SENG 550 Final Project: Patrol Optimizer

Samuel Tomecek
University of Calgary
Calgary, Canada
samuel.tomecek@ucalgary.ca

Ajay Arumugam 30113455
University of Calgary
Calgary, Canada
ajay.arumugam@ucalgary.ca

***Preamble*—**Contribution of team members: Ajay Arumugam (50%), Samuel Tomecek (50%). Ajay Arumugam and Samuel Tomecek declare that we contributed equally towards this project. Link to repository of our project:

https://github.com/AjayArumugam07/PatrolOptimizer-BigDataAnalysis

***Abstract*—**The Toronto Police Service (TPS) is the primary law enforcement agency in Toronto, Ontario, Canada. This report presents a comprehensive analysis of crime patterns in Toronto, focusing on the application of the K-Means Clustering Algorithm to identify distinct crime clusters. The crime data is organised into 6 different clusters and visualised on a map. Subsequent analyses include temporal patterns of Major Crime Indicators (MCI) offences, aiding officers in scheduling and resource allocation. Additionally, the distribution of crime categories across clusters is examined, providing localised insights for law enforcement strategies. The findings presented in this report serve as a valuable resource for the Toronto Police Service to enhance the effectiveness of crime prevention and response strategies tailored to specific geographic and temporal contexts within the city.

I. Introduction

The Toronto Police Service (TPS) is the primary law enforcement agency in Toronto, Ontario, Canada, serving a diverse urban population of around 5 million people. With crime in the city increasing every year, patrolling the city can be highly optimised by analysing historical crime data. We obtained a historical crime dataset from the Toronto Police Webpage containing all Major Crime indicators by reported date and related offences reported between 2014 to 2022. It includes many fields such as, date and time of crime, location of crime, offence type, etc.

There are quite a few existing projects that analyse Toronto Crime data and highlight general patterns about various neighbourhoods in the city. Our project takes a unique approach of using our analysis to help the Toronto Police Service patrol the city more efficiently. Some of the data analytics questions we seek to answer with this project are the following: How can the city be divided into jurisdictions for the Police to patrol? When and where should police presence be increased? What types of crime are prevalent in the different jurisdictions? We decided to analyse this data using a K-Means Clustering Algorithm and various data visualisations in order to find patterns that could be useful for the Toronto Police Service to patrol the city. We decided to analyse this data using a K-Means Clustering Algorithm and divide the city into 6 jurisdictions. Various data visualisations were performed in order to find patterns across jurisdictions and answer our data analytics questions.

II. Background and Related Work

A. Technical Background

To understand the report, a good grasp of Unsupervised Machine Learning Algorithms like the K-Means Clustering Algorithm would be helpful.

B. Related Work

A couple other projects also analyse the Toronto Crime Dataset. Here is a link to a project from Susan Li that is relevant to our work:

https://towardsdatascience.com/exploring-clustering-and-mapping-torontos-crimes-96336efe490f.

In this project, Susan analyses major crime trends in Toronto by neighbourhood. This study can be especially useful to newcomers to Toronto who are looking for a neighbourhood to live in. Our project takes a different twist, and uses the data to help the Toronto Police Service better patrol the city.

III. Methodology

A. Project Setup

The project was run via Jupyter Notebook set up in a GitHub project. Our project has a basic file structure. The jupyter notebook was in the root of the file system, with a folder that would contain any resources, including the dataset. There was a second folder that would contain the cleaned dataset. The following python packages are required to run the project: pyspark, scikit-learn, matplotlib, pandas, folium and seaborn.

B. Data Cleaning

Before any processing, our data needed to be cleaned to avoid any errors and false data. Firstly, we imported the data and read it to a variable. We then decided to first drop any fields that we did not need for our application. Any remaining columns were renamed to be easier to work with in the future. The fields that were chosen to be kept for our application were dayOfTheWeek, hourOfTheDay, offence, latitude and longitude It was noticed that the days of the week sometimes had a trailing whitespace that could affect our sorting later in the project, so those were removed. Any entries in our dataset that lacked data for latitude and longitude were found to be 0. Our clustering needs valid longitude and latitude data, so any entries with data values of 0 were removed. Next, the hour of occurrence was saved as an integer from 0-23 representing the 24 hr clock. Any entries that had an invalid hour were removed. Finally, in the case that any days of the week were invalid, any entries that did not match our list of days were removed.

