Final Project

Miraj Ismail, Yann Gbaguidi

29/03/2022

Introduction

Safety is one of the most important aspects of human well-being. When choosing a house to buy, or rent, we always prefer houses in neighborhoods we deem to be safe. We want our kids to go to schools in safe neighborhoods, we prefer working in safe environments. This increasing importance of the sense of safety in human well-being raised a question that this paper will try to investigate: Is Toronto a safe city?

To answer this question, we will focus our study on the two main types of offenses (break-and-entering and robbery) in the city of Toronto, and by distinguishing between the city's rich and poor neighborhoods.

To examine the different aspects aforementioned, we will ask the following questions: (a) How did crime numbers (break-and-entering and robbery) in the city have evolved between 2014-2020? (b) Do most attempted robberies (break-and-entering) result in robberies? (c) Is Toronto a more violent city than other big cities in the world? (d) Are there disparities in the amount of crimes between Toronto's rich and poor neighborhoods

Data Description

To evaluate the relative safety of Toronto, we retrieved a data set from the Toronto Police Service's Public Safety Portal which contained the different crime rates across all of Toronto's neighborhoods. We also retrieved the average income across each of these neighborhoods from the 2016 Canada Census. We then joined both data set by neighborhoods, and because our investigation only took observed break-and-entering crimes as well as robberies, we filtered the other types of offenses of the data set. Our final data set contained the rates of break-and-entering, and robberies across each of Toronto's neighborhoods, as well as the average income in each of these neighborhoods.

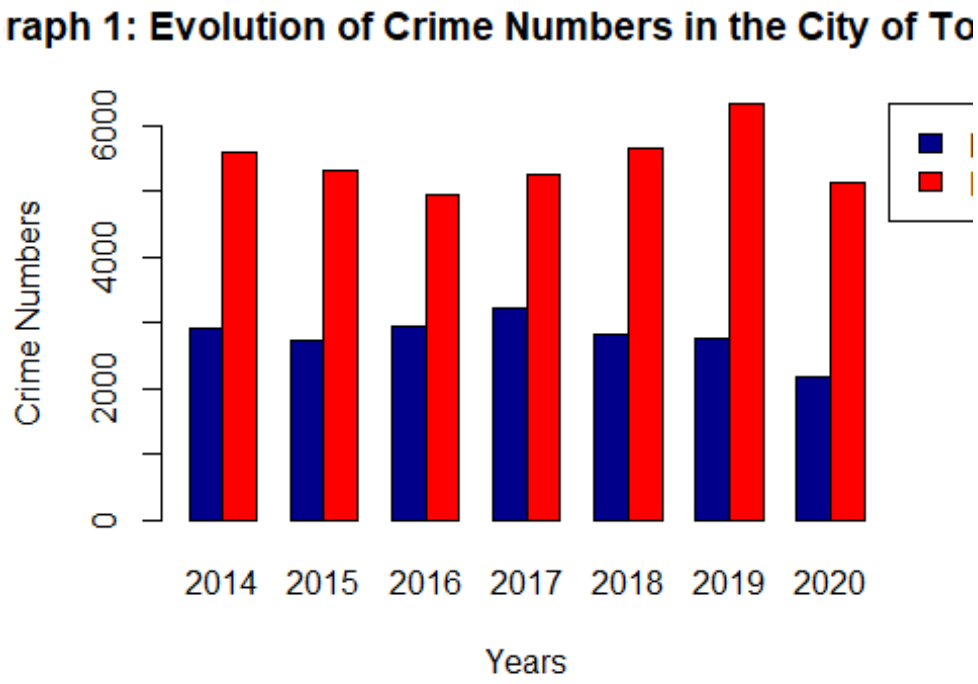
Data Analysis

(a) How did crime numbers (break-and-entering and robbery) in the city have evolved between 2014-2020?

The following graph shows the progression of break-and-entering offenses and robberies across Toronto from 2014 to 2020

```
barplot(tab, main = "Graph 1: Evolution of Crime Numbers in the City of
Toronto",
        xlab = "Years", ylab = "Crime Numbers", col = c("darkblue", "red"),
```

```
        legend.text = rownames(tab), args.legend = list(x = "topright", inset
= c(-0.5,0)), beside = TRUE)
```



From this bar graph, we can observe that the overall break-and-entering and robbery rates have decreased from 2014 to 2020 in Toronto. We could therefore imply that Toronto was a much "safer" city in 2020 than it was in 2014, however, this assumption could be refuted because of the lockdown imposed by the COVID-19 Pandemic as crime rates where back on the rise beforehand. Although graph one shows there was a decrease in cases of robbery and break-and-entering, it doesn't yet tell us how safe it is as compared to other cities. Question 2 attempts to better answer this question.

