My title*

My subtitle if needed

Jin Zhang

December 1, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/KrystalJin1/Toronto_Bike_Theft_Analysis.git

1 Introduction

Overview paragraph

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 startup folder created by (**rohan?**). It is very helpful in keeping everything organized. I also referenced many of the examples and techniques provided by Telling Stories with Data (**alexander2023telling?**), which helped me understand how to visualize the data and communicate the findings effectively. In addition, the `tidyverse` (**citetidyverse?**) packages are essential for simplifying data cleaning and analysis.`styler` (**citestyler?**) is used in arranging code style. These tools help to organize and present data efficiently. I also plotted the graphs with `ggplot2` (**citeggplot2?**) and read the CSV file with `here` (**citehere?**). We also use the `comma` function from the `scales` (**citescales?**) library to format the data, thereby avoiding scientific notation in our charts. We used the `knitr` (**citeknitr?**) to show the table.

2.1 Data Source

The dataset used for this analysis originates from the Toronto Police Service's public data (Toronto Police Service 2024). It aims 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 raw dataset provides 35 varibles with 37178 obeservations related to the incidents of bike theft, capturing various aspects of each theft that are crucial for understanding patterns and determining effective measures. The variables that I chose from the whole dataset is `STATUS`, `OCC_MONTH`, `OCC_HOUR`, `NEIGHBOURHOOD_158`, `PREMISES_TYPE`, `BIKE_COST`. However, using these variables directly makes analyzing more difficult. For example, the original data of `OCC_MONTH` was recorded as month names (in its character form, like "January"), which would complicate temporal analysis. Also, the `NEIGHBOURHOOD_158` included a large amount of neighborhoods' name, making it difficult to derive meaningful and useful insights from the neighborhood data.

After simply understand the raw data, the next step was data cleaning and to transform it into a more suitable format for analysis. The first action involved renaming key variables for better clarity and consistency, ensuring that each variable was descriptive and aligned with the analysis goals. The renaming and explanation of each varible included in the following:

- `Occurrence_Hour`(`OCC_HOUR`): The hour of the day (0-23) when the theft occurred.
- `Premises_Type`(`PREMISES_TYPE`): The type of location where the theft took place (e.g., residential, commercial).
- `Bike Cost`(`BIKE_COST`): The reported monetary value of the stolen bicycle.

Further specially state:

`OCC_MONTH` variable was originally recorded as month names (e.g., January, February). It was converted into numeric format `Occurrence_Month`, with each month represented by an integer from 1 to 12 (e.g., January = 1, February = 2). `STATUS` was converted to a binary variable named `Theft_Status` to simplify analysis by distinguishing whether a theft occurred (1) or not (0) and for further model analysis `NEIGHBOURHOOD_158` was converted to `Region`. According to the `neighbourhood_mapping`, the neighborhood data was grouped into broader regions (e.g.,

Downtown, Midtown, Scarborough) to simplify the analysis. Each neighborhood was mapped to a major region based on its geographic location.

The remaining modified variables are explained in the following:

- `Occurrence_Month`(`OCC_MONTH`): The month in which the bicycle theft occurred.
- `Theft Status`(`STATUS`): Indicates whether the reported bicycle theft occurred or other(e.g., unfound, recovery). (1 = stolen, 0 = Other).
- `Region`(`NEIGHBOURHOOD_158`): The broader geographical area where the theft occurred, grouped from specific neighborhood names (e.g., Downtown, Midtown, Scarborough).

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

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

Table 2: 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
1	0	12	Apartment	1200
1	0	7	Other	600
1	0	12	Apartment	1500
1	18	1	Apartment	200

Table 3: 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 3 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 679.9, with a wide variation from 0 to 2034, 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. Each entry in the dataset corresponds to a real-world burglary event and includes attributes such as when the burglary occurred, the type of premises (e.g., residential, commercial), and the neighborhood or district. 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 provides a month-by-month breakdown of bicycle theft incidents in different types of premises and is colored according to the status of each incident. The bar chart clearly shows

