STAA57 Project

Factors That Lead to Bikes Being Stolen Outside

Ze Jun Guan

Departmetn of Statistics
University of Toronto Scarborough
Canada
April 9, 2023

Introduction

The data I used in this project is called **bicycle thefts**, it's a data set published by the Toronto Police Services. It can be found from ___https://open.toronto.ca/___. This data set contains occurrences of bicycle thefts from 2014 to 2022 and it details the time and place at which the bicycle was stolen. The research question I am trying to answer in this report are "What is the probability of a bicycle being stolen outside?",and "Does this probability change across years?"By outside, it means whether or not your bicycle is located at your living place.As long as the bicycle is not stolen from house or apartment, it is considered to be outside.

Data Description

After cleaning the data by removing observations with NAs and removing variables that are not interested in. The number of observations in the data set has decreased from 30154 to 24653, and the number of variables has decreased from 31 to 15. The following table describes the variables included in our final data set:

Variable_Names	Description
Occurrence_Year	Year of Occurrence
Occurrence_Month	Month of Occurrence
Occurrence_DayOfWeek	Day of week theft occurred
Occurrence_Hour	Hour theft occurred
Division	Police Division where event occurred
Hood_ID	City of Toronto Neighbourhood identifier
Premises_Type	Premises type of occurrence
Bike_Type	Bicycle Type
Bike_Speed	Bicycle Speed
Colour	Colour of bicycle
Cost_of_Bike	cost of bicycle
Status	Statue of event
Outside	1 if bicycle was stolen outside from living space, 0 otherwise
house	1 if bicycle was stolen from house, 0 otherwise
Apartment	1 if bicycle was stolen from apartment, 0 otherwise

Tables and Graphs

```
## `geom_smooth()` using formula = 'y ~ x'
```

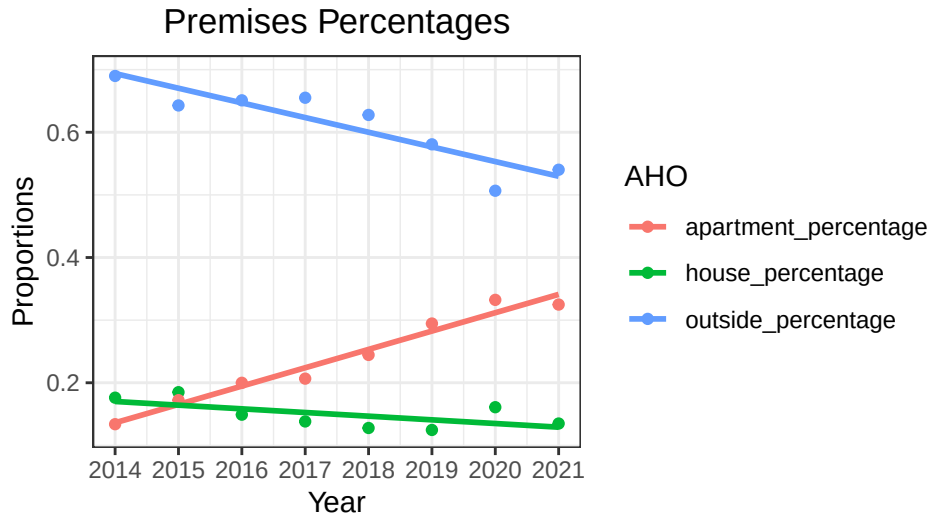

Figure 1

Figure 1 consists of three trends. The blue line represents the percentage of bicycles stolen from outside each year. The red line represents the percentage of bicycles stolen from apartments each year. The green line represents the percentage of bicycles stolen from houses each year. As we can see from Figure 1, proportion for bicycles stolen outside or from houses has a overall decreasing trend, but proportion for bicycles stolen from apartments are rapidly increasing.

