Executive Summary

Our team will be using a combination of three datasets retrieved from the Toronto Police Service and the City of Toronto Open Data portal: Auto Theft Open Data, Break & Enter Open Data, and Neighborhood Open Data. Using the crime datasets, we are investigating the correlations related to auto-theft cases. We will be primarily comparing auto theft data to break-and-enter cases and neighbourhood statistics. This will allow us to make present inferences, as well as future predictions of auto theft cases in Toronto. The five techniques that we will utilize for our analysis are technique-significant merging of data, feature engineering, predictive modelling, visualizing data for insight generation, and executive dashboarding.

Technique-Significant Merging/Joining of Data

This is an essential step for any analysis project. It involves collecting and consolidating data from various sources using standard fields or attributes to tell a richer story. The technique involves understanding the role of primary and foreign keys in a database. This technique is essential for our data because we are trying to analyze crimes and understand the patterns and root causes of auto crimes in different parts of Canada. To do this, we need a wide variety of behavioural and situational data that would not typically be a part of one-directional crime data. The main steps that we must take to apply this technique are identifying data sources, understanding data structures, determining common fields, cleaning data, selecting appropriate merging techniques, performing the merge, and validating the data. This technique will help us carry out better analyses that can lead to top-tier applicable insights on auto theft crimes in Canada. Our team has strong SQL expertise and will be leveraging SQL for this step; we do not envision any problems/limitations in carrying out this step.

Feature Engineering

After the data merging and integration step, feature engineering will be the next significant aspect of our analysis work. In this stage, our principal task is to extract and construct the features from the datasets, which would be necessary tools to make our predictive model. By using these features, we aim to find the key elements related to auto theft and break-and-enter incidents and the data supporting predictions of future trends from these types of crimes.

1. *Temporal Features:* We plan to extract basic elements like year, month, day, and compare and analyze patterns of criminal activity during specific days, such as holiday periods, weekends, and weekdays.
2. *Geographic Location:* For geographic information, our strategy is to identify hotspot locations of crime first. Then, we will test the relationship between locations and surrounding community attributes. For example, the number and type of community facilities and distance to transportation hubs might affect crime rates.
3. *Deep Integration of Community Data:* We will research Neighborhood Open Data to find out how community characteristics, like demographic structure and economic levels, impact crime rates. Meanwhile, factors within the community, such as education resources and recreational facilities, could also potentially be connected to crime occurrence.

Through these steps, our goal is to build a comprehensive feature set, including temporal, geographic, and community background characteristics, which will be the foundation of building an accurate predictive model. This process contributes to more precise prediction and understanding of city crimes and uses data to support crime prevention and control efforts for the City of Toronto. The main limitation that we may encounter is lacking feature engineering skill and experience in our team. We will aim to overcome this limitation through self-learning.

Predictive Modelling

This predictive analysis aims to gain insights into the contributing factors of crimes. One of the tables produced from the collected datasets contains the following: number of crimes, neighbourhood name, neighbourhood number, total age group of the populations, married/common-law rate, education rate, employment rate, average age, and average total household income (2020). Utilizing this data, we will perform a regression analysis to assess whether the model performed well and if it violated any of the assumptions central to regression modelling as per below:

$$\text{Number\_of\_crime} = \beta_0 + \beta_1\text{*Neighbourhood\_Name} + \beta_2\text{*Neighbourhood\_Number} + \beta_3\text{*Total\_Age\_groups\_of\_the\_population} + \beta_4\text{*Married/CommonLaw\_Rate} + \beta_5\text{*Education\_Rate (Bachelor\_or\_higher)} + \beta_6\text{*Employment\_rate} + \beta_7\text{*Average\_age} + \beta_8\text{*Average\_total\_income\_in\_2020} + \varepsilon$$

We will investigate the following questions:

1) Does the model have any predictive power? (Check the F-Test)
2) Do the variables we have included belong in the model? (Check the T-Tests)
3) Have we violated any regression assumptions (Check the plots)
    a. Residual vs Fitted Value – to check if there are any patterns.
    b. Normal QQ plots and density plots – to check for multivariate normality.
    c. Scale-Location - to check for homoscedasticity/heteroskedasticity.
    d. Residual vs. Leverage – to identify observations with high leverage.
4) Is the R^2 sufficient for business requirements?

This technique is essential as it gives us in-depth insights into our data: fitness of the model, outliers, and whether the current variable used to model is significant or not. We might run into a problem of having a low R-square. We will perform different models with the variables by removing insignificant variables. Also, we may achieve transformation if significant outliers are found.

Visualizing data for insight generation

It’s much easier for a common man or even for leaders to understand a story than crunching numbers, statistics, and p-values. As we know, humans tend to be visual creatures. Data Visualization is an art that allows representation of the data through charts, plots, and graphs. It allows communication of complex data relationships, patterns, and data-driven insights in a way that is easy to understand. Using Tableau, one of the reports we are planning to develop will be an executive overview of the anticipated or predicted crimes in the next month, breaking down by premises as either Commercial or Residential. The other will be a line graph showing the predicted auto thefts in a neighbourhood for the next 6 months. This technique is appropriate for our dataset as we can use our predictive modelling outputs to generate visual graphs to better interpret our results. A limitation that we may come across when visualizing data is not being able to fit all essential data into a single figure. We will have to choose which data to visualize based on priority and may have to come up with multiple graphs.

Executive Dashboard

Utilizing our datasets, we will generate an executive dashboard to provide the Toronto police force with an overview of the correlations between auto thefts, break-and-enter cases, and neighbourhoods. The dashboard can also be made available to the general public to raise awareness and educate Torontonians further on this topic. Our main goal is to use Tableau to present our data in a simple manner using a geographical map, bar graphs, and statistical banners. A geographical interactive map of Toronto can allow individuals to see what neighbourhoods have the most auto theft cases, as well as the statistical value to go with it. Bar graphs can be used to compare two different variables, such as the number of auto thefts and break-and-enter cases over a particular time period. This technique is most appropriate for the data as our intent is to present it to an audience with a non-technical background (police force and the public). Having more visuals and an easy-to-navigate dashboard will make it efficient for our targeted audience to derive key points. A limitation that we may face is producing simple visuals given a large dataset. We can overcome this limitation by carefully choosing which constraints we should set in place to prioritize data that is most important for our purpose. Additionally, the majority of our team members have not used Tableau before, and this will be a learning experience for all.

Our Datasets

Auto Theft Open Data – Link:
https://data.torontopolice.on.ca/datasets/95ab41aee16847dba8453bf1688249d6_0/explore?location=18.083507%2C-39.819624%2C3.00
Break & Enter Open Data - Link:
https://data.torontopolice.on.ca/datasets/040ead448df2412da252cfbb532e77ac_0/explore
Neighborhood Open Data - link: https://open.toronto.ca/dataset/neighbourhood-profiles/