

# Changing Dynamics of Urban Bike Thefts: Understanding Trends to Inform Future Prevention Efforts\*

A Replication Quantitative Analysis Using Toronto Police Data (2014-2024)

Jin Zhang

December 2, 2024

This study utilizes a decade of Toronto Police Service data spanning from 2014 to 2024 to examine the dynamics of bicycle thefts, focusing specifically on monthly and hourly variations, as well as the influence of different locations. Analysis reveals that the theft rates are higher in summer months and during late evening hours under the apartment buildings. This research highlights critical insights into urban crime patterns, providing valuable information for enhancing security measures and planning urban safety strategies effectively.

## Table of contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Data</b>	<b>3</b>
2.1	Data Source . . . . .	3
2.2	Data Overview and Cleaning Outcomes . . . . .	4
2.3	Measurement . . . . .	5
2.4	Data visualization . . . . .	6
2.5	Outcome variables . . . . .	7
2.6	Predictor variables . . . . .	8
<b>3</b>	<b>Model</b>	<b>9</b>
3.1	Model set-up . . . . .	9
3.1.1	Model justification . . . . .	10

---

\*Code and data are available at: [https://github.com/KrystalJin1/Toronto\\_Bike\\_Theft\\_Analysis.git](https://github.com/KrystalJin1/Toronto_Bike_Theft_Analysis.git)

3.1.2	API for Predictive Analysis . . . . .	11
3.1.3	Model validation . . . . .	11
<b>4</b>	<b>Results</b>	<b>12</b>
<b>5</b>	<b>Discussion</b>	<b>13</b>
5.1	Summary of Findings . . . . .	13
5.2	Understanding About the World . . . . .	14
5.2.1	Limitations . . . . .	15
5.2.2	Future Steps . . . . .	16
	<b>Appendix</b>	<b>17</b>
<b>A</b>	<b>Idealized Methodology of Toronto Police Service</b>	<b>17</b>
A.1	Methodology Framework . . . . .	17
A.1.1	Introduction to the Bike Theft Analysis Methodology . . . . .	17
A.1.2	Survey and Sampling Methodology . . . . .	17
A.1.3	Recruitment Process . . . . .	18
A.2	Data Collection and Analysis . . . . .	19
A.2.1	Data Collection Protocol . . . . .	19
A.2.2	Statistical Analysis of Survey and Police Data . . . . .	20
A.3	Implementation and Community Engagement . . . . .	21
A.3.1	Budget Allocation and Resource Management . . . . .	21
A.3.2	Community Engagement and Feedback Integration . . . . .	22
A.4	Reporting and Conclusion . . . . .	22
A.4.1	Transparency and Reporting . . . . .	22
A.4.2	Conclusion . . . . .	22
<b>B</b>	<b>Data details</b>	<b>23</b>
<b>C</b>	<b>Model details</b>	<b>23</b>
C.1	Posterior prediction check . . . . .	23
C.2	Posterior distribution of betas . . . . .	23
C.3	Diagnostics . . . . .	23
C.3.1	Trace plot . . . . .	23
C.3.2	Rhat Plot . . . . .	23
	<b>References</b>	<b>29</b>

# 1 Introduction

Overview paragraph   google scholar 3   pdf   other   literature

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

## 2 Data

We used the R programming language (R Core Team 2023) to analyze the data and plot the graphs for this paper. The folder structure for this paper follows the starter folder created by (Wickham et al. 2019a). I also referenced many of the examples and techniques provided by Telling Stories with Data (Alexander 2023), which helped me understand how to visualize the data and communicate the findings effectively. In addition, the `tidyverse` (Wickham et al. 2019b) packages are essential for simplifying data cleaning and analysis. `styler` (Müller and Walthert 2024) is used in arranging code style. I also plotted the graphs with `ggplot2` (Wickham 2016) and read the CSV file with `here` (Müller 2023). We used the `knitr` (Xie 2014) to show the table. The `rstanarm` package (Goodrich et al. 2022) allowed us to fit Bayesian regression models using Hamiltonian Monte Carlo. For improving the appearance and functionality of our tables, we employed the `kableExtra` package (Zhu 2024). Data import and processing were facilitated by the `readr` package (Wickham et al. 2023) for fast and efficient reading of tabular data, and the `arrow` package (Richardson et al. 2024) for reading and writing data in various formats including Parquet. The `modelsummary` package (Arel-Bundock 2024) provided tools to summarize and present regression model outputs in a convenient and customizable way. Visualizations of Bayesian models were created using the `bayesplot` package (Gabry et al. 2024) and `styler` (Müller and Walthert 2024) used in arranging code style. We also use the `plumber` (Allen 2024) to build the API.

### 2.1 Data Source

The dataset used for this analysis originates from the Toronto Police Service’s public data (Toronto Police Service 2024). It tries to promote transparency and allow the public to analyze and understand various types of crime in Toronto. The dataset provides comprehensive records of bike theft incidents reported across different regions of the city. It serves as a valuable resource for analyzing trends and developing preventive strategies to combat bike theft. In addition to the Toronto Police Service’s bicycle theft dataset, other Canadian cities and countries provide similar data. For example, the Ottawa Police Service offers a dataset

detailing bicycle theft occurrences from 2015 to 2020. But this dataset is limited to Ottawa and may not reflect the unique patterns and trends present in Toronto, which means there were no similar datasets that could have been used in this research.

## 2.2 Data Overview and Cleaning Outcomes

The initial raw dataset comprises 35 variables across 37,178 observations, detailing various aspects of bicycle theft incidents. This extensive collection includes data pivotal for discerning theft patterns and devising preventative strategies. Key variables were selected for their relevance and were initially presented in a format not conducive to straightforward analysis. For instance, time data was originally recorded as month names like “January,” which complicates temporal analysis due to its non-numeric format.

To refine the dataset for more effective analysis, the initial step involved a thorough data cleaning process. This process included renaming key variables to ensure clarity and consistency across the dataset, making each variable’s purpose clear and directly aligned with the objectives of the analysis. Moreover, the original premises type data was varied and broad, leading to diluted analytical outcomes when examining location-specific trends. To address this, less frequent categories such as Educational, Commercial, and Transit locations were consolidated under a singular ‘Other’ category. This regrouping allowed for a more focused analysis on prominent categories—Apartment, House, and Outdoors—each treated as distinct groups. This restructuring was crucial for achieving a more targeted and meaningful interpretation of the data, allowing for a clearer understanding of theft patterns by location type. To more clearly display and understand the analysis data, Table 4 is a preview of the cleaned data table.(see Appendix Section B)

The explanation of each cleaned variable with definition included in the following:

Table 1: Variables and their Definitions for the Bike Theft Model.

Variable	Definition
<b>Theft Status</b>	Binary variable where 1 indicates a bike was stolen and 0 indicates no theft.
<b>Occurrence Hour</b>	Hour of the day when the bike theft incident occurred, ranging from 0 to 23.
<b>Occurrence Month</b>	Month of the year when the bike theft occurred, ranging from 1 (January) to 12 (December).
<b>Premises Type</b>	Categorizes the location of the bike theft as House, Apartment, Outdoors, or Other.
<b>Bike Cost</b>	The monetary value of the bike, used as a continuous variable.

To more clearly display and understand the analysis data, Table 4 is a preview of the cleaned data table.

Table 2: Summary Statistics of Cleaned Bike Theft Data

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
Theft_Status	2	0	1.0	0.1	0.0	1.0	1.0
Occurrence_Hour	24	0	13.4	6.5	0.0	14.0	23.0
Occurrence_Month	12	0	7.1	2.5	1.0	7.0	12.0
Bike_Cost	1813	0	715.1	511.9	0.0	600.0	2330.0

The summary statistics in Table 2 show key insights about the cleaned bike theft data. The variable ‘Theft Status’ has only two unique values, with a mean of 1.0, indicating that the majority of bikes were stolen, as the variable is binary (0 or 1). ‘Occurrence Hour’ has a mean value of 13.4, suggesting that bike thefts tend to happen during afternoon hours, and it has values ranging from 0 to 23, representing all hours of the day. The ‘Occurrence Month’ variable has 12 unique values, covering the entire year, with an average value of 7, indicating more thefts occurring around mid-year. The average bike cost is around 715.1, with a wide variation from 0 to 2330, showing the diverse range of bikes affected.

## 2.3 Measurement

The measurement process for obtaining bicycle theft data involves transforming real-world events into a structured dataset, such as the Toronto Police Service’s bicycle theft records. This dataset is a product of the process by which bicycle thefts are reported to the police, and is primarily obtained through online reporting tools or direct communication with law enforcement.

The reporting process for bicycle thefts under 5,000 dollars requires the victim to provide as much detailed information as possible, including the serial number, customizations (e.g., unique colors or features), and the estimated value of the bicycle. This information is critical to ensure accuracy and traceability of individual cases in the dataset. Bicycles reported must have a value of less than 5,000 dollars or they will be reported through a different channel that differentiates between minor and major thefts.

