CASE STUDY REPORT - DETERMINANTS OF THE REPORTED CRIMES IN TORONTO

Emin Rza – 1006991560, Jason Richard – 1007706412, Saahil Thukral - 1007165434

University of Toronto
April 7, 2023

Summary of Data

```
## Rows: 2,701
## Columns: 8
## $ `_id`         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17…
## $ ObjectId      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17…
## $ ReportedYear <dbl> 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 201…
## $ GeoDivision  <chr> "D11", "D11", "D11", "D11", "D11", "D11", "D11", "D11", "…
## $ Category      <chr> "Controlled Drugs and Substances Act", "Crimes Against Pr…
## $ Subtype       <chr> "Other", "Auto Theft", "Break & Enter-Apartment", "Break …
## $ Count_        <dbl> 201, 124, 85, 58, 89, 23, 232, 628, 36, 1774, 648, 1, 182…
## $ CountCleared <dbl> 195, 43, 37, 18, 34, 7, 83, 230, 12, 790, 496, 1, 112, 7,…
```

The dataset is retrieved from the City of Toronto Open Dataset from https://open.toronto.ca/dataset/police-annual-statistical-report-reported-crimes/ that was last refreshed on November 18, 2022, regarding the comprehensive overview of reported crimes from the year 2014-2019 in Toronto. This dataset has a total of 2701 reported entries with 8 variables that are important to analyze the different types of crimes based on year, division, and subtype.

Here is a breakdown of what each variable represents:

- X_id: (int) Unique row identifier for Open Data database
- ObjectId : (int) Unique identifier from the source system
- ReportedYear: (int) Year crime was reported
- GeoDivision : (chr) Geographic division where the crime took place
- Category: (chr) Crime category
- Subtype: (chr) Crime category subtype
- Count: (int) Total number of crimes
- CountCleared: (int)Total number of crimes identified as cleared

Background and Significance

The Annual Statistical Report (ASR) of the Toronto Police Service provides a detailed overview of various police-related statistics such as reported crimes, victims, personnel, budget, and other administrative information. The dataset contains all reported criminal offenses between 2014 and 2019, aggregated by division and reported date, including crimes that were later deemed unfounded, occurred outside Toronto's limits or had no verified location.

The Toronto Police Service has cooperated with the Municipal Freedom of Information and Protection of Privacy Act to protect the privacy of people who are associated with the reported occurrences to make sure that no personal information pertaining to any of the persons involved will be disclosed. The information has been divided into groups according to year, category, subtype, and region.

If an occurrence involves multiple offense types, it will be included in multiple categories. The count presented *does not* indicate the number of distinct incidents.

Data Collection
The Toronto Police Service usually gathers information through a variety of tools, such as incident reports, service requests, officer notes, and other police-related paperwork. First, the data is used to produce reports and statistics and then it is entered into a central database to make the Annual Statistical Report. Police officers manually enter data or use automated methods to do so depending on the type of information. Toronto Police Service also collaborates with other agencies such as hospitals and community organizations to collect data.

```
##      Count_       CountCleared
##  Min.   :   1   Min.   :   0.0
##  1st Qu.:  53   1st Qu.:  15.0
##  Median : 134   Median :  56.0
##  Mean   : 374   Mean   : 161.3
##  3rd Qu.: 380   3rd Qu.: 153.0
##  Max.   :7256   Max.   :2207.0
```

The mean value of Count and CountCleared is 374 and 161.3 respectively, which is higher than the median, indicating that there are some very high values in both that are increasing the mean. The distribution of Count and CountCleared appears to be heavily skewed to the right, as the maximum value is much larger than the mean and the median, suggesting that there are many outliers in the data.

Overall Research Question: How do the reported crime rates and cleared crimes vary across different *GeoDivisions* in Toronto every year?

Analysis

```
## `summarise()` has grouped output by 'ReportedYear'. You can override using the
## `.groups` argument.

## # A tibble: 8 × 4
## # Groups:   ReportedYear [8]
##   ReportedYear Category                Count_ CountCleared
##          <dbl> <chr>                    <dbl>        <dbl>
## 1         2019 Crimes Against Property  94996        22346
## 2         2018 Crimes Against Property  93015        23622
## 3         2017 Crimes Against Property  80132        22707
## 4         2021 Crimes Against Property  77756        12190
## 5         2020 Crimes Against Property  77630        13757
## 6         2016 Crimes Against Property  73851        22715
## 7         2015 Crimes Against Property  69630        22060
## 8         2014 Crimes Against Property  68819        24096
```

data1 suggests that Crimes against property are the highest crimes that are committed and have been consistently increasing from 2014 to 2019. This rising trend gathered from the data indicates that

