Motor Vehicle Risks in Toronto: Uncovering Collision and Theft Patterns*

Risk Peaks in the Morning on Dark, Snow-Packed Roads and in High Urban Density Areas

Yingke He

December 14, 2024

This study examines motor vehicle theft and traffic collisions in Toronto, focusing on spatial and temporal patterns,and environmental factors. A composite risk score model is used and highlights that collisions occur more frequently under unfavorable road conditions, such as poor lighting and wet or icy surfaces, and in areas with inadequate traffic control measures. These findings emphasize the need for targeted interventions, including enhanced road infrastructure and lighting in collision-prone zones. Such measures aim to improve safety and security for motorbike riders and owners, contributing to safer 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. Furthermore, studies have shown that socioeconomic conditions and local enforcement of traffic laws significantly influence the incidence of motor theft and collisions, highlighting the need for multi-faceted strategies to address these risks (Charron 2009; Law and Petric 2024).

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 motor collisions, and a theft index is calculated which are then combined with the likelihood of motor collisions into a single, interpretable composite risk score. This metric 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) theft risks vary based on temporal factors, such as time of day and day of the week, with mornings exhibiting higher susceptibility to theft, while early mornings have lower theft risks; (2) Collision risks are strongly influenced by environmental and situational conditions, with risks being particularly high under poor visibility. Factors such as wet or icy road surfaces and inadequate traffic controls further significantly increase the likelihood of incidents; and (3) neighborhood characteristics play a critical role in shaping both theft and collision risks, with high-collision-risk neighborhoods primarily clustered in high-traffic zones and a notable concentration of collision incidents in downtown Toronto, reflecting the impact of dense urban environments and heavy traffic flows on risk levels.

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 probability of collision, as well as the method

used to derive the theft index and combine these probabilities into a composite risk score. Section 4 then presents the main findings, including understandings 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

This project is motivated and guided by Rohan Alexander and his book (Alexander 2023). Data used in this paper was cleaned, analyzed and modeled with the programming language R (R Core Team 2023). Also with support of additional packages in R: `readr` (Wickham and Hester 2023), `ggplot2` (Wickham 2016), `tidyverse` (Wickham, Averick, et al. 2023), `dplyr` (Wickham, François, et al. 2023), `here` (Müller 2022), `knitr` (Xie 2023), `kableExtra` (Zhu 2023), `palmerpenguins` (Horst, Hill, and Eynden 2023), `performance` (Lüdecke et al. 2023), `tidygraph` (Pedersen 2023b), `ggraph` (Pedersen 2023a), `ggridges` (Wilke 2021), `lubridate` (Grolemund and Wickham 2011), `sf` (Pebesma 2018), `osmdata` (Padgham and Super 2023), and `car` (Fox and Weisberg 2023).

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

2.1 Source

This study utilized two datasets published by the Toronto Police Service, available from Open Data Toronto (Gelfand 2020). 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 solid geographic analyses of theft 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. An interactive visualization of the final risk index can be found in Appendix A.

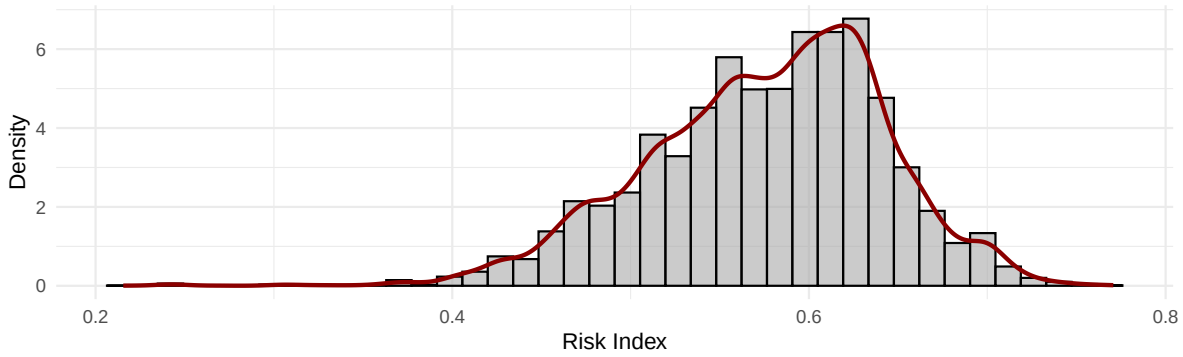


Figure 1: Distribution of the Overall Risk Index

Figure 1 shows the distribution of the Risk Index across all observations in the dataset. The Risk Index shows 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 indicates 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 increase the risks in specific cases. This idea can inform targeted interventions, focusing on the outliers with high Risk Index values to mitigate the most significant risks.

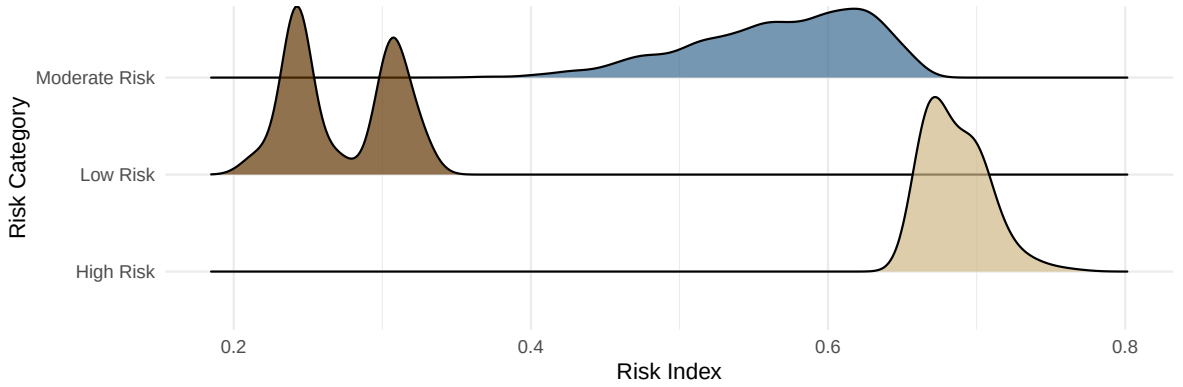


Figure 2: Risk Index Distribution by Risk Category

Building on the distribution of the Risk Index in Figure 2 highlights the density of the Risk

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

The High Risk category demonstrates a single, narrow peak at relatively higher Risk Index values, indicating a concentrated group of neighborhoods with consistently elevated risk levels. This density suggests a homogeneity in the high-risk group, where most neighborhoods share similar risk characteristics. Conversely, the Low Risk category shows a bimodal distribution, with two distinct peaks at much lower Risk Index values. This indicates variability within this category, with some neighborhoods experiencing very low risk and others clustering near the moderate range. The Moderate Risk category displays a more diffuse distribution, with a broad range of Risk Index values extending into both the High Risk and Low Risk ranges. This variability suggests that neighborhoods in the Moderate Risk category experience diverse risk profiles, likely shaped by a combination of environmental, temporal, 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

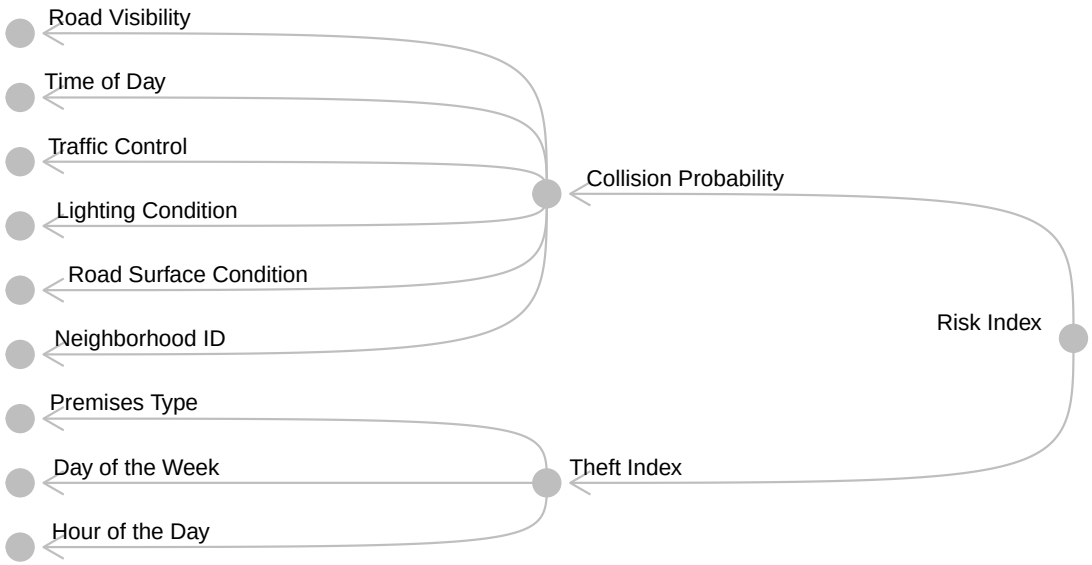


Figure 3: Hierarchical relationship between risk factors contributing to theft and collision probabilities.