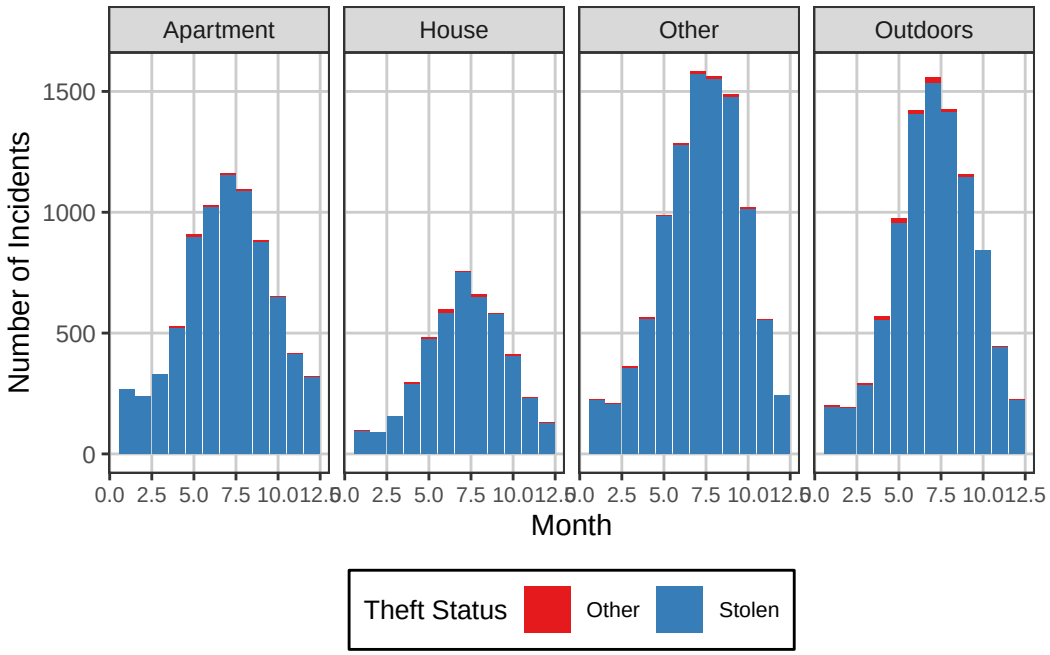


Figure 1: Seasonal Patterns of Bicycle Theft Across Different Premises Types

that thefts tend to peak in the warmer months, a trend that can be attributed to increased bicycle usage. A closer look at the chart reveals that a large percentage of cases remain unsolved (Status = Stolen), while only a small portion of cases are in unknown or recovered status (Status = Other). In addition, locations such as "Outdoors" areas and "Other" have a particularly high number of bicycle thefts, suggesting that these locations are more likely to have bicycle thefts compared to other locations.

Figure 2 presents the frequency of bike thefts by hour throughout the day, categorized by theft status. Two colors distinguish the theft statuses: teal for stolen bikes (Theft Status = 1) and red for other statuses (Theft Status = 0). This visualization shows that theft incidents peak during the late evening hours, particularly around 18:00 to 23:00, suggesting a higher risk of bike theft during these hours. The lowest incidence of thefts tends to occur in the early morning hours, from midnight to early morning.

Figure 3 shows the distribution of bike costs for bicycles reported as stolen. We can see that the majority of stolen bikes fall in the lower to mid-range price categories, with a concentration around $200 to $800. There are also a few bikes with higher values, indicating that expensive bikes, though stolen less frequently, are still targets. This distribution suggests that thieves generally target moderately priced bikes, likely because they are more common and potentially easier to steal.

