BDA 106
CAPSTONE PROJECT - FINAL REPORT

CAN MAJOR CRIME INDICATORS (MCI) BE PREDICTED?

Team 2

[redacted]

Anup Samuel

[redacted]

Contents

EXECUTIVE SUMMARY

A recent report from Ryerson University's Centre for Urban Research and Land Development (CUR) has highlighted the continued growth of Toronto and the surrounding region. Toronto and the Greater Toronto Area (GTA) have been hailed as the fastest growing city across the entire United States and Canada. Out of the four largest metro areas in the U.S. and Canada, Toronto is the only one with population growth recorded during the 12-month period ending in July 2019. Within the GTA, Toronto's population is projected to rise from 2.96 million in 2018 to 4.27 million in 2046, an increase of 44.5 per cent.[1]

With an increase in population, the number of crimes committed in Toronto as measured by the Crime Severity Index, which takes into account the severity and number of crimes committed, has exhibited a steady upward trend from 45.07 in 2014 to 53.64 in 2018.[2]

Toronto Police Services (TPS) is the largest municipal police service in Canada and is the primary agency responsible for providing law enforcement and policing services in Toronto. It has the onus of providing a trusted, community focused policing in Toronto. It has engaged with and has been inclusive of the diversity of the city and has evolved continually to meet the changing needs of the city.

This report attempts to establish a relationship between different demographic attributes and Major Crime Indicators (MCI). The report has also tried to identify patterns and insights on criminal behaviour and identify major crime hotspots within Toronto. A comparative analysis of crimes between Toronto and other major cities in North America has also been included in the report and discussed with the Analytics and Innovation Department of the TPS.

The majority of analyses are based on data for major crime indicators and Toronto open source demographic data. A well-structured and comprehensive dataset was derived using these data. Several statistical and machine learning tools such as correlation, feature selection, Lasso regularization etc. were employed for analyses. The report consists of several visualizations highlighting different crime trends.

The number of inhabitants in specific neighbourhoods plays a key role in the prevalence of crime. Our statistical analyses indicate positive correlation between increasing population and crime rates, as well as negative correlation between average neighbourhood income and crime rates.

Based on analysis of crime data, assault cases per capita are significantly higher in Toronto compared to a similar city such as New York. Additionally, the occurrence of homicides- especially by shooting- in Toronto exhibits a strong positive upward trend.

PROBLEM DESCRIPTION

There has been a consistent and gradual increase in crimes in Toronto as exhibited by the increasing Crime Severity Index (CSI). The answer to conventional police service delivery is not limited to continuously increasing public funding. Using data analytics and by leveraging technology, we can enhance the ability of police services to respond to crimes encountered daily without incurring additional costs. This also allows police services to judiciously assign increasingly limited police resources.

This report presents a comprehensive analysis to identify Major Crime Indicators and establish relationships with demographic features such as population, age, income etc. The report also includes a comparative analysis of crime indicators vis-a-vis other major North American cities with similar demographic features such as New York, Los Angeles, and Chicago. In addition to these, we have also mapped MCI data to crime hotspots in Toronto.

DATA DESCRIPTION

Considering the sensitive information in crime and demographic databases, we have been working with various open-source datasets. The primary datasets we used were the crime datasets released by TPS in their public safety data portal.[3] These datasets cover various crime data like MCI, Homicide, Traffic etc. The datasets are clean and easy to use. The datasets also include information about the time, date, weapon type, neighbourhood, and latitude and longitude of incidents of crime. Most of these datasets cover crime data from 2014 - 2019. The MCI database has around 200,000 rows of data.

In order to link the crime data with the demographic information, we were able to use the neighbourhood level demographic dataset available in Toronto's open data website.[4] This dataset is extensive in its scope and covers various facets of demographic data like population, income, education, housing etc. for the 140 neighbourhoods of the city. By using the common neighbourhood number, we were able to join the MCI and the neighbourhood datasets and understand more about the role of demographic features in crimes.

For the comparative analysis of crime numbers between Toronto and other cities, we used the open data found in the websites of the cities and the police departments of New York, Chicago & Los Angeles.[5]

DATA PREPARATION

The TPS data was well-formatted and did not have data quality issues. There was no necessity to perform data cleaning on the TPS data. The rows with null values were dropped from all open-source city databases and no values were replaced.

Pearson correlations were obtained between various demographic variables and major crime indicators to highlight which variables would be most predictive of MCI. The following table shows the demographic attributes with strong Pearson correlation coefficients ($r > 0.65$) with MCI:

Variable Name	Correlation Coefficient (r)
Working Age (25-54 years)	0.74
Total - Highest certificate, diploma or degree	0.71
Population	0.69
Persons living alone (total)	0.69
Total - Immigrant	0.68
Total - Low-income status	0.68

Table 1 – Correlation coefficients of key demographic attributes

Many of the variables highly correlated with MCI were found to be also highly collinear with each other, such that they did not provide much useful information independently. This issue highlighted the need for the use of other feature selection methods for determining the most important demographic attributes in relation to prediction of MCI.

