NYPD Shooting Incident Data Report

1: Introduciton

The dataset used in this report records every shooting incident in New York City from 2006 through to the end of the previous calendar year. The data is manually extracted, reviewed and then published on the NYPD website. Each record includes details about a shooting incident such as the location, time, and demographics of both suspects and victims. The dataset is intended to help the public analyze and understand shooting and criminal activities in NYC.

Libraries needed:

```
library(tidyverse)
library(ggplot2)
library(dplyr)
```

Importing the data:

```
shooting_data <- read_csv(
  "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

2: Tidying the Data

Here we are cleaning up the dataset by changing appropriate types and getting rid of any columns not needed. Any rows that include NA will be removed from the analysis.

```
# Converting Date
shooting_data$OCCUR_DATE <- as.Date(shooting_data$OCCUR_DATE, format = "%m/%d/%Y")
shooting_data$Month_Year <- format(shooting_data$OCCUR_DATE, "%Y-%m")

# Removing Columns
shooting_data <- shooting_data[c('Month_Year', 'STATISTICAL_MURDER_FLAG', 'BORO',
                                 'PRECINCT', 'VIC_AGE_GROUP', 'VIC_SEX', 'VIC_RACE')]

# Removing NA
shooting_data <- shooting_data %>% drop_na()
shooting_data <- shooting_data %>% filter(VIC_AGE_GROUP != '1022')
shooting_data <- shooting_data %>% filter(VIC_AGE_GROUP != 'UNKNOWN')
shooting_data <- shooting_data %>% filter(VIC_SEX != 'U')

# Summary
summary(shooting_data)
```

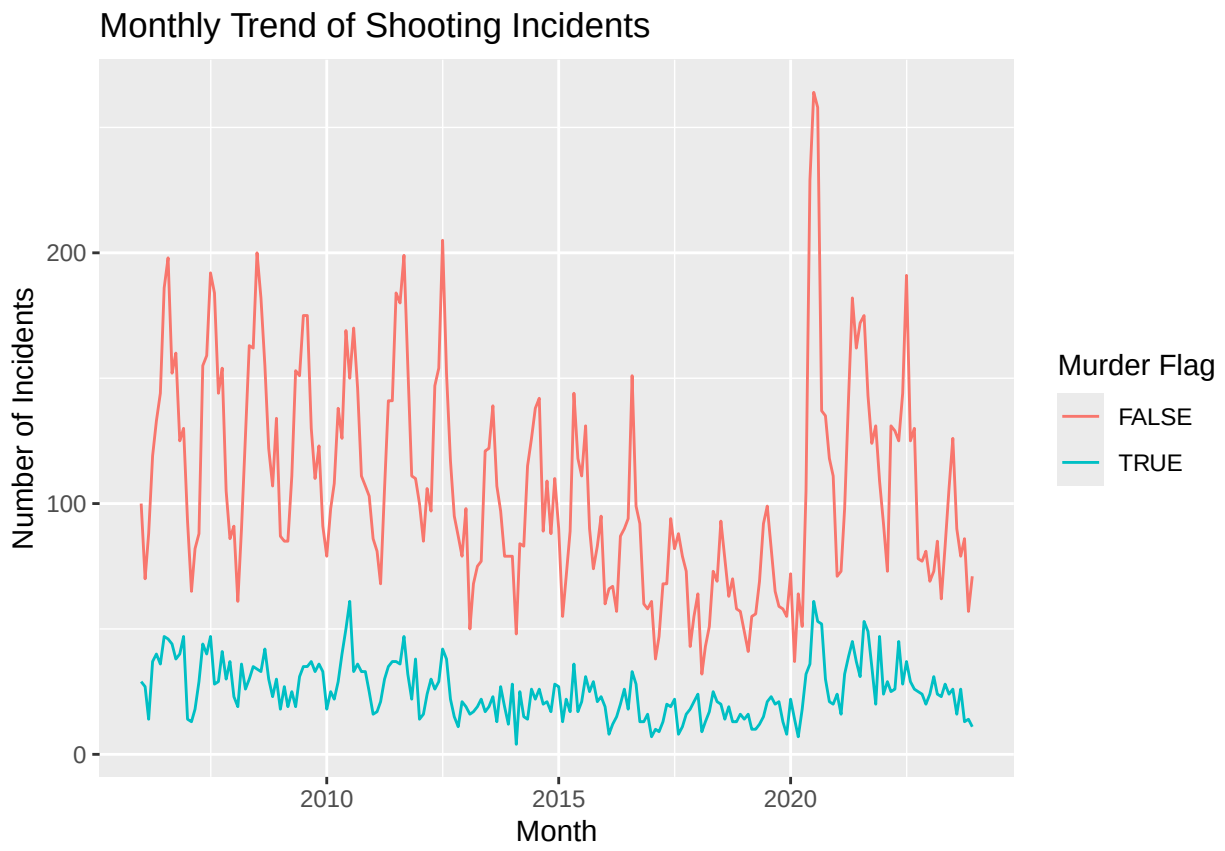
```
##   Month_Year        STATISTICAL_MURDER_FLAG     BORO              PRECINCT
##  Length:28491       Mode :logical           Length:28491       Min.   :  1.00
##  Class :character   FALSE:22981             Class :character   1st Qu.: 44.00
```

