Examining Bicycle Thefts in Toronto and the Benefit of Riding a Cheap Red BMX Bike During Winter*

From 2014 to 2019

Ken Lee

31 January 2021

Abstract

Bicycle theft is a problem that every city experiences, but is overlooked even when it generates a wide array of benefits to the community. This research paper will focus on exploring the patterns of bicycle theft in Toronto from 2014 to 2019, while identifying factors influencing the theft and recovery of bikes in the city. More specifically, it will conduct an exploratory analysis on the primary offences, date, price (cost), bike type, and bike colour of bike-theft occurrences. Insights from this study will not just be useful for the public's understanding of bike thefts, but also for the development of bike-theft classification and predictive models focused on diving deeper into the causes.

1 Introduction

Biking is a phenomenon that is shared across the world, bringing benefits such as exercise, money-saving, entertainment, reduced-carbon emissions, and much more. However, there are many challenges that limit its wider usage. One of those challenges is bike theft, which is a common problem that occurs in many major cities from San Francisco all the way to Toronto. However, as stated by the article "The Problem of Bicycle Theft" (Shane D. Johnson 2008), bike theft is usually considered a low-priority, meaning this problem is not actively being solved. Hence, this is why it is important for us to dive deeper into this problem. After all, seeing the many advantages this hobby or means of transportation brings to the world, it is vital to understand the patterns and causes for bike-related thefts.

This paper will focus on documented cases of bike thefts in the city of Toronto, examining bicycle thefts and bike-related crimes such as breaking and entering and shoplifting. The main purpose of this study is to explore the data available and understand the patterns of bike theft and the factors that could be causing or encouraging them. We will be observing the crimes over time, the years, months, and days from 2014 to 2019, considering their frequencies and how they differ by primary offence (the type of theft/crime). Additionally, factors bike users have control over, such as the cost of the bike, the type, and the color, will be explored for their occurrence frequencies and recovery rates. On the other hand, due to the limitations and biases of the sample data set, we will not be focusing on regression analysis and other forms of classification or prediction models. This paper will first discuss the source of the information, and its potential drawbacks and biases, followed by an exploratory analysis of the Toronto bicycle thefts from 2014 to 2019, examining offence patterns throughout the years, and factors such as the bike cost, type, and their recovery rates.

[*] Code and data are available at: https://github.com/kenllee97/kenllee97-Examining_Toronto_Bicycle_Theft_2014_2019.

2 Data

For this project, we use R (R Core Team 2020) and packages such as "Tidyverse" (Wickham et al. 2019) to analyze the data. Additionally, R packages like "janitor" (Firke 2021) to clean the data, and "knitr" (Xie 2015) to create this PDF file. AT last, "ggplot2" (Wickham 2016) and "KableExtra" (Zhu 2020) were also used to create graphs and tables.

2.1 Data Source

The data we are using for this report comes from the R package "opendatatoronto" (Gelfand 2020). This package helps us obtain the data sourced from Toronto's Open Data Portal, which is the official source for data collected from Toronto's divisions and agencies. The data set we will be focusing on is "Bicycle Thefts" ("Bicycle Thefts" 2020), which was published under the Open Government License by Toronto Police Services and last updated on Aug 18, 2020 (data set refreshes annually). The data was collected from Toronto police crime reports regarding bicycle incidents. The methodology in which this data was collected was from police crime reports filed through the phone, in-person, or online. Hence, because they are police reports, a strong aspect of the data set is the abundant information that it provides. On the other hand, the weaker aspect is the fact that not all information may be accurate, given as they are accounts filed by individuals and the fact that it leaves out unreported cases.