C. Cluster Preparation

Since our data had coordinate data of the crime, we decided to run K-Means Clustering Algorithm which is an unsupervised machine learning algorithm that groups similar data points into clusters. In the context of our project, it is used to figure out the major clusters, or hotspots of crime in Toronto. Since it is hard to visualise and determine the number of clusters to use, we decided to use the elbow technique to determine the hyperparameter k. To use this technique, the Within Cluster Sum of Squares needs to be computed for each k value. The WCCS values of k from 1 to 19 have been plotted in the graph below. Since the graph seems to level out after 6, we decided k=6 would be the most optimal number of clusters for our data.

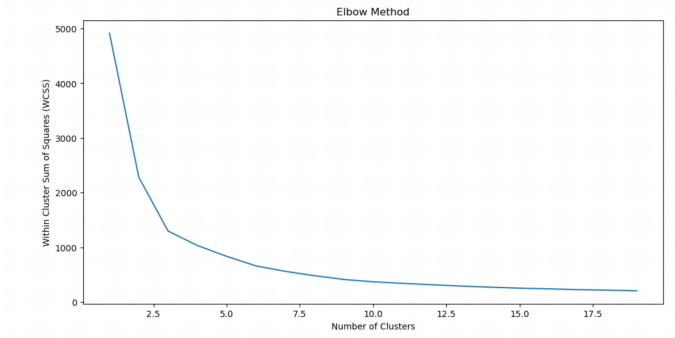


Figure 1 - Elbow Method Results

D. Cluster Model

In order to fit the data into the KMeans model, the data columns were first scaled using the StandardScaler. The output of the StandardScaler was then used to fit the KMeans model with k=6 determined earlier. The final centroids of all 6 clusters after running the model were then obtained and plotted along with the other data points.

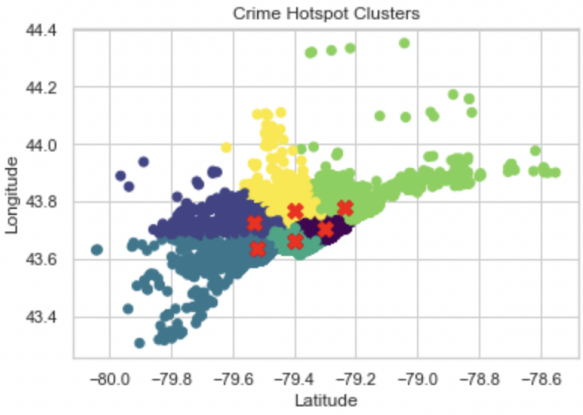


Figure 2 - Positions of MCI Data Sorted by Cluster

We also overlaid the above graph on top of the map of Toronto using the Folium geoplotting library. It is an interactive graph that allows you to zoom in and out to view the different data points in each cluster. The centroids of each cluster are shown with a red marker. Since the folium map is very large, we could not push the output to Github. However, the code in our repository can be run locally to view the folium map and interact with the data.

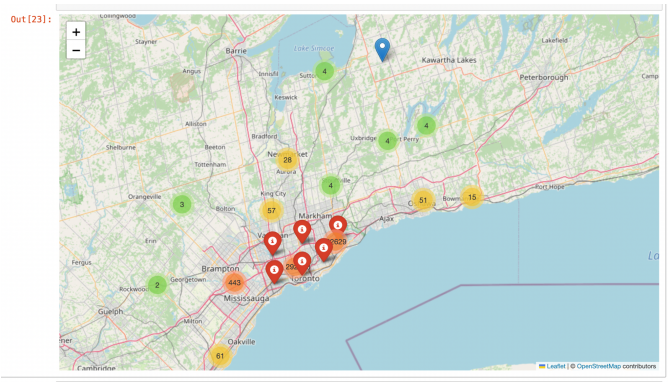


Figure 3 - Folium Data Representation Zoomed Out

When zoomed in, the markers can be clicked to identify what cluster they are a part of.