To get a better understanding of which variables are most important for prediction of MCI, we performed backward elimination, recursive feature elimination and lasso regularization in Python. The features selected by the backward elimination and recursive feature elimination methods are shown in the following table:

Backward Elimination	Recursive Feature Elimination
Population Working Age (25-54 years) Seniors (65+ years) Average household size Persons living alone (total) Total -Income statistics Total - Immigrant Total - Highest certificate, diploma or degree	Population Working Age (25-54 years) Seniors (65+ years) Average household size Total -Income statistics Total - Immigrant Total - Highest certificate, diploma or degree

Table 2 – Key attributes based on machine learning techniques.

The backward elimination and recursive feature elimination methods both selected similar features as having the most importance in prediction of MCI.

Lasso regularization works by assigning a penalty value of 0 to features it deems irrelevant and dropping them from the model. All features with a value greater than zero are considered to be important and all features with values lower than zero are considered unimportant in prediction of MCI. The following figure shows the result of the lasso regularization method:

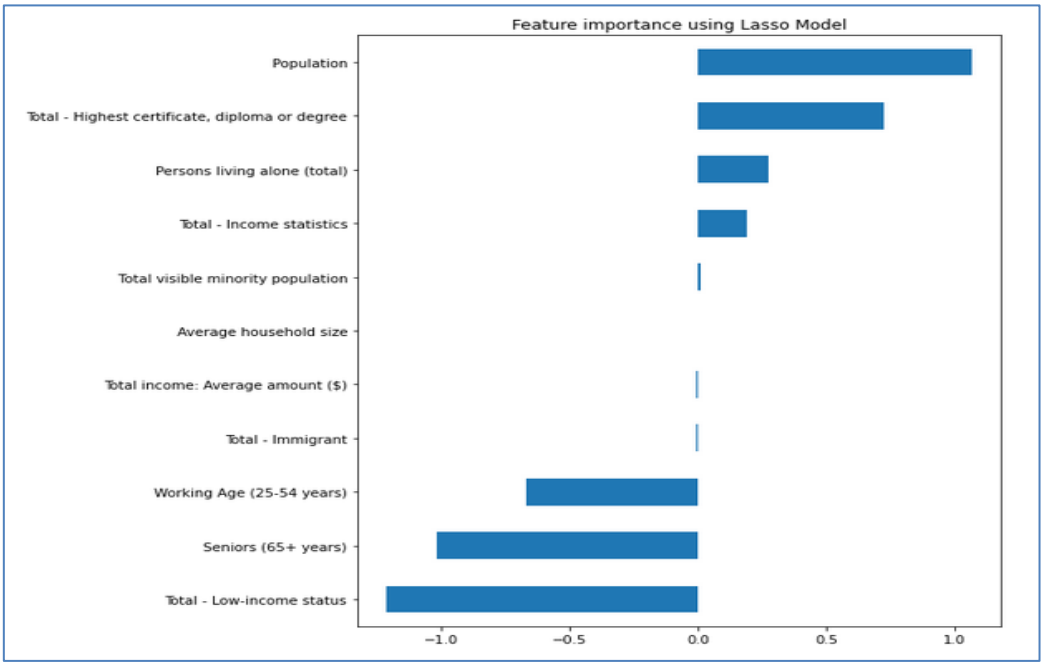

Figure 1 – Key demographic attributes using Lasso regularization

The lasso regularization method selected fewer features as important in prediction of MCI than both backward elimination and recursive feature elimination, and had an accuracy score of 78%. Some of the features that are highly collinear with each other such as number of working age persons and seniors with population, and total low-income status with total income statistics are considered insignificant by the lasso method. The variables selected by the lasso regularization method are therefore more informative

than both variables with the highest correlation coefficient with MCI, and variables selected by backward elimination and recursive feature elimination.

DATA ANALYSIS

COMPARATIVE ANALYSIS

The main aim of this part was to compare the crime statistics between Toronto and other major North American cities. This relational analysis would help in understanding the trend of crimes in Toronto and whether similar patterns can be observed in other cities having similar demographics.

Since the 4 cities considered (Toronto, New York, Chicago & Los Angeles) have different populations and therefore different crime numbers, we have compared the cities on a crimes per capita basis (total number of crimes divided by the population of the city).