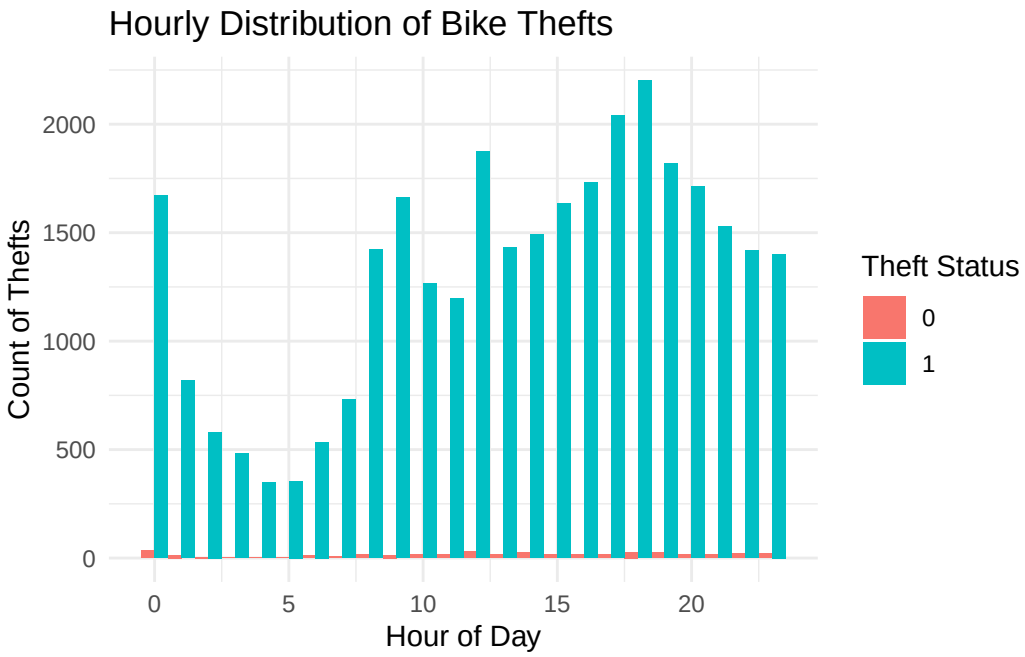


Figure 2: Hourly Distribution of Bike Thefts

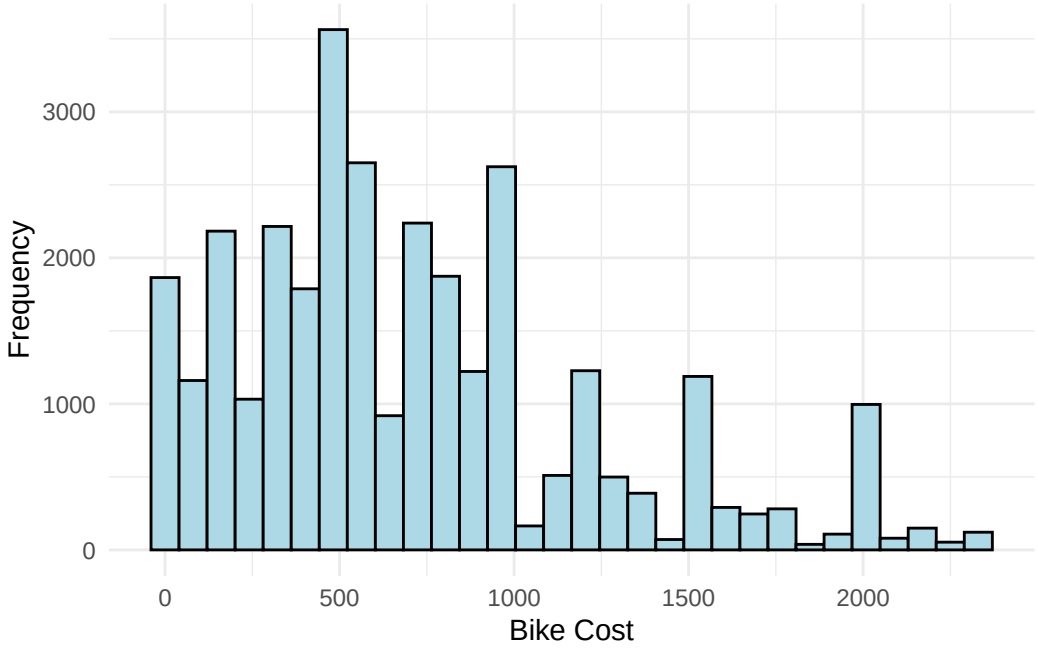


Figure 3: Distribution of Stolen Bicycle Costs

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.

2.6 Predictor variables

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

Occurrence hour captures the time of day the bike theft incident occurred, ranging from 0 to 23. Certain hours may be more prone to thefts due to lower public visibility or peak criminal activity times. Occurrence month reflects the month when the theft took place, from 1 (January) to 12 (December). Seasonal variations could affect theft rates, with possibly higher incidents in warmer months when bikes are more frequently used and accessible. Premises type categorized into "Residential," "Outdoors," and "Other," this variable helps in understanding the influence of location on theft probability. "Residential" includes places like apartments and houses, "Outdoors" covers areas such as educational, commercial, and transit, and "Other" encompasses miscellaneous or less common locations that do not fit into the first two categories. Different locations may have varying risks associated with bike theft, influenced by factors like population density, surveillance, and accessibility. Bike cost is a continuous variable represents the monetary value of the bike, under the premise that more expensive bikes might be more attractive targets for thieves due to their higher resale value.