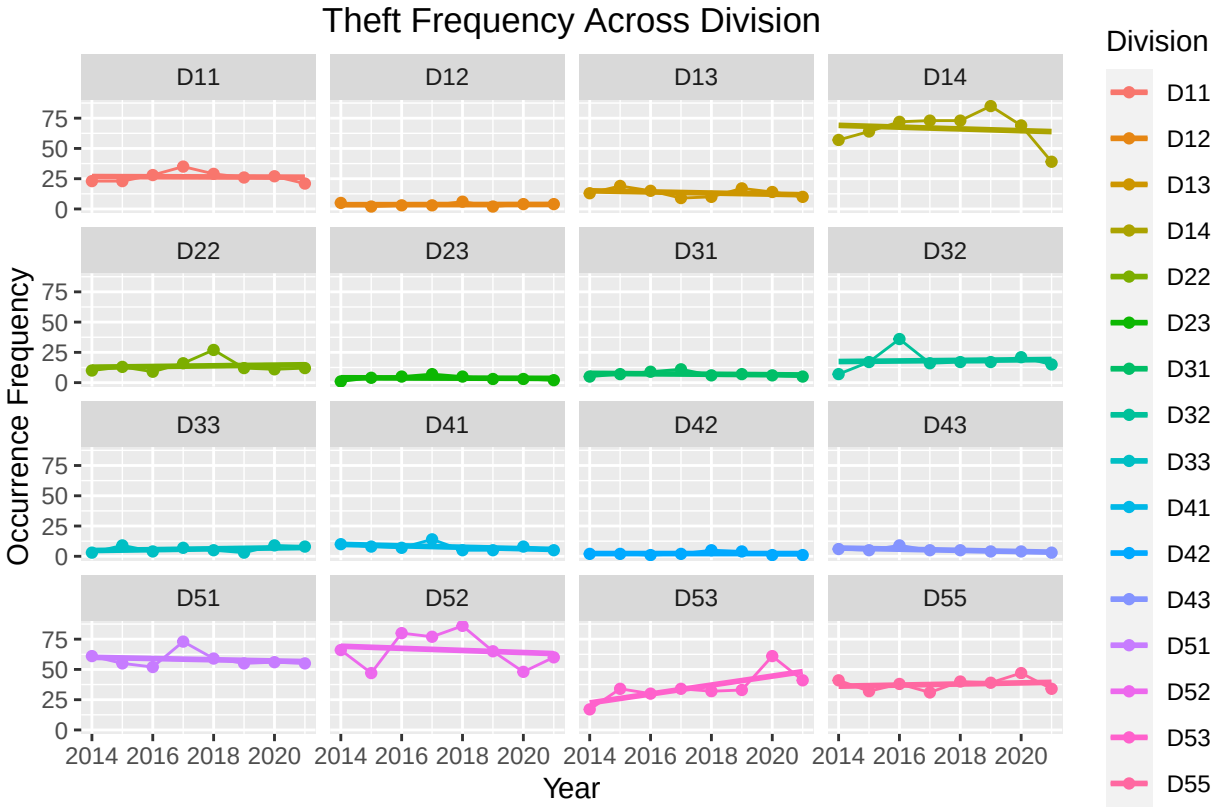

Figure 2

There's always some areas in the city with higher crime rate than others. In Figure 2,a graph of theft frequency was plotted across division. It's obvious that for most divisions, frequency of theft were under 20

across years.However, divisions **D14**, **D51**, **D52** , **D53** and **D54** appears to have high volume of thefts. The linear trend for these divisions might be decreasing at first, but if you take a close look, there's a sudden drop in frequency from 2020 to 2022. These drops leads the trend downward, and this could be due to the outbreak of COVID-19.