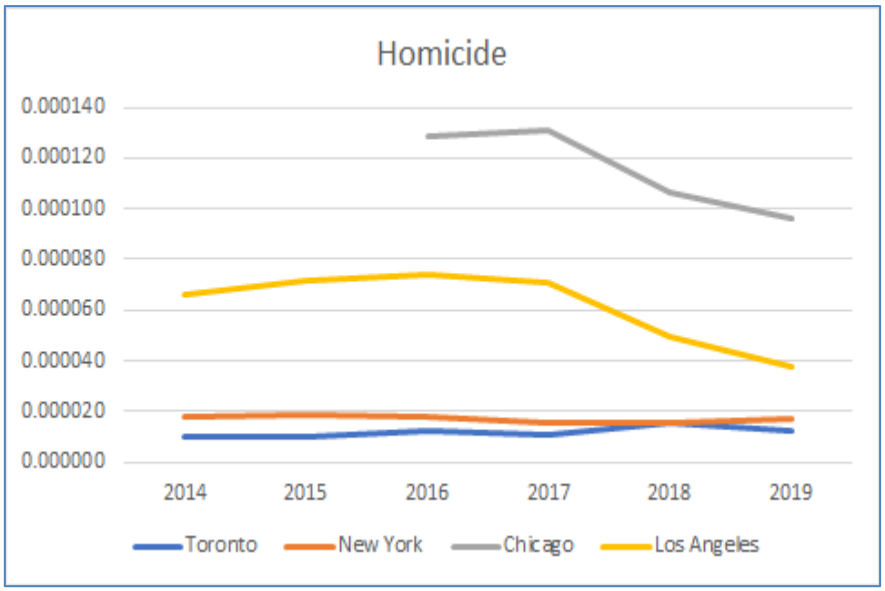

Figure 2 - Per capita count of homicides for the 4 cities

From figure 2, we can observe that the homicide count for Toronto has been relatively constant from 2014-2019 and the per capita homicide numbers are similar to that of New York. The per capita homicide numbers for Chicago vary from around 5 to 6 times that of Toronto.

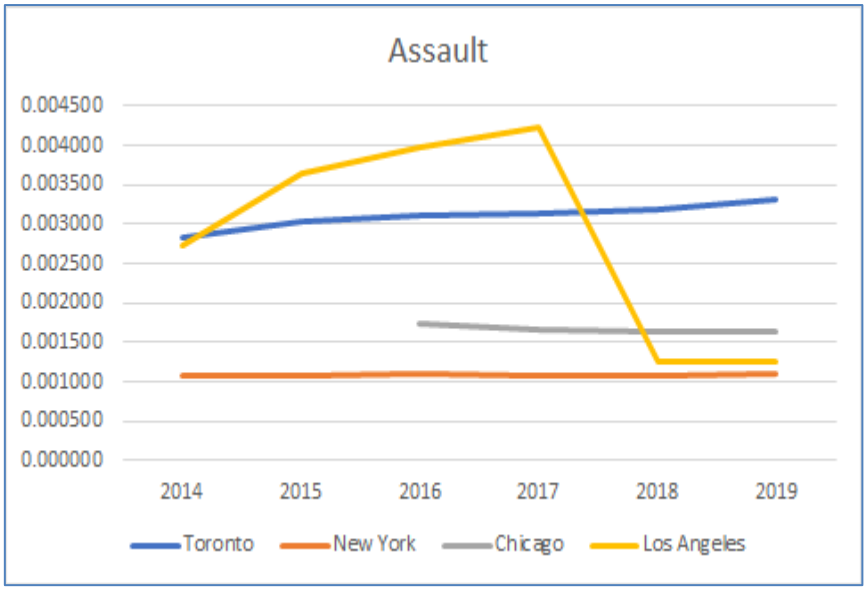

Figure 3 - Per capita count of assault cases for the 4 cities

Figure 3 gives the per capita count of assault cases. In 2019, the per capita number of assault cases in Toronto was more than 3 times the number of New York City and the upwards trend's continuation is worrying.

An interesting outcome of this comparative analysis was trying to understand the drastic fall in crime numbers for the city of Los Angeles. Across all the 4 categories (Homicide, Assault, Robbery & Auto Theft), there has been a significant drop in crime numbers during 2018 and 2019. According to the city's police chief, neighbourhood level community policing is a major factor for this.[6]

VISUALIZATION OF MCI DATA

The aim of this section was to analyze the Major Crimes Indicators dataset of the Toronto Police Service and use it to understand various trends in crime numbers. By analyzing these trends, we can provide suggestions which in turn help in making better decisions.