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

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

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

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

Here's the revised description for your Bayesian logistic regression model where "Apartment" is specified as the baseline category for the premises type variable:

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, β_1 relates to $X_{\text{Occurrence Hour},i}$, exploring how different hours affect the log-odds of theft, while β_2 for $X_{\text{Occurrence Month},i}$ examines how theft likelihood fluctuates monthly. The coefficients β_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 β_3 for "House" would imply higher log-odds of theft in house settings compared to apartments. Lastly, β_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 β 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 Goodrich et al. (2022). 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.

3.1.2 Model validation

Figure 5 compares the observed data (y) with the replicated data generated by the Bayesian logistic regression model (y_{rep}). The alignment between the observed and replicated values indicates that the model reasonably captures the overall pattern of theft occurrences. However, minor deviations at lower probabilities highlight areas where the model's fit could be improved, suggesting potential limitations in capturing the little difference of the data.

Figure 9 visualizes the Markov Chain Monte Carlo (MCMC) sampling for each parameter in the Bayesian logistic regression model across four chains. It shows consistent mixing and convergence, as indicated by the overlapping and steady oscillations of the chains within a stable range. For parameters such as the coefficient of occurrence hour and occurrence month, the chains demonstrate good mixing, suggesting that the model has adequately explored the parameter space. Similarly, for the coefficients for variables "premises type: outdoors" and "premises type: residential," the chains are stable, indicating reliable sampling. The absence of divergent transitions or extreme fluctuations across all parameters confirms the MCMC algorithm's proper functioning, implying that the posterior estimates are credible and the model is well-calibrated.

?@fig-rhat 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 D.

4 Results

Table 4 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 4: 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

?@tbl-modelsummary-results provides insights into the factors influencing bike theft occurrences based on the Bayesian logistic regression model. Overall, premises type emerges as a significant determinant of theft likelihood. Bikes parked in commercial and outdoor settings

show a lower risk of theft, while educational premises indicate a potential increase in risk, though the precision of this estimate is lower. Temporal factors, such as the occurrence hour and month, have minimal impact on predicting theft occurrences. The model suggests that where a bike is parked is more crucial in determining theft risk than when it is parked.

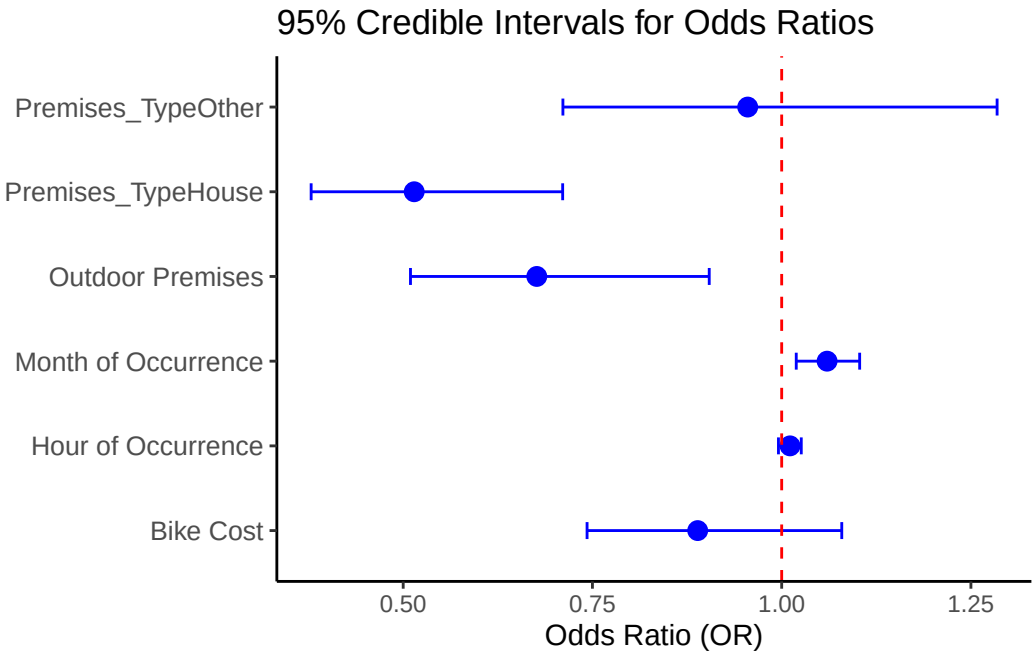


Figure 4: Credible Interval of Bike Theft

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Idealized Methodology of Totonto Police Service

The methodology employed by the Toronto Police Service (TPS) for analyzing bike theft is rooted in rigorous data collection and meticulous validation processes (Toronto Police Service 2024). Initially, data is gathered from police reports submitted by the community and officers in the field. This preliminary data, which is subject to change, captures various aspects of bike theft incidents, including location, time, and details of the theft.