The "Bicycle Theft" data set contains bicycle theft occurrences from 2014 to 2019, with a sample of 21,854 unique reports and 26 features (columns). The data set consists of the following features: *id, Index*, event_unique_id, Primary_Offence, Occurrence_Date, Occurrence_Year, Occurrence_Month, Occurrence_Day, Occurrence_Time, Division, City, Location_Type, Premise_Type, Bike_Make, Bike_Model, Bike_Type, Bike_Speed, Bike_Colour, Cost_of_Bike, Status, Hood_ID, Neighbourhood, Lat, Long, ObjectId, and geometry. Upon cleaning the data, there were no instances of duplicates, but there were a significant amount of redundancies from the unique identifiers such as id and Index to the premise type and location type. In the exploratory phase, we reviewed all the features frequencies and their potential relationships (which can be found in the scripts folder). However, we will be focusing on the features that had the most potential impact for an exploratory analysis, as many of the features were limited by biases and the need for additional data sets to create valuable and significant insights. The primary features used are "Primary Offence" (the type of offence charged/filed), "Occurrence Date" (when it occurred), "COst of Bike" (the cost of the occurrence), "Bike Type" (type of bike), "Bike Colour" (colour of the bike), and "Status" (whether it was recovered).

You can find the code on how we retrieved the data on Toronto bicycle thefts in the scripts folder.

2.2 Data Biases

Before summarizing the data, it is important for us to review the potential biases and problems of this data set that may affect the internal and external validity of this paper's findings. One of the main biases to consider is the fact that this data set only includes information on reported bike thefts, disregarding unreported ones. The data set also contains the year, month, day, and time of the occurrence, but should not be taken solemnly as some of the victims may not recall the date and time of the incident accurately (especially when there are no NA values for these fields). For instance, a victim may have left a bicycle unattended for a couple of days before finding out it was stolen. Hence, the individual would not be able to tell exactly when the incident took place. Speaking of unintentional false data recollection, there may also be intentional fake reports. The reason being that individuals may have many reasons for creating a false report such as claiming an insurance policy (affecting the accuracy of the bike price). Additionally, the data set is also biased because it only consists of items that were stolen, leaving out items that were not stolen. This would ultimately limit the use of classification or predictive models. In other words, one cannot use this data set to determine the many features leading to bike thefts.

At last, another aspect of the data that may affect the statistical significance and validity of this paper's findings is the fact that locations (latitude and longitude) were deliberately offset to the nearest road intersection for ethical reasons, protecting the privacy of the parties involved. Nevertheless, it is also vital to have accurate information on neighborhoods of the incidents, as biased data may create biased patterns that could result in more police patrols in certain areas and reduced traffic to certain neighborhoods, affecting local businesses. All in all, the data set may have potential biases and inaccuracies which can affect the paper's validity and involve ethical implications regarding the use of its discovered insights.

2.3 Exploratory Analysis

Upon the initial exploratory analysis, we have created a graph showing the frequency of bike theft-related incidents from 2014 to 2019 (**Figure 1**). From the graph, we can clearly see a cyclical pattern that seems to follow the seasonal changes of the city of Toronto. More specifically, the number of bike theft occurrences seems to reach its peak during the summer (warm periods), and its troughs during the winter (cold periods). This also seems to reflect the findings of another research study, "Breaking into bicycle theft:Insights from Montreal, Canada" (Lierop 2015), depicting summer as the peak of bike thefts. All in all, bike thefts in Toronto seem to follow a logical cyclical weather pattern as there will be more bike sage during the summer (leading to higher chances of bike theft), and less bike usage during the winter (decreasing the chances of bike thefts).