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

The **Hour of the Day** variable is used as a predictor to analyze temporal trends in motorbike theft occurrences. This variable helps identify whether certain hours are associated with elevated or reduced theft risks. Contributing factors may include decreased monitoring during nighttime hours, increased activity in high-risk areas during specific times, or patterns related to commuter schedules. Figure 4 outlines the distribution of bicycle theft incidents across different hours of the day in a 24-hour format. This temporal variable helps the model capture time-dependent patterns in theft occurrences, identifying periods of heightened risk, such as late morning to early afternoon (10:00 to 15:00), and lower-risk periods, such as early morning hours (02:00 to 06:00).By including this variable, the theft index model utilizes the observed trend of thefts being more frequent during morning and early afternoon hours (7:00 to 13:00) while being significantly lower during the early morning hours (02:00 to 06:00). This temporal variability highlights specific periods of increased theft risk, enabling the model to capture these patterns effectively. Understanding this trend allows for better predictions of theft dynamics and offers meaningful ideas for designing targeted interventions and optimizing resource allocation based on the time of day.

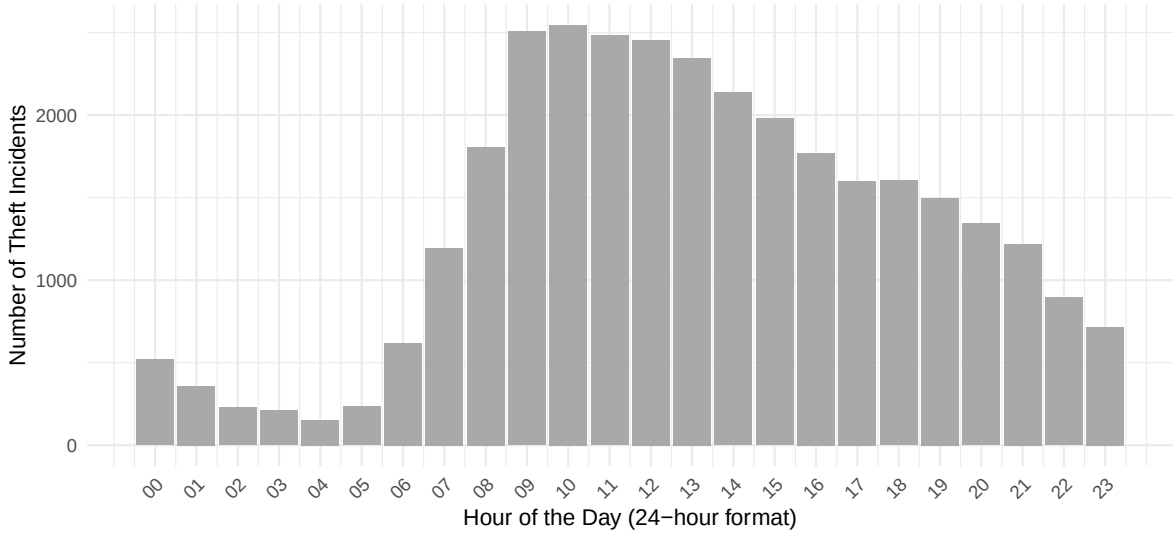


Figure 4: Theft Incidents by Hour of the Day

2.4.1.2 Day of the Week

The **Day of the Week** variable is included as a predictor to analyze weekly trends in motorbike theft incidents. Figure 5 visualizes the distribution of thefts across the week, revealing that incidents tend to peak on Monday and Tuesday, with slightly lower occurrences on weekends, particularly Saturday and Sunday. This predictor captures potential variations tied to weekly routines, such as higher theft risks during the start of the workweek when urban activity is higher, or reduced risks over the weekend when monitoring or exposure may change. By incorporating this variable, the model can better account for these temporal patterns, enhancing its ability to predict theft incidents based on the day of the week.

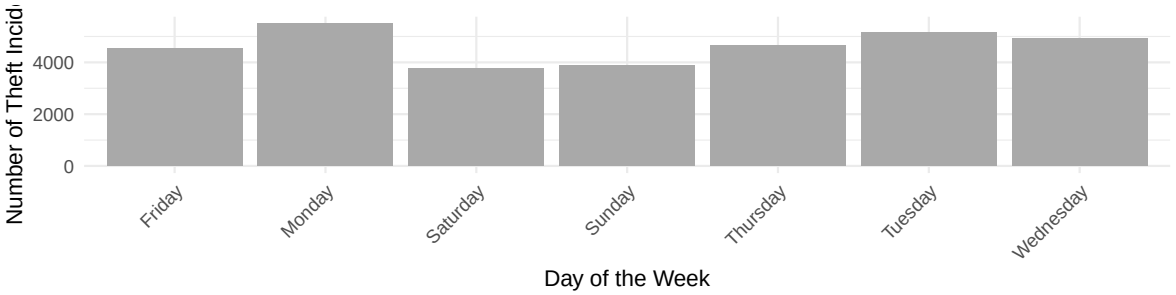


Figure 5: Theft Incidents by Day of the Week

2.4.1.3 Premises Type

The **Premises Type** variable is included as a predictor to examine the contextual environments where motorbike theft incidents are most prevalent. This variable categorizes thefts by location type, such as outdoor spaces, residential areas, or commercial properties, to identify settings associated with higher or lower risks. Figure 6 highlights that outdoor locations account for the majority of thefts, followed by residential areas like houses and apartments, while locations such as educational and transit premises illustrates much lower incident counts. These patterns suggest that factors such as accessibility, lack of surveillance, and the density of parked motorbikes significantly influence theft risks across different premises types.

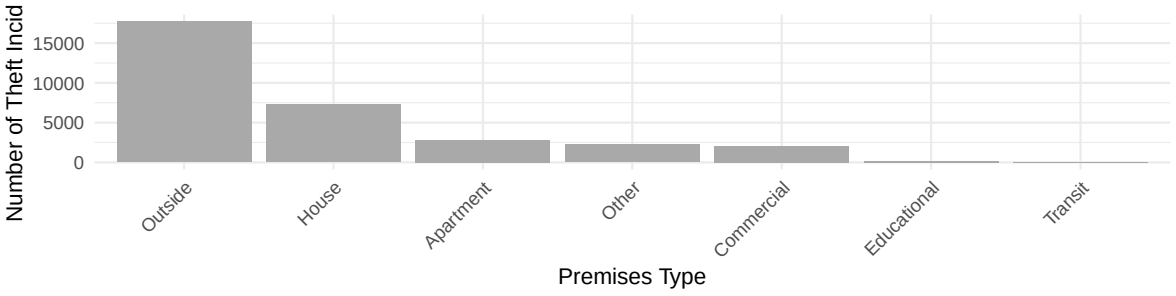


Figure 6: Theft Incidents by Day of the Week

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 provides an understanding into the contextual factors that contribute to traffic incidents, capturing the dynamic interaction between environmental factors and human behavior.

2.4.2.1 Neighborhood ID

The **Neighborhood ID** variable serves as a categorical predictor, uniquely identifying neighborhoods within the City of Toronto. This variable captures spatial differences in motorbike theft and collision risks, facilitating the identification of localized patterns and trends. Table 1 provides examples of Neighborhood IDs and their corresponding incident counts, illustrating the variation in theft incidents across different areas. This data highlights neighborhoods with high or low incident rates, enabling the model to incorporate spatial variability and improve predictions by understanding how collision and theft risks are distributed geographically.

Table 1: 10 Examples of Neighborhood Incidents Count

Neighborhood ID	Incident Count
1	597
170	376
119	361
70	353
85	304
140	25
173	25
29	20
67	20
114	18

2.4.2.2 Road Surface Condition

The **Road Surface Condition** variable serves as a categorical predictor, detailing the state of the road at the time of an incident. Categories include dry, wet, loose snow, slush, ice, packed snow, and others. This variable enables the analysis to assess how different surface conditions contribute to motorbike collision risks. Figure 7 illustrates the number of incidents associated with different road surface conditions. Most incidents occur on dry roads, followed by wet roads, likely due to their higher frequency of use. In contrast, conditions such as loose snow, ice, and packed snow account for a smaller share of incidents, possibly because they occur less frequently.

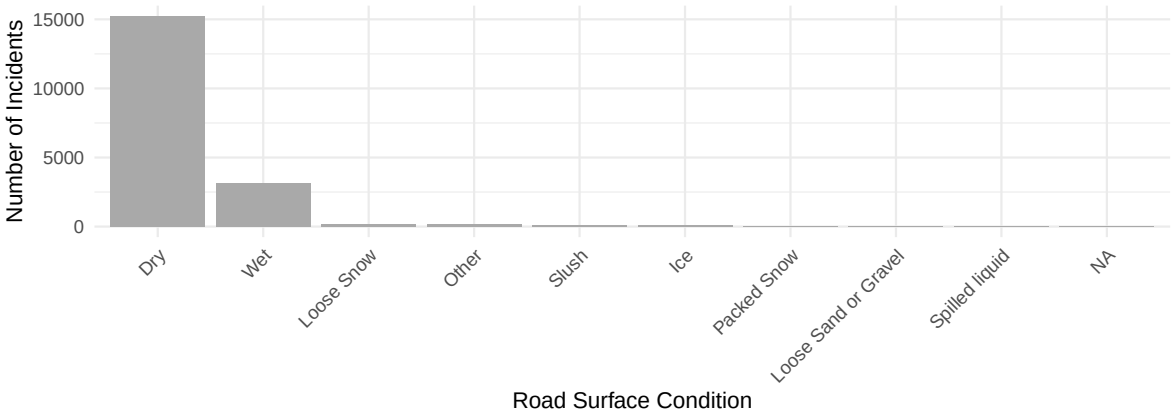


Figure 7: Incidents by Road Surface Condition

2.4.2.3 Lighting Condition

The **Lighting Condition** variable is a categorical predictor that describes the lighting environment at the time of an incident, including categories such as daylight, darkness, artificial lighting, and transitional periods like dawn and dusk. This variable helps analyze how visibility and lighting conditions impact the likelihood of motorbike thefts and collisions.