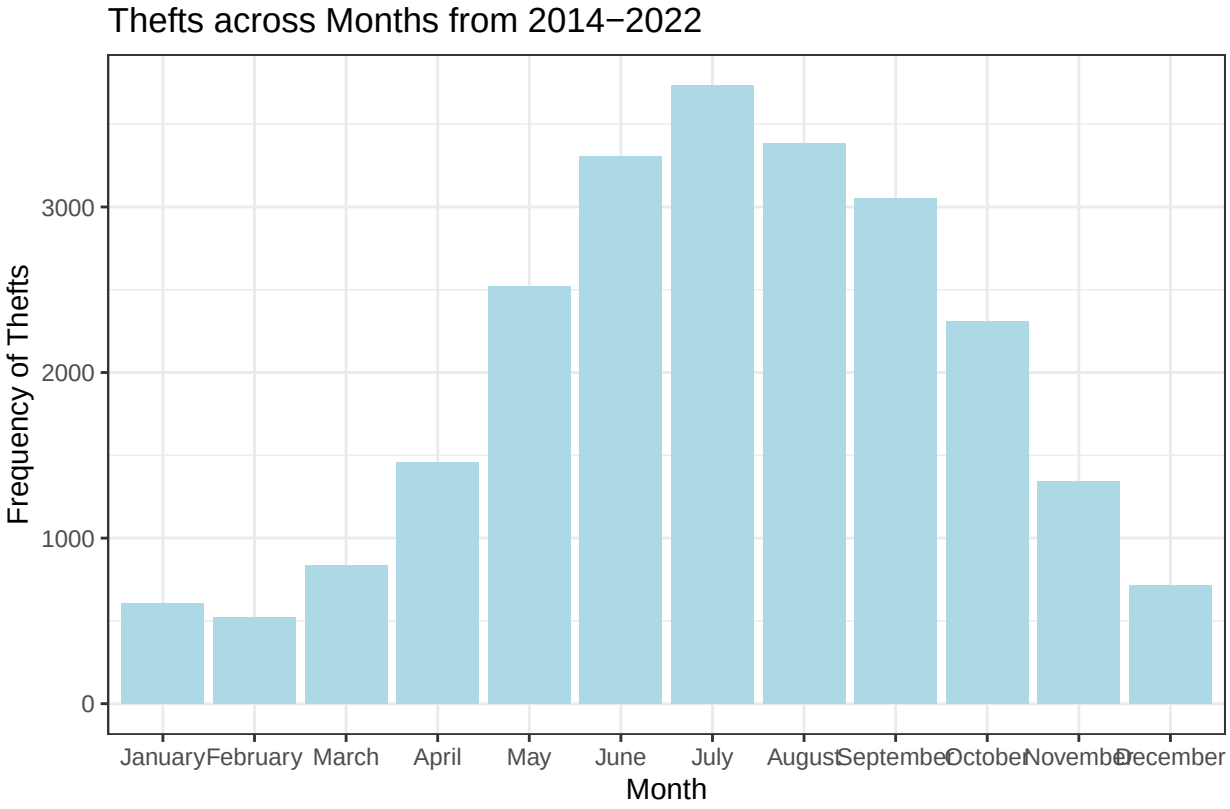

Figure 3

Theft happens all the time and it never stops, but do they get a break just like regular employees. Figure 3 may not answer the whole question, but it seems like some thiefs do take a break. Figure 3 is a bar graph of the frequency of thefts across months from 2014-2022. Starting from January, frequency gradually increases until it reaches the peak in July, and the frequency will taper off after July. Majority of the counts occurs between April and October.

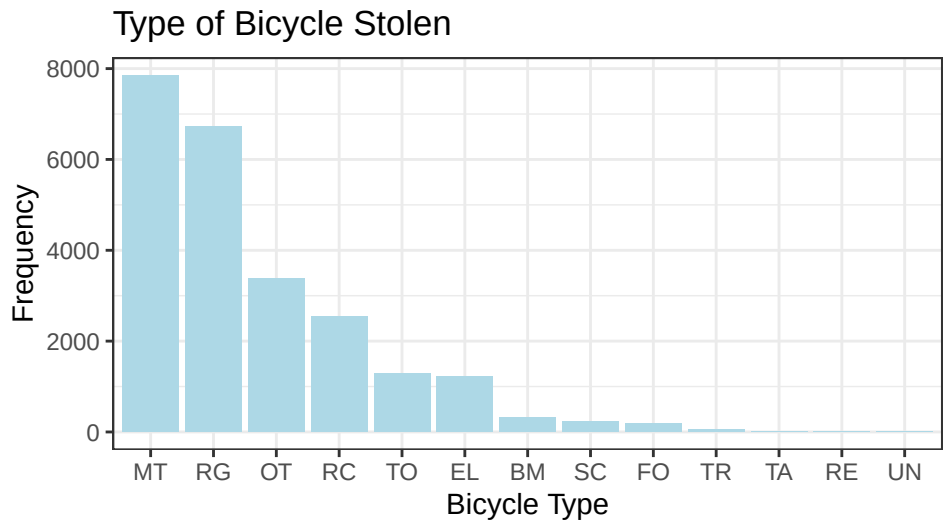

Figure 4

Table 2: Type of Bicycles Stolen(Proportion wise)