Toronto Police should address this as a major concern and should implement preventive measures to keep the community safe.

```
## # A tibble: 8 × 3
##   ReportedYear Count_ CountCleared
##          <dbl>  <dbl>        <dbl>
## 1         2019 144507        57625
## 2         2018 143170        59126
## 3         2017 130036        60016
## 4         2016 122667        59271
## 5         2021 120244        40656
## 6         2020 118528        41661
## 7         2015 117470        58388
## 8         2014 113506        58821
```

The above table (referred to as **data2** in the Appendix) was created to provide the number of crimes committed each year. The highest number of crimes committed was in 2019 with a total of 144507 crimes and the lowest was in 2014 with 113506. This indicates an overall increase in crimes throughout the years.

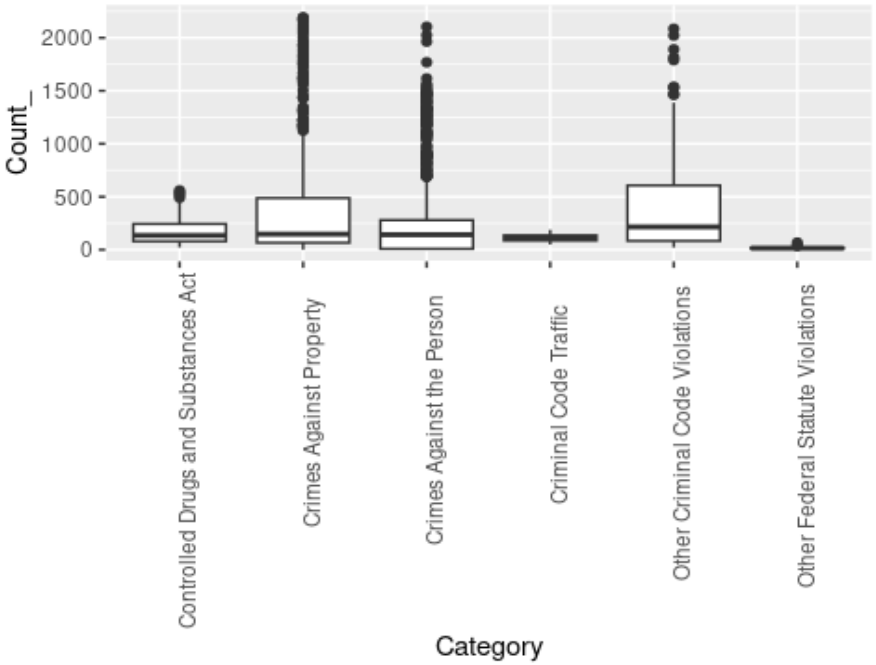
```
## # A tibble: 6 × 3
##   ReportedYear CountType    Totals
##          <dbl> <chr>         <dbl>
## 1         2014 Count_         8805
## 2         2014 CountCleared   4114
## 3         2015 Count_         8876
## 4         2015 CountCleared   3955
## 5         2016 Count_        10244
## 6         2016 CountCleared   4736
```

We manipulated **data3** by pivoting it longer to change the structure to make **data3.pivoted** and focused on GeoDivision D51, allowing us to create a grouped bar chart showing trends of every year.