Figure 8 illustrates that the majority of incidents occur during daylight hours, likely due to higher traffic volumes and activity levels during the day. Incidents under "Dark" and "Dark, artificial" conditions are less frequent but still significant, highlighting the increased risks associated with poor visibility. Transitional conditions like "Dawn" and "Dusk" contribute minimally to the total number of incidents. These findings emphasize the need for targeted safety measures, such as enhancing artificial lighting and promoting driver vigilance in low-visibility conditions, to reduce collision risks effectively.

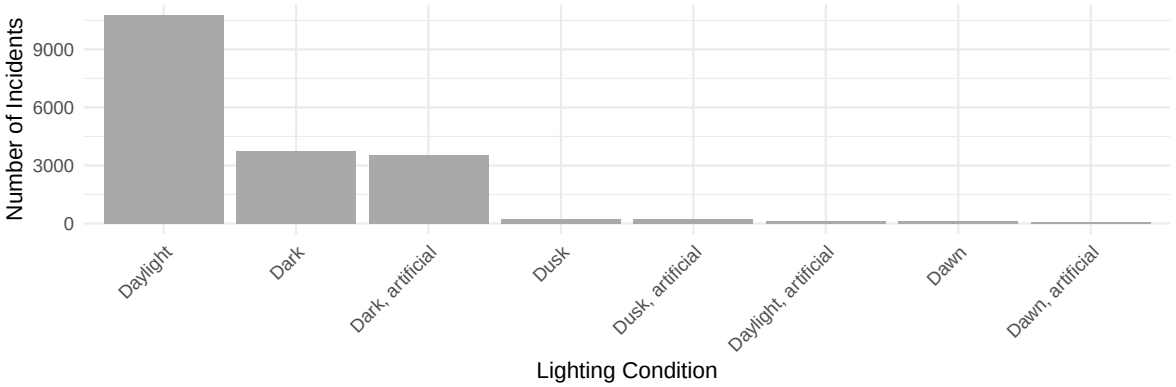


Figure 8: Incidents by Lighting Conditionk

2.4.2.4 Traffic Control

The **Traffic Control** variable categorizes the type of traffic management present at the site of each incident, such as "No Control," "Traffic Signal," or "Stop Sign." This variable provides ideas into how various traffic control measures influence the likelihood of motorbike collisions.

As shown in Table 2, the majority of incidents occur in areas with "No Control" (9,021 incidents), indicating that the absence of traffic regulation significantly contributes to collision risks. Areas with "Traffic Signal" also account for a substantial number of incidents (8,035), likely due to the higher traffic volumes typically associated with signalized intersections. Other traffic control types, such as "Stop Signs" (1,464 incidents) and "Pedestrian Crossovers" (208 incidents), are associated with considerably fewer incidents, reflecting their more localized application.

Table 2: Summary of Traffic Control

Traffic Control	Incident Count
No Control	9021
Traffic Signal	8035
Stop Sign	1464
Pedestrian Crossover	208
Traffic Controller	108
NA	75
Yield Sign	21
Streetcar (Stop for)	16
Traffic Gate	5
Police Control	2
School Guard	2

2.4.2.5 Time of Day

The **Time of Day** variable records the exact hour and minute when a collision occurred, represented in a 24-hour format. For example, "2216" corresponds to 10:16 PM, and "807" indicates 8:07 AM. This variable enables the analysis of temporal patterns in collision risks, identifying times of higher or lower incident probabilities.

Factors influencing these patterns may include reduced visibility during nighttime hours, increased traffic during peak commuting times, or variations in driver behavior throughout the day. Table 3 illustrates examples of incident counts at specific times, providing insights into how collisions are distributed across the day. Analyzing these temporal trends helps to identify high-risk periods, which can guide the implementation of targeted safety measures, such as increased monitoring or public awareness campaigns during specific hours.

Table 3: Examples of Different Times of Day with Incident Count

Time of Day	Incident Count
2004	12
1018	19
2337	5
313	5
2102	11

2.4.2.6 Visibility

The **Visibility** variable is a categorical predictor that captures environmental visibility conditions at the time of a collision, such as clear, rain, snow, fog, or strong winds. This variable is essential for understanding how varying visibility levels influence collision risks. Reduced visibility conditions, such as those caused by rain or snow, can impair drivers' ability to detect road hazards or other vehicles, increasing the likelihood of incidents. Conversely, clear conditions often correlate with safer driving environments.

Figure 9 summarizes the distribution of collisions across different visibility conditions, revealing that the vast majority of incidents occur during clear conditions. This is likely due to the overall higher frequency of clear weather compared to adverse conditions. Rain accounts for the second-highest number of incidents, while other conditions, such as snow, fog, and strong winds, contribute to a much smaller share of collisions.

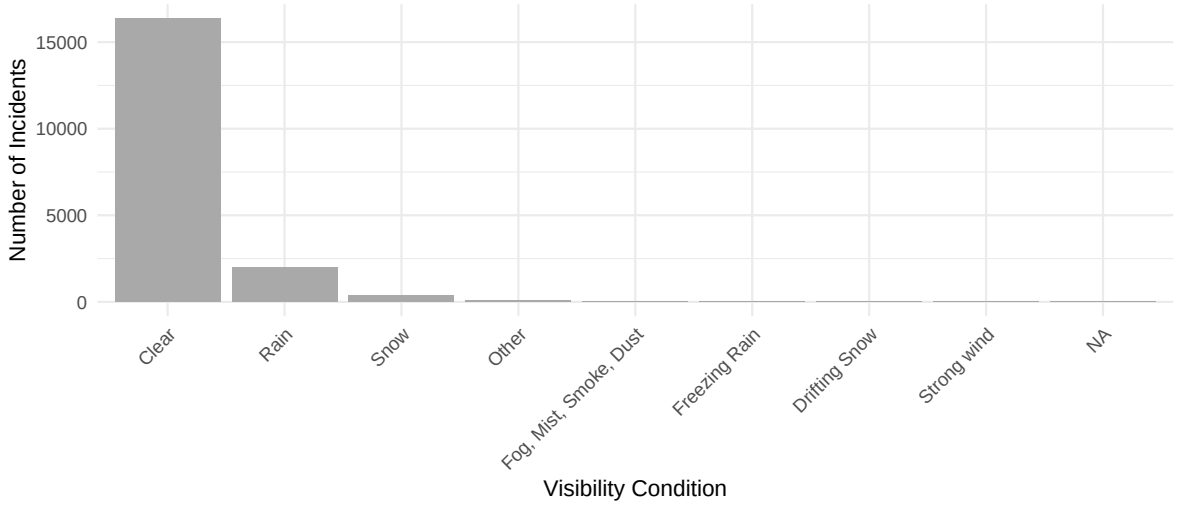


Figure 9: Incidents by Visibility Condition

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.

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 scheme of 0.3 for theft and 0.7 for collision, emphasizing collision risk while ensuring theft risk is considered. By combining these components, the Risk Index offers a unified measure of motorbike-related risks, enabling stakeholders to evaluate and compare safety conditions across different contexts.

The model assumes independence between theft and collision risks and relies on proportional theft representations, which may limit its ability to capture interactions or account for unmeasured confounders. **Alternative methods** like generalized additive models or machine learning techniques could address these limitations but were not used to prioritize interpretability and simplicity.

This modeling approach is justified by its ability to adapt to the data's constraints while maintaining statistical rigor and interpretability. The sub-index approach for theft risk is designed to align with the characteristics of the dataset, avoiding assumptions about unobserved cases, and the use of logistic regression for collision risk ensures solid and reliable predictions. Together, these components form a 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

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.To validate this approach, the calculated values of the theft sub-index were compared against historical crime data trends, confirming that the observed proportions align with real-world patterns of theft distribution across temporal and spatial categories.

3.3.2 Composite Risk Index

For the assumption of the composite risk index, it is assumed that the weights assigned to the collision probability and theft component reflect their relative importance in contributing to the overall risk, based on the severity and frequency of these events in real-world scenarios. 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:

$$w_1 \text{ and } w_2$$

to evaluate the 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 remains valid for estimating probabilities of severe collisions and provides a reliable contribution 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 (Stoltzfus 2011). 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 thoughts into whether the relationship is approximately linear.

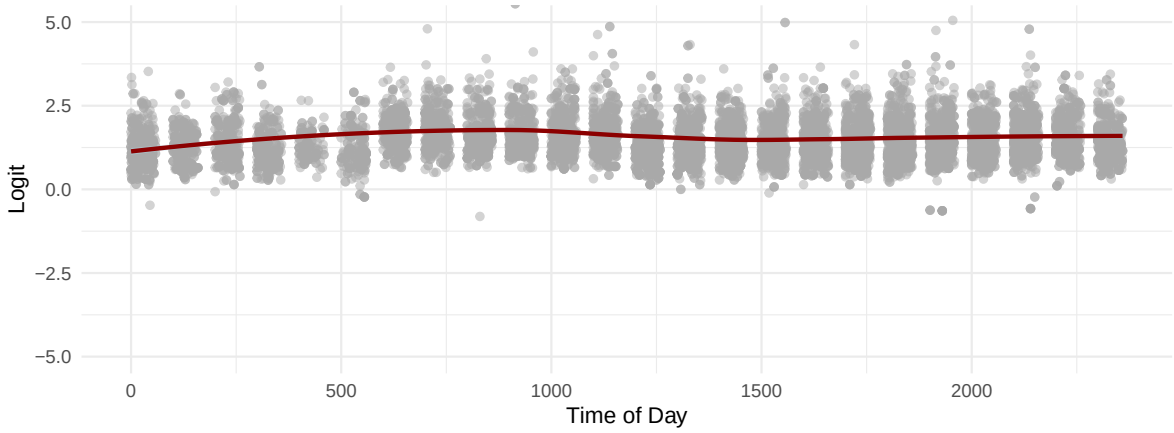


Figure 10: Logit Plot for Time of Day in the Collision Probability Model

Figure 10 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 (Stoltzfus 2011).

