Understanding Bicycle Theft in High-Risk Neighborhoods: Key Predictors for Targeted Risk Assessment*

An Analysis of Theft Patterns and Risk Factors in Toronto's High-Theft Areas

Tommy Fu

December 2, 2024

This study explores patterns of bicycle theft in Toronto, focusing on thefts in high-risk neighborhoods using logistic regression. The analysis identifies bike cost, premises type, and time of occurrence as significant predictors of thefts in high-risk neighborhoods. Results show that high-cost bikes are disproportionately stolen in public spaces and during evening hours, with specific neighborhoods exhibiting elevated theft rates. These findings highlight risk factors and provide actionable insights for targeted interventions, urban planning, and improved security measures in vulnerable areas.

Table of contents

1 Introduction **3**
- 1.1 Overview . . . 3
- 1.2 Estimand . . . 3
- 1.3 Results Summary . . . 3
- 1.4 Why this paper matters . . . 4
- 1.5 Paper Structure: . . . 4

2 Data **4**
- 2.1 Overview . . . 4
- 2.2 Measurement . . . 5
- 2.3 Data Processing . . . 5
- 2.4 Outcome variables . . . 6

*Code and data are available at: https://github.com/YichengFu/bike_thefts_analysis.git.

2.5 Predictor variables . . . 7
2.6 Data Visualizations . . . 7

3 Model **9**
3.1 Model Overview . . . 9
3.2 Model set-up . . . 9
3.3 Model justification . . . 10

4 Results **10**
4.1 Model Validation . . . 11

5 Discussion **13**
5.1 The Role of Premises in Shaping Bike Theft Risks . . . 13
5.2 Timing Matters: How Time of Day Influences Bike Thefts . . . 13
5.3 The Price of Vulnerability: Impact of Bike Cost on Theft Likelihood . . . 13
5.4 Limitations and Challenges in Understanding Theft Patterns . . . 14
5.4.1 Addressing the Exclusion of Location Type as a Predictor . . . 14
5.4.2 Balancing Generalizability Across Neighborhood Contexts . . . 14
5.5 Enhancing Prevention: Directions for Future Research and Policy . . . 14

Appendix **15**

A Idealized Survey Methodology **15**
A.1 Survey Objectives . . . 15
A.2 Implementation Strategy . . . 15
A.2.1 Sample Size and Target Audience . . . 15
A.2.2 Recruitment Channels . . . 16
A.2.3 Budget Allocation . . . 16
A.3 Survey Design Elements . . . 16
A.3.1 Opening Message . . . 16
A.3.2 Key Questions . . . 16
A.3.3 Closing Message . . . 18
A.4 Bias Mitigation Strategies . . . 18
A.5 Tradeoffs and Limitations . . . 18

B Additional data details **19**
B.1 Raw data variables . . . 19
B.2 Data visualization . . . 20

C Model details **22**
C.1 Posterior predictive check . . . 22
C.2 Markov chain Monte Carlo Convergence Check . . . 24
C.2.1 90% Credibility Interval . . . 24

References **28**

1 Introduction

1.1 Overview

Bicycle theft is a persistent urban problem with significant economic and emotional impacts on individuals and communities. In cities like Toronto, where cycling plays an increasingly vital role in promoting sustainable transportation, understanding the factors contributing to bicycle theft is essential for designing effective preventive measures. While previous studies have examined general theft trends, few have focused specifically on identifying patterns in high-risk neighborhoods, where targeted interventions could have the greatest impact. In this paper the bike thefts data from Toronto Police Open Data will be utilized . This paper seeks to address this gap by exploring the spatial, temporal, and contextual factors associated with bicycle thefts in Toronto.

1.2 Estimand

The estimand of this study is the likelihood of thefts occurring in high-risk neighborhoods compared to others, given key predictors such as bike cost, premises type, and time of theft. The estimand focuses on understanding the characteristics of thefts in these neighborhoods, allowing us to identify significant factors that differentiate high-risk areas from others. By using a logistic regression model, the analysis aims to quantify these relationships and provide actionable insights.

1.3 Results Summary

The findings of this study reveal significant links between theft occurrences in high-risk neighborhoods and factors such as bike cost, premises type, and time of day. High-cost bicycles are more frequently stolen, reflecting their appeal as valuable targets. Public spaces, including streets and parks, show higher odds of theft compared to more secure environments like houses or garages, likely due to easier access and fewer security measures.

Thefts are more likely to occur during evening hours compared to morning or daytime periods, a trend that aligns with reduced visibility and activity levels during these hours. Additionally, some neighborhoods consistently report higher theft rates, pointing to localized factors such as infrastructure, socioeconomic conditions, or enforcement levels. These findings suggest the need for tailored strategies, including improved lighting, enhanced surveillance, and increased public awareness, to reduce theft risks in vulnerable areas and times.

1.4 Why this paper matters

This research provides actionable knowledge to enhance theft prevention in urban settings. By identifying the circumstances under which thefts are more likely to occur, the study offers evidence to inform targeted interventions, such as improved security measures in vulnerable areas. These findings contribute to creating safer urban spaces for cyclists.

1.5 Paper Structure:

The remainder of this paper is structured as follows. In Section 2, the overview of the data used in this study and the variables of interests will be introduced. Further the data normalization will be discussed in details. Section 3 illustrates the Bayesian logistic regression model built in our analysis, some details include model set up, assumptions and justification. Section 4 highlights the result of the model visualizing using tables and graphs. Lastly, Section 5 contains discussion of the analysis based on findings, the limitations of the model and the suggestion for future research. Section A contains information about the survey as well as its design and limitations. Section B contains more information about the original data and its visualizations. Section C shows more plots and information about the model built for this research.

2 Data

2.1 Overview

Our dataset is sourced from Toronto Police Open Data (Toronto Police Service 2024), specifically the "Bicycle Thefts" dataset, which provides comprehensive details on bicycle thefts reported across the Greater Toronto Area. This dataset includes variables that capture information about the stolen bicycles, such as their reported cost, make, and type, as well as contextual details about the thefts, including the date, time, premises type, and neighborhood. Additionally, geospatial data, including latitude and longitude coordinates, enables spatial analysis to explore patterns in theft occurrences across different areas. Heatmap of Toronto area by theft counts and other informative plots will be shown in Section 2.6. The heatmap utilized shapefile data sourced from University of Toronto Map Library (University of Toronto n.d.).

The dataset's level of granularity allows for an in-depth examination of theft dynamics, facilitating an analysis of how temporal, spatial, and contextual factors interact to influence theft risk. For example, information on premises type provides insights into whether thefts are more common in public or private spaces, while bike cost highlights economic factors associated with thefts. A detailed list of the variables, along with sample values, is provided in the appendix Section B (Table 3) to offer additional context for understanding the dataset. This structured

data enables a robust analysis aimed at identifying patterns and predictors of bicycle theft across Toronto.

2.2 Measurement

This study utilizes data from Toronto Police open data portal (Toronto Police Service 2024), specifically focusing on detailed records of bicycle theft incidents reported across the city. The dataset includes a rich variety of variables that capture theft characteristics, such as the reported value of the bike, the date and time of the theft, and the theft's status, alongside spatial details like the neighborhood and premises type. Additionally, geospatial coordinates, including latitude and longitude, enable detailed mapping and spatial analysis of theft hotspots. These features make the dataset well-suited for understanding theft dynamics in high-risk neighborhoods, where theft patterns may differ based on socioeconomic, environmental, or infrastructural factors.

The dataset focuses on stolen bicycles, eliminating the noise of other types of crimes and allowing for a more precise examination of the contextual and temporal factors influencing bicycle thefts. Key variables, such as BIKE_COST, provide insights into economic patterns of theft, while OCC_DATE and OCC_HOUR allow for the identification of temporal trends. Variables like PREMISES_TYPE add further depth by categorizing thefts based on their environmental and spatial contexts, helping to discern whether certain locations, such as streets, parks, or garages, are more vulnerable than others.

