Analysis of NYPD shooting incident data

Detailed analysis on day of week and time of occurrence of shooting incidents

Rajesh Rao

Contents

Objective of analysis and description of dataset . . . 2

- Description of dataset . . . 2
- Objective . . . 2

Pre-processing and preparation . . . 2

- Setting up of R options . . . 2

Data Import . . . 3

Data Exploration and Data Tidying . . . 3

- Exploratory Analysis . . . 3
- Data Tidying . . . 4
    - Removal of attributes not related to our analysis . . . 4

Data Analysis . . . 5

- Analysis of day of shooting incidents (Weekend vs Weekdays) . . . 5
- Historical trend of shooting incidents per year . . . 5
- Shooting incidents by time of the day . . . 5

Data Visualization . . . 6

- Setting the theme for visualization . . . 6
- Visualization of historical trend in shooting incidents by year and victim gender . . . 6
- Visualization of day of occurence shooting incidents . . . 7
- Shooting Incidents by Hour of the day . . . 8

Data Model: Analyzing the day of occurence of shooting events . . . 9

- Creating weekend variable . . . 9
- Creating a linear model and statistical significance . . . 9
- Creating a new dataset to compare linear model vs actual shooting incident . . . 10

Data bias and Limitations of the Study . . . 11

Conclusion . . . 12

Objective of analysis and description of dataset

Description of dataset

1. List of every shooting incident that occurred in NYC going back to 2006 through the end of the 2021.
2. Prepared and published by New York City
3. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence.
4. Around 26 thousand records and 19 columns.

Objective

1. Analyze yearly trend in shooting incidents in New York city
2. Analyze what time of the day shooting incidents occur the most
3. Analyze what day of the week shooting incidents occur the most

Pre-processing and preparation

Setting up of R options

```
knitr::opts_chunk$set(echo = TRUE)
```

Initializing Session Information and Loading R packages

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.31   lifecycle_1.0.3 magrittr_2.0.3  evaluate_0.19
##  [5] rlang_1.0.6     stringi_1.7.8   cli_3.5.0       rstudioapi_0.14
##  [9] vctrs_0.5.1     rmarkdown_2.19  tools_4.2.2     stringr_1.5.0
## [13] glue_1.6.2      xfun_0.36       yaml_2.3.6      fastmap_1.1.0
## [17] compiler_4.2.2  htmltools_0.5.4 knitr_1.41
```

```
library(dplyr)
library(ggplot2)
library(forcats)
library(tidyverse)
library(lubridate)
```

Data Import

Importing historical data from city of New York Website

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

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Data Exploration and Data Tidying

Exploratory Analysis

Gather summary statistics of the data imported

```
summary(nypd_shooting)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA’s   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Investigating sample data

```
head(nypd_shooting, n=5)
```

```
## # A tibble: 5 x 19
##   INCIDE~1 OCCUR~2 OCCUR~3 BORO  PRECI~4 JURIS~5 LOCAT~6 STATI~7 PERP_~8 PERP_~9
##      <dbl> <chr>   <time>  <chr>   <dbl>   <dbl> <chr>   <lgl>   <chr>   <chr>
## 1   2.36e8 11/11/~ 15:04   BROO~      79       0 <NA>    FALSE   <NA>    <NA>
## 2   2.31e8 07/16/~ 22:05   BROO~      72       0 <NA>    FALSE   45-64   M
## 3   2.31e8 07/11/~ 01:09   BROO~      79       0 <NA>    FALSE   <18     M
## 4   2.38e8 12/11/~ 13:42   BROO~      81       0 <NA>    FALSE   <NA>    <NA>
## 5   2.24e8 02/16/~ 20:00   QUEE~     113       0 <NA>    FALSE   <NA>    <NA>
## # ... with 9 more variables: PERP_RACE <chr>, VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>,
## #   Latitude <dbl>, Longitude <dbl>, Lon_Lat <chr>, and abbreviated variable
## #   names 1: INCIDENT_KEY, 2: OCCUR_DATE, 3: OCCUR_TIME, 4: PRECINCT,
## #   5: JURISDICTION_CODE, 6: LOCATION_DESC, 7: STATISTICAL_MURDER_FLAG,
## #   8: PERP_AGE_GROUP, 9: PERP_SEX
```

Data Tidying

Removal of attributes not related to our analysis For our analysis, we are interested in date and time occurrence of shooting incidents.

