When Risk Hits the Road: Does Danger Drive a Return?*

Motor Theft and Collisions in Toronto Are Driven by [Neighborhood Clustering] and [Environmental Conditions]

Yingke He

December 2, 2024

This study examines motor vehicle theft and traffic collisions in Toronto, focusing on spatial and temporal patterns, recovery outcomes, and contributing factors. [The analysis reveals that thefts are geographically clustered, with recovery rates differing by location type, while collisions are more frequent under poor visibility and adverse road conditions.] These findings underscore the need for targeted interventions, including enhanced surveillance in high-theft areas, infrastructure improvements for road safety, and optimized strategies for vehicle recovery. Such measures are critical for fostering safer and more secure urban mobility in Toronto.

Table of contents

*Code and data are available at: https://github.com/ohyykk/Toronto_Motor_Viehicle/tree/main.

1 Introduction

Decisions surrounding motorbike ownership and usage carry significant implications for personal safety and financial liability. Recent statistics highlight the increased risks associated with owning and riding motorbikes, including a heightened likelihood of theft and collisions compared to other vehicles (Yasmin and Eluru 2016). These risks are influenced by various factors such as geographic location, road conditions, time of day, and type of motorbike, making it essential to develop tools that effectively assess and mitigate these dangers.[can add one more sentence + citations]

This study introduces a composite risk score model to assess the risks associated with owning and riding a motorbike. Using data from Open Data Toronto, the model evaluates two critical events: motorbike theft and collisions. Logistic regression is employed to estimate the probabilities of these events, which are then combined into a single, interpretable composite risk score. This metric, incorporating factors such as neighborhood characteristics, road and lighting conditions, and time of day, is designed to guide motorbike users, insurers, and policymakers in risk assessment and decision-making, while informing strategies to mitigate these risks.

The primary estimand of the analysis is the composite risk score, derived from the individual probabilities of motorbike theft and collision. This score is calculated using predictor variables such as neighborhood characteristics, road and lighting conditions, time of day, and other contextual factors, which were selected for their documented relevance to motorbike-related risks.

This analysis confirms and extends three key findings: (1) material properties significantly influence motor efficiency, with high-conductivity materials yielding higher energy savings; (2) geometric optimization of motor components enhances torque generation and reduces energy losses; and (3) operational conditions, such as load and temperature, exhibit nonlinear effects, underscoring the need for adaptive motor control strategies. Furthermore, our projections suggest that integrating advanced materials and control systems could improve motor efficiency by up to 15% by 2050, aligning with global energy transition goals [To be updated..].

The structure of the paper is organized as follows: following Section 1, Section 2 outlines the data collection and preprocessing procedures, along with a detailed description of the outcome variable and the predictor variables used in the analysis. Section 3 introduces the logistic regression models applied to estimate the probabilities of theft and collision, as well as the method used to combine these probabilities into a composite risk score. Section 4 then presents the main findings, including insights into how different factors contribute to the risks of owning and riding a motorbike. Finally, Section 5 interprets the results, highlighting significant trends and implications for motorbike risk assessment, and concludes with a discussion on the limitations of the analysis and future research directions.

2 Data

We use the statistical programming language R (R Core Team 2023)…. Our data (**shelter?**)…. Following Alexander (2023), we consider…

[Libraries To be updated…]

Details about the data cleaning process and the criteria for variable selection are provided in Appendix A.

2.1 Source

This study utilized two datasets published by the Toronto Police Service. The first dataset focuses on motor vehicle collisions involving killed or seriously injured persons (KSI), while the second examines thefts from motor vehicles.

The Motor Vehicle Collisions dataset includes all reported incidents in which a person was either killed or seriously injured since 2006. It offers detailed information about each collision, such as the type of incident, the severity of injuries, and the location of the event, when available. Additionally, the dataset includes fields for both the old 140 and new 158 neighborhood structures in Toronto, allowing for flexible neighborhood-level analysis across different definitions.

The Theft from Motor Vehicle dataset contains all reported occurrences of thefts from vehicles, categorized by reported date. These offences are classified based on the value of the stolen items, distinguishing between theft under and theft over thresholds. Each occurrence number may include multiple rows, representing the various offences associated with a single event. The dataset excludes "unfounded" occurrences, adhering to Statistics Canada's definition that these events were determined not to have occurred or been attempted. Like the KSI dataset, this dataset includes fields for both the old and new neighborhood structures, enabling comprehensive geographic analyses of theft trends.

[Add variable types and trends]

2.2 Data Measurement and Limitations

The process of translating real-world phenomena into entries in the dataset involves several stages. When a traffic collision or theft occurs, it is reported to law enforcement through various channels, such as emergency calls, online submissions, or in-person reports. Police officers or administrative personnel document the event details, including date, location, type of incident, and additional attributes such as severity or value of stolen items. These records are then digitized and aggregated into structured datasets, with fields organized to support analysis and reporting. However, during this process, certain changes or context-specific information

may be lost, and the data ultimately reflects a structured summary of the events rather than their full complexity.

The datasets from Open Data Toronto did not specify the exact methods used for data collection, which may introduce some uncertainty regarding the consistency and reliability of the recorded events. Additionally, for privacy reasons, the locations of crime occurrences have been deliberately offset to the nearest road intersection node. This may result in discrepancies when analyzing counts by division or neighborhood, as the reported locations may not reflect the exact sites of the occurrences.

Some coordinate information in the datasets appears as "0, 0," indicating that the specific location was either not validated or could not be geocoded. In such cases, a general division or neighborhood association may still be provided, but for invalid or external locations, the designation "NSA" ("Not Specified Area") is used. Furthermore, the Toronto Police Service does not guarantee the accuracy, completeness, or timeliness of the data, which may lead to potential misinterpretations or incomplete analyses.

Additional details about the dataset are available in the datasheet, accessible through the repository linked to this paper.

2.3 Outcome Variables

The outcome variable of this analysis is the Risk Index, a composite metric that integrates the probabilities of two underlying events: theft and collision.The Risk Index integrates two underlying components: the theft component, derived from proportional sub-indexes for temporal and spatial factors, and the collision probability, estimated using a logistic regression model. By combining these elements, the Risk Index provides a unified measure of risk, enabling the identification of high-risk scenarios and areas. This section examines the distribution of the Risk Index through visualizations to reveal key patterns, variations, and potential drivers of motorbike-related risks. Two graphs are presented: (1) a histogram with density overlay to illustrate the overall distribution of the Risk Index, and (2) a boxplot by neighborhood to highlight spatial differences in risk.