While other datasets, such as police crime reports or neighborhood demographic data, could have supplemented this analysis, their limited accessibility and lack of detailed information on bike-specific incidents made them unsuitable for the current study. The chosen dataset's specificity and granularity ensure that the analysis remains focused and relevant to the objective of understanding theft patterns in high-risk neighborhoods. These features enable a rigorous exploration of the relationships between bike attributes, theft characteristics, and spatial factors, supporting the development of targeted interventions to reduce theft risks in vulnerable areas.

2.3 Data Processing

The raw dataset, sourced from Toronto Police open data portal (Toronto Police Service 2024), underwent comprehensive processing steps to ensure it was accurate, relevant, and ready for analysis. One of the initial steps involved addressing missing values. Variables with a high proportion of missing data, such as BIKE_MODEL, BIKE_SPEED, and BIKE_COLOUR, were removed as they offered limited analytical value. Additionally, observations with missing critical values like BIKE_COST or BIKE_MAKE were filtered out to maintain data completeness and integrity. This ensured the dataset included only records with sufficient detail for analysis. More details available in Section B

The dataset was further refined by excluding theft incidents where the STATUS variable was "UNKNOWN" or "RECOVERED," as these cases were not directly relevant to the study's focus on thefts. Temporal variables, including OCC_DATE (the occurrence date of the theft) and REPORT_DATE (the date the theft was reported), were standardized to a uniform Date format, facilitating the analysis of trends over time, including seasonal and hourly patterns.

A crucial variable, is_high_risk_neighborhood, was constructed to flag thefts occurring in the top 10 neighborhoods with the highest theft frequencies. This variable was derived by counting incidents per neighborhood and identifying the areas most affected by theft, enabling the study to focus on high-risk locations. The geospatial coordinates LONG_WGS84 and LAT_WGS84 were retained to enable mapping and spatial analysis of theft incidents, providing a foundation for visualizing theft hotspots. Contextual variables, such as PREMISES_TYPE (e.g., house, garage, public area), were also preserved to explore environmental factors influencing theft risk.

The cleaned dataset was saved as a Parquet file, chosen for its efficient storage and compatibility with downstream modeling and visualization workflows. This format allowed for fast read-write operations and seamless integration with analysis tools. The cleaning and preparation steps ensured that the dataset was robust and aligned with the study's goal of identifying theft patterns and predictors in high-risk neighborhoods.Packages used in this paper are 'tidyverse' (Wickham et al. 2019), 'here' (Müller 2020),'arrow' (Richardson et al. 2024), 'lubridate'(Grolemund and Wickham 2011), 'testthat' (Wickham 2011), 'rstanarm' (Brilleman et al. 2018), 'sf'(Pebesma 2018), 'knitr' (Xie 2014), 'forcats' (Wickham 2023) and 'kableExtra' (Zhu 2024).

2.4 Outcome variables

The dataset includes several important variables that serve as predictors in this study. These include `BIKE_COST`, a numeric variable representing the reported value of stolen bicycles, which is crucial for understanding how bike value influences theft patterns. `PREMISES_TYPE` is categorical variables providing contextual details about where thefts occurred, such as public spaces or residential areas, and their environmental settings like streets or parks. Temporal details are captured through variables such as `OCC_HOUR`, representing the hour of the day the theft occurred, and `OCC_DATE`, which allows for trends and seasonal patterns to be explored. Geographic variables like `NEIGHBOURHOOD_140` and the corresponding longitude and latitude coordinates provide spatial context, enabling an examination of how theft patterns vary across Toronto neighborhoods. Together, these variables form the foundation for identifying significant factors associated with bicycle theft patterns.

Table 1: Preview of the Cleaned Data

Occ Date	Location	Premises	Bike Cost
12/26/2013 5:00:00 AM	Other Commercial / Corporate Places (For Profit, Warehouse, Corp. Bldg	Commercial	1300
12/30/2013 5:00:00 AM	Streets, Roads, Highways (Bicycle Path, Private Road)	Outside	500
9/30/2013 5:00:00 AM	Apartment (Rooming House, Condo)	Apartment	750
12/25/2013 5:00:00 AM	Apartment (Rooming House, Condo)	Apartment	1500
12/25/2013 5:00:00 AM	Streets, Roads, Highways (Bicycle Path, Private Road)	Outside	400

2.5 Predictor variables

The variable `is_high_risk_neighborhood` is a binary indicator designed to classify thefts based on whether they occurred in one of the top 10 neighborhoods with the highest theft frequencies. This variable was constructed to focus the analysis on areas most impacted by bicycle theft. A value of 1 indicates that the theft occurred in a high-risk neighborhood, while 0 indicates all other neighborhoods. This classification highlights localized clusters of theft activity and enables the study to identify key predictors of theft in these vulnerable areas.

2.6 Data Visualizations

The majority of reported bicycle thefts occur in a small subset of neighborhoods, demonstrating significant clustering in high-risk areas. Figure 1 illustrates the distribution of thefts by risk classification with different premises types comparison in top ten risky neighborhoods . Figure 2 visualizes bicycle theft counts by neighborhood across Toronto, highlighting spatial patterns in theft frequency. More detailed visualizations of the data are located in Section B.2

Figure 1 illustrates the distribution of bike thefts across different premises types within high-risk neighborhoods. Despite encompassing only 10 neighborhoods, these areas contribute significantly to the overall theft count, underscoring their vulnerability. The figure also highlights variability in premises types, with public and commercial spaces such as “Outside” and “Commercial” showing the highest theft counts, likely due to high foot traffic and lower security measures. Conversely, private premises like “House” exhibit fewer incidents, indicating

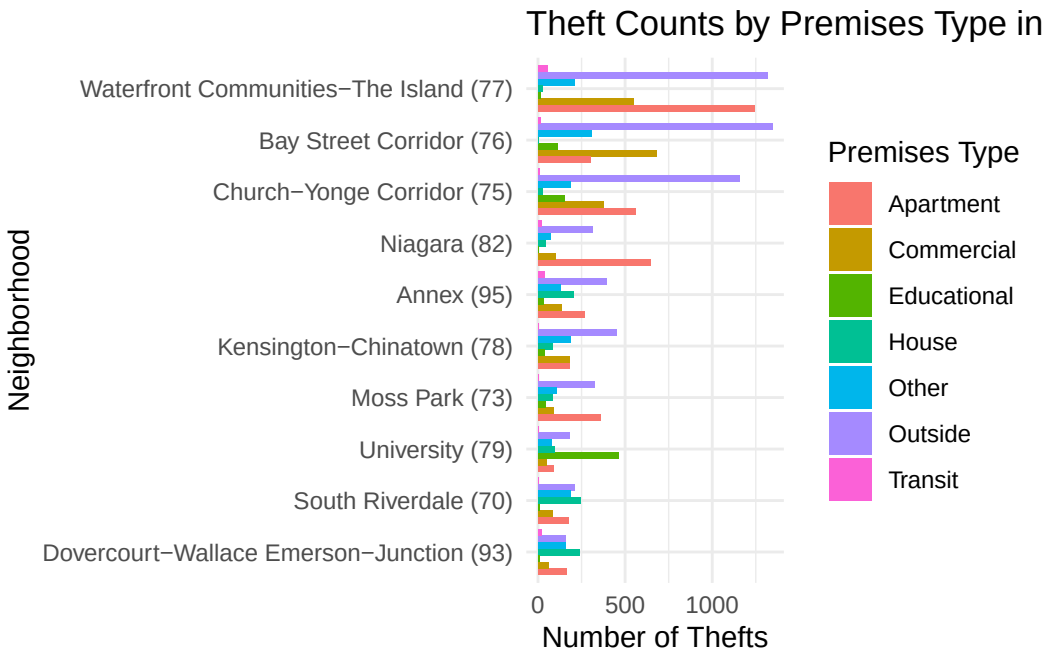


Figure 1: Theft Counts by Premises Type in High-Risk Neighborhoods

better security controls. These findings emphasize the necessity of tailored intervention strategies, such as enhanced surveillance and secure bike storage options, to mitigate theft risks in these high-priority areas.

