FINAL PROJECT

Data Detectives

OCTOBER 15, 2024

Rajvir Kaur, Nydual Makuach, Matias Totz, Farhad Ahmady, Marcelino Rodriguez

Contents

Crime Statistics in Calgary.

Introduction

When living in a large city, there are many important things to know such as the rent price you can expect, cost of living, salaries, the culture, and many more. According to the website youmoveme, some of the largest concerns people should keep in mind are school reviews, commute time, and safety [1]. For this project, we will be analyzing crime in Calgary since it may become the place where we live for the next few years. Of all the items listed above to know about a city, it is a bit harder to tell the safety of a city unless you see the crimes in person. The significance of this project is to visualize crime in Calgary to help people either moving to Calgary or thinking of making it their long-term home.

Understanding the safety of a large city can be tough unless you see the crimes happening in front of you. However, Open Calgary [2], has data sets on important topics within Calgary such as crime. However, how often do they occur? What crimes are happening more frequently? And where are those crimes taking place? These are some of the questions that someone might have when think about crimes happening in Calgary. This document centers its attention to analyze the data set Community_Crime_Statistics [3] from open Calgary and we aim to understand the behaviors of crimes occurring. By the end of the analysis, we will be able to present our findings to anyone thinking of moving to or living in Calgary to help them understand everything they need to know about crime in Calgary from the years 2018 to 2023. Our domain for this project will mainly be Calgary.

Dataset

The data set we will use in this project is called Community_Crime_Statistics [3], which is found on Open Calgary [2]. The mentioned data set holds information about crimes that occurred in Calgary from the years 2018 to the current year. Our data set has data from the current month but since 2024 is not fully complete, we will mainly be focusing on comparing the years from 2018 to 2023. The data set contains five key columns:

Community: This column captures the specific Calgary community where the crime has occurred. There are over 200 communities in Calgary and most of them are shown in this data set. Not only are we going to find the safest and most unsafe communities, but we will also be putting each community from our data set into sectors such as northeast, south, west, etc. When you live in a city, you will spend the most amount of time in your community and those nearest to yours so understanding the safety of the sector you are in is very important. For us to identify which community belongs in what sector, we will be using a secondary data set called Community_Sectors [4], also obtained from Open Calgary.

Category: This column contains the description of crimes. Currently, there are nine crime categories, which are: Assault (Non-domestic), Break & Enter – Commercial, Break & Enter – Dwelling, Break & Enter - Other Premises, Commercial Robbery, Street Robbery, Theft FROM Vehicle, Theft OF Vehicle, Violence 'Other' (Non-domestic). These have been identified as violent crime offences by the Centre for Canadian Justice excluding domestic violence. Most of these categories are self-explanatory, however it should be noted that the category break & enter includes attempts as well.

Count: This column contains the crimes counts that happened per each data row entry. Since our data set is updated monthly, the count captures if there are multiple of the same crimes in each month. Rather than having each row represent a new crime, each row represents a specific category of a month.

Month: This column contains the month when the crime occurred. Its values are numeric, and it goes from 1 to 12 where 1 is January and 12 is December.

Year: This column contains the year when the crime occurred. Its values are numeric, and it goes from 2018 to 2024. For the current year, the months are updated till August.

The data set contains 71,906 rows which means that at a minimum, there have been 71,906 violent crimes (excluding domestic assault) from 2018 to August 2024. However, since each row can contain multiple crimes, there are many more crimes.

When visiting the data set on Open Calgary's website, they have a section on terms of use under the license/attribution section. It is stated that we are encouraged to use the information as long as we acknowledge the source of the information. We are free to modify, distribute, and use the information in any way for any purpose by doing so.

Pre-work:

In this section, the chosen dataset chosen for analysis is loaded into R to exploration

```
#install.packages("dplyr")
#install.packages("ggplot2")

library(dplyr)

##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

crimes_df <- read.csv("Community_Crime_Statistics_sectors.csv")
#print(crimes_df)
```

From the previews code we can have a taste of the data. we can see the 5 original columns plus the added Sector column that links a community to its respective sector. It also shows the distribution of the data, and it is easy to identify that the data set have different years, months, locations, communicates and crimes therefore we can study the behavior of crimes over all these variables. In the following chapters the analysis of the data set will be distributed over this column to have a better understanding of how the crimes behave

Importing data

After mapping each community to a particular sector, the above is the final data structure for our project.

```
crime_data <- read.csv("Community_Crime_Statistics_Sectors.csv")

head(crime_data)
```

```
##   Community                   Category Crime.Count Year Month    Sector
## 1       01B     Assault (Non-domestic)           1 2022    11 NORTHWEST
## 2       01B Break & Enter - Commercial           1 2019     6 NORTHWEST
## 3       01B Break & Enter - Commercial           1 2019     8 NORTHWEST
## 4       01B Break & Enter - Commercial           2 2020     3 NORTHWEST
## 5       01B Break & Enter - Commercial           2 2020     7 NORTHWEST
## 6       01B Break & Enter - Commercial           1 2020     8 NORTHWEST
```

Guiding Questions.

Crime impacts both quality of life and economic growth [5]. By analyzing these questions, we can gain insights into how law enforcement monitors high-crime areas, examines crime patterns to guide public policy and enhances the allocation of resources. Additionally, it raises awareness among Calgary residents about their safety and helps anticipate future challenges, including those caused by COVID-19.

Question 1) Crimes in Calgary Sectors

There are seven sectors in the city of Calgary: CENTRE, EAST, NORTH, NORTHEAST, NORTHWEST, SOUTH, SOUTHEAST, and WEST.

Determine which community has the highest crime for a particular sector.

- There are over 200 communities in Calgary and most of them are shown in this dataset. We have two files for this project that is Community_Crime_Statistics and Community Sectors. We have created a data structure by mapping each community to its sector.

Identify the most common type of crime committed in each sector of Calgary.

- There are nine crime categories, which are: Assault (Non-domestic), Break & Enter – Commercial, Break & Enter – Dwelling, Break & Enter -1 Other Premises, Commercial Robbery, Street Robbery, Theft FROM Vehicle, Theft OF Vehicle, Violence ‘Other’ (Non-domestic).

Find which area has the highest and lowest Crime.

There are seven sectors in the city of Calgary which are CENTRE, EAST, NORTH, NORTHEAST, NORTHWEST, SOUTH, SOUTHEAST, and WEST. In this, we will find which sector of Calgary has the highest and lowest crime.

Below is the code for the analysis

```
library(dplyr)

##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

sector_crime_summary <- crime_data %>%
  group_by(Sector) %>%
  summarise(Total_Crime_Count = sum(Crime.Count, na.rm = TRUE))

print(sector_crime_summary)

## # A tibble: 8 × 2
##   Sector    Total_Crime_Count
##   <chr>                 <int>
## 1 CENTRE                67521
## 2 EAST                  17106
## 3 NORTH                 13984
## 4 NORTHEAST             38904
## 5 NORTHWEST             17965
## 6 SOUTH                 24708
## 7 SOUTHEAST             12353
## 8 WEST                  11764

# Finding the sector with the lowest crime count
lowest_crime_sector <- sector_crime_summary %>%
  filter(Total_Crime_Count == min(Total_Crime_Count))

# Finding the sector with the highest crime count
highest_crime_sector <- sector_crime_summary %>%
  filter(Total_Crime_Count == max(Total_Crime_Count))

cat("Sector with the lowest crime count:\n")
```

```
## Sector with the lowest crime count:

print(lowest_crime_sector)

## # A tibble: 1 × 2
##   Sector Total_Crime_Count
##   <chr>              <int>
## 1 WEST               11764

cat("\nSector with the highest crime count:\n")

##
## Sector with the highest crime count:

print(highest_crime_sector)

## # A tibble: 1 × 2
##   Sector Total_Crime_Count
##   <chr>              <int>
## 1 CENTRE             67521
```

In the above, we have calculated the total crime count for each sector. We can see that the sector CENTRE has the highest crime count i.e. 67521 and the WEST sector has the lowest crime with the value of 11764 as compared to the other sectors.

