My title*

My subtitle if needed

Jin Zhang

November 28, 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

Data analysis is performed by using statistical programming language R (R Core Team 2023)……………..

2.1 Data Source

The dataset used for this analysis originates from the Toronto Police Service's public data (**torontodataset?**)(cite). 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: Preview of Cleaned Bike Theft Data

Theft_Status	Occurrence_Hour	Occurrence_Month	Premises_Type	Bike_Cost	Region

1	0	9	Apartment	750	Downtown
1	10	12	Outside	400	Downtown
1	0	7	Commercial	600	Midtown
1	20	1	House	900	Toronto
1	20	1	Apartment	479	Downtown
1	9	1	Outside	800	Downtown
1	21	1	Outside	750	Downtown
1	20	1	Outside	100	Downtown
1	12	1	Outside	1800	Downtown
1	17	1	Outside	1100	Downtown

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	7.0	2.5	1.0	7.0	12.0
Bike_Cost	603	0	679.9	479.8	0.0	600.0	2034.0

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 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

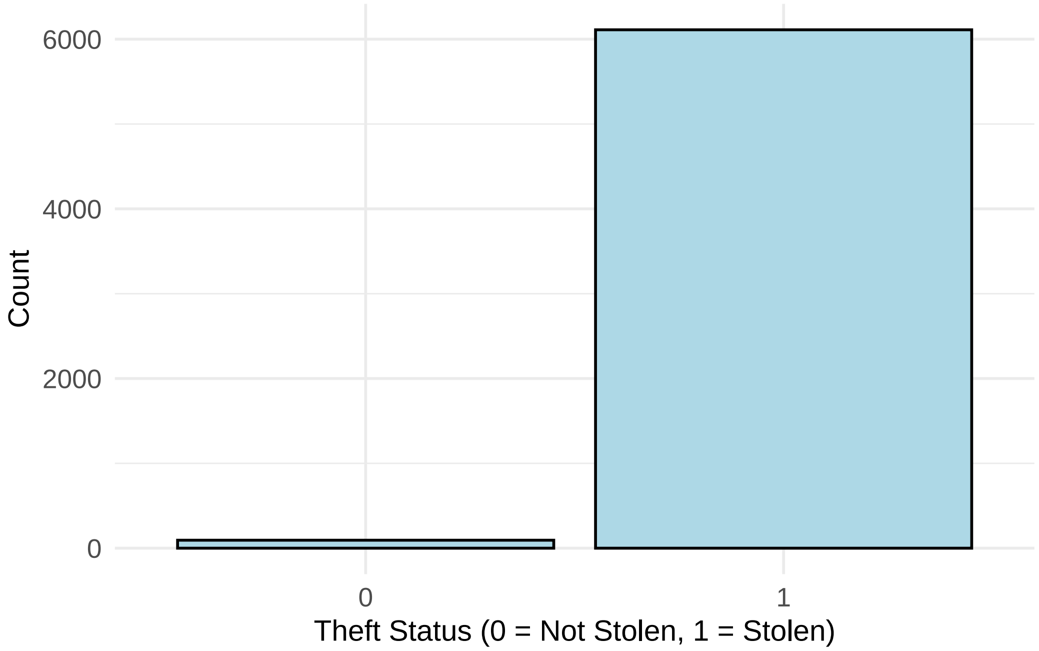


Figure 1: Count of Different Premises Types

Figure 2 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 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 = Oher). In addition, locations such as “outdoor” areas and “apartments” have a particularly high number of bicycle thefts, suggesting that these locations are more likely to have bicycle thefts compared to other locations.