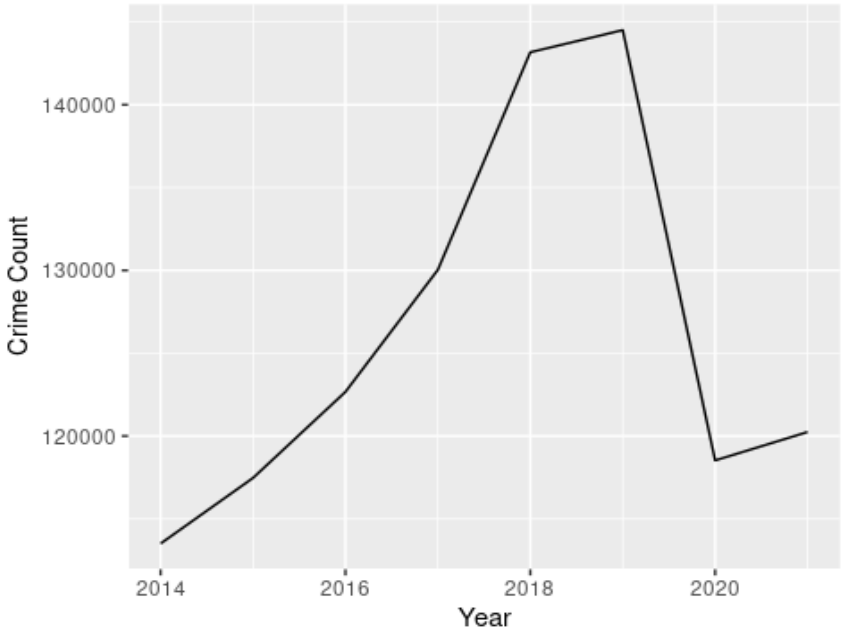
```
## # A tibble: 48 × 4
## # Groups:   ReportedYear [8]
##    ReportedYear Category                          Avg_Count Avg_CountCleared
##           <dbl> <chr>                                 <dbl>            <dbl>
##  1         2019 Crimes Against Property                621.            146.
##  2         2018 Crimes Against Property                608.            154.
##  3         2017 Crimes Against Property                524.            148.
##  4         2021 Crimes Against Property                508.             79.7
##  5         2020 Crimes Against Property                507.             89.9
##  6         2019 Other Criminal Code Violations         494.            431.
##  7         2016 Crimes Against Property                486.            149.
##  8         2018 Other Criminal Code Violations         485.            432.
##  9         2017 Other Criminal Code Violations         468.            427.
## 10         2015 Crimes Against Property                455.            144.
## # … with 38 more rows
```

The table (**data4**) shows the average reported incidents of the different crimes committed in the different years. The highest average number of reported crimes was in 2019 for Crimes Against Property, while the lowest average count of reported crimes was in 2021 for Other Federal Statute Violations.

Graphical Representations and Plots:

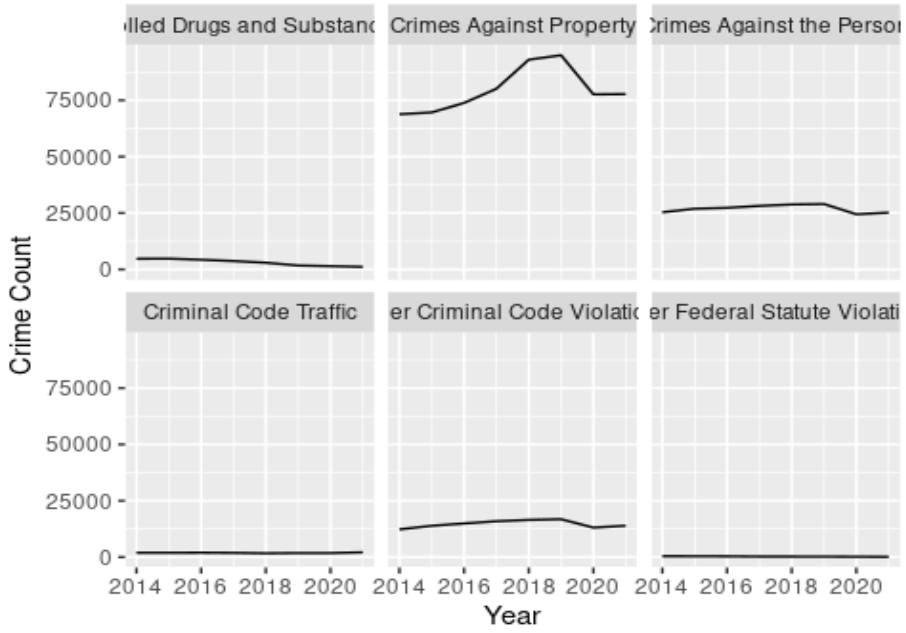


The figure above consists of boxplots, indicating the number of crimes based on the types of crimes being committed. Based on the figure above, we can see that Crime Against Property shows the highest outlier compared to other categories. Overall, since the outliers of the data are so extreme, the normality of the data is in question, as the boxplots indicate extreme skewness to right. Therefore, it would be recommended to use techniques such as bootstrapping, as it does not depend on the distribution of the data.