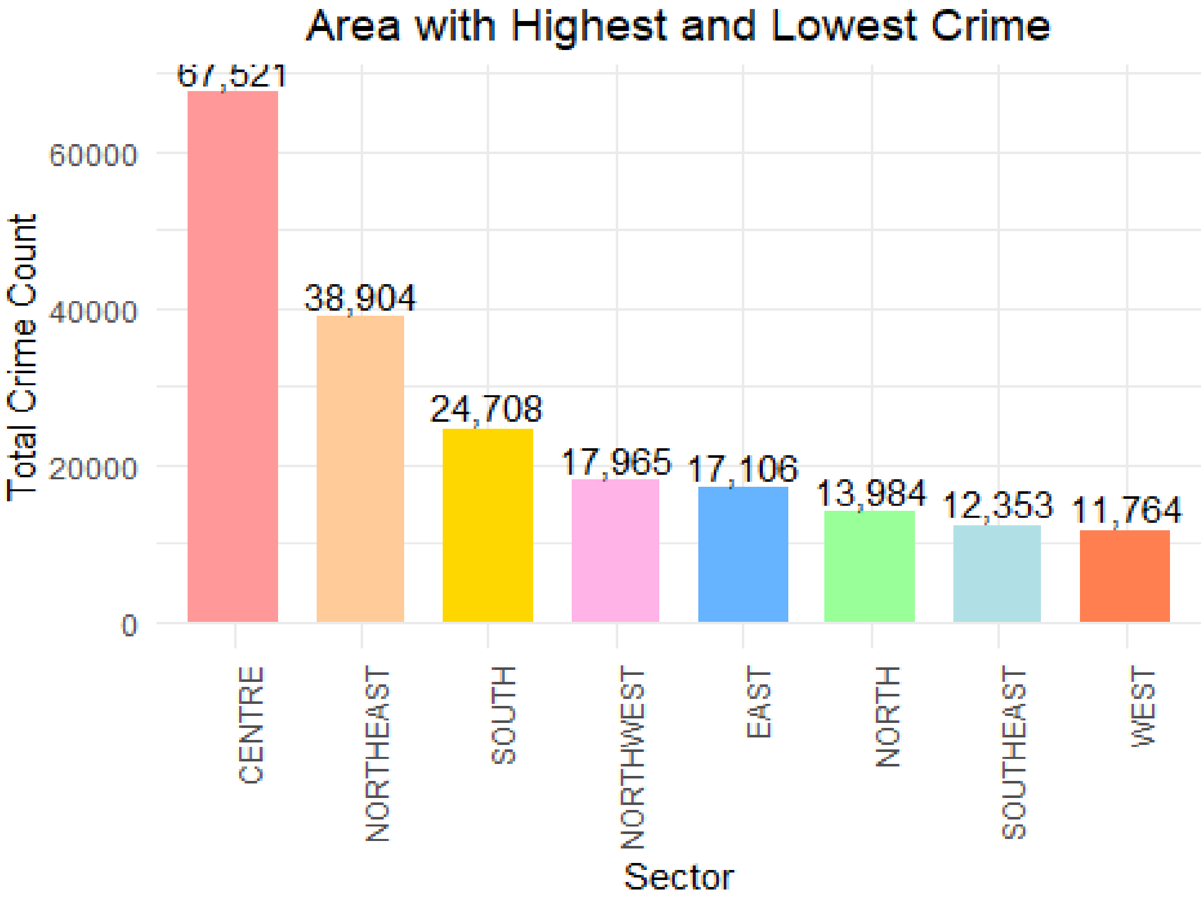
Visualization of Dataset:

We have used two types of charts i.e. bar chart and pie chart for the visualization of the above.

Below is the code for the analysis

Bar Chart

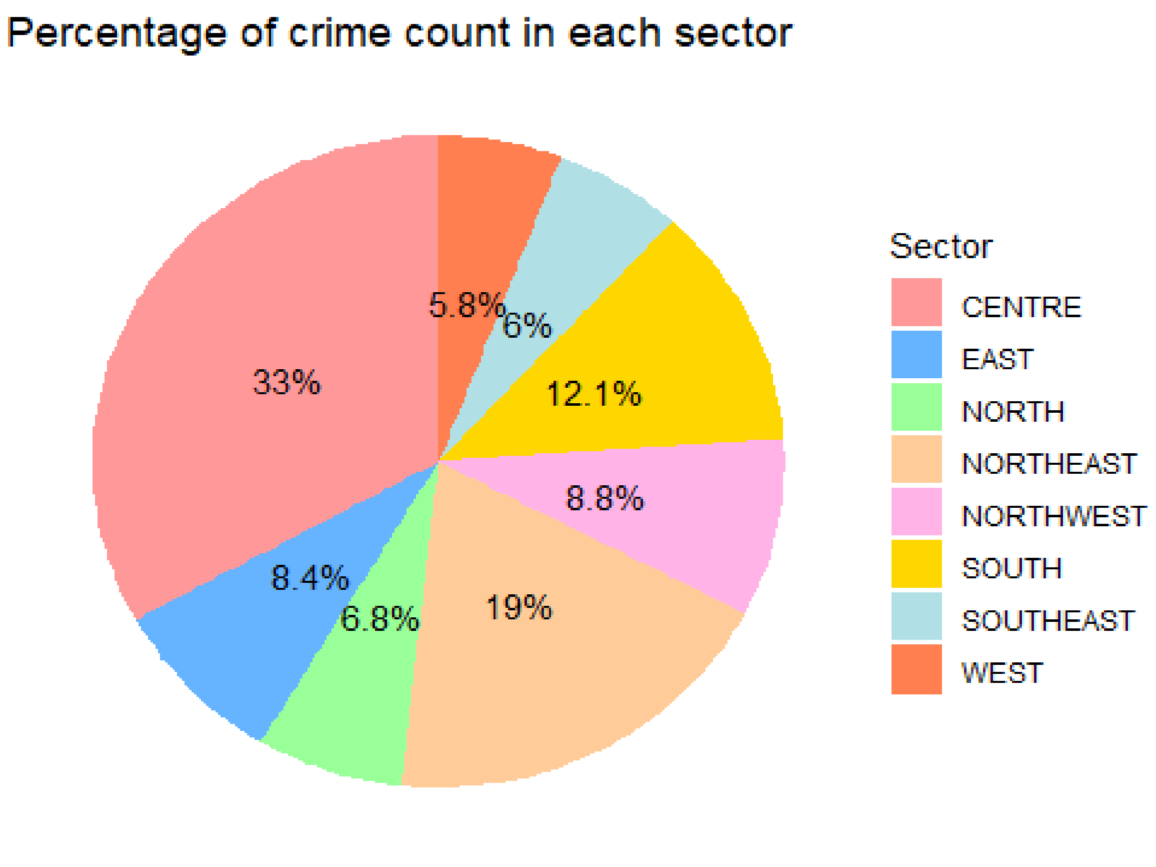
```
library(ggplot2)
library(dplyr)
library(scales)

sector_crime_summary <- crime_data %>%
  group_by(Sector) %>%
  summarize(Total_Crime = sum(Crime.Count, na.rm = TRUE))

ggplot(sector_crime_summary, aes(x = reorder(Sector, -Total_Crime), y = Total
_Crime, fill = Sector)) +
  geom_bar(stat = "identity", width = 0.7) +  # Adjust bar width
  geom_text(aes(label = comma(Total_Crime)), vjust = -0.3, color = "black") +
# Add formatted text labels
  scale_fill_manual(values = c("#FF9999", "#66B3FF", "#99FF99", "#FFCC99", "#
FFB3E6", "#FFD700", "#B0E0E6", "#FF7F50")) +  # Light colors
  labs(title = "Area with Highest and Lowest Crime", x = "Sector", y = "Total
Crime Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        plot.title = element_text(hjust = 0.5),  # Center the title
        legend.position = "none")  # Remove legend if not necessary
```



Below is the code for the analysis

Pie Chart

```
library(ggplot2)
library(dplyr)

sector_crime_summary <- crime_data %>%
  group_by(Sector) %>%
  summarize(Total_Crime = sum(Crime.Count, na.rm = TRUE))

ggplot(sector_crime_summary, aes(x = "", y = Total_Crime, fill = Sector)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +  # Transform to pie chart
  geom_text(aes(label = paste0(round(Total_Crime / sum(Total_Crime) * 100, 1)
, "%")),
            position = position_stack(vjust = 0.5), color = "black") +  # Add
percentage labels
  scale_fill_manual(values = c("#FF9999", "#66B3FF", "#99FF99", "#FFCC99", "#
FFB3E6", "#FFD700", "#B0E0E6", "#FF7F50")) +
  labs(title = "Percentage of crime count in each sector") +
  theme_void()
```



The above chart represents the percentage of crime count for each sector. We can see in the below chart that 33% of crime has occurred in the sector CENTRE and 5.8% in the WEST sector.

Determine which community has the highest crime for the sector.

The below code counts the total number of crimes count for each community that comes under the sector CENTRE.