```
##  Mode  :character   TRUE :5510              Mode  :character   Median : 67.00
##                                                                Mean   : 65.49
##                                                                3rd Qu.: 81.00
##                                                                Max.   :123.00
##  VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:28491       Length:28491       Length:28491
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

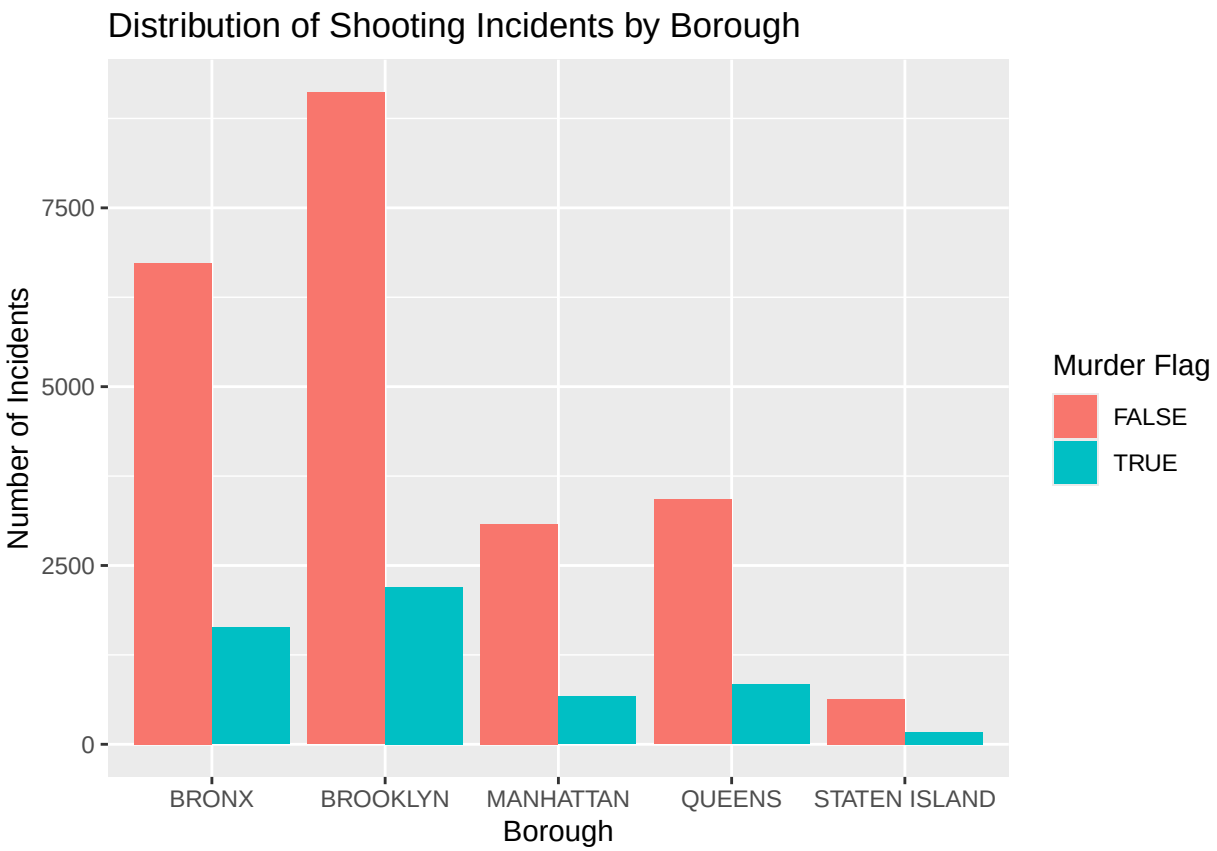
3: Visualization and Analysis

Lets look at the distribution of shootings by three fields: Month, Borough and Precint:

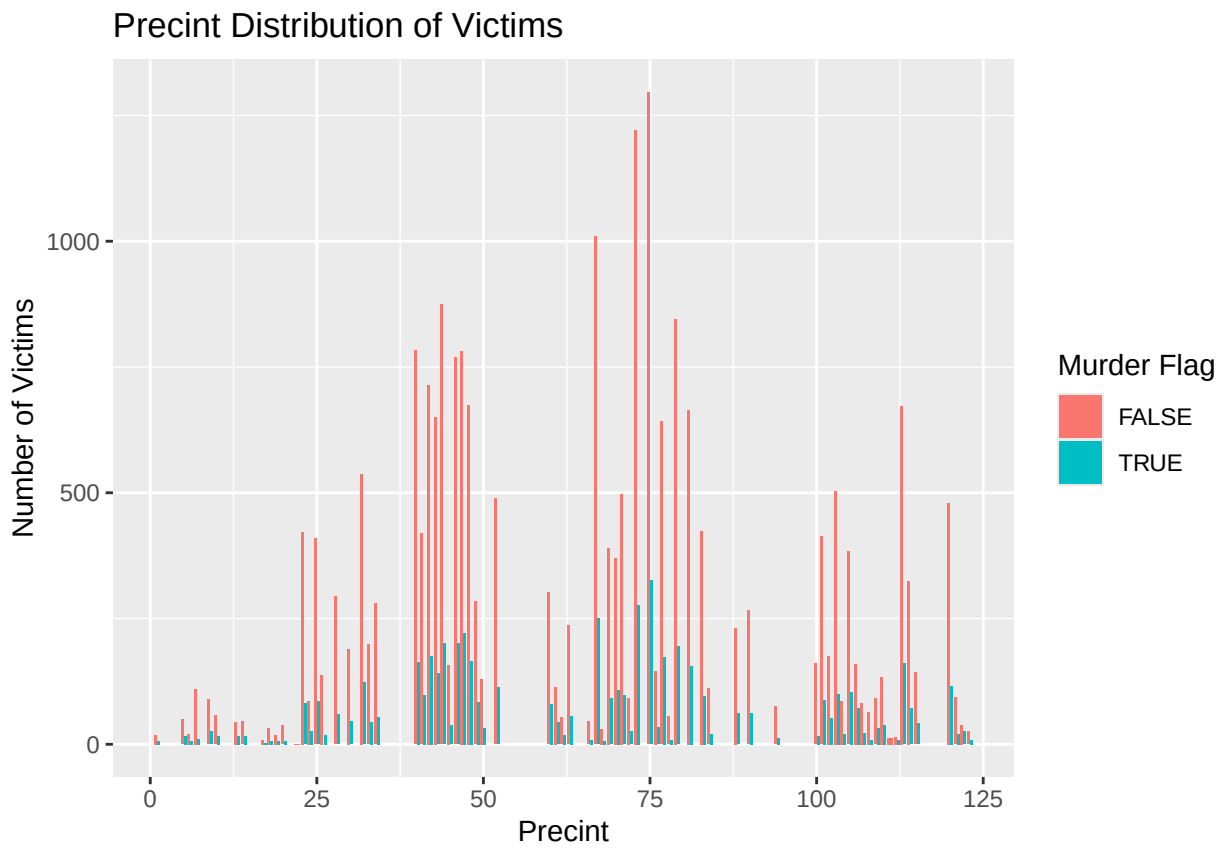
```
monthly_trend_count <- shooting_data %>%
  group_by(Month_Year, STATISTICAL_MURDER_FLAG) %>%
  summarise(Count = n(), .groups = 'drop')

ggplot(monthly_trend_count, aes(x = as.Date(paste0(Month_Year, "-01")),
                                y = Count, color = STATISTICAL_MURDER_FLAG)) +
  geom_line() +
  labs(title = "Monthly Trend of Shooting Incidents",
       x = "Month",
       y = "Number of Incidents",
       color = "Murder Flag")
```




```
ggplot(shooting_data, aes(x = BORO, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of Shooting Incidents by Borough",
       x = "Borough",
       y = "Number of Incidents",
       fill = "Murder Flag")
```

