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,and environmental factors. A composite risk score model is used and reveals 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

1 Introduction **3**

2 Data **4**
- 2.1 Source . . . 4
- 2.2 Data Measurement and Limitations . . . 4
- 2.3 Outcome Variables . . . 5
- 2.4 Predictor Variables . . . 7
  - 2.4.1 Theft index . . . 8
  - 2.4.2 Collision Probability . . . 11

3 Model **16**
- 3.1 Model Set-Up . . . 17
  - 3.1.1 Theft Risk Sub-Indexes . . . 17

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

3.1.2 Collision Probability Model . . . 17
3.1.3 Risk Index Calculation . . . 18
3.2 Model Justification . . . 18
3.3 Model Assumptions and Validations . . . 19
3.3.1 Theft Sub-Index . . . 19
3.3.2 Composite Risk Index . . . 20
3.3.3 Collision Model . . . 20

4 Results **24**
4.1 When Risk Peaks: Temporal Trends in the Index . . . 24
4.1.1 Time-of-Day Analysis . . . 24
4.1.2 Time Series Analysis . . . 25
4.2 Conditions of Danger: Environmental Drivers of Risk . . . 26
4.3 Mapping the Risk: Neighborhood-Level Insights . . . 29
4.3.1 Neighborhood Risk Distribution . . . 29
4.3.2 Maps . . . 30

5 Discussion **31**
5.1 Temporal Risk Patterns and Behavioral Recommendations . . . 31
5.2 Environmental and Situational Influences on Risk . . . 31
5.3 Policy Implications of Spatial Risk Disparities . . . 32
5.4 Limitations . . . 32
5.5 Future Research . . . 32

Appendix **33**

A Additional Data Details **33**
A.1 Data Cleaning . . . 33

B Model details **33**
B.1 Posterior predictive check . . . 33
B.2 Diagnostics . . . 33

C Idealized Methodology for a Survey on Motor Risk **34**
C.1 Survey Overview . . . 34
C.2 Sampling Approach . . . 34
C.3 Survey structure . . . 34
C.3.1 Question Types . . . 35
C.3.2 Question List . . . 35
C.4 Recruitment Strategy . . . 35
C.5 Linkage to Literature . . . 36

References **36**

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 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 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, available from Open Data Toronto (**OpendataToronto?**). 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.

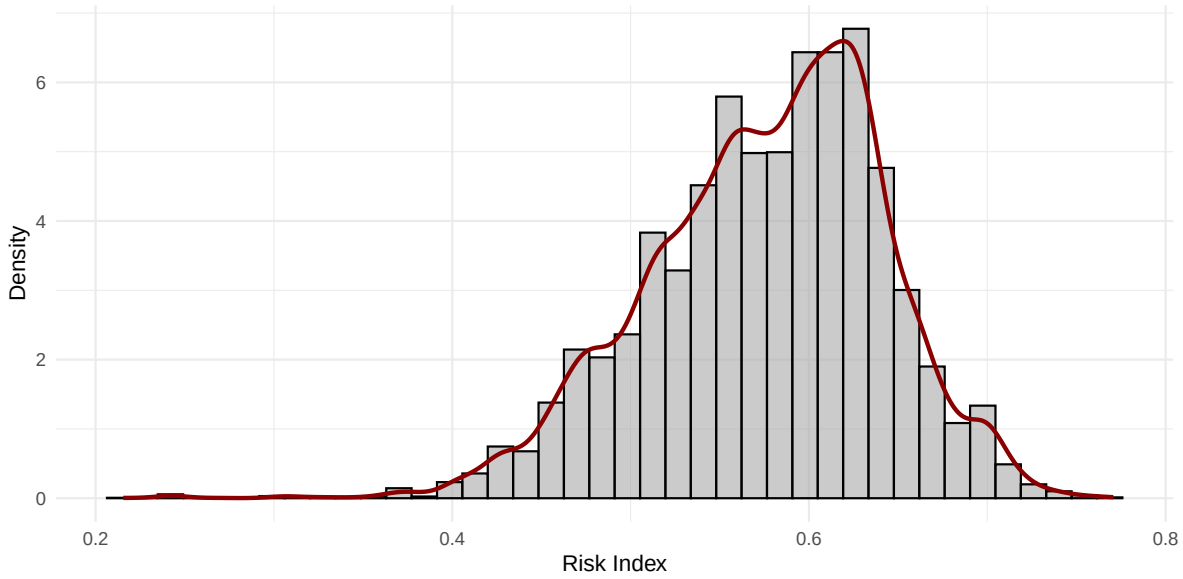


Figure 1: Distribution of the Overall Risk Index

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 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 increase the risks in specific cases. This insight can inform targeted interventions, focusing on the outliers with high Risk Index values to mitigate the most critical risks.

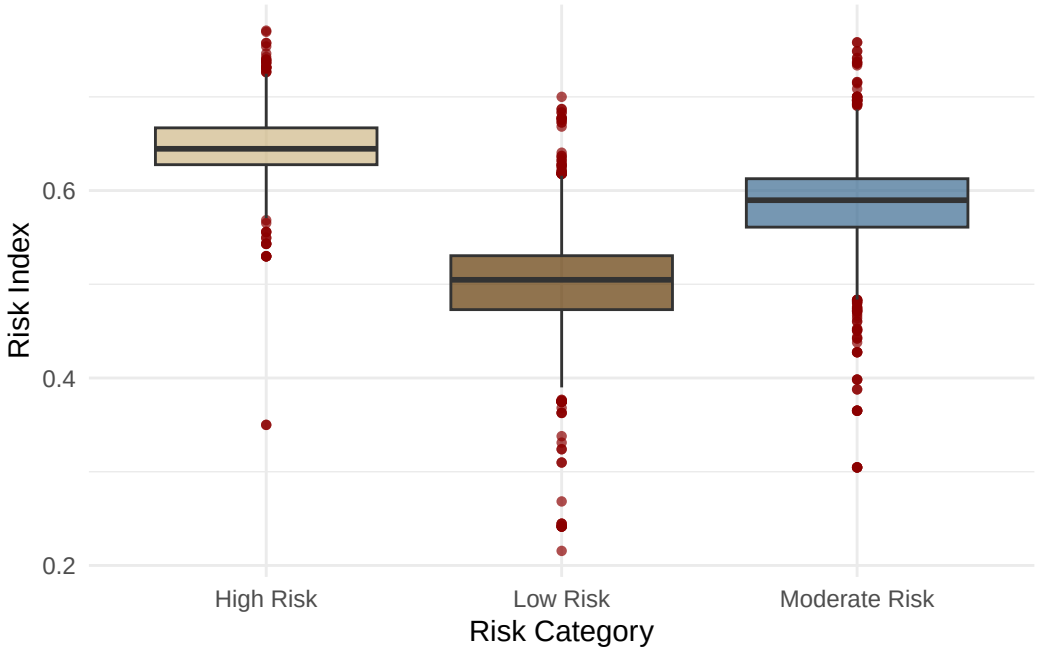


Figure 2: Risk Index Distribution by Risk Category

Building on the distribution of the Risk Index in Figure 1, 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 illustrates 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