To ensure the accuracy and reliability of the data, TPS employs a multi-stage validation process. This involves cross-referencing the initial reports with follow-up investigations and updates from the field. The methodology acknowledges the potential for mechanical and human errors and clarifies that the data is preliminary and not suitable for comparative historical analysis over time due to these possible updates and corrections.

Statistical methods are then applied to analyze the validated data. These methods include descriptive statistics to understand the frequency and distribution of bike thefts, and more complex analytical techniques like trend analysis and hot spot mapping to identify patterns and areas with high theft rates. This analytical approach is transparent about its limitations and openly discusses the challenges associated with data accuracy and timeliness.

The methodology also incorporates community feedback mechanisms. TPS engages with local communities to validate findings and gather additional insights through surveys and public meetings. This feedback is crucial for refining data interpretation and ensuring the analysis reflects real-world conditions and community concerns.

Moreover, TPS emphasizes transparency in its methodology. Detailed documentation is provided to the public and stakeholders, explaining each step of the data collection and analysis process. This documentation helps build trust and accountability, ensuring that the community understands how data about bike theft is handled and used for policing and safety improvements.

The survey designed by TPS to engage the community in addressing bike theft is comprehensive and structured to capture a wide range of insights. It begins with an introduction that explains the purpose of the survey and its importance in shaping police strategies and community safety measures. The survey includes detailed contact information to ensure transparency and accessibility, allowing participants to reach out with questions or for further engagement.

The questions in the survey are carefully constructed to gather detailed information about the respondents' experiences and perceptions of bike theft. These include demographic questions to understand who is most affected by bike theft, questions about the frequency and circumstances of theft, and inquiries about the effectiveness of current police responses and community preventive measures.

The survey employs a variety of question types to ensure a rich collection of data. Likert scales measure perceptions of safety and police performance, multiple-choice questions assess awareness of preventive measures, and open-ended responses gather personal stories and suggestions for improvement. This mixture of question types helps in collecting both quantitative data and qualitative insights, providing a holistic view of the community's experience with bike theft.

Finally, the survey concludes with a thank-you section, expressing gratitude to the respondents for their time and contributions. This closing reinforces the value of community input and encourages ongoing engagement with TPS initiatives.

This idealized survey and methodology reflect a comprehensive and thoughtful approach to addressing bike theft in Toronto, highlighting the TPS's commitment to community collaboration and data-driven policing.