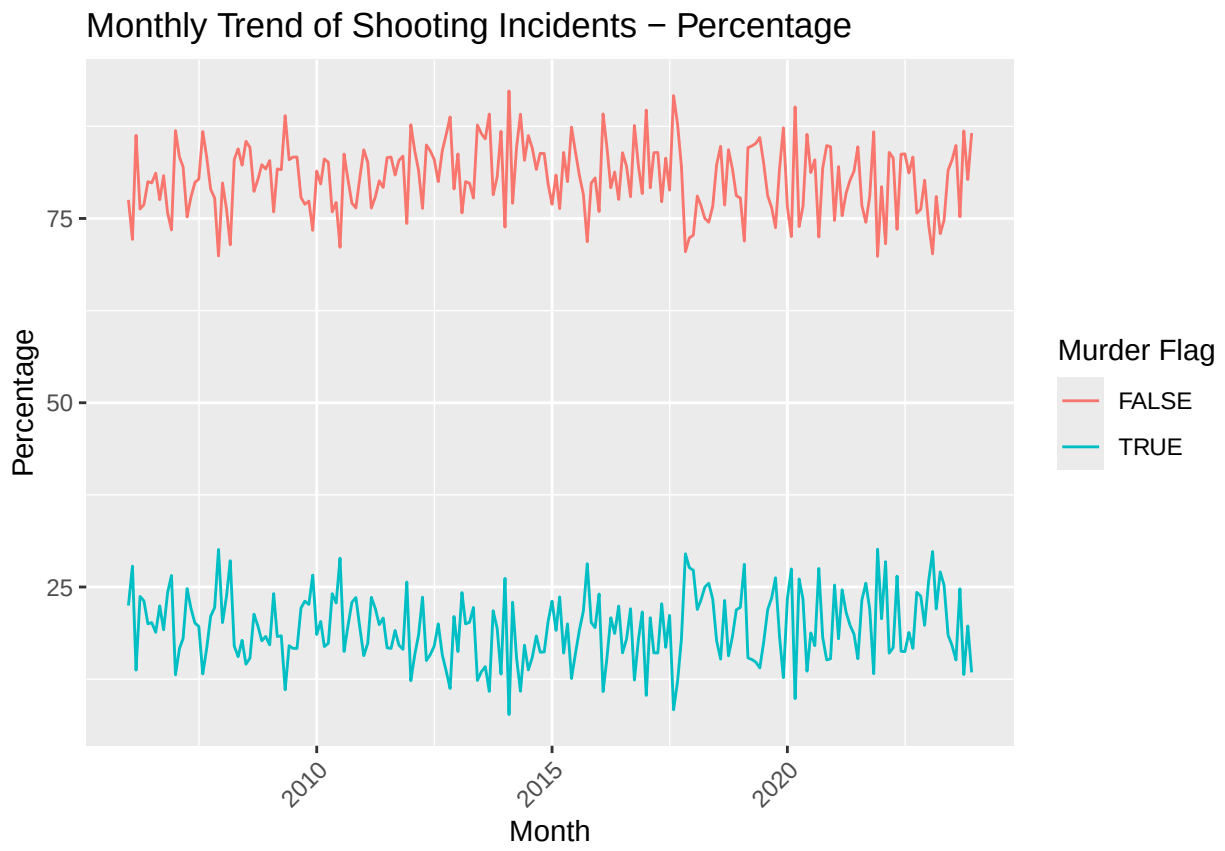



```
ggplot(shooting_data, aes(x = PRECINCT, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(position = "dodge") +
  labs(title = "Precint Distribution of Victims",
       x = "Precint",
       y = "Number of Victims",
       fill = "Murder Flag")
```