Below is the code for the analysis

```
library(dplyr)

# Filter for the CENTRE sector and summarize total crime count by Community
total_crime_northwest <- crime_data %>%
  filter(Sector == "CENTRE") %>%
  group_by(Community) %>%
  summarize(Total_Crime = sum(Crime.Count, na.rm = TRUE))

print(total_crime_northwest)
```

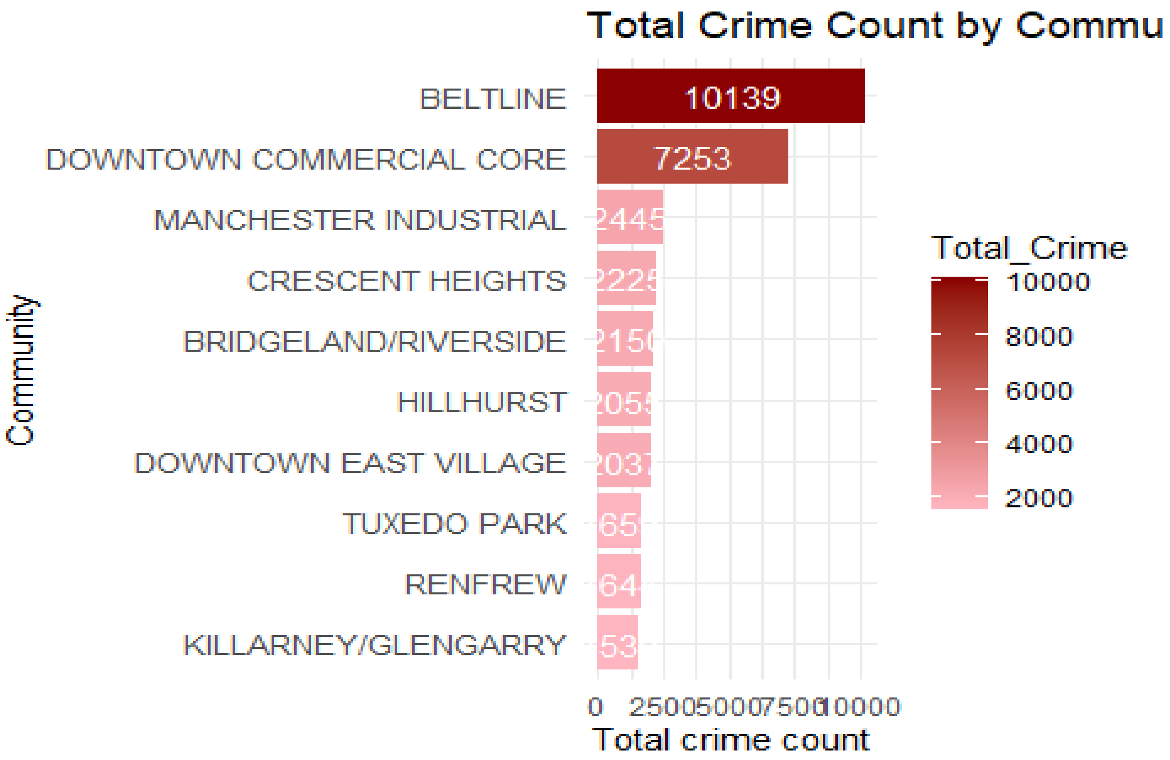
```
## # A tibble: 61 × 2
##    Community            Total Crime
##    <chr>                      <int>
##  1 ALTADORE                    1051
##  2 ALYTH/BONNYBROOK             578
##  3 BANFF TRAIL                 1325
##  4 BANKVIEW                    1490
##  5 BEL-AIRE                      57
##  6 BELTLINE                   10139
##  7 BRIDGELAND/RIVERSIDE        2150
##  8 BRITANNIA                    137
##  9 BURNS INDUSTRIAL             434
## 10 CAMBRIAN HEIGHTS             287
## # ℹ 51 more rows
```

Visualization using a bar chart

This chart represents the total crime count of the top 10 communities that come under the sector CENTRE.

Below is the code for the analysis

```
library(ggplot2)
library(dplyr)

top_communities_northwest <- crime_data %>%
  filter(Sector == "CENTRE") %>%
  group_by(Community) %>%
  summarize(Total_Crime = sum(Crime.Count, na.rm = TRUE)) %>%
  arrange(desc(Total_Crime)) %>%  # Sort in descending order
  slice_head(n = 10)  # Get top 10 communities

colors <- scales::gradient_n_pal(c("darkred", "red", "lightpink"))(seq(0, 1,
length.out = nrow(top_communities_northwest)))

ggplot(top_communities_northwest, aes(x = reorder(Community, Total_Crime), y
= Total_Crime, fill = Total_Crime)) +
  geom_bar(stat = "identity") +  # Bar color
  scale_fill_gradient(low = "lightpink", high = "darkred") +
  labs(title = "Total Crime Count by Communities in CENTRE",
       x = "Community",
       y = "Total crime count") +
  coord_flip() +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10)) +
  geom_text(aes(label = Total_Crime), position = position_stack(vjust = 0.5),
color = "white")
```



As we can see the BELTLINE community has the highest crime rate it10139 as compared to the other communities.

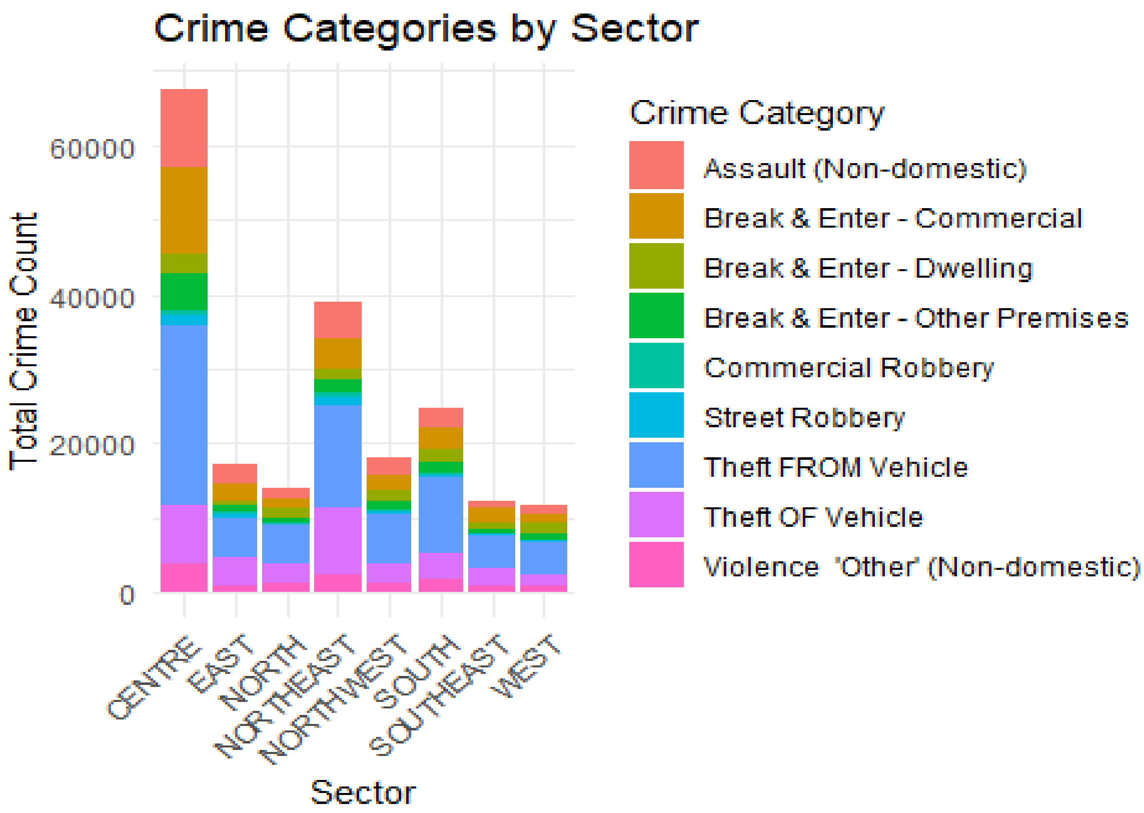
We have done only for sector CENTRE. We can also calculate the crime count for other communities that come under different sectors of Calgary to check which community has the highest crime.

The most common type of crime committed in each sector in Calgary.

Currently there are nine crime categories, which are: Assault (Non-domestic), Break & Enter – Commercial, Break & Enter – Dwelling, Break & Enter -1 Other Premises, Commercial Robbery, Street Robbery, Theft FROM Vehicle, Theft OF Vehicle, violence ‘Other’ (Non-domestic). In this, we will identify the most common type of crime committed in each sector of Calgary.

