Bicycle Theft Report

What is the scope of bike theft in Toronto over the past few years?

Chris Yan 1009096746, Aaron Dong 1008961232, Armaan Rehman Shah 1009641309

2023-04-07

Introduction

The report concerns with the data “Bicycle Thefts Open Data” which contains occurrences related to bicycle thefts reported to the Toronto Police Service. These occurrences are related to a variety of offences where the theft of a bicycle was included. We obtained this dataset from the City of Toronto Open Data portal. It records all of the bike thefts in Toronto starting from as early as 2013. The dataset is published by the Toronto Police services and is refreshed annually all the way up until 2022. The dataset contains large amount of entries and multiple information variables.

Purpose

The purpose of our report is to inform the public about the bike safety concerns in Toronto and its surrounding areas with regards to the precedence of stolen bikes in certain areas. We want to make it known in our report that the public should avoid parking bikes in hot-spots where bikes are often stolen. A map of the density of the stolen bike occurrences will show the precise locations where one should avoid leaving their bike. With our research we hope to provide a thorough investigation into these hot-spots and also provide such information of stolen bike occurrences to the Toronto Police.

Dataset Description

Table 1: Dataset Observations, Data types, and Descriptions

```
##
##
## |Variable              |Types     |Description                            |
## |:---------------------|:---------|:--------------------------------------|
## |X_id                  |integer   |unique entry id order                  |
## |event_unique_id       |character |unique event id unsorted               |
## |Primary_Offence       |character |type of crime committed                |
## |Occurrence_Date       |character |date of crime committed                |
## |Occurrence_Year       |integer   |year of crime committed                |
## |Occurrence_Month      |character |month of crime committed               |
## |Occurrence_DayOfWeek  |character |days of week crime (Mon-Sun)           |
## |Occurrence_DayOfMonth |integer   |date of month crime (1-31)             |
## |Occurrence_DayOfYear  |integer   |day of year crime (1-365)              |
## |Occurrence_Hour       |integer   |hour of day crime (1-24)               |
## |Report_Date           |character |incident reported to authorities       |
## |Report_Year           |integer   |year of incident reported              |
## |Report_Month          |character |month of incident reported             |
## |Report_DayOfWeek      |character |day of week incident (Mon-Sun)         |
## |Report_DayOfMonth     |integer   |day of month incident (1-31)           |
## |Report_DayOfYear      |integer   |day of year incident (1-365)           |
## |Report_Hour           |integer   |hour of day incident                   |
## |Division              |character |division of location                   |
## |City                  |character |city name in Toronto                   |
## |Hood_ID               |character |hood id for neighborhoods (out of 158) |
## |NeighbourhoodName     |character |neighborhood name in Toronto           |
## |Location_Type         |character |broad crime location                   |
## |Premises_Type         |character |description of bike theft location     |
## |Bike_Make             |character |bike brand in theft                    |
## |Bike_Model            |character |bike model in theft                    |
## |Bike_Type             |character |type of bike stolen                    |
## |Bike_Speed            |integer   |bike color in theft                    |
## |Bike_Colour           |character |bike speed in theft                    |
## |Cost_of_Bike          |numeric   |bike cost in theft                     |
## |Status                |character |current bike status                    |
## |geometry              |character |precise crime location                 |
```

As with all the data, let us do a preliminary safety removal of any data that has NA entries. In addition, let us print out a summary of crimes by year to see if there are any outliers.

Table 2: Total Number of Bike Thefts Each Year

```
##
## 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
##    1    2    3    2   28 2475 2611 3132 3253 3413 3115 3253 2591  886
```