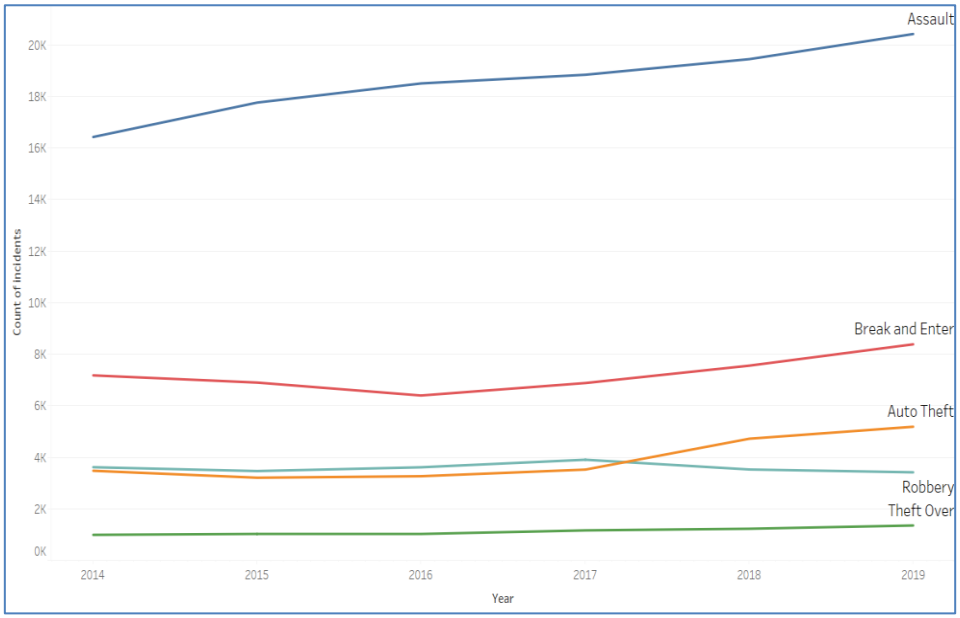

Figure 4 - Total crime numbers by MCI category (Toronto 2014-2019)

From figure 4, we can observe that Assault is the most prominent category of MCI with the highest number of cases and the number of cases has been steadily increasing. Except for robbery, every other category of MCI has seen an increased or constant number of cases.

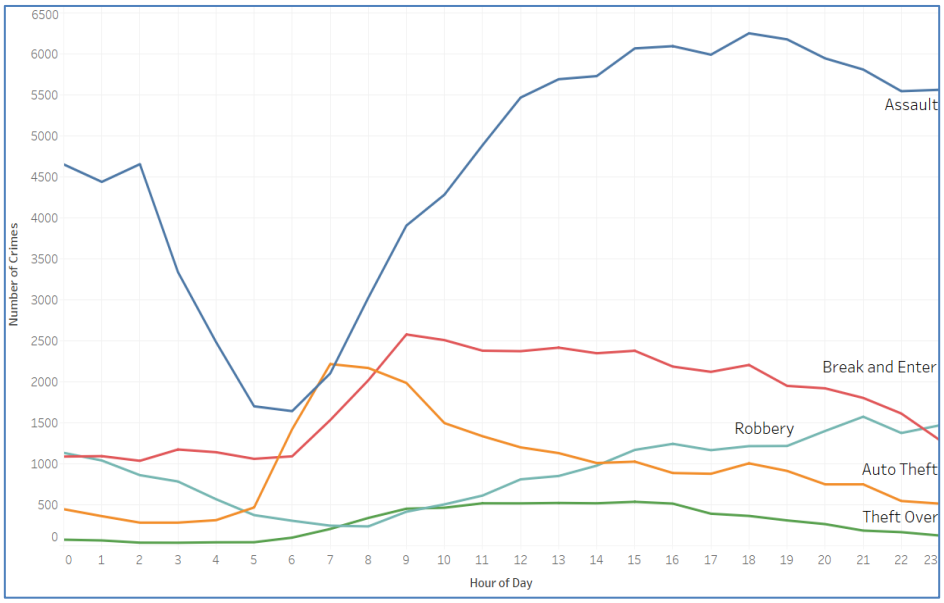

Figure 5 - Distribution of crimes by hour of day (Toronto 2014-2019)

Figure 5 shows how the counts of different crime categories vary by the hour of the day. This information helps to make better decisions regarding crime prevention and allocation of police resources. There seems to be a drastic drop in the number of assault cases in the early morning hours (between 2 to 7 AM). This could be due to fewer people on the streets. At the same time, there is an increase in cases of auto theft during these hours.