Figure 2 ranges from dark purple (low theft count) to bright yellow (high theft count), indicates the relative number of bicycle thefts in each neighborhood. The visualization reveals that bicycle thefts are highly concentrated in certain neighborhoods, particularly in the downtown core and surrounding areas, as indicated by the bright yellow and orange regions. These areas may be more susceptible to theft due to higher population density, greater cycling activity, or specific environmental factors such as the availability of parking infrastructure or security measures. This map underscores the spatial disparity in theft incidents and suggests that targeted interventions, such as improved bike parking security and increased awareness campaigns, may be particularly beneficial in high-theft neighborhoods. This visualization also provides a foundation for further analysis, such as exploring the relationship between theft hotspots and socioeconomic or infrastructure variables.

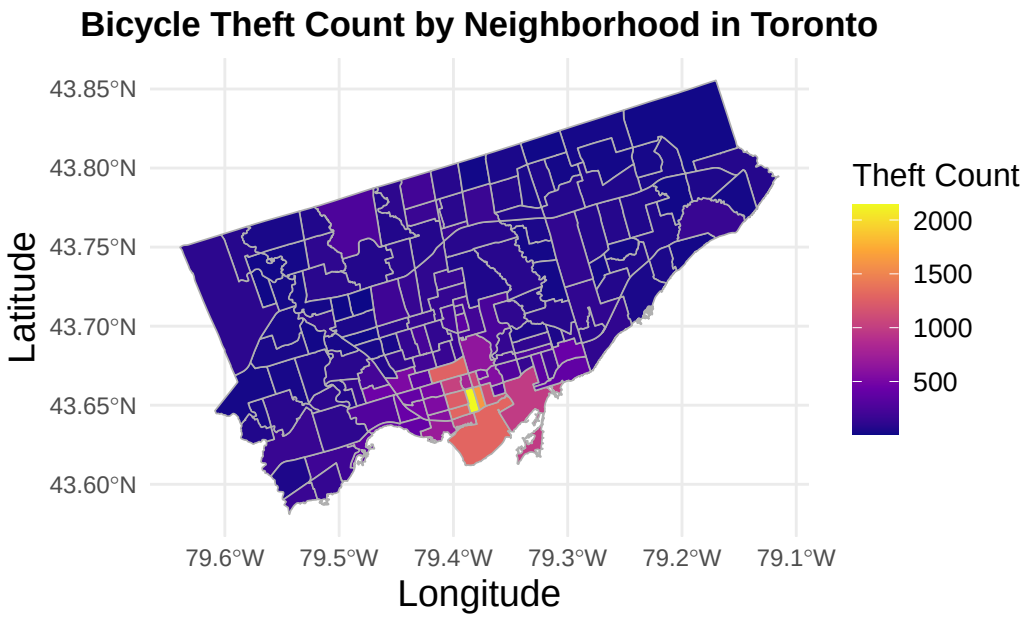


Figure 2: Bicycle Theft Count by Neighborhood in Toronto

3 Model

3.1 Model Overview

In this analysis, we utilized a Bayesian logistic regression model to examine the factors associated with bicycle thefts in high-risk neighborhoods. The dependent variable is `is_high_risk_neighborhood`, a binary indicator identifying whether a theft occurred in one of the top 10 neighborhoods with the highest theft counts. Detailed model diagnostics and background information are available in Appendix Section C.

3.2 Model set-up

The model is specified as follows:

$$y_i \mid \pi_i \sim \text{Bern}(\pi_i) \tag{1}$$

$$\begin{aligned}\text{logit}(\pi_i) = \alpha + \beta_1 \times \text{LOG\_BIKE\_COST}_i + \beta_2 \times \text{PREMISES\_TYPE}_i \\ + \beta_3 \times \text{OCC\_HOUR}_i\end{aligned} \tag{2}$$

$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1, \beta_2, \beta_3 \sim \text{Normal}(0, 2.5) \tag{4}$$

Here, y_i represents the binary outcome variable indicating whether a theft occurred in a high-risk neighborhood. The probability of this event, (π_i) is modeled using a logistic link function. Predictors in the model include the reported BIKE_COST, the PREMISES_TYPE of the theft, and the time of day (OCC_HOUR) when the theft occurred.Weakly informative priors were used to regularize the model. Specifically, all parameters(α,β_1,β_2,β_3) were assigned normal prior distributions with a mean of 0 and a standard deviation of 2.5. Sampling for the model was conducted using Markov Chain Monte Carlo (MCMC) methods implemented in the rstanarm package Brilleman et al. (2018) in R(R Core Team 2023). To optimize runtime, a random sample of 1000 data entries was used, with a seed of 215 to ensure reproducibility. Model diagnostics, including convergence checks and posterior summaries, are presented in Appendix Section C.

3.3 Model justification

The Bayesian logistic regression was chosen for its suitability for binary outcome variables and its ability to incorporate prior information. This approach enables a probabilistic interpretation of results, allowing for uncertainty quantification in parameter estimates. The use of weakly informative priors helps stabilize the model and prevents overfitting, particularly given the limited sample size.

Regarding the predictors, we hypothesize the following relationships:

-BIKE_COST: Higher-cost bicycles are more likely to be targeted in thefts due to their resale value and desirability. -PREMISES_TYPE: Theft risk may be higher in public or semi-public spaces, such as streets and parks, where security measures are limited. -OCC_HOUR: Theft likelihood is expected to increase during evening hours when visibility and public activity are reduced.

The Bayesian logistic regression model provides a robust framework for examining these relationships, offering insights into the factors influencing thefts in high-risk neighborhoods.

4 Results

Our results are summarized in Table 2. The intercept, representing the baseline log-odds of a theft occurring in such neighborhoods when all other predictors are held constant, is estimated at -0.386. This negative value suggests that, in the absence of other influencing factors, the likelihood of thefts in high-risk areas is relatively low. The effect of bike cost, represented as the log-transformed variable (log_bike_cost), is minimal and statistically insignificant (0.031). This indicates that, after accounting for premises type and time of day, the cost of the bike alone does not strongly affect the likelihood of thefts occurring in high-risk neighborhoods. While high-cost bikes may attract theft in general, this result implies that contextual factors, rather than cost, play a more significant role in shaping theft patterns in these areas.

Premises type emerges as an important factor with notable variability in its impact. Theft likelihood is higher in "Commercial" (0.935) and "Educational" (0.630) premises compared to the reference category, suggesting that these locations may be more accessible or targeted by thieves. Conversely, thefts are less likely to occur at "House" (−1.034) and "Other" (−0.473) premises, indicating these settings may offer more security or deterrents. "Outside" premises (0.662) show a moderate positive association, while "Transit" (−0.575) premises display a slight negative effect, suggesting varying levels of theft risk depending on location type. Time of day, categorized into "Morning," "Afternoon," "Evening," and "Night," shows relatively small effects on theft likelihood. Compared to "Afternoon" (the reference category), "Evening" (−0.120), "Night" (−0.132), and "Morning" (−0.162) exhibit slightly reduced log-odds. These findings suggest minimal differences in theft risk across times of the day within high-risk neighborhoods, potentially reflecting consistent levels of vigilance or opportunity.

The model's fit, as indicated by an R^2 value of 0.100, shows that 10% of the variation in theft likelihood is explained by the included predictors. While this value suggests the presence of unaccounted-for factors, the model's log-likelihood and information criteria (LOOIC: 1299.7, WAIC: 1299.6) indicate reasonable performance for this type of analysis.

Credible intervals for the predictors, visualized in Figure 11, reveal that several premises types, such as "Commercial" and "House," show statistically significant effects, as their intervals do not cross zero. These results highlight the importance of contextual factors like premises type in determining theft likelihood and underscore the need for targeted strategies to mitigate risks in high-risk neighborhoods.

Overall, the findings emphasize that while bike cost may not significantly influence thefts in high-risk areas, the type of premises and broader contextual factors play a pivotal role. These insights can inform more effective theft prevention strategies by focusing on specific high-risk premises and addressing vulnerabilities in these environments.

4.1 Model Validation

For posterior predictive checks, Figure 6 demonstrates that the posterior distribution from our Bayesian logistic regression model aligns closely with the observed data on bike thefts in high-risk neighborhoods. This suggests that the model captures the underlying patterns of theft occurrences accurately, supporting the robustness of its predictions. Similarly, Figure 7 compares the posterior to the prior distributions, highlighting parameter changes such as those for "log_bike_cost" and "time_of_day." These shifts indicate that the observed data provided substantial information, refining our prior beliefs about these predictors.