Bike_Type	count	prop
MT	7847	0.330
RG	6716	0.282
OT	3376	0.142
RC	2545	0.107
TO	1280	0.054
EL	1211	0.051
BM	311	0.013
SC	236	0.010
FO	178	0.007
TR	52	0.002
TA	20	0.001
RE	11	0.000
UN	8	0.000

Figure 4 and Table 1 tells the type of bicycles thiefs in favor of. More than 80% of bicycles stolen came from type **MT**, **RG**, **OT**, and **RC**.

Hypothesis Testing

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x[c(1, 8)] out of n[c(1, 8)]
## X-squared = 121.44, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1273158 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.6899314 0.5401662
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  y[c(1, 8)] out of n[c(1, 8)]
## X-squared = 16.329, df = 1, p-value = 2.662e-05
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.0242943 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.1762014 0.1349426
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  z[c(1, 8)] out of n[c(1, 8)]
## X-squared = 266, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.1718078
## sample estimates:
##    prop 1    prop 2
## 0.1338673 0.3248912
```

From Figure 1, we knew that the percentage for bicycles stolen outside and from house has a decreasing trend. Inversely, bicycles stolen from apartments has an increasing percentage. In the above three hypothesis testings, I have set 2014 as the base year, and compared it with the proportion of stolen bicycles in 2022. Although we have already knew the trend, but more importantly, we would like to know if there are a statistical significant change in proportions? From the result of the first two tests, the null hypothesis are $H_0 : \pi_{2014,outside} = \pi_{2022,outside}$ and $H_0 : \pi_{2014,house} = \pi_{2022,house}$ respectively. The alternative hypothesis are $H_1 : \pi_{2014,outside} > \pi_{2022,outside}$ and $H_0 : \pi_{2014,house} > \pi_{2022,house}$. Both results have p-value much less than 0.05, and 95% confidence intervals that does not include zeroes. We have strong evidence to say that the proportion in 2014 for houses/outside are greater than the proportion in 2022. There's a decrease in proportion and it's statistically significant. The last test has $H_0 : \pi_{2014,apartmetn} = \pi_{2022,apartment}$, and alternative hypothesis $H_0 : \pi_{2014,apartment} < \pi_{2022,apartment}$. The result of the test has a p-value less than $2.2e-16$ and a 95% confidence interval from [**-1,-0.1718078**]. We have enough evidence to say that proportion in 2014 is significantly smaller compared to the proportion in 2022 for bicycles stolen from apartments.

Regression

```
##
## Call:
## glm(formula = outside ~ Occurrence_Month + Occurrence_Hour +
##     Division + Bike_Type + Bike_Speed + Cost_of_Bike, family = "binomial",
##     data = bicycle2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2673  -1.2226   0.6295   1.0440   2.8164
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -7.459e-02  8.877e-02  -0.840 0.400758
## Occurrence_MonthAugust     2.397e-01  6.792e-02   3.529 0.000417 ***
```