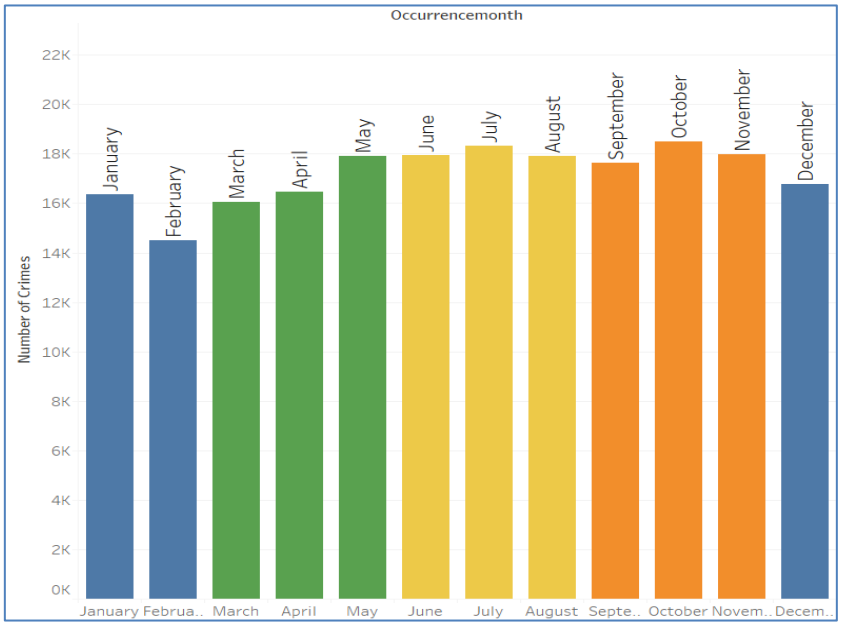

Figure 6 - Seasonal distribution of crimes (Toronto 2014-2019)

To understand the seasonal distribution of crime, we plotted figure 6, which gives the total crime count for each month across all categories. As can be observed, the number of crimes tend to peak during late spring and summer and tend to be at its lowest during the winter months.

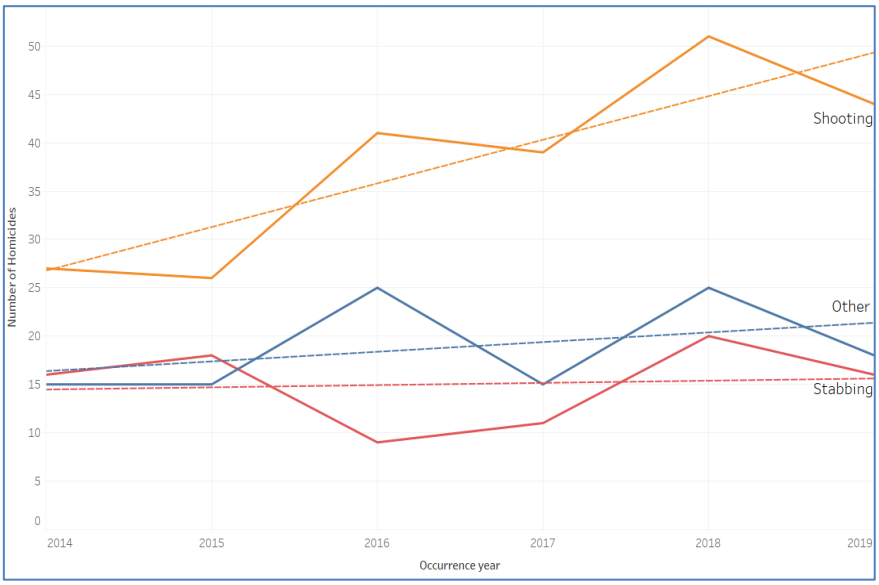

Figure 7 - Distribution of count of different homicide categories (Toronto 2014-2019)

Controlling the number of homicides is an essential part of policing. This is where figure 7 helps in visualizing the trends in homicide numbers based on the type of weapon used. The figure clearly indicates that there has been a steady increase in the number of homicides due to shooting. The number of stabbing cases have remained constant while homicides due to other means (like poisoning, blunt force trauma etc.) have also steadily increased.

HOT-SPOT MAPPING OF CRIME DATA

An important aspect of analyzing crime data is to map the crime numbers to different neighbourhoods in order to understand the count as well as annual trends over the years. This will help decision makers to make informed decisions regarding new strategies and allocation of police resources.