The trace plots in Figure 8 & Figure 9 show stable and horizontal chains with adequate mixing across iterations, suggesting no convergence issues in the Markov chain Monte Carlo process. Furthermore, the Rhat plot in Figure 10 confirms this conclusion, as all Rhat values are close to 1 and well below the threshold of 1.05. This demonstrates strong convergence and reliability of the posterior estimates.

Table 2: Bike Thefts in High Risk Area(n=1000)

	Bike Thefts in High Risk Area
(Intercept)	−0.386
	(0.304)
log_bike_cost	0.031
	(0.041)
PREMISES_TYPECommercial	0.935
	(0.248)
PREMISES_TYPEEducational	0.630
	(0.346)
PREMISES_TYPEHouse	−1.034
	(0.238)
PREMISES_TYPEOther	−0.473
	(0.233)
PREMISES_TYPEOutside	0.662
	(0.172)
PREMISES_TYPETransit	−0.575
	(0.461)
time_of_dayEvening	−0.120
	(0.169)
time_of_dayMorning	−0.162
	(0.189)
time_of_dayNight	−0.132
	(0.218)
Num.Obs.	1000
R2	0.102
Log.Lik.	−640.567
ELPD	−651.8
ELPD s.e.	10.1
LOOIC	1303.6
LOOIC s.e.	20.2
WAIC	1303.6
RMSE	0.47

Additional details and supporting figures can be found in Appendix Section C.

5 Discussion

5.1 The Role of Premises in Shaping Bike Theft Risks

Our analysis highlights the significant influence of premises type on the likelihood of bike thefts. Public spaces, including "Commercial" and "Outside" premises, exhibit the highest estimated coefficients, reinforcing their status as high-risk areas. This finding is consistent with the assumption that public and commercial spaces, with their high foot traffic and limited security, offer increased opportunities for theft. Conversely, "House" premises have a negative coefficient, indicating a reduced likelihood of theft when bikes are stored at private residences. This reflects the higher level of control and security afforded by private properties.

An unexpected finding pertains to "Transit" premises, which also display a negative coefficient. Although transit areas are often perceived as high-risk due to the volume of commuters, the results suggest that targeted measures, such as surveillance cameras and police patrols, may be deterring theft in these zones. However, the wider confidence intervals around this estimate indicate some uncertainty, highlighting the need for further investigation into specific security practices in transit areas.

5.2 Timing Matters: How Time of Day Influences Bike Thefts

The timing of bike thefts also emerges as an important factor. Our model categorizes thefts into "Morning," "Afternoon," "Evening," and "Night" periods. Although the differences are subtler compared to premises types, theft likelihood increases slightly during "Night" hours, as reflected by its positive coefficient. This finding aligns with the widely held belief that lower visibility and reduced pedestrian presence at night create favorable conditions for theft.

"Morning" and "Afternoon" periods, in contrast, exhibit lower coefficients, suggesting reduced theft activity during these times. This could be attributed to the higher visibility and increased presence of commuters, shoppers, and general public activity during daylight hours, which may act as a deterrent. These findings underscore the potential for targeted prevention measures, such as enhanced lighting and increased surveillance, particularly during evening and night hours.

5.3 The Price of Vulnerability: Impact of Bike Cost on Theft Likelihood

The value of the bike, as measured by its log-transformed cost, plays a key role in predicting theft likelihood. The positive coefficient indicates that more expensive bikes are more likely to be targeted. This aligns with expectations, as high-value bikes offer greater potential resale

value for thieves. However, the relatively small effect size suggests that factors like ease of theft and accessibility may often outweigh bike cost in determining theft likelihood.

This result underscores the need for increased awareness among bike owners, particularly those with high-value bikes, regarding secure storage practices. It also highlights the potential benefits of policy interventions, such as secure parking facilities in theft-prone areas, to mitigate risks associated with high-cost bicycles.

5.4 Limitations and Challenges in Understanding Theft Patterns

5.4.1 Addressing the Exclusion of Location Type as a Predictor

Due to multicollinearity issues, location type was excluded from the final model. This limits the granularity of our findings, particularly in understanding theft risks associated with specific public areas, such as parks or transit stations. While premises type captures broader contextual differences, a more refined inclusion of location-based variables could provide a deeper understanding of the spatial dynamics of bike theft.

5.4.2 Balancing Generalizability Across Neighborhood Contexts

The model focuses on thefts occurring in high-risk neighborhoods, which may limit its applicability to areas with lower theft incidence. While this approach allows for targeted insights into urban centers and densely populated regions, expanding the dataset to include a more diverse range of neighborhoods could enhance generalizability and provide a fuller picture of theft dynamics across different settings.

5.5 Enhancing Prevention: Directions for Future Research and Policy

Future research should aim to incorporate additional contextual variables, such as weather conditions, population density, and local crime rates, to improve predictive accuracy. Enhanced geospatial analyses, including hot spot mapping within neighborhoods, could further refine our understanding of theft patterns. Moreover, investigating the effectiveness of interventions such as secure parking zones, surveillance technologies, and community engagement programs could provide actionable insights to reduce bike theft. These findings could inform policies and practices aimed at improving bike security and safety in urban areas.

Appendix

A Idealized Survey Methodology

A.1 Survey Objectives

The primary objective of this survey is to understand the factors influencing bike theft patterns in Toronto, with a focus on high-risk neighborhoods. The survey aims to capture data that cannot be directly observed through traditional theft reports, such as:

- Causal Relationships: Identifying how premises type, time of day, and bike value influence theft likelihood.
- Behavioral Insights: Exploring bike owners' security habits and their perception of theft risks.
- Perception and Awareness: Assessing public understanding of theft risks and secure parking availability.
- Counterfactual Scenarios: Investigating how potential changes in infrastructure or security options might alter individual behavior.

This survey is designed to directly inform strategies for targeted interventions and policy recommendations by capturing both perceptions and actual behaviors. By comparing responses from high-risk and low-risk neighborhoods, we aim to uncover nuanced insights into theft patterns and prevention opportunities.

A.2 Implementation Strategy

A.2.1 Sample Size and Target Audience

The survey will focus on three key groups of participants:

- Bike Theft Victims: Individuals who have experienced theft within the last five years.
- Current Cyclists: Bike owners who actively use their bikes for commuting or recreation.
- Residents Near High- and Low-Risk Areas: To compare theft patterns across neighborhoods.

The target sample size is set at **3,000 respondents**, divided as follows:

- 1,000 participants from **high-risk neighborhoods**.
- 1,000 participants from **low-risk neighborhoods**.
- 1,000 individuals with **no prior experience of bike theft** to gather perceptions of theft risk.

A.2.2 Recruitment Channels

To ensure diverse participation, we employ the following recruitment strategies:

1. **Digital Targeting**: Social media advertisements on platforms like Instagram, Facebook, and X (Twitter), geo-targeted to areas with reported theft incidents and high bike usage.
2. **Community Outreach**: Partnerships with local cycling groups, bike repair shops, and advocacy organizations to distribute the survey link.
3. **Email Distribution**: Outreach to residents through neighborhood associations and cycling networks.
4. **Incentives**: Participants receive a $20 gift card for completing the survey.

A.2.3 Budget Allocation

The survey has a total budget of **$100,000**, allocated as follows:

- **$20,000**: Digital advertising and outreach campaigns.
- **$15,000**: Partnerships with cycling organizations for outreach.
- **$65,000**: Participant incentives, ensuring a high response rate.

A.3 Survey Design Elements

A.3.1 Opening Message

Participants will be welcomed with the following message:

"Thank you for participating in our survey on bike thefts in Toronto. We aim to understand theft patterns and develop strategies to improve security for cyclists. Your input is invaluable and will help inform policies to reduce bike thefts. This survey will take approximately 5 minutes to complete, and your responses will remain anonymous. As a thank-you, you will receive a $20 gift card though your email within 3 business days."