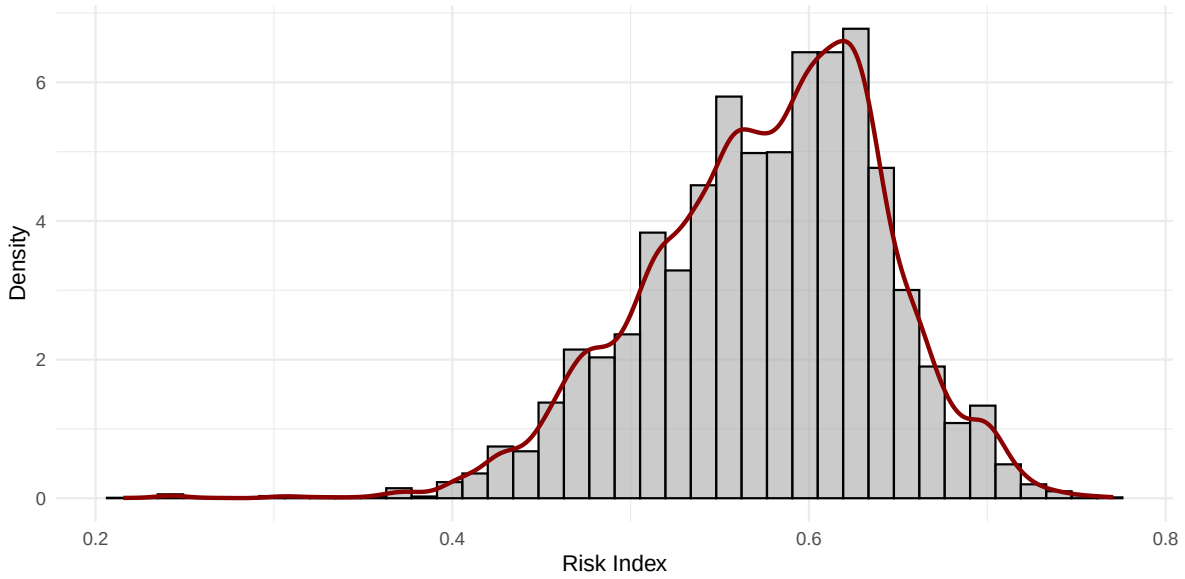


Figure 1: Distribution of the Overall Risk Index

Figure 1 shows the distribution of the Risk Index across all observations in the dataset. The Risk Index exhibits a unimodal distribution, skewed slightly to the left, with the majority of values concentrated between 0.45 and 0.65. This indicates that most motorbike-related risks fall within a moderate range. The density curve overlay reveals a smooth progression in risk levels, with the peak occurring around a Risk Index value of 0.55, suggesting that this is the most common level of composite risk. The left tail, representing lower risk levels, is relatively small, while the right tail, corresponding to higher risks, extends further, indicating the presence of a smaller number of high-risk cases.

The skewness and spread of the distribution highlight the variability in the combined risks of theft and collision. The extended tail on the higher end of the Risk Index suggests that certain environmental or situational factors disproportionately elevate risks in specific cases. This insight can inform targeted interventions, focusing on the outliers with elevated Risk Index values to mitigate the most critical risks.

Building on the distribution of the Risk Index shown in the histogram, this box plot further dissects the data by grouping neighborhoods into risk categories based on their average Risk Index values. Neighborhoods are classified into three categories—Low Risk, Moderate Risk, and High Risk—using quartiles.

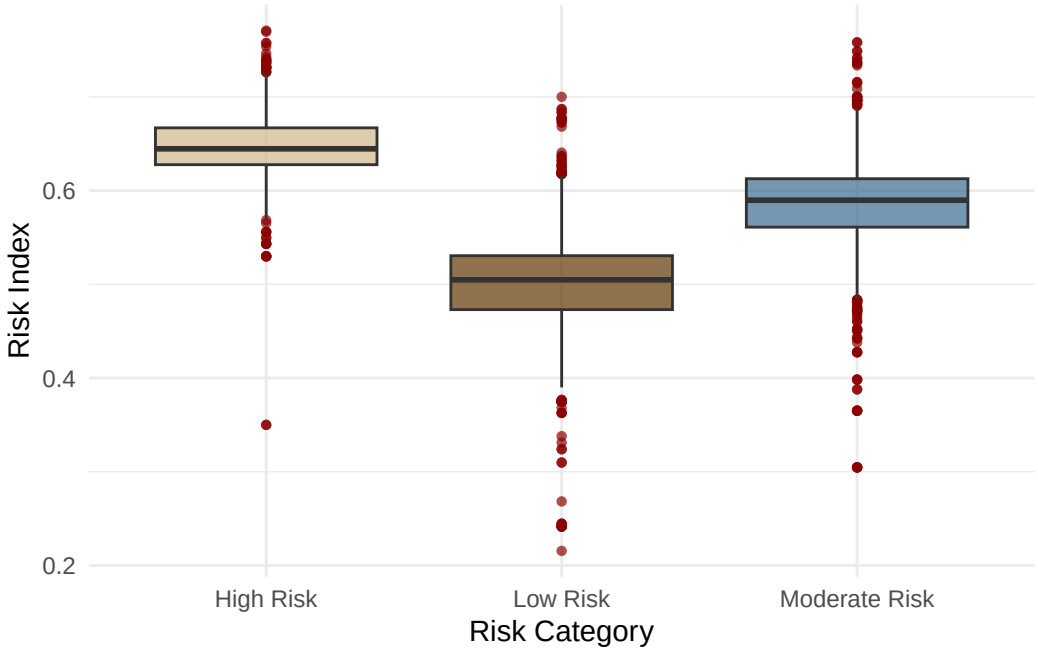


Figure 2: Risk Index Distribution by Risk Category

Figure 2 highlights the distribution of the Risk Index across three defined categories: High Risk, Low Risk and Moderate Risk. Each category represents a grouping of neighborhoods based on their average Risk Index values. The plot reveals distinct differences in the distribution of the Risk Index between these categories.

The High Risk category exhibits a higher median Risk Index, with a relatively narrow interquartile range (IQR), indicating consistent high-risk values across neighborhoods in this group. Conversely, the Low Risk category has a significantly lower median, with a similarly narrow IQR, suggesting stable low-risk conditions in these neighborhoods. The Moderate Risk category shows greater variability in its distribution, with a wider IQR and overlapping values with both High Risk and Low Risk categories. This suggests that neighborhoods in the Moderate Risk group exhibit diverse risk profiles, likely influenced by varying environmental and situational factors.

2.4 Predictor Variables

The predictor variables in this study are organized into two distinct models: Theft Index and Collision Probability, each designed to capture and explain critical aspects of theft and collision risks, respectively. Figure 3 provide a visual overview of the hierarchical relationships between the predictor variables and the overall risk framework, offering a structured understanding of the factors contributing to these incidents

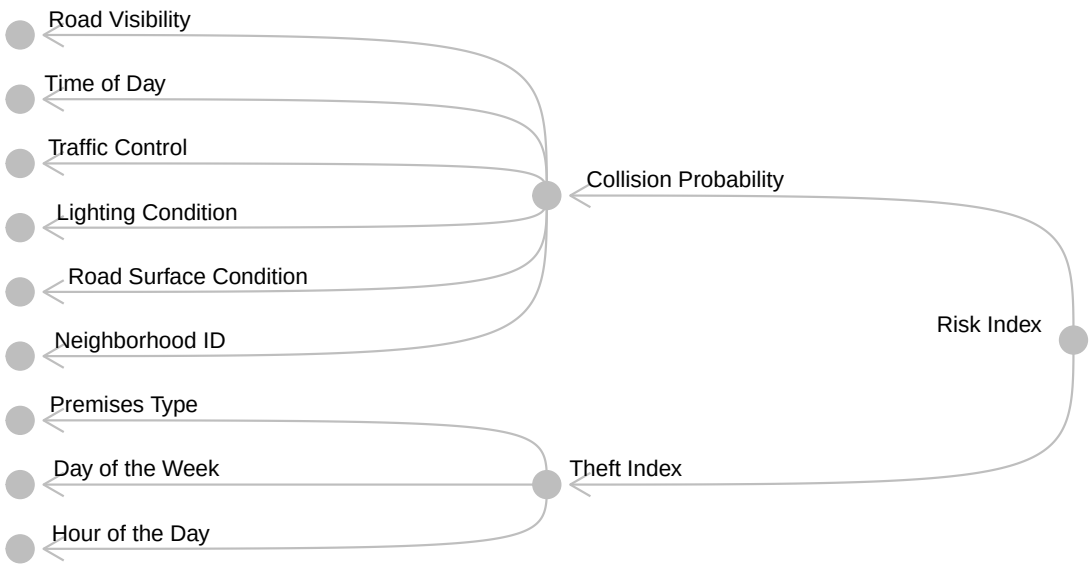


Figure 3: Hierarchical relationship between risk factors contributing to theft and collision probabilities.