Below is the code for the analysis

```r
# Load necessary libraries
library(ggplot2)
library(dplyr)

# Load the dataset
crime_data <- read.csv("Community_Crime_Statistics_Sectors.csv")

# Summarize the data: count of crimes by sector and category
crime_summary <- crime_data %>%
  group_by(Sector, Category) %>%
  summarise(Total_Crime_Count = sum(Crime.Count, na.rm = TRUE)) %>%
  ungroup()

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

# Create the stacked bar chart
ggplot(crime_summary, aes(x = Sector, y = Total_Crime_Count, fill = Category)
) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Crime Categories by Sector",
       x = "Sector",
       y = "Total Crime Count",
       fill = "Crime Category") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The above-stacked bar chart shows that “Theft From a Vehicle” is the most common type of crime that occurs almost in every sector.

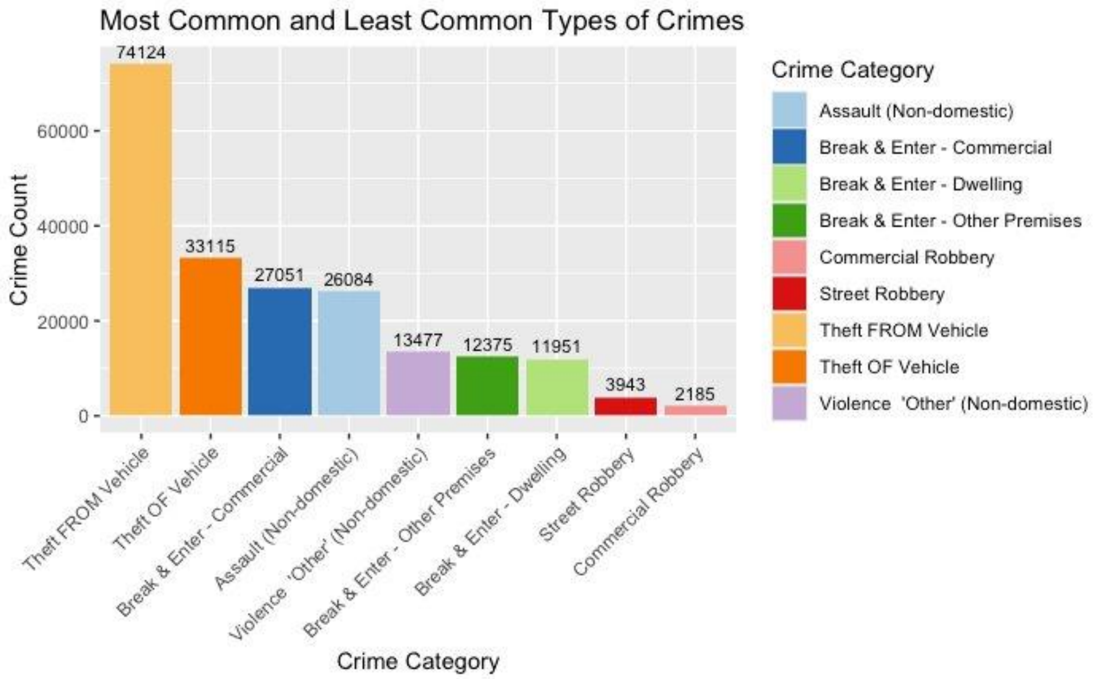
Question 2) Most common and least common types of crimes in Calgary

Understanding the most common crimes helps residents take proactive steps to protect themselves. For instance, if a crime like theft from a vehicle is the most common, people might invest in better vehicle security systems for their cars and take measures when parking in public areas. Moreover, knowing which crimes are least common can ease some anxieties about certain dangers, enabling a more balanced perspective of safety in the city. Lastly, it can influence law enforcement campaign strategies. For instance, if theft from a vehicle was the most common crime, they would invest in programs that prevent vehicle crimes.

Identify the most common and least common types of crimes in the dataset

Below is the code for the analysis

```
# Count the total number of crimes for each category
crime_counts <- crime_data %>%
  group_by(Category) %>%
  summarise(Frequency = sum(Crime.Count)) %>%
  arrange(desc(Frequency))

# Show the most common and least common types of crimes
most_common_crime <- crime_counts[1, ]
least_common_crime <- crime_counts[nrow(crime_counts), ]

print("Most Common Crime:")

## [1] "Most Common Crime:"

#print(most_common_crime)

## # A tibble: 1 × 2
##   Category           Frequency
##   <chr>                  <int>
## 1 Theft FROM Vehicle     74124

print("Least Common Crime:")

## [1] "Least Common Crime:"

print(least_common_crime)

## # A tibble: 1 × 2
##   Category           Frequency
##   <chr>                  <int>
## 1 Commercial Robbery      2185

# Visualize the data using a bar chart
ggplot(crime_counts, aes(x = reorder(Category, -Frequency), y = Frequency, fi
ll = Category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Frequency), vjust = -0.5, size = 3) +
  labs(title = "Most Common and Least Common Types of Crimes",
       x = "Crime Category",
       y = "Crime Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Paired") +
  guides(fill = guide_legend(title = "Crime Category"))
```



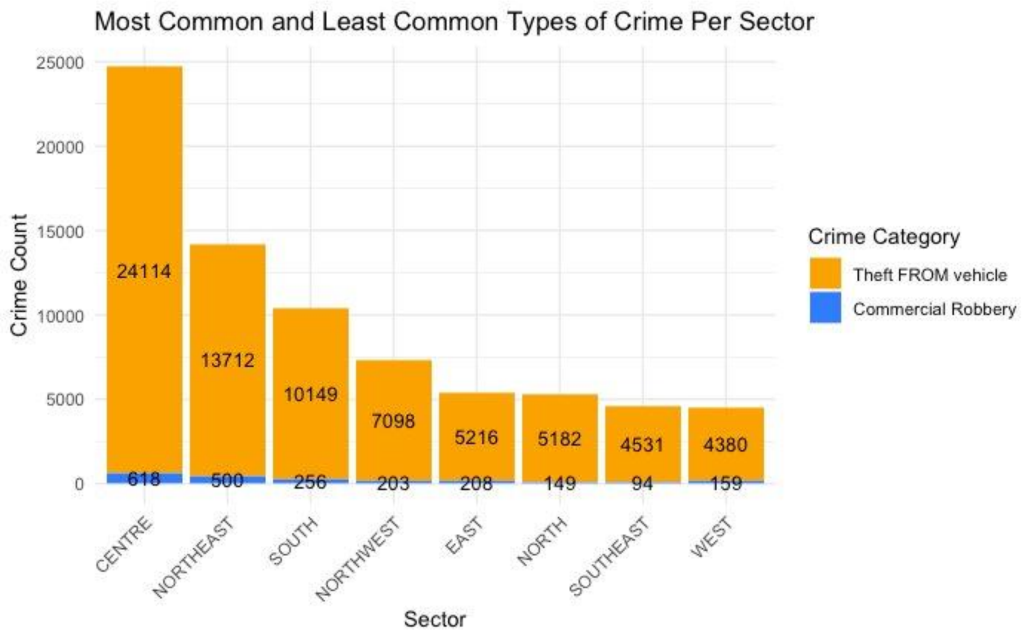
The bar chart above shows a visual rep representation of crime categories per crime count As illustrated, Theft FROM vehicle is the most common crime with 74,124 counts and the least common crime is commercial Robbery with 2,185 counts.

Analyze the most common and the least common types of crime per sector

Below is the code for the analysis

```
merged_data <- crime_data
merged_data$Category <- tolower(merged_data$Category)

# Calculate the total crime count per sector and category
crime_counts_per_sector <- merged_data %>%
  group_by(Sector, Category) %>%
  summarise(Frequency = sum(Crime.Count)) %>%
  ungroup()

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

```r
# Correct category names from the dataset (assuming correct case-sensitive names)
correct_category_names <- crime_counts_per_sector %>%
  mutate(Category = case_when(
    Category == "theft from vehicle" ~ "Theft FROM vehicle",
    Category == "commercial robbery" ~ "Commercial Robbery",
    TRUE ~ Category
  ))

# Filter the data to include only 'Theft FROM vehicle' and 'Commercial Robbery'
filtered_crimes <- correct_category_names %>%
  filter(Category %in% c("Theft FROM vehicle", "Commercial Robbery"))

# Set the factor levels for Category to ensure proper stacking order
filtered_crimes$Category <- factor(filtered_crimes$Category,
                                   levels = c("Theft FROM vehicle", "Commercial Robbery"))

# Plot the data as a stacked bar chart with the correct stacking order
ggplot(filtered_crimes, aes(x = reorder(Sector, -Frequency), y = Frequency, fill = Category)) +
  geom_bar(stat = "identity", position = "stack") +
  geom_text(aes(label = Frequency), position = position_stack(vjust = 0.5), size = 3.5, color = "black") +
  labs(title = "Most Common and Least Common Types of Crime Per Sector",
       x = "Sector",
       y = "Crime Count") +
  # Use the desired colors: Theft FROM vehicle in orange, Commercial Robbery in blue
  scale_fill_manual(values = c("Commercial Robbery" = "dodgerblue", "Theft FROM vehicle" = "orange")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  guides(fill = guide_legend(title = "Crime Category"))
```



As illustrated in the bar chart above, the most common crime which is Theft From a vehicle is highest in the CENTER sector with 24,114 counts and lowest in the WEST sector with 4,380 counts.

For the least common crime which is Commercial Robbery is highest in the CENTER with 618 counts and lowest in the SOUTHEAST with 94 counts.

Question 3) Analyze the behaviour of crimes over the years:

There are 7 years in this data set, However, there are only 6 complete years since the year 2024 is incomplete. For this project, the analysis will be study only for the complete years, in other words, the year 2024 will be excluded. However, the current trend in 2024 will be analyzed to have a preview of the behaviour of the crimes in contrast with the previous years.