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 (Stoltzfus 2011). 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 4: 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 4. 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 solid 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

4.1.1 Time-of-Day Analysis

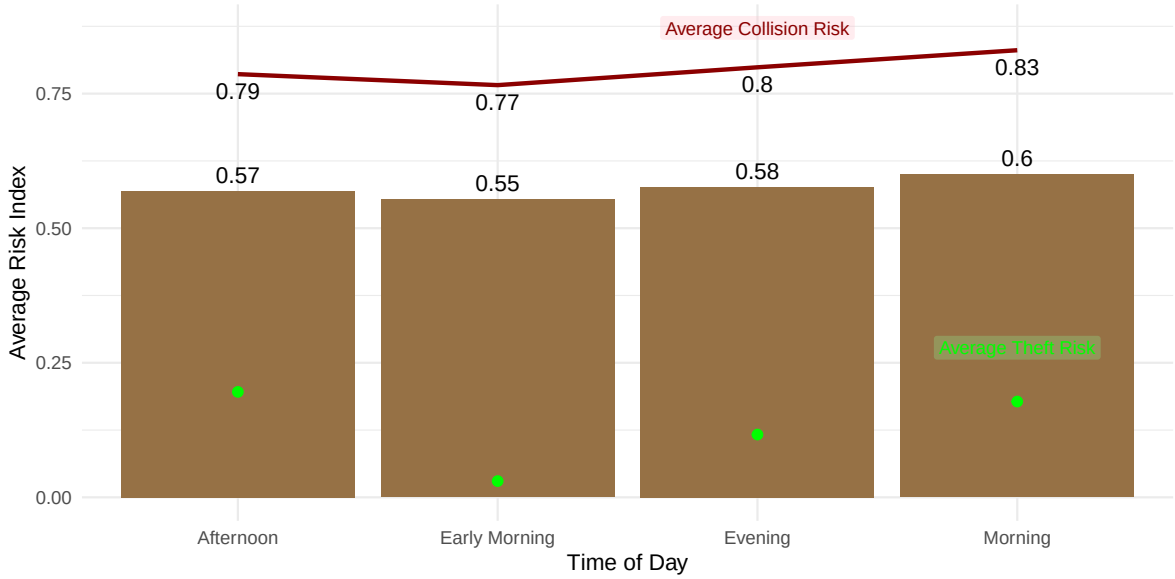


Figure 11: Risk Index by Time of Day, with bars showing Overall Risk Index, the red line showing average Collision Risk, and green dots showing average Theft Risk.

Figure 11 illustrates the temporal variations in the overall Risk Index, divided into four time-of-day segments: Early Morning, Morning, Afternoon, and Evening. Collision risk illustrates a clear temporal pattern, with peaks during commuting hours—specifically in the Morning (7–9 AM) and Evening (5–7 PM), corresponding to periods of high traffic density. This indicates that collision risk is closely tied to traffic patterns, greater caution should be imposed during these periods to reduce the likelihood of collisions and enhance overall road safety. In contrast, theft risk remains relatively stable across all time bins, with no significant fluctuations. For better visualization, theft risk values have been scaled by a factor of 10, highlighting its consistently smaller contribution to the overall Risk Index.

The highest overall Risk Index is observed in the Morning bin, driven by elevated collision risk during this time. Afternoon and Evening bins show slightly lower overall Risk Index values, reflecting moderate variations in collision risk. The Early Morning bin has the lowest Risk Index, corresponding to minimal traffic activity and theft incidents.

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 the year 2020 during covid outbreaks. 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 year 2020.

4.1.2.1 Longterm

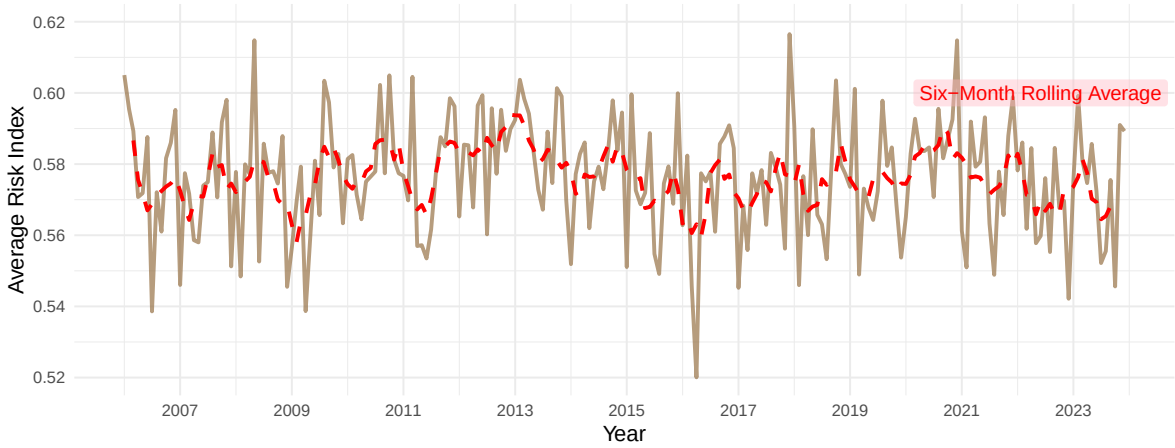


Figure 12: Long-Term Temporal Trends in Risk Index (Mid-2016 to Mid-2024) with Rolling Average

Figure 12 depicts the long-term trends in the Average Risk Index, measured daily and displayed from mid-2006 to mid-2023. The brown solid line represents the daily Average Risk Index, while the red dashed line illustrates a six-month rolling average, providing a smoother depiction of overall patterns.

The daily Average Risk Index fluctuates significantly, reflecting changes in factors such as collision and theft occurrences over time. These short-term variations could be associated with seasonal patterns, weather changes, or traffic conditions. Meanwhile, the rolling average highlights broader, more consistent patterns, indicating stable long-term trends without dramatic increases or decreases. Periodic dips and peaks, such as those observed in late 2016 and mid-2018, emphasize the importance of addressing temporary shifts in risk to support safer conditions.

4.1.2.2 Covid Period

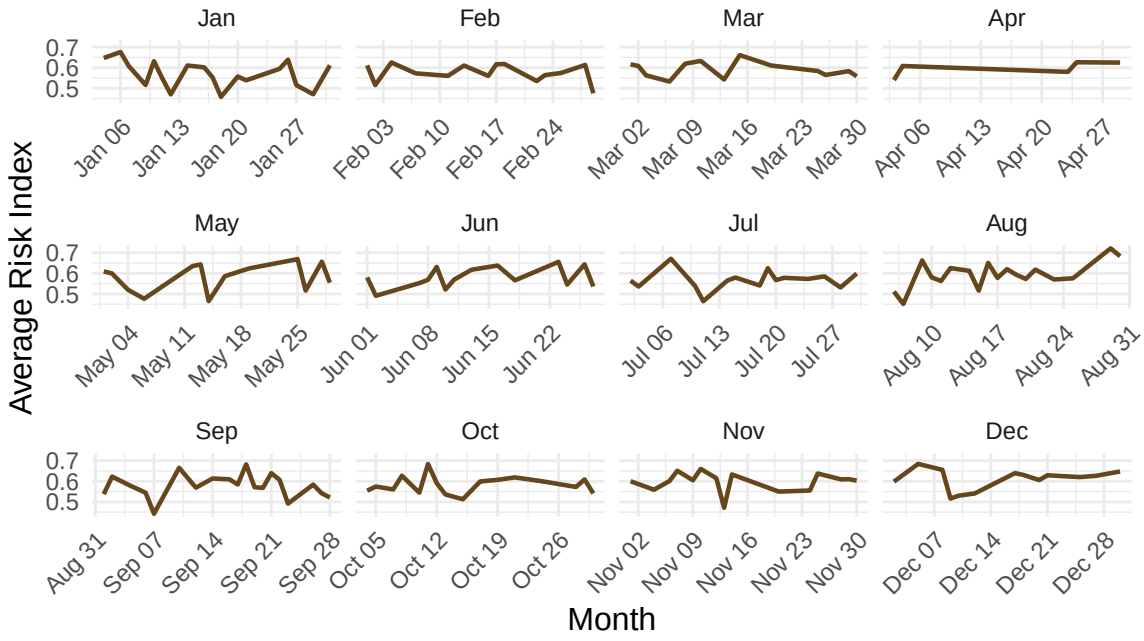


Figure 13: Risk Index Variations in 2020 during Covid-19

Figure 13 presents a detailed, month-by-month breakdown of the Risk Index for the year 2020. The graph highlights variations in risk levels across months, which may correspond to factors such as weather conditions, holidays, or other contextual events. For instance, dips in early September and November could relfect changes in traffic volume, seasonal behaviors, or shifts in theft or collision dynamics approaching the end of the year.