Once a report is submitted, it is reviewed by the appropriate authorities and, if verified, the incident becomes an entry in the dataset used for analysis. Researchers can then analyze this data to identify patterns and trends, examining variables such as time of theft and location types. This methodical approach effectively links real-world theft events with their analytical representations, enabling a thorough examination of the dynamics and risk factors tied to different locations and premises within the dataset. This structured data transformation process is essential for generating insights into bicycle theft, providing a foundation for strategies to

mitigate these incidents effectively. These attributes help to understand the context of each burglary and allow for further analysis to identify burglary patterns or high-risk areas.

## 2.4 Data visualization

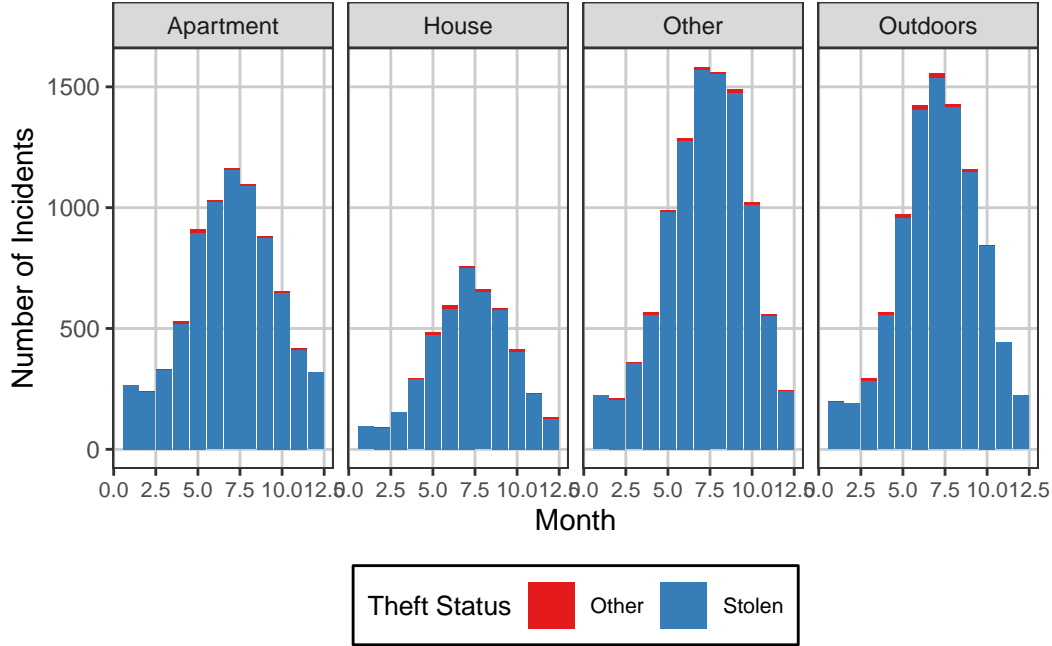


Figure 1: Seasonal Patterns of Bicycle Theft Across Different Premises Types

Figure 1 provides a detailed month-by-month breakdown of bicycle theft incidents across various premises, categorized by theft status. It reveals a clear seasonal trend, with thefts peaking during the warmer months, likely due to increased bicycle usage. The data show that a vast majority of incidents are classified as ‘Stolen’, suggesting most cases remain unresolved, with few bicycles with the status ‘unknown’ and ‘recovered’. Notably, locations categorized as ‘Outdoors’ and ‘Other’ record higher theft frequencies, indicating these are more vulnerable to theft. This analysis is important for making the effective theft prevention strategies, such as enhancing security measures and increasing surveillance during peak periods, to mitigate theft risks in identified hotspots.

Figure 2 shows the frequency of bike thefts categorized by hour of the day and delineated by theft status. It uses two colors to represent different statuses: blue for confirmed stolen bikes (Theft Status = 1) and red for bikes that were either recovered or not confirmed as stolen (Theft Status = 0). The data show a notable increase in theft occurrences during the late evening, peaking between 18:00 and 23:00, which underscores a heightened risk of theft during these hours. Conversely, the early morning hours from midnight onwards display the

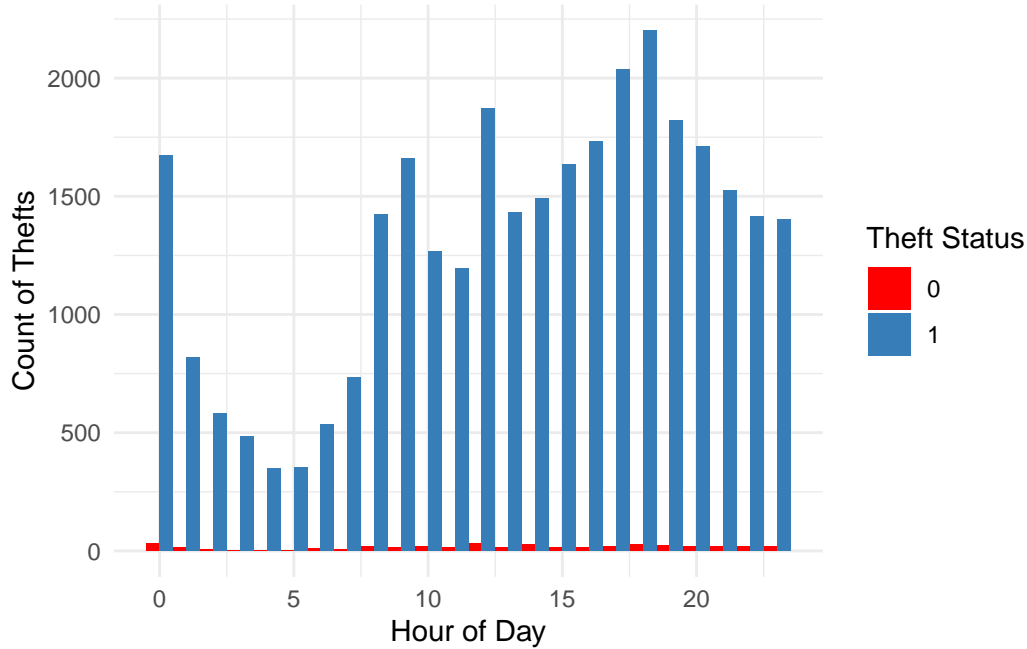


Figure 2: Time-Specific Patterns in Bike Thefts: Hourly Distribution Analysis

lowest frequencies of theft, indicating minimal activity. This pattern suggests that preventative measures, such as enhanced security or surveillance, could be most effectively targeted during the evening hours to mitigate the risk of bike theft.

Figure 3 illustrates the frequency of bicycle thefts at various price points. The histogram shows a prominent peak in thefts for bikes priced between \$200 and \$800, highlighting that moderately priced bicycles are the most frequent targets. This is potentially due to their prevalence and easier resale value compared to high-priced bicycles, which, although targeted less frequently, still represent a significant portion of thefts. The chart provides insight into thief preferences and suggests that increased security measures may be particularly necessary for bikes within this price range to mitigate theft risks.

## 2.5 Outcome variables

The outcome variable for the model is theft status, a binary indicator where 1 signifies that a bike theft occurred and 0 indicates no theft. This variable is essential as it directly represents the event the model tries to predict. By defining the theft status in binary terms, the model can apply logistic regression techniques to estimate the probability of theft based on various conditions and timings, which is vital for understanding patterns and implementing preventive measures.

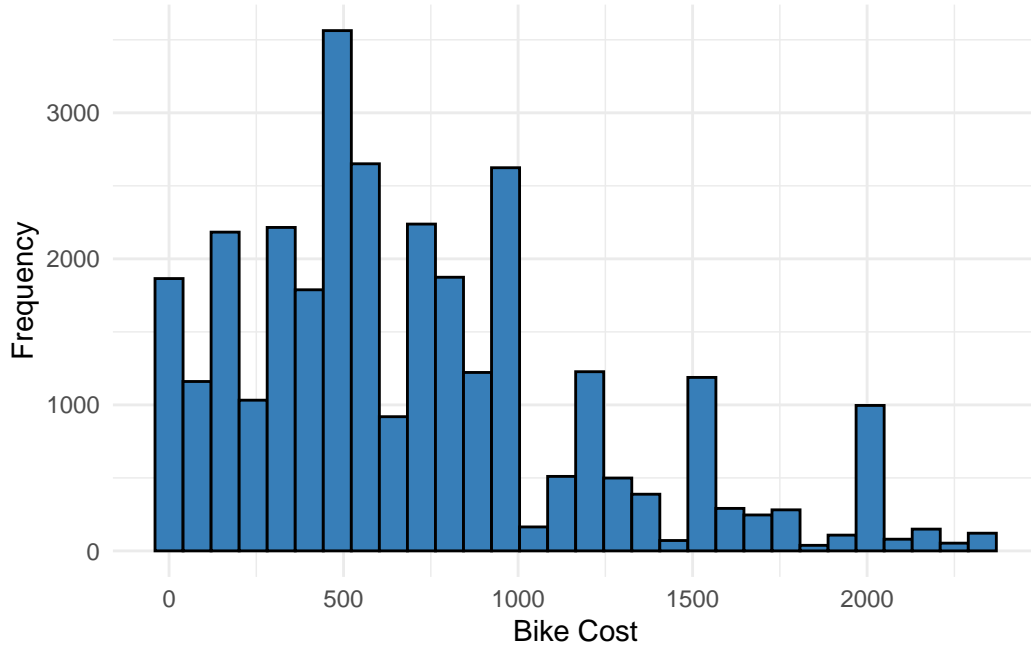


Figure 3: Distribution of Stolen Bicycle Costs

## 2.6 Predictor variables

The predictor variables are carefully chosen to capture the diverse factors that could influence the likelihood of bike theft:

The predictor variables for this analysis are thoughtfully selected to address various factors that might influence bike theft probability. ‘Occurrence Hour’ tracks the time of day the theft occurs, ranging from 0 to 23, to identify potential high-risk hours often associated with decreased public visibility and peak criminal activities. ‘Occurrence Month’ captures the month of the theft from January to December, reflecting seasonal variations that may see increases during warmer months when bicycles are used more frequently. ‘Premises Type’ is divided into “House,” “Apartment,” “Outdoors,” and “Other.” This categorization helps determine how different locations influence theft risks, with residential areas like houses and apartments generally being more secure, while public spaces categorized under “Outdoors” are more vulnerable. “Other” includes various non-standard locations such as educational and commercial areas, each carrying distinct security challenges. ‘Bike Cost’, as a continuous variable, represents the value of the bike, positing that more expensive bikes might be targeted more frequently due to their higher resale value. This comprehensive approach aims to uncover patterns and insights to help develop effective theft prevention strategies tailored to specific environmental and temporal factors.



### 3 Model

The goal of the model is to determine how different factors impact the probability of bicycle theft while incorporating prior knowledge to improve parameter estimates, especially with limited or imbalanced data, and to provide insights into the uncertainty of these estimates. Ultimately, it goes to predict theft likelihood and inform effective interventions to reduce theft rates.

#### 3.1 Model set-up

$$y_i \mid p_i \sim \text{Bernoulli}(p_i) \tag{1}$$

$$\begin{aligned} \text{logit}(p_i) = & \beta_0 + \beta_1 \times X_{\text{Occurrence Hour},i} + \beta_2 \times X_{\text{Occurrence Month},i} \\ & + \beta_3 \times X_{\text{Premises Type},i} + \beta_4 \times X_{\text{Bike Cost},i} \end{aligned} \tag{2}$$

$$\begin{aligned} \beta_0 & \sim \text{Normal}(0, 2.5) \\ \beta_1 & \sim \text{Normal}(0, 2.5) \\ \beta_2 & \sim \text{Normal}(0, 2.5) \\ \beta_3 & \sim \text{Normal}(0, 2.5) \\ \beta_4 & \sim \text{Normal}(0, 2.5) \end{aligned} \tag{1}$$

The Bayesian logistic regression model predicts the likelihood of a bike theft ( $y_i$ ), where ( $y_i = 1$ ) indicates a theft and ( $y_i = 0$ ) otherwise. The model includes several predictors:  $X_{\text{Occurrence Hour},i}$  captures the hour of the day (0–23) to identify temporal patterns in theft likelihood, and  $X_{\text{Occurrence Month},i}$  represents the month (1–12) to account for seasonal variations.  $X_{\text{Premises Type},i}$  categorizes the theft locations into “Apartment” (the baseline category), “House”, “Outdoors”, and “Other”, examining how location influences theft probability. Additionally,  $X_{\text{Bike Cost},i}$  is a continuous variable representing the bike’s reported value, hypothesizing that higher values may increase theft likelihood due to perceived value.

Specifically,  $\beta_1$  relates to  $X_{\text{Occurrence Hour},i}$ , exploring how different hours affect the log-odds of theft, while  $\beta_2$  for  $X_{\text{Occurrence Month},i}$  examines how theft likelihood fluctuates monthly. The coefficients  $\beta_3$  for  $X_{\text{Premises Type},i}$  show the variation in theft log-odds for “House”, “Outdoors”, and “Other” compared to the baseline “Apartment”. For example, a positive  $\beta_3$  for “House” would imply higher log-odds of theft in house settings compared to apartments. Lastly,  $\beta_4$  for  $X_{\text{Bike Cost},i}$  quantifies how the bike’s cost impacts theft risk, suggesting that more expensive bikes are more likely targets.

These  $\beta$  coefficients are crucial for understanding how variables such as  $X_{\text{Occurrence Hour},i}$ ,  $X_{\text{Occurrence Month},i}$ , and  $X_{\text{Premises Type},i}$  influence the likelihood of a reported bike theft. Analyzing these relationships helps pinpoint specific patterns in theft occurrences, revealing whether certain times of day, types of premises, or bike cost ranges significantly alter theft risk.

We run the model in R (R Core Team 2023) using the `rstanarm` package of (`rstanarm?`). We use the default priors from `rstanarm`.

### 3.1.1 Model justification

The Bayesian logistic regression model was chosen based on prior knowledge about factors influencing bike theft. For instance, premises type and residential status are expected to have different theft risks based on environmental characteristics, while bike cost reflects the value of the stolen item, which might influence the likelihood of theft. Bayesian regression allows us to incorporate this prior understanding through weakly informative priors, enhancing the model’s ability to capture these relationships. Additionally, variables like occurrence hour and premises type may exhibit complex patterns that the Bayesian approach can better account for by combining prior knowledge with observed data. The model provides not only point estimates but also posterior distributions, enabling a clearer understanding of uncertainty in each predictor’s impact. But the model assumes a linear relationship between predictors and the log-odds of theft, as well as independence between theft events. However, these assumptions may be limiting. For example, if theft events are highly correlated spatially or temporally—such as clusters of thefts occurring in the same area or during specific times—the model may not accurately capture these patterns.

Model comparison indicates that both the Bayesian logistic regression model and the frequentist logistic regression model achieved identical accuracy scores of 98.57% on the test dataset. This high accuracy suggests that both models effectively captured the relationship between the predictor variables (e.g., `Occurrence_Hour`, `Occurrence_Month`, `Premises_Type`, `Bike_Cost`) and the binary outcome (`Theft_Status`). The process of comparing these models involved splitting the dataset into training (80%) and testing (20%) subsets to evaluate their predictive performance on unseen data. Both models were trained on the training set, with their predictions compared against the actual values in the test set. Classifications were made using a threshold of 0.5 for predicted probabilities, and accuracy was computed as the proportion of correctly classified cases. These results demonstrate that both modeling approaches are highly effective for this particular dataset, yielding near-identical predictive performance.

While both the Bayesian and frequentist logistic regression models achieved identical accuracy scores, the Bayesian approach offers several advantages that make it particularly valuable. Bayesian models provide a probabilistic framework that incorporates prior knowledge into the analysis, allowing for more informed and interpretable results. For example, priors can regularize estimates, particularly in cases where the dataset is small or certain predictors are sparse, reducing the risk of overfitting. Additionally, Bayesian models generate full posterior

distributions for parameter estimates, rather than single point estimates, which offers richer insights into the uncertainty and variability of the parameters. This is especially beneficial for decision-making processes that rely on quantifying uncertainty, as it provides credible intervals rather than just confidence intervals, which are more interpretable in probabilistic terms. Thus, even though the predictive performance of both models is comparable in this instance, the Bayesian approach offers additional interpretative benefits and robustness, particularly in scenarios with limited or uncertain data.

### 3.1.2 API for Predictive Analysis

In addition to the statistical methods applied in our study, we utilized a specifically developed API to predict bike theft incidents based on several variables including the hour of the day, month of the year, premises type, and the cost of the bike. This API, built using the `plumber` and `rstanarm` packages, leverages a Bayesian logistic regression model trained on historical data from 2014 to 2024, as sourced from the Toronto Police Service. The API's predictive functionality allows us to dynamically assess the risk of bike theft by integrating real-time data inputs, offering a practical application for our research findings.

The API processes input variables such as the occurrence hour and month, type of premises, and bike cost to estimate the probability of a theft. This is done by calculating the odds of theft using the model's posterior samples for each input factor. The resulting predictions provide insights into the patterns of theft across different times and locations, demonstrating how these factors influence bike theft risk. This predictive tool not only validates our analytical models but also serves as a bridge between theoretical research and practical, real-world application, enhancing the utility of our findings for urban safety and crime prevention strategies.