A.3.2 Key Questions

The survey includes structured questions targeting bike theft experiences and perceptions:

1. Have you experienced bike theft in the last five years?

- Yes/No

2. If yes, where did it occur?

- Home

- Commercial Area
- Transit Station
- Public Park
- Other

3. Did you report the theft to authorities?

- Yes/No/Other

4. Approximate value of the stolen bike?

- Open-ended

5. Select the word “Blue”.

- Red
- Blue
- Orange
- Green

6. How do you secure your bike? Check all that apply:

- U-lock
- Chain lock
- Secure parking
- Other

7. Where do you usually park your bike?

- Street
- Garage
- Apartment Building
- Transit Station
- Other

8. On a scale of 1–5, how confident are you in your bike’s security measures?

- 1 = Not confident, 5 = Very confident

9. Which time of day do you perceive as most vulnerable for bike thefts?

- Morning
- Afternoon
- Evening
- Night

10. Which neighborhoods do you consider high-risk for bike thefts?

- Open-ended

11. Are you aware of secure bike parking facilities in your area?

- Yes/No

A.3.3 Closing Message

"Participants receive a confirmation message:"Thank you for completing the survey! Your input will help us understand and reduce bike thefts in Toronto. Your $20 gift card will be emailed within three business days.

For any questions about the survey or your gift card, please contact: tommy.fu@mail.utoronto.ca
"

Idealized Survey can be found here: https://forms.gle/TEMMyXt31XbYMW6c7

A.4 Bias Mitigation Strategies

To ensure accurate and reliable data, the survey implements the following bias mitigation strategies:

1. **Anonymity and Neutral Question Design**: Responses will remain confidential, and questions are designed to avoid leading participants toward specific answers.
2. **Survey Length and Engagement**: The survey is limited to 5 minutes, with a mix of question types to maintain participant interest.
3. **Monitoring Selection Bias**: Detailed non-response data will be collected, and statistical corrections will be applied to account for any potential bias.

A.5 Tradeoffs and Limitations

Despite careful design, the survey faces several limitations:

- **Selection Bias**: Participants may disproportionately include individuals who have experienced theft or are particularly concerned about security.
- **Hypothetical Scenarios**: Responses to hypothetical questions may not fully align with real-world behaviors.
- **Generalizability**: Findings may reflect Toronto-specific dynamics and may not be directly applicable to other urban areas.

Despite these limitations, this survey is a critical tool for exploring theft dynamics and informing effective prevention strategies. The insights gained will complement other data sources and contribute to a deeper understanding of bike theft patterns in Toronto.

B Additional data details

B.1 Raw data variables

Table 3: The list of variables and a sample value from the raw data

Table 3: List of Variables in the Raw Dataset with Sample Values

Variable Name	Sample Value
OBJECTID	1
EVENT_UNIQUE_ID	GO-20141261431
PRIMARY_OFFENCE	THEFT UNDER
OCC_DATE	1/1/2014 5:00:00 AM
OCC_YEAR	2014
OCC_MONTH	January
OCC_DOW	Wednesday
OCC_DAY	1
OCC_DOY	1
OCC_HOUR	7
REPORT_DATE	1/1/2014 5:00:00 AM
REPORT_YEAR	2014
REPORT_MONTH	January
REPORT_DOW	Wednesday
REPORT_DAY	1
REPORT_DOY	1
REPORT_HOUR	7
DIVISION	D14
LOCATION_TYPE	Apartment (Rooming House, Condo)
PREMISES_TYPE	Apartment
BIKE_MAKE	SUPERCYCLE
BIKE_MODEL	NA
BIKE_TYPE	MT
BIKE_SPEED	10
BIKE_COLOUR	NA
BIKE_COST	NA
STATUS	STOLEN
HOOD_158	085
NEIGHBOURHOOD_158	South Parkdale (85)
HOOD_140	085
NEIGHBOURHOOD_140	South Parkdale (85)
LONG_WGS84	-79.4436451187837
LAT_WGS84	43.6376571871944

Variable Name	Sample Value
x	-8843626.12140861
y	5409538.95619472

B.2 Data visualization

Figure 3 is the Ranked Bar Plot for Top 10 High-Risk Neighborhoods. Figure 4 shows the bike thefts count by the time of the day. Figure 5 shows the bike thefts count by different type of premises.

