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

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

Tommy Fu

November 30, 2024

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

Table of contents

1 Introduction **2**
1.1 Overview . . . 2
1.2 Estimand . . . 3
1.3 Results Summary . . . 3
1.4 Why this paper matters . . . 3
1.5 Paper Structure: . . . 3

2 Data **3**
2.1 Overview . . . 3
2.2 Measurement . . . 4
2.3 Data Processing . . . 4
2.4 Outcome variables . . . 4

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

2.5 Predictor Variable 5
2.6 Predictor variables 5

3 Model 5
3.1 Model set-up 6
3.1.1 Model justification 6

4 Results 6

5 Discussion 6
5.1 First discussion point 6
5.2 Second discussion point 6
5.3 Third discussion point 8
5.4 Weaknesses and next steps 8

Appendix 9

A Additional data details 9
A.1 Raw data variables 9

B Model details 10
B.1 Posterior predictive check 10
B.2 Diagnostics 10

References 11

1 Introduction

1.1 Overview

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

1.2 Estimand

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

1.3 Results Summary

The analysis highlights significant relationships between theft occurrences in high-risk neighborhoods and predictors such as bike cost, premises type, and time of day. Findings suggest that high-cost bicycles are frequently targeted in public spaces during evening hours, with some neighborhoods experiencing disproportionately higher theft rates.

1.4 Why this paper matters

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

1.5 Paper Structure:

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

2 Data

2.1 Overview

Our data is sourced from Toronto Police Open Data(Toronto Police Service 2024) specifically on the bike thefts dataset which captures details in bike thefts reported in Great Toronto Area.

The bike thefts dataset including details about the bicycles, such as cost, make, and type, as well as information about the thefts, such as the date, time, location type, premises type, and neighborhood. Geospatial data, including latitude and longitude, is also included to allow for spatial analysis. A sample of the data variables can be find Table 3 in the appendix.

2.2 Measurement

```
This study uses data from Toronto’s open data portal, which provides detailed records of bicy
```

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Data Processing

The raw dataset, sourced from Toronto’s open data portal, underwent several processing steps to ensure accuracy and relevance for the study. Missing values were addressed by removing variables with significant proportions of missing data, such as BIKE_MODEL, **‘BIKE_SPEED’**, and **‘BIKE_COLOUR’**, and filtering out observations with missing **‘BIKE_COST’** or **‘BIKE_MAKE’**. Theft incidents with a STATUS of “UNKNOWN” or “RECOVERED” were excluded to focus on relevant cases. Date variables **‘OCC_DATE’** and ‘REPORT_DATE’ were standardized to a Date format for temporal analysis. A key variable, is_high_risk_neighborhood, was created to identify thefts in the top 10 neighborhoods with the highest frequencies, enabling targeted investigation. Geographic coordinates **‘LONG_WGS84’** and **‘LAT_WGS84’** and contextual variables, such as **‘LOCATION_TYPE’** and **‘PREMISES_TYPE’**, were retained for spatial and environmental analysis. The cleaned dataset was saved as a Parquet file for efficient use in modeling and visualization workflows.

2.4 Outcome variables

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

Table 1: Preview of the Cleaned Data

Table 1: First 5 Rows of the Cleaned Dataset

BIKE_COST	OCC_DATE	LOCATION_TYPE	PREMISES_TYPE	STATUS
1300	12/26/2013 5:00:00 AM	Other Commercial / Corporate Places (For Profit, Warehouse, Corp. Bldg	Commercial	STOLEN
500	12/30/2013 5:00:00 AM	Streets, Roads, Highways (Bicycle Path, Private Road)	Outside	STOLEN
750	9/30/2013 5:00:00 AM	Apartment (Rooming House, Condo)	Apartment	STOLEN
1500	12/25/2013 5:00:00 AM	Apartment (Rooming House, Condo)	Apartment	STOLEN
400	12/25/2013 5:00:00 AM	Streets, Roads, Highways (Bicycle Path, Private Road)	Outside	STOLEN

2.5 Predictor Variable

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.6 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.