B Model details

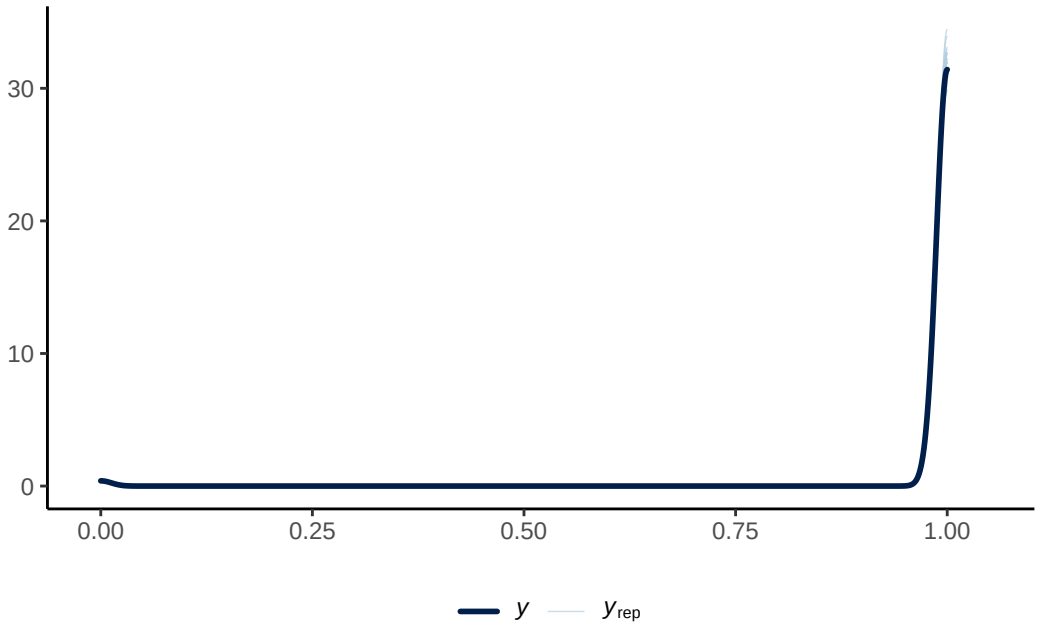


Figure 5: Posterior prediction checr

?@fig-posterior displays the posterior distributions of all exponentiated model coefficients to reflect the effect of predictors on the outcome on the odds ratio scale.