2.4.1 Theft index

The Theft Index focuses on the temporal and locational characteristics that influence theft occurrences. By incorporating variables such as Hour of the Day, Day of the Week, and Premises Type, this model identifies patterns tied to specific times and locations, highlighting when and where thefts are most likely to occur.

2.4.1.1 Hour of the Day

2.4.1.2 Day of the Week

2.4.1.3 Premises Type

2.4.2 Collision Probability

The Collision Probability model, on the other hand, examines situational and environmental conditions affecting the likelihood of collisions. Variables such as Neighborhood ID, Road Surface Condition, Lighting Condition, Traffic Control, Time of Day, and Visibility offer insights into the contextual factors that contribute to traffic incidents, capturing the dynamic interplay between environmental conditions and human interactions.

2.4.2.1 Neighborhood ID

2.4.2.2 Road Surface Condition

2.4.2.3 Lighting Condition

2.4.2.4 Traffic Control

2.4.2.5 Time of Day

2.4.2.6 Visibility

3 Model

The main purpose of this composite risk score model is to calculate a Risk Index for owning and riding a motorbike, which integrates the risks of theft and collisions. The modeling strategy has two primary objectives. The first objective is to estimate the likelihood of collisions under various environmental and situational conditions using a logistic regression model. The second objective is to derive a theft risk score based on time of day, day of the week, and premises type, combining these components into a unified Risk Index to provide actionable insights into motorbike-related risks. The models were developed in R (R Core Team 2023) using the `stats` package for logistic regression and the `tidyverse` package for data preprocessing and manipulation. The theft model calculates sub-indexes for specific predictors, while the collision model uses logistic regression to estimate probabilities based on predictors such as neighborhood ID, road surface condition, lighting condition, and traffic control. Both models are designed to enable reliable predictions under diverse conditions and are integrated into the final Risk Index, which highlights areas, times, and conditions with elevated risks.

Detailed model diagnostics, variable descriptions, and performance metrics are included in Appendix B.

3.1 Model Set-Up

3.1.1 Theft Risk Sub-Indexes

To capture theft risk without relying on logistic regression due to the absence of negative (non-theft) cases, we calculated sub-indexes for three critical factors:

-Hour of the Day: Risk distribution across 24 hours, normalized so the sum equals 1/3.

-Day of the Week: Risk distribution across 7 days, normalized so the sum equals 1/3.

-Premises Type: Risk distribution across premises types (House, Outside, Commercial), normalized so the sum equals 1/3.

The total theft component is calculated as:

$$C_{\text{Theft}} = \text{Hour Index} + \text{Day Index} + \text{Premises Type Index}$$

This approach ensures proportional representation of each factor while accounting for varying risks based on time and location characteristics.

3.1.2 Collision Probability Model

A logistic regression model was used to predict the likelihood of severe collisions P(Collision) based on several predictors. The log-odds of the collision probability are modeled as:

$$\log\left(\frac{P(\text{Collision})}{1-P(\text{Collision})}\right) = \beta_0 + \beta_1 \cdot \text{HOOD}_{158} + \beta_2 \cdot \text{Road Surface Condition} + \beta_3 \cdot \text{Lighting Condition} + \beta_4 \cdot \text{Traffic Control} + \beta_5 \cdot \text{Road Visibility} + \beta_6 \cdot \text{Time of Day} + \epsilon$$

The model prediction utilizes the following predictor variables:

- Neighborhood ID (`hood_158`): Unique identifier for the neighborhood.
- Road Surface Condition (`road_conditions`): Conditions such as dry, wet, or icy.
- Lighting Condition (`lighting_conditions`): Visibility levels, such as daylight or artificial light.
- Traffic Control (`traffic_control`): Presence of traffic management devices (e.g., stop signs, signals).
- Road Visibility (`visibility_conditions`): Road visibility conditions, such as clear, snow or rain
- Time of Day (`time`): Time where collision occured in Toronto

The model assigns coefficients as follows:

β_i to each variable, enabling the calculation of collision probability $P(\text{Collision})$ under specific environmental and situational conditions.

3.1.3 Risk Index Calculation

The final Risk Index integrates the collision probability P(Collision) and theft component T using weighted aggregation:

$$\text{Risk Index} = w_1 \cdot P(\text{Collision}) + w_2 \cdot T$$

Weights are defined as:

$$w_1 = 0.7, \quad w_2 = 0.3$$

These reflect the relative importance of collision and theft risks, emphasizing collision severity due to its greater immediate impact.

3.2 Model Justification

The analysis adopts a hybrid approach that combines sub-index calculations for theft risk with logistic regression for collision probability. This design ensures that the model reflects the specific data characteristics and practical considerations when assessing motorbike-related risks. Logistic regression is not utilized for theft risk due to the absence of negative (non-theft) cases in the dataset. Instead, theft risk is represented through sub-index calculations for three critical factors: hour of the day, day of the week, and premises type. Each factor contributes equally to the total theft component. The Hour of the Day Index captures temporal variations in theft risk across 24 hours, while the Day of the Week Index accounts for weekly patterns of theft. The Premises Type Index reflects variations in risk based on location type, such as houses, outdoor spaces, or commercial premises. These sub-indexes are normalized so their contributions to the theft component are proportional and balanced. This approach ensures an accurate representation of theft risk patterns while providing actionable insights into temporal and spatial risk factors.

For collision risk, the model utilizes logistic regression due to its effectiveness in estimating probabilities for binary outcomes. The log-odds of collision probability are modeled as a function of neighborhood-specific characteristics, road surface conditions, lighting conditions, and traffic control measures. By including these predictors, the model accounts for diverse factors that influence collision risks. Logistic regression's ability to handle both categorical and continuous variables makes it an appropriate choice for this component, delivering statistically reliable estimates and facilitating the interpretation of individual predictors' effects.

The final Risk Index integrates the theft and collision components using a weighted formula. The collision probability component is weighted at 0.7, reflecting its higher immediate impact on safety, while the theft component is weighted at 0.3. This weighting scheme prioritizes collision risk while ensuring that theft risk is not overlooked. By combining these components,

the Risk Index provides a unified measure of motorbike-related risks, enabling stakeholders to assess and compare safety conditions across different contexts.