Display counts of crimes over data set years (**2018-2024**):

Below is the code for the analysis

```
grouped_data <- crimes_df %>%
  group_by(Year) %>%
  summarize(Total_Crimes = sum(Crime.Count), .groups = 'drop')

# View the result
print(grouped_data)
```

```
## # A tibble: 7 × 2
##    Year Total_Crimes
##   <int>        <int>
## 1  2018        35437
## 2  2019        38299
## 3  2020        31918
## 4  2021        28573
## 5  2022        33014
## 6  2023        27973
## 7  2024         9091
```

The code previously executed shows the total number of crimes including the incomplete year 2024 and the year 2024 could make the data behave strangely due to the low number of crime occurrences, therefore it was excluded from the dataset for better analysis

Display of crimes counts of the interested years (**2018-2023**):

Below is the code for the analysis

```r
# Filter out the year 2024
crimes_no_2024 <-  crimes_df %>%
  filter(Year != 2024)



# Group the crimes by year excluding 2024
grouped_data_no_2024 <- crimes_no_2024 %>%
  group_by(Year) %>%
  summarize(Total_Crimes = sum(Crime.Count), .groups = 'drop')

# replace values 1 to 6 to the years
grouped_data_no_2024$Year_Index <- 1:6


# View the result
print(grouped_data_no_2024)

## # A tibble: 6 × 3
##    Year Total_Crimes Year_Index
##   <int>        <int>      <int>
## 1  2018        35437          1
## 2  2019        38299          2
## 3  2020        31918          3
## 4  2021        28573          4
## 5  2022        33014          5
## 6  2023        27973          6
```

From the information above it is possible to calculate some statistics. What has been the year with most crimes and what has been the year with the lowest crimes over that timeline?

Below is the code for the analysis

```r
total_crimes_count <- sum(grouped_data_no_2024$Total_Crimes)
total_crimes_count

## [1] 195214

# Probabilities
grouped_data_no_2024$Percentage <- (grouped_data_no_2024$Total_Crimes / total
_crimes_count) * 100
#grouped_data_no_2024

year_max_crimes <- grouped_data_no_2024[which.max(grouped_data_no_2024$Percen
tage), ]
```

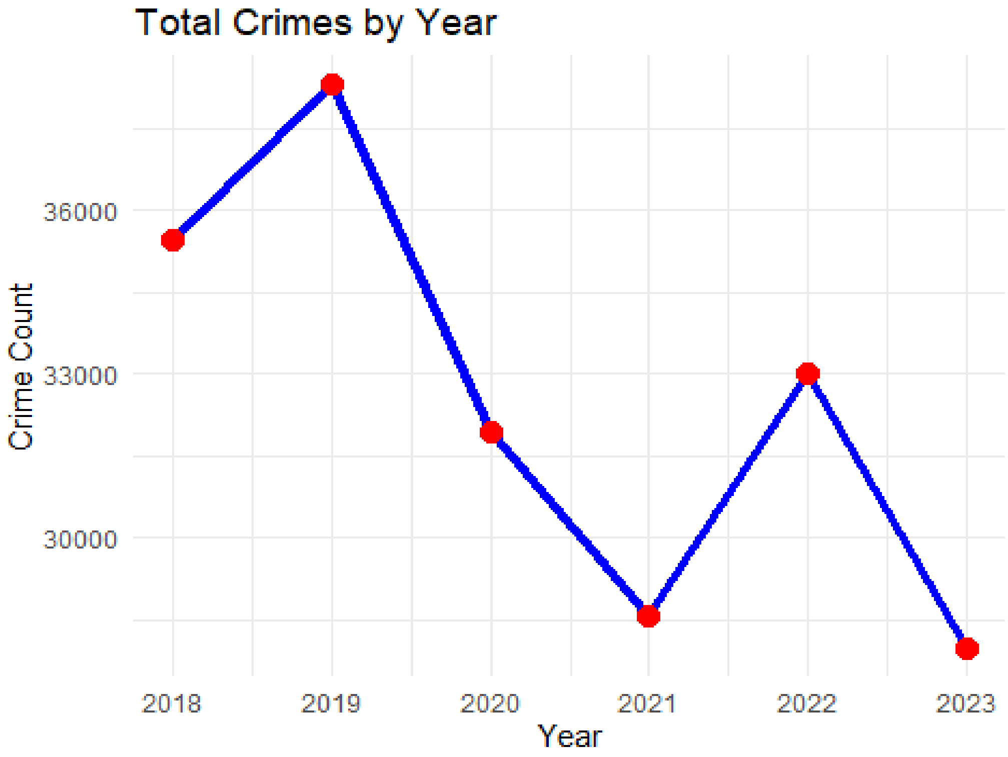
```
year_min_crimes <- grouped_data_no_2024[which.min(grouped_data_no_2024$Percen
tage), ]


year_max_crimes

## # A tibble: 1 × 4
##    Year Total_Crimes Year_Index Percentage
##   <int>        <int>      <int>      <dbl>
## 1  2019        38299          2       19.6

year_min_crimes

## # A tibble: 1 × 4
##    Year Total_Crimes Year_Index Percentage
##   <int>        <int>      <int>      <dbl>
## 1  2023        27973          6       14.3
```

After running the previews code, it is clear to see that the highest year with most crimes committed was **2019** with **38299** out of **195214** having a total percentage of **19.62%** of the total timeline from **2018** to **2023**. On the other hand, we can see that the year of **2023** has the lowest crime rate with **27973** out of **195214** having a total percentage of **14.33** of the total timelines. This can be interpreted as if the crimes were decreasing over the years, but is this true?

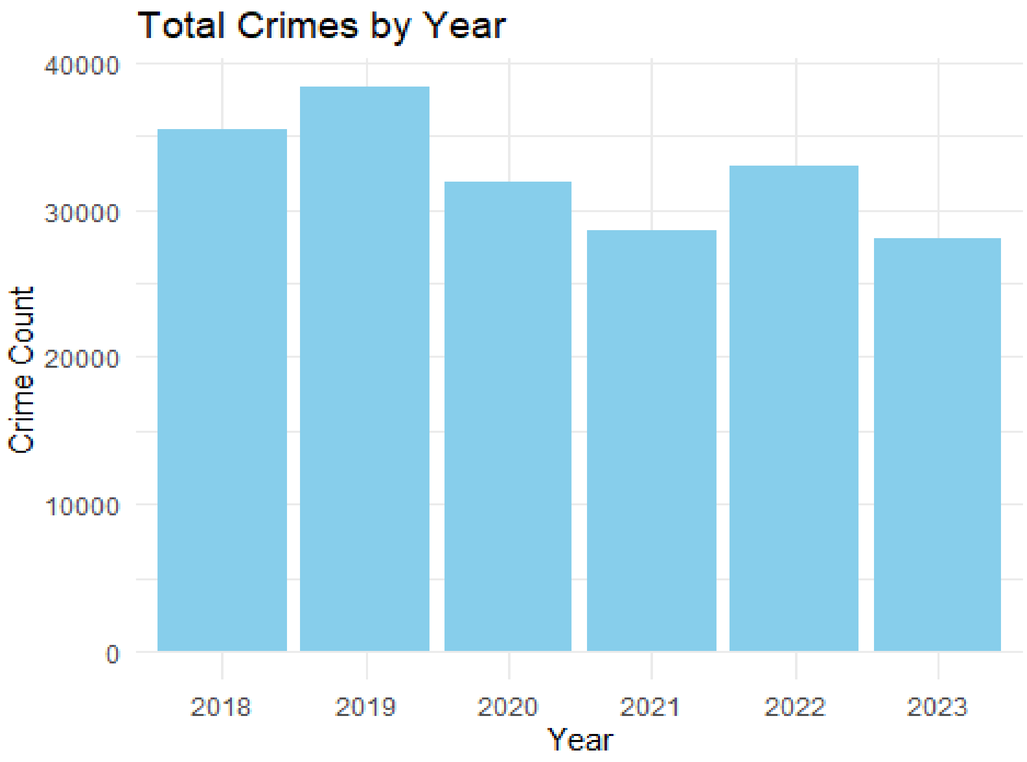
Let's represent the crime counts in a graph where it could be more evident its behavior.

Below is the code for the analysis

```
library(ggplot2)

#Excluding 2024 Line
ggplot(grouped_data_no_2024, aes(x = Year, y = Total_Crimes)) +
  geom_line(color = "blue", size = 1.5) +
  geom_point(color = "red", size = 3.5) +  # Optional: add points on the line
  labs(title = "Total Crimes by Year",
       x = "Year",
       y = "Crime Count") +
  theme_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle:last_lifecycle_warnings()` to see where this warning was
## generated
```



```r
# Bar chart
ggplot(grouped_data_no_2024, aes(x = factor(Year), y = Total_Crimes)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Total Crimes by Year",
       x = "Year",
       y = "Crime Count") +
  theme_minimal()
```



In the charts, we can see the behaviour of crimes over time. It is clear how the overall crime count behaves, so it is faster to identify the highest and lowest points in the crime counts, once again displaying the highest point in the crime count in **2019** and the lowest in **2023**.

The results take us to another question have the crimes decreased over the years, is there a relation between the years and the crimes? is it behaving lineal?