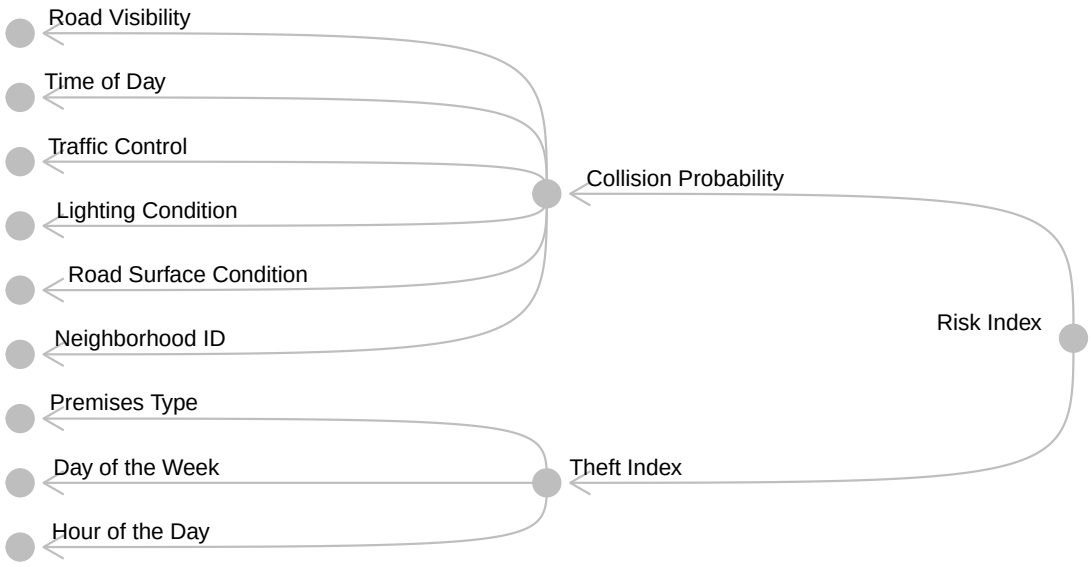


Figure 3: Hierarchical relationship between risk factors contributing to theft and collision probabilities.

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

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. (**plt-hour_summary?**) outlines the distribution of theft incidents across different hours, offering insights into periods with disproportionately high or low occurrences, which could impact the understanding of temporal theft risk dynamics.

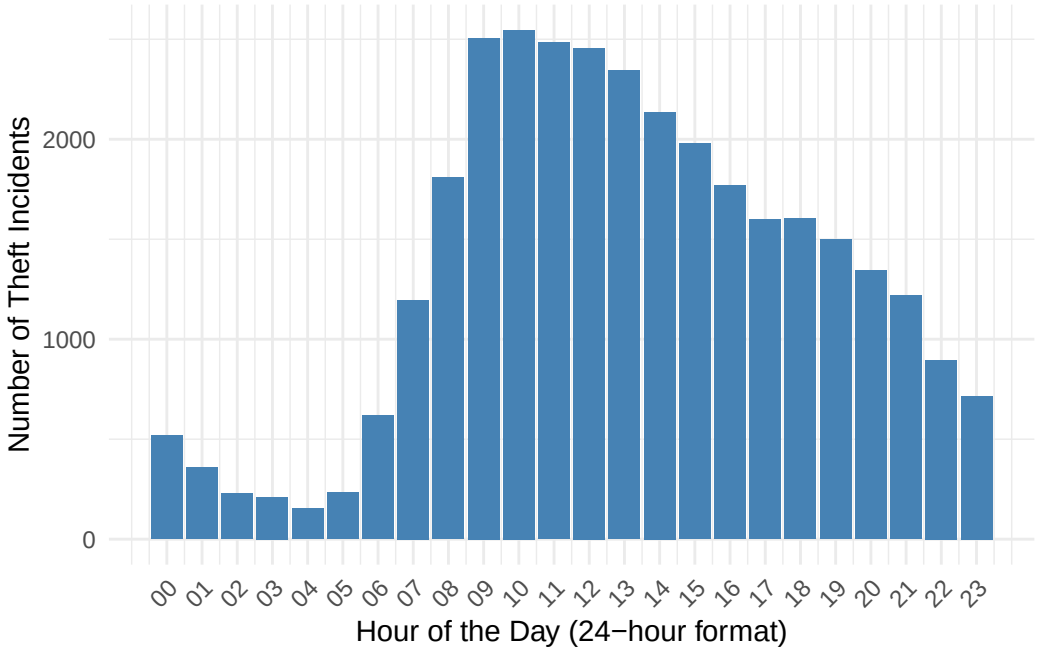


Figure 4: Theft Incidents by Hour of the Day

2.4.1.2 Day of the Week

The **Day of the Week** variable is included as a predictor to examine weekly patterns in motorbike theft incidents. This variable helps determine whether theft risks vary across different days of the week. Factors such as increased activity on weekends, reduced monitoring during certain weekdays, or patterns linked to routine schedules may influence these variations. (**plot-day_summary?**) summarizes the distribution of theft incidents by day of the week, highlighting any days with notably high or low occurrences.