### 3.1.3 Model validation

Figure 4 shows a posterior predictive check, which is a crucial diagnostic tool in Bayesian analysis used to compare the data observed in the real world ( $y$ ) with data simulated from the model ( $y_{rep}$ ). This comparison is fundamental for assessing whether the Bayesian logistic regression model is accurately capturing the underlying patterns of theft occurrences. In this analysis, the general alignment between the observed and replicated values across most of the probability range suggests that the model adequately captures the central trend of theft occurrences. This implies that the model's assumptions and the chosen priors are appropriate for the bulk of the data, providing a reasonable foundation for making inferences or predictions based on the model. However, the graph also shows minor deviations, particularly noticeable at higher probability values. These deviations signal areas where the model could be further refined. Such discrepancies might indicate that the model does not fully account for all nuances in the data, possibly due to oversimplified assumptions or constraints that fail to capture less frequent but plausible outcomes.

Figure 9 provides a detailed visualization of the Markov Chain Monte Carlo (MCMC) sampling for various parameters in the Bayesian logistic regression model that tries to predict bike theft incidents. Each subplot specifically demonstrates the trace of sampled values across thousands of iterations for key model parameters: the intercept, hourly occurrence, monthly occurrence, premises type, and bike cost. These trace plots are essential for evaluating the convergence of the chains and the thoroughness with which the sampling process explores the posterior distributions of the parameters. The trace plots for ‘Hourly Occurrence’ and ‘Monthly Occurrence’ reveal the dynamic influence of time on the likelihood of bike thefts, showing fluctuations that suggest possible patterns or anomalies in theft occurrence throughout different hours of the day and months of the year. The plots for ‘Outdoor Premises’, ‘Residential Premises’, and ‘Bike Cost’ reflect on how each of these variables potentially alters the risk of theft, with the trace lines providing insight into the stability and significance of their effects within the model.

In Bayesian analysis, the quality of parameter estimation is gauged by the degree of overlap and the density of the trace lines across multiple chains. Here, the dense and intertwined nature of the traces across all parameters indicates effective mixing and suggests that the chains have reached equilibrium. This supports the model’s statistical reliability and the robustness of the inferences that can be drawn about the factors influencing bike theft. Moreover, the visual consistency observed across different chains in each plot confirms that the posterior estimates are not sensitive to the initial values of the chains, further validating the model’s findings.

Figure 10 displays the diagnostic, which assesses the convergence of Markov Chain Monte Carlo (MCMC) simulations for each parameter in the Bayesian logistic regression model. All  $\hat{R}$  values are clustered around 1.00 and well below the critical thresholds of 1.05 or 1.1, indicating strong convergence across all parameters. This suggests that the multiple chains have mixed well and the posterior distributions are reliably estimated. As a result, the model’s output can be trusted for further inference and analysis.

Here we briefly describe the evidence and the detailed graph and plot for model validation and inspection, which are included in Appendix C.

## 4 Results

Table 3 has provided the following results for the predictors of bike theft: The odds of a bike theft occurring increase slightly with each additional hour of the day, as evidenced by the odds ratio (OR) for Hour of Occurrence at 1.02 with a 95% credible interval ranging from 1.00 to 1.04. Similarly, the odds ratio for Month of Occurrence is 1.01, with its credible interval spanning from 0.97 to 1.06, suggesting only a marginal change in theft likelihood across months.

For location-based predictors, thefts are significantly more likely to occur at Outdoor Premises with an OR of 1.51, and a credible interval from 1.01 to 2.26, indicating a clear increase in

risk compared to other premises. Conversely, Residential Premises show an OR of 0.96, with a credible interval of 0.71 to 1.31, suggesting a slightly lower likelihood of thefts at these locations relative to others. The influence of Bike Cost per \$1000 increase on theft likelihood is relatively neutral, with an OR nearly at 1 (0.99) and a credible interval between 0.78 and 1.27, indicating that the bike’s cost does not significantly alter the odds of it being stolen.

These results highlight the importance of the theft location, particularly outdoor settings, in predicting bike theft occurrences, while temporal factors such as the hour and month of the incident, and the cost of the bike show minimal to no significant impact on the likelihood of theft.

Table 3: Odds Ratios and 95% Credible Intervals for Bike Theft Model Predictors, with Bike Cost Adjusted per \$1000 Increase

Variable	Odds Ratio	Lower 95% CI	Upper 95% CI
Hour of Occurrence	1.01	1.00	1.03
Month of Occurrence	1.06	1.02	1.10
House Premises	0.51	0.38	0.71
Other Premises	0.96	0.71	1.28
Outdoor Premises	0.68	0.51	0.90
Bike Cost per \$1000	0.89	0.74	1.08

Figure 5 shows a side-by-side comparison of prior and posterior distributions for each parameter in our Bayesian logistic regression model. This helps us see how initial assumptions (priors) change when we use real data (posteriors). For example, the shifts from priors to posteriors for parameters like the Occurrence Hour and Bike Cost show that the model makes significant adjustments based on the data it receives. These shifts are important because they tell us that the model is effectively updating its predictions to better match what’s actually happening. Parameters like Occurrence Month and different premises types—like House, Other, and Outdoor—show varying levels of change, indicating their different impacts on the likelihood of bike theft. It shows that the model is capable of learning from the data, making it a reliable tool for understanding the factors that influence bike theft. This helps us trust the model more and use it to make better decisions or further analyses.

## 5 Discussion

### 5.1 Summary of Findings

In this paper, we conducted an in-depth analysis of bike theft data from Toronto using a Bayesian logistic regression model. The primary goal was to understand which factors most

significantly affect the likelihood of a bike being stolen. To achieve this, we utilized data provided by the Toronto Police Service, focusing on several key variables such as the time of day, month, type of location (house, apartment, outdoors, or other), and the cost of the bike. The initial step involved cleaning and structuring the raw data to ensure it was suitable for analysis. We then built a Bayesian logistic regression model to estimate the probability of theft based on these different factors. This approach enabled us to uncover specific trends and identify high-risk areas for bike theft, providing valuable insights into the conditions under which thefts are most likely to occur. By understanding these patterns, we can propose targeted interventions to reduce bike thefts in the city.

Our data cleaning process involved dealing with missing values, reformatting categorical data, and aggregating certain location types to make the analysis more manageable and meaningful. We consolidated categories such as educational, commercial, and transit locations into a broader ‘Other’ category to ensure the model could focus on the primary types of locations where bikes are stored. This process was crucial for making the dataset more practical for analysis and helped improve the reliability of the conclusions drawn from the model. The Bayesian approach allowed us to incorporate prior knowledge into the analysis, which was particularly useful given the limitations in the dataset, such as the imbalance between theft and non-theft cases. Overall, our analysis provided a structured method to determine the impact of different environmental and temporal factors on bike theft.

## 5.2 Understanding About the World

One of the key insights from this paper is that the type of location has a significant impact on the likelihood of bike theft. Our results showed that bikes parked outdoors are at a much higher risk of being stolen compared to those parked in residential areas like houses or apartments. This finding highlights the importance of location security in preventing bike theft. Public spaces, such as streets and parks, often lack adequate security measures, making them attractive targets for thieves. Therefore, to reduce the incidence of bike thefts, it is crucial to improve security measures in these outdoor areas. This could include installing more surveillance cameras, adding secure bike racks, and promoting the use of locks that are more difficult to break. Engaging local businesses and municipalities to create safer parking environments for cyclists could also play an important role in addressing this issue.

Another important insight from our analysis is the role of timing in bike theft. The data showed that bike thefts are more likely to occur in the afternoon and evening. This suggests that the risk of theft is higher during periods when people are more active and bikes are more likely to be left unattended for short periods, such as during errands or social visits. This insight is valuable because it helps to determine when theft prevention measures would be most effective. For example, increasing police patrols or community surveillance during these high-risk hours could help deter potential thieves. Additionally, public awareness campaigns could be launched to educate cyclists about the higher risk during these times, encouraging them to take extra precautions such as using better locks or parking in more secure locations.

Understanding the temporal patterns of theft can also help city planners and law enforcement allocate resources more effectively to mitigate theft risks.

The analysis also revealed that the cost of the bike is a relevant factor, though its effect is more nuanced. While expensive bikes are valuable targets, mid-range bikes are also frequently stolen, possibly because they are easier to resell and less likely to have advanced security features. This suggests that theft prevention strategies should not only focus on high-end bikes but also on making security accessible and practical for all cyclists, regardless of the bike's value. Encouraging the use of multiple types of locks, offering secure parking areas, and subsidizing bike registration services are some of the strategies that could be employed to protect bikes across different price ranges.

### **5.2.1 Limitations**

While our analysis provides valuable insights, it is not without its limitations. One major weakness is the assumption that the relationship between predictors and the likelihood of theft is linear. This assumption may not fully capture the complexity of real-world factors that influence bike theft. In reality, there could be non-linear relationships or interactions between variables that our model does not account for. For example, the effect of bike cost might vary depending on the time of day or the type of location, indicating an interaction effect that our linear model cannot capture. Future studies could explore more sophisticated models, such as non-linear regression or machine learning approaches, to better capture these complex relationships.