The year 2020 coincides with the beginning of the COVID-19 pandemic, which had a strong impact on urban mobility patterns. Lockdowns, social distancing measures, and reduced economic activities likely led to significant decreases in traffic volumes and changes in behavior related to motor vehicle usage (Buehler and Pucher 2021). However, the Risk Index remains relatively consistent throughout the year, with only minor fluctuations across months. This relative stability may suggest that while pandemic-related restrictions likely influence traffic and mobility patterns, these changes did not significantly impact the overall Risk Index for the year. The consistent trends could reflect a balance between reductions in traffic collisions due to decreased activity during lockdown periods and a stable baseline for theft and other risk factors.

4.2 Conditions of Danger: Environmental Drivers of Risk

This section examines how factors such as road surface conditions, lighting conditions, and traffic control types contribute to changes in the Risk Index. Categorizing the data based on these factors illustrates specific conditions under which the likelihood of collisions or theft increases.

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 understandings into how these factors may contribute to the overall risk.

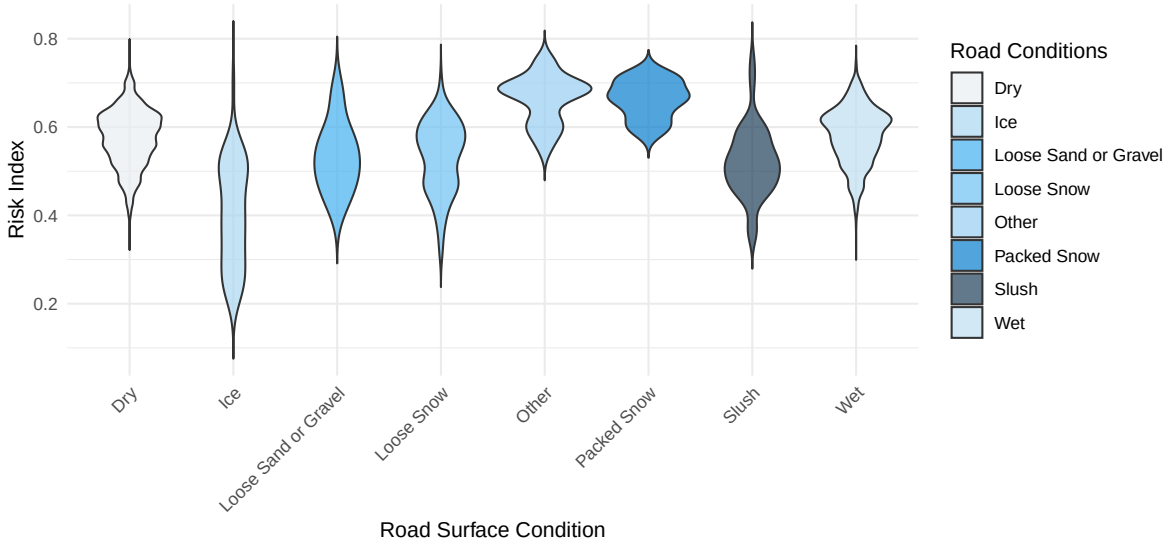
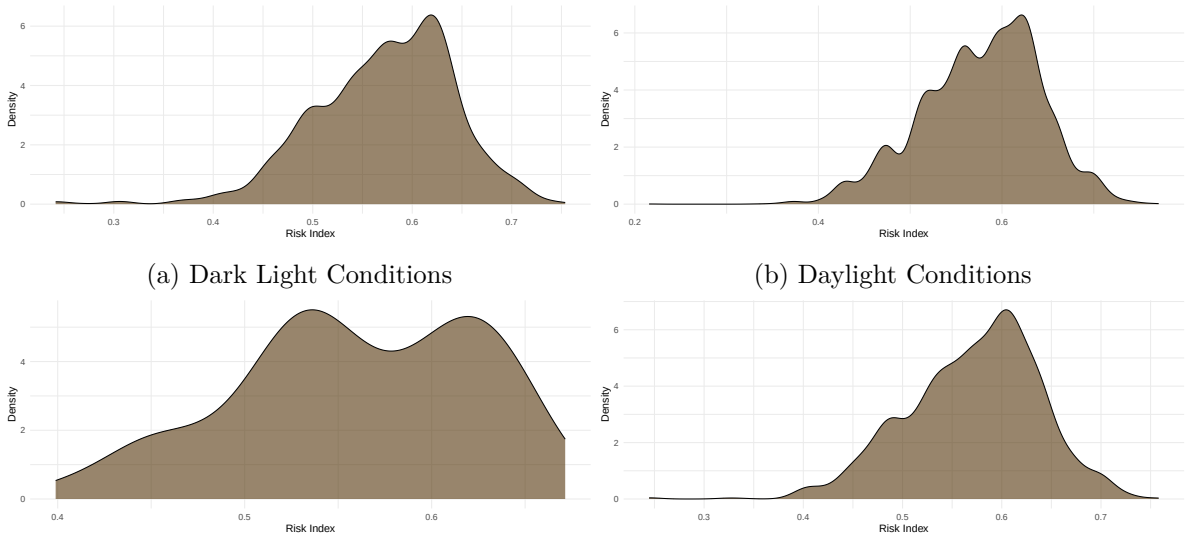


Figure 14: Risk Index Distribution by Road Surface Conditions

Figure 14 highlights notable differences in the Risk Index distribution across various road surface conditions. Wet and icy conditions, in particular, show higher variability in the Risk Index, suggesting that these surfaces may lead to more unpredictable and elevated risk levels. Dry conditions, on the other hand, generally exhibit a lower and more consistent risk. These findings highlight the significant impact that road surface conditions can have on traffic-related risks, underscoring the importance of considering these factors in risk management and safety measures.



(a) Dark Light Conditions

(b) Daylight Conditions

(c) Dawn Light

(d) Artificial Lighting Conditions

Figure 15: Risk Index by Lighting Conditions

Figure 15 presents the distribution of the Risk Index across four distinct lighting categories: Dark Conditions, Daylight Conditions, Dawn Conditions, and Artificial Lighting Conditions.

Dark Light Conditions in Figure 15a highlights a concentration of Risk Index values in the moderate-to-high range, indicating that low visibility during dark conditions significantly contributes to elevated collision risks. These findings underscore the importance of improved street lighting or reflective road markers in mitigating risks under dark conditions.

Daylight Conditions distrubition in Figure 15b shows a peak at slightly lower Risk Index values compared to dark conditions, reflecting the safety advantages of enhanced visibility. Daylight conditions reduce collision risks, likely due to clear visibility and the predictability of traffic patterns during the day.

The Risk Index distribution (Figure 15c) in dawn conditions displays a broader spread, indicating transitional risk levels. The variability in light conditions during dawn may lead to inconsistent visibility, increasing collision risks for certain hours. Targeted measures, such as adaptive lighting, may help reduce risks during this time.

Artificial Conditions in Figure 15d shows a moderate-to-high concentration of Risk Index values, suggesting that artificial lighting, while helpful, does not fully mitigate risks associated with low visibility. Enhanced lighting technologies and maintenance of artificial lights could further improve safety outcomes under these conditions.

4.3 Mapping the Risk: Neighborhood-Level Insights

4.3.1 Neighborhood Risk Distribution

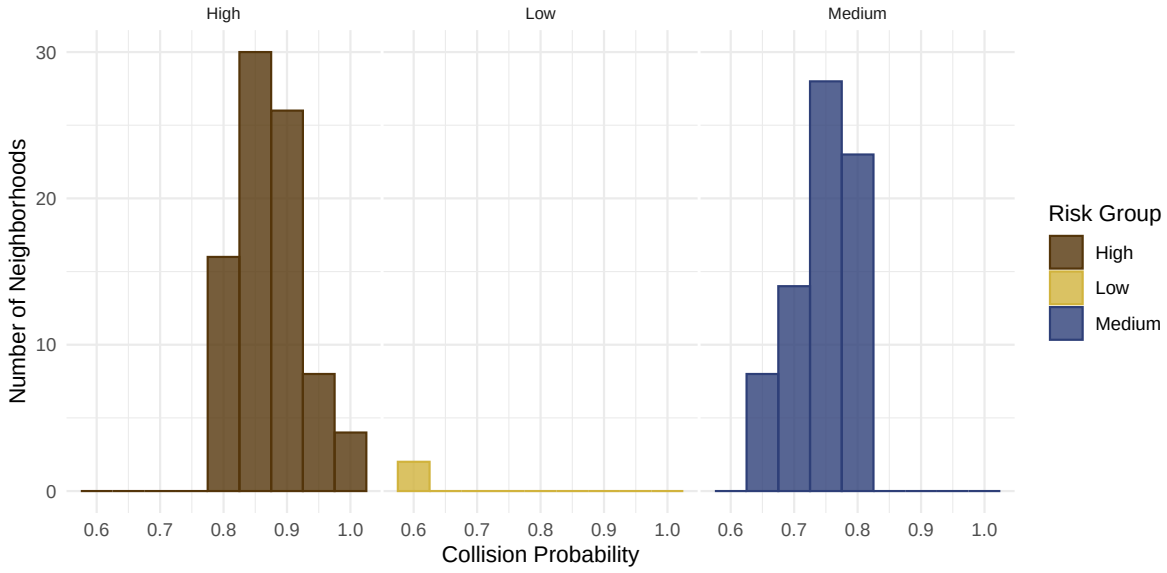


Figure 16: Neighborhoods grouped by average collision probabilities into High, Medium, and Low risk categories.

