**An analysis of neighbourhood crime in the city of Toronto**

*Naresh Vempala*

CAPSTONE MILESTONE REPORT

**1. Background and Questions to be answered:**

Toronto is considered to be a safe city in comparison to other big cities. In an article in the Economist (2015), Toronto was ranked as the safest major city in North America and the eighth safest major city in the world, as cited in Wikipedia.

https://en.wikipedia.org/wiki/Crime\_in\_Toronto

Despite being a relatively safer city, Toronto has its fair share of crime. The city consists of 140 officially recognized neighbourhoods along with several other unofficial, smaller neighbourhoods. As is the case with any big city, some neighbourhoods are considered to be less safe than others. Several reasons are attributed to higher crime – lower income, lower literacy and access to education, unemployment leading to illegal drug activity etc. An analysis of crime and neighbourhood data within Toronto will provide us with a good understanding of how many of these assumptions are true and to what degree. It might additionally reveal hidden patterns, trends or relationships between some independent variables and our dependent variable (e.g. major crimes) that would not be obvious.

For this project, I my focus is on crime in Toronto’s 140 official neighbourhoods. I am exploring the following topics in no particular order.

1) Providing a summarized visualization of all the major crimes in Toronto. I might have to do some type of normalization (e.g. divide the number of crimes in a neighbourhood by the population of that neighbourhood).

2) Comparing 3-5 most crime prone neighbourhoods against 3-5 least crime prone neighbourhoods.

3) Finding the difference in these neighbourhoods as regards median household income and education? Which is the most prominent age group of people? Does this in any way affect crime?

4) My dataset provides me with data only for two years – 2008 and 2011. I plan to compare crime data for both years. Has anything changed from 2008 to 2011?

5) What is the neighbourhood with the most change?

6) What could be the reasons for this change? Can the data give us an answer?

7) Finally, I will try to come up with something predictive (using machine learning). This would be more speculative as the data does not afford any type of validation, given that I have only two years of data. But this would still be helpful as a predictive tool in the current climate of ongoing gentrification projects in Toronto since the last 4-5 years, coinciding with a significant increase in the construction of large-to-midsize condominium buildings in various neighbourhoods, undoubtedly changing the social demographics of the city.

**2. Potential clients:**

There are different types of clients that would be interested in the findings from this project.

(i) Canadian/US online and print media that cover socio-economic and urban issues: These clients are magazines that take an active interest in stories that are driven by socially relevant issues and are backed by data analytics, for creating awareness within the public while simultaneously enhancing the quality of their readership.

Canadian magazines such as *The Walrus* and *THIS Magazine*.

(ii) I also anticipate interest from the following clients – Government funded bodies and non-profits offering job placement services, subsidized education services, housing and relocation services for low-income communities and immigrants

**3. Datasets used, data wrangling and approach:**

The City of Toronto has an Open Data portal, which consists of over 200 datasets.

http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=9e56e03bb8d1e310VgnVCM10000071d60f89RCRD

These datasets are organized into 15 different categories. I am using 2 main datasets from this Open Data portal consisting of safety and demographics data for two years – 2011 and 2008. So, there are 4 datasets in total (2 safety datasets and 2 demographics datasets). I might use a 3rd dataset pertaining to economics.

The raw datasets were provided as excel sheets. I converted these into csv files, and imported them as pandas dataframes. I checked to see if there were any missing values in these dataframes. The column names were very long with spaces in between. So, I used a dictionary to rename all the columns into smaller, single-word names. This makes any data operations more manageable.

Both datasets have all the 140 neighbourhoods of Toronto listed alphabetically with an ID for each neighbourhood. The neighbourhood IDs match for the safety and demographics datasets. The safety datasets (the ones I am most interested in), have different crimes distributed by their respective column names for 2011 and 2008. I noticed some differences between the 2011 and the 2008 dataset. There was one column for all major crimes (i.e., sum of 8 different types of crimes) in the 2011 dataset but not in the 2008 dataset. So, I computed this value for the 2008 dataset and added a new column.

I noticed that the demographics dataset for 2011 has 39 columns while the one for 2008 has 85 columns. This was because in 2008 the city of Toronto collected data in each neighbourhood on language and ethnicity of different groups whereas in 2011 the city only collected language data. This could be a challenge for making comparisons between 2008 and 2011. However, the language data should be sufficient to come up with any insights revealed by the data.

I realized while generating some initial plots that population would be a major confounding variable. In other words, a neighbourhood might just have more major crimes occurring because of higher population density. To avoid this, I could look at crime per capita. So, I divided the crime statistics in each neighbourhood with the population of that neighbourhood to compute crime per 1000 people.

One main limitation of this data is that I only have safety data available for two years 2008 and 2011. So, if I intend to come up with a predictive model on crime for later years, I do not have much data. One way of working around this challenge is to ignore the year and just combine all neighbourhoods for both years together to isolate important features that could predict the overall crime for any neighbourhood. In other words, what are the most important or salient contributors of crime in a neighbourhood, irrespective of year?