Another limitation is that our dataset only includes reported bike thefts. This means that unreported cases, which could constitute a significant portion of all thefts, were not included in the analysis. The decision to report a theft may depend on various factors, such as the value of the bike, the perceived likelihood of recovery, or personal experiences with law enforcement. This reporting bias could lead to an incomplete picture of bike theft in Toronto, potentially skewing our findings. For example, lower-value bikes might be underrepresented in the data if owners feel it is not worth reporting their theft, even though these bikes are frequently targeted by thieves. Addressing this issue would require additional data collection efforts, such as surveys to estimate the number of unreported thefts.

Lastly, the model assumes independence between theft events, which may not always be the case. There could be clusters of thefts in certain areas or during specific times, suggesting that one theft might increase the likelihood of another nearby. This spatial and temporal correlation is not captured in our current model, which could lead to an underestimation of theft risk in certain hotspots. Including spatial analysis or temporal clustering in future models could provide a more comprehensive understanding of theft patterns and help in developing more targeted prevention strategies.

### 5.2.2 Future Steps

There are still many questions left unanswered by our analysis, and there are several directions for future research that could help address these gaps. One area for further study is the inclusion of additional variables that might influence bike theft, such as the availability and quality of bike parking, neighborhood crime rates, and socioeconomic factors. Including these variables could provide a more detailed understanding of the contextual factors that contribute to theft risk. For example, neighborhoods with higher crime rates may see more bike thefts, and understanding these dynamics could help in tailoring interventions to specific areas.

Future research could also benefit from using more sophisticated models that consider potential non-linearities and interactions between variables. Machine learning models, such as decision trees or random forests, could be used to explore complex relationships in the data that are not well captured by traditional regression models. These approaches could help identify hidden patterns and provide more accurate predictions of theft risk under different conditions. Additionally, integrating spatial data and using geographic information system (GIS) tools could help identify hotspots for bike theft and provide insights into how environmental factors, such as proximity to public transit or major roads, affect theft risk.

Another important direction for future research is to gather more data on unreported thefts. Conducting surveys or working with community organizations to estimate the number of unreported thefts could provide a more complete picture of the true scale of the problem. This data could help address the reporting bias in our current analysis and lead to more accurate conclusions about the factors that influence bike theft. Additionally, partnerships with bike shops, community groups, and law enforcement could help in collecting richer data and implementing targeted interventions based on the findings.

Finally, our study suggests several practical measures that could be implemented to reduce bike theft. These include increasing security in outdoor areas, providing better infrastructure for secure bike parking, and running awareness campaigns to educate cyclists about theft risks and prevention strategies. Collaborating with city authorities, law enforcement, and community organizations will be key to making these interventions effective. Future research could evaluate the impact of these interventions to determine which strategies are most successful in reducing bike theft and could be scaled up for broader implementation.



## **Appendix**

### **A Idealized Methodology of Toronto Police Service**

#### **A.1 Methodology Framework**

##### **A.1.1 Introduction to the Bike Theft Analysis Methodology**

The Toronto Police Service (TPS) has developed a comprehensive methodology to analyze bike thefts, grounded in rigorous data collection, community engagement, and evidence-based practices. The approach aims to identify patterns, improve preventive measures, and foster community trust through transparency and collaboration. By combining statistical rigor with public feedback, TPS ensures that its strategy for addressing bike theft is not only data-driven but also reflective of the real-world experiences of Toronto residents.

##### **A.1.2 Survey and Sampling Methodology**

At the core of TPS's methodology is a meticulously crafted survey designed to gather detailed insights into the experiences, perceptions, and demographic factors related to bike theft within Toronto's community. The survey is structured to capture a diverse range of information, including demographic data (e.g., age, gender, income level, and ethnicity), theft-specific details (e.g., theft locations, times, and preventive measures used), and perceptions of the effectiveness of police interventions. This rich dataset allows TPS to identify the groups most affected by bike theft, assess the effectiveness of existing strategies, and pinpoint areas for targeted improvements in theft prevention and response measures.

To ensure the survey results are representative of Toronto's diverse population, TPS employs a stratified random sampling methodology. This approach involves segmenting the population into distinct strata based on key variables, such as geographic location (e.g., neighborhoods or police precincts), demographic characteristics (e.g., age brackets, income levels, or primary language spoken), and exposure to bike theft (e.g., prior victims versus non-victims). The stratification process is informed by both census data and historical bike theft reports to ensure the sample captures the complexity of Toronto's demographic and geographic landscape. For instance, neighborhoods with higher recorded theft rates, such as areas near transit hubs or downtown districts, are proportionally represented in the sample to ensure their unique conditions and challenges are adequately reflected.

Once the strata are defined, participants are randomly selected within each group, maintaining the proportional representation of each stratum relative to its size in the overall population. Randomization is conducted using computer-generated random numbers to eliminate selection bias and ensure methodological rigor. For smaller or underrepresented strata, such as residents in lower-income neighborhoods or recent immigrants, oversampling is implemented to

ensure their experiences and perspectives are adequately captured. Oversampling adjusts for potential non-response or underrepresentation, ensuring these groups have sufficient weight in the analysis without distorting the overall population characteristics.

To further enhance the validity and reliability of the results, TPS incorporates pre-survey weighting adjustments to align the sample distribution with known population benchmarks from census and administrative data. Post-survey weighting is then applied to account for response rate variations across strata, ensuring the final dataset accurately mirrors Toronto's population. For example, if a particular age group or geographic region exhibits lower response rates, their survey responses are weighted more heavily during analysis to maintain proportionality.

The survey instrument itself is designed following best practices in survey methodology. Questions are crafted to minimize respondent fatigue and bias, employing a mix of closed-ended (e.g., Likert scales, multiple choice) and open-ended formats to capture both quantitative data and qualitative insights. Pilot testing is conducted with a small, diverse sample of participants to evaluate the clarity, relevance, and comprehensiveness of the survey questions. Feedback from the pilot test informs revisions to the survey, ensuring that it captures the nuances of community experiences while remaining accessible and comprehensible to a wide range of respondents.

To ensure inclusivity, the survey is translated into multiple languages commonly spoken in Toronto, such as French, Chinese, Tamil, and Punjabi, enabling participation from non-English-speaking residents. Both online and in-person administration methods are used to further broaden accessibility. Online surveys are distributed via secure platforms, while in-person surveys are conducted by trained personnel in high-theft neighborhoods and community centers, ensuring participation from individuals who may lack internet access.

This rigorous approach to survey design and sampling methodology ensures that the data collected is robust, representative, and methodologically sound, providing a reliable foundation for TPS's analysis of bike theft and the development of targeted, data-driven interventions.

### **A.1.3 Recruitment Process**

Recruiting participants for the survey involves a combination of online and in-person strategies to reach a wide audience. TPS uses its existing communication channels, such as email newsletters and social media platforms, to inform the public about the survey and encourage participation. Flyers are distributed at bike shops, community centers, and public libraries to reach individuals who may not be actively engaged online. Partnerships with local cycling advocacy groups further enhance recruitment efforts by leveraging their networks to connect with frequent bike users.

To incentivize participation, TPS offers small rewards, such as \$10 gift cards for bike accessories or discounts at local bike shops. Additionally, participants are entered into a draw for larger

prizes, such as premium bike locks or helmets. This strategy not only boosts participation rates but also demonstrates TPS’s commitment to supporting the cycling community.

## **A.2 Data Collection and Analysis**

### **A.2.1 Data Collection Protocol**

The data collection process employed by TPS is meticulously designed to prioritize accuracy, inclusivity, and reliability, ensuring the data reflects the diverse experiences and perspectives of Toronto’s population. A mixed-method approach is utilized to accommodate varying preferences and technological access among participants. Online surveys are hosted on secure and user-friendly platforms, such as Qualtrics or SurveyMonkey, enabling respondents to complete the survey at their convenience using computers, tablets, or smartphones. Security measures, including encryption and unique access codes, are implemented to protect respondent privacy and prevent unauthorized access. Notifications and reminders are sent via email or SMS to encourage timely completion of the survey.

In-person surveys are conducted in strategically chosen locations, particularly in neighborhoods with high bike theft rates or areas identified as underrepresented in online responses. These surveys are administered by trained personnel equipped with portable devices, such as tablets or laptops, to facilitate data entry. The in-person approach ensures that individuals without reliable internet access or those who may feel more comfortable engaging face-to-face are able to participate. Survey locations include community centers, libraries, transit hubs, and bike shops, chosen for their accessibility and relevance to the target population. Efforts are made to schedule these sessions during hours that maximize reach, such as evenings or weekends, to accommodate varied work schedules.