Figure 16 presents the distribution of Toronto neighborhoods based on their average collision probabilities, categorized into three risk levels: High, Medium, and Low. The High-risk category mainly includes neighborhoods with collision probabilities between 0.8 and 1.0, indicating a concentration of areas with increasing risk. The Medium-risk category includes neighborhoods with probabilities between 0.6 and 0.8, reflecting a moderate level of collision likelihood. The Low-risk category consists of neighborhoods with probabilities below 0.6, suggesting a lower collision risk in these areas. This distribution underscores the uneven spread of collision risks across Toronto neighborhoods, providing valuable insights into spatial patterns. It highlights the potential for geographically targeted interventions, such as implementing road safety measures or infrastructure improvements in neighborhoods with higher collision probabilities to address localized risks effectively.

4.3.2 Maps

Figure 17: Neighborhood Collision Points in Toronto Highlighting High Collision Risk Areas

Figure 17 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 increased. In contrast, the yellow points represent areas where collisions have taken place but with lower collision risks. While these points are more dispersed across the map, they show a notable concentration in downtown Toronto. This clustering in the downtown area suggests that even neighborhoods with moderate or lower collision risks can experience a high frequency of incidents due to the dense urban environment, heavy traffic flows, and increased walkers and cyclist activity typical of central urban areas. This observation highlights the importance of urban traffic management strategies across risk levels to improve safety in both high- and low-risk neighborhoods.

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

This study highlights distinct temporal variations in the motorbike theft and collision Risk Index, providing valuable insights for both behavioral adaptations and policy interventions. Collision risk shows a clear temporal pattern, peaking during Morning (7–9 AM) and Evening (5–7 PM) periods, which correspond to high traffic density during commuting hours. This reinforces the critical link between collision risk and traffic patterns, emphasizing the need for heightened caution and enhanced safety measures during these high-risk periods to reduce the likelihood of collisions and improve road safety. In contrast, theft risk remains stable across all time bins, showing no significant fluctuations. To enhance visibility of theft risk in the data, its values were scaled by a factor of 10, demonstrating its consistently smaller contribution to the overall Risk Index.

The Morning time segment exhibits the highest overall Risk Index, driven by elevated collision risk. The Afternoon and Evening segments show slightly lower Risk Index values, reflecting moderate variations in collision risk. Conversely, the Early Morning segment has the lowest Risk Index, corresponding to minimal traffic activity and theft incidents. These patterns underscore the importance of adopting time-sensitive strategies for mitigating risks during specific periods.

5.1.1 Behavioral Recommendations for Motorbike Owners

Motorbike owners can take several proactive measures to mitigate risks based on the identified temporal patterns. These include avoiding high-risk hours, strengthening security measures, and practicing defensive driving. For theft prevention, owners should limit motorbike use during late-night hours or park in secure locations with surveillance systems. Utilizing anti-theft devices such as GPS trackers or motion-sensitive alarms can effectively deter theft, particularly during vulnerable periods. To reduce collision risks, motorbike owners should exercise extra caution during peak traffic hours, adhere to speed limits, and maintain safe distances to ensure safer driving conditions.

5.1.2 Policy Recommendations for Stakeholders

Policymakers and law enforcement agencies can implement targeted strategies informed by these insights. Increasing law enforcement presence during high-theft hours through time-sensitive patrols can serve as a deterrent and ensure quicker response times. Enhanced traffic

monitoring during rush hours, using surveillance cameras and traffic officers, can help manage congestion and reduce collision risks. Additionally, public awareness campaigns and educational initiatives can inform motorbike owners about high-risk periods and promote safety practices tailored to temporal risk patterns, fostering a safer environment for all road users.

5.2 Environmental and Situational Influences on Risk

Environmental and situational factors play a pivotal role in shaping the Risk Index, emphasizing their impact on motorbike thefts and collisions. This study identified key predictors, including road surface conditions, lighting environments, and traffic control mechanisms, that collectively contribute to variations in risk levels. Poor road conditions, inadequate lighting, and the absence of effective traffic control measures are identified as major contributors to elevated risks.

Wet or icy road surfaces significantly heighten collision risks due to reduced traction and extended braking distances. Although dry roads are the most common, they are not devoid of risk, often correlating with higher traffic volumes that increase collision likelihood. Low-light conditions, particularly at night or under artificial lighting, further amplify risks by reducing visibility and reaction times. While daylight conditions offer improved visibility, high traffic density during peak hours introduces additional risks. Moreover, areas lacking traffic control measures, such as stop signs or traffic signals, experience elevated incident rates, highlighting the critical role of such infrastructure in mitigating risks.

5.2.1 Recommendations for Stakeholders

1. Road Maintenance and Infrastructure Enhancements: Regularly inspect and maintain road surfaces to reduce hazards like potholes, standing water, or ice. Implement adaptive measures, such as anti-skid materials on high-risk roads, particularly in areas prone to wet or icy conditions.
2. Lighting Improvements: Install and maintain streetlights in poorly lit areas to improve visibility and safety for drivers and pedestrians. Explore smart lighting systems that adjust brightness based on environmental conditions, enhancing visibility during low-light hours.
3. Traffic Control Measures: Increase the presence of traffic signals, stop signs, and pedestrian crossings in high-risk zones. Deploy dynamic traffic control systems, such as real-time traffic lights that adapt to congestion levels, reducing collision risks.

5.2.2 The Interaction of Environmental Factors

Environmental factors often interact in ways that amplify risks. For example, wet roads combined with poor lighting conditions create a dual challenge for drivers, reducing both traction and visibility. Similarly, areas lacking traffic controls are particularly vulnerable during nighttime or in adverse weather conditions. Addressing these intersections through multi-faceted strategies can yield substantial safety improvements. By understanding and addressing these environmental and situational risks, policymakers and urban planners can implement targeted interventions that enhance road safety and reduce motorbike-related incidents. This approach not only mitigates immediate risks but also contributes to building resilient and sustainable urban environments.

5.3 Policy Implications of Spatial Risk Disparities

This study highlights significant spatial disparities in the Risk Index across neighborhoods, revealing that certain areas are disproportionately affected by motorbike thefts and collisions. These disparities often correlate with socioeconomic factors, infrastructure quality, and population density, underscoring the need for targeted interventions.

The analysis identifies several neighborhoods with consistently high Risk Index values. These areas are often characterized by limited law enforcement presence, inadequate lighting, and poor road conditions, making them hotspots for theft and collisions. Conversely, neighborhoods with lower Risk Index values typically have better infrastructure, higher socioeconomic status, and a stronger law enforcement presence, which collectively contribute to enhanced safety.

5.3.1 Recommendations for Policymakers

1. Infrastructure Upgrades: Improve street lighting in high-risk neighborhoods to enhance visibility and deter criminal activity; address road surface issues, such as potholes and uneven pavements, to reduce collision risks.
2. Law Enforcement Strategies: Increase patrols in theft-prone areas during high-risk times to enhance deterrence and response capabilities; employ community policing initiatives to build trust and encourage collaboration between residents and law enforcement.
3. Localized Awareness Campaigns: Conduct education campaigns in high-risk neighborhoods to raise awareness about theft prevention measures and safe driving practices; provide subsidized access to anti-theft devices for residents in vulnerable areas.

4. Equitable Urban Planning: prioritize investments in underserved neighborhoods to address systemic inequities that contribute to heightened risk levels; develop multi-stakeholder strategies involving local governments, law enforcement, and community organizations to address both immediate and long-term safety concerns.

5.3.2 Broader Implications

The observed spatial disparities in risk not only highlight immediate safety concerns but also reflect broader socioeconomic inequalities. Addressing these disparities through targeted policies and equitable resource allocation can foster safer communities and reduce overall motorbike-related incidents. By adopting these evidence-based interventions, policymakers can mitigate spatial risk disparities, ensuring that safety measures are distributed equitably across neighborhoods and tailored to the unique needs of each area.

5.4 Limitations

The datasets used in the analysis contain missing or incomplete entries, particularly for variables such as road surface conditions and lighting. These limitations may introduce bias or underrepresent specific risk factors, potentially affecting the overall realiability of the findings. Additionally, the Risk Index relies on fixed weights for theft and collision components, which may oversimplify the relative importance of these factors. This static approach might not fully capture variations in risk across different contexts or over time.

The temporal and geographic scope of the study is another limitation. Since the analysis is limited to a specific region and time frame, its findings may not generalize to other locations or periods with differing traffic patterns or environmental conditions. Furthermore, while the study identifies spatial disparities in risk, it does not fully integrate socioeconomic data, such as income levels or educational attainment, which could provide a deeper understanding of the factors contributing to these disparities. Lastly, the analysis does not extensively explore interactions between risk factors. For instance, the combined impact of poor lighting and wet road conditions may amplify risks in ways that were not accounted for in this study.

5.5 Future Research

Building on the findings and limitations of this study, future research should aim to enhance the understanding of motorbike theft and collision risks through several approaches. One direction is the development of dynamic risk models that integrate real-time data, such as weather conditions, traffic density, and law enforcement activity. These models could provide more immediate and accurate assessments of risk, enabling proactive interventions. Another area for research is the impact of seasonal and event-based variations. Examining how risk

patterns shift during holidays, festivals, or different seasons would allow for more refined time-sensitive recommendations for both motorbike owners and policymakers.