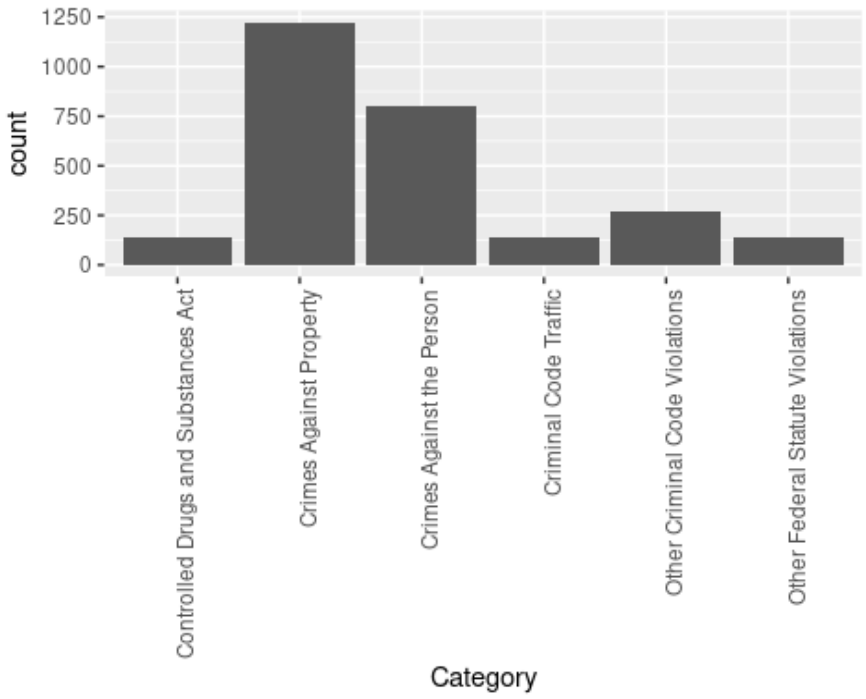


The line graph above shows the trend in crimes being committed from 2014 to 2021. From 2014 to 2018, there has been a significant increase at a fast pace. From 2018 to 2019, there is still an increase, however the rate is relatively slower compared to before. From 2019 to 2020, however, the number of crimes declines extremely rapidly, with the number of crimes in 2020 being lower than

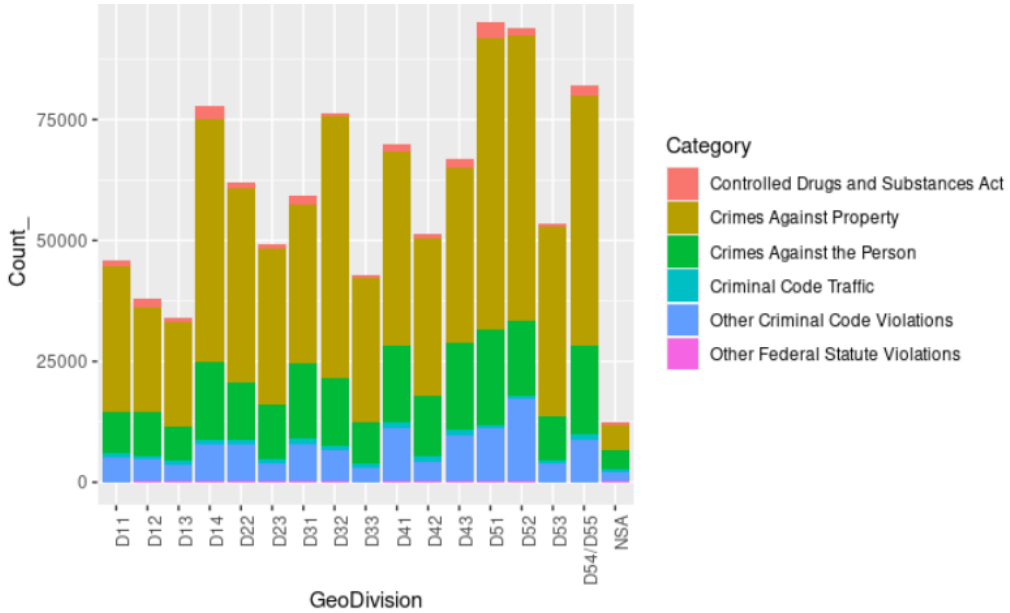
2016. This might be due to Toronto Police addressing the concerns of the high count of committed crimes and showing a significant improvement in the city's safety. Nevertheless, the overall trend is positive where the crime rates get reduced over time although the number of crimes did grow slightly by a few thousand from 2020-2021 during the pandemic. This analysis is valuable because it shows that over the years, there has been a significant improvement in making the City of Toronto safer.



The graph above depicts crime trends from 2014 to 2021 by Crime Categories. The Controlled Drugs and Substance Act exhibits a steady linear decline in the number of crimes. Conversely, Crimes Against Property show an increase at a steeper rate, peaking in 2019 and dropping significantly in 2020. The remaining categories such as Crimes Against Person, Criminal Code Traffic, Other Criminal Code Violations, and Other Federal Statute Violations exhibit stable horizontal trend lines with slight fluctuations. Overall Federal Statute Violations have the lowest impact and Crimes Against Property have the biggest impact on the number of crimes committed.