```
## Occurrence_MonthDecember  -2.534e-01  9.842e-02  -2.575 0.010020 *
## Occurrence_MonthFebruary  -9.831e-02  1.089e-01  -0.902 0.366859
## Occurrence_MonthJanuary   -1.277e-01  1.046e-01  -1.220 0.222480
## Occurrence_MonthJuly       2.421e-01  6.691e-02   3.619 0.000296 ***
## Occurrence_MonthJune       2.112e-01  6.819e-02   3.097 0.001953 **
## Occurrence_MonthMarch      4.190e-02  9.331e-02   0.449 0.653385
## Occurrence_MonthMay       -5.411e-02  7.099e-02  -0.762 0.445941
## Occurrence_MonthNovember  -3.345e-02  8.147e-02  -0.411 0.681386
## Occurrence_MonthOctober    2.259e-01  7.290e-02   3.099 0.001944 **
## Occurrence_MonthSeptember  2.620e-01  6.899e-02   3.797 0.000146 ***
## Occurrence_Hour           -4.629e-03  2.406e-03  -1.924 0.054403 .
## DivisionD12                4.636e-01  1.564e-01   2.964 0.003036 **
## DivisionD13               -2.994e-01  8.799e-02  -3.403 0.000666 ***
## DivisionD14                9.606e-02  5.972e-02   1.608 0.107736
## DivisionD22                3.438e-02  8.257e-02   0.416 0.677173
## DivisionD23                2.612e-01  1.566e-01   1.668 0.095345 .
## DivisionD31                2.127e-01  1.245e-01   1.709 0.087516 .
## DivisionD32                2.303e-01  7.973e-02   2.888 0.003877 **
## DivisionD33                1.961e-01  1.188e-01   1.651 0.098741 .
## DivisionD41                2.461e-01  1.191e-01   2.066 0.038826 *
## DivisionD42                3.974e-01  1.599e-01   2.486 0.012923 *
## DivisionD43                6.899e-01  1.324e-01   5.211 1.88e-07 ***
## DivisionD51                6.151e-01  6.166e-02   9.976  < 2e-16 ***
## DivisionD52                1.603e+00  6.635e-02  24.161  < 2e-16 ***
## DivisionD53                3.753e-01  6.769e-02   5.544 2.95e-08 ***
## DivisionD55                1.417e-01  6.555e-02   2.162 0.030630 *
## Bike_TypeRG               -8.477e-02  3.717e-02  -2.281 0.022569 *
## Bike_TypeOT                1.485e-01  4.623e-02   3.213 0.001315 **
## Bike_TypeRC               -1.336e-01  5.072e-02  -2.634 0.008433 **
## Bike_TypeTO               -9.402e-03  6.543e-02  -0.144 0.885735
## Bike_TypeEL                7.029e-01  7.643e-02   9.198  < 2e-16 ***
## Bike_TypeBM                1.037e-01  1.288e-01   0.805 0.420861
## Bike_TypeSC                4.871e-01  1.537e-01   3.169 0.001529 **
## Bike_TypeFO                2.335e-01  1.727e-01   1.352 0.176451
## Bike_TypeTR               -5.748e-01  2.974e-01  -1.933 0.053243 .
## Bike_TypeTA                3.419e-01  4.981e-01   0.687 0.492389
## Bike_TypeRE                4.757e-01  7.143e-01   0.666 0.505450
## Bike_TypeUN               -1.185e-01  7.100e-01  -0.167 0.867437
## Bike_Speed                 2.883e-03  1.513e-03   1.906 0.056671 .
## Cost_of_Bike              -4.233e-05  1.206e-05  -3.511 0.000447 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30197  on 22605  degrees of freedom
## Residual deviance: 28534  on 22564  degrees of freedom
##   (1185 observations deleted due to missingness)
## AIC: 28618
##
## Number of Fisher Scoring iterations: 4
```

$$logit(p) = \beta_0 + \beta_1 x_{Occurrence\ Month} + \beta_2 x_{Occurrence\ Hour} + \beta_3 x_{Division} + \beta_4 x_{Bike\ Type} + \beta_5 x_{Bike\ Speed} + \beta_6 x_{Cost\ of\ Bike}$$

Above is the summary of the table.The baseline category is at April, Occurrence_Hour equal to 1 and Division equal to D11. Variable Occurrence_Month, Occurrence_Hour, Division and Bike_Type are significant variables to the model, although some of the categories within each variable aren't important. For example, June-October, and December are important categories and other months are not. Which means these are the months that thieves in favor of. Likewise, time between 6-23 are also important. Divisions with high frequency shown in Figure 2 are significant to the model as well. One thing to notice about is that the cost of the bike has a negative relationship with the odds ratio. This means the more expensive your bike is, it's less likely to be stolen from outside. Interesting fact that it is the opposite of what we generally think.

For categorical variable Occurrence Month, the regression coefficient is equal to 0.2709 in August. This means that the odds of having a bicycle stolen from outside is $e^{0.2709} = 1.31$ or 31% higher than stolen from houses or apartments. Interpretations are same for other categories. For quantitative variable like **Cost of Bike**, it has regression coefficient equal to -0.00003857. This means that holding everything else constant, one dollar increase in the price of the bicycle, the odds of having it stolen from outside decrease by $1 - e^{-0.00003857} = 0.000038569$ or 0.0038569%.