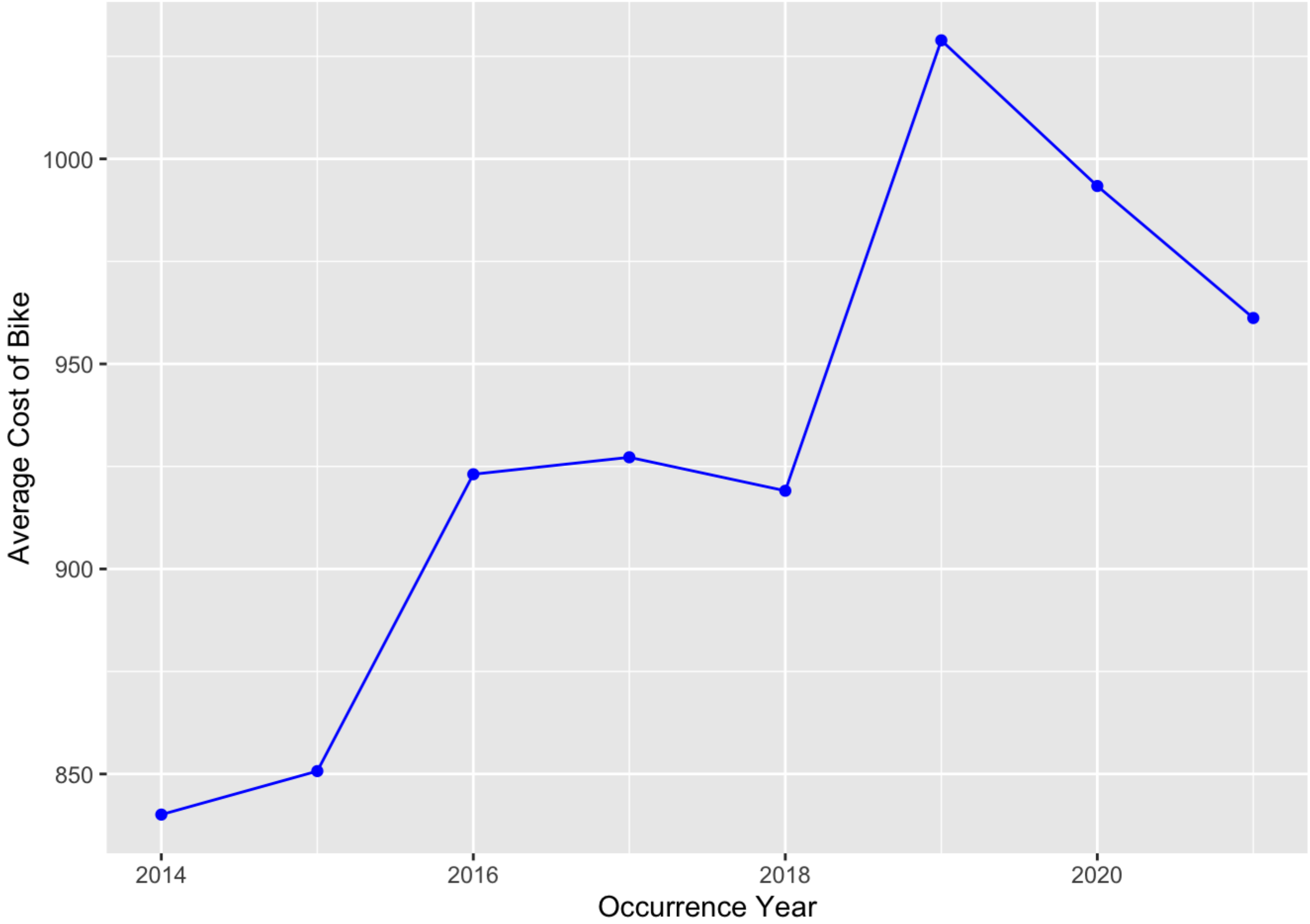
We can see that there are some outliers from 2009-2013. In addition, the 2022 data is not complete. Let us filter them out and obtain a complete data, the one we will be using for our report below.

Table 3: Data After Filtering Incomplete Years

```
##   Year Theft_Count
## 1 2014        2475
## 2 2015        2611
## 3 2016        3132
## 4 2017        3253
## 5 2018        3413
## 6 2019        3115
## 7 2020        3253
## 8 2021        2591
```

