Clustering of Toronto's neighborhoods based of firearms shootings

Lorenzo Rappuoli

July 16, 2021

1. Introduction

1.1 Background

Toronto is among the ten richest metropolises in the world and its level of wealth is such that it attracts thousands of people from all over the world every year.

Below are its scores:

- Overall rating (out of 100): 97.2.

- Stability: 100.0.

- Health care: 100.0.

- Culture and environment: 97.2.

- Education: 100.0.

- Infrastructure: 89.3.

The city consists of 140 neighborhoods, which differ in population, housing density, and average income of its inhabitants.

The increase in population has caused an increase in housing prices and overcrowding has led to an increase in housing densities, with all the consequences that can be imagined.

1.2 Problem

The increase in population has not been uniform across the city. Some districts have been more affected by the increase in population with direct consequences on elements such as safety and crime rates.

This project seeks to group neighborhoods into homogenous groups that indicate the likelihood of being a victim of a firefight based on data collected by the police and data on population, population density and average income of the inhabitants.

1.3 Interest

Toronto citizens will be interested in knowing how their neighborhoods rank based on risk, as well as those who want to move to Toronto and are looking for housing.

In addition, government and law enforcements will also be interested in knowing the status of the city's neighborhoods.

2. Data acquisition and cleaning

2.1 Data sources

The data sources for the project are the following ones:

- **Demographics of Toronto neighbourhoods**: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

  In this Wikipedia page there is a list of demographic data on each Toronto neighbourhood from the Canadian census. It is useful because it contains the information about population, density and average income for every neighborhood.

- **Shootings & Firearm Discharges:** https://open.toronto.ca/dataset/shootings-firearm-discharges/

This dataset contains all shooting-related occurrences reported to the Toronto Police Service.

The injury Levels are:

1. Death: Where the injured person (as defined above) has died as a result of injuries sustained from a bullet(s).

2. Injuries: Where the injured person (as defined above) has non-fatal physical injuries as a result of a bullet(s).

Data features:

Column	**Description**
_id	Unique row identifier for Open Data database
Event_Unique_ID	Occurrence Number
Occurrence_Date	Date Shooting Occurred
Occurrence_year	Year Shooting Occurred
Month	Month Shooting Occurred
Day_of_week	Day of Week Shooting Occurred
Occurrence_Hour	Hour Shooting Occurred
Time_Range	Time Range Shooting Occurred
Division	Police Division where Shooting Occurred
Death	Count of Deaths Resulted from Shooting-related event
Injuries	Count of Injuries Resulted from Shooting-related event
Hood_ID	Identifier of Neighbourhood where Homicide Occurred
Neighbourhood	Name of Neighbourhood where Homicide Occurred
ObjectId	Autogenerated unique record identifier
geometry	Latitude and Longitude

2.2 Data cleaning

To do this analysis, I chose to use only data for 2020 so that I could give an objective representation of the current situation.

The starting table was from the police data where I selected Occurrence_year=2020 and then grouped the values by neighborhood.

The average gave a good idea of the number of injuries and deaths per shooting and gave a satisfactory result in indicating the latitude and longitude of the neighborhoods.

There were some problems in cross-referencing the results from the web with the police dataset. This was due to the different names of the neighborhoods in the two data sources.

To solve this problem, I preferred to substitute the average of the column values for the NANs, assuming that this choice would affect the final result less than leaving things as they were.

2.3 Feature selection

As features I decided to use the following values from the Toronto Police dataset: Occurrence_year, Death, Injuries, Neighbourhood and geometry.

I then selected the year 2020 and grouped the values by neighborhood. The metric used was the average, which allowed for indices showing the average number of deaths and injuries per shooting.