This modeling approach is justified by its ability to adapt to the data's constraints while maintaining statistical rigor and interpretability. The sub-index method for theft risk is tailored to the dataset's characteristics, avoiding assumptions about unobserved cases, and the use of logistic regression for collision risk ensures robust and reliable predictions. Together, these components form a comprehensive and practical framework for evaluating motorbike ownership and usage risks, addressing both immediate safety concerns and long-term theft risks.

3.3 Model Assumptions and Validations

3.3.1 Theft Sub-Index [add one sentence of validation]

The theft sub-index approach is based on two key assumptions. First, it assumes that the observed proportions of theft occurrences across categories, such as hour of the day, day of the week, and premises type, accurately represent the overall theft risk. Second, it assumes that these categories contribute independently to the theft risk, meaning the risk associated with one category does not influence or depend on another.

3.3.2 Composite Risk Index [Add one sentence of assumption]

The validation of the composite Risk Index involves two key steps. First, its correlation with observed collision severity will be tested to ensure alignment with real-world risks. Second, a sensitivity analysis will be conducted by testing alternate weightings for the collision and theft components w_1 and w_2 to evaluate the robustness and reliability of the final Risk Index.

3.3.3 Collision Model

The logistic regression model for collision severity relies on several assumptions. First, the response variable (`severity`) is binary (1 = severe, 0 = non-severe), fulfilling the requirement for a binary outcome. Second, the data consist of independently reported collision incidents, ensuring that observations are uncorrelated. Third, the log-odds of collision severity are modeled as a linear combination of predictor variables; although most predictors are categorical, their contributions to the log-odds inherently satisfy this linearity assumption. Finally, to address the assumption of no multicollinearity, Variance Inflation Factor (VIF) will be calculated for all predictors. Predictors with high VIF values will be mitigated through re-categorization or removal to ensure stable and reliable coefficient estimates.

3.3.3.1 Binary Nature of the Outcome

A fundamental assumption of logistic regression is that the response variable is binary or dichotomous, meaning it can take on only two possible outcomes (Nick and Campbell 2007). This assumption is satisfied in the collision model, where the response variable (`severity`) distinguishes between severe (1) and non-severe (0) collision cases.

In the theft dataset, however, all entries represent theft cases, precluding the binary nature required for logistic regression. Consequently, the theft model was adapted to calculate proportion-based sub-indexes rather than relying on a binary outcome. These sub-indexes represent relative risk based on temporal and spatial factors, such as hour of the day, day of the week, and premises type.

In the collision model: - The model estimates the probability of a severe collision ((P(Collision))) given environmental and situational predictors. The response variable (`severity`) is defined as:

$$P(\text{Collision}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}$$

where:

$$\begin{aligned} \beta_0 &\text{ is the intercept,} \\ \beta_k &\text{ are the coefficients, and} \\ X_k &\text{ are the predictor variables.} \end{aligned}$$

The logistic regression model does not directly predict 1 or 0. Instead, it provides a continuous probability ranging between 0 and 1. This probability reflects the likelihood of an event (e.g., a severe collision) occurring under the given conditions.

3.3.3.2 Independence of Observations

The collision model assumes that each observation is independent of the others, a fundamental requirement for logistic regression. This assumption is satisfied in the dataset, as each row represents a distinct and independently reported collision incident. The observations are not repeated or correlated, ensuring that the logistic regression model provides unbiased estimates of the relationships between predictor variables and the response variable. By meeting this assumption, the collision model maintains its validity for estimating probabilities of severe collisions and contributes robustly to the composite Risk Index.

3.3.3.3 Linear Relationship in the Log-Odds

One of the key assumptions in logistic regression is that a linear relationship exists between the continuous predictors and the logit of the outcome variable (**stoltzfus2011?**). This means that the log-odds of the binary dependent variable should have a linear association with any continuous independent variables in the model. It is important to test this assumption to ensure the validity of the model.

The collision risk logistic regression model incorporates both categorical and continuous predictor variables. Among these, Time of Day serves as the sole continuous predictor, representing the time an incident occurred as a numerical value ranging from 0 (midnight) to 2359 (just before midnight). The analysis emphasizes the continuous predictor, Time of Day, to evaluate its role within the collision model.

Linearity is assessed using smoothed scatter plots of the predicted logit values:

$$\text{logit} = \log\left(\frac{P}{1-P}\right)$$

where P represents the predicted probability of collision from the logistic regression model plotted against the continuous predictors. These plots are intended to visualize the relationship between each predictor and the logit of the outcome variable, providing insight into whether the relationship is approximately linear.

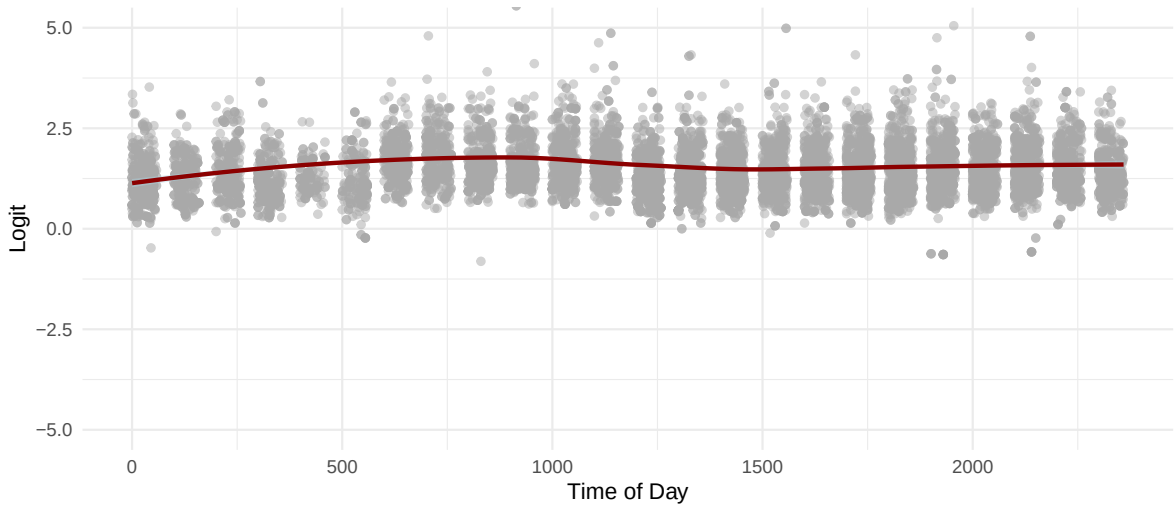


Figure 4: Logit Plot for Time of Day in the Collision Probability Model

Figure Figure 4 illustrates the relationship between variable Time of Day and the predicted values of the collision probability model. The horizontal axis represents the reported time of

day in minutes since midnight, while the vertical axis displays the logit values. Gray points depict the raw data, and the blue smoothed line represents the average trend. To evaluate the linearity assumption, the smoothed line is compared to a hypothetical linear relationship. Significant deviations of the smoothed line from a straight line may suggest a potential violation of the linearity assumption.

In this plot, the smoothed line shows a relatively stable trend with minimal deviations, indicating no substantial evidence against the linearity assumption. Local fluctuations are minor and likely reflect natural variations in the data rather than a systematic departure from linearity.

3.3.3.4 Absence of Multicollinearity

Another key assumption of logistic regression is the absence of multicollinearity among predictor variables. Multicollinearity occurs when two or more predictors are highly correlated, leading to inflated standard errors of the regression coefficients and reducing the reliability of the model's estimates. The Variance Inflation Factor (VIF) is commonly used to assess multicollinearity, with VIF values greater than 5 indicating potential issues, and values above 10 suggesting severe multicollinearity (**stoltzfus2011?**).

