Executive Summary

We will use the Traffic Collisions Open Data set obtained from the Toronto Police Service data portal. Our goal is to help our audience visualize the predicted number of traffic collisions for the next 12 months and categorize them based on automobile collisions, motorcycle collisions, and passenger or pedestrian accidents. In doing so, we can help the Police force visualize what types of collisions are most common and where proactive measures should be implemented. We will follow five main processes to execute our project: Data processing and aggregation, feature engineering, classification (logistic regression, KNN, and decision tree), forecasting (time series, ARIMA), and dashboarding.

Data processing and aggregation

The first step of the project will be data processing and aggregation. We will explore the data and ensure that the quality of the dataset is good enough for our modelling and deep analysis.

1. Data Cleaning: A quick quality check will be processed to address missing values, duplicated values, and unit standardization.
2. Collisions' data will be aggregated on different temporal bases, such as daily, weekly, and monthly counts. This will help identify trends over time, which are crucial for time series forecasting.

We do not expect to encounter any major problems during the data processing and aggregation as we have prior experience from our last project.

Feature Engineering

After cleaning and aggregating the data, feature engineering will be conducted to summarize information from detailed data to a higher level suitable for temporal analysis and forecasting. Time-related features, such as the four seasons, weekdays or weekends, and rush hour, will be extracted and transformed. This step is designed to maximize the analytical value of the data, ensuring that subsequent models are both robust and insightful. As per our 860 project, we have obtained a good understanding and experience of feature engineering and do not expect to come across many limitations.

Classification (Logistic Regression, KNN, and Decision Tree)

We aim to use classification methods to gain insights into the Traffic Collisions Open Data obtained from the Toronto Police Service. Our goal is to classify the severity of injuries sustained from traffic collisions; non-fatal and fatal, based on the categorical variables given in the dataset:
Location (Neighborhood), Hour (Rush hour or not), Day (Weekday/Weekend), Types of vehicles involved (Automobile, motorcycle, bicycle), and People (Passenger/Pedestrian).

Below are the three methods that we will experiment with and will use the one that yields the best results:

Logistic Regression

We will perform the following steps:

- Normal or standardize the data
- Split the data into training and testing data sets
- Set penalty=I2, C=1e42, solver=liblinear, multi_class= multinomial as we have a multiclass prediction problem
- Evaluate the models and AIC for model performance
- Perform Confusion Matrix for accuracy measure
- Plot the Gains and Lift Chart

The advantage of Logistic Regression is that it is suitable for binary classification problems, which aligns perfectly with our goal for this prediction with just an overview. It behaves as a linear regression and produces coefficients for each variable, which gives the influence of each variable to the predictor. This is also a disadvantage for Logistic Regression since we need to assume the relationship between predictors and target are linear.

KNN
We will perform the following steps:

- Normal or standardize the data
- Split the data into training and testing data sets
- Tune the hyperparameter, K, through the grid search method to pick the appropriate value
- Evaluate the models using the Confusion Matrix for accuracy measure

The advantage of KNN is its effectiveness. It utilizes the k 'nearest data points to identify the class to which the observations belong. The limitation for KNN is the choice of K. The estimate can be increased by increasing the value of K, but this would also need to include a wider range of data points into the sample. There needs to be a trade off between k and quality.

Decision Trees
We will perform the following steps:

- Split the data into training and testing data sets
- Tune the hyperparameters: max_depth, min_samples_split, min_impurity_decrease through Grid Search to find the right parameters
- Set random_state, max_depth, min_samples_split, and min_impurity_decrease
- Evaluate the models using regressionSummary

The advantage of a Decision Tree is that data does not need to be normalized/standardized.

Classification is essential to our prediction because the model result gives training and testing sets accuracy. It leverages historical/existing data points to predict new data points and to identify areas of high risk in collision/traffic. A limitation for decision tree is the ability to overfit the data. The decision rules split the observations into smaller subgroups until it cannot be split anymore.

Forecasting (time series/ARIMA)
We would use the time series forecasting technique to exploit patterns in the data over time. ARIMA is appropriate for both trend and difference stationary times series analysis. We propose carrying out two types of forecasting: short-term (within the next year) and long-term forecasting (within the next five years). The outcome of this analysis is to develop strategies for effective traffic management and highlight high-risk periods or some factors contributing to traffic accidents.

To build the model, we would follow the steps below;

1. Identify the appropriate ARIMA model parameters (p, d, q) through graphical and statistical tests.
2. Use auto ARIMA to determine the optimal (p, d, q) and select the model with the best fit based on the evaluation.
3. Evaluate the model's accuracy and effectiveness using the Root Mean Squared Error (RMSE).
4. Analyze and derive a trend to predict traffic accidents over one and five-year periods based on factors such as season, location, vehicle brand, and driver age to pinpoint specific times and places where targeted traffic management strategies are most needed.

An example of a trend we would derive from this is predicting traffic accidents by season, location, vehicle brand, driver age, etc. A common limitation that is faced during this technique is overfitting. We aim to avoid this limitation by using a dataset with a large sample.

Dashboard

For our final step, we will create a dashboard to portray our analytical results as a story using Power BI. We will use various graphs to show the relation between different variables such as the number of collisions over the seasons, the severity of injuries sustained from traffic collisions, and automobile collisions versus motorcycle collisions. Additionally, we can incorporate a map of Toronto displaying the number of collisions per region, which can be filtered by year. Creating a dashboard to visualize our findings is the best technique when taking our audience into consideration, as individuals who do not come from an analytical background can still follow our story effectively. Furthermore, it is also a concise way of displaying our analysis. A limitation we may face is that the team members working on the dashboard have no prior experience using Power BI. However, we aim to use this as a self-learning experience and will reach out to other team members who have experience using the application if any issues arise.