To enhance inclusivity, TPS offers the survey in multiple languages, reflecting the linguistic diversity of Toronto. Languages such as French, Mandarin, Tamil, and Punjabi are prioritized based on demographic data and community feedback. Trained bilingual survey administrators are present during in-person sessions to assist respondents in their preferred language, ensuring that language barriers do not impede participation. This multilingual approach is complemented by culturally sensitive training for administrators, enabling them to engage effectively with diverse populations.

Throughout the data collection process, TPS emphasizes rigorous quality control measures. Responses are cross-verified against police reports and administrative records to confirm the validity of reported theft incidents. For example, if a respondent claims to have reported a theft, the corresponding police report is checked for consistency in details such as location, time, and circumstances of the theft. Any discrepancies are flagged for further investigation, and respondents may be contacted to clarify their responses if necessary.

Duplicate entries are identified and removed using advanced data processing techniques. Each respondent is assigned a unique identifier, which is combined with metadata such as timestamps

and IP addresses (for online surveys) to detect and eliminate redundant responses. For in-person surveys, administrators collect basic contact information to ensure each participant is counted only once. Surveys that are incomplete or contain implausible responses, such as contradictory answers or extreme outliers, are flagged during the initial review. These responses undergo additional scrutiny, and where appropriate, the affected entries are corrected or excluded from the final dataset.

To ensure data consistency and reduce potential biases introduced by human error, TPS employs standardized protocols for both online and in-person data collection. These protocols include detailed instructions for survey administrators, calibration of data entry tools, and predefined response validation rules embedded within the survey software. For instance, certain questions may include range checks to prevent unrealistic inputs, such as negative values for theft counts or implausible time ranges for theft occurrences.

After the initial data collection phase, a dedicated quality assurance team reviews the dataset for completeness and coherence. This includes checking for missing responses in critical fields, ensuring uniform formatting across variables, and applying statistical techniques to identify anomalies or patterns that may indicate data integrity issues. The cleaned and validated dataset is then prepared for analysis, ensuring that it provides a reliable basis for identifying trends, patterns, and actionable insights into bike theft in Toronto.

By combining advanced technology, multilingual accessibility, rigorous quality controls, and community-centric practices, TPS's data collection process sets a high standard for reliability and inclusivity. This robust methodology ensures that the data collected is both comprehensive and accurate, forming a solid foundation for evidence-based policy recommendations and community safety initiatives.

### **A.2.2 Statistical Analysis of Survey and Police Data**

Once the data is collected and validated, TPS employs a comprehensive analytical framework to extract meaningful insights into bike theft patterns. The process begins with descriptive statistics, summarizing the frequency, distribution, and scale of theft incidents across Toronto. Metrics such as theft counts by neighborhood, time of occurrence, and bike cost provide an overview of the issue, while cross-tabulations reveal relationships between variables like theft location, time of day, and premises type. For example, such analysis may show that thefts are more frequent in high-traffic areas during evening hours or that high-value bikes are disproportionately targeted.

Hot spot mapping is a key component, using geographic information systems (GIS) to identify areas with consistently high theft rates. These maps highlight priority locations for resource allocation, such as increased patrols near transit hubs or public bike racks, and can incorporate temporal data to pinpoint peak theft periods. This spatial analysis is complemented by advanced statistical modeling to identify factors that significantly influence theft likelihood. Logistic regression models, for instance, quantify how variables like location type, time of day,

and bike value impact the odds of theft, offering data-driven insights into risk factors. Interaction terms further refine the analysis, revealing nuanced relationships such as the heightened risk of outdoor thefts during late-night hours.

Temporal and predictive modeling also play a crucial role. Time-series models identify seasonal trends, such as spikes in thefts during summer, allowing TPS to anticipate and proactively address future risks. Predictive models forecast high-risk locations and times, guiding the strategic deployment of resources and preventive measures. Cluster analysis and decision trees group similar incidents and visually represent key predictors, aiding in both interpretation and communication with stakeholders.

Throughout the process, TPS ensures analytical rigor by validating model assumptions, testing sensitivity to input changes, and cross-referencing findings with historical data. The insights derived not only inform internal strategies, such as patrol scheduling and resource allocation, but also support community outreach by providing clear, actionable recommendations. This multifaceted approach ensures TPS's response to bike theft is data-driven, targeted, and effective.

## **A.3 Implementation and Community Engagement**

### **A.3.1 Budget Allocation and Resource Management**

TPS allocates its resources strategically to ensure the effectiveness of the methodology while maintaining fiscal responsibility. The survey design phase is allocated \$15,000, which includes \$5,000 for developing and translating questions into multiple languages to ensure accessibility, \$6,000 for conducting pilot tests to refine the survey instrument, and \$4,000 for printing and distribution of survey materials. An additional \$10,000 is allocated for training survey administrators, ensuring they are well-equipped to manage both in-person and online data collection efforts effectively. The purchase of equipment, such as tablets for data entry during in-person surveys, is budgeted at \$8,000. Licensing software for advanced data analysis and visualization tools, such as statistical packages and mapping software, accounts for another \$7,000.

The majority of the budget, approximately \$30,000, is dedicated to data collection efforts. This includes \$12,000 for travel expenses for survey personnel conducting in-person data collection in high-theft areas and \$10,000 for participant incentives, such as \$10 gift cards for bike accessories or discounts at local bike shops. Another \$8,000 is allocated to logistical support, including transportation of equipment and setting up survey booths at community centers, libraries, and bike shops.

Additional funds of \$15,000 are set aside for community engagement activities to ensure TPS maintains an ongoing dialogue with stakeholders. This includes \$7,000 for hosting public

meetings in high-theft neighborhoods, \$5,000 for creating accessible summary reports and visualizations of findings, and \$3,000 for communication materials, such as flyers and newsletters, to share results and gather further feedback from residents.

By clearly defining and adhering to this detailed \$70,000 budget, TPS ensures that the project remains financially sustainable while delivering high-quality, actionable results to reduce bike theft and enhance community safety.

### **A.3.2 Community Engagement and Feedback Integration**

Community engagement is a cornerstone of TPS's bike theft methodology. Throughout the process, TPS actively seeks input from residents to validate findings and refine strategies. Public meetings are held in neighborhoods with high theft rates to share preliminary results and gather feedback. These sessions provide an opportunity for residents to voice their concerns, share their experiences, and suggest solutions.

The feedback collected during these sessions is integrated into the analysis to ensure that the findings reflect the community's lived experiences. For instance, residents might highlight specific locations or times of day that are particularly prone to theft, which can then be incorporated into the hot spot mapping analysis. This iterative process strengthens the validity of the results and fosters trust between TPS and the community.

## **A.4 Reporting and Conclusion**

### **A.4.1 Transparency and Reporting**

To build trust and accountability, TPS places a strong emphasis on transparency throughout the methodology. Detailed documentation is prepared at each stage of the process, explaining how data is collected, validated, and analyzed. This documentation is made publicly available, along with summary reports that highlight key findings and actionable recommendations.

Reports are designed to be accessible and engaging, featuring clear visualizations such as maps and charts to communicate complex information effectively. These reports are shared with stakeholders, including policymakers, cycling advocacy groups, and the general public, to ensure that the findings inform decision-making and contribute to community safety initiatives.

### **A.4.2 Conclusion**

The idealized methodology of the Toronto Police Service for analyzing bike theft exemplifies a comprehensive, community-centered approach to data-driven policing. By combining rigorous statistical analysis with robust community engagement and transparent reporting,

TPS not only enhances its understanding of bike theft patterns but also builds trust with the communities it serves. This methodology serves as a model for other cities seeking to address similar challenges and demonstrates the value of integrating data, community input, and evidence-based strategies into policing efforts.

## B Data details

Table 4: Preview of Cleaned Bike Theft Data

Theft_Status	Occurrence_Hour	Occurrence_Month	Premises_Type	Bike_Cost
1	19	12	Other	1300
1	0	9	Apartment	750
1	16	12	Apartment	1500
1	10	12	Outdoors	400
1	17	12	Outdoors	500
1	12	1	Apartment	1019

