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

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

Tommy Fu

December 1, 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

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

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

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

1.4 Why this paper matters

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

1.5 Paper Structure:

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

2 Data

2.1 Overview

Our dataset is sourced from Toronto Police Open Data (Toronto Police Service 2024), specifically the "Bicycle Thefts" dataset, which provides comprehensive details on bicycle thefts reported across the Greater Toronto Area. This dataset includes variables that capture information about the stolen bicycles, such as their reported cost, make, and type, as well as contextual details about the thefts, including the date, time, location type, premises type, and neighborhood. Additionally, geospatial data, including latitude and longitude coordinates, enables spatial analysis to explore patterns in theft occurrences across different areas. Heatmap of Toronto area by theft counts will be shown in Section 2.2 using shapefile data sourced from University of Toronto Map Library (University of Toronto n.d.).

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

2.2 Measurement

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

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

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

2.3 Data Processing

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

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

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

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

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

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 variables

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

2.5.1 Distribution of High-Risk Neighborhoods

The majority of reported bicycle thefts occur in a small subset of neighborhoods, demonstrating significant clustering in high-risk areas. Figure Figure 1 illustrates the distribution of thefts by risk classification with different premises types comparison.

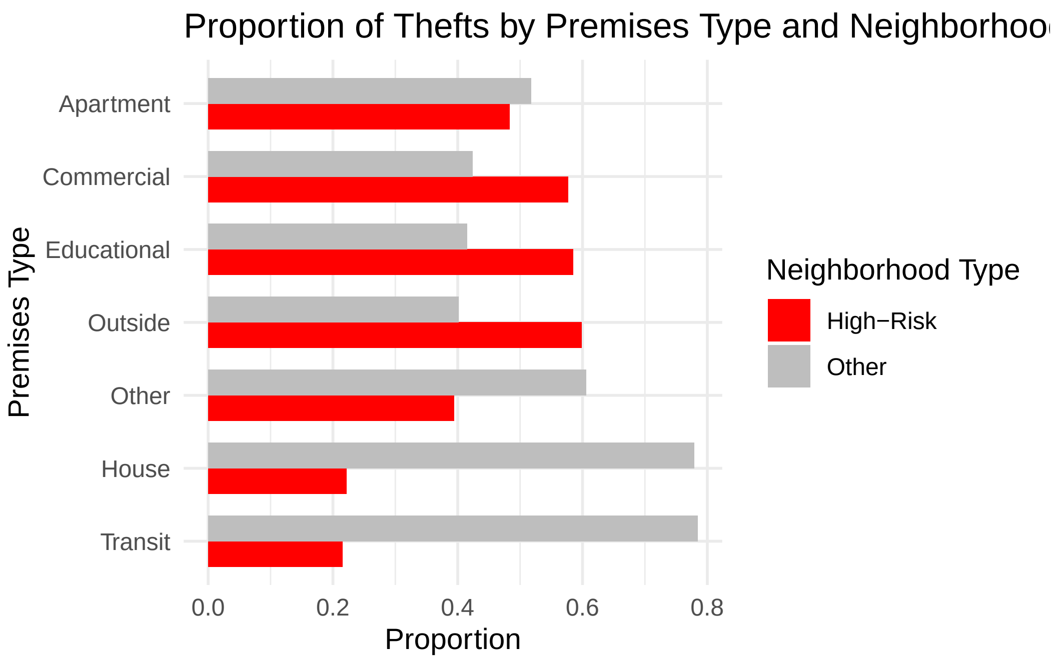


Figure 1

The bar chart shows that high-risk neighborhoods account for a significant proportion of thefts, despite representing only 10 neighborhoods out of the total. This concentration highlights the importance of focusing on these areas for targeted interventions.

2.5.2 Characteristics of High-Risk Neighborhoods

Table Figure 2 provides a summary of the top 10 high-risk neighborhoods, including the total number of thefts recorded in each.