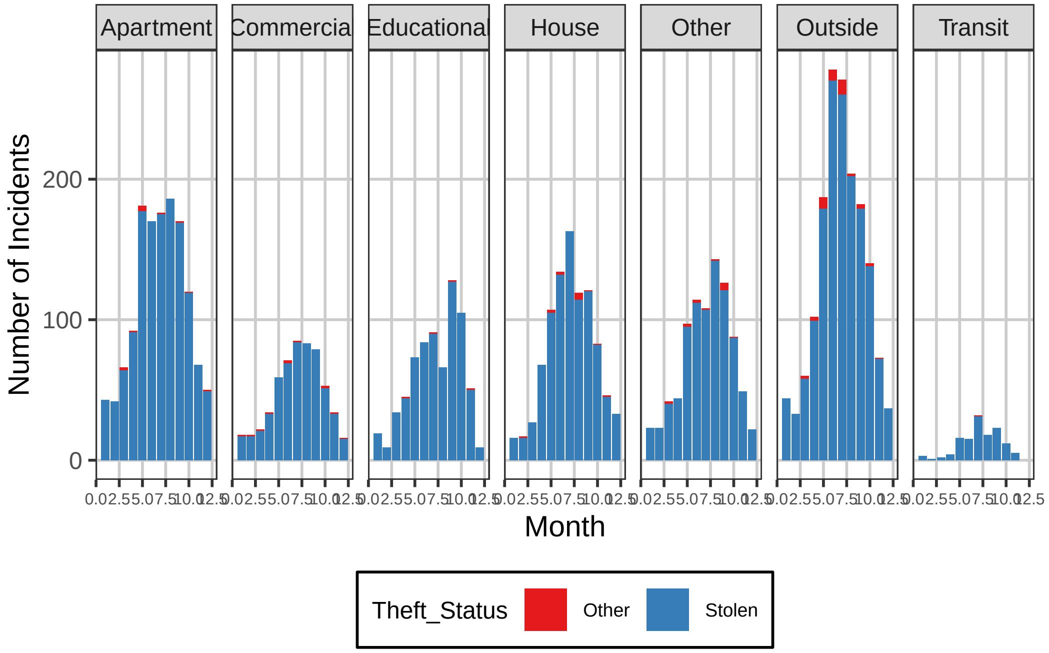


Figure 2: Seasonal Patterns of Bicycle Theft Across Different Premises Types

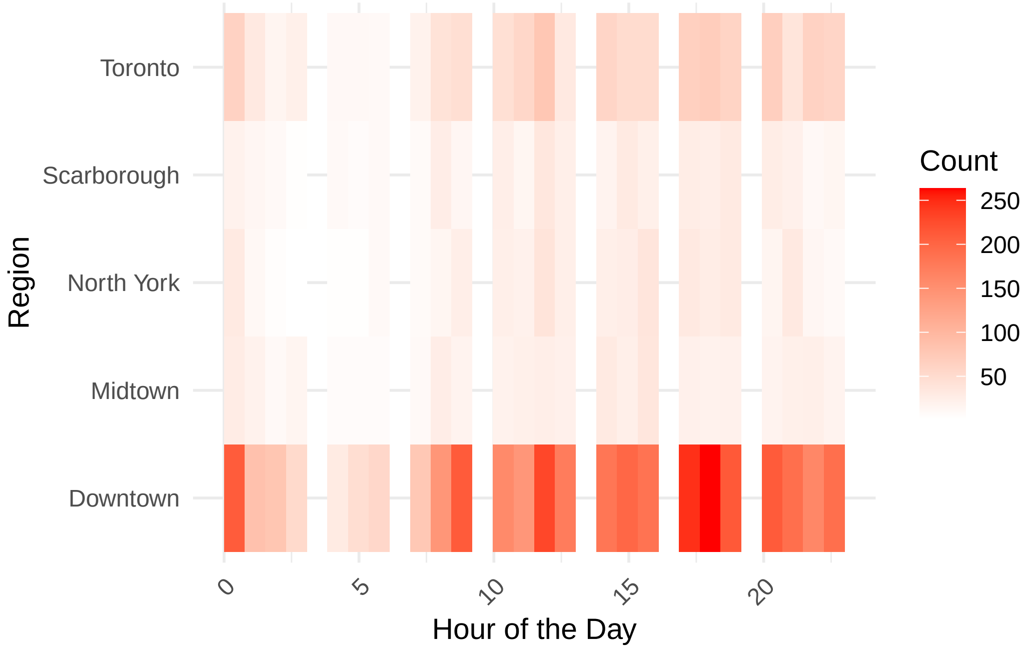


Figure 3: Hourly Bicycle Theft Incidents by Region

This heat map Figure 3 visualizes the frequency of bicycle thefts by the hour of the day across different regions of the city. The intensity of the red color represents a higher number of incidents. Downtown and Midtown are shown to have more frequent thefts, particularly during the late morning to early afternoon hours. This pattern suggests that these central areas experience more bicycle thefts during times when there is likely more activity.