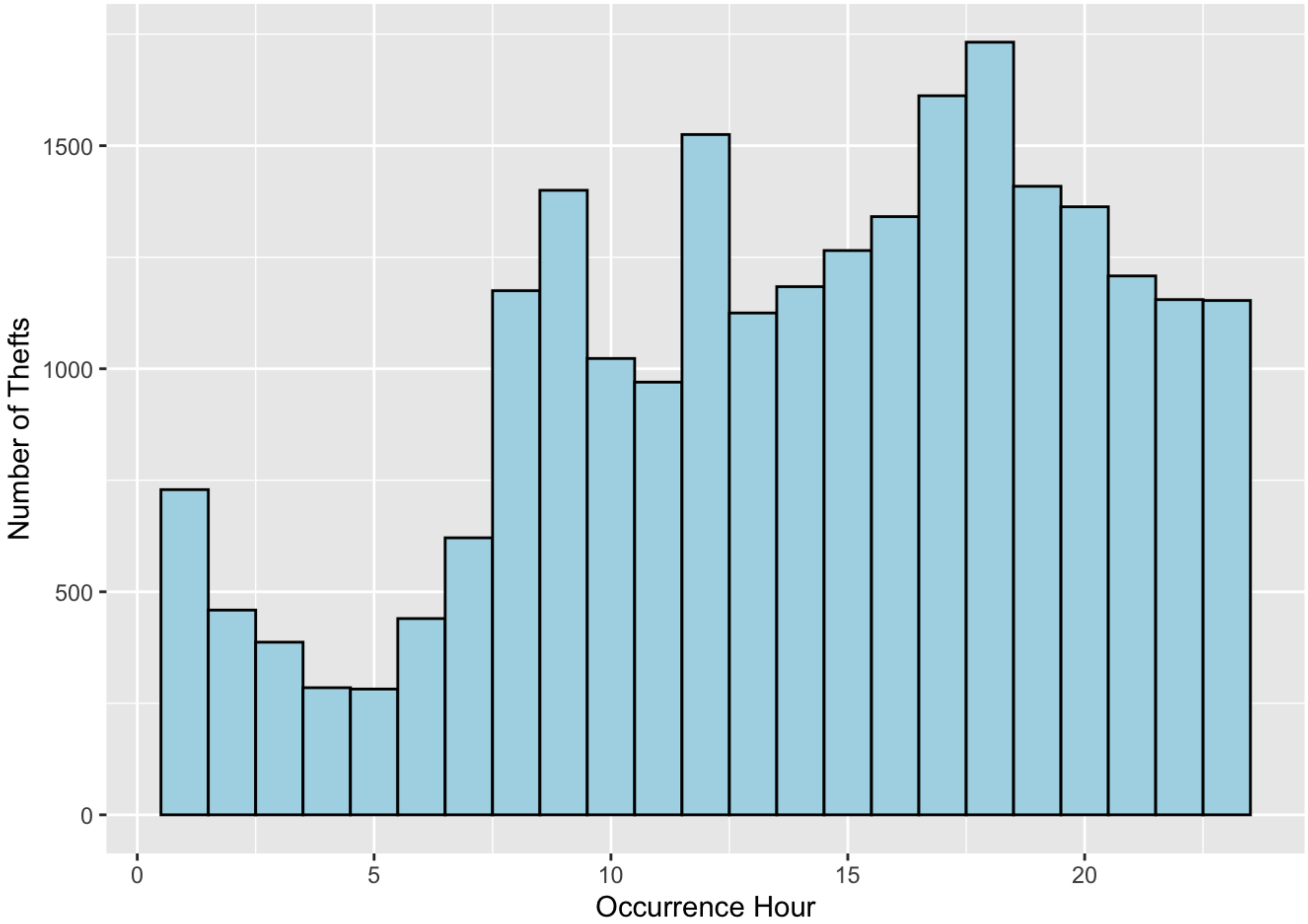
For our first observation, let us use grouping, summarizing to create a simple plot

Graph 1: Average Cost of Stolen Bikes by Occurrence Year



From the graph, we can observe that there was a upward trend of the bike pricing until the 2020. This could suggest a decrease in bike quality stolen since the pandemic started.
For our next observation, let us utilize the hour variable given to determine the high traffic time for bike thefts.

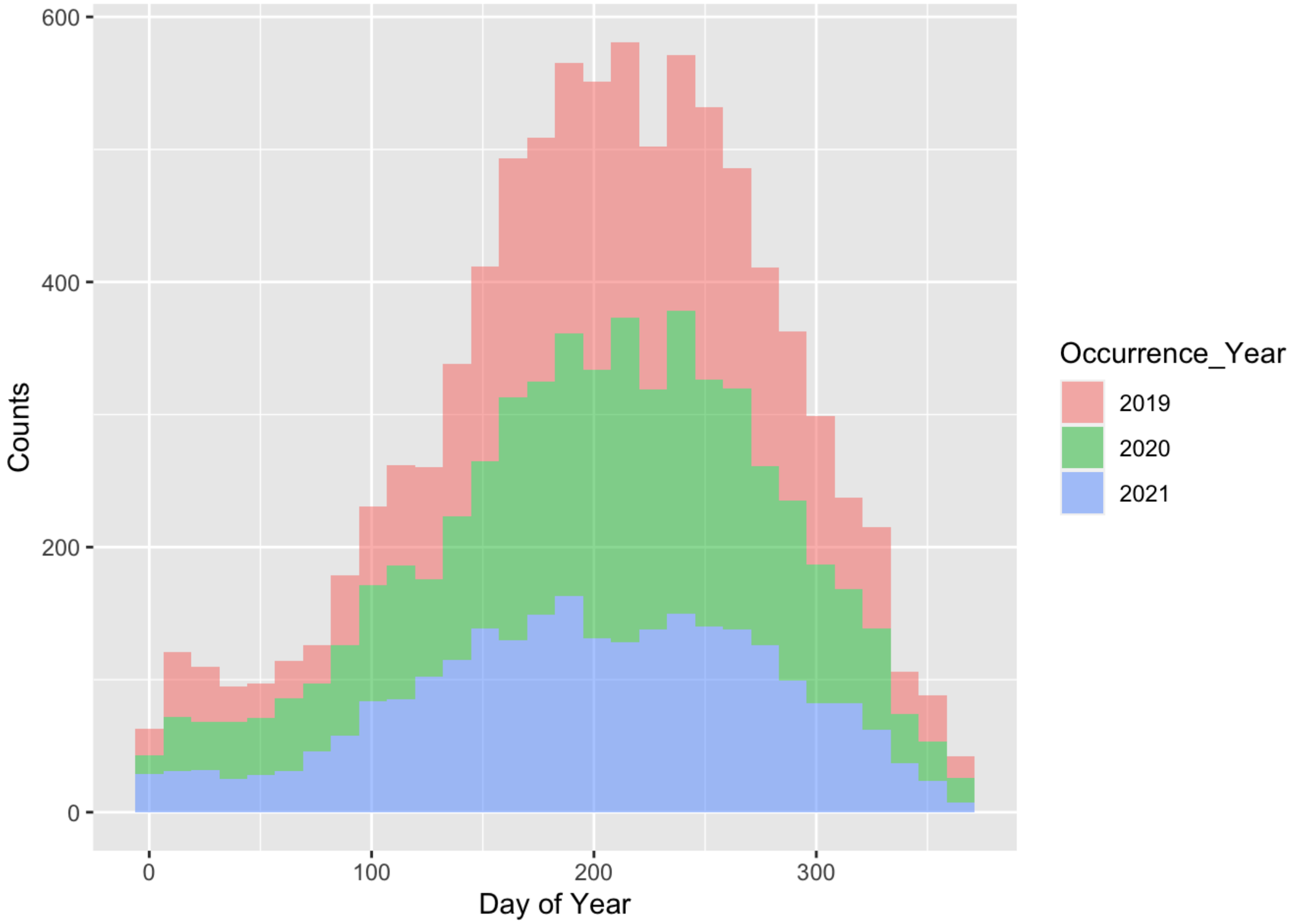
Graph 2: Number of Thefts During each Hour