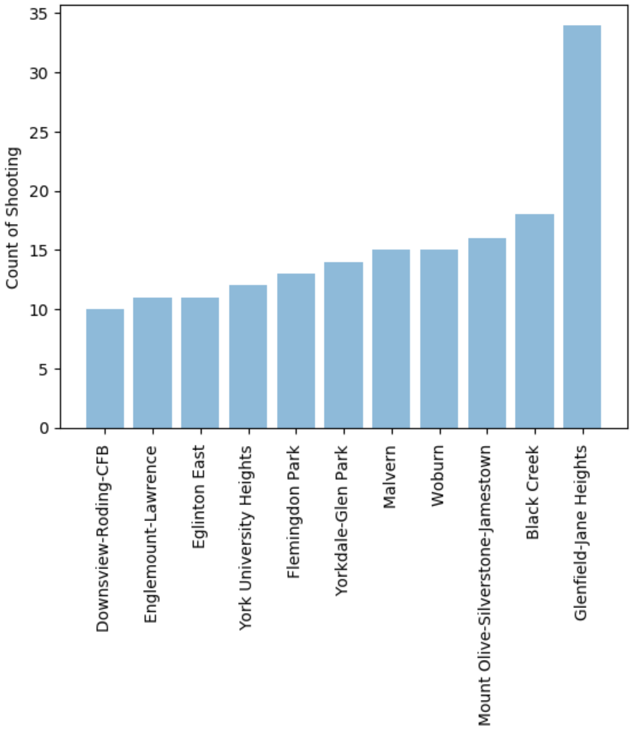
I then added a column with the shooting count for each neighborhood and merged the table with data from Wikipedia.

For those values that were null because the neighborhoods were spelled differently in the two data sources, I averaged them out.

3. Exploratory Data Analysis

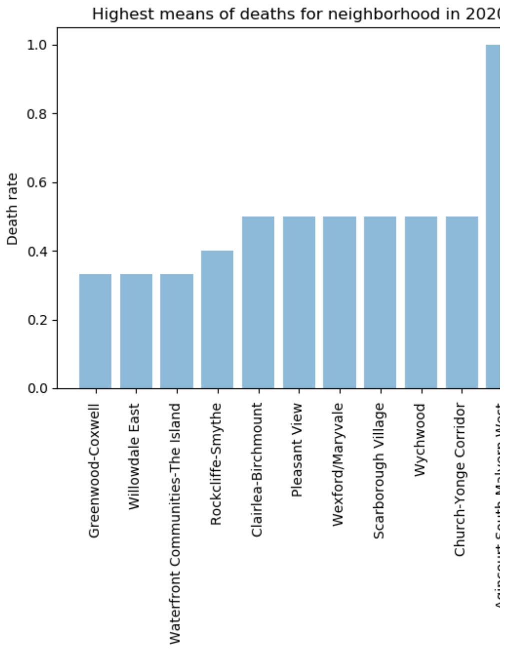
3.1 Counting of shootings per neighborhood

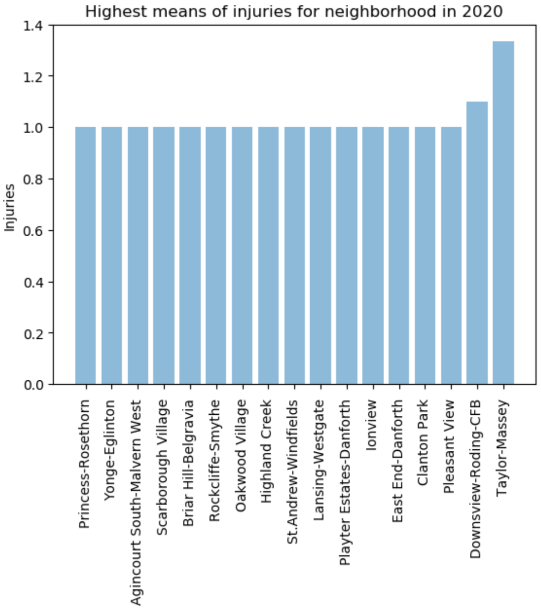
As can be seen, some neighborhoods turn out to have more shootings than others. If we focus the analysis on neighborhoods that had more than ten shootings in 2020, the result is as follows:



From this we can see that only a small fraction of the 140 neighborhoods had a substantial number of shootings, less than 10%.

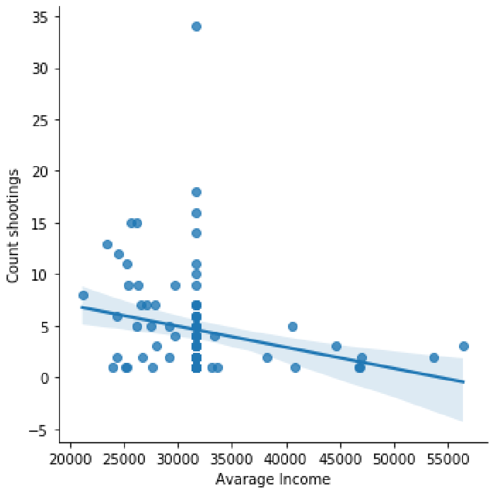
However, if we check the neighborhoods that had the most deaths and injuries, we can see that only a few of the neighborhoods with the most shootings fall into these lists.