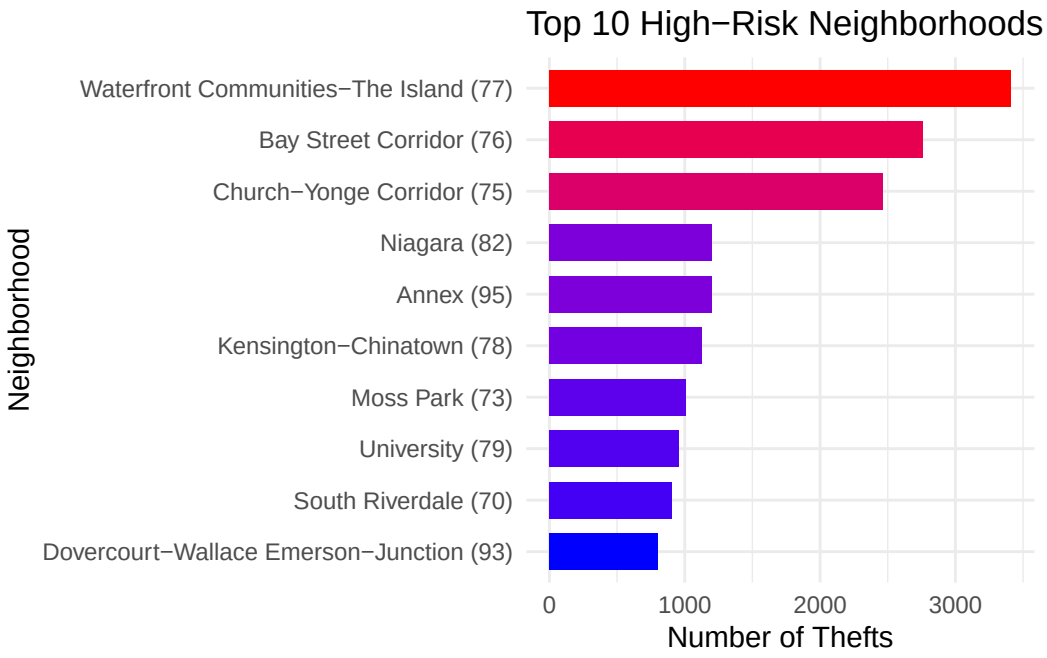


Figure 3: Ranked Bar Plot for Top 10 High-Risk Neighborhoods

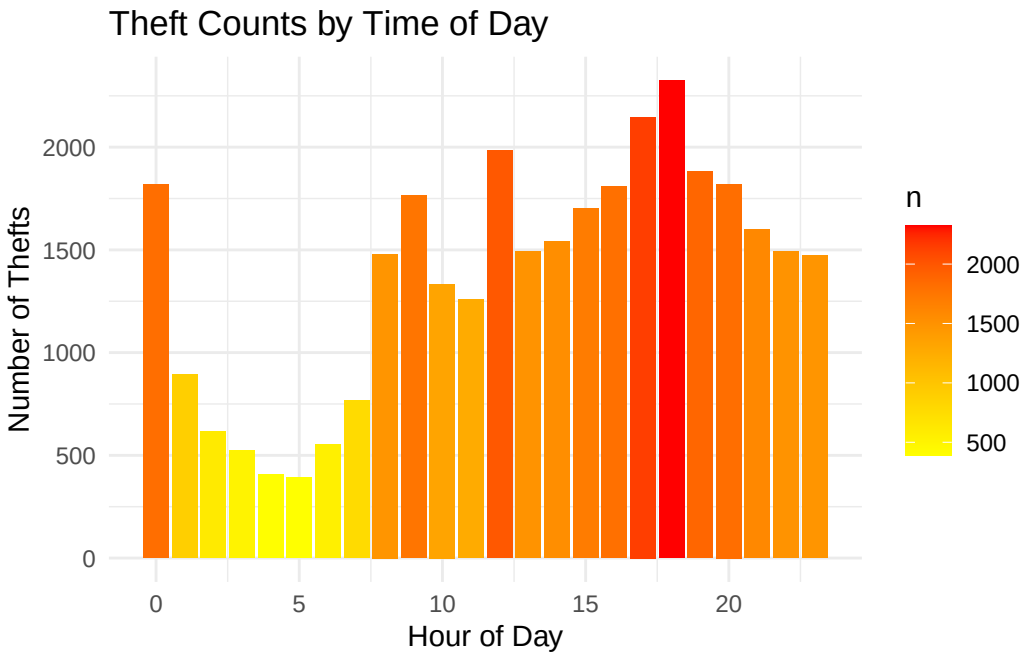


Figure 4: Theft Counts by hour of the day

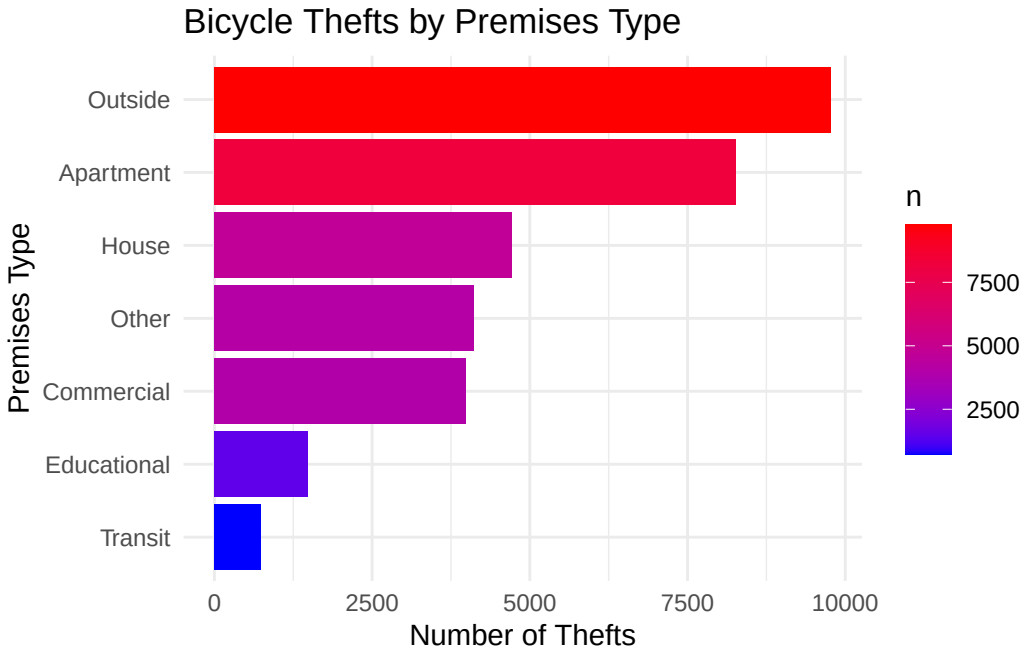


Figure 5: Bicycle Thefts by Premises Type

C Model details

C.1 Posterior predictive check

Figure Figure 6 shows the posterior predictive distribution, comparing the observed data (actual theft occurrences in high-risk neighborhoods) with predictions generated by the posterior distribution of our Bayesian logistic regression model. The close alignment between observed and predicted data indicates that the model is well-calibrated and effectively captures patterns in the data. This provides confidence that the model represents the underlying processes driving theft patterns and is capable of producing reliable inferences.

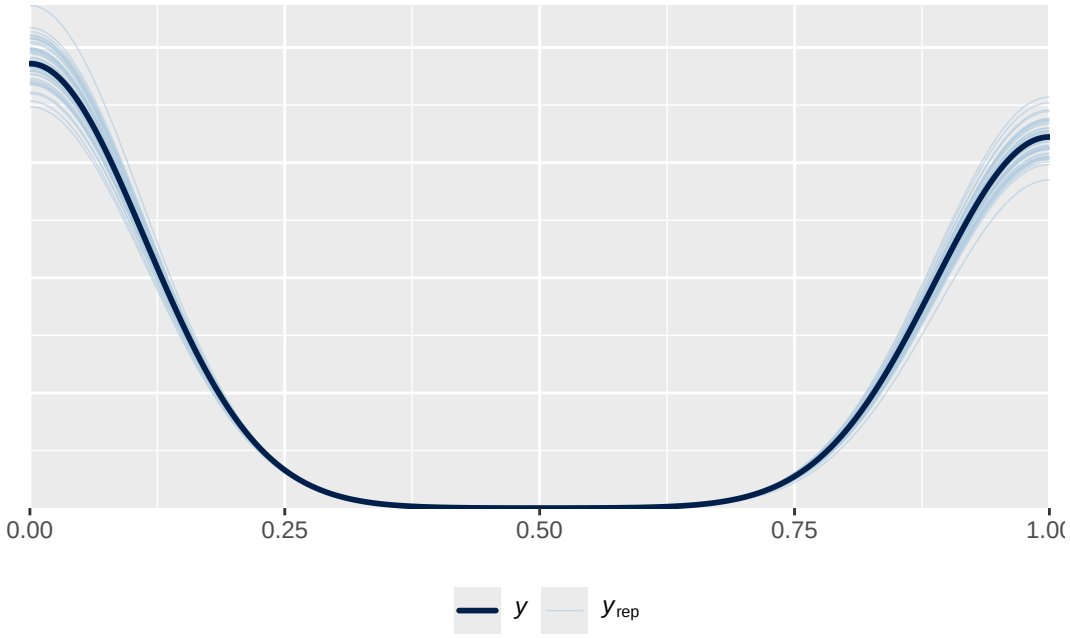


Figure 6: Posterior distribution for logistic regression model

Figure Figure 7 compares the posterior and prior distributions of the model parameters. For some predictors, such as thefts in high-risk neighborhoods involving specific premises types or during certain times, the posterior distributions shift significantly compared to the priors. This indicates that the observed data strongly influenced the parameter estimates, updating our prior beliefs. For others, the posterior and prior distributions overlap substantially, suggesting that the observed data aligns well with prior expectations.