1. We will be removing additional fields like Latitude,Longitude, Jurisdiction etc.

```
nypd_shooting_tidy <- nypd_shooting %>% select(-c(X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat))
```

2. Additionally, we found perpetrator's demographics data is largely missing. Since, this is not relevant to our analysis. We will be removing these attributes instead of imputing the data points

```
nypd_shooting_tidy <- nypd_shooting_tidy %>% select(-c(PERP_AGE_GROUP,PERP_SEX,PERP_RACE)) %>%
  select(-c(LOCATION_DESC,JURISDICTION_CODE,LOCATION_DESC))
```

3. Occurrence Date is character format. We will be converting to date.

```
nypd_shooting_tidy$OCCUR_DATE <- as.Date(nypd_shooting_tidy$OCCUR_DATE, format = "%m/%d/%Y")
```

Data Analysis

Analysis of day of shooting incidents (Weekend vs Weekdays)

1. We will be analyzing the relationship between day of shooting and number of shooting incidents. For this, we will be creating a new variable which indicates day of actual shooting incident and it will be derived from OCCUR_DATE. Start of the week is Monday.

```
nypd_shooting_dayofweek <- nypd_shooting_tidy %>%
  mutate(day_of_week = wday(OCCUR_DATE, week_start = 1))
```

2. A new table will be created aggregate shooting incidents on day of occurrence

```
nypd_shooting_dayofweek_agg <- nypd_shooting_dayofweek %>%
  group_by(day_of_week) %>% summarize(count = n())
```

Historical trend of shooting incidents per year

1. We will be analyzing if there are any historical trends in number of shooting incidents based on year. For this, we will be deriving year of occurence of shooting incident form OCCUR_DT

```
nypd_shooting_year <- nypd_shooting_tidy %>%
  mutate(year = year(OCCUR_DATE))
```

2. A new table will be created aggregate shooting incidents per year and gender

```
nypd_shooting_year_agg <- nypd_shooting_year %>%
  group_by(victim_gender = VIC_SEX, year = year(OCCUR_DATE)) %>% summarize(count = n())
```

```
## ‘summarise()‘ has grouped output by ’victim_gender’. You can override using the
## ‘.groups‘ argument.
```

Shooting incidents by time of the day

1. We will be analyzing what time of the day the shooting incidents occur the most. For this, we will be creating a new attribute hour of the incident which is derived form OCCUR_TIME

```
nypd_shooting_hour <- nypd_shooting_tidy %>%
  mutate(hour = format(as.POSIXct(OCCUR_TIME,format="%H:%M:%S"),"%H"))
```

2. A new table will be created aggregate shooting incidents per hour of occurrence

```
nypd_shooting_hour_agg <- nypd_shooting_hour %>%
  group_by(hour) %>% summarize(count = n())
```