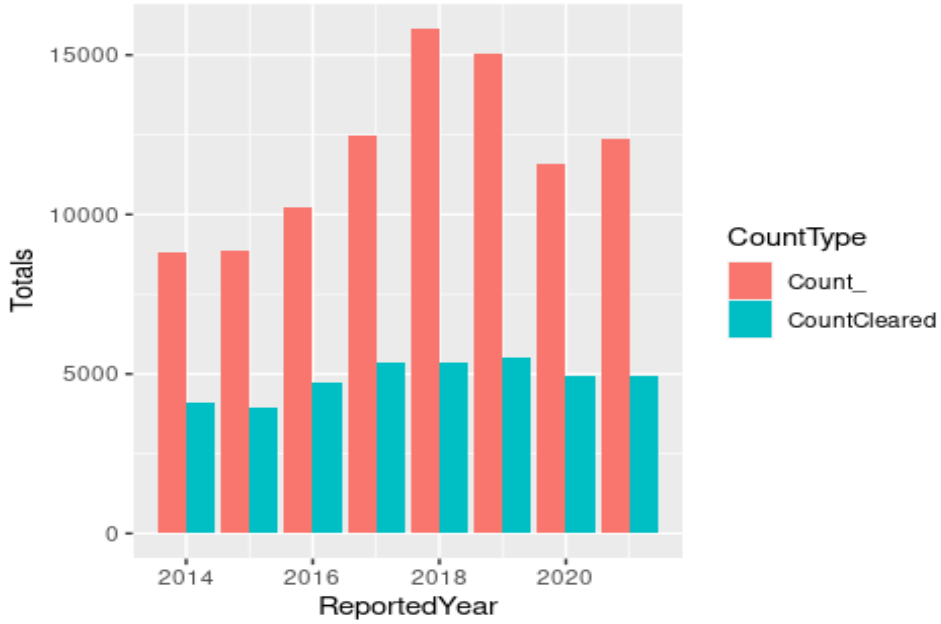


The bar chart above illustrates crimes committed based on the Crime Categories. Crimes Against Property appears to have the biggest impact on committed crimes, followed by Crime Against Person and Other Criminal Code Violations. This indicates that property crimes are the most pervasive form of criminal activity in Toronto. When considering the data across multiple years, this trend remains consistent and these findings indicate that Toronto Police should prioritize prevention strategies aimed at reducing property and personal crimes to improve the city's safety.



The stacked bar chart above provides insight into the committed crimes across different GeoDivision and Crime Categories in Toronto. The data shows that the highest committed crime is located at GeoDivision D51, with Crimes Against Property being the dominant category. Conversely, NSA is revealed to have the lowest number of committed crimes among all GeoDivisions. The Toronto

Police can use this information to target specific areas and provide more resources and protections to address the high incidence of committed crimes, especially Crime Against Property in GeoDivision D51.



The graph above illustrates a detailed analysis of a grouped bar chart based on the number of committed crimes and cleared crimes during the year in *GeoDivision D51*, the most dangerous GeoDivision in Toronto. There is a steady increase in the counts of crimes from 2014 and peaking in 2018 where the overall trend shows a decrease in crime from 2018 to 2021. Although the number of crimes has been fluctuating, CountCleared has been at the relatively same level throughout the year, indicating that the ratio between CountCleared and committed crimes is becoming larger, and is a matter of concern.

Confidence intervals using t.test and prop.test

Finding the confidence interval of *Count_* (which is considered dependent variable):

```
##
##  One Sample t-test
##
## data:  data$Count_
## t = 29.419, df = 2700, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  349.0564 398.9095
## sample estimates:
## mean of x
##   373.983
```

The code has been run and the range of the mean value is between 349.0564 and 398.9095, with the mean of the data being 373.983.

However as seen from the boxplots of the data, the normality of the data appears to be questionable, therefore a bootstrap confidence interval is being done:

```
##     2.5%    97.5%
## 349.9980 399.2958

##
##  One Sample t-test
##
## data:  data$Count_
## t = 29.419, df = 2700, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  349.0564 398.9095
## sample estimates:
## mean of x
##   373.983
```

The code has been run and the range of the mean value is between 348.4253 and 397.7442, which is quite similar to the confidence interval found using the t-test. However, since the data does not appear to be normal, a bootstrap test is a much better test, as it works, irrespective of the distribution of the data, as opposed to the t-test, which works for normally distributed data.

