Project on Crime data Analysis

Rajvir Kaur

2024-10-15

For the first section, we have focused on the below points for the crime data analysis in the city of Calgary.

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, WEST.

Determine which community has the highest crime for particular sector.

- There are over 200 communities in Calgary and most of them are shown in this dataset. As we have two files for this project i.e Community_Crime_Statistics and Community_Sectors. We have created a data structure by mapping each community to their particular 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).

Import data

```
setwd("/Users/rajvirkaur/Downloads")

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
```

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

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, WEST. In this we will find which sector of the calgary has the highest and lowest crime.

```
library(dplyr)
```

```
##
## Attaching package: '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 x 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 x 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 x 2
##   Sector Total_Crime_Count
##   <chr>              <int>
## 1 CENTRE             67521
```

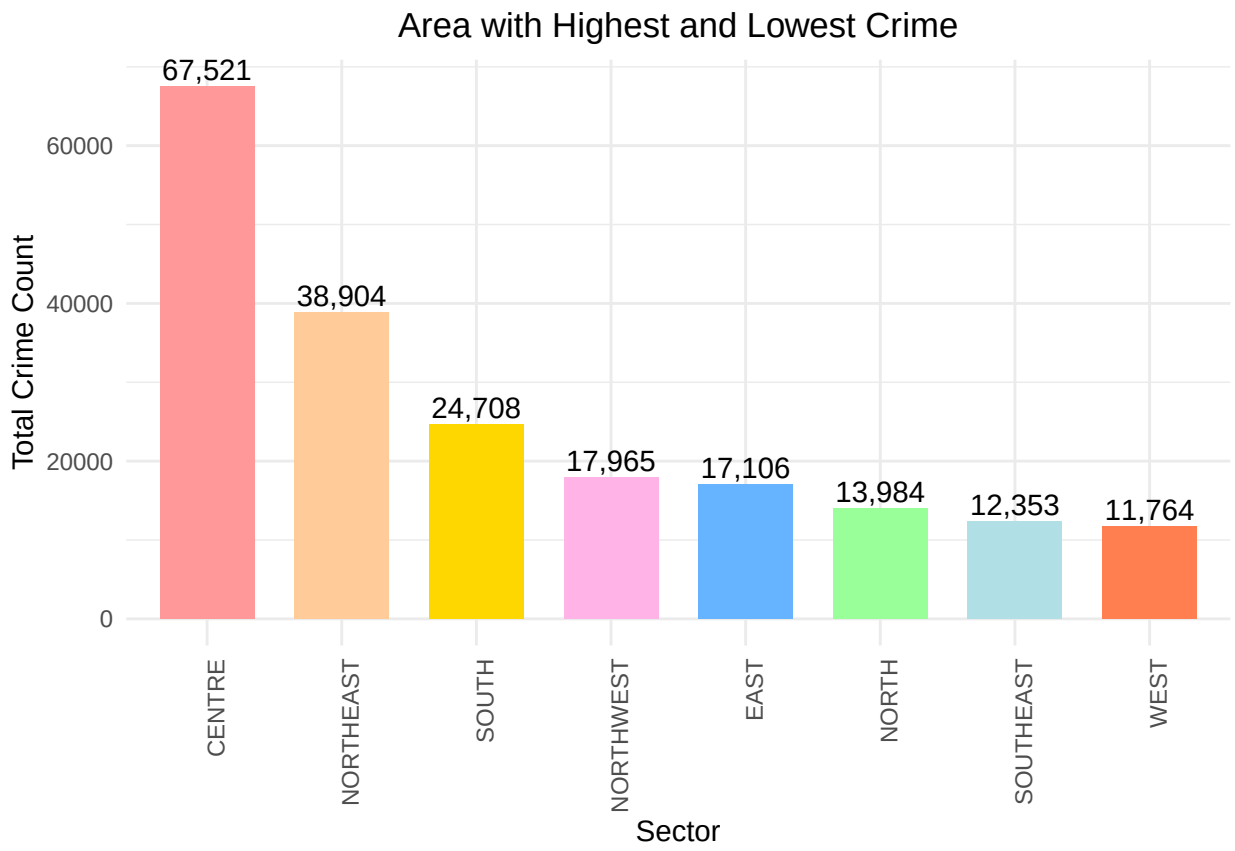
In the above, we have calculated the total crime count for for each sector. We can see that the sector CENTRE has the highest crime count i.e 67521 and the WEST sector has lowest crime with value 11764 as compare to the other sectors.