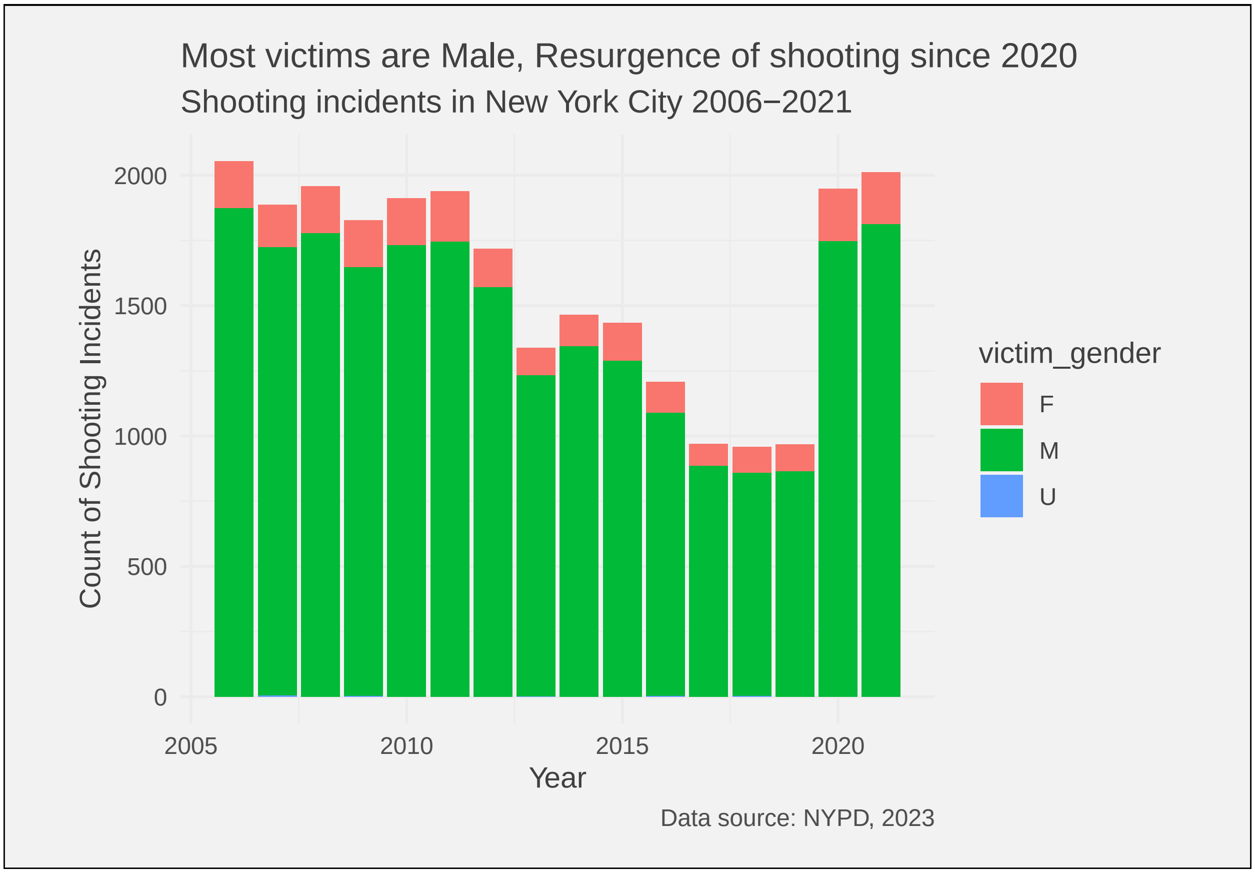
Data Visualization

Setting the theme for visualization

```
theme_shooting <- function() {
    theme_minimal() +
theme(
    text = element_text(color = "gray25"),
    plot.subtitle = element_text(size = 12),
    plot.caption = element_text(color = "gray30"),
    plot.background = element_rect(fill = "gray95"),
    plot.margin = unit(c(5, 10, 5, 10), units = "mm")
  )
}
```