Observing this histogram, we can see that after midnight, there is a significant decrease in thefts. The theft occurrences spikes again after 7am before peaking at 6pm. This observation make sense as most citizens secure their bikes during nighttime, while incidents are more likely to occur during high traffic times in the morning and evening.

For our next observation, let us create a shaded histogram to see if we can observe anything related to day of the year and bike thefts For this graph, we have decided to take the 2019-2021 as it is recent and having more year would overcrowd the graph.

Graph 3: Bike Thefts Count in 2019-2021 by day of year



We can deduce that overall, bike thefts count start to spike starting day 100, which is roughly starting April. It peaks during day 200 or so, which is roughly the midst of summer. This observation makes sense as more people are likely to take out their bikes and bike during the warmer season compared to the winter. In addition, we see a decrease in bike thefts after 2019. This could be the reason to our downward trend in Figure 3, where the lower average prices is resulted by this decrease in theft count.

It is also interesting for us to see what type of bike are the Toronto population riding. Let us create a summary table to find out what is the most popular bike colour and make in 2021.

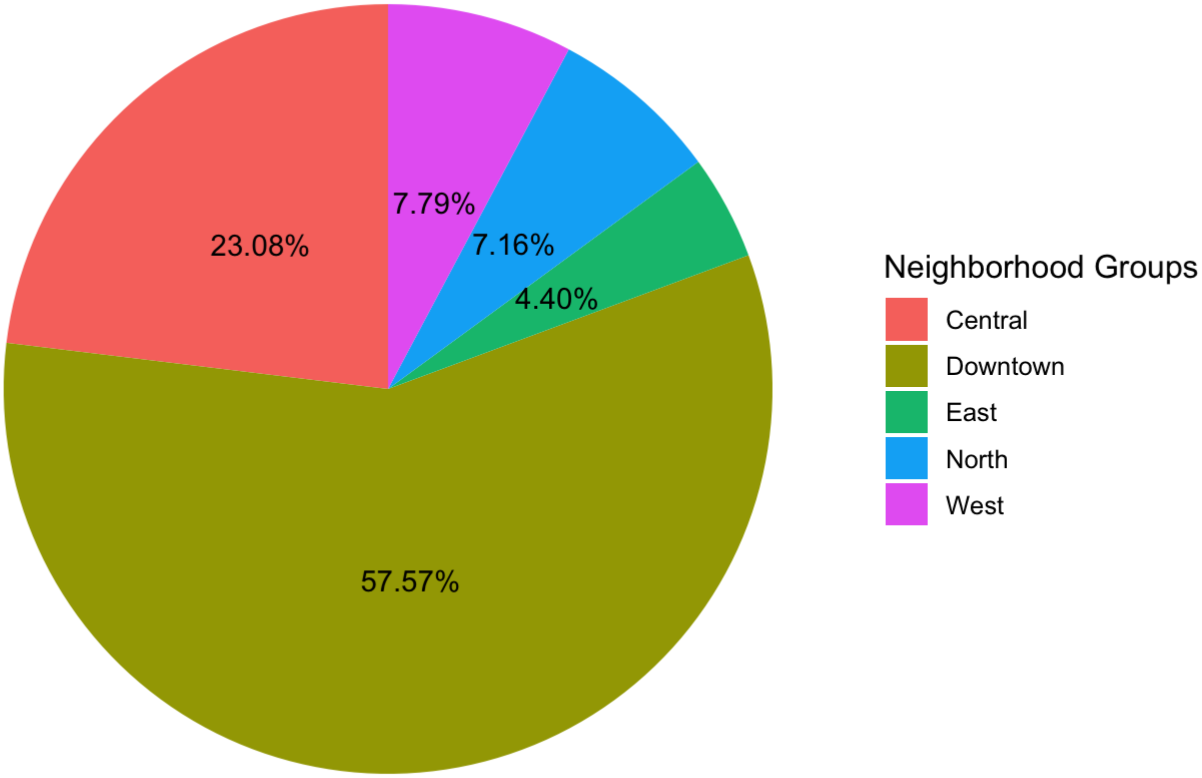
Table 4: Bike Make and Colour of stolen bikes in Toronto 2021

```
##    Occurrence_Year       Bike_Make Bike_Colour total
## 1             2021 OT                   BLK      186
## 2             2021 UK                   BLK      135
## 3             2021 TR                   BLK       61
## 4             2021 OT                   GRY       58
## 5             2021 UK                             55
## 6             2021 OT                   BLU       48
## 7             2021 OT                   WHI       46
## 8             2021 GI                   BLK       45
## 9             2021 OT                             37
## 10            2021 GI                   GRY       34
```