```
## [1] 0.6508204 0.6379107 0.6498619 0.6348020
```

```
## [1] 0.6433488
```

The result shown above are the concordance index c or AUC value after cross validation.This test shows the trade-off between sensitivity and specificity. The concordanceindex c, which estimates the probability that the predictions and the outcomes are concordant.The results showed that our model has a average concordance index of 0.644, which means the probability of predictions and the outcomes are concordant is 64.4%. This value isn't high, but based on the limited data we have, I would say the concordance index is reasonable for our model.

```
##        5%       95%
## 0.6865998 0.7037271
```

It's interesting to know whether or not the probability of having your bicycle stolen outside has decreased across years. Using the same model as above, I bootstrapped the observation from 2014 and found the average probability between each trial. From those averages, I calculated the 95% confidence interval. The result is $[0.6864277, 0.7038608]$. This means that if you have your bike parked outside, the chance of it being stolen is between 68.6% to 70.4%. Which is terrific.

```
##        5%       95%
## 0.5332921 0.5491111
```

For simplicity, I only compared the 95% confidence interval between 2014 and 2021. The 95% confidence interval for 2021 is $[0.5332302, 0.5488628]$. A huge difference compare to the confidence interval in 2014.

Conclusion

From the results in the regression model, we see that Month, bike type, division,cost of bike are important factors that influence our model. Other variables like **hour**(the time the bike was stolen) and bike speed has p-values close to but greater than 0.05. This indicates that these two variables may also be important and we shall include them in the model. Just like how I predicted in the graphs, thieves like to steal during certain months and there are certain types of bikes that they like to steal. From our graphs, it showed that theft peak happens between July to October, and July to October are the statistically significant months in our model. Theieves also like to steal from some specific areas, as some police division are statistically significant and some are not. Most interesting fact is that bike with higher cost are less likely to be stolen, perhaps this relates to some of the laws in Canada. Different value of stolen items may lead to different penalties if you get caught, this might be the reason why expensive bikes are less likely to be stolen.

Another research question we'd like to answer is "Did the probability decrease over time?" From the bootstrap function, the 95% confidence interval for 2014 and 2021 has difference of 15~16%. Therefore, we are proud to

say that the probability has gone down, and it gone done by a lot. Although I didn't do testings among these two confidence intervals, but I think it's obvious that they're statistically different.

Limitations and Improvements

Like I said earlier, our model is limited and which is why I say a concordance index of 0.644 is reasonable and not bad. Although we have data of the bikes and the time it occurred, but the causes of crime are composed of many aspect. For example, education, religion, family condition, economic condition of the area etc. More stereotypical speaking, race is also an important factor. Hence, if we would like to improve the accuracy of the model, we should collect all kinds of different data.

Appendix

```
library(tidyverse)
library(knitr)
```

```
# read data
bicycle <- read.csv("bicycle-thefts - 4326.csv")
```

```
#remove variables that are not in interest
bicycle %>% select(Occurrence_Year:Occurrence_DayOfWeek,Occurrence_Hour,Division,
                Hood_ID,Premises_Type,Bike_Type:Status) ->bicycle2
```

```
#Set the first three letters of Bike_colour as it's colour, otherwise
#there will be too many colors involved
bicycle2 %>% mutate(colour = str_sub(Bike_Colour, start = 1,end = 3)) %>%
  filter(colour != "" | colour == "18") ->bicycle2

# remove NA values
bicycle2 %>% filter(Bike_Speed !=" NA") %>% select(-c(Bike_Colour))->bicycle2
```

```
# remove rows with NAs in Occurrence_hour column
bicycle2 %>% filter(Occurrence_Hour != "NA" ) ->bicycle2
```

```
# remove Division that in unknown
bicycle2 %>% filter(Division != "NSA") ->bicycle2
bicycle2<- bicycle2 %>% filter(Occurrence_Year!=2022) #remove year 2022, does not contain full year dat
```