## C Model details

### C.1 Posterior prediction check

### C.2 Posterior distribution of betas

### C.3 Diagnostics

#### C.3.1 Trace plot

#### C.3.2 Rhat Plot

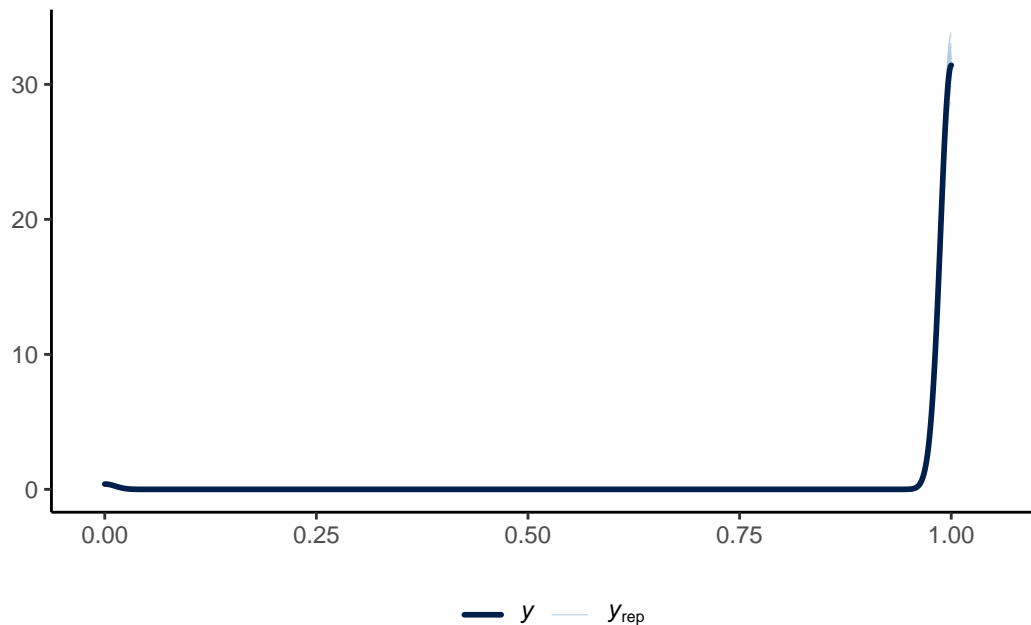


Figure 4: Posterior prediction check

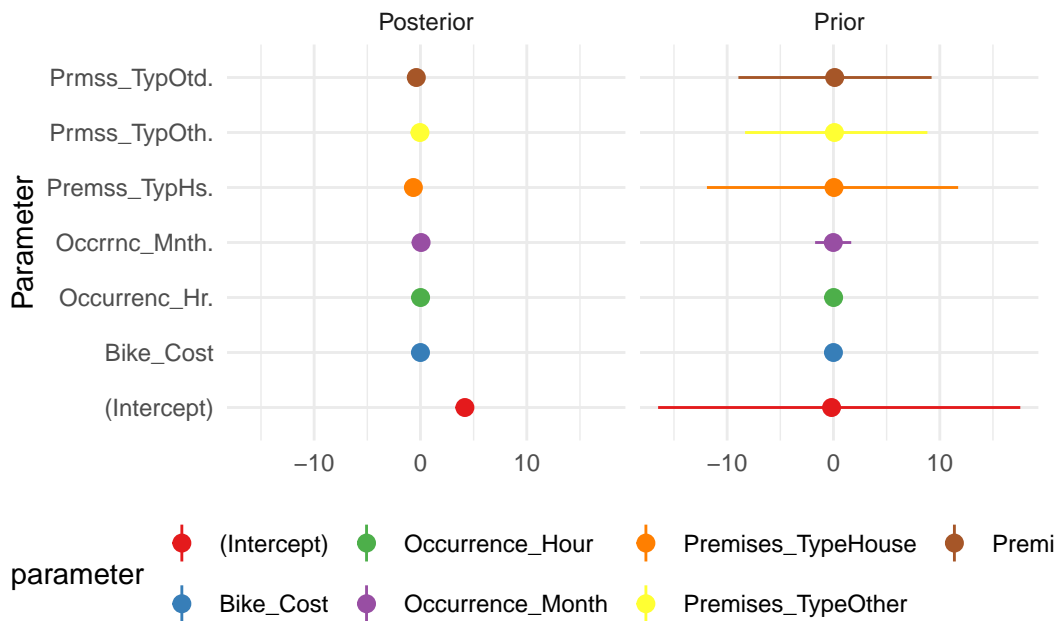


Figure 5: Comparison of Posterior and Prior Distributions for Model Parameters



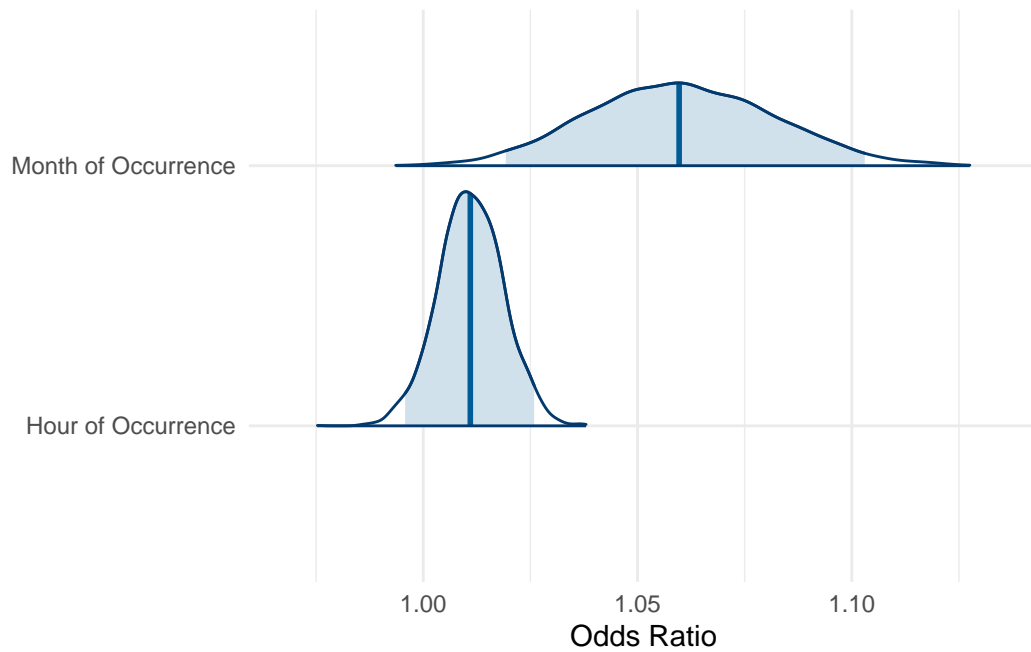


Figure 6: Posterior distribution of betas

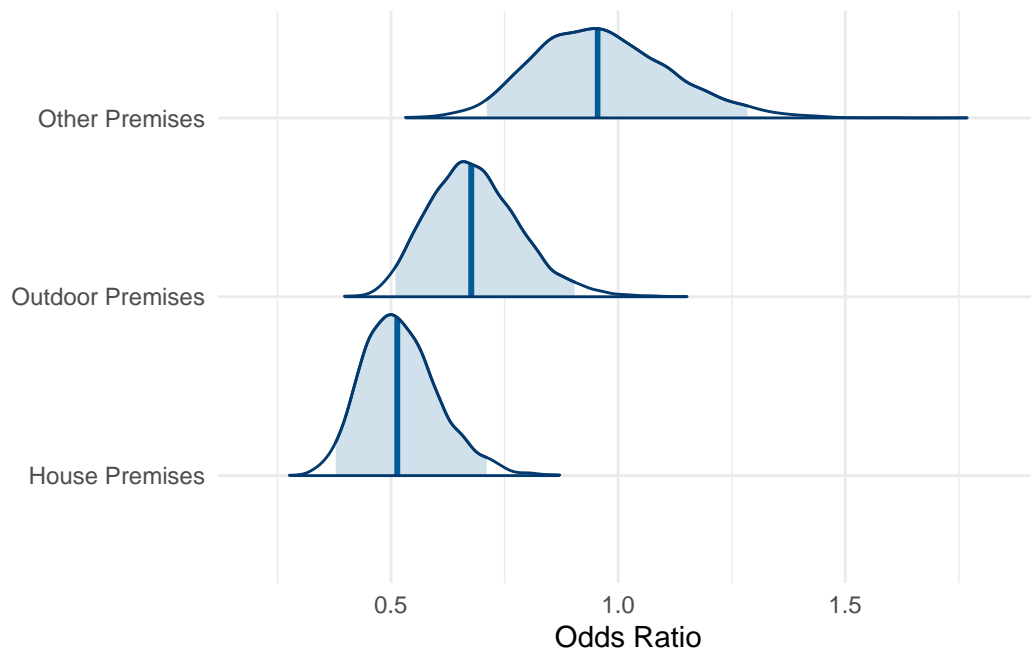


Figure 7: Posterior distribution of betas

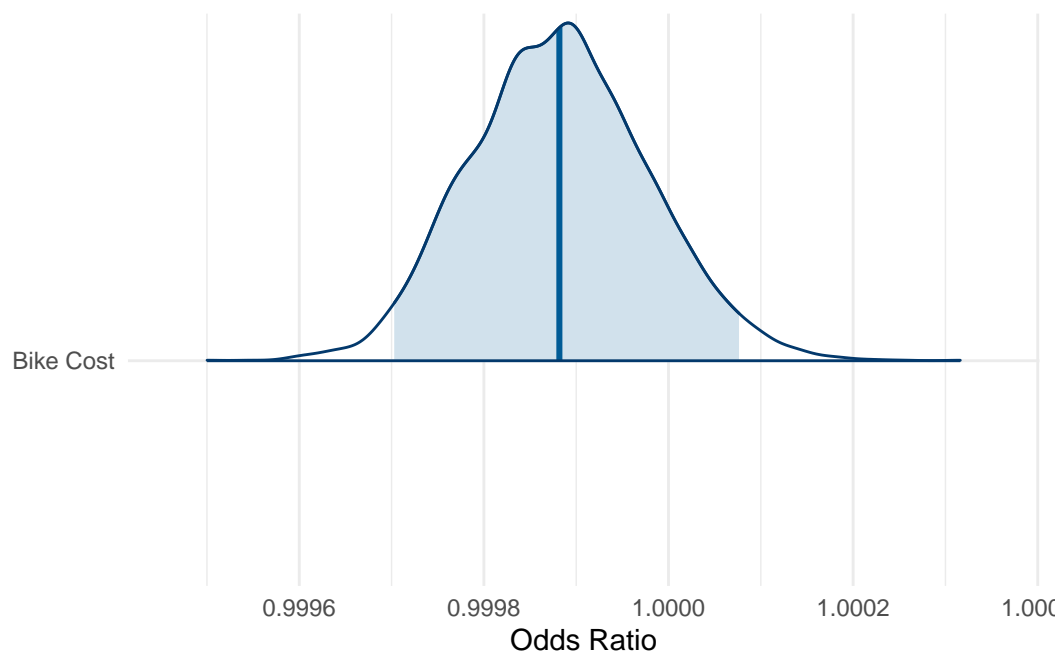


Figure 8: Posterior distribution of betas

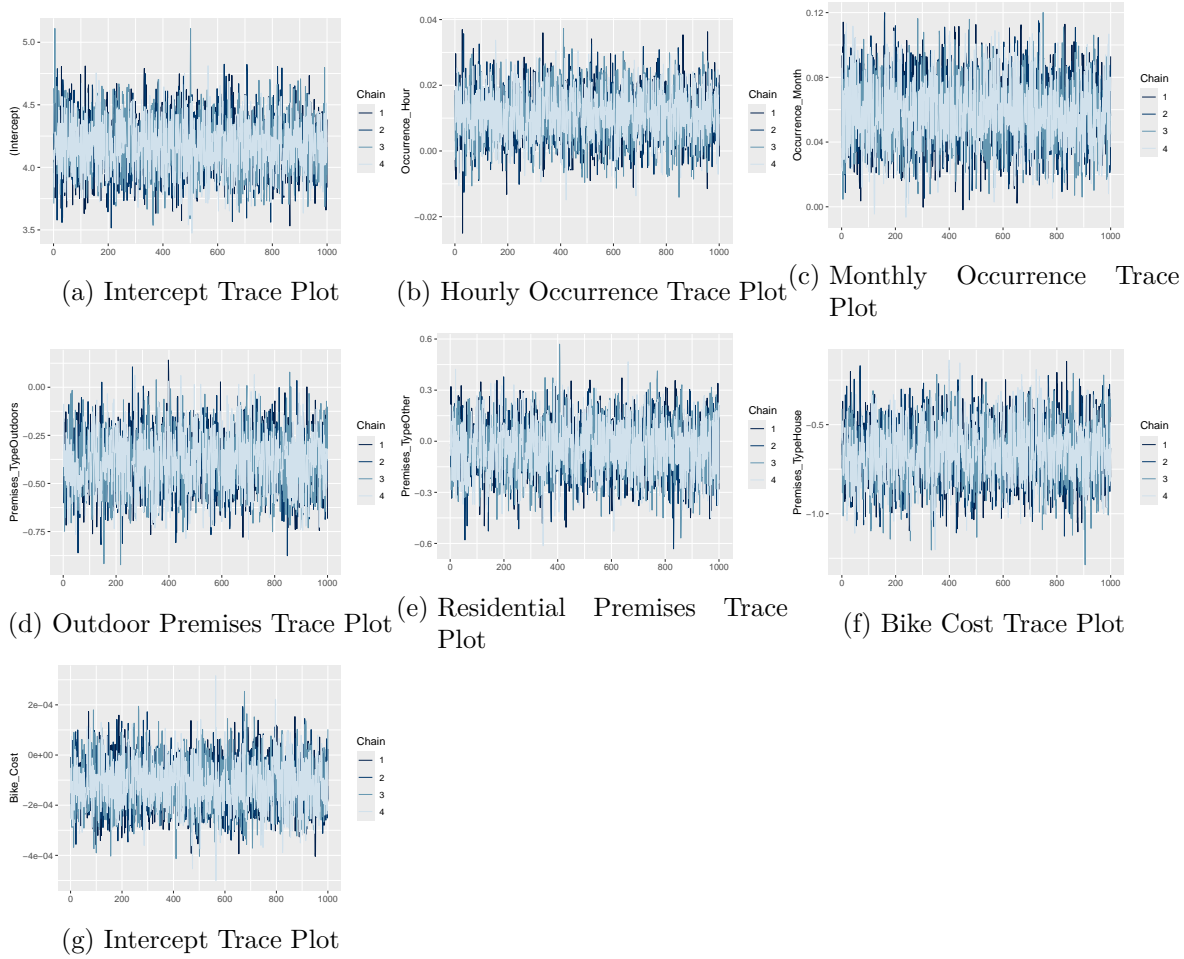


Figure 9: Trace plot of Bike Theft

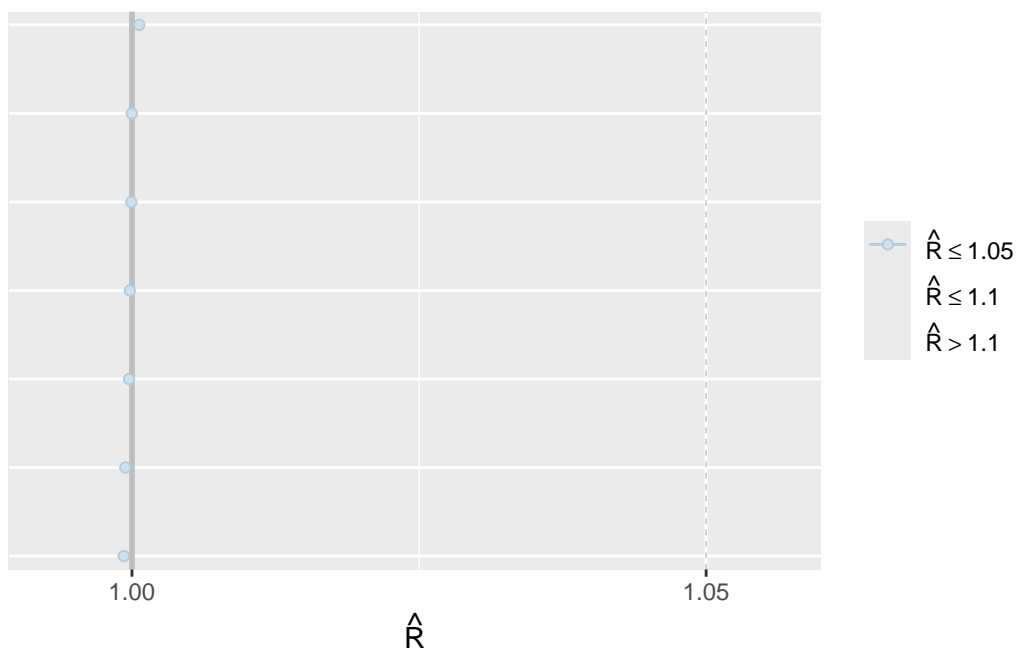


Figure 10: Rhat Plot of Bike Theft

## References

- Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC.
- Allen, Jeffrey. 2024. *Plumber: An API Generator for r*. <https://CRAN.R-project.org/package=plumber>.
- Arel-Bundock, Vincent. 2024. *modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. <https://CRAN.R-project.org/package=modelsummary>.
- Gabry, Jonah et al. 2024. *bayesplot: Plotting for Bayesian Models*. <https://CRAN.R-project.org/package=bayesplot>.
- Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. “rstanarm: Bayesian applied regression modeling via Stan.” <https://mc-stan.org/rstanarm/>.
- Müller, Kirill. 2023. *Here: A Simpler Way to Find Your Files*. <https://CRAN.R-project.org/package=here>.
- Müller, Kirill, and Lorenz Walthert. 2024. *styler: Non-Invasive Pretty Printing of R Code*. <https://CRAN.R-project.org/package=styler>.
- 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 et al. 2024. *arrow: Integration to 'Apache' 'Arrow'*. <https://CRAN.R-project.org/package=arrow>.
- Toronto Police Service. 2024. “Toronto Police Service Public Safety Data Portal.” <https://data.torontopolice.on.ca/pages/c78364ab031747359fa8afb78febdd3d>.
- Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley et al. 2023. *readr: Read Rectangular Text Data*. <https://CRAN.R-project.org/package=readr>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019a. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- , et al. 2019b. “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>.