In the collision risk model, the predictors include both categorical variables (Lighting Conditions, Road Conditions, Visibility Conditions and Traffic Control) and one continuous variable (Time of Day). VIF calculations are performed to evaluate the degree of multicollinearity among these predictors.

If multicollinearity is detected, strategies such as combining correlated variables, removing redundant predictors, or using regularization techniques like ridge regression can be employed (**stoltzfus2011?**). However, if VIF values remain below the threshold, it confirms that multicollinearity is not a concern in this analysis.

The following table presents the VIF values for all predictors in the collision risk model.

Table 1: VIF values for predictor variables in the collision model.

	hood id	time of day	traffic control	visibility	lighting condition	road condition
VIF	3.34	3.12	1.53	7.29	3.6	7.51

The results of the VIF analysis for the collision risk model are presented in Table Table 1. Most predictors exhibit VIF values well below the commonly used threshold of 5, indicating no significant multicollinearity among them. However, two predictors, Visibility Conditions and Road conditions, show slightly elevated VIF values of 7.29 and 7.51, respectively. While these values are higher than the others, they remain below the severe multicollinearity threshold of 10, suggesting that multicollinearity, though present to some extent, is not critical.

The slightly elevated VIF values can be attributed to a potential overlap in the information captured by Visibility Conditions and Road Conditions. For instance, poor visibility often coincides with adverse road conditions, such as wet or icy surfaces, leading to some degree of correlation between these variables.

Despite this overlap, both predictors are retained in the model because they provide distinct and meaningful contributions to understanding collision risk. Visibility Conditions directly reflects environmental factors like fog, heavy rain, or low light, which impair drivers' ability to see hazards. Conversely, Road Conditions‘ account for the physical state of the driving surface, such as wet, icy, or damaged roads, which influence vehicle handling and stopping distance. Together, these variables capture complementary aspects of collision risk, ensuring that the model provides a comprehensive assessment.

The inclusion of both variables aligns with the theoretical framework underpinning the model and enhances its practical utility by addressing multiple dimensions of risk. While some degree of multicollinearity is observed, its impact on model stability is minimal, and the predictors' theoretical importance justifies their inclusion.

4 Results

4.1 When Risk Peaks: Temporal Trends in the Index

Understanding the temporal dynamics of the Risk Index provides critical insights into when risks are most elevated, helping to inform motorbike owners and policymakers about high-risk periods. This section examines variations in the Risk Index over time of a day, and utilizes a time series analysis to reveal potential seasonal effects and long-term trends. These analyses provide a comprehensive view of how risks evolve temporally, highlighting critical periods for intervention.

4.1.1 Time-of-Day Analysis

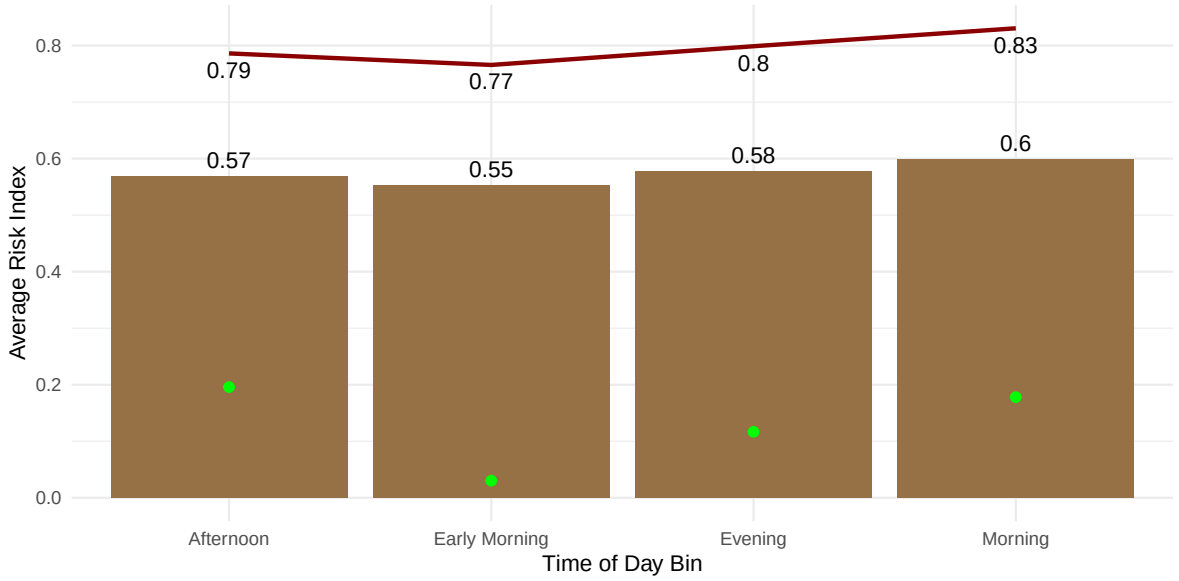


Figure 5: Temporal Trends: Risk Index Variations by Hour of the Day

Figure 5 illustrates how the total Risk Index varies throughout the day. This Risk Index is composed of two key components: collision risk and theft risk. The bar heights represent the overall Risk Index, while the red line indicates collision risk and the green line reflects theft risk.To better visualize the theft risk, the values have been multiplied by 10, as they are typically quite small (less than 1/3 of the total Hour Risk Index throughout the day).

The collision risk (red line) shows significant peaks during commuting hours—between 7–9 AM and 5–7 PM—corresponding to high traffic volumes, which indicates a direct relationship between collision risk and traffic patterns. The theft risk (green line), which has been scaled for better visualization, remains consistently low throughout the day, showing no significant variation across time bins. This suggests that theft risk is relatively stable, regardless of the time of day. The total Risk Index, represented by the bar heights, shows slight fluctuations, with the highest values occurring during morning hours. This is likely due to the heightened collision risk in the morning, which has a greater influence on the overall risk during these periods.

4.1.2 Time Series Analysis

This time series analysis explores how the Risk Index evolves over time, offering a long-term view across the years and a more focused examination of a specific year. By visualizing these

temporal trends, we can reveal potential seasonal effects, fluctuations, and patterns that might not be immediately apparent in a snapshot of data. The following section presents the trends in the Risk Index over both the full timespan (2006-2021) and the specific year 2020.