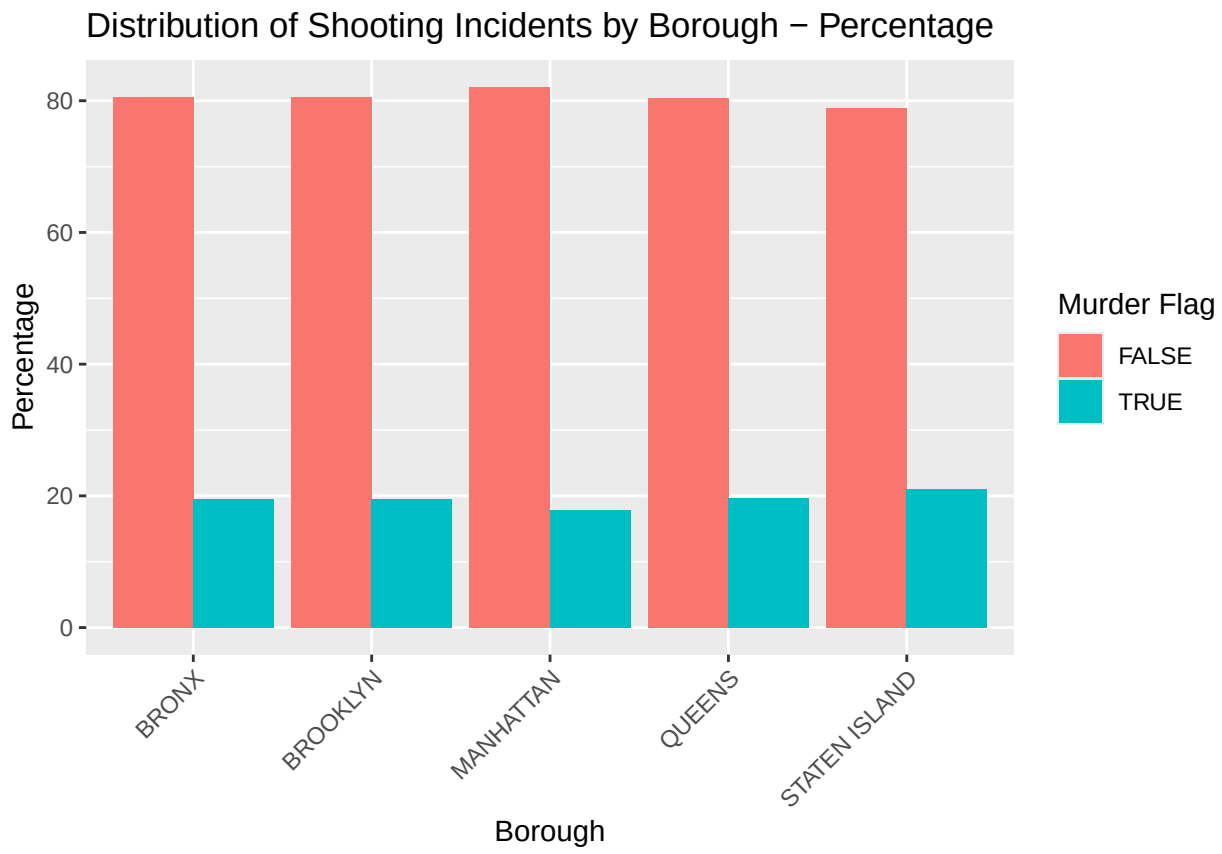


We can see that there are certain time, districts and precincts that have a lot higher numbers of shootings than others. Lets see if there is a difference in how deadly shootings are by these fields:

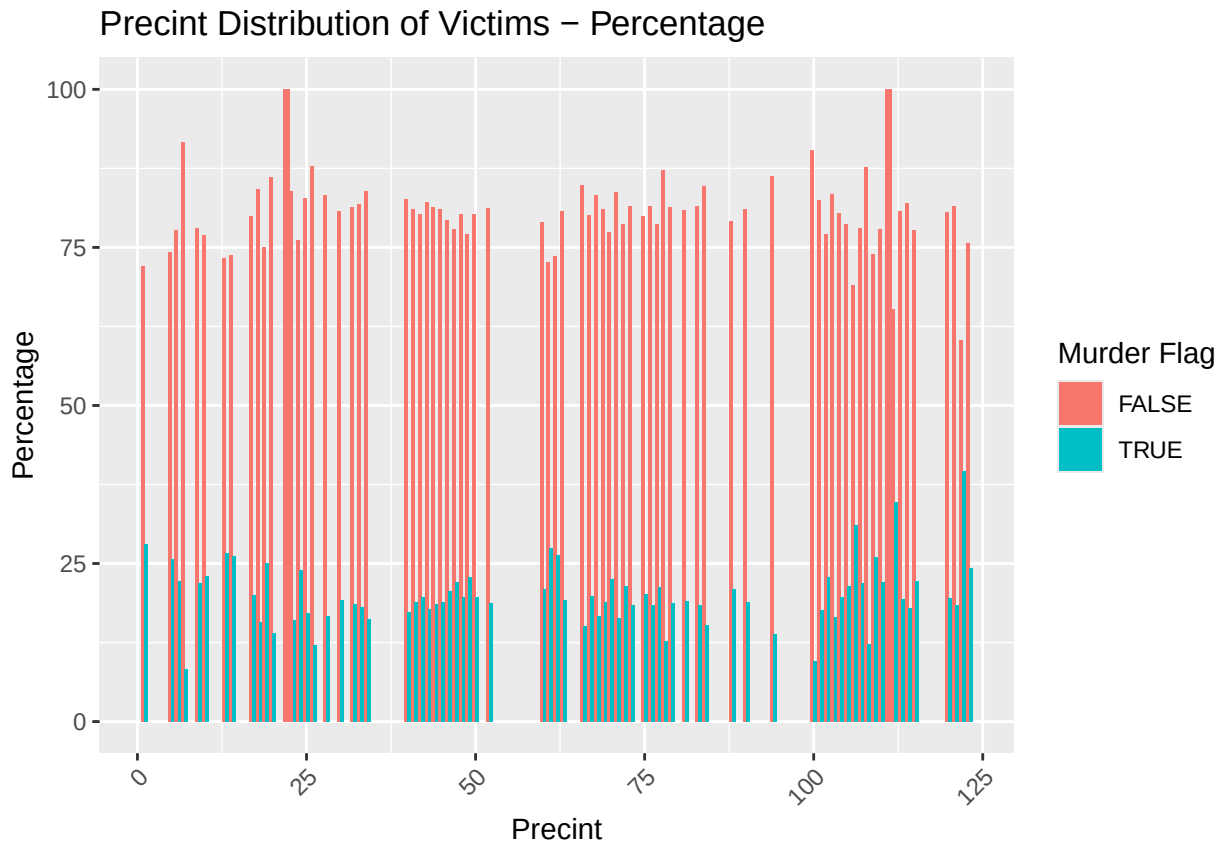
```
monthly_trend <- shooting_data %>%
  group_by(Month_Year, STATISTICAL_MURDER_FLAG) %>%
  summarise(Count = n(), .groups = 'drop') %>%
  group_by(Month_Year) %>%
  mutate(Percentage = (Count / sum(Count)) * 100)

ggplot(monthly_trend, aes(x = as.Date(paste0(Month_Year, "-01")),
                          y = Percentage, color = STATISTICAL_MURDER_FLAG)) +
  geom_line() +
  labs(title = "Monthly Trend of Shooting Incidents - Percentage",
       x = "Month",
       y = "Percentage",
       color = "Murder Flag") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
borough_distribution <- shooting_data %>%
  group_by(BORO, STATISTICAL_MURDER_FLAG) %>%
  summarise(Count = n(), .groups = 'drop') %>%
  group_by(BORO) %>%
  mutate(Percentage = (Count / sum(Count)) * 100)

ggplot(borough_distribution, aes(x = BORO, y = Percentage, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Shooting Incidents by Borough - Percentage",
       x = "Borough",
       y = "Percentage",
       fill = "Murder Flag") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



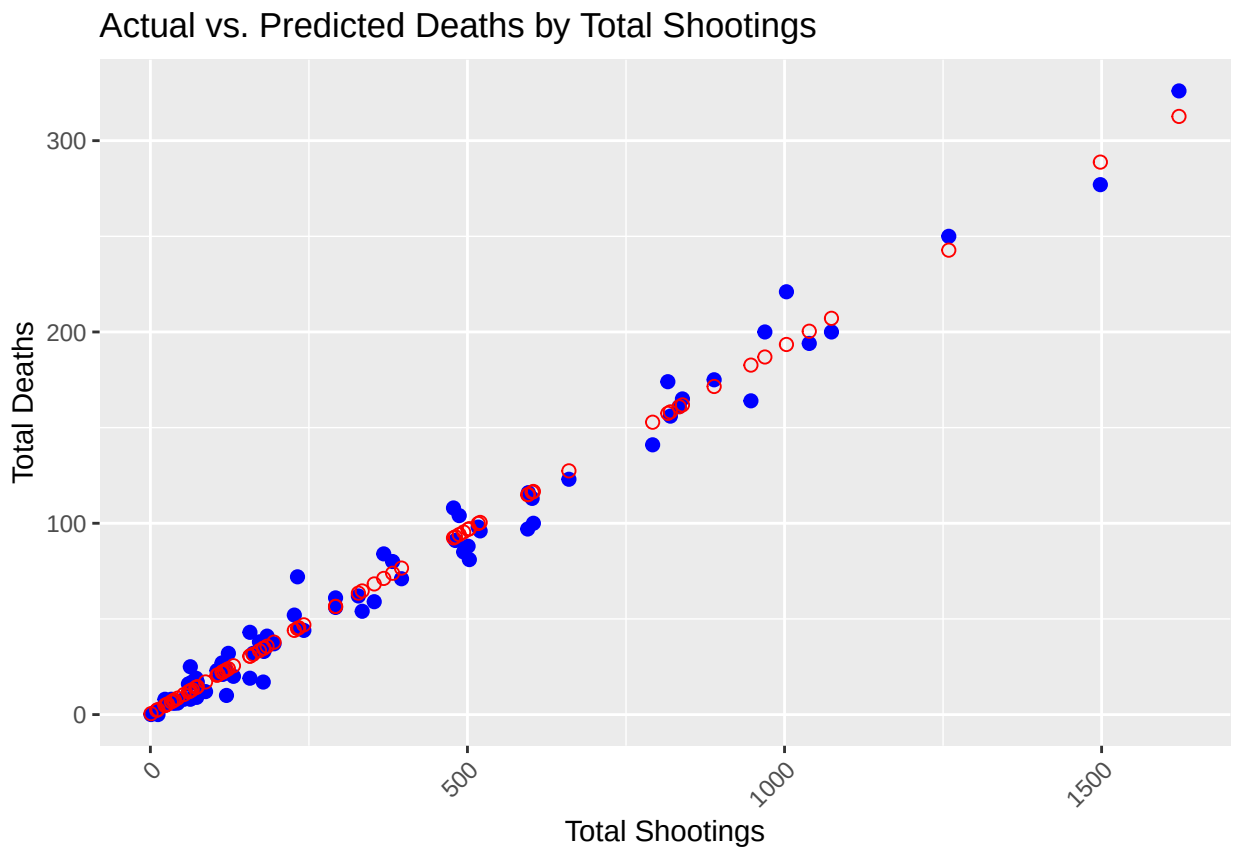