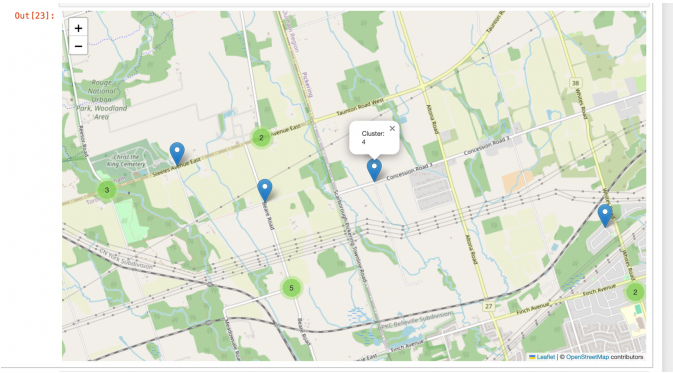


Figure 4 - Folium Data Representation Zoomed In

E. Visualisation of Cluster Data

After the data was classified into separate clusters that would each have a patrol unit, analysis was done to find how each cluster differs from one another and all the data as a whole. Firstly, we created a plot for each day of the week for each cluster. This resulted in 42 graphs that displayed the number of crimes per hour of the day for each day of the week in each cluster. These graphs would help officers to identify any trends of when MCI offences are most likely to occur throughout the day for a given cluster location. Another set of 7 was created for all 6 clusters combined, ie. the total dataset. Officers would be able to compare a specific day of the week in their cluster to the entire Toronto Police dataset. Only the charts for the total data and cluster 0 are shown. The charts for clusters 1-5 are generated and are in the jupyter notebook.

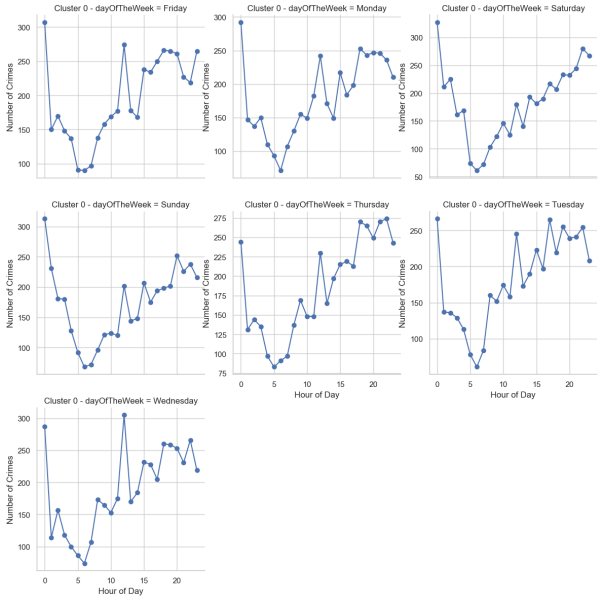
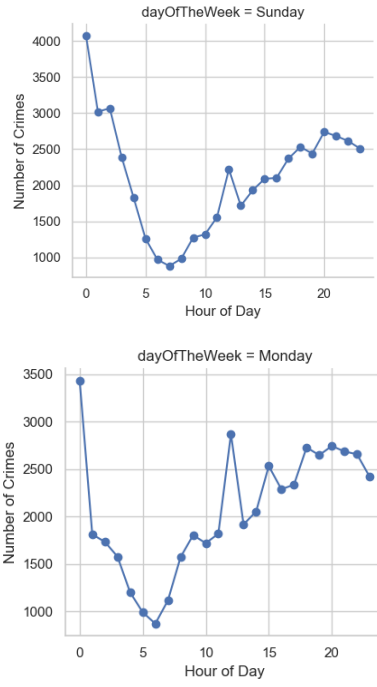
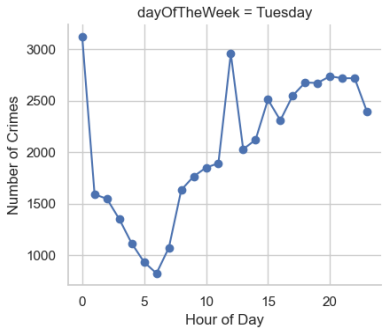
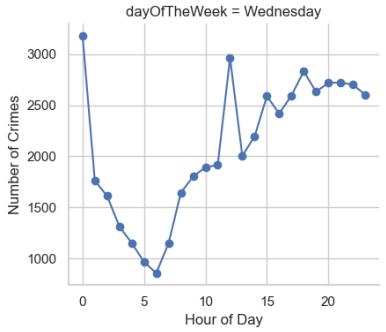
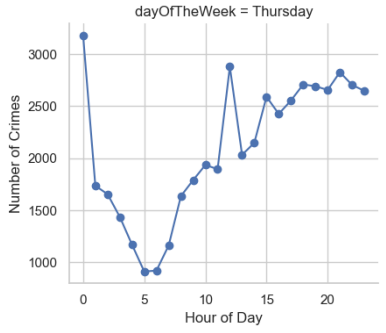