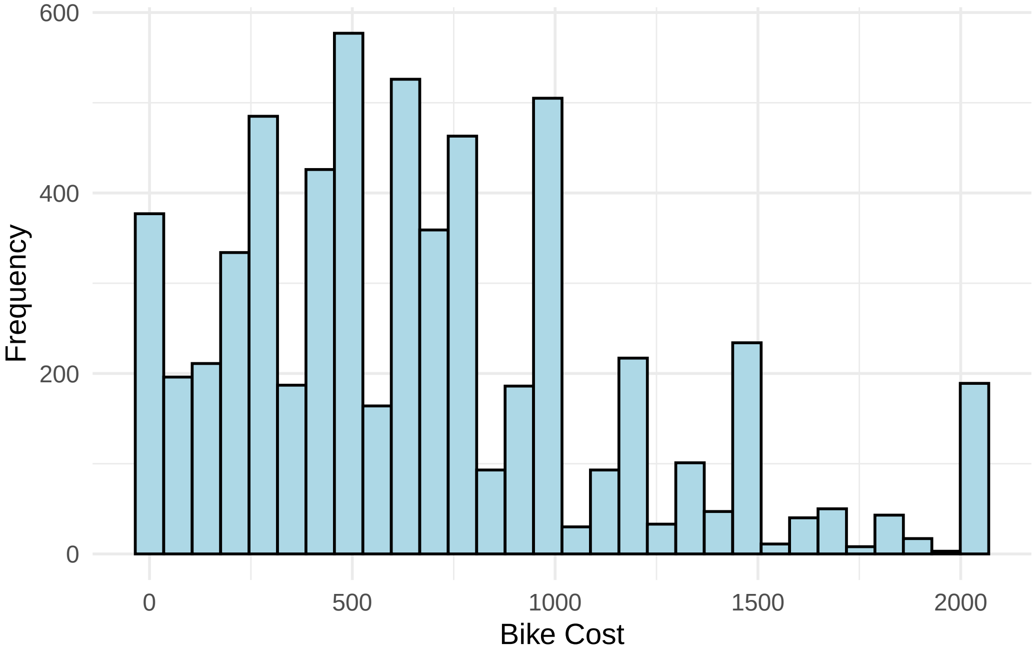


Figure 4: Distribution of Stolen Bicycle Costs

Figure 4 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.

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

2.5 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta &\sim \text{Normal}(0, 2.5) && (4)\\
\gamma &\sim \text{Normal}(0, 2.5) && (5)\\
\sigma &\sim \text{Exponential}(1) && (6)
\end{aligned}$$

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

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 3.

Table 3: Explanatory models of flight time based on wing width and wing length

	Bike_Theft_model
(Intercept)	4.33
	(0.50)
Occurrence_Hour	0.02
	(0.02)
Occurrence_Month	0.05
	(0.04)
Premises_TypeCommercial	−0.88
	(0.43)
Premises_TypeEducational	0.44
	(0.60)
Premises_TypeHouse	−0.52
	(0.43)
Premises_TypeOther	−0.73
	(0.42)
Premises_TypeOutside	−1.18
	(0.35)
Premises_TypeTransit	0.61
	(1.21)
Bike_Cost	0.00
	(0.00)
RegionMidtown	0.71
	(0.54)
RegionNorth York	0.47
	(0.48)
RegionScarborough	−0.79
	(0.32)
RegionToronto	0.31
	(0.32)
Num.Obs.	6205
R2	0.008
Log.Lik.	−469.896
ELPD	−485.5
ELPD s.e.	40.4
LOOIC	971.1
LOOIC s.e.	80.9
WAIC	969.3
RMSE	0.12

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 Additional data details

B Model details

B.1 Posterior predictive check

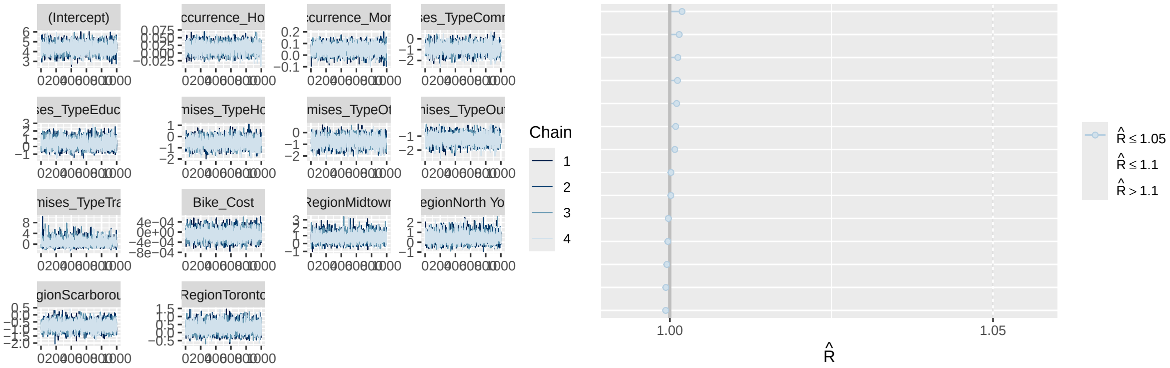
In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

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

B.2 Diagnostics

Figure 5a is a trace plot. It shows... This suggests...

Figure 5b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

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