```
# create dummy variables outside, house and apartment
# if bicycle is stolen outside/house/apartment, it's equal to 1, otherwise 0
bicycle2 %>% mutate(outside = case_when(Premises_Type == "Outside"|Premises_Type=="Other"|
                                          Premises_Type=="Educational"|Premises_Type=="Commercial"|
                                          Premises_Type=="Transit"~1,
                                        TRUE ~ 0 ),
                    house = case_when(Premises_Type =="House"~1,
                                      TRUE ~0),
                    Apartment = case_when(Premises_Type =="Apartment"~1,
                                          TRUE~0)) ->bicycle2
```

```
Description = data.frame(
  "Variable_Names" = c("Occurrence_Year","Occurrence_Month","Occurrence_DayOfWeek",
                       "Occurrence_Hour","Division","Hood_ID","Premises_Type",
                       "Bike_Type","Bike_Speed","Colour","Cost_of_Bike",
                       "Status","Outside","house","Apartment"),
  "Description" = c("Year of Occurrence",
                    "Month of Occurrence",
                    "Day of week theft occurred",
                    "Hour theft occurred",
                    "Police Division where event occurred",
                    "City of Toronto Neighbourhood identifier",
                    "Premises type of occurrence",
                    "Bicycle Type",
                    "Bicycle Speed",
                    "Colour of bicycle",
                    "cost of bicycle",
                    "Statue of event",
                    "1 if bicycle was stolen outside from living space, 0 otherwise",
```

```
                    "1 if bicycle was stolen from house, 0 otherwise",
                    "1 if bicycle was stolen from apartment, 0 otherwise")
)
kable(Description)
```

```
bicycle2 %>% group_by(Occurrence_Year) %>% summarise(outside_percentage = mean(outside),
                                                     house_percentage =mean(house),
                                                     apartment_percentage = mean(Apartment)) %>%
  slice(6:13) -> Type_percentage
Type_percentage %>% pivot_longer(cols = c(outside_percentage,house_percentage,
                                          apartment_percentage), names_to = "AHO") ->Type_percentage
```

```
# percentage of bikes being stolen outside, house, and in apartment
Type_percentage %>% ggplot(aes(x= Occurrence_Year, y= value,colour = AHO))+
                           geom_point()+geom_smooth(method = lm,se = FALSE)+
  scale_x_continuous(breaks = 2014:2021) +theme_bw() +
  labs(x = "Year",y="Proportions",title = "Premises Percentages",caption = "Figure 1")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
# amount of bike stole in each division across years
bicycle2 %>% filter(Occurrence_Year==c(2014,2015,2016,2017,2018,
                                       2019,2020,2021)) %>%
  group_by(Occurrence_Year,Division) %>% summarise(count = n()) %>%
  ggplot(aes(x = Occurrence_Year, y = count, colour = Division))+geom_point()+
  geom_line()+facet_wrap(~Division,ncol = 4) +geom_smooth(method = lm,se=FALSE)+
  labs(x = "Year",y = "Occurrence Frequency", title = "Theft Frequency Across Division",
       caption = "Figure 2")+
  theme(plot.title = element_text(hjust = 0.5))

# amount of bikes stolen in each month
bicycle2  %>%
  ggplot(aes(x=Occurrence_Month)) +geom_bar(fill="light blue" )+scale_x_discrete(limits = month.name)+
  labs(x = "Month", y= "Frequency of Thefts",title = "Thefts across Months from 2014-2022",
       caption = "Figure 3") +theme_bw()

#type of bike thiefs like to steal the most
bicycle2 %>% group_by(Bike_Type) %>% summarise(count =n()) %>%
   mutate(prop = prop.table(count)) ->overall_bike_type
overall_bike_type %>% arrange(desc(prop)) %>% pull(Bike_Type) %>% unique ->bike_type_order
bicycle2$Bike_Type <- factor(bicycle2$Bike_Type, levels = bike_type_order)