Visualization of Dataset:

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

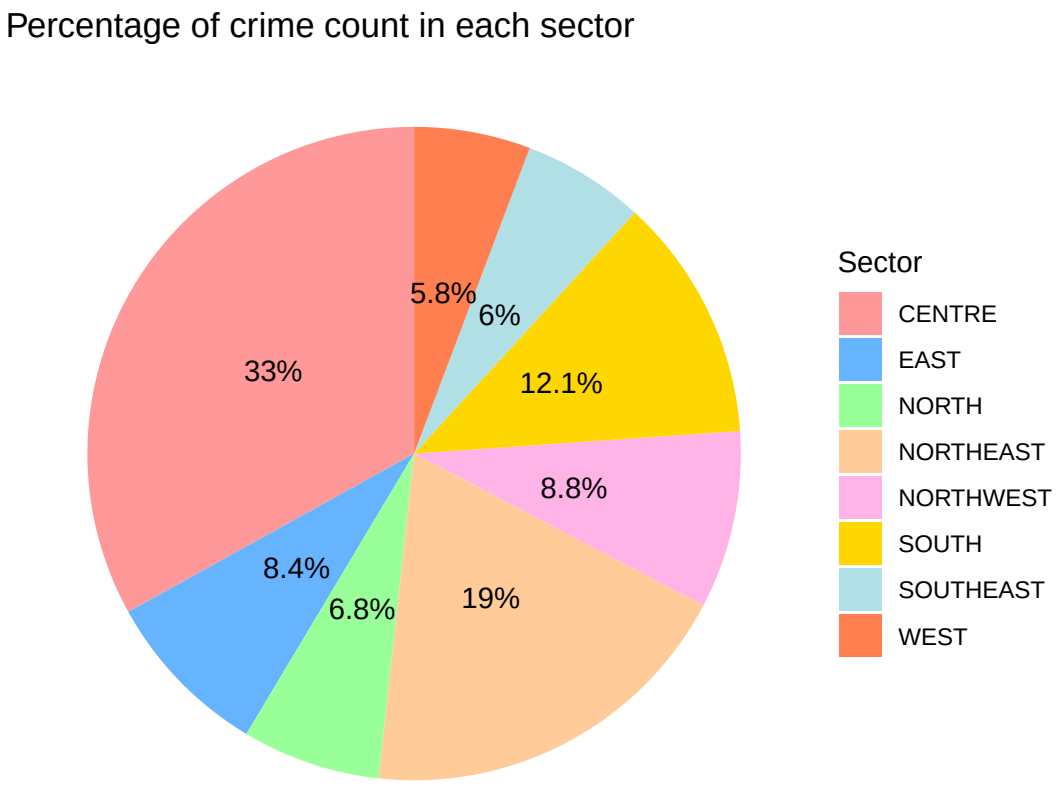
Bar Chart

```
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 lab
  scale_fill_manual(values = c("#FF9999", "#66B3FF", "#99FF99", "#FFCC99", "#FFB3E6", "#FFD700", "#B0E0
  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
```



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", "#B0E0
  labs(title = "Percentage of crime count in each sector") +
  theme_void()
```



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

Determine which community has the highest crime for particular sector.

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