4.1.2.1 Longterm

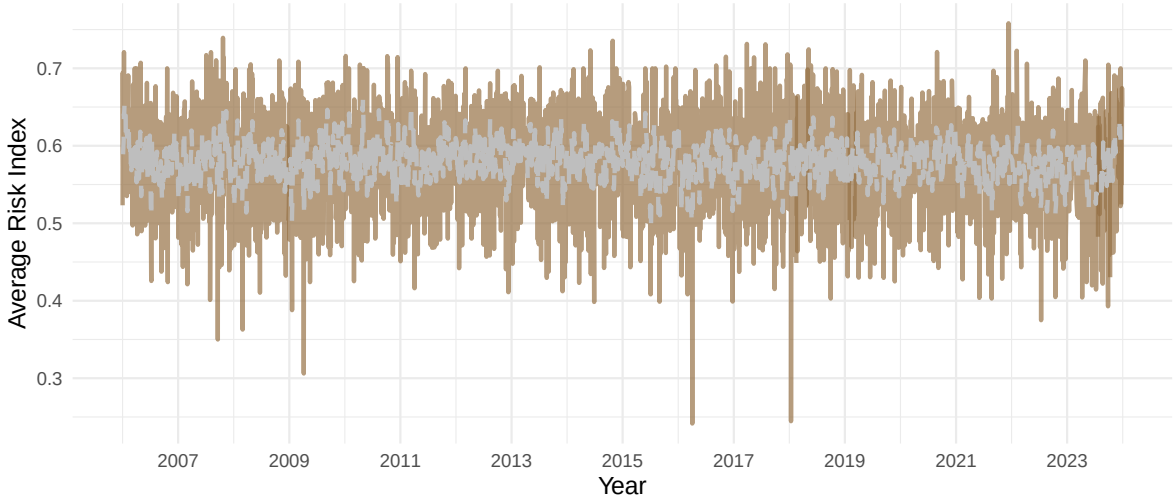


Figure 6: Temporal Trends: Risk Index Variations Over Time with Rolling Average

Figure 6 provides a broad overview of the Risk Index over the entire period from 2006 to 2021, with the average Risk Index for each day represented by the blue line. To facilitate clearer interpretation, a 6-month rolling average is included, shown by the dashed red line. The plot highlights long-term trends in risk levels, showing periods of heightened and reduced risk. Although there are some fluctuations, the rolling average smooths these variations, enabling us to observe the general trend over time. The x-axis labels every three years for better clarity, making it easier to track shifts in risk levels. From this, we can discern if certain years or periods saw notable increases or decreases in the Risk Index, possibly suggesting underlying external factors like policy changes or events that affected the risk.

4.1.2.2 Year 2020

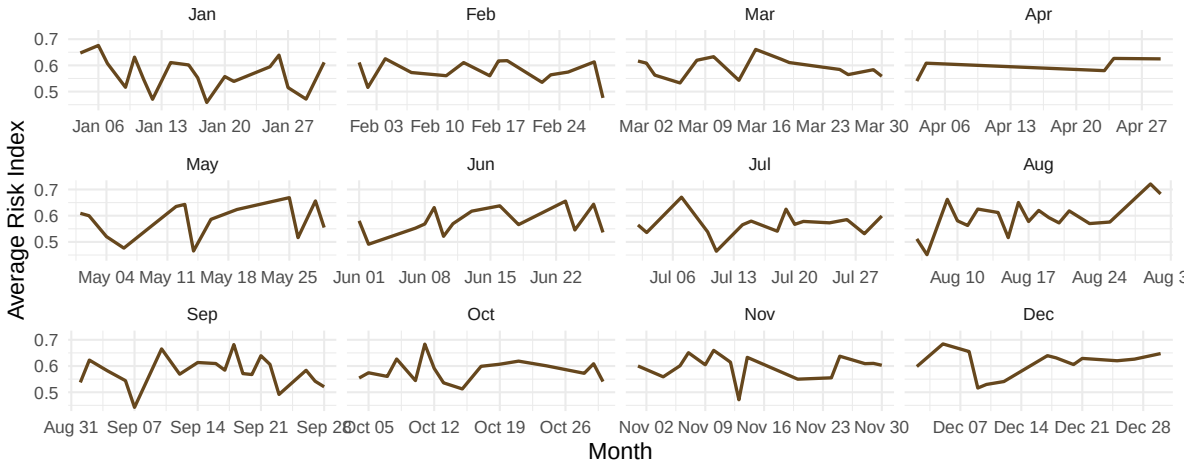


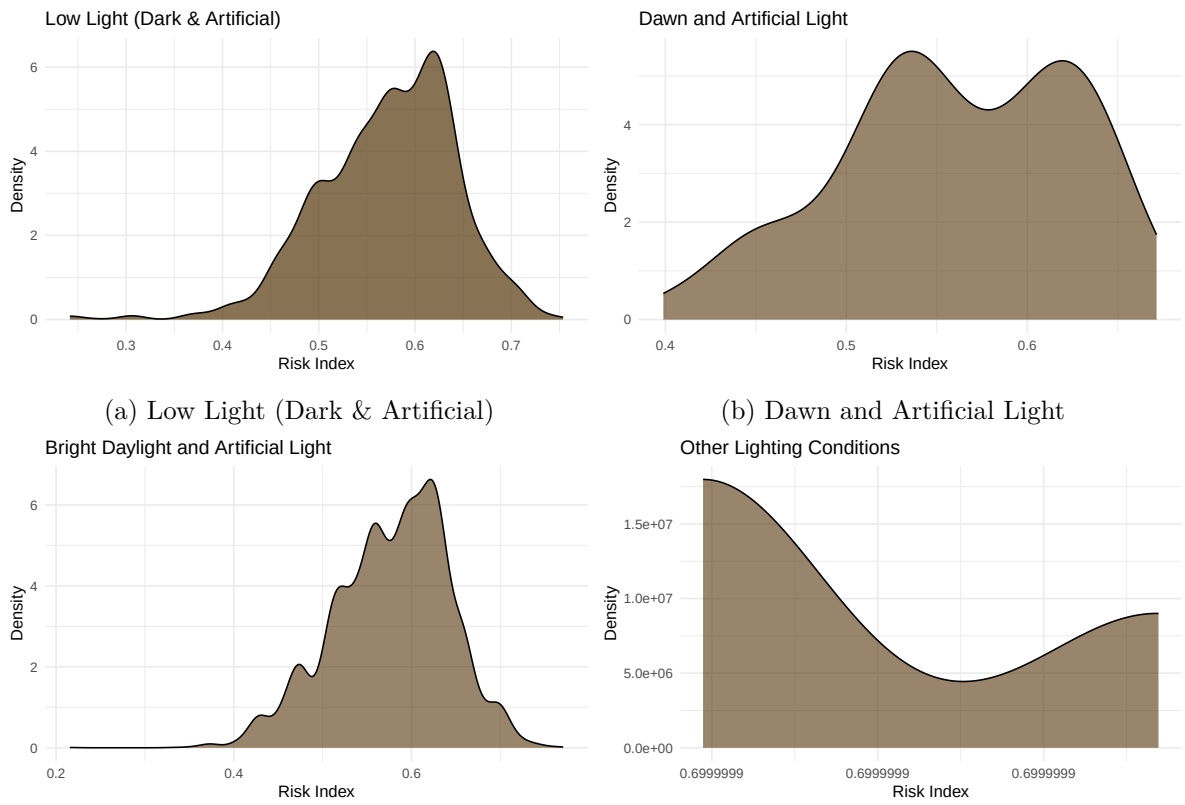
Figure 7: Risk Index Variations in 2020

Figure 7 provides a detailed, month-by-month breakdown of the Risk Index for the year 2020. This faceted layout offers a clear view of short-term fluctuations and potential seasonal patterns by presenting trends for each month separately. The graph highlights variations in risk levels across the months, making it easier to identify distinct periods of higher or lower risk. These trends may correspond to factors such as weather conditions, holidays, or other contextual events. For instance, spikes or dips in risk within specific months could reflect changes in traffic volume, seasonal behaviors, or other dynamic factors influencing motorbike-related incidents throughout the year.