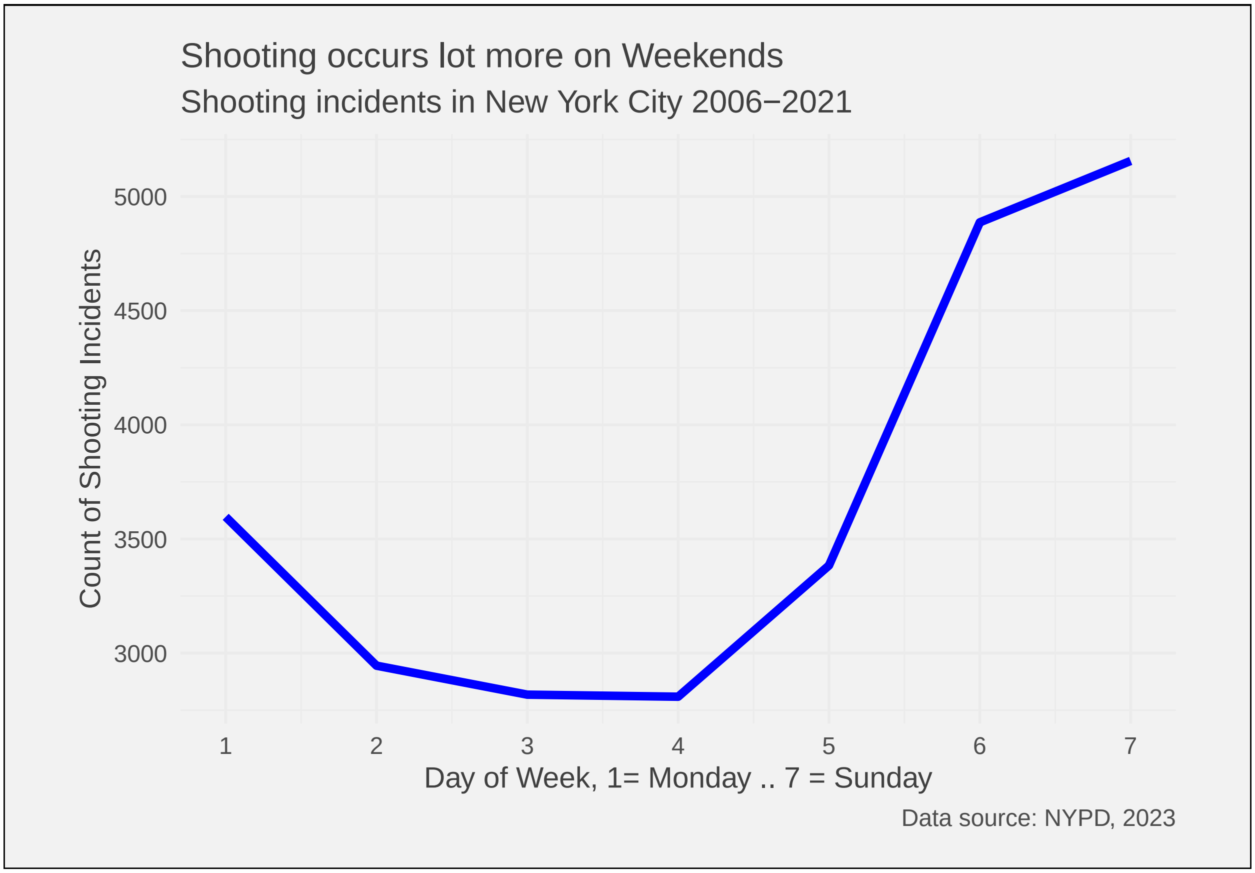
Visualization of historical trend in shooting incidents by year and victim gender

```
ggplot(data = nypd_shooting_year_agg,
       aes(x = year, y = count, fill=victim_gender)) + geom_bar(stat = "identity") + labs(
    x = "Year",
    y = "Count of Shooting Incidents",
    title = "Most victims are Male, Resurgence of shooting since 2020",
    subtitle = "Shooting incidents in New York City 2006-2021",
    caption = "Data source: NYPD, 2023"
  ) + theme_shooting()
```