From the above table, we can notice that the most frequent type of bike stolen in 2021 was a black model OT. We can also observe that, in general, Toronto prefer black bikes above the rest.

We can identify whether a hood is in the East, West, North, Central or Downtown by looking at the HOOD_ID Let us create a graph of percentages.

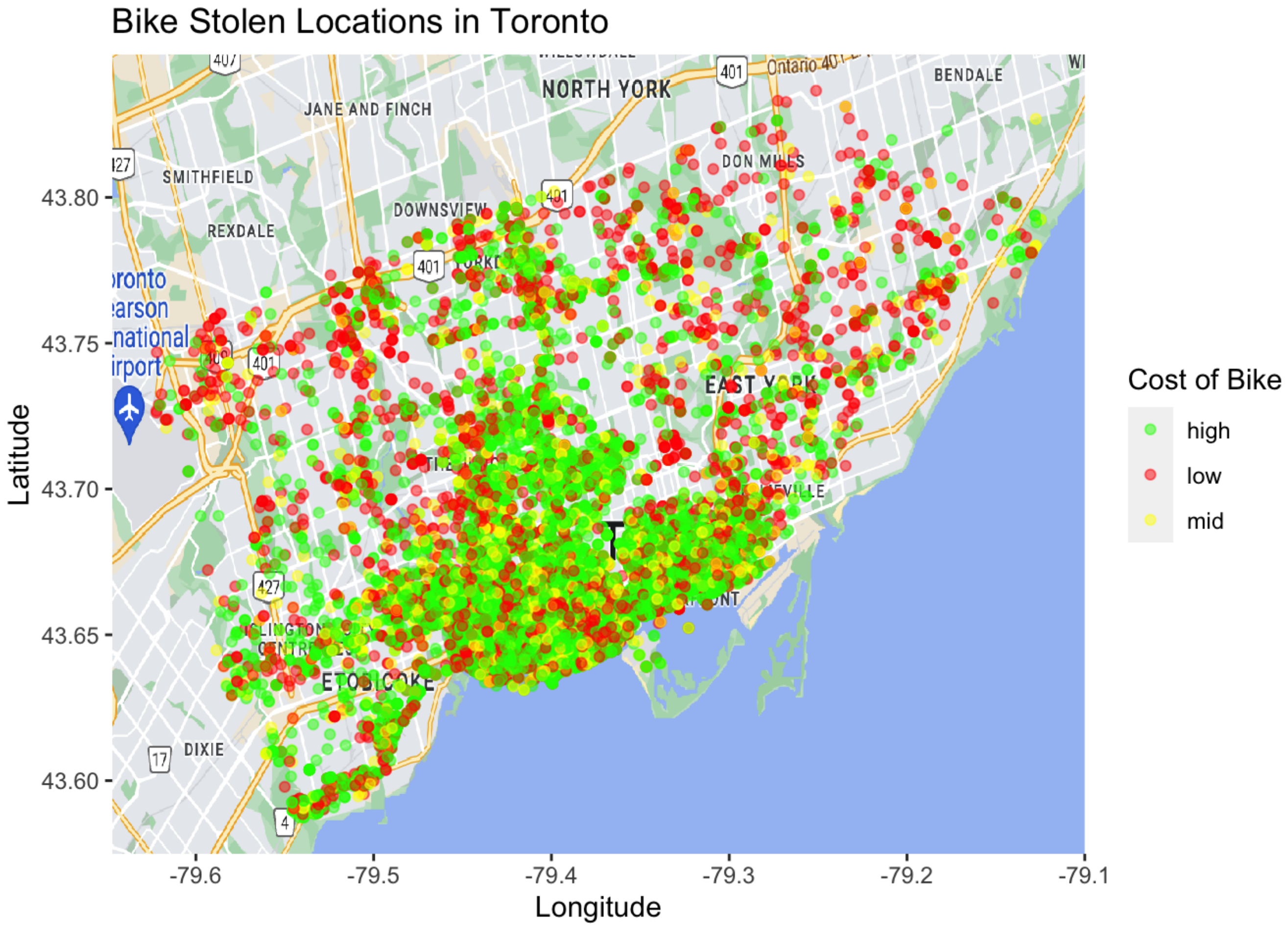
Figure 1: Neighborhood Group Counts



Notice from this pie chart graph that the majority of bike thefts occur downtown. These sections are divided by the different hood id's where we can see on the open Toronto website that each neighborhood is given an numerical id from 1 to 140. By separating into 5 sections we get different sections to analyze. The result is expected since downtown is the most busy place in Toronto and has a bigger concentration of bikers on a daily basis.

Let's map out the occurrences of such bike thefts onto a scatter plot to visualize the span of the thefts across Toronto.