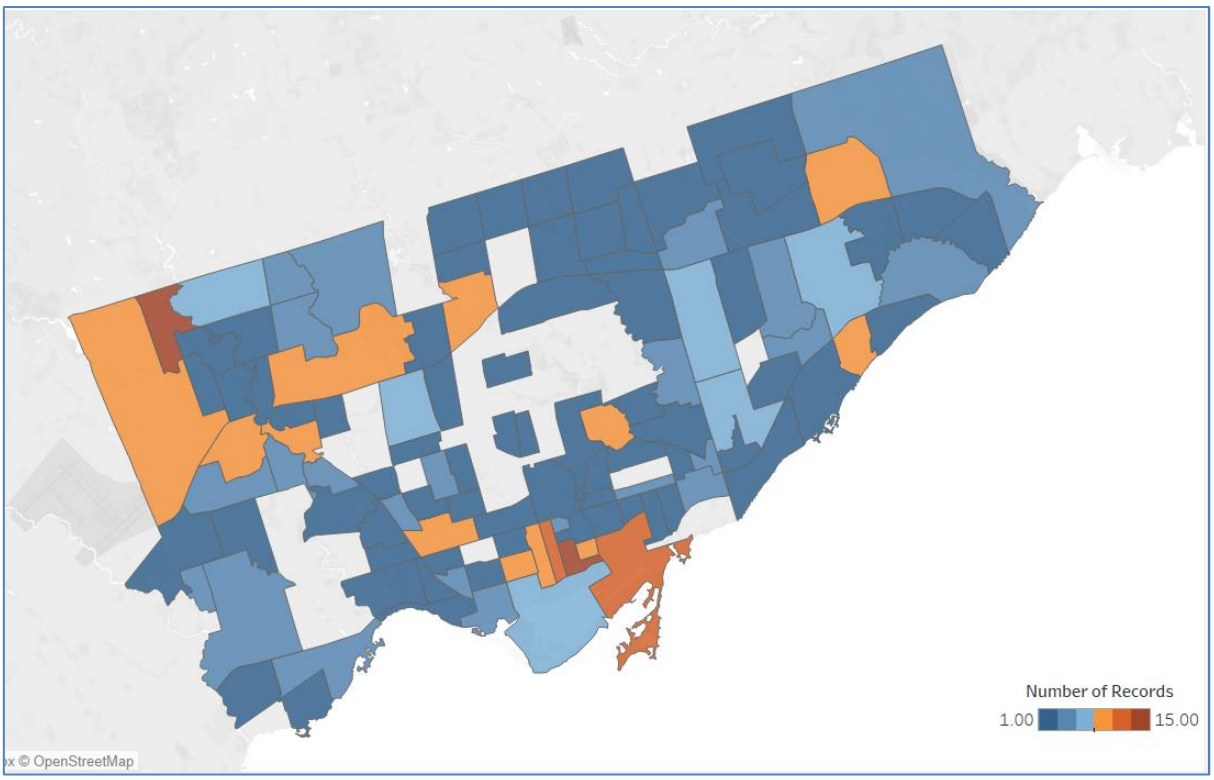

Figure 8 - Neighbourhood level distribution of homicides (Toronto 2014 - 2019)

The visualization above, created using Tableau, shows the neighbourhood-wise distribution of homicides in Toronto from 2014-2019. As can be clearly seen, most of the cases were either in the western neighbourhoods (West Humber, Mount Olive, Kingsview Village etc.) or downtown (South Riverdale, Regent Park, Moss Park etc.).

The final part of the project was the creation of an interactive dashboard (using Tableau) to show the hot-spot mapping of crimes throughout Toronto for the years 2014-2019. This dashboard can be filtered based on MCI category, year of occurrence and month of occurrence. This can be used to observe both the number and trends of crimes occurring in each neighbourhood and is a valuable tool in framing and implementing better policing decisions.