4.2 Conditions of Danger: Environmental Drivers of Risk

Understanding how environmental and situational factors contribute to variations in the Risk Index is crucial for identifying the conditions under which motorbike risks are heightened. This section explores how different factors, such as road surface conditions, lighting conditions, and traffic control types, influence the Risk Index. By breaking down the data into categories based on these factors, we can reveal specific conditions that might increase the likelihood of accidents or theft. The following visualizations provide insight into the distribution of risk under varying environmental conditions, highlighting the factors that contribute to motorbike-related incidents.

To better understand how environmental factors influence the Risk Index, we examine the distribution of the Risk Index under various road surface conditions. This analysis explores whether different surface types—such as dry, wet, icy, and others—affect the level of risk associated with traffic incidents. The following violin plot visually compares the Risk Index across different road surface conditions, providing insight into how these factors may contribute to the overall risk.



(a) Low Light (Dark & Artificial)

(b) Dawn and Artificial Light

(c) Bright Daylight and Artificial Light

(d) Other Lighting Conditions

Figure 8: Risk Index by Lighting Conditions

Figure 8 display the distribution of the Risk Index across various lighting conditions. The data is broken down into four main lighting categories: Low Light (Dark & Artificial), Dawn and Artificial Light, Bright Daylight and Artificial Light, and Other Lighting Conditions.

Figure 8a highlights the risk distribution during low-light conditions, including both "Dark" and "Dark, Artificial" categories. The density curve shows that motorbike risks are relatively higher in these conditions, particularly in the moderate-to-high range of the Risk Index, indicating that poorer visibility and the presence of artificial lighting increase the likelihood of incidents.

Figure 8b captures the risk profile during dawn and artificial lighting. The distribution shifts slightly, showing a lower concentration of higher risk indices compared to the low-light category. While there is still some risk during dawn, the overall risk index appears to be less extreme compared to the "Low Light" category, possibly due to the improving visibility as the sun rises.

Figure 8c shows a much lower density of high-risk events. This suggests that the highest risk does not generally occur during daylight hours, as visibility is much better, and traffic conditions tend to be more predictable.

Figure 8d captures a small number of cases that don't fit into the above categories. The risk distribution in this category is less pronounced, likely because it includes a variety of edge cases, such as unusual lighting conditions or missing data.

4.3 Mapping the Risk: Neighborhood-Level Insights

4.3.1 Neighborhood Risk Distribution

Figure 9 illustrates the distribution of neighborhoods in Toronto grouped by their average collision probabilities into three risk categories: High, Medium, and Low. Each panel represents one of the risk categories and shows the range of collision probabilities (x-axis) and the corresponding number of neighborhoods (y-axis) in that category. For example, the High-risk group predominantly consists of neighborhoods with collision probabilities between 0.8 and 1.0, indicating a clustering of high probabilities in these areas. The Medium-risk group primarily includes neighborhoods with probabilities between 0.6 and 0.8, while the Low-risk group has neighborhoods with probabilities below 0.6.

This visualization aligns with the study's focus on spatial patterns, as it highlights that collision risks are not evenly distributed across neighborhoods. The results suggest targeted interventions could be more effective when tailored to the risk level of specific areas, such as road safety measures in neighborhoods with high collision probabilities.

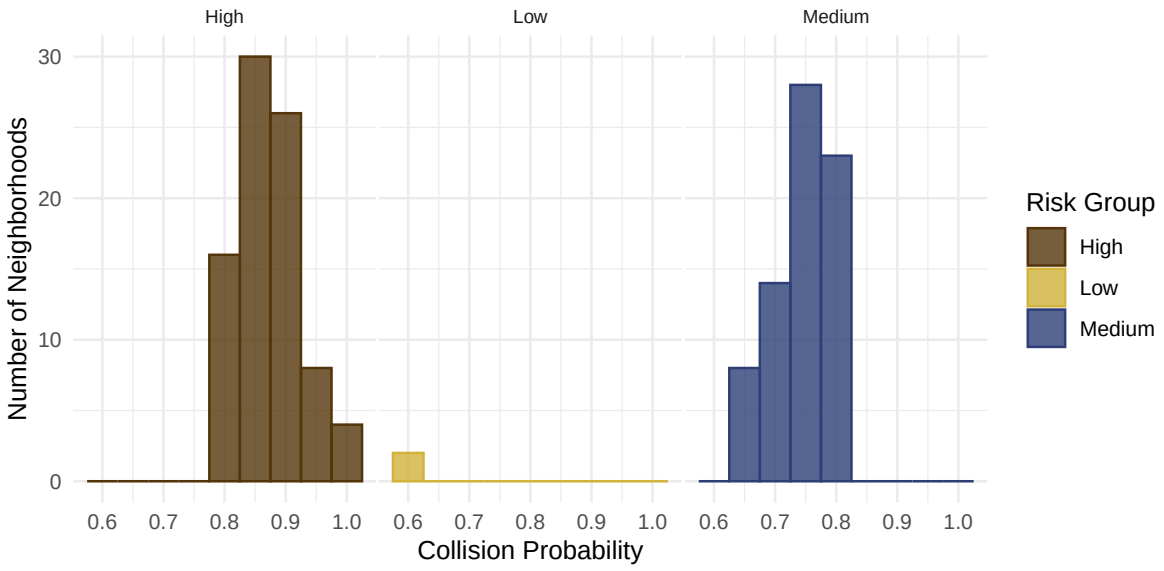


Figure 9: Neighborhoods grouped by average collision probabilities into High, Medium, and Low risk categories.

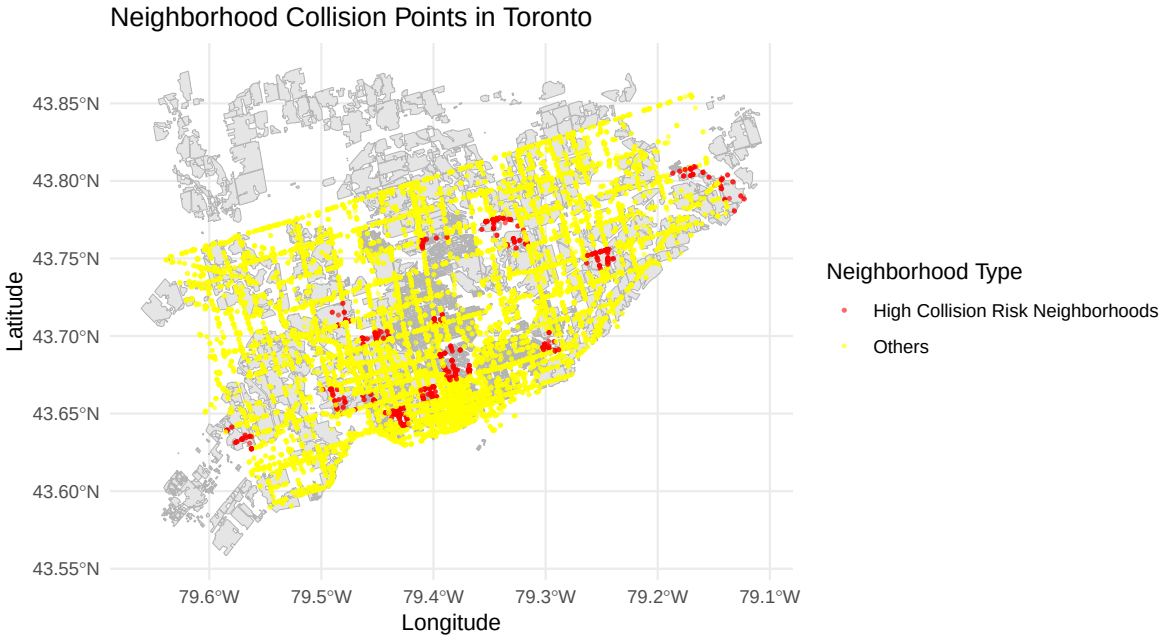


Figure 10: Neighborhood Collision Points in Toronto Highlighting High Collision Risk Areas