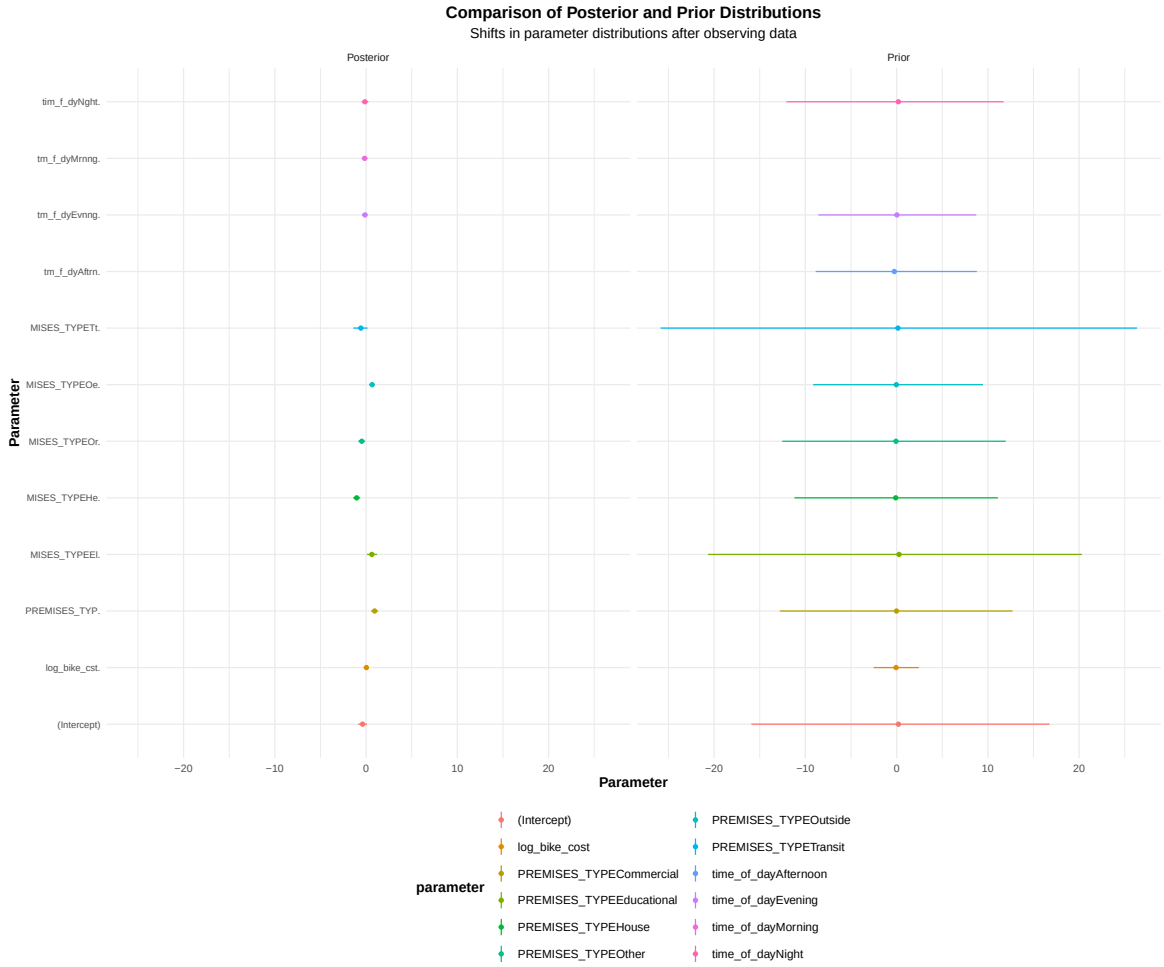


Figure 7: Comparison of Posterior and Prior Distributions

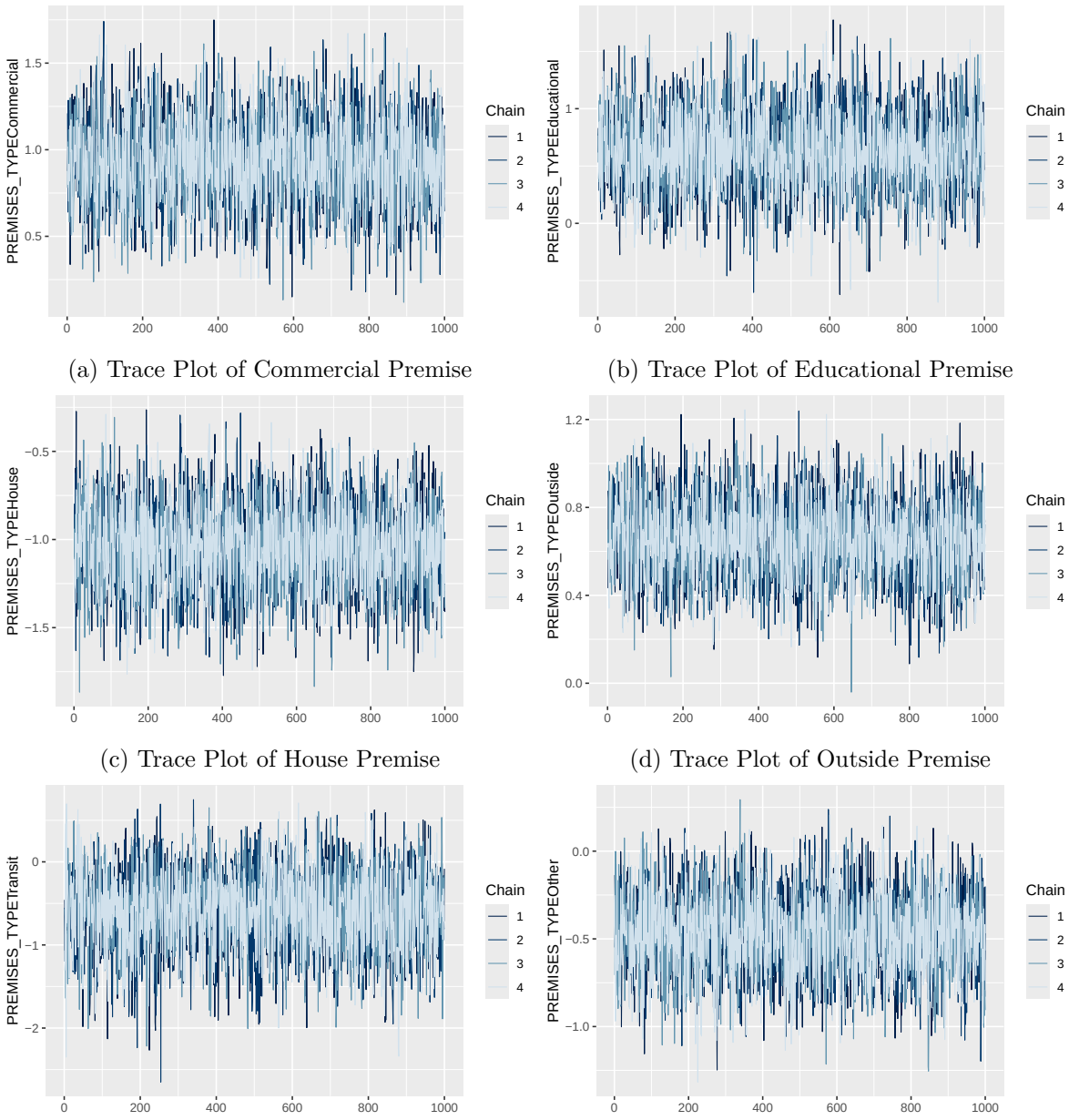
C.2 Markov chain Monte Carlo Convergence Check

Figures Figure 8 and Figure 9 display the trace plots for our Bayesian logistic regression model, providing a detailed assessment of the convergence of the Markov Chain Monte Carlo (MCMC) sampling process. These plots visualize the parameter chains across iterations for key predictors selected based on their relevance to theft patterns. Specifically, parameters like **Intercept**, **Log Bike Cost**, and representative **Premises Type** are shown due to their strong potential to influence theft likelihood. The trace plot reveal that the chains oscillate horizontally, overlap between chains, and lack any noticeable divergence, indicating stable sampling and effective convergence. This stability ensures that the model provides reliable estimates for analyzing theft risk factors in high-risk neighborhoods.

Figure Figure 10 presents the Rhat diagnostic for the model parameters, a key measure of convergence in MCMC sampling. The Rhat statistic compares the variability within each chain to the variability between chains, providing insight into whether the chains have mixed well. In this model, all Rhat values are very close to 1 and remain well below the threshold of 1.05, indicating that the chains have converged effectively. This result reinforces the reliability of the posterior estimates and confirms that the sampling process has stabilized, ensuring the model's robustness for inference.

C.2.1 90% Credibility Interval

Figure Figure 11 displays the 90% credible intervals for the predictors in the Bayesian logistic regression model. These intervals illustrate the uncertainty surrounding each parameter estimate, allowing us to assess their significance and direction of effect. Parameters whose intervals do not cross zero are considered statistically significant, as they demonstrate a consistent association with the likelihood of bike theft in high-risk neighborhoods. For example, certain premises types show strong, consistent effects, while others, such as specific time-of-day variables, have wider intervals reflecting greater uncertainty. This visualization provides a clear summary of the model's results and highlights the most impactful predictors of theft risk.