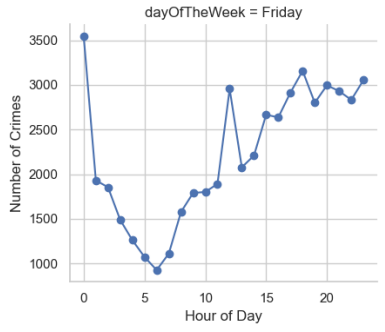
Figure 5 - MCI Offences Per Hour Per Day For Cluster 0











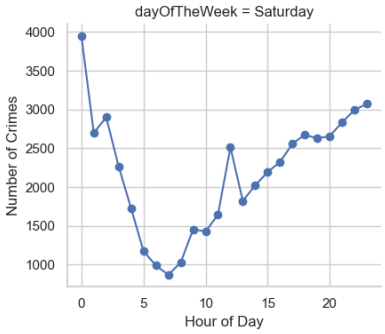


Figure 6 - MCI Offences Per Hour Per Day For All Data

We then created a combined graph for each cluster that showed all 7 days for each cluster on the same graph. This would help officers compare days of the week for their cluster to create a schedule that could apply to multiple days at once. Only the chart for cluster 0 is shown. The charts for clusters 1-5 are generated and displayed in the jupyter notebook.

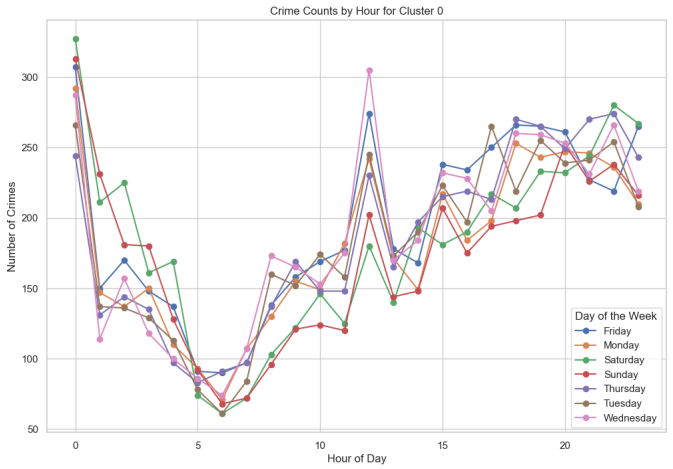


Figure 7 - Crime Counts by Hour for Cluster 0

Next, a distribution of crime offences for the entire dataset was created. This is an average of all six clusters and can be used as a comparison tool for the cluster data. It is a pie chart showing the proportionality of all MCI categories. Next, we created more pie charts for each of the clusters. Now officers can gain a greater insight to their specific cluster and learn

what crimes they should expect, and when compared to another cluster, or the entire city, what MCI offence is more or less likely to occur. Only the charts for the total data and cluster 0 are shown. The charts for clusters 1-5 are generated and are in the jupyter notebook.