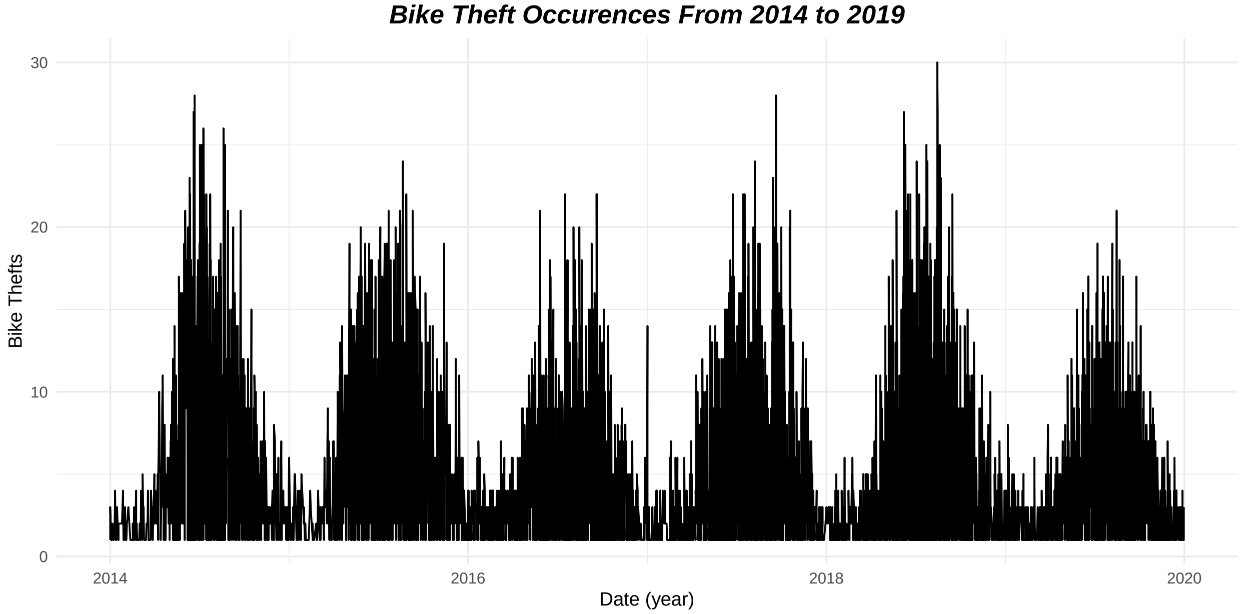

Figure 1: Bike Theft Occurences From 2014 to 2019

Nevertheless, if we dice the Occurrences into their primary offences, we can derive more detailed insights into the patterns of bike thefts in Toronto from 2014 to 2019. **Figure 2** shows us the top 10 most frequent primary offences, denoting general thefts under $5,000 and bicycle thefts under $5,000. This goes to show us that most of the thefts are on normal bikes and their parts. Additionally, **Figure 3** illustrates the frequency of the top ten primary offences from 2014 to 2019. From this figure, we can identify two unique events and patterns. The first one is the change in primary offences, as thefts under $5,000 dominate from 2014 to 2016, but then shifts to the theft of bicycle under $5,000 from 2016 to 2019 (Illustrating how thieves began to steal whole bicycles instead of just their parts). The second insight is the fact that there was a spike of property found in the beginning of 2017, which seems to make sense, as the previous year is the period where more whole bikes were stolen, making it easier to be found. Nevertheless, it does not further explain why the number of property found decreased by mid-2017 (Maybe the perpetrators learned and adapted their methodologies, obscuring the tracking of stolen items).