The inclusion of socioeconomic indicators in future studies is also critical. Variables such as income levels, unemployment rates, and population density could offer a more thorough view of the root causes of spatial disparities in risk. Additionally, future research should explore the interaction effects between environmental and situational factors. For example, investigating how wet roads combined with poor lighting conditions affect risk levels could inform more nuanced interventions.

Qualitative research with residents and stakeholders in high-risk neighborhoods could also provide valuable insights into localized challenges and solutions. By engaging directly with affected communities, researchers can design and evaluate community-centered interventions that address both motorbike theft and collision risks. Finally, longitudinal studies would enable a better understanding of how risk patterns evolve over time and provide a means to evaluate the long-term effectiveness of policy measures and infrastructure improvements. These future directions collectively offer a roadmap for advancing the study of motorbike risks and enhancing road safety.

Appendix

A Shiny Application

The Risk Index by neighborhood can be visualized here.

B Additional Data Details

B.1 Cleaning Methods

The goal of the data cleaning process was to import raw theft and collision data, and refine the necessary columns to prepare cleaned datasets for analysis. The process began with loading two CSV files: "Motor Vehicle Collisions with KSI Data.csv" and "theft-from-motor-vehicle.csv", which contained information about products and their associated transactions. Examples of raw data are illustrated in Table 5 and Table 6.

Table 5: Examples of Raw Collision Data

Variable Name	Example Value
_id	1
ACCNUM	893184
DATE	2006-01-01
TIME	236
STREET1	WOODBINE AVE
STREET2	O CONNOR DR
OFFSET	None
ROAD_CLASS	Major Arterial
DISTRICT	Toronto and East York
ACCLOC	Intersection Related
TRAFFCTL	No Control
VISIBILITY	Clear
LIGHT	Dark
RDSFCOND	Wet
ACCLASS	Non-Fatal Injury
IMPACTYPE	Approaching
INVTYPE	Passenger
INVAGE	50 to 54
INJURY	Major
FATAL_NO	None

INITDIR	None
VEHTYPE	None
MANOEUVER	None
DRIVACT	None
DRIVCOND	None
PEDTYPE	None
PEDACT	None
PEDCOND	None
CYCLISTYPE	None
CYCACT	None
CYCCOND	None
PEDESTRIAN	None
CYCLIST	Yes
AUTOMOBILE	None
MOTORCYCLE	None
TRUCK	None
TRSN_CITY_VEH	None
EMERG_VEH	Yes
PASSENGER	Yes
SPEEDING	Yes
AG_DRIV	None
REDLIGHT	Yes
ALCOHOL	None
DISABILITY	60
HOOD_158	Woodbine-Lumsden
NEIGHBOURHOOD_158	60
HOOD_140	Woodbine-Lumsden (60)
NEIGHBOURHOOD_140	D55
DIVISION	{”type”: ”MultiPoint”, ”coordinates”: [[-79.318797, 43.699595]]}
geometry	N/A

Table 6: Examples of Raw Theft Data

Variable Name	Example Value
_id	1
EVENT_UNIQUE_ID	GO-20141261501
REPORT_DATE	2014-01-01
OCC_DATE	2014-01-01
REPORT_YEAR	2014

REPORT_MONTH	January
REPORT_DAY	1
REPORT_DOY	1
REPORT_DOW	Wednesday
REPORT_HOUR	8
OCC_YEAR	2014
OCC_MONTH	January
OCC_DAY	1
OCC_DOY	1
OCC_DOW	Wednesday
OCC_HOUR	8
DIVISION	D51
LOCATION_TYPE	Single Home, House (Attach Garage, Cottage, Mobile)
PREMISES_TYPE	House
UCR_CODE	2142
UCR_EXT	200
OFFENCE	Theft From Motor Vehicle Under
MCI_CATEGORY	NonMCI
HOOD_158	73
LONG_WGS84	-79.37453055088850
LAT_WGS84	43.65706729617110
geometry	{"type": "MultiPoint", "coordinates": [[-79.3745305, 43.6570672]]}

To prepare the raw collision data for analysis, a structured cleaning process was carried out with a focus on maintaining consistency and accuracy. Essential libraries were loaded at the start of the process, including `tidyverse` for data manipulation, `janitor` for standardizing column names, `here` for managing file paths, and `arrow` for saving the cleaned data in an efficient Parquet format. These tools provided a solid framework for handling various cleaning tasks effectively and ensuring the dataset was ready for further use.

The raw traffic data was loaded from a CSV file using `read_csv()`. An initial step involved checking for the presence of a `geometry` column, which was removed if found, as it was not relevant for the analysis. Column names were standardized using `janitor::clean_names()`, ensuring they were formatted in snake_case for consistency. Additional formatting steps replaced dots with underscores and removed non-alphanumeric characters, making the column names more usable and easier to interpret.

Duplicates were removed from the dataset using the `distinct()` function, ensuring each record was unique. Missing values were handled by replacing placeholder text like `"None"` with `NA` across all character columns. This step helped standardize the handling of missing data, allowing it to be appropriately identified and excluded in subsequent analyses.

Key columns were converted to appropriate data types to facilitate accurate analysis. The `invage` column, representing age, was converted to numeric, while the `injury` column was transformed into a factor to categorize injury types effectively. These type conversions ensured that the data could be analyzed and interpreted meaningfully.

To improve readability and maintain consistency with other datasets, several columns were renamed. For example, `accnum` was renamed to `accident_id`, `date` to `report_date`, and `traffctl` to `traffic_control`. This renaming improved the clarity of the dataset and aligned it with conventions used in related datasets, such as the cleaned crime data.

The dataset's geographic data was validated by ensuring latitude and longitude values fell within acceptable ranges. Latitude values were checked to be between `-90` and `90`, and longitude values were validated to be between `-180` and `180`. Records with invalid geographic coordinates were excluded, ensuring the spatial data was accurate and reliable.

Finally, the cleaned dataset was saved in Parquet format using `write_parquet()`. The Parquet format was chosen for its storage efficiency and compatibility with data processing tools, making it an ideal format for large datasets. Additionally, a summary of the cleaned dataset was generated using the `summary()` function, providing an overview of the dataset's structure and content.

In summary, this cleaning process ensured the collision data was free of duplicates and invalid entries, featured consistent and intuitive column names, and had validated geographic information. The resulting cleaned dataset show in Table 7 was stored in an optimized format, making it ready for downstream analysis tasks.

Table 7: Example of Cleaned Collision Data

Variable Name	Example Value
id	1
accident_id	893184
report_date	2006-01-01
time_of_day	236
street_primary	WOODBINE AVE
street_secondary	O CONNOR DR
offset	NA
road_type	Major Arterial
district	Toronto and East York
accident_location	Intersection Related
traffic_control	No Control
visibility_conditions	Clear
lighting_conditions	Dark
road_conditions	Wet
accident_classification	Non-Fatal Injury

impact_type	Approaching
invtype	Passenger
invage	NA
injury	Major
fatal_no	NA
initdir	NA
vehicle_type	NA
maneuver	NA
driver_action	NA
driver_condition	NA
pedtype	NA
pedact	NA
pedcond	NA
cyclistype	NA
cycact	NA
cyccond	NA
pedestrian	NA
cyclist	NA
automobile	Yes
motorcycle	NA
truck	NA
trsn_city_veh	NA
emerg_veh	NA
passenger	Yes
speeding	Yes
ag_driv	Yes
redlight	NA
alcohol	Yes
disability	NA
hood_158	60
neighborhood	Woodbine-Lumsden
hood_140	60
neighbourhood_140	Woodbine-Lumsden (60)
division	D55

The data cleaning process for the raw theft data show in Table 6 was designed to ensure the dataset is accurate, consistent, and ready for analysis. First, the required libraries such as `tidyverse` for data manipulation, `janitor` for standardizing column names, `here` for handling file paths, and `arrow` for managing Parquet files were loaded. Additionally, a directory

structure was established to store the cleaned data in a designated folder. This setup provided a solid foundation for the cleaning process.

To validate the data, predefined lists of valid values for certain fields were established. Specifically, a set of valid division codes (e.g., "D51", "D42") and offense categories (e.g., "Theft From Motor Vehicle Under", "Other Theft") was created to filter out any unrecognized or irrelevant entries. This ensured that only records meeting specific criteria would be retained in the final dataset.

The raw crime data was loaded from a CSV file, and column names were standardized using the `clean_names()` function from the `janitor` package. Key columns, such as `report_date` and `occ_date`, were converted into date formats, while latitude and longitude fields were cast to numeric values for validation. The `report_hour` column was converted into integers, and a new column, `report_dow`, was added to capture the day of the week from the `report_date`.

Rows with missing critical values, such as `event_unique_id`, `report_date`, `division`, `offence`, or geographic coordinates, were removed to maintain data quality. Additionally, latitude and longitude values were validated to ensure they fell within acceptable ranges (latitude: -90 to 90, longitude: -180 to 180). Text fields such as `division` and `offense` were standardized by trimming extra spaces and converting values to uppercase. Records with division codes or offenses not included in the predefined valid lists were excluded from the dataset. To further ensure data integrity, duplicate rows based on `event_unique_id` were removed.