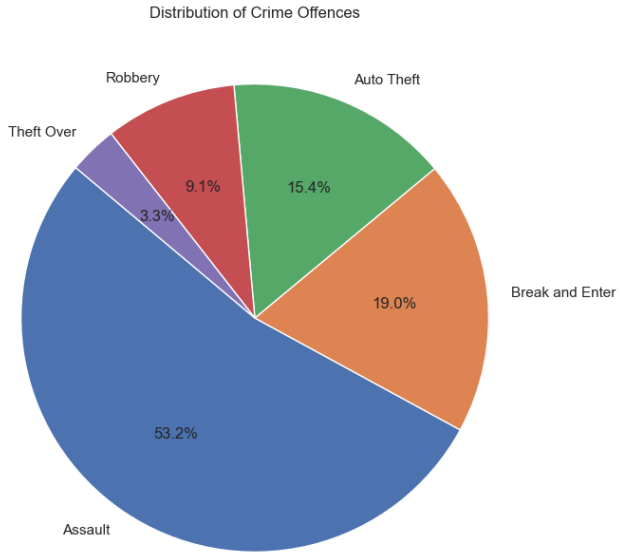


Figure 8 - Distribution of Crime Offences for All Data

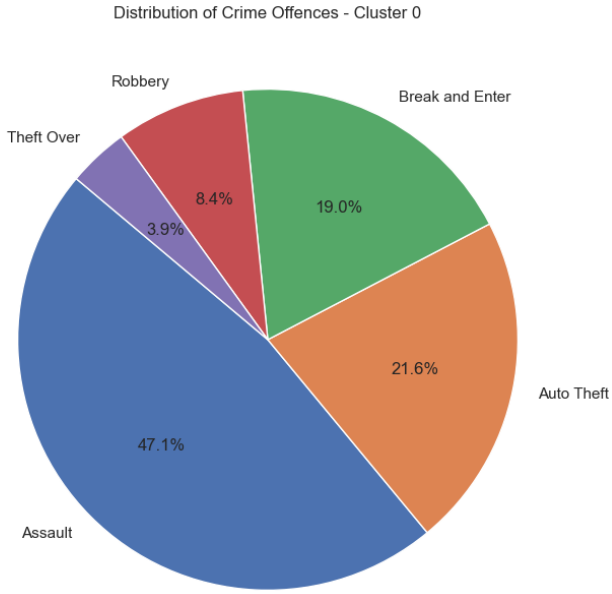
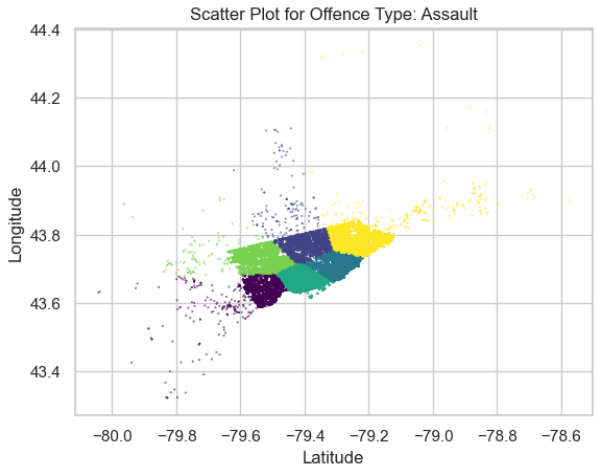
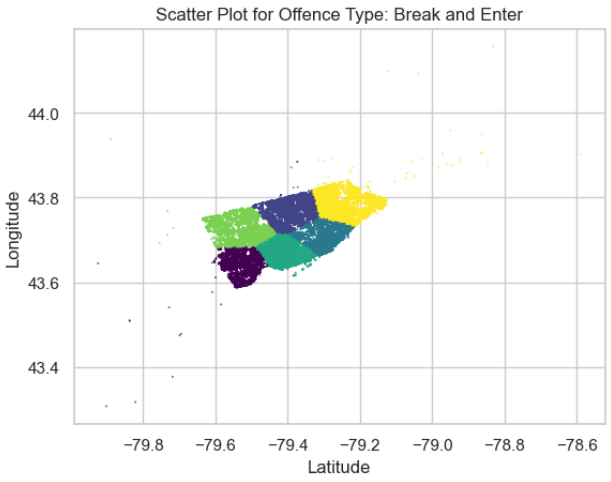
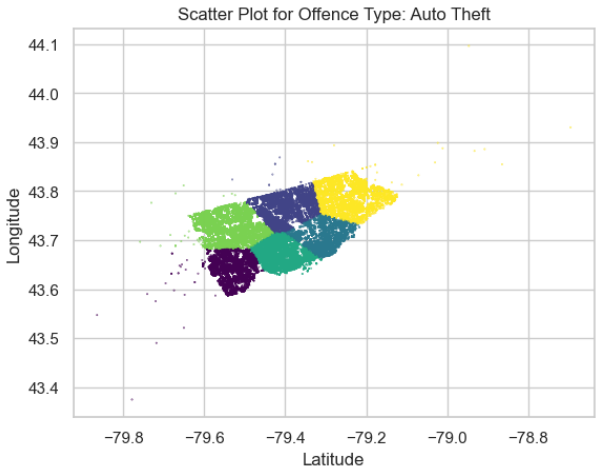


Figure 9 - Distribution of Crime Offences for Cluster 0

Along with the pie charts, each MCI category was mapped and colour coded by cluster to show a visual representation of where each offence occurred, so officers could better ascertain where each type of crime occurs in their cluster. For example, they could check if robberies are generally isolated to downtown, or it has an even spread amongst the city.