Figure 2: Bike Stolen Locations in Toronto



Once again, we can see mass overlapping at downtown area. Also, note that due to the inability to access the google API, the map background is implemented by hand, therefore it was offset by a bit.

Let’s now take a sample of 1000 observations from the dataset and construct a 97% confidence interval for the proportion of bikes stolen on a Monday. Let’s also construct a 97% Bootstrap proportion of bikes stolen on a Monday:

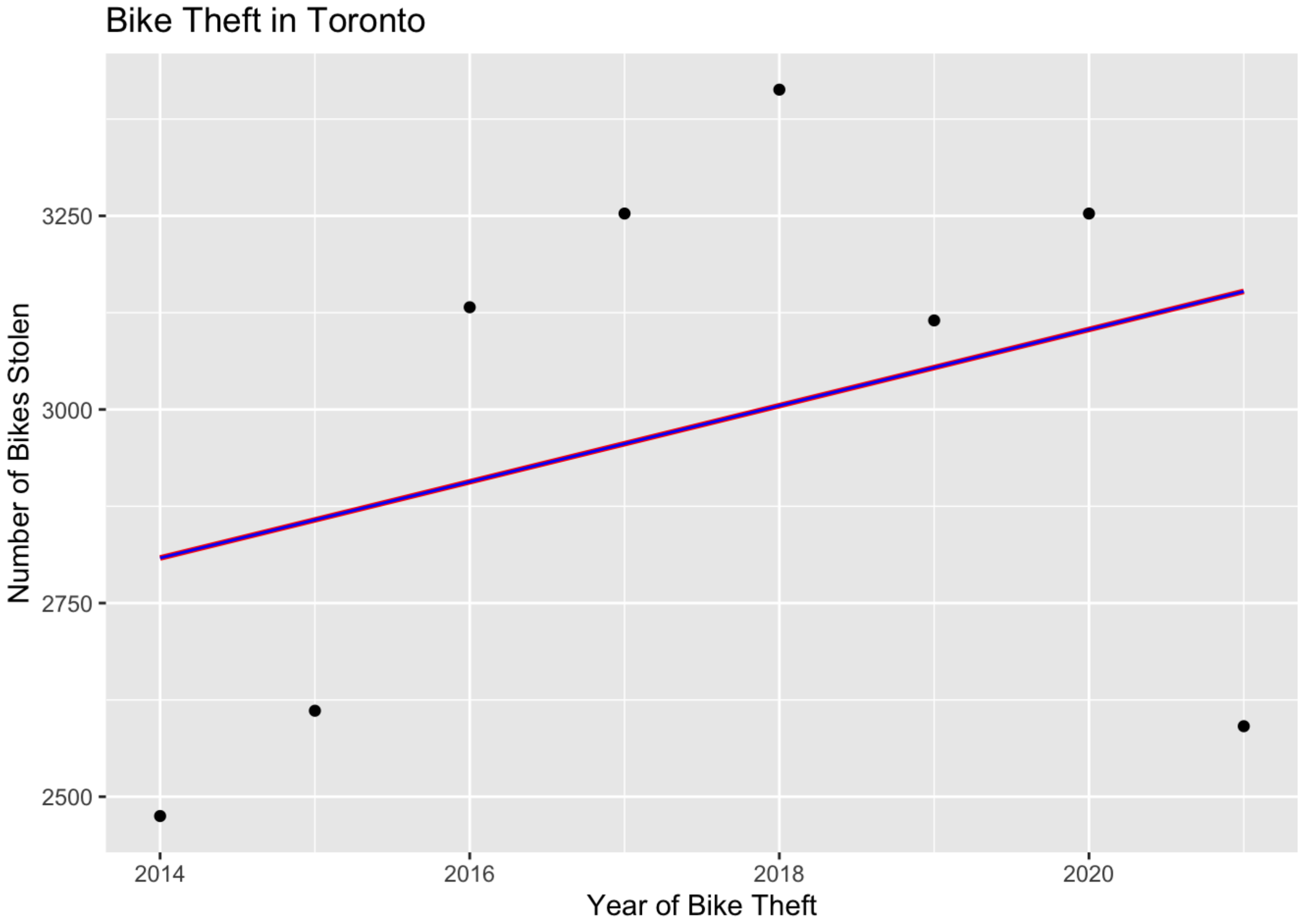
```
## 
##  1-sample proportions test with continuity correction
## 
## data:  sum(samp$Occurrence_DayOfWeek == "Monday") out of length(samp$Occurrence_DayOf
Week), null probability 0.5
## X-squared = 466.49, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 97 percent confidence interval:
##  0.1341145 0.1851553
## sample estimates:
##     p 
## 0.158
```

```
##   10%   90% 
## 0.143 0.172
```