(a) Trace Plot of Commercial Premise

(b) Trace Plot of Educational Premise

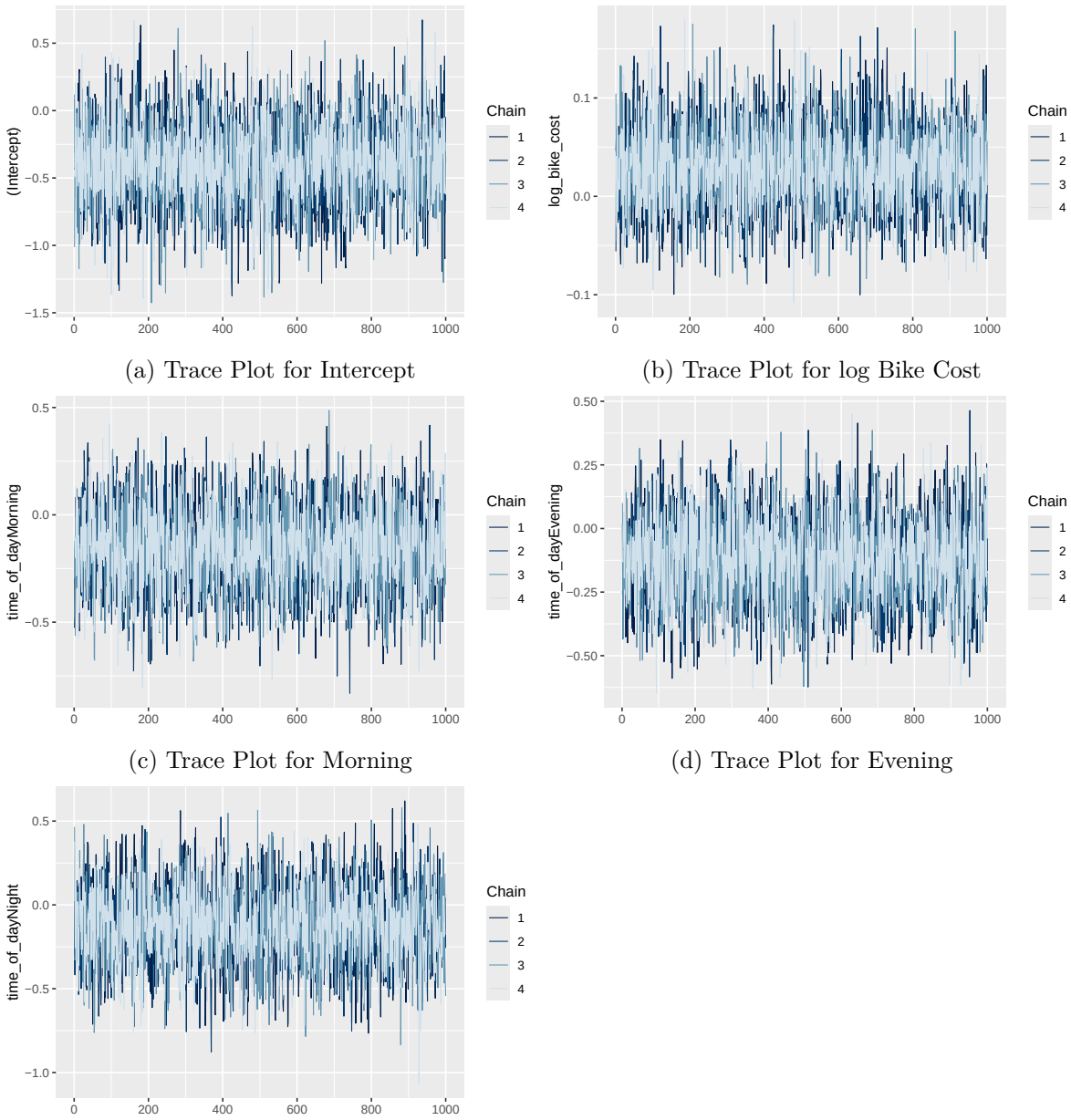
(c) Trace Plot of House Premise

(d) Trace Plot of Outside Premise

(e) Trace Plot of Transit Premise

(f) Trace Plot of Other Premise

Figure 8: Trace plot of premises type



(a) Trace Plot for Intercept

(b) Trace Plot for log Bike Cost

(c) Trace Plot for Morning

(d) Trace Plot for Evening

(e) Trace Plot for Night

Figure 9: Trace plots for time_of_day, intercept, and log_bike_cost

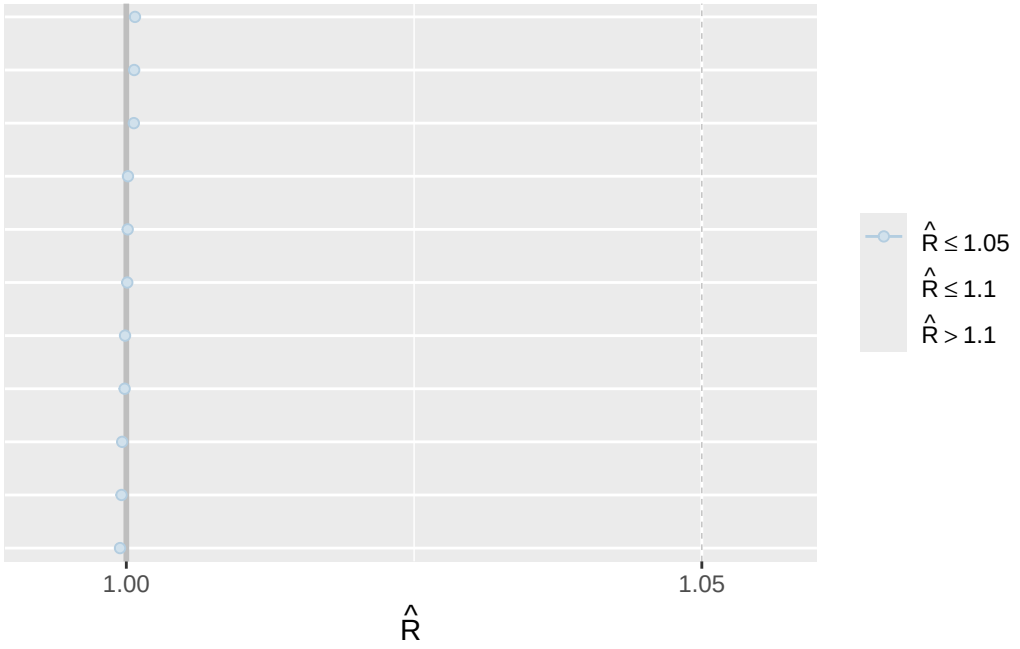


Figure 10: Rhat plot

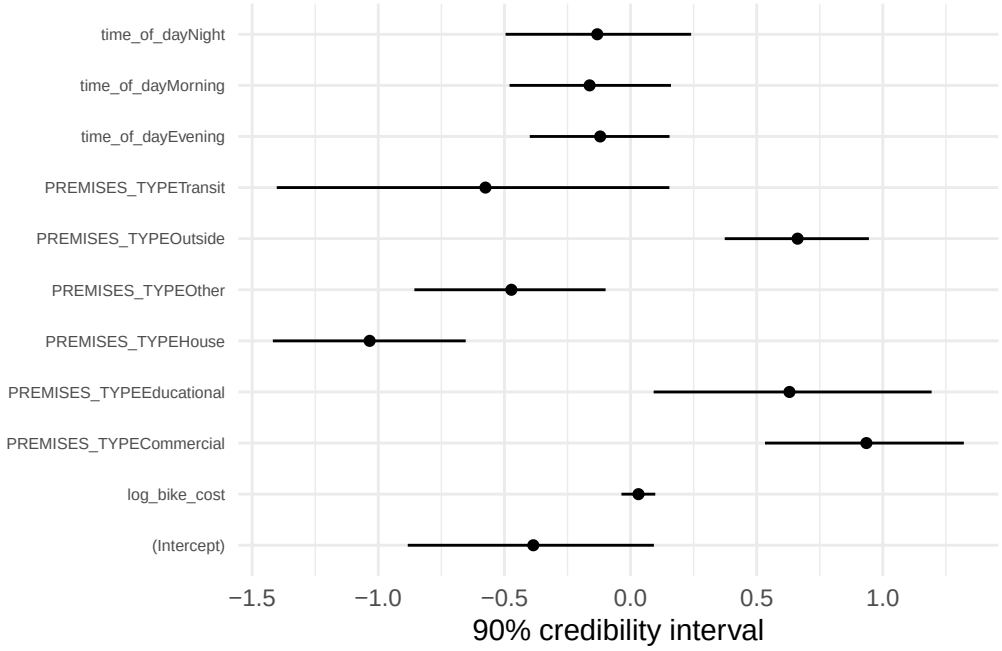


Figure 11: Credible intervals for predictors of positive poverty status

References

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Toronto Police Service. 2024. "Bicycle Thefts Open Data." Toronto, Canada. https://data.torontopolice.on.ca/datasets/TorontoPS::bicycle-thefts-open-data/about.

University of Toronto. n.d. "Introduction to GIS Using r." https://mdl.library.utoronto.ca/technology/tutorials/introduction-gis-using-r.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2023. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.