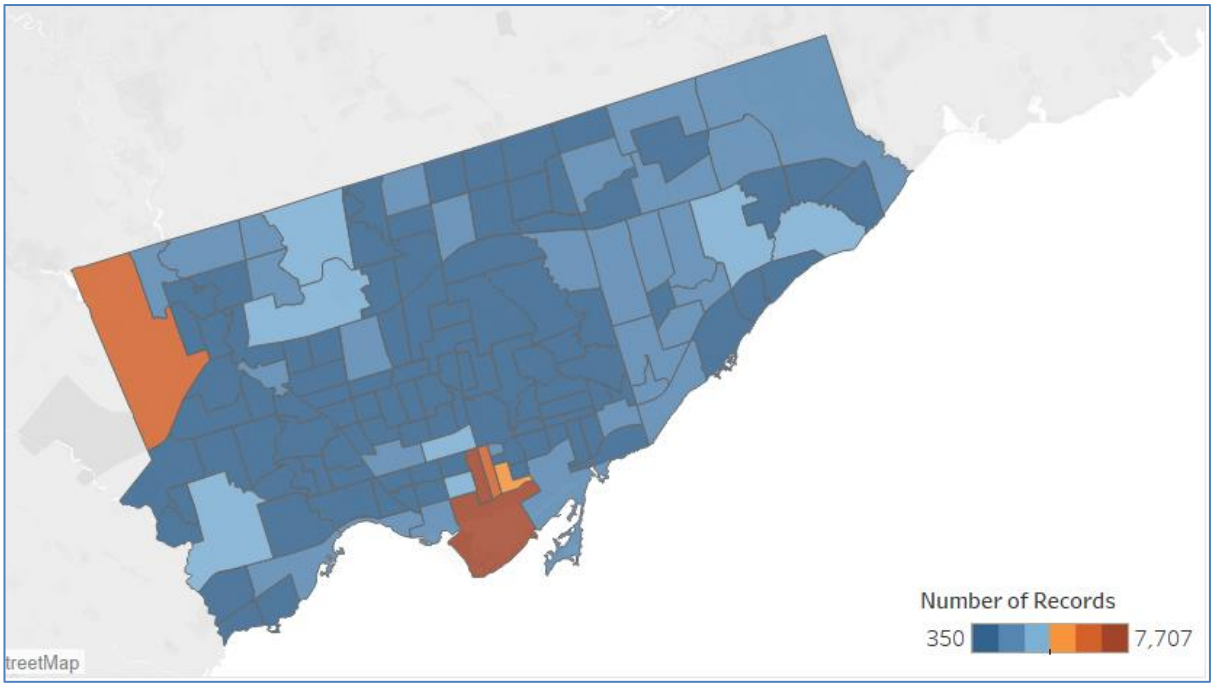

Figure 9 - Neighbourhood level hotspot mapping of total crimes (Toronto 2014-2019)

Figure 9 is a screenshot of an interactive dashboard which can be used to determine the total number as well as trends over time for various MCI categories in the 140 neighbourhoods of Toronto. The figure has colour coded each neighbourhood and it is clearly visible that the largest number of crimes occurred downtown (Waterfront Communities - The Island) with 7707 cases between 2014 and 2019 as well as in the western neighbourhood of West Humber - Clairville with 5680 cases.

CONCLUSION

Two main demographic attributes that have been selected as important in all feature selection methods are population and income. Population is highly positively correlated with crime, such that the neighbourhoods with a higher population have a higher prevalence of crime. Income is weakly negatively correlated with crime, such that neighbourhoods with a lower average income had a higher prevalence of crime. These results indicate that police resources should be concentrated more in neighbourhoods with higher population, and to a lesser degree, to neighbourhoods with lower average resident income.

Crime rate in Toronto is lower during the colder months and is higher during the warmer months. The count of assaults, which dominates crime cases in Toronto, is highest during the afternoon to evening hours and is at its lowest during the early morning hours. The number of homicide cases involving shootings has seen a sharp increase in recent years. The TPS can take steps to curtail gang activities and work towards preventing the proliferation of illegal firearms within the city to reduce the number of shootings.

Using the interactive MCI dashboard shown in figure 9, we were able to observe the trends over time in crime data for different MCI categories. This analysis can be used to suggest police resource allocation to combat specific challenges.

APPENDIX

REFERENCES

1. https://urbantoronto.ca/news/2020/06/toronto-now-fastest-growing-city-and-metro-area-us-and-canada
2. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510002601
3. https://data.torontopolice.on.ca/pages/open-data
4. https://open.toronto.ca/dataset/neighbourhood-profiles/
5. https://opendata.cityofnewyork.us/
6. LA crime dropped in 2019, for the 2nd year in a row
7. https://home.chicagopolice.org/wp-content/uploads/2020/07/1_PDFsam_CompStat-Public-2020-Week-28.pdf
8. http://assets.lapdonline.org/assets/pdf/cityprof.pdf