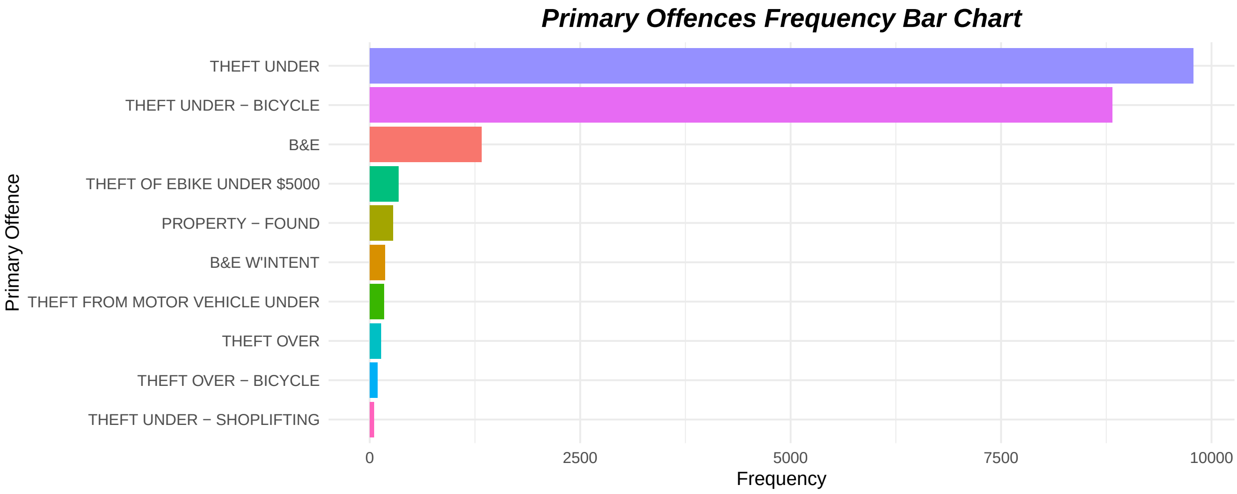

Figure 2: Frequency of Primary Offences

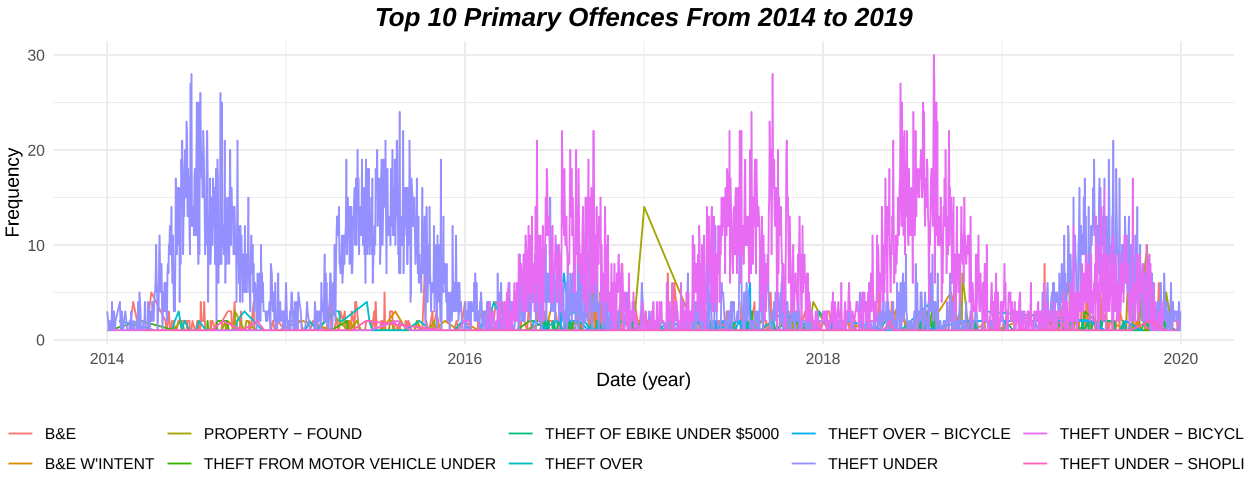

Figure 3: Top 10 Primary Offences From 2014 to 2019

Now that we know the cyclical pattern of bicycle thefts and the changing trend of having the whole bicycle stolen, what can be done to reduce its chances of happening, or at least increasing the chances to recover them? **Figure 4** shows us a point histogram of all the bike costs. The first aspect we notice is that many of the prices ranged between $0 and $1,000. There were also some outliers as Table **1** denotes a maximum of $120,000. Additionally, it also shows us that the mean is 938, with a median of 600, highlighting the right skewness of the data.

Hence, based on this data, we can tell that the majority of the thefts happen at the lower end of the cost, and this may be because there much more lower-valued bikes and items, and more expensive properties may be secured more safely (reducing chances of theft). After all, as highlighted by the study "Breaking into bicycle theft: Insights from Montreal, Canada" (Lierop 2015), more expensive bikes were less likely to be stolen. However, a histogram is not enough, as one would need other techniques such as logistic regression models and less biased data to determine the correlation between cost and chances of theft. Additionally, there would also be other challenges limiting the analysis and correlation of these features. One of them is that they are not distilled into groups of primary offences, meaning the cost covers a wide array of items from bicycles to electric bikes. Therefore, it would be hard to make such a general conclusion, as there would be different factors at play for distinct items such as a bicycle and a store. Additionally, most of the data on stolen bikes have blank cost values, further complicating the analysis.