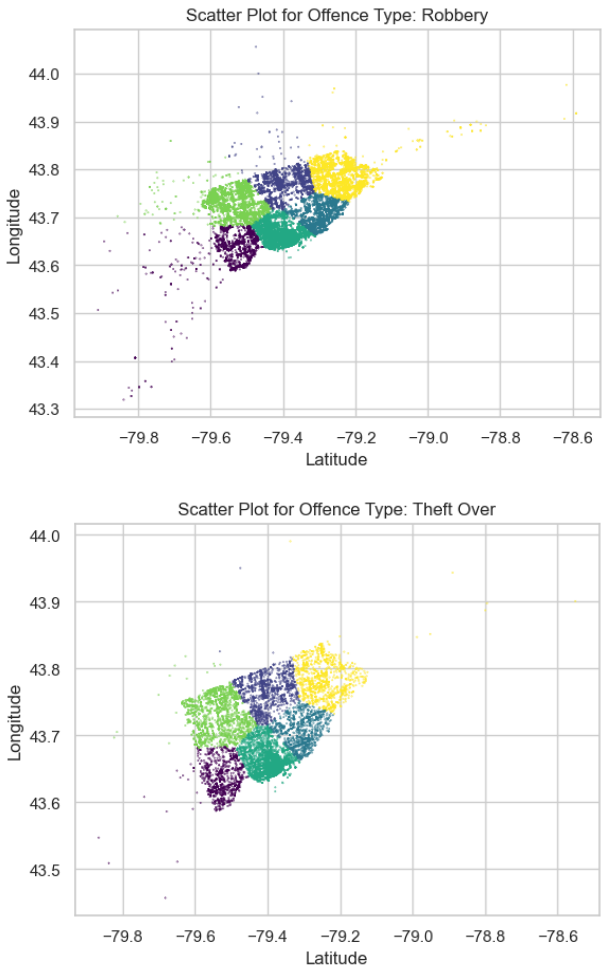


Figure 10 - Maps of Offence Data

IV. Results

1. *Analysis of Cluster Data*

We will be performing this analysis from the point of view of officers in cluster 0. Our results and graphs can be used to gain insights about a specific cluster including generating a schedule and learning what crime to expect. Firstly, we can analyse Figure 7. It can be seen that a few patterns emerge. Firstly, we can see that all the weekdays are very similar. There is a high at midnight and drops off quickly until it reaches a daily low at 6am. Then MCI crime rates steadily increase until noon before a small drop and another slow climb back until midnight. The weekends are also very similar, with a higher peak at midnight, but generally lower crime rates throughout the waking hours of the day when compared to the weekdays. The police in cluster 0 may decide to have 2 different schedules, one for weekdays and another for the weekend. They could have more officers taking their days off on the weekend, due to its lowered crime rate compared to the weekdays. They could also schedule more officers during daylight on weekdays, with similar numbers at night. Human bias can also be seen in the data. There are unnatural peaks in crime rates at both midnight and noon. It is our belief that these numbers are inflated and could be spread to its neighbours slightly. Because Canada runs on the 24 hour clock it is more likely that if someone is not sure what time a crime occurred, they would report a round number which is commonly 12.

Next, the officers could analyse the distribution of crime offences. When compared to the overall data, it can be seen that cluster 0 has largely the same representation of crime when compared to the overall data, with a few exceptions. Firstly in the overall data seen in Figure 8, Assaults were far and away the most likely crime at 53.2% but in cluster 0, it drops to 47.1% as seen in FIgure 9. Robbery, theft over and break and enter were at very similar rates when compared to the overall data, but cluster 0 has a significantly higher rate of auto theft when compared to Toronto as a whole. The rate jumped from 15.4% up to 21.6%. This would inform officers to have more resources allocated to auto theft in cluster 0, and officers could send out localised alerts to warn about auto theft.

This process could be repeated for all clusters to gain localised knowledge that would influence actions taken in each cluster. For our project, it was decided to use 6 clusters due to the results of the elbow method. Our model divides Toronto into clusters with an even crime rate and a fair distribution, allowing easier allocation of officers to each district. If it is found that each cluster is too generalised, our model could be adjusted for any number of clusters, which will inherently give more specialised data that officers could use for their given district.

2. *Conclusions and Future Work*

Our project came down to exploring the dataset, data cleaning, k-means clustering, data representation and data analysis.

By visualising the results of k-clusters, police would be able to more efficiently serve their community by having the most pertinent information available. The number of clusters that we chose may not be ideal for the number of existing departments in Toronto, so the model may need to be adjusted to suit the needs of the existing police force. This could easily be changed to suit their needs, and to help officers learn about their community.

In the future, the results of the clusters could have more data to be displayed to officers and more in depth analysis could be done. Not only could more conclusions be drawn, but

more fields could be added to the clustering model. Models could be made by altering our current clustering model to account for MCI offence type, time of year and much more. Our result offers flexibility for additional models and analysis to be done in the future.