable rate, and unemployment rate

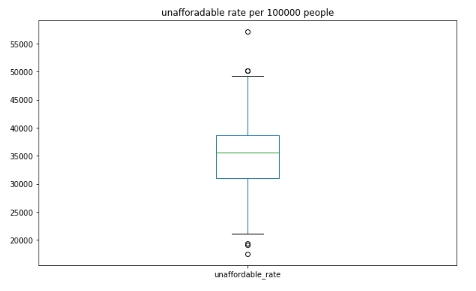
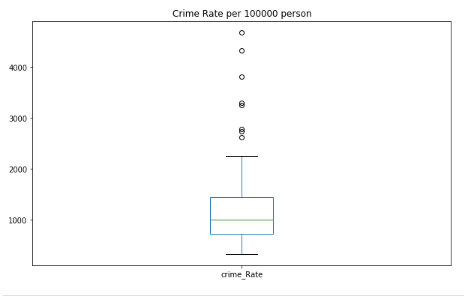


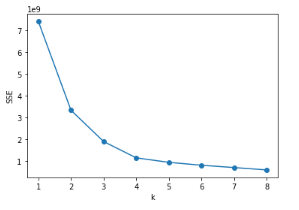


Table and figures above show the distribution of our data.The minimum ,median and maximum value of each factor will help us to distinguish clusters of our data in the data clustering stage.

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：

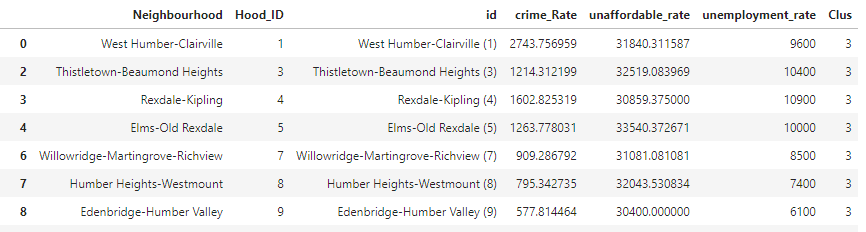


Cluster1：



Cluster2：

Cluster3：



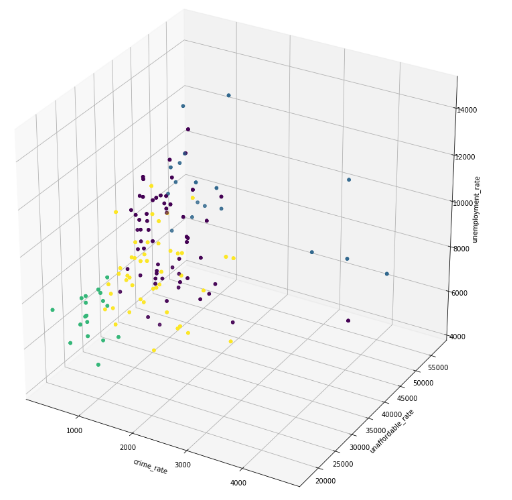
In cluster 0，neighbourhoods have higher on average unaffordable rate，high or low crime rate，and medium to high unemployment rate. The overall performance is not good.

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.

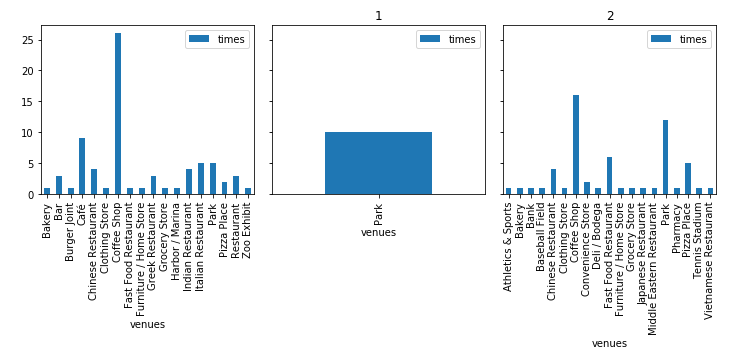
3.4Vennues Exploration

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:



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”

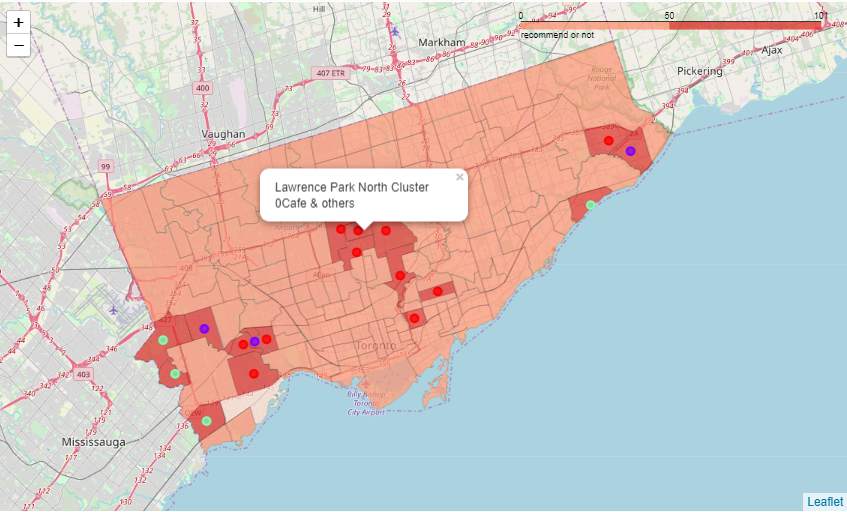
1. **Results**

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:



1. **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.

1. **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.

1. **References**

[1] Wikipedia-List of neighbourhoods in Toronto

[2] Toronto city-government neighbourhood-profiles

[3] Forsquare API

[4] Crime data on Toronto Police website