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

*Naresh Vempala*

**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 will focus only on crime in Toronto’s 140 official neighbourhoods, and explore the following topics in no particular order.

1) I will provide 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) I will compare 3-5 most crime prone neighbourhoods against 3-5 least crime prone neighbourhoods.

3) What is 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. Data 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 will use 2-3 datasets from this Open Data portal consisting of safety, economics, and demographics data. The raw datasets are provided as excel sheets. I will convert these into csv files, and import them as pandas dataframes. Any type of data wrangling, clean up etc. will be done in python using pandas and numpy.

To create a compelling data story, I will likely use visualizations involving scatter plots, box/violin plots, and histograms. Some dataframes will likely be merged and specific columns will be sliced. For the speculative, predictive component, I will be using an appropriate machine learning method from scikit-learn.

**4. Deliverables:**

My deliverables will include (1) a narrative of the project in the form of a paper, and (2) code that will be made available as part of a github repository. I might possibly include a slide deck, time permitting.