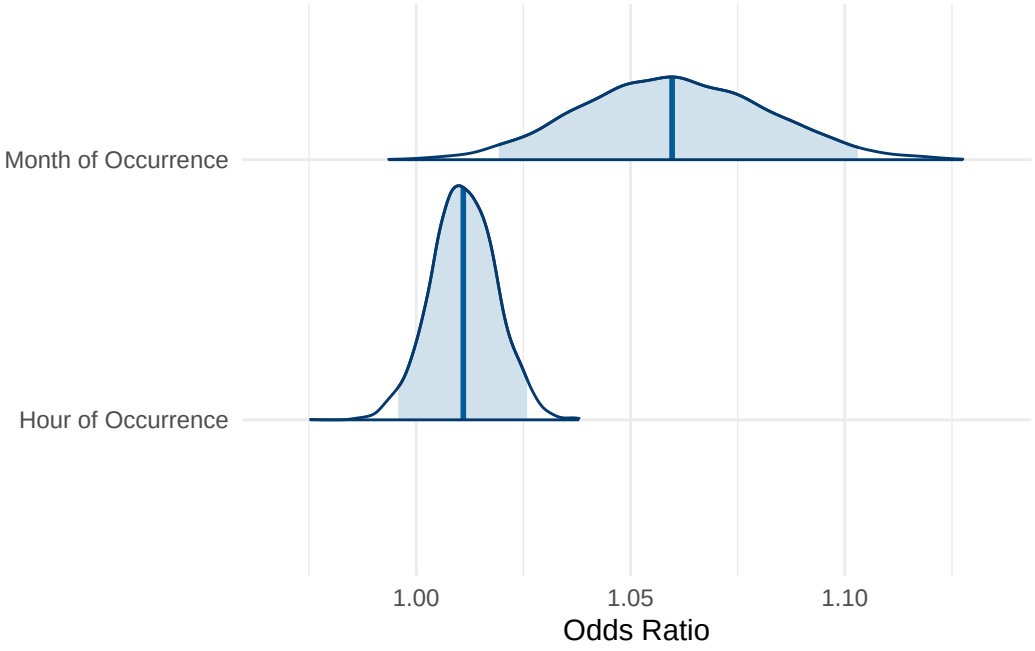


Figure 6: Posterior of betas

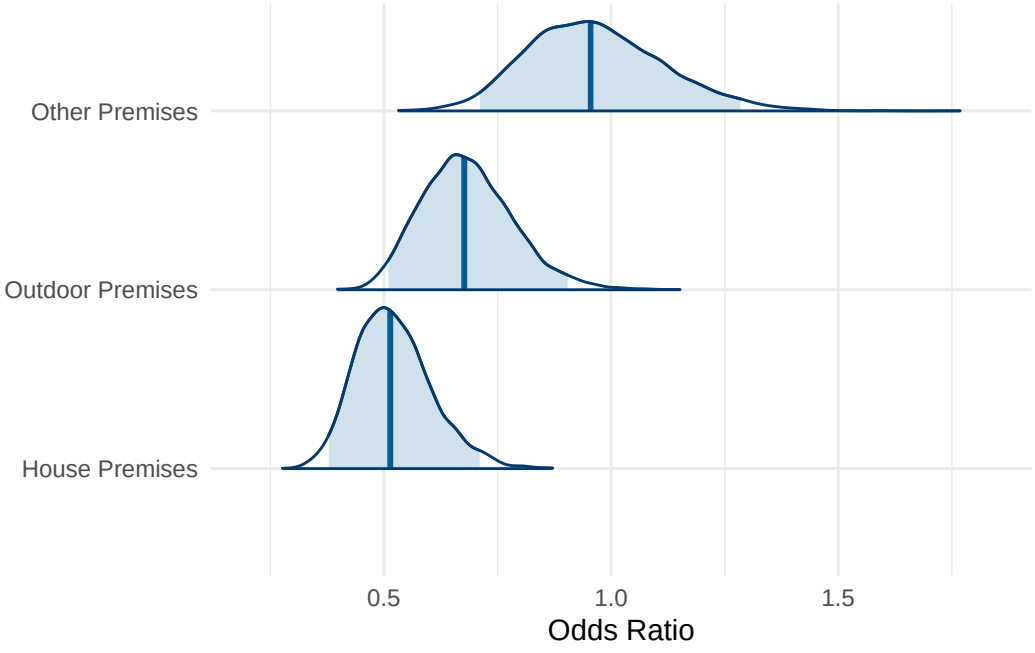


Figure 7: Posterior of betas

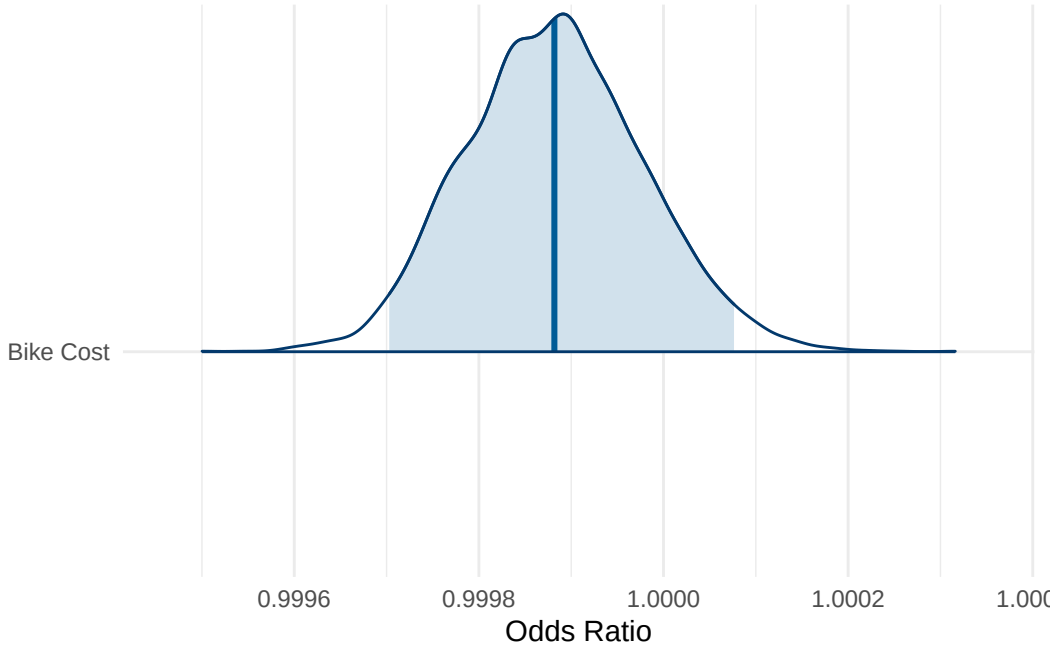


Figure 8: Posterior of betas

C Additional data details

D Model details

D.1 Posterior predictive check

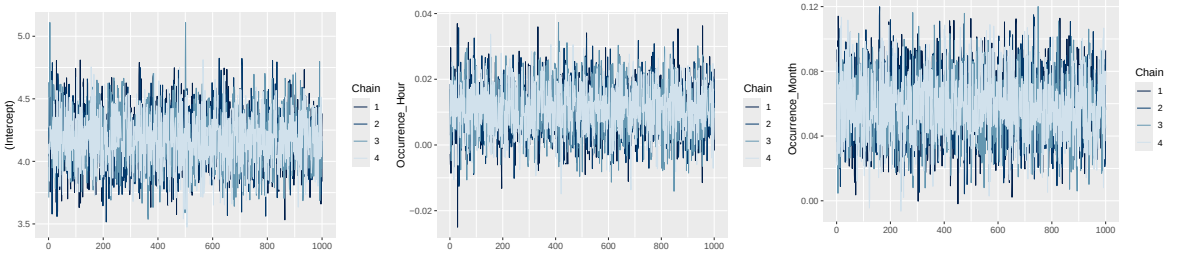
In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

