Examining Bicycle Thefts in Toronto*

From 2014 to 2019

Ken Lee

27 January 2021

Abstract

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

1 Introduction

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

2 Data

2.1 Data Source

The data we are using for this report comes from the R package opendatatoronto. This package helps us obtain the data sourced from Toronto's Open Data Portal, which is the official source for data collected from the city's divisions and agencies. The data set we will be focusing on is "Bicycle Thefts," which was published by Toronto Police Services and last updated on Aug 18, 2020 (data set refreshes annually). Additionally, the data set is under the Open GOvernment License - Toronto.

The "Bicycle Theft" data set contains bicycle theft occurrences from 2014 to 2019, containing 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.

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 from 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. Additionally, the data set also includes 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, whether locked or not, 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, this data may also suffer from an intentional false data recollection. 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).

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 it's discovered findings.

2.3 Exploratory Analysis

3 References