95% confidence interval of the number of crimes that have been cleared:

```
##
##  One Sample t-test
##
## data:  data$CountCleared
## t = 31.332, df = 2700, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  151.1683 171.3523
## sample estimates:
## mean of x
##  161.2603
```

When this test is run several times, there is a 95% confidence that the true mean falls between 151.1683 and 171.3523.

Proportion test

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(data$CountCleared) out of sum(data$Count_), null probability 0.5
## X-squared = 19127, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4302308 0.4321634
## sample estimates:
##         p
## 0.4311968
```

Since p-value < 2.2e-16, there is strong evidence that the proportion of crimes cleared is significantly different from 0.5. The sample proportion of crimes cleared *p* is reported as 0.4311968 with a 95% confidence interval of (0.4302308, 0.4321634). This means that we can be 95% confident that the true proportion of crimes cleared falls within this range.

Hypothesis Testing

Testing whether the average number of crimes is greater than 300:

```
##
##  One Sample t-test
##
## data:  data$Count_
## t = 5.8199, df = 2700, p-value = 3.292e-09
## alternative hypothesis: true mean is greater than 300
## 95 percent confidence interval:
##  353.0662      Inf
## sample estimates:
## mean of x
##   373.983
```

Looking at the p-value, which is much smaller than $\alpha = 0.05$, we can reject the null hypothesis and therefore come to the conclusion that the mean count of cases across the years is greater than 300. With the skewness of the data, the t-test may not be the best option. Therefore, a bootstrapping test has also been conducted below.

Bootstrapping: an alternative to the t-test used for hypothesis testing.

```
## [1] 0.969
```

The result from the bootstrapping test is displayed above. In comparison to the t-test, the p-value is much higher than $\alpha = 0.05$, indicating that the null hypothesis is not rejected, and there is a possibility that the mean value is 300.

Regression Model

The total number of crimes (Count_) as the dependent variable:

```
##
## Call:
## lm(formula = Count_ ~ ReportedYear + GeoDivision + Category +
##     Subtype + CountCleared, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1263.2   -56.0    -1.4    50.6  3529.6
##
## Coefficients:
```

```
##                                                 Estimate Std. Error
## (Intercept)                                   -3.585e+04  3.617e+03
## ReportedYear                                   1.765e+01  1.791e+00
## GeoDivisionD12                                -3.584e+01  2.368e+01
## GeoDivisionD13                                -1.637e+01  2.373e+01
## GeoDivisionD14                                 7.637e+01  2.374e+01
## GeoDivisionD22                                 4.735e+01  2.366e+01
## GeoDivisionD23                                 4.975e+01  2.368e+01
## GeoDivisionD31                                 1.435e+01  2.374e+01
## GeoDivisionD32                                 1.059e+02  2.369e+01
## GeoDivisionD33                                 2.688e+01  2.365e+01
## GeoDivisionD41                                -5.430e+00  2.380e+01
## GeoDivisionD42                                 3.798e+01  2.364e+01
## GeoDivisionD43                                -3.627e+01  2.387e+01
## GeoDivisionD51                                 1.134e+02  2.393e+01
## GeoDivisionD52                                 5.906e+01  2.403e+01
## GeoDivisionD53                                 6.887e+01  2.383e+01
## GeoDivisionD54/D55                             7.189e+01  2.380e+01
## GeoDivisionNSA                                -5.907e+01  2.399e+01
## CategoryCrimes Against Property                4.579e+02  2.577e+01
## CategoryCrimes Against the Person              1.162e+02  2.561e+01
## CategoryCriminal Code Traffic                  4.035e+01  2.567e+01
## CategoryOther Criminal Code Violations        -2.669e+02  2.978e+01
## CategoryOther Federal Statute Violations       1.002e+02  2.603e+01
## SubtypeAttempt Murder                          1.007e+02  3.253e+01
## SubtypeAuto Theft                             -1.829e+01  4.249e+01
## SubtypeBreak & Enter-Apartment                -1.788e+02  4.252e+01
## SubtypeBreak & Enter-Commercial               -1.657e+02  4.237e+01
## SubtypeBreak & Enter-House                    -1.504e+02  4.259e+01
## SubtypeBreak & Enter-Other                    -2.272e+02  4.296e+01
## SubtypeFraud                                   1.702e+02  4.128e+01
## SubtypeOther                                   1.098e+02  2.970e+01
## SubtypeOther Criminal Violations - Offensive Weapons  4.365e+02  4.888e+01
## SubtypeRobbery-Financial                       9.916e+01  3.251e+01
## SubtypeRobbery-Other                           1.500e+02  3.089e+01
## SubtypeSexual Violation                        1.177e+02  3.092e+01
## SubtypeTheft Over $5000                       -1.942e+02  4.279e+01
## SubtypeTheft Under $5000                       1.127e+03  3.631e+01
## CountCleared                                   1.693e+00  3.020e-02
##                                                t value Pr(>|t|)
## (Intercept)                                     -9.912  < 2e-16 ***
## ReportedYear                                     9.851  < 2e-16 ***
## GeoDivisionD12                                  -1.514 0.130246
## GeoDivisionD13                                  -0.690 0.490558
## GeoDivisionD14                                   3.217 0.001310 **
## GeoDivisionD22                                   2.001 0.045442 *
## GeoDivisionD23                                   2.101 0.035723 *
## GeoDivisionD31                                   0.604 0.545788
## GeoDivisionD32                                   4.471 8.10e-06 ***
## GeoDivisionD33                                   1.137 0.255800
## GeoDivisionD41                                  -0.228 0.819568
## GeoDivisionD42                                   1.607 0.108209
## GeoDivisionD43                                  -1.520 0.128729
## GeoDivisionD51                                   4.737 2.28e-06 ***
## GeoDivisionD52                                   2.458 0.014041 *
## GeoDivisionD53                                   2.890 0.003886 **
## GeoDivisionD54/D55                               3.021 0.002542 **
## GeoDivisionNSA                                  -2.462 0.013887 *
## CategoryCrimes Against Property                 17.768  < 2e-16 ***
## CategoryCrimes Against the Person                4.539 5.90e-06 ***
## CategoryCriminal Code Traffic                    1.572 0.116114
## CategoryOther Criminal Code Violations          -8.965  < 2e-16 ***
## CategoryOther Federal Statute Violations         3.850 0.000121 ***
## SubtypeAttempt Murder                            3.097 0.001973 **
## SubtypeAuto Theft                               -0.430 0.666930
## SubtypeBreak & Enter-Apartment                  -4.205 2.69e-05 ***
## SubtypeBreak & Enter-Commercial                 -3.911 9.41e-05 ***
## SubtypeBreak & Enter-House                      -3.531 0.000420 ***
## SubtypeBreak & Enter-Other                      -5.288 1.34e-07 ***
## SubtypeFraud                                     4.124 3.84e-05 ***
## SubtypeOther                                     3.699 0.000221 ***
## SubtypeOther Criminal Violations - Offensive Weapons   8.930  < 2e-16 ***
## SubtypeRobbery-Financial                         3.050 0.002312 **
## SubtypeRobbery-Other                             4.856 1.27e-06 ***
```