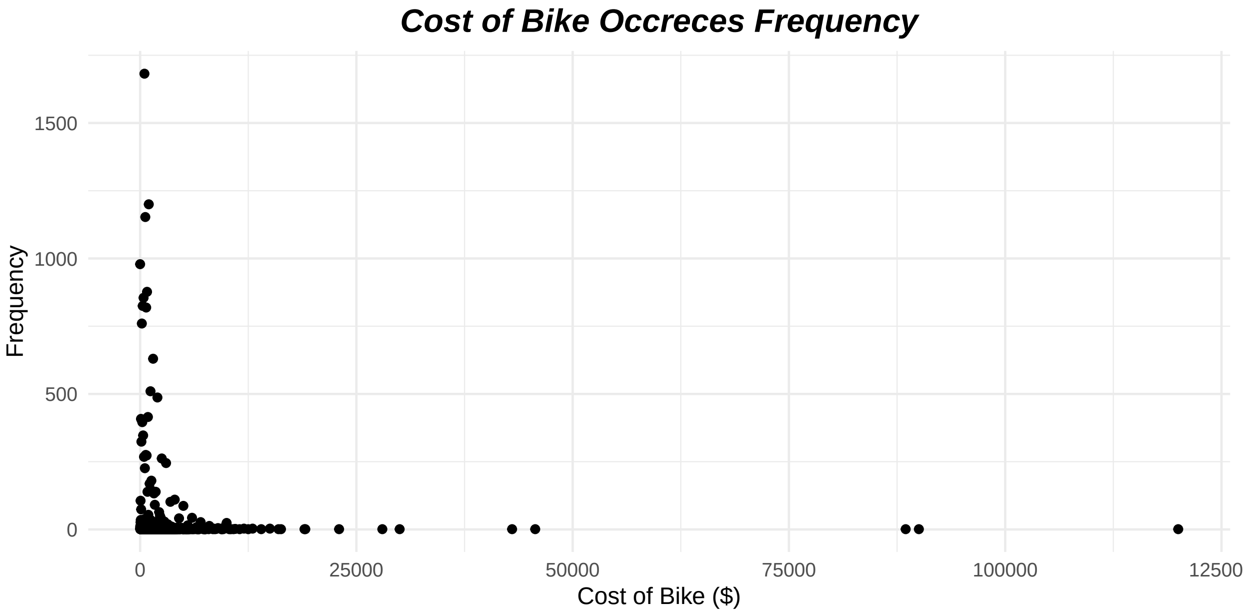

Figure 4: Cost of Bike Occreces Frequency

Table 1: Cost of Bike Statistical Summary

	Cost of Bike
	Min. : 0
	1st Qu.: 350
	Median : 600
	Mean : 938
	3rd Qu.: 1000
	Max. :120000

Nonetheless, although we may not be able to distinctively identify the causation of theft from the prices, we can examine the recovery of thefts. **Figure 5** illustrates the cost of the recovered thefts, highlighting how higher cost thefts may have lower chances of being recovered, seeing as most of the recovered items are at the low end. However, we cannot clearly pin whether lower costs can lead to higher chances of recovery, as many other thefts that were not recovered were also in the same range.