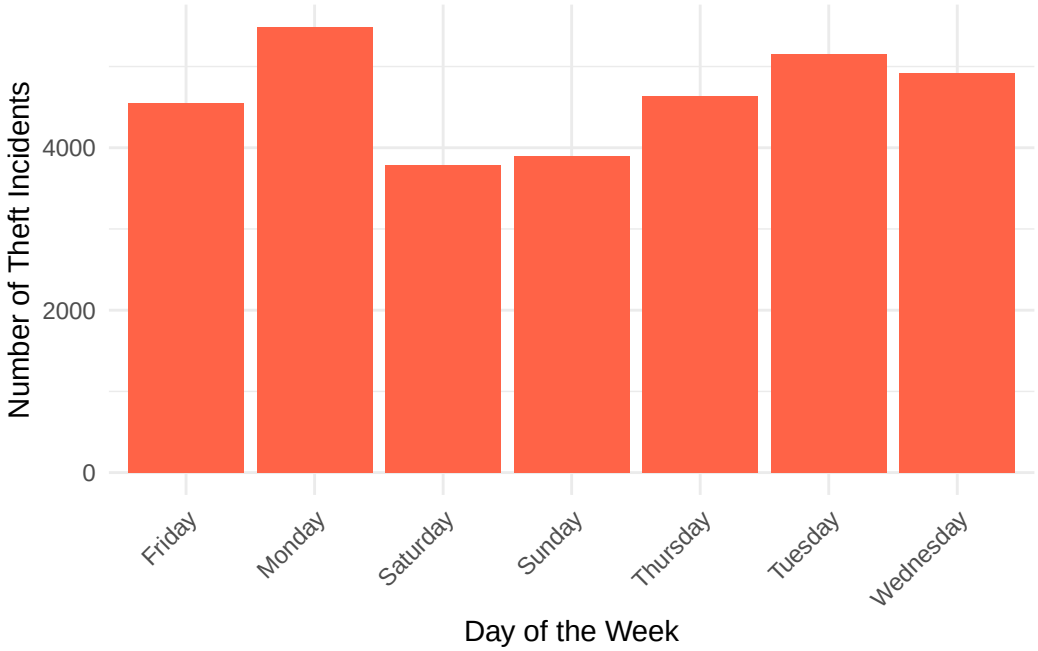


Figure 5: Theft Incidents by Day of the Week

2.4.1.3 Premises Type

The **Premises Type** variable serves as a predictor to analyze the contextual settings where motorbike theft incidents occur. This variable helps identify whether certain types of premises, such as residential areas, commercial establishments, or public parking lots, are associated with higher or lower theft risks. Factors like the availability of surveillance, the density of parked motorbikes, or accessibility to offenders may contribute to variations across premises types. (**plt-premises_summary?**) provides a breakdown of the number of theft incidents by premises type, highlighting locations with elevated or reduced risks.

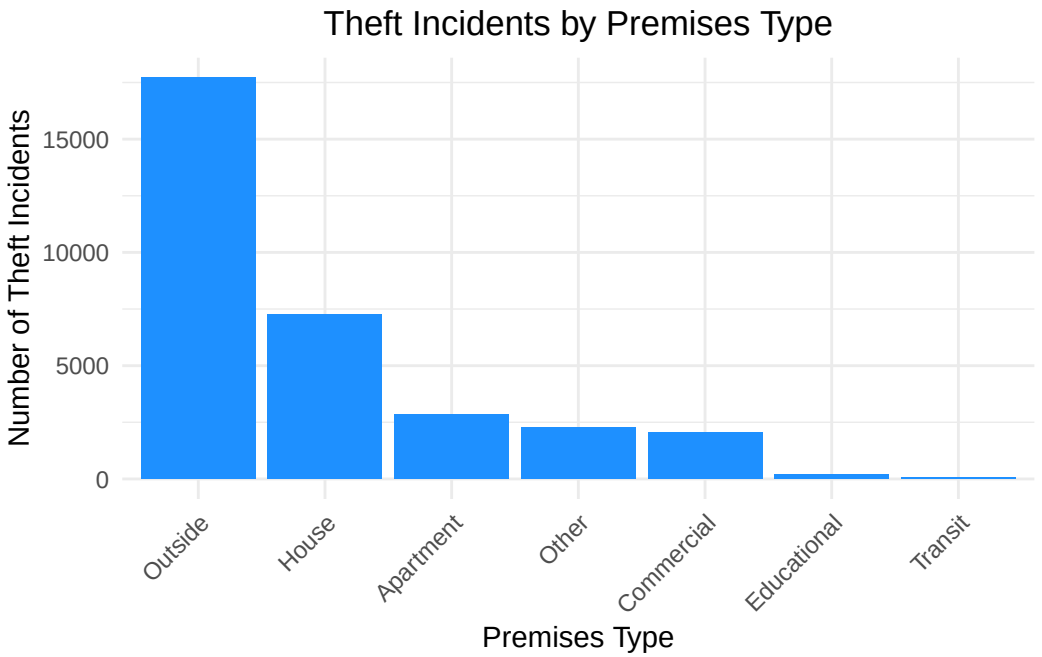


Figure 6: Theft Incidents by Day of the Week

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 offer insights 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 is a categorical predictor that represents unique identifiers for neighborhoods in the City of Toronto. This variable captures spatial variations in motorbike theft and collision risks, enabling the analysis to identify neighborhood-specific patterns and trends. Table 1 provides a list of example Neighborhood IDs, illustrating how this variable is used to represent different areas in the dataset. This summary helps contextualize the role of neighborhoods in understanding collision risk distributions.

Table 1

Neighborhood ID	Incident Count

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

10 Examples of Neighborhood Incidents Count

2.4.2.2 Road Surface Condition

The Road Surface Condition variable is a categorical predictor that describes the state of the road at the time of an incident, including categories such as dry, wet, loose snow, slush, ice, packed snow, gravel, spilled liquid, and others. This variable helps analyze how varying surface conditions influence motorbike collision risks. (**plot-road_surface_summary?**) illustrates the number of incidents across all these road surface conditions, revealing that the majority of incidents occur on dry roads, followed by wet roads. Other surface conditions, such as snow, ice, or gravel, account for a much smaller proportion of incidents. This distribution suggests that while dry conditions are most common for collisions, likely due to higher traffic volumes, wet and less stable surfaces may pose proportionally greater risks. These insights emphasize the importance of considering all road conditions when designing targeted road safety measures, particularly during adverse weather, to mitigate collision risks effectively.