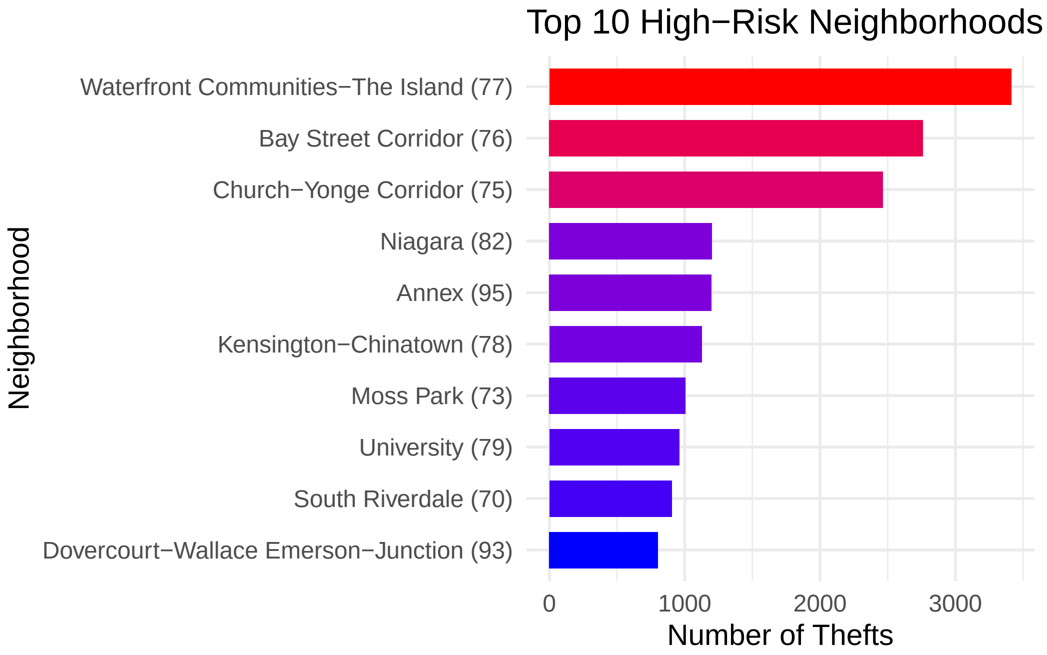
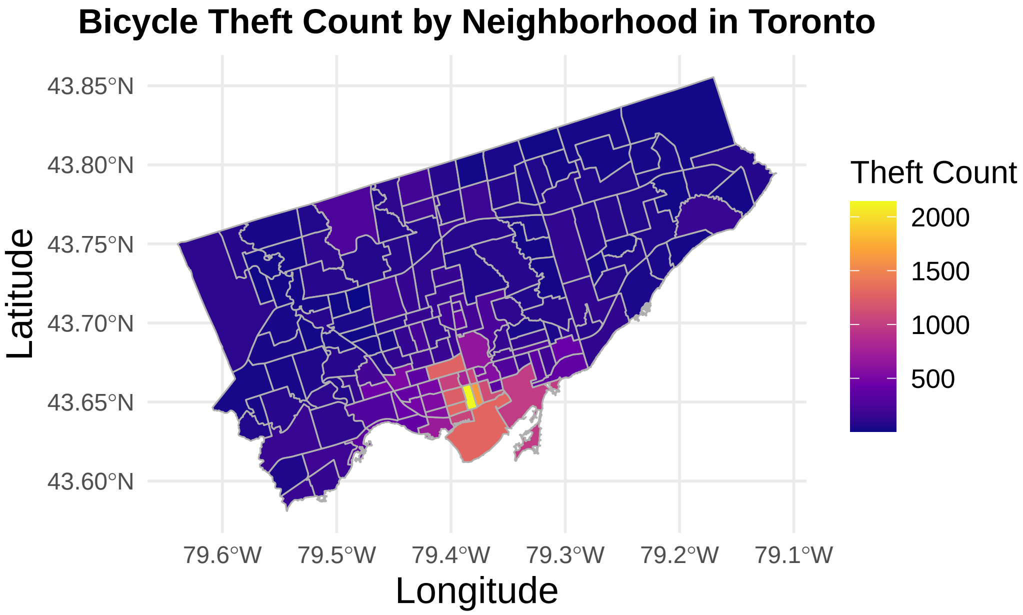


Figure 2

2.6 Data Visualizations



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 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.

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 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

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…

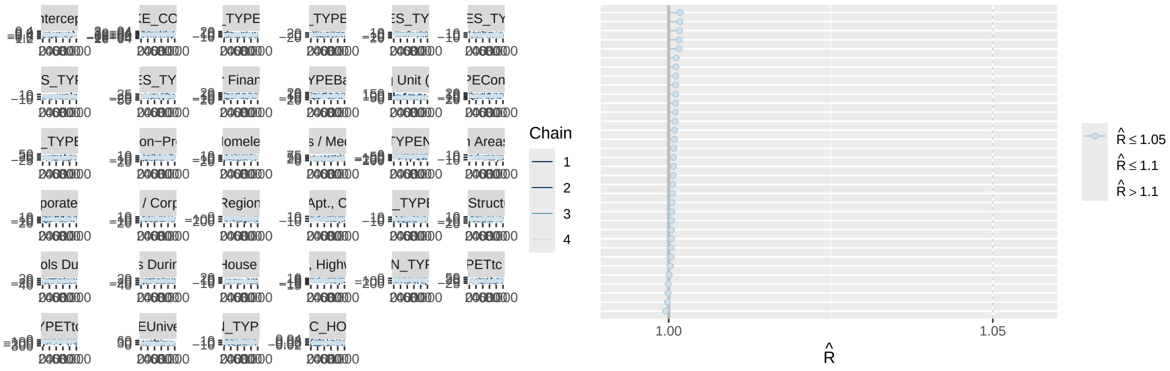
In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

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

B.2 Diagnostics

Figure 3a is a trace plot. It shows… This suggests…

Figure 3b is a Rhat plot. It shows… This suggests…



(a) Trace plot

(b) Rhat plot

Figure 3: 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/.

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

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

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

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

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

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

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

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

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

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