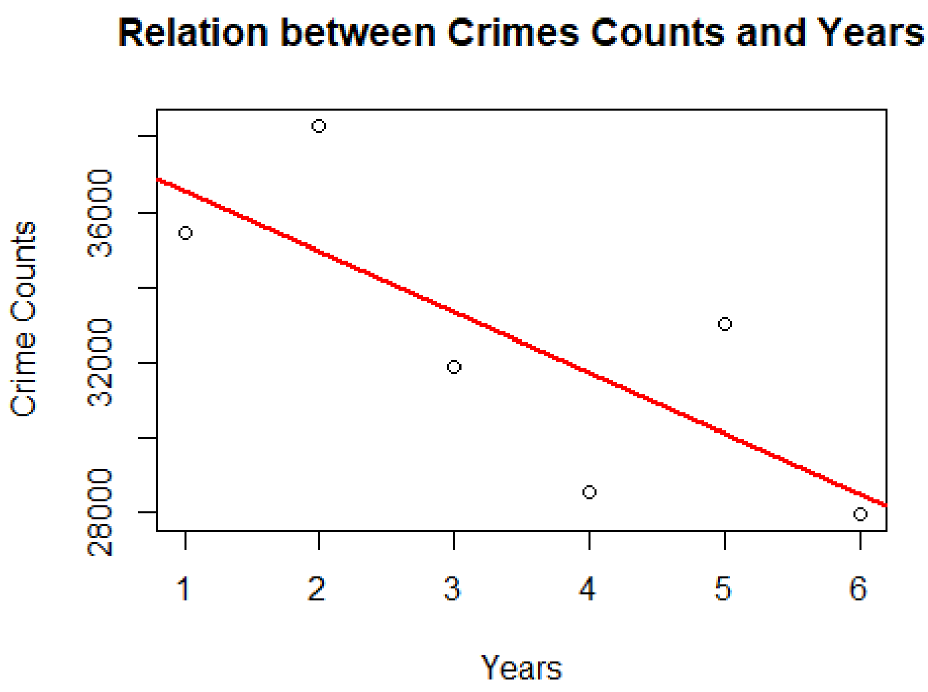
Regression

```
# Regression model
crime_model <- lm( Total_Crimes~ Year_Index, data = grouped_data_no_2024)



# Summary of the regression model
summary(crime_model)

##
## Call:
## lm(formula = Total_Crimes ~ Year_Index, data = grouped_data_no_2024)
##
```

```
## Residuals:
##       1       2       3       4       5       6 
## -1135.8  3341.0 -1425.1 -3155.2  2900.6  -525.5 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  38187.7     2679.2  14.254 0.000141 ***
## Year_Index   -1614.9      687.9  -2.347 0.078742 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2878 on 4 degrees of freedom
## Multiple R-squared:  0.5794, Adjusted R-squared:  0.4742 
## F-statistic:  5.51 on 1 and 4 DF,  p-value: 0.07874
```

To obtain the regression line we apply the following code:

```
plot(grouped_data_no_2024$Year_Index, grouped_data_no_2024$Total_Crimes ,xlab
="Years",
     ylab="Crime Counts",main ="Relation between Crimes Counts and Years")
#plot the regression line
abline(lm(Total_Crimes~ Year_Index, data= grouped_data_no_2024),col="red",lwd
=2)
```



According to the graph and the regression line we tend to decrease.

To investigate the strength between variables year and crime count to define how strung is the dependency of variables, we calculate the correlation value:

```
Correlation_Val = cor(grouped_data_no_2024$Year_Index, grouped_data_no_2024$T
otal_Crimes)
Correlation_Val

## [1] -0.7611798
```

The correlation value is **-0.7611**, this means that there is a strong negative correlation between the variables year and crime counts.

Let’s see how the crimes behave in a plot:

```
# beta0 / beta1
# Regression model
R_year <- lm( Total_Crimes~ Year_Index , data = grouped_data_no_2024)
R_year

##
## Call:
## lm(formula = Total_Crimes ~ Year_Index, data = grouped_data_no_2024)
##
## Coefficients:
## (Intercept)   Year_Index
##       38188        -1615
```

We got the values: **38188** for β_0 and **-1615** for β_1. In other words, our minimal crimes when there is no trend will have a value of **38188** and a decremented value of **-1615**

So now we can try to predict the crimes that we could have in 2024.

$$Y = \beta_0 + \beta_1 * X + E$$

```
predic_crimes_2024_decremental <- 38188 -1615 * (7)
predic_crimes_2024_decremental

## [1] 26883
```

We got a value of **26883**, which means that there will be a decrease in crimes in 2024. Running the above we got the result is **26883**, in other words, the prognosticated crime counts for 2024 will be the ones calculated keeping in mind that this is calculated with a **-0.7611** correlation and having in consideration that we have a Multiple R-squared of **0.5794** in other words is a **57.94%** efficiency.

Hypothesis Test 1

Accordingly, to the regression line and interpreting as a negative correlation we test the hypothesis as β_0 less that 0, to verify that in fact the data behaves negatively.

$$H_o: \beta_1 = 0 \hspace{ .5cm } VS \hspace{ .5cm } H_a: \beta_1 < 0 $$

From the calculation before:

```
summary(crime_model)

##
## Call:
## lm(formula = Total_Crimes ~ Year_Index, data = grouped_data_no_2024)
##
## Residuals:
##       1       2       3       4       5       6
## -1135.8  3341.0 -1425.1 -3155.2  2900.6  -525.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38187.7     2679.2  14.254 0.000141 ***
## Year_Index   -1614.9      687.9  -2.347 0.078742 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2878 on 4 degrees of freedom
## Multiple R-squared:  0.5794, Adjusted R-squared:  0.4742
## F-statistic:  5.51 on 1 and 4 DF,  p-value: 0.07874
```

Running the summary, we got the t_{value} or $t_{observed}$, and the df.

```
t_value <- -2.347
df <- 4
```

```r
p_value <- pt(-2.347, 4)
p_value

## [1] 0.03938657
```

Having the p_value equal to **0.039** approximately **0.04**. Then we can say that the **p_value** is < **alpha** (0.05), therefore we reject the null hypothesis in favor of the alternative, and then we can say that the behaviour of the crimes has a negative linear tendency.

In conclusion, we can say that the yearly crimes have decreased, an approximately **-1615** over the years with a **58%** present of model accuracy and a **-0.7611** strong negative linear relationship (correlation value) between years and crimes. In other words, this model is only 58% right, is not the best to predict the actual change over the years.

Question 4) Analyze the behaviour of crimes over the years:

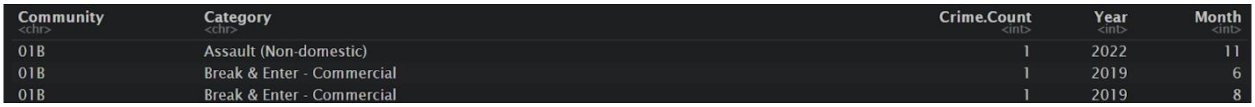
The objective of this analysis will be to conclude if the lockdown during 2020 and 2021 had any impact on the crime rates in Calgary. Specifically, we will be looking at three different lockdown phases which were March 2020 to May 2020, November 2020 to January 2021, and April 2021 to June 2021.

The reason why we are interested in the question is to explore the impacts crime of in Calgary. If we were to encourage Calgarians to stay indoors and significantly reduce the number of people travelling into Calgary, we would probably assume that total crime rates are going to decrease. However, people who were doing crime before the lockdown would likely continue to do crime during lockdown. So, this raises a good question of whether crime rates changed during lockdown. Analyzing the answer to this question could help policymakers create better rules around future lockdowns. For example, suppose we see that crime rates increase during lockdowns. In that case, we can then plan and make sure that during any future lockdowns, government officials can remind Calgarians to lock their doors and to frequently check up on their vehicles for possible break-ins.

EDA and Data Visualization:

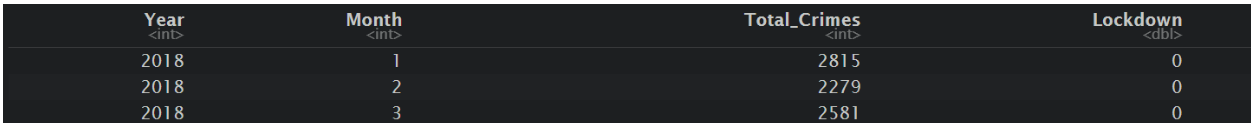
Before we begin any analysis, we should first prepare our data for any future steps. Currently, our dataset does not have any identification of whether a month had a lockdown or not. So, we need to create a new data frame with the three lockdown phases mentioned in the introduction and then loop through each row of our dataset to tag each month. After this step, we will then have a dataset with an extra column at the end that identifies our lockdown months. We also need to combine our data, so we only have one monthly row.