ADDITIONAL VISUALIZATIONS

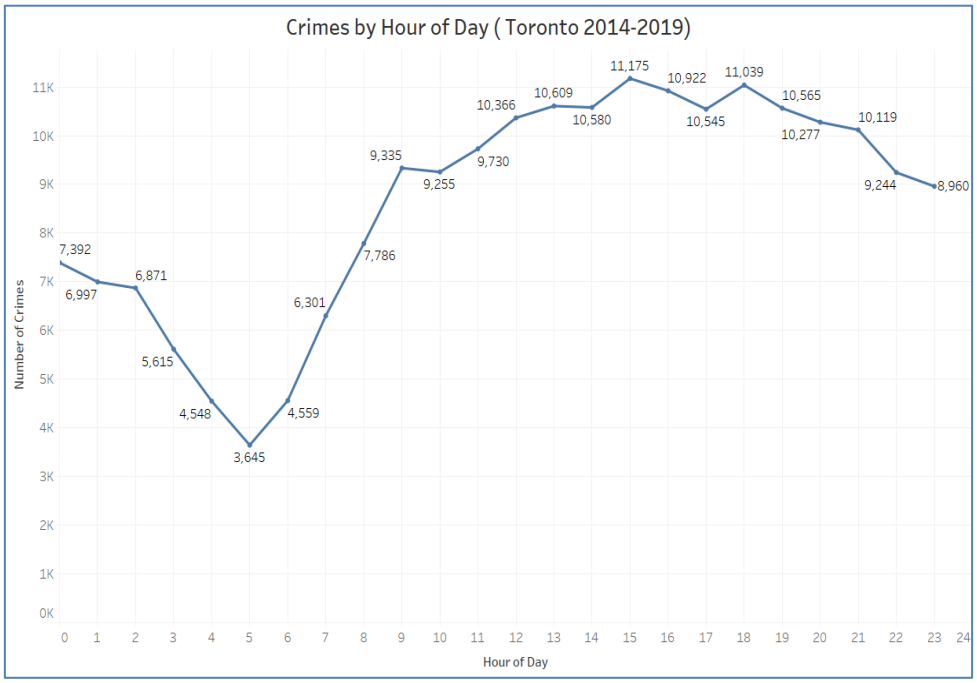

Figure 10 - Total number of crimes by hour of day

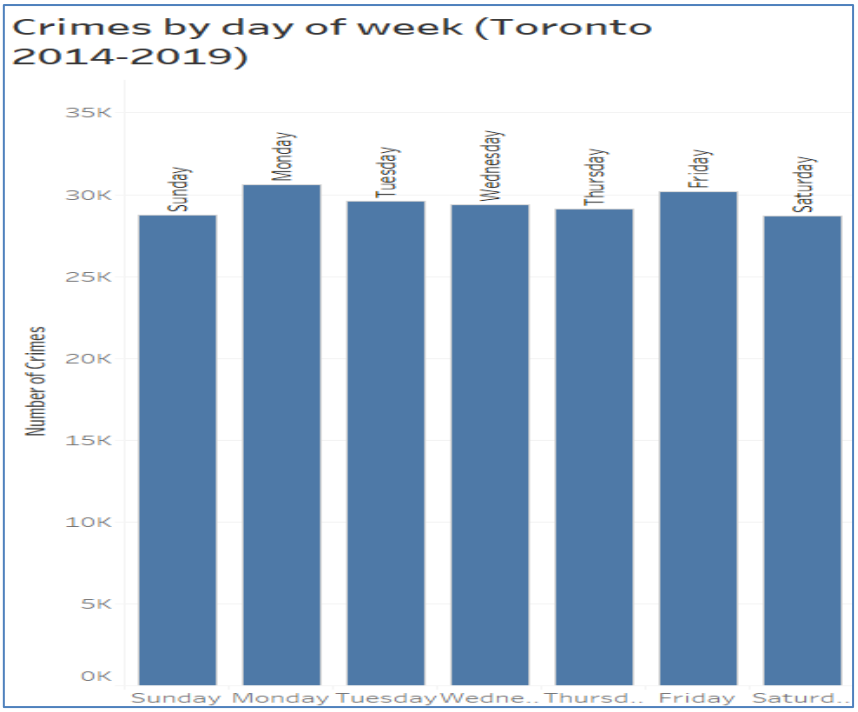

Figure 11 - Total crime numbers by day of the week

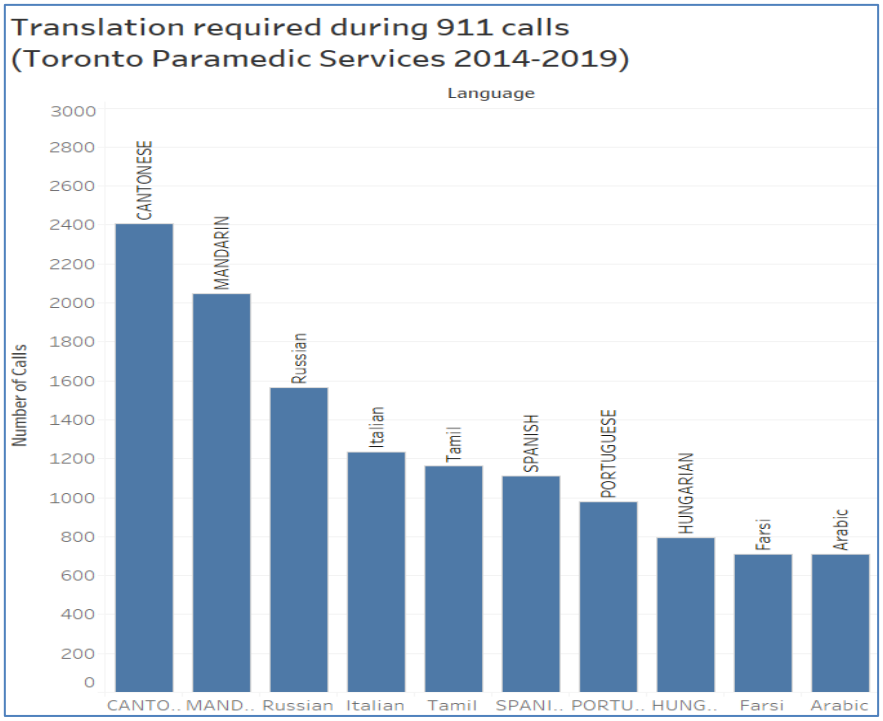

Figure 12 - Number of 911 calls requiring translation