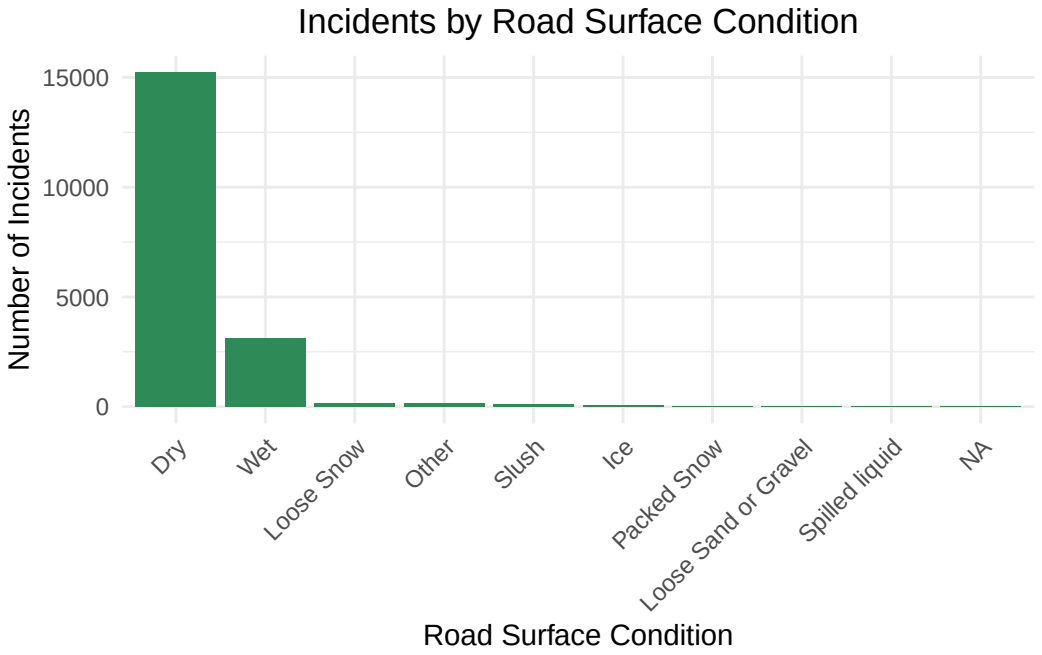


Figure 7: Incidents by Road Surface Condition

2.4.2.3 Lighting Condition

The Lighting Condition variable is a categorical predictor that represents the type of lighting at the time of an incident, including conditions such as daylight, darkness, artificial lighting, or dawn. This variable is crucial for assessing how visibility levels and lighting environments influence the likelihood of motorbike thefts and collisions. The bar plot (**plt-lighting_summary?**) shows that the majority of incidents occur during daylight conditions, which likely reflects higher traffic volumes during daytime hours. Incidents under "Dark" and "Dark, artificial" lighting conditions are significantly fewer but still notable, emphasizing the increased risks associated with reduced visibility. Other conditions, such as "Dawn" and "Dusk," contribute minimally to the total incidents. These insights highlight the importance of targeted safety interventions, such as improved artificial lighting and driver awareness during nighttime and transitional lighting conditions, to mitigate collision risks.

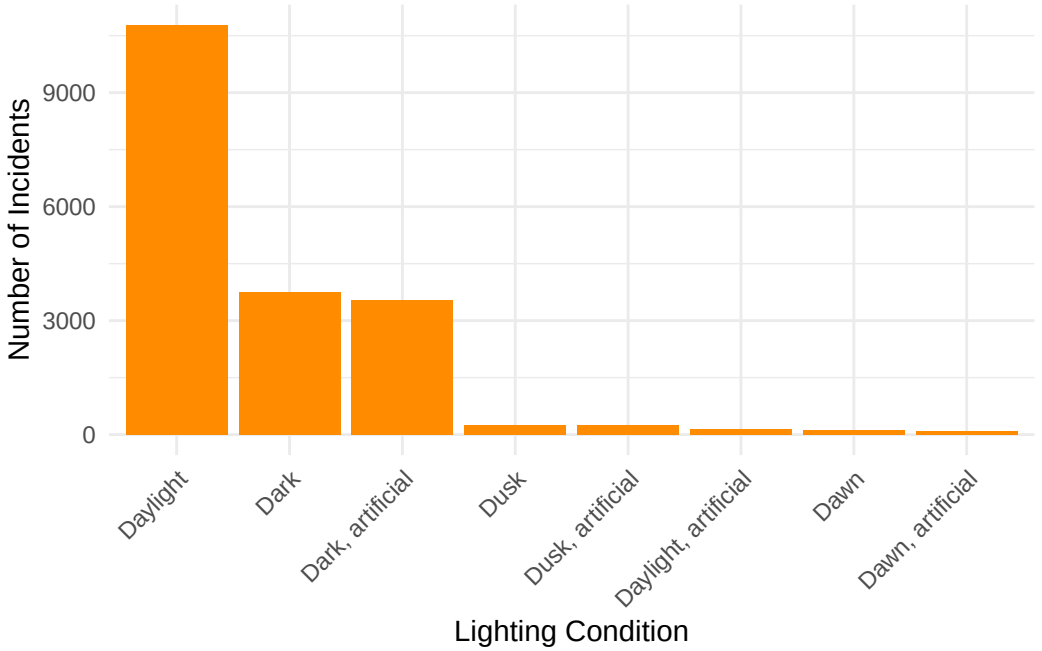


Figure 8: Incidents by Lighting Conditionk

2.4.2.4 Traffic Control

The Traffic Control variable categorizes the type of traffic management present at the location of each incident, such as "No Control," "Traffic Signal," or "Stop Sign." This variable enables an analysis of how different traffic control measures correlate with collision risks. As shown in Table 2, the majority of incidents occur in areas with "No Control" (9,021 incidents), followed closely by areas with "Traffic Signal" (8,035 incidents). This suggests that the absence of traffic control mechanisms, as well as high traffic volumes at signalized intersections, may contribute significantly to collision probabilities. Less frequent traffic control types, such as "Stop Signs" (1,464 incidents) or "Pedestrian Crossovers" (208 incidents), are associated with considerably fewer incidents. These findings underscore the importance of implementing and optimizing traffic control measures, particularly in areas without existing controls, to reduce collision risks effectively.