(b)Do most attempted robberies (break-and-entering) result in robberies??

The following tests assess whether or not there is a positive correlation between break-and-entering cases and robberies across Toronto from 2018 to 2020. We will create a 95% confidence interval to test whether there is a positive correlation most cases of break-and-entering result in robberies, and repeat the same process, but this time, using bootstrapping.

```
pop_mean
```

```
## [1] 155.7545
```

```r
qnorm(c(0.05, 0.95), mean = sample_mean,
      sd = sqrt(sample_var/length(pop_sample)))
```

```
## [1] 145.0208 196.5507
```

```r
boot_function = function(){
  s = sample(pop_sample, size = length(pop_sample), replace = T)
  return(mean(s))
}

boot_x_bar = (replicate(1000, boot_function()))
quantile(boot_x_bar, c(0.05, 0.95))
```

```
##       5%      95%
## 145.7679 198.0179
```

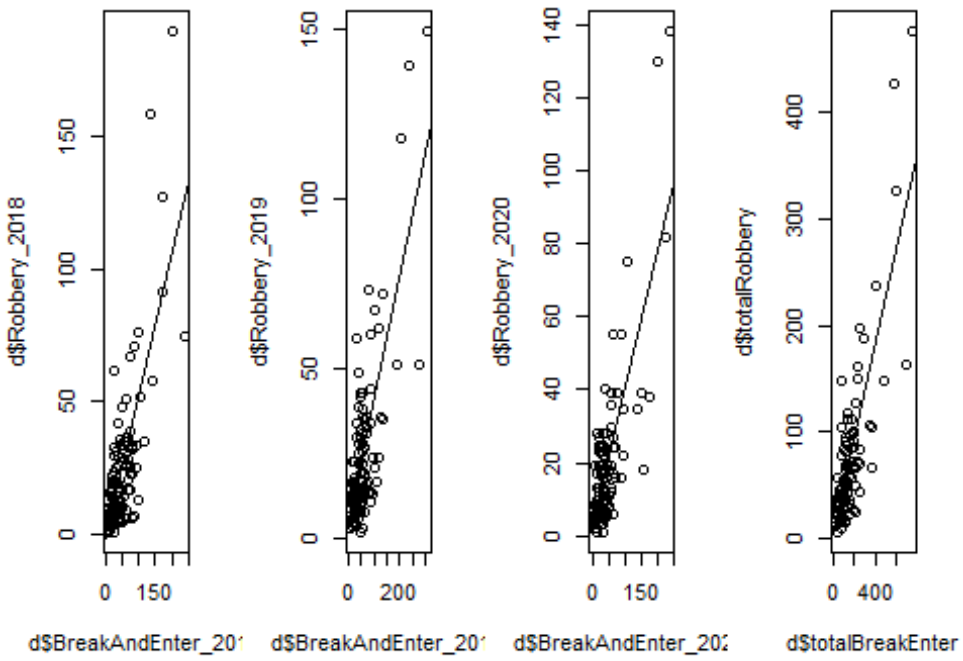
We can observe that the mean number of breaking and entering cases at the population level falls within the confidence interval when using a normal distribution as well as bootstrapping. / /