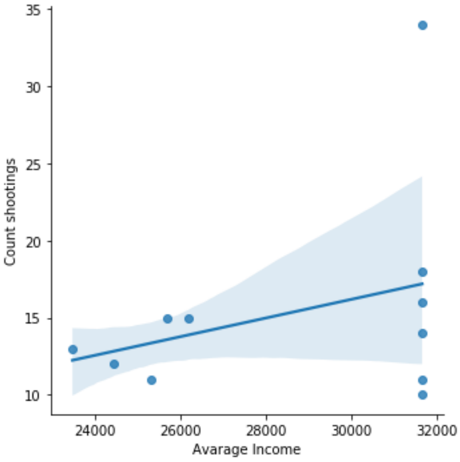


This indicates that more shootings do not always result in more deaths and injuries.

When correlating in number of shootings with the average income of the neighborhood, there is a slightly negative correlation. This perhaps may indicate that poorer neighborhoods could be more dangerous.

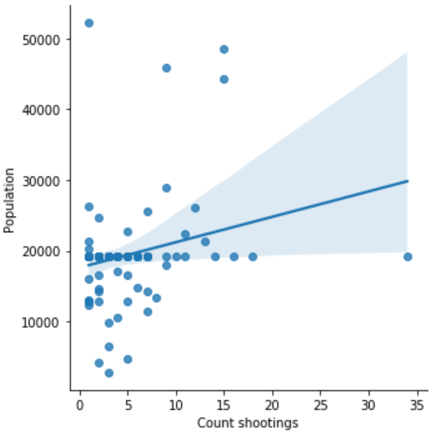


If we limit the analysis, however, to neighborhoods with the most shootings, the correlation becomes positive:



This could indicate that many shootings are related to crimes such as robbery and extortion.

On the other hand, the correlation between the number of inhabitants and the number of shootings is positive: the more people who live in a place, the more likely gun crimes are:



4. Predictive Modeling

In order to group neighborhoods according to their dangerousness, clustering algorithms must be used.

In addition, a classification algorithm could have been added after the clustering algorithm, so as to create a tool to make predictions about which class a given neighborhood would belong to in the future. However, this is a later step in our study.