Table 2

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

Summary of Traffic Control

2.4.2.5 Time of Day

The **Time of Day** variable captures the specific hour and minute when a collision occurred, represented in a 24-hour format. For instance, "2210" corresponds to 10:10 PM, while "315" indicates 3:15 AM. This variable enables a detailed analysis of temporal patterns in collision risks, helping to identify whether certain times of the day are associated with higher or lower probabilities of incidents.

Contributing factors may include reduced visibility during nighttime hours, increased traffic density during peak commuting periods, or varying driver behaviors at different times. Table 3 provides examples of time data, offering a comprehensive view of how collision occurrences are distributed throughout the day. Understanding these temporal trends can inform targeted interventions to enhance road safety during high-risk periods.

Table 3

Time of Day
115
1032
909
340
1027

5 Examples of Different Times of Day

2.4.2.6 Visibility

The Visibility variable is a categorical predictor that describes the environmental visibility conditions at the time of a collision, such as clear, foggy, or reduced due to rain or snow. This variable is critical for assessing how different visibility levels impact the likelihood of collisions.

Poor visibility conditions, such as heavy fog, snow, or rain, can obscure drivers' ability to perceive road hazards or other vehicles, increasing the probability of incidents. Conversely, clear visibility conditions often correlate with lower collision risks.

(**plot-visibility_summary?**) provides a barplot summarizing the number of collision incidents under each visibility condition, illustrating how varying environmental factors contribute to changes in collision probabilities. The plot shows that the majority of incidents occur under clear conditions, with a significantly higher incident count than other conditions. Rain, the second most frequent condition, accounts for a smaller but notable proportion of incidents, while other conditions like snow, fog, and strong winds contribute minimally. These findings emphasize the need for proactive measures such as improving road drainage and visibility aids during adverse weather to mitigate collision risks effectively.

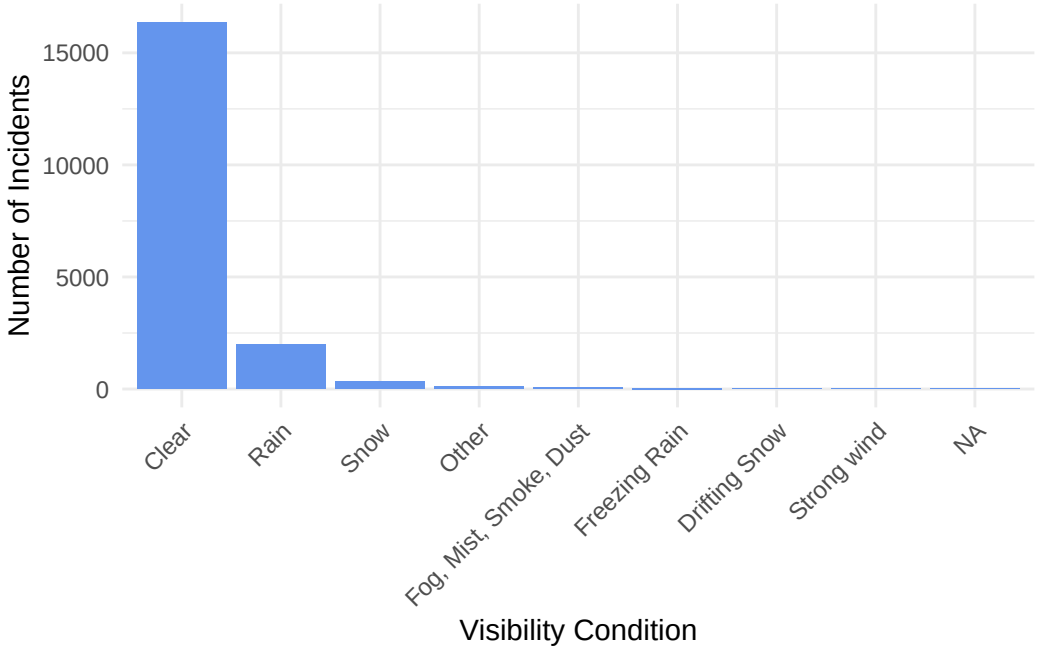


Figure 9: Incidents by Visibility Condition

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.

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:

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

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:

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

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

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 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.

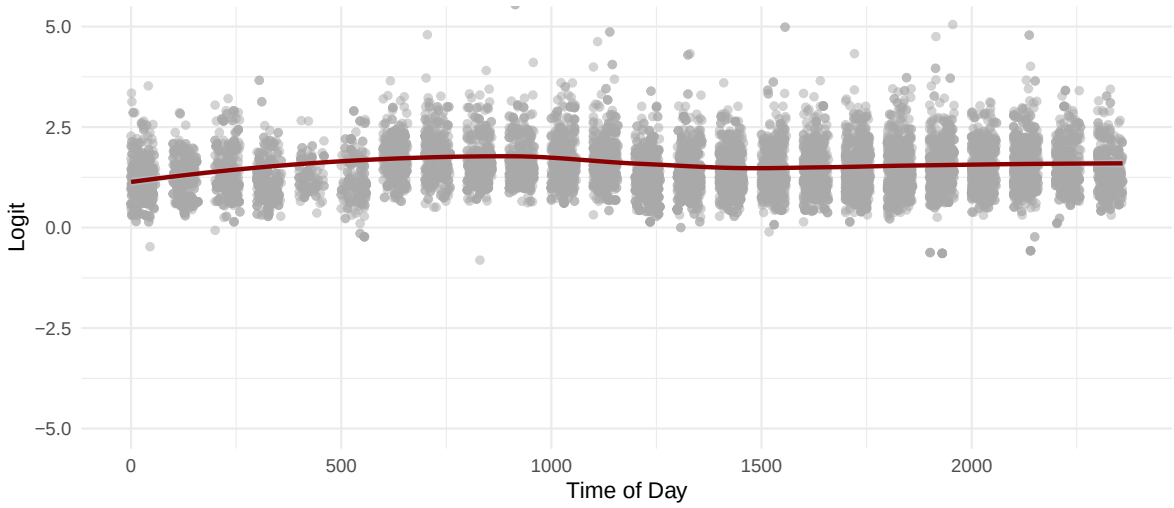


Figure 10: Logit Plot for Time of Day in the Collision Probability Model

Figure 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 (**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 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 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 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