```
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)
```

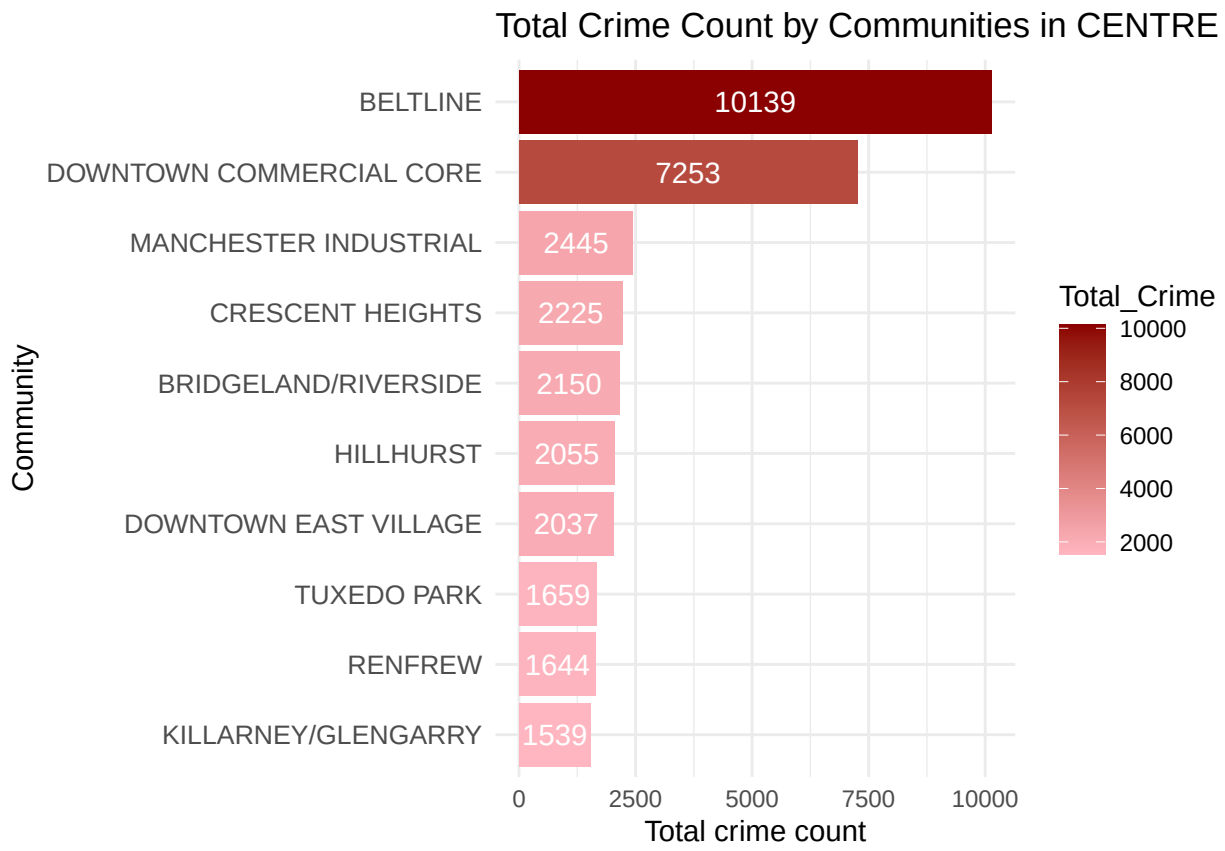
```
## # A tibble: 61 x 2
##    Community            Total_Crime
##    <chr>                      <int>
##  1 ALTADORE                    1051
##  2 ALYTH/BONNYBROOK             578
##  3 BANFF TRAIL                 1325
##  4 BANKVIEW                    1490
```