The dataset was further refined by selecting only relevant columns, such as `event_unique_id`, `report_date`, `division`, `location_type`, and geographic coordinates. Columns were renamed to make them more intuitive, such as renaming `event_unique_id` to `event_id` and `report_date` to `report`. This helped to simplify and clarify the structure of the data.

Finally, the cleaned dataset was saved as a Parquet file (`cleaned_crime_data.parquet`) using the `write_parquet()` function. The Parquet format was chosen for its efficiency in storage and processing, making the dataset ready for further analysis. Examples of cleaned theft data is presented in Table 8. This cleaning process resulted in a high-quality dataset that is well-prepared for any downstream tasks.

Table 8: Example of Cleaned Theft Data

Variable Name	Example Value
event_id	GO-20141261501
report	2014-01-01
occurrence	2014-01-01
hour	8
day_of_week	Wednesday
division	D51

location	Single Home, House (Attach Garage, Cottage, Mobile)
premises_type	House
offense	Theft From Motor Vehicle Under
mci_category	NonMCI
hood_158	73
longitude	-79.37453055088850
latitude	43.65706729617110

C Idealized Methodology for a Survey on Motor Risk

C.1 Survey Objectives

The primary goal of this survey is to investigate the factors contributing to motorbike risks, including theft and collisions, by collecting data from motorbike owners and riders. This research aims to understand the interplay of demographic, behavioral, and environmental factors that influence risk levels. Insights gained from the survey will help develop targeted interventions to enhance motorbike safety and security.

C.2 Sampling Methodology

The target population for this survey includes licensed motorbike riders and owners across diverse geographic regions. The survey will focus on individuals who have used motorbikes within the past year to ensure relevance and recency of responses.

A stratified random sampling approach will be employed to capture diverse perspectives and ensure statistical validity. Stratification will be based on:

Geographic regions: Urban, suburban, and rural areas to understand variations in risks and behaviors. Demographic characteristics: Age, gender, income, and education levels to capture socioeconomic diversity. Riding experience: Novice, intermediate, and experienced riders to assess how skill and familiarity influence risks. A sample size of approximately 1,000 respondents is proposed, distributed proportionally across the strata to ensure adequate representation for comparative analysis.

C.3 Survey Structure and Content

Link to the survey: Motorbike Safety and Risk Assessment Survey

C.3.1 Questionnaire Design

The questionnaire will include a mix of closed-ended and open-ended questions. Closed-ended questions will use Likert scales, multiple-choice options, and ranking formats, while open-ended questions will provide opportunities for respondents to share detailed thoughts.

1. Closed-Ended Questions

These questions offer predefined answer choices and are used to gather quantitative data. They are easy to analyze and are suitable for identifying patterns and trends.

2. Open-Ended Questions

Open-ended questions allow respondents to express their thoughts and experiences in detail. They are useful for collecting qualitative data and identifying informatiosn

3. Matrix Questions

Matrix questions allow multiple related items to be rated on the same scale, making it easier to assess various factors within a single question.

4. Ranking Questions

Ranking questions require respondents to prioritize options based on their preferences or perceived importance. They are effective for understanding preferences and trade-offs.

5. Dichotomous Questions

These questions have only two answer choices and are typically used to gather straightforward, binary data.

C.4 Recruitment Strategy

To ensure a diverse and representative sample, participants will be recruited through a combination of online and offline channels, targeting motorbike riders and owners across various demographics, regions, and riding experiences. Online recruitment will utilize motorbike-related communities and platforms, including forums, social media groups, and email lists maintained by motorbike organizations. These platforms allow direct engagement with individuals who are likely to be interested in the survey, ensuring efficient outreach to a large pool of potential respondents. Social media advertisements and posts in popular groups dedicated to motorbike enthusiasts will also be utilized to broaden the reach and attract participants from different backgrounds.

Offline recruitment will complement the online efforts by targeting physical spaces frequented by motorbike riders. Flyers and posters will be distributed at motorbike dealerships, repair

shops, riding schools, and popular gathering spots for riders. These locations are strategic, as they attract individuals actively engaged in motorbike use and maintenance, making them ideal candidates for the survey. To further incentivize participation, small rewards such as gift cards or entries into a raffle will be offered. These incentives are designed to encourage participation while demonstrating appreciation for the respondents' time and effort. The combination of online and offline recruitment strategies ensures that the survey captures a wide range of perspectives, improving the reliability and generalizability of the findings.

C.5 Linkage to Literature

The design of this survey is grounded in a solid structure of literature on motor vehicle risk perception, road safety, and crime prevention. Studies on environmental and situational predictors of road accidents have informed the inclusion of variables such as time of day, weather conditions, road surface quality, and lighting (Ling et al. 2020). Research has consistently highlighted these factors as critical determinants of collision risk, guiding the formulation of survey questions aimed at understanding how riders perceive and respond to these risks in their daily activities (Fridman et al. 2020).

Additionally, studies on theft prevention measures have shaped the focus on security practices and riders' awareness of theft risks. Prior research has emphasized the importance of behavioral and environmental factors in preventing motorbike theft, such as the use of locks, secure parking spaces, and community awareness programs (Anderson and Linden 2014). These ideas have been incorporated into the survey's sections on preventive measures and risk awareness, ensuring the collection of data relevant to both individual behaviors and broader community-based interventions.

Finally, the survey's sampling methodology and design are informed by established literature on survey-based risk assessments. Stratified random sampling, a method widely regarded as effective for ensuring representativeness, was chosen based on evidence from prior studies (Harlow 1988). The survey structure, combining closed-ended, open-ended, and conditional questions, is modeled on best practices in survey design to capture both quantitative and qualitative data. By integrating findings from previous research, the survey aims to contribute to the ongoing discourse on motorbike safety and theft prevention, while addressing gaps in the literature specific to the Toronto context.

This careful alignment with existing literature not only validates the survey design but also ensures its findings will be relevant and valuable for developing targeted interventions to enhance motorbike safety and security. By linking the survey to established research, the study builds on a foundation of evidence, contributing new thoughts into the complex interplay of factors influencing motorbike risks.

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in R*. Chapman; Hall/CRC.

Anderson, Jeff, and Rick Linden. 2014. "Why Steal Cars? A Study of Young Offenders Involved in Auto Theft." *Canadian Journal of Criminology and Criminal Justice* 56 (2): 241–60.

Buehler, Ralph, and John Pucher. 2021. "COVID-19 Impacts on Cycling, 2019–2020." *Transport Reviews* 41 (4): 393–400.

Charron, Mathieu. 2009. *Neighbourhood Characteristics and the Distribution of Police-Reported Crime in the City of Toronto*. Statistics Canada Ottawa.

Fox, John, and Sanford Weisberg. 2023. *car: Companion to Applied Regression*. https://CRAN.R-project.org/package=car.

Fridman, Liraz, Rebecca Ling, Linda Rothman, Marie Soleil Cloutier, Colin Macarthur, Brent Hagel, and Andrew Howard. 2020. "Effect of Reducing the Posted Speed Limit to 30 Km Per Hour on Pedestrian Motor Vehicle Collisions in Toronto, Canada-a Quasi Experimental, Pre-Post Study." *BMC Public Health* 20: 1–8.

Gelfand, Sharla. 2020. *opendatatoronto: Access the City of Toronto Open Data Portal*. https://cran.r-project.org/package=opendatatoronto.

Grolemund, Garrett, and Hadley Wickham. 2011. *lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Harlow, Caroline Wolf. 1988. *Motor Vehicle Theft*. US Department of Justice, Bureau of Justice Statistics.

Horst, Allison, Alison Hill, and Kristin Van Den Eynden. 2023. *palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://CRAN.R-project.org/package=palmerpenguins.

Law, Jane, and Alexander T Petric. 2024. "Monitoring Day and Dark Traffic Collisions in Toronto Neighbourhoods with Implications for Injury Reduction and Vision Zero Initiatives: A Spatial Analysis Approach." *Accident Analysis & Prevention* 207: 107728.

Ling, Rebecca, Linda Rothman, Marie-Soleil Cloutier, Colin Macarthur, and Andrew Howard. 2020. "Cyclist-Motor Vehicle Collisions Before and After Implementation of Cycle Tracks in Toronto, Canada." *Accident Analysis & Prevention* 135: 105360.

Lüdecke, Daniel et al. 2023. *performance: Assessment of Regression Models Performance*. https://CRAN.R-project.org/package=performance.

Müller, Kirill. 2022. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Nick, Todd G., and Kathleen M. Campbell. 2007. "Logistic Regression." In *Topics in Biostatistics*, edited by Alvan R. Walker, 273–301. New York, NY: Springer. https://doi.org/10.1007/978-1-59745-530-5_14.

Padgham, Mark, and Joe Super. 2023. *osmdata: OSM Data in R*. https://CRAN.R-project.org/package=osmdata.

Pebesma, Edzer. 2018. *sf: Simple Features for R*. https://CRAN.R-project.org/package=sf.

Pedersen, Thomas. 2023a. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. https://CRAN.R-project.org/package=ggraph.

———. 2023b. *tidygraph: A Tidy API for Graph Manipulations*. https://CRAN.R-project.org/package=tidygraph.

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

Stoltzfus, Jill C. 2011. "Logistic Regression: A Brief Primer." *Academic Emergency Medicine* 18 (10): 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, et al. 2023. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2023. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wilke, Claus O. 2021. *Ggridges: Ridgeline Plots in 'Ggplot2'*. https://CRAN.R-project.org/package=ggridges.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.

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.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.