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

We run the model in R (R Core Team 2023) using the `rstanarm` package of Brilleman et al. (2018). 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 2.

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.

Table 2: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	−0.39
	(0.21)
BIKE_COST	0.00
	(0.00)
PREMISES_TYPECommercial	0.98
	(6.01)
PREMISES_TYPEEducational	3.69
	(9.23)
PREMISES_TYPEHouse	−0.54
	(4.82)
PREMISES_TYPEOther	0.18
	(5.15)
PREMISES_TYPEOutside	0.34
	(3.76)
PREMISES_TYPETransit	−4.57
	(11.58)
LOCATION_TYPEBank And Other Financial Institutions (Money Mart, Tsx)	0.94
	(6.01)
LOCATION_TYPEBar / Restaurant	0.51
	(6.08)
LOCATION_TYPECommercial Dwelling Unit (Hotel, Motel, B & B, Short Term Rental)	31.10
	(26.73)
LOCATION_TYPEConvenience Stores	0.37
	(6.14)
LOCATION_TYPEGo Station	3.71
	(11.50)
LOCATION_TYPEGroup Homes (Non-Profit, Halfway House, Social Agency)	0.24
	(5.38)
LOCATION_TYPEHomeless Shelter / Mission	0.22
	(5.39)
LOCATION_TYPEHospital / Institutions / Medical Facilities (Clinic, Dentist, Morgue)	19.53
	(13.91)
LOCATION_TYPENursing Home	−39.16
	(33.69)
LOCATION_TYPEOpen Areas (Lakes, Parks, Rivers)	−0.74
	(3.78)
LOCATION_TYPEOther Commercial / Corporate Places (For Profit, Warehouse, Corp. Bldg	−0.29
	(6.01)
LOCATION_TYPEOther Non Commercial / Corporate Places (Non-Profit, Gov'T, Firehall)	−1.36

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

A.1 Raw data variables

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

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

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

Variable Name	Sample Value
NEIGHBOURHOOD_140	South Parkdale (85)
LONG_WGS84	-79.4436451187837
LAT_WGS84	43.6376571871944
x	-8843626.12140861
y	5409538.95619472

B Model details

B.1 Posterior predictive check

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

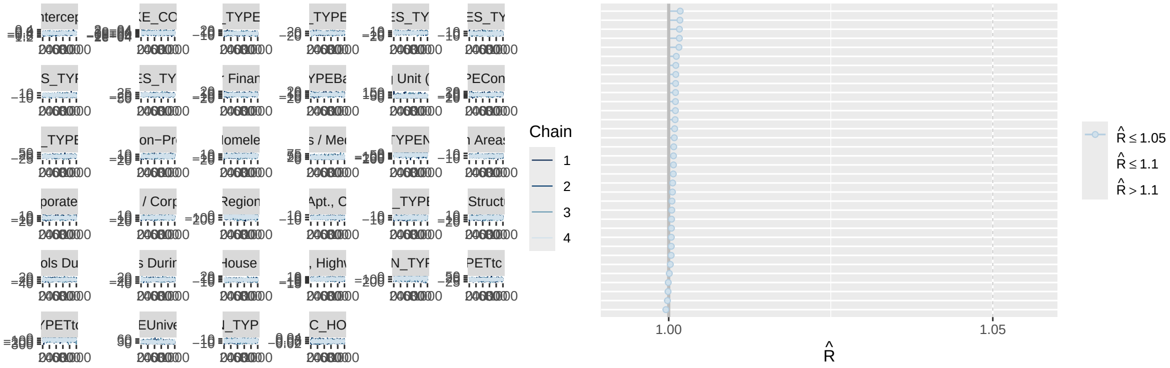
In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

Figure 1a is a trace plot. It shows… This suggests…

Figure 1b is a Rhat plot. It shows… This suggests…



(a) Trace plot

(b) Rhat plot

Figure 1: Checking the convergence of the MCMC algorithm

References

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

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