4.1.1 Time-of-Day Analysis

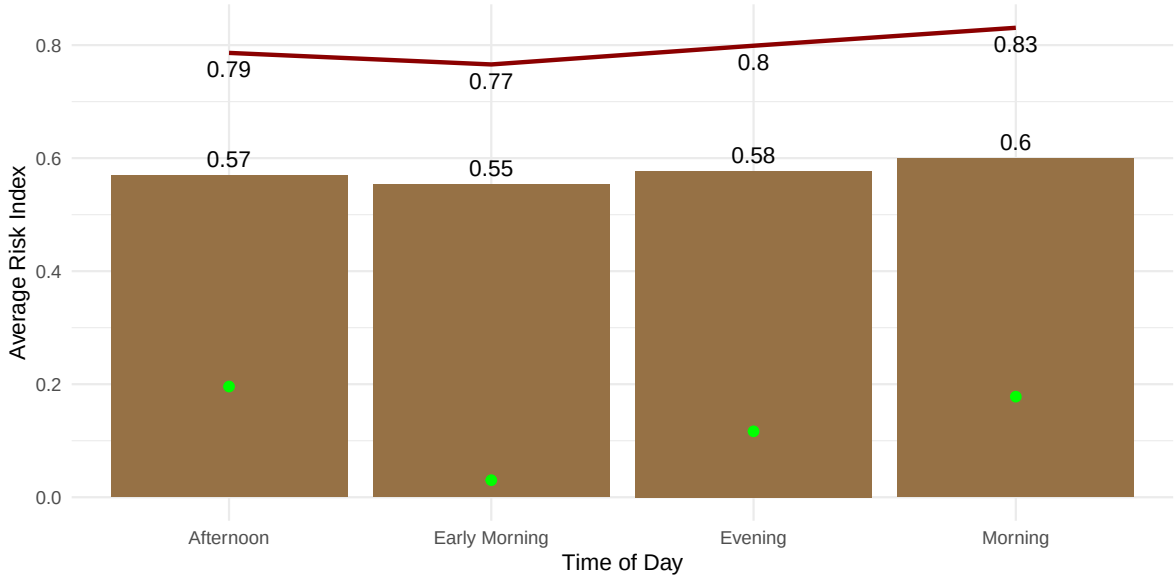


Figure 11: Risk Index by Time of Day, with bars showing Overall Risk Index, the red line showing Collision Risk, and green dots showing Theft Risk.

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 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

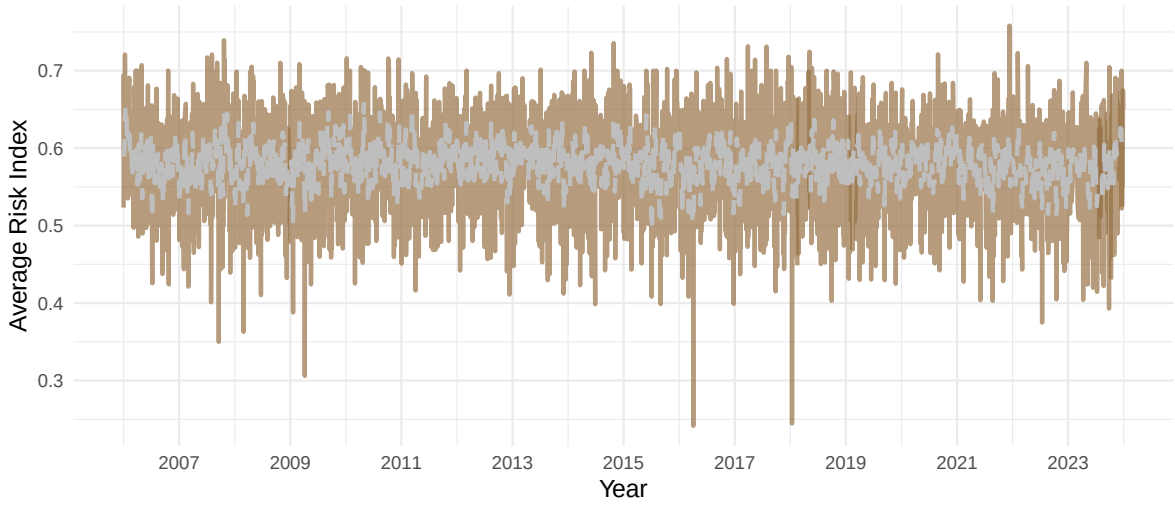


Figure 12: Long-Term Temporal Trends in Risk Index (Mid-2016 to Mid-2024) with Rolling Average

Figure 12 illustrates the long-term temporal trends in the Average Risk Index, measured daily and visualized from mid 2006 to mid 2023. The solid line represents the daily Average Risk Index, while the grey dashed line shows a six-month rolling average, providing a smoothed representation of the underlying trends.

The daily Average Risk Index illustrates considerable variability, reflecting fluctuations in risk factors such as collision and theft incidents over time. These short-term variations may be influenced by factors like weather conditions, traffic density, or seasonal changes. In contrast, the rolling average highlights more stable, long-term patterns, suggesting relatively stable long-term trends in the Risk Index, with no sudden increases or decreases over the years. However,

several periodic dips and peaks in late 2016 and mid 2018 underscore the importance of addressing short-term change in risk to enhance overall safety.

4.1.2.2 Covid Period

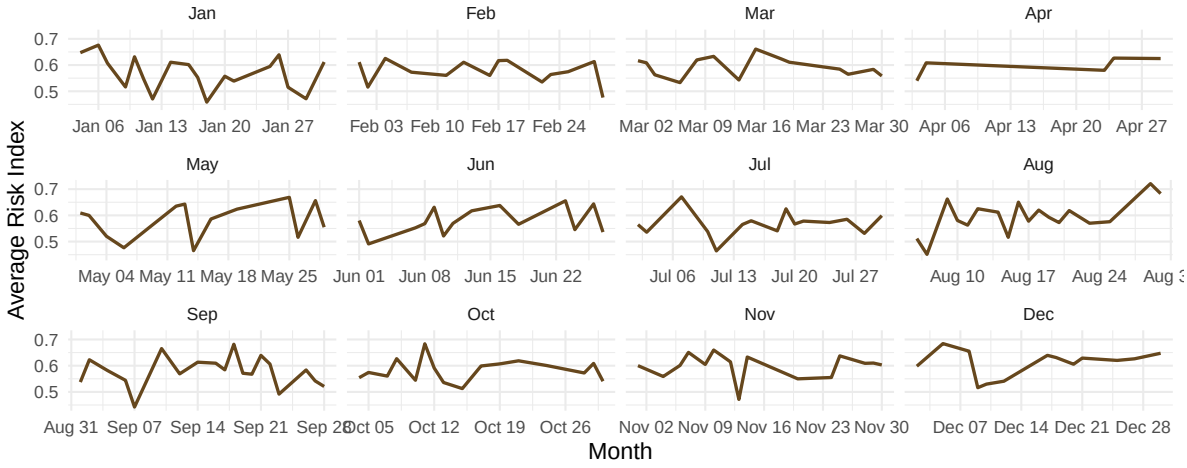


Figure 13: Risk Index Variations in 2020 during Covid-19

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 profound 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 reveals 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 insight into how these factors may contribute to the overall risk.