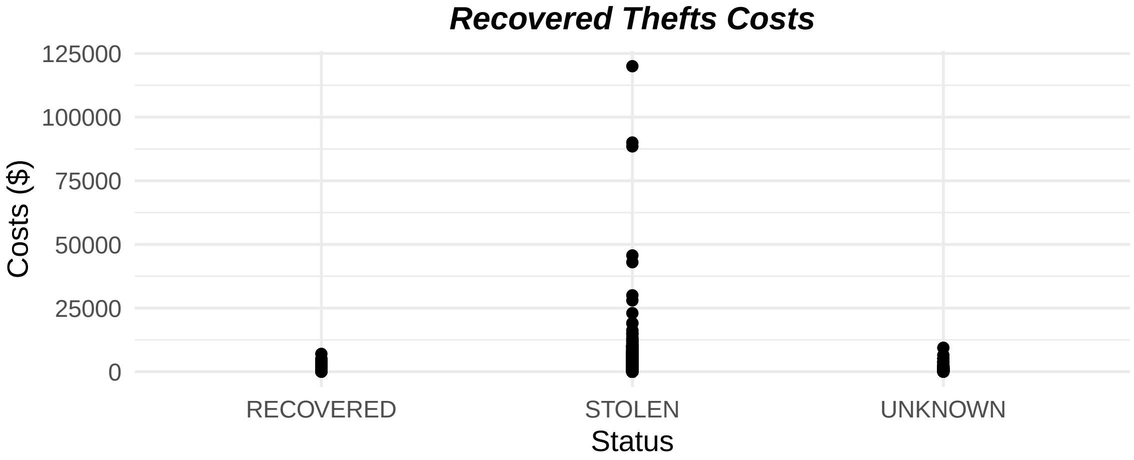

Figure 5: Recovered Thefts Costs

At last, we also explored other factors individuals have control over, such as the colour of the bike, to determine which ones had the highest chance for recovery. **Table 2** illustrates the thefts of bikes grouped by the ten most reported colours, denoting their frequency and recovery date. From the table, we can tell that black was the colour that was stolen the most. This makes sense, as it is probably the most used colour. However, we also cannot conclude that this colour will increase the chances of theft, as we do not have the data on all the other bikes that were not stolen. Nevertheless, we can recognize that the color red bike has the highest chances of recovery, as it had the highest recovery rate.

Table 2: Bike Colour Frequency and Recovery Rate

Bike Color	Frequency	Recovered	Recovery Rate
RED	1533	22	0.0143509
GRN	588	8	0.0136054
WHI	1691	23	0.0136014
PLE	397	5	0.0125945
BLU	1960	24	0.0122449
SIL	1010	12	0.0118812
BLK	6212	60	0.0096587
GRY	1798	15	0.0083426
ONG	398	2	0.0050251
DBL	341	1	0.0029326

The final factor we explored was the type of bike. **Table 3** shows the frequency in which different types of bikes were stolen and their recovery rates. Here we can see that mountain bikes (MT) were the most stolen, but as mentioned before, it can not be said that it increases the chances of theft as we have no data on the bikes that were not stolen. For all we know, they might just have been the most stolen because of their higher proportion of mountain bikes in the population. Nevertheless, we can tell that BMX bikes had the highest recovery rate.

Table 3: Bike Type Freqency and Recovery Rate

Bike Type	Frequency	Recovered	Recovery Rate
BM	324	7	0.0216049
SC	250	5	0.0200000
RC	2384	36	0.0151007
EL	1098	16	0.0145719
TO	1127	14	0.0124224
MT	6915	83	0.0120029
RG	5733	56	0.0097680
OT	3513	34	0.0096783
FO	160	1	0.0062500

All in all, this paper may have uncovered the cyclical theft pattern of bikes, the changing trend of stealing bikes, and the idea that a cheap red BMX bike may have the highest chance of recovery, but it certainly does not explain the causes for the theft. After all, this paper was just an exploratory examination of the Toronto bike theft data set. To further understand the features affecting thefts and how to reduce this problem, more data sets on the bike populations (especially ones that have not been stolen) and other aspects such as the number of bike parking spots need to be examined simultaneously (like what other studies like “Breaking into bicycle theft: Insights from Montreal, Canada” (Lierop 2015) have done). Only then would we be able to implement more techniques from logistic regression to deep learning to further understand the causes of this problematic phenomenon.

References

"Bicycle Thefts." 2020. Toronto Police Services; City of Toronto. https://open.toronto.ca/dataset/bicycle-thefts/.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal.* https://CRAN.R-project.org/package=opendatatoronto.

Lierop, Grimsrud van, D. 2015. "Breaking into Bicycle Theft: Insights from Montreal, Canada." *International Journal of Sustainable Transportation.*

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Shane D. Johnson, Adam Thorpe, Aiden Sidebottom. 2008. "The Problem of Bicycle Theft." https://popcenter.asu.edu/content/bicycle-theft-0.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

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. 2015. *Knitr: A Comprehensive Tool for Reproducible Research in r.* Chapman; Hall/CRC.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.