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

D.2 Diagnostics

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

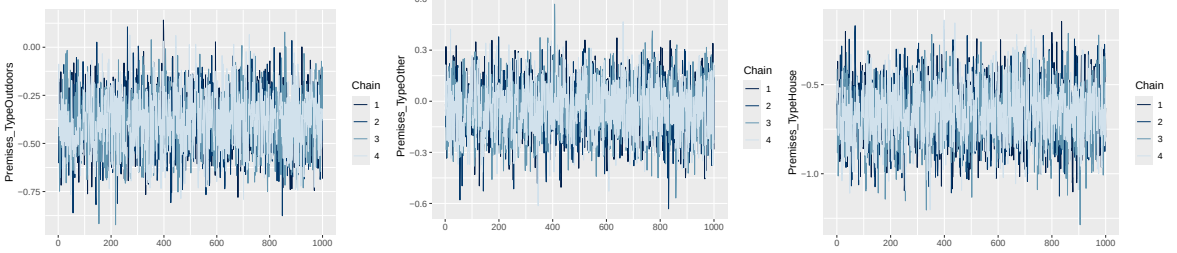
?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...



(a) Intercept Trace Plot

(b) Hourly Occurrence Trace Plot

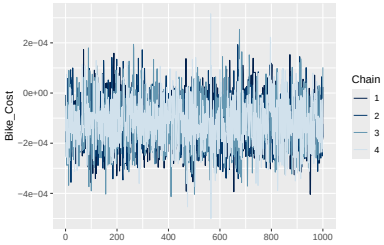
(c) Monthly Occurrence Trace Plot



(d) Outdoor Premises Trace Plot

(e) Residential Premises Trace Plot

(f) Bike Cost Trace Plot



(g) Intercept Trace Plot

Figure 9: Trace plot of Bike Theft

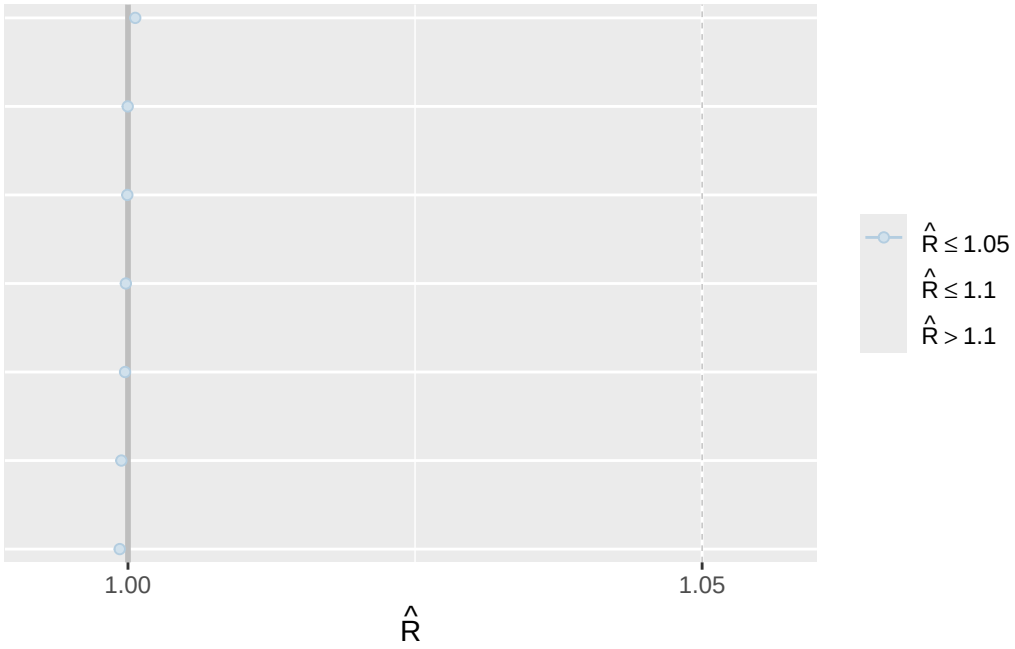


Figure 10: Rhat Plot of Bike Theft

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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

Toronto Police Service. 2024. "Public Safety Data Portal." https://data.torontopolice.on.ca/.