From this, we know that the proportion of bikes stolen on a Monday in the sample of 1000 observations is estimated to be 0.146, with a 97% confidence interval of 0.1229641 to 0.1724226. Using bootstrapping, the 10th percentile and 90th percentile of the bootstrapped proportions of bikes stolen on a Monday are estimated to be 0.132 and 0.160, respectively. This means there is no explicit relationship between thefts occurring on Mondays implying thiefs don’t specifically target Mondays to steal bikes.

```
##
## Call:
## lm(formula = Theft_Count ~ Year, data = training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -561.5 -268.1  105.3  243.3  408.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -96237.39  115001.47  -0.837    0.435
## Year            49.18      57.00   0.863    0.421
##
## Residual standard error: 369.4 on 6 degrees of freedom
## Multiple R-squared:  0.1104, Adjusted R-squared:  -0.03791
## F-statistic: 0.7443 on 1 and 6 DF,  p-value: 0.4214
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :
## There were missing values in resampled performance measures.
```



In this code, we are analyzing bike theft data in Toronto using linear regression and cross-validation techniques. We created a scatterplot of the data and overlay the line of best fit for the training and testing data. Notice that they overlap, hence being a good predictive model. The linear regression model is validated using cross-validation, which helps us to assess its accuracy and prevent overfitting.

Summary

Upon analysis of our dataset producing intriguing results with regression analysis, graphs, tables, bootstrapping, etc., we show that for the past few years in Toronto, bike theft is seen to be most common during summertime afternoons and in particular downtown in location. In addition, our analysis shows that there is both a decrease in the cost of the bikes stolen and in occurrences of such theft in general since 2020, with the latest data recorded being 2021. Furthermore, from the description of the bikes, we notice that the most common bike stolen in Toronto came to be a black OT bike model. Our purpose, as stated at the beginning of our report, includes the intent to inform the public of bike safety concerns in Toronto and surrounding areas. From our produced results we see that certain areas have higher probability of bike thefts to occur than others. Downtown Toronto, as seen in our scatterplot graph of the occurrences of bike theft in the area, shows that it is a place where the public should be extra wary of their bikes’ safety. Our findings hope to inform the public and also help the Toronto Police and government in enacting commitment to help solve such problems, as the number of stolen bikes every year, despite decreasing, is still not a small number.

Appendix

Table 2

```
data <- na.omit(origin)
summary <- table(data$Occurrence_Year)
print(summary)
```

Table 3

```
data = data %>% filter(Occurrence_Year >= 2014) %>%
  filter(Occurrence_Year < 2022)
summary_table = data %>% count(Occurrence_Year) %>%
  rename(Year = Occurrence_Year, Theft_Count = n)
print(summary_table)
```

Graph 1

```
# Calculate average cost of bike by year
avg_cost_by_year <- data %>%
  group_by(Occurrence_Year) %>%
  summarize(avg_cost = mean(Cost_of_Bike))
# Create plot of average cost of bike by year
ggplot(avg_cost_by_year, aes(x = Occurrence_Year, y = avg_cost)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(x = "Occurrence Year", y = "Average Cost of Bike")
```

Graph 2

```
ggplot(data = data, aes(x = Occurrence_Hour)) +
  geom_histogram(binwidth = 1, color = "black", fill = "lightblue") +
  labs(x = "Occurrence Hour", y = "Number of Thefts")
```