4.3.2 Maps

Figure 10 illustrates the distribution of collision points across Toronto, emphasizing neighborhoods categorized as "High Collision Risk" in red and other areas in yellow. The spatial extent spans latitudes from approximately 43.55°N to 43.85°N and longitudes from -79.6°W to -79.1°W, covering most of Toronto's urban landscape.

The high collision risk neighborhoods are clustered primarily in specific areas, as denoted by the red points. These regions often correspond to densely populated or high-traffic zones, where the likelihood of traffic collisions is significantly elevated. In contrast, the yellow points represent areas with lower collision risks, which are more dispersed across the map, particularly in less congested parts of the city.

This map provides a critical spatial perspective, highlighting the geographic disparities in traffic collision risks within Toronto. It underscores the need for targeted safety interventions, such as traffic calming measures and enhanced infrastructure, in the high-risk neighborhoods. The visualization also serves as a tool for urban planners and policymakers to identify and prioritize areas where safety improvements could significantly reduce the frequency and severity of traffic collisions, fostering a safer urban environment.

5 Discussion

5.1 Temporal Risk Patterns and Behavioral Recommendations

Focus: This section can interpret the temporal trends in the Risk Index and their implications for motorbike owners and policymakers. For instance, if risks peak at certain hours or days, actionable recommendations can be made for owners to avoid high-risk times or take precautions.

Key Points: Discuss high-risk times of day or days of the week for both theft and collisions. Recommendations for motorbike owners to reduce risk exposure, such as avoiding specific time periods or enhancing security measures during peak theft hours. Potential scheduling adjustments for law enforcement patrols to align with high-risk periods..

5.2 Environmental and Situational Influences on Risk

Focus: This section can examine how road, lighting, and traffic conditions contribute to the Risk Index, linking these findings to actionable changes in infrastructure or urban planning.

Key Points: Discuss the significant environmental predictors of risk, such as poor road surface conditions or inadequate lighting. Recommend urban planning measures, such as improved road maintenance or installation of streetlights, to mitigate risks. Highlight how different

environmental factors interact to amplify risks, suggesting multi-faceted approaches to improve safety.

5.3 Policy Implications of Spatial Risk Disparities

Focus: This section can discuss the significant disparities in the Risk Index across neighborhoods. Highlight how certain areas are disproportionately affected by theft or collisions and the potential socioeconomic or infrastructure factors contributing to these risks.

Key Points: Prioritizing high-risk neighborhoods for targeted interventions, such as improved street lighting or traffic control measures. Suggestions for law enforcement strategies to reduce theft in hotspots. Need for localized awareness campaigns in high-risk areas.

5.4 Limitations

5.5 Future Research

Future Work1: Explore how temporal risk patterns change with seasons or special events (e.g., holidays, festivals) to refine recommendations further.

Future Work2: Develop predictive models integrating weather or real-time traffic data to dynamically assess and communicate motorbike risks.

Future Work3: Investigate the correlation between neighborhood-level Risk Index values and broader social or economic indicators, such as income levels or population density.

Appendix

A Additional Data Details

A.1 Data Cleaning

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

C Idealized Methodology for a Survey on Motor Risk

C.1 Survey Overview

This survey aims to assess factors contributing to motorbike risks, including theft and collisions, by gathering data from motorbike owners and riders. The survey will explore demographic characteristics, riding behavior, motorbike usage patterns, and perceptions of environmental and situational risks.

C.2 Sampling Approach

The survey will use a stratified random sampling method to ensure diverse representation across:

- Geographic regions (urban, suburban, rural areas).
- Demographic groups (age, gender, income, and education levels).
- Riding experience (novices, intermediate, and experienced riders).

The target population includes licensed motorbike riders and owners. A sample size of approximately 1,000 respondents is proposed to ensure statistical validity across strata.

C.3 Survey structure

The survey will consist of five main sections:

- Demographics: Basic information on respondents (e.g., age, gender, income, education, geographic location).
- Riding Behavior: Frequency, duration, and purpose of motorbike usage.
- Risk Awareness and Perception: Personal experiences with collisions or theft and perceptions of environmental risks.
- Situational Factors: Details of riding conditions such as time of day, weather, road surface, and lighting.
- Preventive Measures: Actions taken by riders to mitigate theft or collision risks (e.g., use of locks, helmets, or avoiding high-risk areas).

C.3.1 Question Types

- Closed-ended questions: For quantitative data (e.g., multiple choice, Likert scales, ranking).
- Open-ended questions: To capture nuanced insights and personal experiences.
- Matrix questions: To evaluate attitudes across multiple dimensions efficiently.

C.3.2 Question List

Here are examples of survey questions:

Demographics: What is your age group? (e.g., 18–24, 25–34, etc.) What is your highest level of education completed?

Riding Behavior: How often do you ride your motorbike? (Daily, Weekly, Monthly, Rarely) For what purposes do you primarily use your motorbike? (Commuting, Recreation, Delivery/Work, Other)

Risk Awareness and Perception: On a scale of 1–5, how would you rate the theft risk in your neighborhood? Have you experienced a motorbike theft or collision in the past year? (Yes/No)

Situational Factors: What time of day do you usually ride? (Morning, Afternoon, Evening, Night) In what weather conditions do you typically ride? (Clear, Rainy, Snowy)

Preventive Measures: What measures do you take to secure your motorbike against theft? (Select all that apply: Lock, Alarm, GPS Tracker, Parking in Secure Locations) Do you avoid specific times or areas due to perceived collision risks? (Yes/No)

C.4 Recruitment Strategy

Participants will be recruited through a combination of online and offline channels:

Online platforms: Motorbike owner forums, social media groups, and email lists of motorbike organizations.

Offline channels: Flyers at motorbike dealerships, repair shops, and riding schools. Incentives such as small gift cards or entry into a raffle may be provided to encourage participation.

C.5 Linkage to Literature

The survey design is informed by prior studies on motor vehicle risk perception, road safety, and crime prevention. Key references include:

Research on environmental and situational predictors of road accidents. Studies on the effectiveness of theft prevention measures. Literature on sampling methods and survey design for risk assessment.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Nick, Todd G., and Kathleen M. Campbell. 2007. "Logistic Regression." In *Topics in Biostatistics*, 273–301. Springer.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Yasmin, Shamsunnahar, and Naveen Eluru. 2016. "Latent Segmentation Based Count Models: Analysis of Bicycle Safety in Montreal and Toronto." *Accident Analysis & Prevention* 95: 157–71. https://doi.org/10.1016/j.aap.2016.06.015.