The following scatter plots examine whether or not there is a positive correlation between break-and-entering and robberies

```r
m1 = lm(Robbery_2018~BreakAndEnter_2018, data = d)
m2 = lm(Robbery_2019~BreakAndEnter_2019, data = d)
m3 = lm(Robbery_2020~BreakAndEnter_2020, data = d)
m4 = lm(totalRobbery~totalBreakEnter, data = d)

par(mfrow = c(1,3))

par(mfrow=c(1,4))
plot(d$BreakAndEnter_2018, d$Robbery_2018)
abline(m1)
plot(d$BreakAndEnter_2019, d$Robbery_2019)
abline(m2)
plot(d$BreakAndEnter_2020, d$Robbery_2020)
abline(m3)
plot(d$totalBreakEnter, d$totalRobbery)
abline(m4)
```




```
summary(m1)

##
## Call:
## lm(formula = Robbery_2018 ~ BreakAndEnter_2018, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.673  -9.255  -1.779   7.094  86.833
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.02202    2.98405  -0.678    0.499
## BreakAndEnter_2018  0.53816    0.04597  11.708   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.09 on 108 degrees of freedom
## Multiple R-squared:  0.5593, Adjusted R-squared:  0.5552
## F-statistic: 137.1 on 1 and 108 DF,  p-value: < 2.2e-16

summary(m2)

##
## Call:
## lm(formula = Robbery_2019 ~ BreakAndEnter_2019, data = d)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.116  -9.318  -1.542   7.342  48.682
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.97829    2.29945    1.73   0.0865 .
## BreakAndEnter_2019  0.36740    0.03008   12.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.87 on 108 degrees of freedom
## Multiple R-squared:   0.58,  Adjusted R-squared:  0.5761
## F-statistic: 149.2 on 1 and 108 DF,  p-value: < 2.2e-16

summary(m3)

##
## Call:
## lm(formula = Robbery_2020 ~ BreakAndEnter_2020, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.771  -6.773  -0.616   7.483  51.969
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.02323    1.86855   1.083    0.281
## BreakAndEnter_2020  0.37815    0.02959  12.780   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.18 on 108 degrees of freedom
## Multiple R-squared:  0.602,  Adjusted R-squared:  0.5983
## F-statistic: 163.3 on 1 and 108 DF,  p-value: < 2.2e-16

#plot(m4)
```

We notice that the P-Value for each of the three years is very small, suggesting there is a significant relationship between the rates of breaking and entering crimes and robberies. This may be due to the fact that these crimes play hand in hand, as breaking and entering crimes can often be motivated by robberies. From the summary, R^2 = 0.602 which indicates that 60.2% variation in Breaking and entering can be explained by our model. This tells us that there is a very strong association between breaking and entering and robbery. / / / **(c)Is Toronto a more violent city than other big cities in the world?** From our research we found out that a city was considered to be fairly unsafe if the number of break-and-entering per year is higher than 70. Our hypothesis was that about 95% to Toronto's neighborhoods had higher rates (Toronto is a majoritarily violent city), and were therefore deemed to be "unsafe" by convention. We used a proportion test to test our hypothesis.

```r
sample_prop = mean(d$totalBreakEnter > 210)

d2 = d %>% summarize(n=n(), x=sum(totalBreakEnter > 210))
prop.test(d2$x,d2$n, p = 0.95)

##
##  1-sample proportions test with continuity correction
##
## data:  d2$x out of d2$n, null probability 0.95
## X-squared = 1350.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.95
## 95 percent confidence interval:
##  0.1171695 0.2692565
## sample estimates:
##         p
## 0.1818182
```

The results of the proportion test show that our p-value significantly smaller than 0.1, therefore we can reject our hypothesis that Toronto is a majoritarily unsafe city. However, this doesn't mean that we have enough evidence to conclude that Toronto is a safe city by convention.

I don't know what to do with this haha

```r
## k fold cross validation

#here we will test the relationship between when totalBreakEnter is greater
than 210
#against the number of robberies
d = d %>% mutate(crimes = ifelse(totalBreakEnter>=210, 1, 0))

k = 4
fold.ind = sample(c(1,2,3,4), size = nrow(d), replace = T)

c.index=vector()

for (i in 1:4){
  d.train = d[fold.ind != i, ]
  d.test = d[fold.ind == i, ]
  logit.mod = glm(crimes~totalRobbery, family = binomial, data=d.train)
  pi_hat = predict(logit.mod, newdata = d.test, type = "response")
  m.roc=roc(d.test$crimes ~ pi_hat)
  c.index[i]=auc(m.roc)
}

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

c.index

## [1] 0.9074074 0.9473684 0.4907407 0.9261364

mean(c.index)

## [1] 0.8179132
```

(d)Are there disparities in the amount of crimes between Toronto's rich and poor neighborhoods The following t-test checks whether there is a difference between the mean number of break-and-entering in poor vs rich neighborhoods (poor neighborhoods are neighborhoods with an average household income less than the national average). Our null hypothesis is that there is no difference between the mean number of break-and-entering in poor vs rich neighborhoods. The alternative hypothesis is that poor neighborhoods have higher break-and-entering rates than rich neighborhoods on average.