Graph 3

```
data2 = data %>% filter(Occurrence_Year %in% c(2019,2020,2021))
ggplot(data2, aes( x=Report_DayOfYear)) +
  geom_histogram(aes(fill=factor(Occurrence_Year)),alpha=0.5)+
  labs(fill="Occurrence_Year") +
  labs(x = "Day of Year", y = "Counts")
```

Table 4

```
data3 = data %>%
  filter(Occurrence_Year == c(2021)) %>%
  group_by(Occurrence_Year, Bike_Make, Bike_Colour) %>%
  summarise(total = sum(n=n())) %>%
  arrange(desc(total))
as.data.frame(data3) %>% head(10)
```

Figure 1

```
new_table <- table(data$NeighbourhoodName)
# create data frame with neighborhood names and counts
new_df <- data.frame(nh = names(new_table), count = as.vector(new_table), hood_id = uniq
ue(data$Hood_ID[match(names(new_table), data$NeighbourhoodName)]))
new_df$hood_id <- as.numeric(new_df$hood_id)
new_df <- new_df %>% mutate(nh_groups = case_when(new_df$hood_id >= 112 ~ 'East',
                                                  (new_df$hood_id<112 & new_df$hood_id>=
84) ~ 'Central',
                                                  (new_df$hood_id<84 & new_df$hood_id>=5
6)~'Downtown',
                                                  (new_df$hood_id<56 & new_df$hood_id>=2
8)~'North',
                                                  TRUE~'West'))
grouped_data <- new_df %>%
  group_by(nh_groups) %>%
  summarize(count = sum(count)) %>%
  mutate(percentage = percent(count / sum(count)))
ggplot(grouped_data, aes(x = "", y = count, fill = nh_groups)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  labs(fill = "Neighborhood Groups") +
  geom_text(aes(label = percentage), position = position_stack(vjust = 0.5)) +
  scale_y_continuous(labels = percent_format()) +
  theme_void() +
  theme(text = element_text(size = 12))
```

Figure 2

```
# Extracting longitude and latitude using regular expressions
data <- data %>%
  mutate(longitude = as.numeric(sub(".*\\((-?\\d+\\.\\d+),.*", "\\1", geometry)),
         latitude = as.numeric(sub(".*,(\\s?-?\\d+\\.\\d+)\\)}", "\\1", geometry))) %>%
  filter(longitude != 0 & latitude != 0 &
         longitude > -79.7 & longitude < -79.1) %>%
  mutate(cost_lost = case_when(Cost_of_Bike>=600~"high",
                               (Cost_of_Bike<600 & Cost_of_Bike>=400)~"mid",
                               TRUE ~"low"))
img <- readPNG("toronto2.png")
ggplot(data, aes(x = longitude, y = latitude, color = cost_lost)) +
  background_image(img) +
  geom_point(alpha = 0.5) +
  scale_color_manual(values = c("low" = "red", "mid" = "yellow", "high" = "green"),
                     name = "Cost of Bike") +
  ggtitle("Bike Stolen Locations in Toronto") +
  xlab("Longitude") +
  ylab("Latitude")
```

Confidence Interval + Bootstrapping

```
samp = data %>% sample_n(1000)
prop.test(x = sum(samp$Occurrence_DayOfWeek == "Monday"), n = length(samp$Occurrence_Day
OfWeek), conf.level = 0.97)
boot_function = function(){
  boot_data = samp  %>% sample_n(nrow(samp), replace = T)
  boot_prop = mean(boot_data$Occurrence_DayOfWeek == "Monday")
  return(boot_prop)
}
quantile(replicate(1000, boot_function()), c(0.1, 0.9))
```

Linear Regression + Cross Validation

```
set.seed(99)
trainIndex <- createDataPartition(summary_table$Theft_Count, p = .8, list = FALSE)
training <- summary_table[trainIndex,]
testing <- summary_table[-trainIndex,]
# Fit a linear regression model to the training data
lm_model <- lm(Theft_Count ~ Year, data = training)
summary(lm_model)
# Fit the linear regression model using 10-fold cross-validation
ctrl <- trainControl(method = "cv", number = 10)
lmFit <- train(Theft_Count ~ Year, data = training, method = "lm",
               trControl = ctrl)
# Predict values for the training and test sets
training$predicted <- predict(lm_model, newdata = training)
testing$predicted <- predict(lm_model, newdata = testing)
# Plot the results with the line of best fit
ggplot(training, aes(x = Year, y = Theft_Count)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red", alpha=0.1) +
  labs(title = "Bike Theft in Toronto",
       x = "Year of Bike Theft",
       y = "Number of Bikes Stolen") +
  geom_line(aes(x = Year, y = predicted), data = training, color = "blue") +
  geom_line(aes(x = Year, y = predicted), data = testing, color = "green")
```