Before data cleaning:



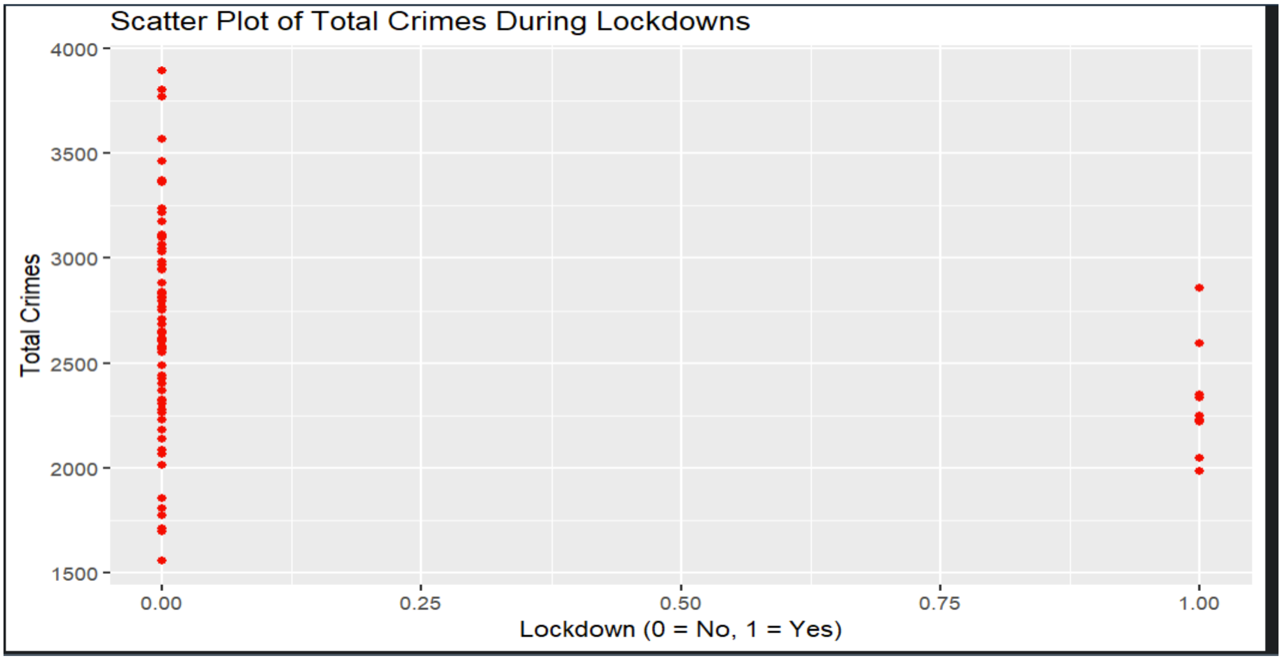
Community <chr>	Category <chr>	Crime.Count <int>	Year <int>	Month <int>
01B	Assault (Non-domestic)	1	2022	11
01B	Break & Enter - Commercial	1	2019	6
01B	Break & Enter - Commercial	1	2019	8

After data cleaning:



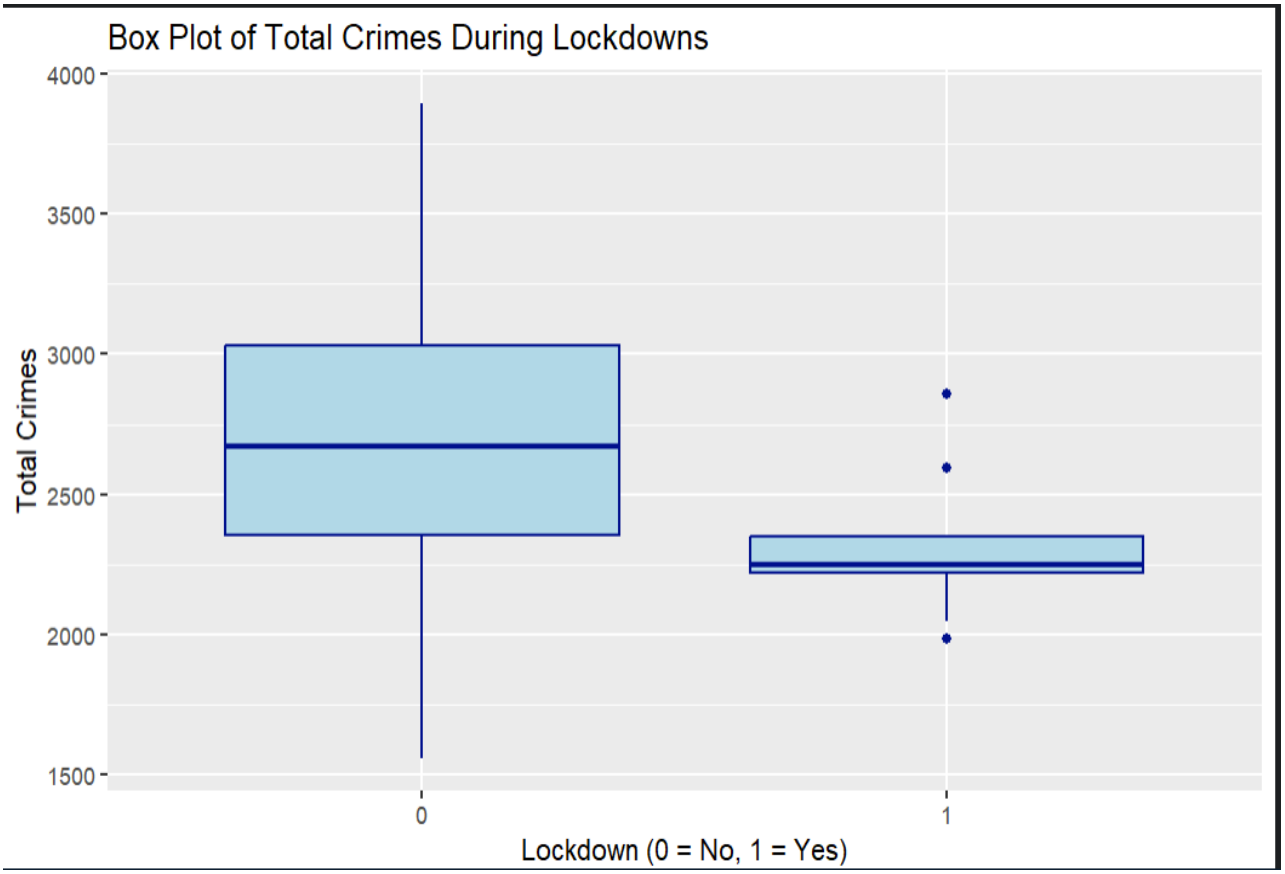
Year <int>	Month <int>	Total_Crimes <int>	Lockdown <dbl>
2018	1	2815	0
2018	2	2279	0
2018	3	2581	0

Now since we are curious if lockdowns had any impact on total crime in Calgary, we should first graph our results to check if there is any sign of a relationship. We will first do a scatter plot comparing the total crimes of months that had lockdowns (1) versus months that did not have lockdowns (0).



It is tough to draw any conclusions about the relationship because the lockdown = 0 group has such a large spread of crime rates. So, let's graph the data in a box plot and see the results. As we can see, there seems to be a stronger relationship shown in the graph where the lockdown = 1 group seems to have fewer total crimes.

Using the correlation function in R Studio, we can conclude that the relationship between lockdown and total crimes is -0.239. However, we are currently comparing years 2018-2024 and most of these years don't have any lockdown months. Therefore, let’s redo our visualizations and see if we get any stronger relationships when we just use the years 2020 and 2021 since these are the only two years that have data for both lockdown and non-lockdown months.



Right away we can see a stronger rift between the lockdown groups. What is even more significant is that our correlation is now -.452, which is more than double the correlation from before. Since our correlation seems to be significantly stronger when we only use years that have both lockdown and non-lockdown months, we will continue this analysis with only 2020 and 2021 data

Hypothesis Test 2

Since we are interested in if Covid-19 lockdowns have an impact on total crimes in Calgary, we can build a hypothesis test using these:

- Null Hypothesis: The COVID-19 lockdowns had no effect on the total crimes in Calgary
- Alternative Hypothesis: COVID-19 did influence the total crimes in Calgary

First, we need to check if our data is normal to decide if we can use a Test or not. Using the Shapiro test where the null hypothesis is that our data is normal and the alternative is that our data is not normal, we get a p-value of 0.584. This p-value is significantly higher than the significance level of 0.05 so we fail to reject the null hypothesis and conclude that our total crimes are normally distributed.

Now using a T.test, we can conclude that our p-value of 0.018 is less than our significance level of 0.05. This means that we reject our null hypothesis and accept our alternative that the months during covid lockdown did have an impact on total crimes in Calgary. We are 95% confident that the true difference in means between non-lockdown months and lockdown months is between 58.65 and 589.22. This means that we are 95% confident that there will be between 59 and 589 more crimes during non-lockdown months compared to lockdown months. Since our 95% confidence interval has both lower and upper limits above 0, we can conclude that lockdown months have a negative relationship with crime meaning that there will be less crime when it is a lockdown.