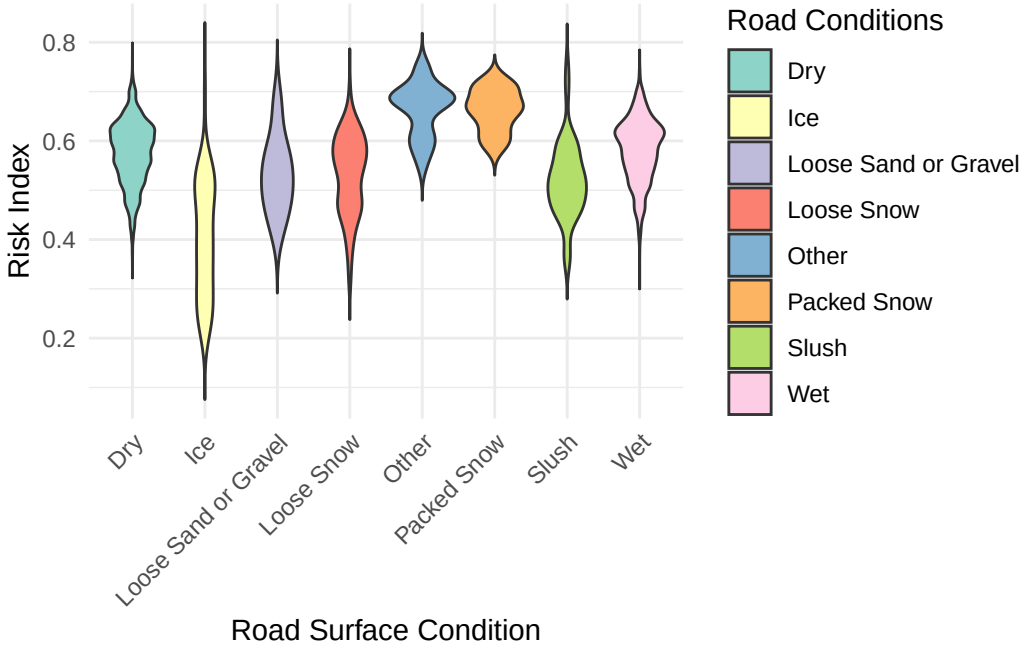
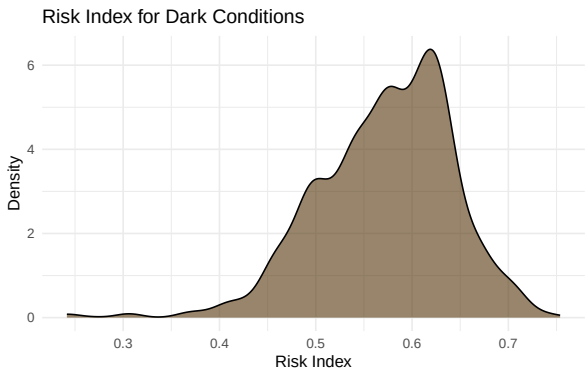


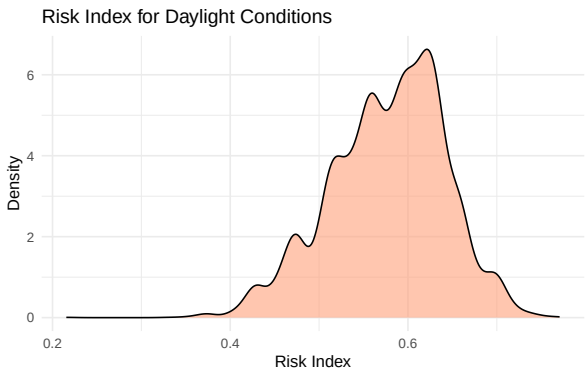
Figure 14: Risk Index Distribution by Road Surface Conditions

Figure 14 reveals 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.

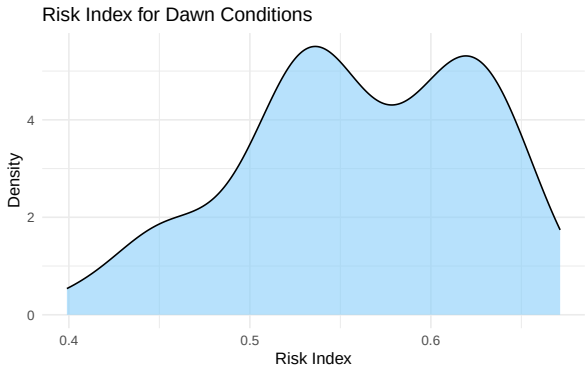
Figure 15 displays the distribution of the Risk Index across various lighting conditions. The data is broken down into four main lighting categories: Low Light (Dark & Artificial), Dawn and Artificial Light, Bright Daylight and Artificial Light, and Other Lighting Conditions.