#proportion of bike type being stolen
bicycle2 %>%
  ggplot(aes(x = Bike_Type))+
  geom_bar(fill="light blue")+ labs(x ="Bicycle Type",y="Frequency",title = "Type of Bicycle Stolen",
                   caption = "Figure 4") +theme_bw()
#proportion of bike type being stolen(table)
overall_bike_type$prop <- round(overall_bike_type$prop,3)
overall_bike_type %>% arrange(desc(prop)) %>% kable(caption = "Type of Bicycles Stolen(Proportion wise)
```

```
# get number of occurrence from 2014-2022 from outside,house or apartment
bicycle2 %>%
  filter(Occurrence_Year >=2014) %>%
  group_by(Occurrence_Year) %>% summarise(frequency_outside = sum(outside),
                                          frequency_house = sum(house),
```

```
                                          frequency_apartment = sum(Apartment)) %>%
  mutate(total_frequency = frequency_outside +frequency_house+frequency_apartment) ->outside_percentage
```

```
#proportion test for year 2014 and 2022
x <- outside_percentage$frequency_outside
y <- outside_percentage$frequency_house
z <- outside_percentage$frequency_apartment
n <- outside_percentage$total_frequency

#proportion test for 2014 and 2021 for different premise type
prop.test(x[c(1,8)],n[c(1,8)],alternative = "greater")
prop.test(y[c(1,8)],n[c(1,8)],alternative = "greater")
prop.test(z[c(1,8)],n[c(1,8)],alternative = "less")
```

```
bicycle2 %>% filter(Occurrence_Year==2014) ->bicycle2014
bicycle2 %>% filter(Occurrence_Year==2015) ->bicycle2015
bicycle2 %>% filter(Occurrence_Year==2016) ->bicycle2016
bicycle2 %>% filter(Occurrence_Year==2017) ->bicycle2017
bicycle2 %>% filter(Occurrence_Year==2018) ->bicycle2018
bicycle2 %>% filter(Occurrence_Year==2019) ->bicycle2019
bicycle2 %>% filter(Occurrence_Year==2020) ->bicycle2020
bicycle2 %>% filter(Occurrence_Year==2021) ->bicycle2021
```

```
# model
overall <-glm(outside ~Occurrence_Month +Occurrence_Hour+Division+Bike_Type+Bike_Speed+
                    Cost_of_Bike,data = bicycle2, family = "binomial")
summary(overall)
```

```
library(pROC)
# corss validation for model and find c.index
k=4
d = bicycle2 %>% mutate(group_ind = sample(c(1:k),size=nrow(bicycle2),replace = T))
c.index = vector()
for (i in 1:k){
d.train = d %>% filter(group_ind != i)
d.test = d %>% filter(group_ind == i)
logit.mod = glm(outside ~ Occurrence_Month + Occurrence_Hour +
    Division + Bike_Type + Bike_Speed + Cost_of_Bike, family = "binomial",
    data = d.train)
pi_hat = predict(logit.mod, newdata=d.test, type = "response")
m.roc=roc(d.test$outside ~ pi_hat)
c.index[i]=auc(m.roc)
}
c.index
mean(c.index)
```

```
#bootstrap samples from 2014 and find predicted probability and
#find average of the probability, and use it to calculate confindence interval
boot_function=function(){
boot_data = bicycle2014 %>% sample_n(size = nrow(bicycle2014), replace = F)
m2 = glm(outside ~ Occurrence_Month + Occurrence_Hour +
    Division + Bike_Type + Bike_Speed + Cost_of_Bike, family = binomial,
    data = boot_data)
s = fitted(m2,type= "response")
return(s)
```

```
}
out = replicate(1000, boot_function())
rowMeans(out) -> rmean2014
quantile(rmean2014,c(0.05,0.95))

#bootstrap samples from 2021 and find predicted probability and
#find average of the probability, and use it to calculate confindence interval
boot_function=function(){
boot_data = bicycle2021 %>% sample_n(size = nrow(bicycle2021), replace = F)
m2 = glm(outside ~ Occurrence_Month + Occurrence_Hour +
    Division + Bike_Type + Bike_Speed + Cost_of_Bike, family = binomial,
    data = boot_data)
s = fitted(m2,type= "response")
return(s)
}
out = replicate(1000, boot_function())
rowMeans(out) -> rmean2021
quantile(rmean2021,c(0.05,0.95))
```