```
##  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 bar chart

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

```
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_commu

ggplot(top_communities_northwest, aes(x = reorder(Community, Total_Crime), y = Total_Crime, fill = Total
  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 that the BELTLINE community has the highest crime rate i.e 10139 as compare to the other communities.

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

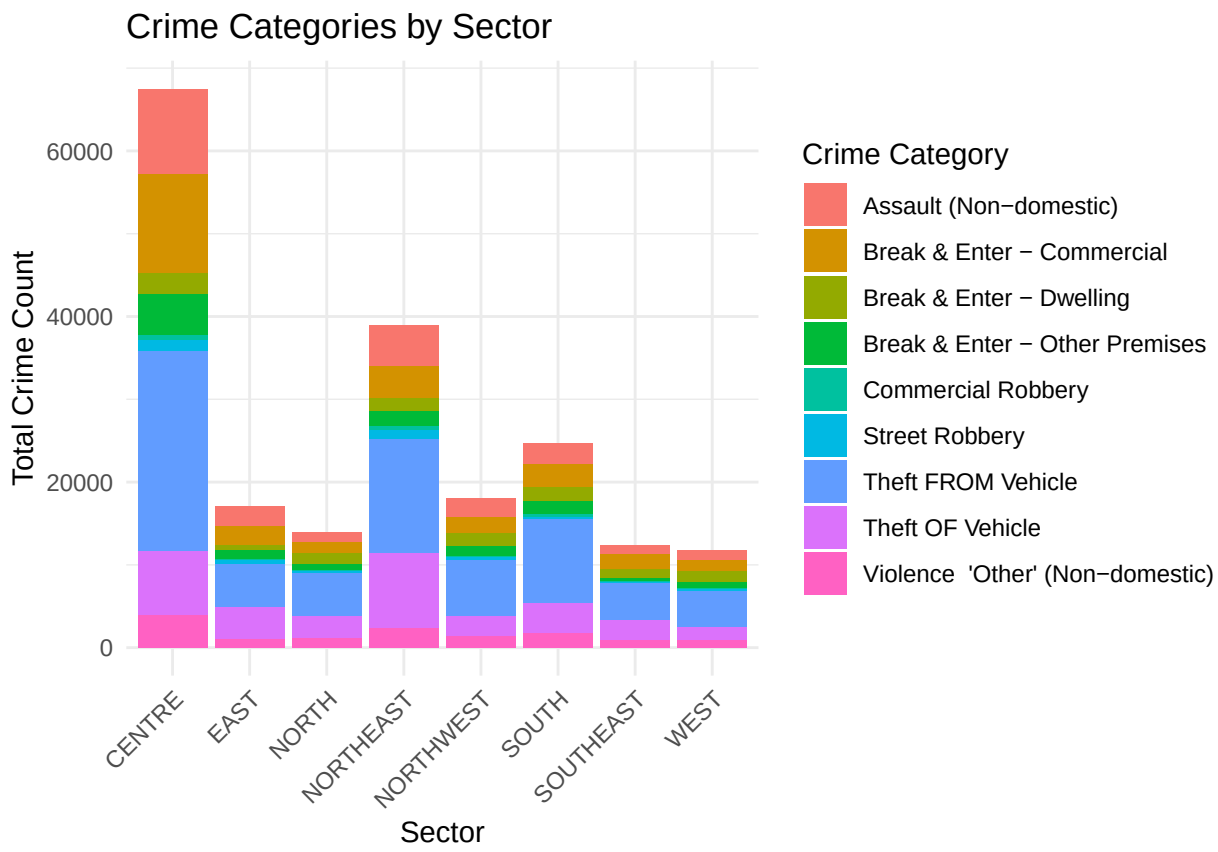
Identify the most common type of crime committed in each sector of Calgary.

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.

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

# Load the dataset
setwd("/Users/rajvirkaur/Downloads")

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 clearly represents that the "Theft From Vehicle" is the most common type of crime that occured almost in every sector.

Apply Linear Regression model to find which sector has the highest crime as compare to other.

```
library(dplyr)

# Convert Sector to a factor: This ensures that R treats Sector as a categorical variable.
```

```r
crime_data$Sector <- as.factor(crime_data$Sector)

# Fit the linear regression model
model <- lm(Crime.Count ~ Sector, data = crime_data)

# View the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = Crime.Count ~ Sector, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -2.572  -1.576  -1.024   0.549 107.428
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.57235    0.02622  136.22   <2e-16 ***
## SectorEAST      -0.54795    0.05464  -10.03   <2e-16 ***
## SectorNORTH     -1.24168    0.05342  -23.24   <2e-16 ***
## SectorNORTHEAST -0.45104    0.04160  -10.84   <2e-16 ***
## SectorNORTHWEST -0.99598    0.05052  -19.72   <2e-16 ***
## SectorSOUTH     -1.30660    0.04336  -30.14   <2e-16 ***
## SectorSOUTHEAST -1.12136    0.05716  -19.62   <2e-16 ***
## SectorWEST      -1.60051    0.05354  -29.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.605 on 71897 degrees of freedom
## Multiple R-squared:  0.02322,    Adjusted R-squared:  0.02312
## F-statistic: 244.1 on 7 and 71897 DF,  p-value: < 2.2e-16
```

- The intercept represents the crime count for the reference sector (e.g., "CENTRE").
- The coefficients for the other sectors (e.g., "EAST", "NORTH") show how much the crime count differs from the reference sector. A negative value means that the crime count in that sector is lower compared to the reference sector.

p-values: These tell you whether the difference in crime count for a particular sector, compared to the reference, is statistically significant.

R-squared: This value indicates how well the model explains the variation in the crime count.