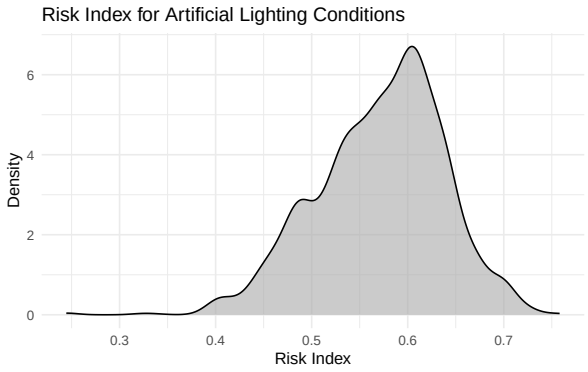
(a) Low Light (Dark & Artificial)



(b) Dawn and Artificial Light



(c) Bright Daylight and Artificial Light



(d) Other Lighting Conditions

Figure 15: Risk Index by Lighting Conditions

Figure 15a highlights that collision risks are higher under poor visibility, with a concentration of Risk Index values in the moderate-to-high range. These findings indicate that insufficient natural light, coupled with artificial lighting, may contribute to a heightened likelihood of incidents.

Figure 15b reflects a mixed risk profile, with the density curve showing a broader spread across the Risk Index. While risk levels are slightly lower compared to low-light conditions, there is still a noticeable presence of moderate-risk incidents, likely influenced by transitioning visibility during dawn hours.

Figure 15c reveals a more concentrated distribution of lower Risk Index values. This suggests that enhanced visibility and predictable traffic patterns during daylight hours contribute to reduced motorbike risks.

Figure 15d captures a unique distribution, with a small number of outlier cases showing a less consistent pattern. These include conditions not explicitly defined by the primary categories or instances where lighting data is unavailable. The density curve indicates relatively stable, moderate-risk levels, likely driven by the diversity of conditions represented in this group.

4.3 Mapping the Risk: Neighborhood-Level Insights

4.3.1 Neighborhood Risk Distribution

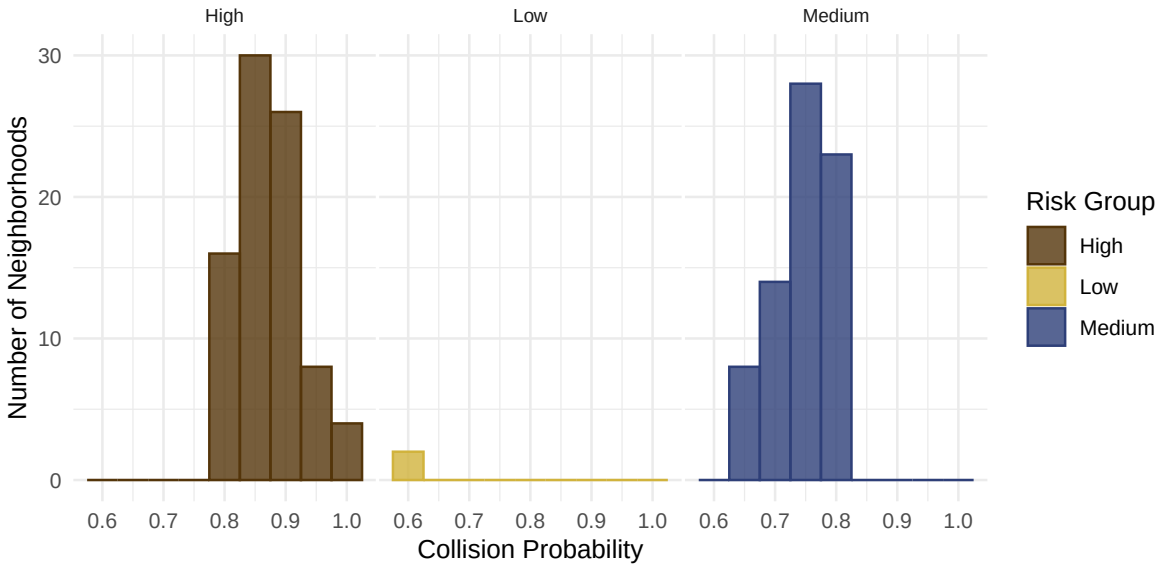


Figure 16: Neighborhoods grouped by average collision probabilities into High, Medium, and Low risk categories.

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

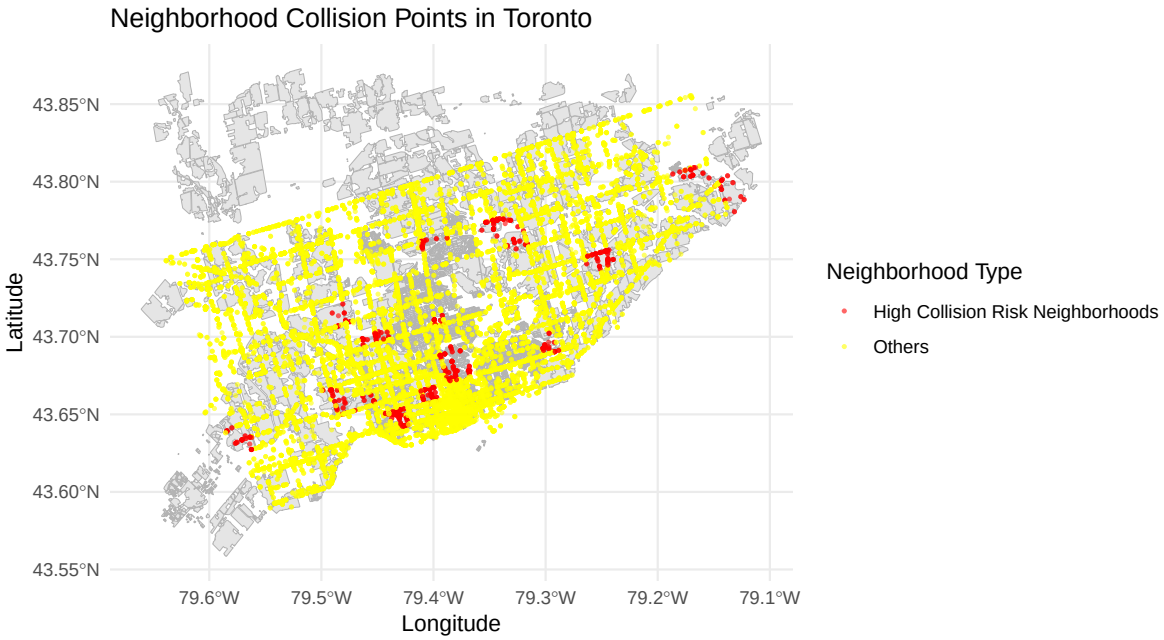


Figure 17: Neighborhood Collision Points in Toronto Highlighting High Collision Risk Areas

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 comprehensive 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

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/.

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

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.