```
## SubtypeSexual Violation                                    3.808 0.000143 ***
## SubtypeTheft Over $5000                                   -4.537 5.94e-06 ***
## SubtypeTheft Under $5000                                  31.046  < 2e-16 ***
## CountCleared                                              56.047  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 211.1 on 2663 degrees of freedom
## Multiple R-squared:  0.8993, Adjusted R-squared:  0.8979
## F-statistic: 642.9 on 37 and 2663 DF,  p-value: < 2.2e-16
```

Important inferences about the model summary:

- $R^2 = 0.8979$: 89.79% variation in *Count_* can be explained by the model. This indicates that our model is very good.
- The coefficient for *ReportedYear* is 17.65, which indicates that with an increase in the year, the average number of reported crimes in a particular category and geographic division increased by 17.65 per year.
- There are several variables in the model that are **not significant** at the $\alpha = 0.05$, including:
  - GeoDivision variables D12, D13, D31, D33, D41, D42, D43: These variables are not significant at $\alpha = 0.05\, level$, which indicates that they do not provide enough explanatory power to the model.
  - CategoryCriminal Code Traffic: p-value = $0.116114 > 0.05$. Thus, the variable is not significant at $\alpha = 0.05\, level$, indicating that it does not provide enough explanatory power to the model.
  - SubtypeAuto Theft: p-value = $0.66693 > 0.05$. Hence, the variable is not significant at $\alpha = 0.05\, level$, indicating that this specific subtype does not provide enough explanatory power to the model.

Cross-validation

Splitting the dataset into two parts: real and test, with a proportion choice of $60\% - 40\%$. respectively

Conducting the **Cross-validation** by checking external validity:

```
## [1] 47347.6
```

```
## [1] 43984.33
```

The results of the Cross-validation are displayed as the MSEs of the *training* and the *test* data respectively. It appears that our model is performing better on the testing dataset compared to the training dataset, which is a positive sign.

Summary

This report analyzes the variation in reported crimes and cleared crimes across GeoDivisions in Toronto annually. The analysis reveals a positive trend of decreasing crime rates over time, as depicted in the second graph. However, the fifth graph indicates that GeoDivisionD51 has the highest crime rates, primarily Crimes Against Property. Using a Linear Regression model, the study finds that certain variables, including SubtypeAuto Theft, CategoryCriminal Code Traffic, and GeoDivision variables D12, D13, D31, D33, D41, D42, and D43, are insignificant at α=0.05. The model's R-squared value of 0.8979 indicates that it is a good fit for the data. Furthermore, cross-validation testing shows better performance on the testing dataset than on the training dataset.

Appendix

Summary of Data

```
library(tidyverse)
data <- read_csv("reportedcrime.csv")
glimpse(data)
```

Section 1: Analysis

```
data2 = data %>% group_by(ReportedYear) %>% summarise(Count_ = sum(Count_), CountCleared
= sum(CountCleared)) %>% arrange(desc(Count_))
data2

data3 = data %>% filter(GeoDivision == "D51") %>% group_by(ReportedYear) %>%
summarise(Count_ = sum(Count_), CountCleared = sum(CountCleared))
data3.pivoted = data3 %>% pivot_longer(Count_:CountCleared, names_to="CountType",
values_to = "Totals", values_drop_na = T)
head(data3.pivoted)

data4 = data %>% group_by(ReportedYear, Category) %>% summarise(Avg_Count =
mean(Count_), Avg_CountCleared = mean(CountCleared)) %>% arrange(desc(Avg_Count))
```

Section 2: Graphical Representations and Plots

```
a1 = ggplot(data, aes(x=Category, y = Count_)) + geom_boxplot() + theme(axis.text.x =
element_text(angle = 90, hjust = 0.8))
a1 + ylim(0,2200)

ggplot(data2, aes(x = ReportedYear, y = Count_)) + geom_line() + labs(x = "Year", y =
"Crime Count")

ggplot(data1, aes(x = ReportedYear, y = Count_)) + geom_line() + facet_wrap(~Category) +
labs(x = "Year", y = "Crime Count")

ggplot(data, aes(x = Category)) + geom_bar() + theme(axis.text.x = element_text(angle =
90, hjust = 1))

data_bar = data %>% group_by(GeoDivision, Category) %>% summarise(Count_ = sum(Count_),
CountCleared = sum(CountCleared))

ggplot(data_bar, aes(x = GeoDivision, y = Count_, fill = Category)) + geom_bar(position
= "stack", stat = "identity") + theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(data3.pivoted, aes(x = ReportedYear, y = Totals, fill = CountType)) +
geom_bar(position = "dodge", stat = "identity")
```

Section 3: Confidence intervals using t.test and prop.test

```
t.test(data$Count_)

boot_data = sample(data$Count_,size = nrow(data))
boot_function=function(){

  boot_s = sample(boot_data,size = nrow(data),replace = TRUE)
  return(mean(boot_s))
}
quantile(replicate(1000,boot_function()), c(0.025,0.975))

#Comparing with t.test
t.test(data$Count_, conf.level = 0.95)
```

95% confidence interval of the number of crimes that have been cleared

```
t.test(data$CountCleared)
```

Proportion test

```
prop.test(sum(data$CountCleared), sum(data$Count_), alternative="two.sided")
```

Hypothesis Testing

```
t.test(data$Count_, mu = 300, alternative = "greater", conf.level = 0.95)
```

Bootstrapping: an alternative to a t-test

```
mu_H0 = 300

x=data$Count_
obs_mean = mean(x)
new_x = x - obs_mean + mu_H0
boot_function= function(){
s= sample(new_x, size=300, replace=T)
return(mean(s))
}
boot_new_x_bar = replicate(1000, boot_function())

mean(boot_new_x_bar<obs_mean)
```

Section 4: Regression Model

Total number of crimes(Count_) as dependent variable:

```
model <- lm(Count_ ~ ReportedYear + GeoDivision + Category + Subtype + CountCleared,
data = data)
summary(model)
```

Cross-validation

```
set.seed(123)
data = data %>% mutate(group_ind = sample(c("train", "test"),
                                          size=nrow(data),
                                          prob=c(0.6, 0.4),
                                          replace = T))
#MSE for the training dataset
m=lm(Count_ ~ GeoDivision + Category + Subtype + CountCleared, data = data %>%
filter(group_ind=="train"))
y.hat=predict(m)

mean((data$Count_[data$group_ind=="train"] - y.hat)^2)

#MSE for the testing dataset
y.hat=predict(m, newdata = data %>% filter(group_ind=="test"))

mean((data$Count_[data$group_ind=="test"] - y.hat)^2)
```