Project Title: Determining Risk Factors for Injuries and Fatalities in Canadian Road Collisions

Contributors: Keegan MacDonald, Sanita Richards, Corey Thorne, Reuben Walker

Date: December 5th, 2023

Course Code: GDAA 1001

**Introduction**

Road collisions remain a significant concern globally, impacting lives and posing a considerable societal challenge. Canada, with its vast network of roadways, experiences its share of incidents, contributing to injuries and fatalities. Understanding the factors associated with these collisions is crucial in formulating effective preventive measures and enhancing road safety protocols.

This report aims to delve into Canada's national collision data from the year 2019 to discern the severity of injuries and fatalities in road collisions. By employing data analytics techniques and modeling in R Studio, this study endeavors to uncover patterns, correlations, and potential causal relationships among various variables involved in these incidents.

**Data Preparation**

Our team embarked on identifying a suitable dataset aligning with the project's objectives: to analyze road collision data in Canada to unveil risk factors associated with injuries and fatalities. We sourced the dataset from National Collision Database (NCDB), ensuring it met the outlined criteria of containing over 100 observations and at least 5 predictor variables, excluding redundant columns such as IDs and geometry. This dataset encompasses a range of categorical variables, including the month, weekday, and hour of collisions, the number of vehicles involved, road type (intersection, passing lane, bridge, etc.), and weather conditions, among 18 others. It's important to note that many variables include extra categories accounting for unknown or missing data, which required careful handling.

Using R, we imported the collision dataset, and conducted an initial assessment to understand its structure, identify variables, and assess data quality. This preliminary step involved scrutinizing for missing values, redundant columns, and potential data cleaning requirements. Our dataset initially comprised 272,301 observations and 23 variables.

The data cleaning process involved several tasks determined by the analysis conducted. In order to ensure the integrity of our analysis, we judiciously removed all observations containing unknown or missing values. Subsequently, six columns deemed irrelevant for machine learning were excluded, including the redundant 'year 2019,' three columns containing unique collision identifiers, a duplicate of the target variable denoting collision severity, and a column detailing protective clothing and safety devices, largely populated with missing values. This meticulous curation resulted in a refined dataset, now comprising 17 columns and 167,803 observations, poised for robust machine learning model development and analysis.

The updated dataset underwent a meticulous cleaning and preparation phase to ensure its suitability for robust analysis. Categorical variables were appropriately grouped and transformed for streamlined analysis in R Studio.

The 'numCars' column was created by categorizing the 'C\_CONF' variable based on collision configurations, distinguishing between single-vehicle incidents and various two-vehicle scenarios. Conversion of several columns into factor types was conducted to enable categorical analysis, including 'P\_ISEV' (Injury Severity), 'C\_MNTH' (Month), 'C\_WDAY' (Weekday), and others deemed vital for categorical assessment.

To enhance readability and interpretability, categorical codes were reassigned meaningful labels. For instance, numerical codes representing injury severity, months, weekdays, and weather conditions were relabeled into descriptive categories. Additionally, column names were updated to ensure clarity and consistency throughout the dataset.

Subsequently, the categorical variables were grouped, optimizing the dataset for analysis and model development. Moreover, a subset of 5000 rows was randomly sampled from the prepared dataset to facilitate manageable yet representative analyses.

This rigorous data preparation process resulted in a refined dataset poised for in-depth analysis, machine learning modeling, and exploration of risk factors associated with injuries and fatalities in road collisions within Canada's 2019 national collision data.

**Exploratory Data Analysis & Visualization**

The visualization strategy adopted in response to the significant disparity in injury severity levels played a pivotal role in enhancing our understanding of the data. The decision to create separate visualizations for fatalities was critical, as it effectively addressed the challenge posed by the overwhelming number of incidents without injury or with minor injuries. By isolating fatalities into distinct plots, we achieved a more nuanced examination of trends specific to this vital category. This tailored approach ensured that the prominence of incidents resulting in no or minor injuries did not overshadow the crucial insights derived from analyzing fatalities. It ultimately facilitated a clearer, more focused representation of trends, enabling a deeper comprehension essential for informed decision-making in accident prevention and mitigation strategies.

The visualizations encompass diverse aspects of collision data, offering insights into temporal patterns, demographics, and road/weather conditions. Analyzing collisions over time, the graphs showcase seasonality, with higher incidents during summer and lower ones in winter. Fatalities, however, show consistency from January to April, differing from other collision types.

Exploring weekly trends, collisions increase through the week, peaking on Fridays, and drop significantly over weekends, a probable outcome of reduced commuting. Fatalities, slightly higher on Mondays and Tuesdays, reach a low on Wednesdays. The density plot concerning collisions throughout the day illustrates spikes during commuting hours, notably around 7 AM and a more prominent one after 3 PM, with reduced incidents between midnight and 5 AM. Fatalities demonstrate less decline during these early hours, potentially tied to reduced visibility.

Demographic analyses based on sex and age reveal distinctions. Male fatalities surpass female ones significantly, while collisions involving women exhibit higher injury proportions. Age-wise, children under 12 display lower injury rates and a reduced share of fatalities, contrasting with higher fatality proportions beyond age 50.

The road and weather condition visualizations initially indicated an abundance of clear/sunny weather and dry road entries. However, proportionate analyses were recommended to unveil insights into conditions contributing to a higher chance of injury or fatality rather than just collision occurrences.

Additionally, histograms dissecting collisions based on variables such as the number of vehicles involved, road configuration, traffic conditions, and vehicle type revealed significant disparities between fatalities and other injury severities. These findings underscored the imperative for separate analyses to discern and understand patterns more effectively.

**Predictive Modelling**

**Evaluation of Results**

**Conclusion**

Utilizing a lollipop graph, we effectively visualized the importance of predictor variables in our classification and regression tree (CART) model. Filtering out factors with an importance greater than zero revealed key variables crucial for predictions. This insight prompted a model refinement, resulting in a notable 3% accuracy improvement in our trimmed dataset, a testament to the significance of these variables in prediction accuracy. Reassuringly, subsequent model runs after further data trimming consistently identified these pivotal variables, affirming their consistent performance. Comparative analyses across multiple models highlighted CART's superiority in accuracy, particularly in predicting fatalities, demonstrating its efficacy in distinguishing fatality outcomes. This model's potential practical application in dispatching emergency responders swiftly in potential fatality cases underscores its real-world relevance. However, discerning between 'No injury' and 'Injury' proved more intricate, suggesting a need for enhanced model precision in distinguishing these categories for practical deployment.

**Reference**

*National Collision Database - Open Government Portal*. (n.d.). https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a