Regression Analysis:

Building off our original question of whether Covid-19 lockdowns had any impact on total crimes in Calgary, we can create these points:

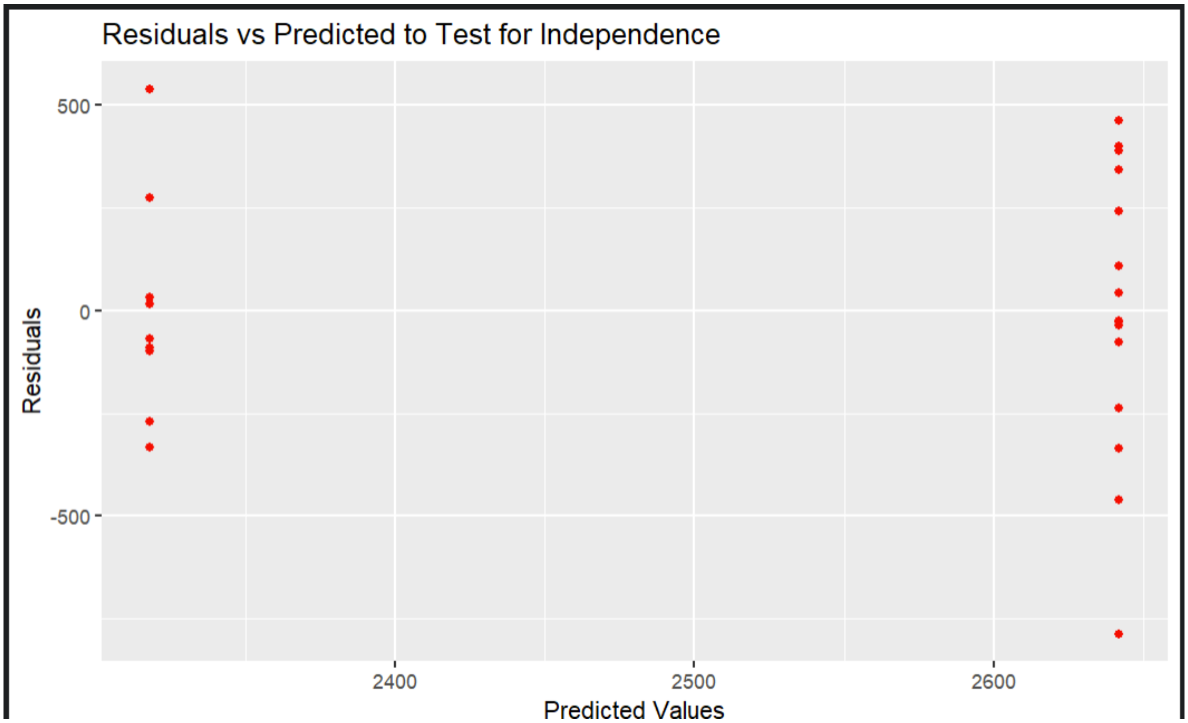
- The Null Hypothesis is that the Covid-19 lockdowns had no effect on the total crimes in Calgary
- Our Alternative Hypothesis is that COVID-19 did have an effect on the total crimes in Calgary
- If our linear regression model has a negative slope for lockdowns, then we would predict that the lockdowns decreased the total crimes
- If our linear regression model has a positive slope for lockdowns, then we would predict that the lockdowns increased the total crimes

Since we already know that there is a relatively decent relationship between lockdowns and total crime rates from our EDA, we can build a model where our y variable is total crimes, and our x variable is lockdowns (where lockdown can either be 0 or 1). Lockdown = 0 means that it was a normal month and lockdown = 1 means it had a lockdown.

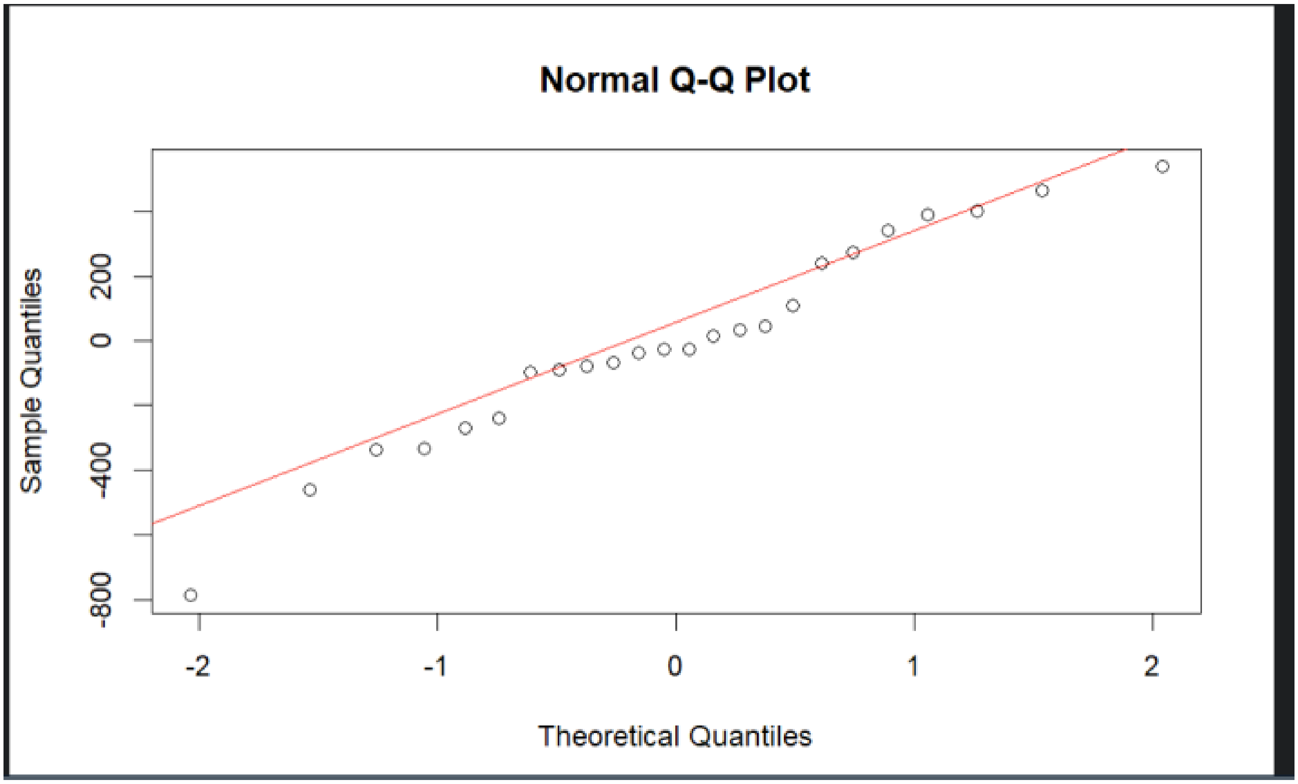
So, our linear regression model is: Total Crimes = 2641.93 - 323.93 * Lockdown

This means that when there is not a lockdown in a month, we predict there to be a total of 2641.93 crimes in Calgary. However, if there is a lockdown, then we predict the total number of crimes in Calgary to be 2318 for that month.

Before we use the p-value gathered from the previous code, we need to check if our model is unbiased. We need to check that the residuals are independent and normally distributed by graphing the residuals versus the predicted y values and using a Q-Q plot.



Since our graph does not indicate any sign of a pattern, we can conclude that the errors are independent.



There seems to be a relatively decent linear relationship so we can conclude it's decently normal.

With a p-value of 0.02674, we can conclude that it is less than our significance level of 0.05 and we reject our null hypothesis and accept our alternative hypothesis that the lockdown months did have an impact on total crimes in Calgary.

Conclusion and Recommendations:

In both our hypothesis test and regression analysis, we concluded that lockdowns did have an impact on crimes in Calgary. We are 95% confident that there will be between 59 and 589 more crimes during non-lockdown months compared to lockdown months. We also concluded that when there is not a lockdown in a month, we predict there to be a total of 2641.93 crimes in Calgary. However, if there is a lockdown, then we predict the total number of crimes in Calgary to be 2318 for that month. So, it seems like months which had lockdowns saw a rough drop of about 300 crimes in that month.

When doing a bit of exploratory analysis on our data before beginning the project, we noticed that the total crime rates from 2018-2019 were significantly higher than crime rates from 2022-2024. It would be interesting to explore why the crime rates have dropped so significantly since COVID-19, and in future, we would like to explore this further. Moreover, we, recommend that policymakers explore why crimes continued to exist when most people were in lockdown during these months. Were people not listening to the policies or did the people who commit most crimes not have a home to be locked down in? It surprising that the total crime rates only dropped about 300 during lockdown months and this indicates that the people who commit most crimes probably are homeless or will commit crimes no matter where they are located. Maybe based on our results, the government should focus more money on rehabilitating people who commit crimes so they don't continue to commit crimes, or the government should spend more money on the homeless. In conclusion, our analysis has illuminated key crime trends and hotspots in Calgary, offering valuable insights into the city's safety overview. We trust that this information will empower policymakers to make informed, data-driven decisions that enhance public safety and contribute to a more secure community for all residents.

References

[1] youmoveme, “6 ways to learn about a new city before you move”, youmoveme.com, May 25, 2023. [Online]. Available: https://youmoveme.com/blog/6-ways-to-learn-about-a-new-citybefore-you-move/. [Accessed Sept. 21, 2024].

[2] The City of Calgary, “City of Calgary’s Open Data Portal,” data.calgary.ca, 2024. [Online]. Available: https://data.calgary.ca/. [Accessed Sept. 21, 2024].

[3] The City of Calgary, “Community Crime Statistics,” https://data.calgary.ca, Sept. 16, 2024. [Online]. Available: https://data.calgary.ca/Health-and-Safety/Community-Crime-Statistics/78ghn26t/about_data. [Accessed Sept. 21, 2024].

[4] The City of Calgary, “Community Sectors,” opencalgary.com, Feb. 1, 2023. [Online]. Available: https://data.calgary.ca/Base-Maps/Community-Sectors/mz2j-7eb5/about_data [Accessed Sept. 21, 2024].