```
PRECINCT_distribution <- shooting_data %>%
  group_by(PRECINCT, STATISTICAL_MURDER_FLAG) %>%
  summarise(Count = n(), .groups = 'drop') %>%
  group_by(PRECINCT) %>%
  mutate(Percentage = (Count / sum(Count)) * 100)

ggplot(PRECINCT_distribution, aes(x = PRECINCT, y = Percentage, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Precint Distribution of Victims - Percentage",
       x = "Precint",
       y = "Percentage",
       fill = "Murder Flag") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Let do a basic model to see the relationship between total shootings and number of deaths:

```
PRECINCT_summary <- shooting_data %>%
  group_by(PRECINCT) %>%
  summarise(
    Total_Shootings = n(),
    Total_Deaths = sum(STATISTICAL_MURDER_FLAG == TRUE , na.rm = TRUE)  # Adjust based on actual flag v
  ) %>%
  filter(Total_Shootings > 0)
```

```
# linear model
model <- lm(Total_Deaths ~ Total_Shootings, data = PRECINCT_summary)


PRECINCT_summary <- PRECINCT_summary %>%
  mutate(Predicted_Deaths = predict(model))

# Scatter plot
ggplot(PRECINCT_summary) +
  geom_point(aes(x = Total_Shootings, y = Total_Deaths), color = "blue", size = 2) +
  geom_point(aes(x = Total_Shootings, y = Predicted_Deaths), color = "red", size = 2, shape = 1) +
  labs(title = "Actual vs. Predicted Deaths by Total Shootings",
       x = "Total Shootings",
       y = "Total Deaths",
       color = "Legend") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



4 Conclusion

In conclusion, our analysis reveals a highly significant correlation between Total Shootings and Total Deaths, indicating a strong linear relationship with low variance. This suggests that Total Shootings is a strong predictor of Total Deaths. However, to enhance the predictive accuracy of the model, incorporating additional factors could be beneficial.

It is important to acknowledge potential limitations in the data. Difference in the data collection methods across different precincts could introduce bias into the resutls. Biased reporting or underreporting in certain ares/ demographics could influence the results. The removal of any rows that have NA may also introduce bias into the results.