4.1 Clustering model

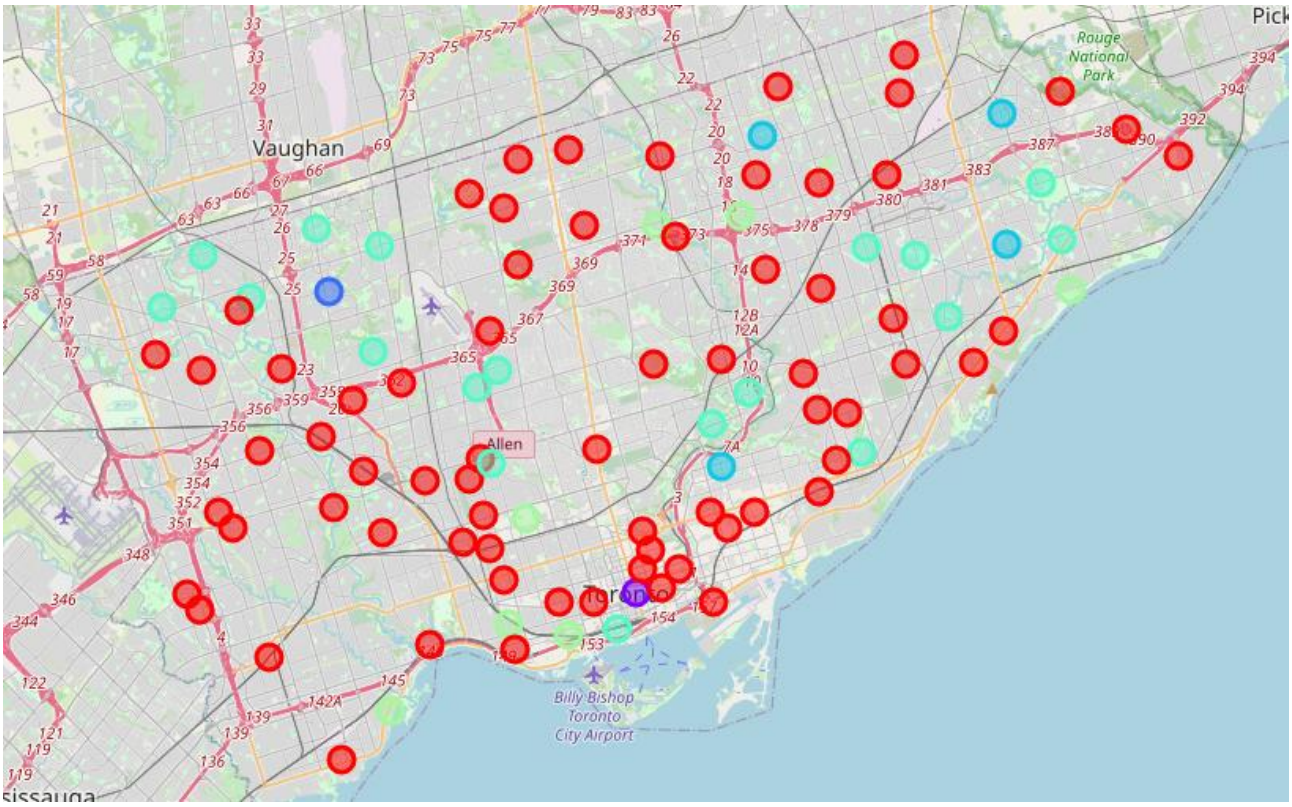
As instructed, I applied the k-means algorithm. Initial analysis suggested that the appropriate number of clusters was 6.

In general, the k-means algorithm is a very efficient iterative algorithm that manages to give very good results with less effort than other clustering algorithms.

It is also particularly well suited for this type of geographic problem because it succeeds in making an immediate response to the reader.

4.1.2 Result of the alghoritm

The result of the algorithm is represented in the following map:



- **Cluster 0 (red):**

  these are all neighborhoods for which we did not have values from Wikipedia and had to use the average. In general, they are neighborhoods with average income but quite quiet, where the number of shootings in 2020 were on average less than 5 with few or no consequences.

- **Cluster 1 (purple):**

  this is the best neighborhood among the low-density, low-income neighborhoods: Bay Street Corridor. The number of shootings in 2020 was 5 but with few or no fatalities.

  Therefore, the neighborhood might be a good choice for those who are looking for an affordable solution.

- **Cluster 2 (dark blue):**

  the Glenfield-Jane Heights neighborhood was by far the worst of all. With 34 firefights in 2020, each of them resulted in either one death or one injury. It still turns out to be an outlier.

- **Cluster 3 (dark green):**

  this cluster contains the neighborhoods where population and density are highest while income per capita is among the lowest. Here shootings are more frequent, as well as higher rates of fatalities and injuries. These are clearly dangerous neighborhoods.

- **Cluster 4 (green):**

  This cluster contains neighborhoods with higher incomes than the city average, but also more shootings and injuries. The population numbers are also quite high, as is the population density.

  While these are not neighborhoods that a person would call risky, they are still places where the danger threshold is high.

- **Cluster 5 (light green):**

  These neighborhoods are some of the quietest and most livable in the city. The housing density is low, the median income high, and the number of shootings, deaths, and injuries borders on zero. They are definitely the best places to live.

5. Conclusions

The results of the study showed that there are varying degrees of dangerousness in the city of Toronto.

Based on the results, the best neighborhoods from a safety perspective were found to be: Bayview Village, Guildwood, Henry Farm, Mimico, Niagara, Roncesvalles, Wychwood.

Among the worst instead are: L'Amoreaux, Malvern, Old East York, Woburn. The worst of all is Glenfield-Jane Heights.

In general, it is not possible to say that the average income and population density are sufficient indices to predict whether a neighborhood is dangerous. But it is possible to say that the higher the average income, the greater the chance of not getting into a firefight.