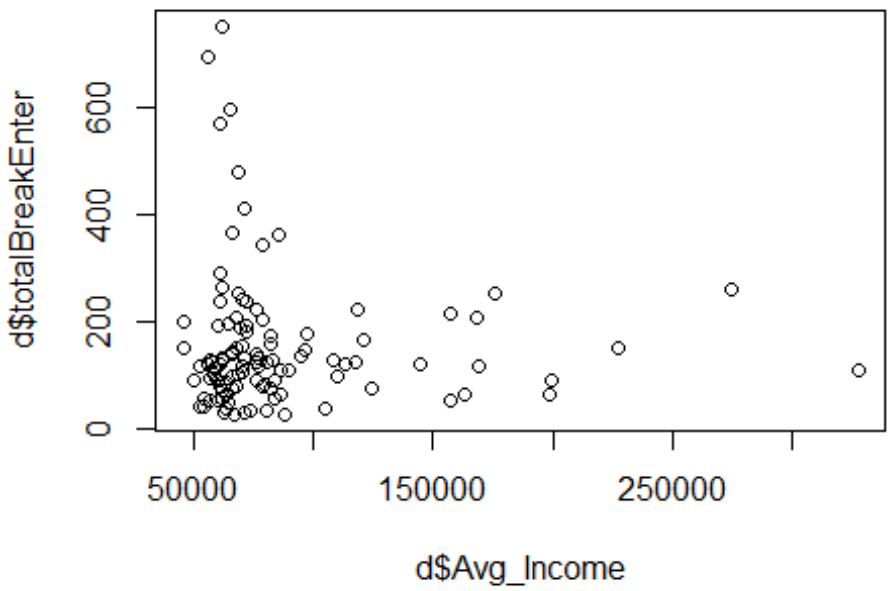
```
poor = d %>% filter(Avg_Income < 62900)
rich = d %>%  filter(Avg_Income > 62900)

t.test(poor$totalBreakEnter, rich$totalBreakEnter,
       mu = 0,
       alternative = "greater",
       conf.level = 0.95)

##
##  Welch Two Sample t-test
##
## data:  poor$totalBreakEnter and rich$totalBreakEnter
## t = 0.75474, df = 38.102, p-value = 0.2275
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -31.72945       Inf
## sample estimates:
## mean of x mean of y
##  174.2258  148.5063
```

The results of our t-test show that our p-value is greater than 0.1, and that our alternative hypothesis is true. To dig further in the subject, we will check if there is a positive correlation between the income and the rates of break-and-entering:

```
m6 = lm(Avg_Income~totalBreakEnter, data = d)
plot(d$Avg_Income, d$totalBreakEnter)
```




```
summary(m6)

##
## Call:
## lm(formula = Avg_Income ~ totalBreakEnter, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40334 -24151 -14249  -1920 240330
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     89351.87    6842.81  13.058   <2e-16 ***
## totalBreakEnter   -18.58      33.96  -0.547    0.585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45530 on 108 degrees of freedom
## Multiple R-squared:  0.002764,   Adjusted R-squared:  -0.00647
## F-statistic: 0.2993 on 1 and 108 DF,  p-value: 0.5854

#the R^2 value is 0.002 meaning there is a very weak relationship between
income and breaking and
#entering rates
```

We can see that there is a very weak relationship between income and breaking-and-entering rates, as the R^2 is 0.002. We cannot infer that the wealth of a neighborhood is a determining factor of that neighborhood's safety.

Summary

Based on the results of question (a), not withstanding the impact of the pandemic's restrictions, the number of break-and-enterings as well as the number of robberies has decreased between 2014-2020. These numbers are also well below the range for the most violent big cities across the world.
Question (b)'s results show that there is a positive correlation between attempted robbery cases (break-and-entering) and robberies: It is therefore safer to affirm that most break-and-entering cases are meant to be robberies. When considering answers to question (a) and (b), since the numbers of break-and-entering cases is almost always twice as much as robberies, we can see that most break-and-entering do not result in robberies, implying that the reactive responses to threats to safety are rather quite efficient.
Regarding question(c), our t-test results show that we can safely reject the hypothesis that the vast majority of Toronto is unsafe, as per convention. That's when cross validation comes in ……..
Question(d) shows that despite the fact that the number of offenses is not "evenly distributed" across the city, there is no positive correlation between a neighborhood's average income and crime rate. Therefore we cannot affirm that one is safer in a rich neighborhood than they would be in a poor one.
Finally, when grouping all the results we have, because we have observed that rates of crimes are decreasing in Toronto, that the the reactive responses to threats to safety are efficient, that the risk of crime is not related to a neighborhood's income (there is no big crime disparity across the city), and that the vast majority of the city's neighborhoods have lower crime rates than that of unsafe cities, we can safely affirm that Toronto is not a violent and unsafe city.