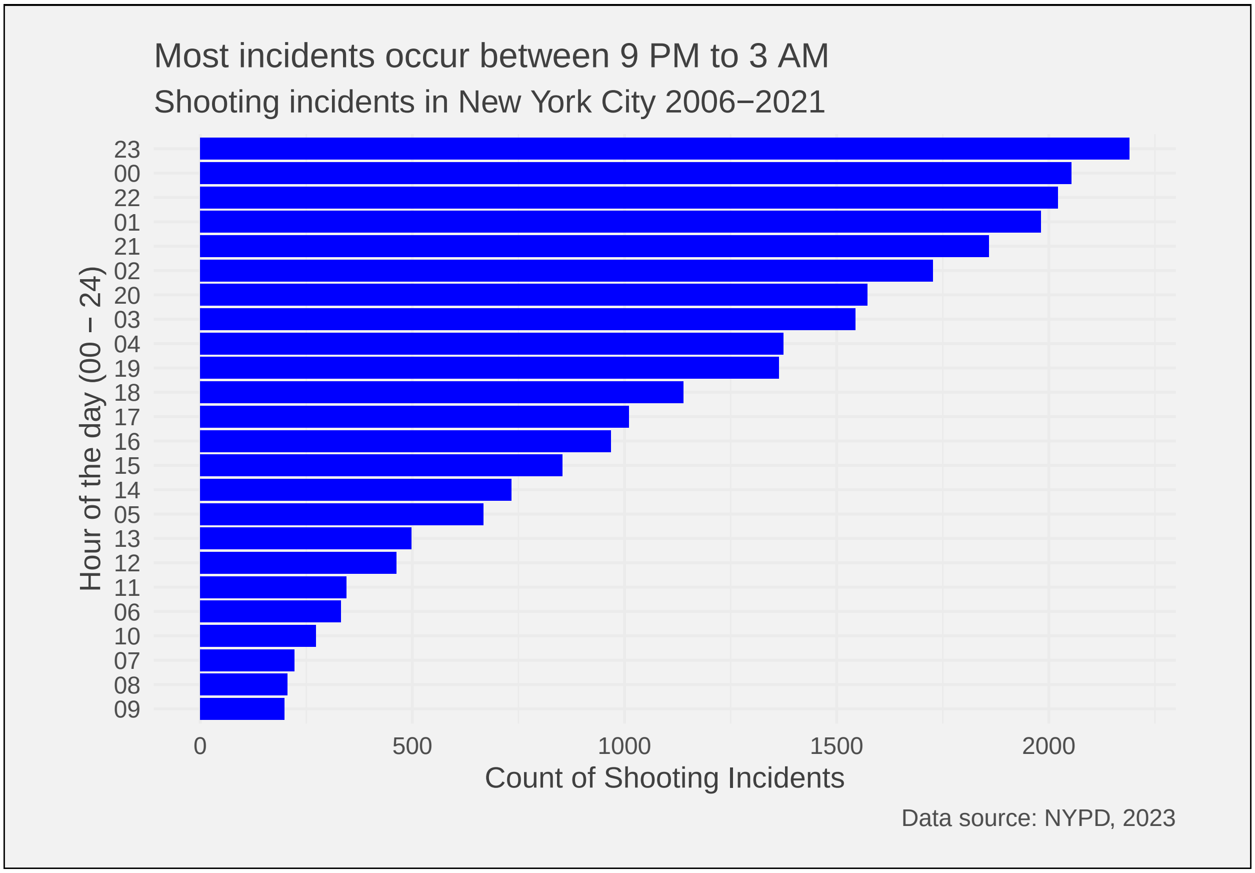
Visualization of day of occurence shooting incidents

```
ggplot(data = nypd_shooting_dayofweek_agg, aes(x = day_of_week, y = count)) + geom_line(color = "blue",
    x = "Day of Week, 1= Monday .. 7 = Sunday",
    y = "Count of Shooting Incidents",
    title = "Shooting occurs lot more on Weekends",
    subtitle = "Shooting incidents in New York City 2006-2021",
    caption = "Data source: NYPD, 2023"
  ) + theme_shooting()
```



Shooting Incidents by Hour of the day

```
ggplot(data = nypd_shooting_hour_agg,
       aes(x = reorder(hour, count), y = count)) + geom_bar(stat = "identity", fill = "blue") + labs(
    x = "Hour of the day (00 - 24)",
    y = "Count of Shooting Incidents",
    title = "Most incidents occur between 9 PM to 3 AM",
    subtitle = "Shooting incidents in New York City 2006-2021",
    caption = "Data source: NYPD, 2023"
  ) + theme_shooting() + coord_flip()
```


Most incidents occur between 9 PM to 3 AM

Shooting incidents in New York City 2006−2021

Data source: NYPD, 2023

Data Model: Analyzing the day of occurence of shooting events

Creating weekend variable

1. We will be trying to develop a model which predicts whether day of week is weekend or not based on the count of shooting incidents occurring on that day. In this way, we can predict if there is a relationship between number of shooting incidents and day of week. A new variable will be created to determine if given day is weekend.

```
nypd_shooting_dow_agg <- mutate(nypd_shooting_dayofweek_agg,
                                is_weekendd = ifelse (nypd_shooting_dayofweek_agg$day_of_week < 6,0,1))
```

Creating a linear model and statistical significance

2. A linear model will be created to predict relationship between is_weekend and count of shooting incidents

```
nypd_shooting_model <- lm(is_weekendd ~ count, data=nypd_shooting_dow_agg)
```

3. Showing the linear model

```
nypd_shooting_model
```

```
##
## Call:
## lm(formula = is_weekendd ~ count, data = nypd_shooting_dow_agg)
##
## Coefficients:
## (Intercept)        count
##  -1.4448546    0.0004733
```

4. Summarizing the performance of linear model

```
summary(nypd_shooting_model)
```

```
##
## Call:
## lm(formula = is_weekendd ~ count, data = nypd_shooting_dow_agg)
##
## Residuals:
##         1         2         3         4         5         6         7
## -0.257521  0.051056  0.111162  0.115421 -0.156713  0.131953  0.004642
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.445e+00  2.593e-01  -5.572 0.002563 **
## count        4.733e-04  6.882e-05   6.877 0.000995 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1653 on 5 degrees of freedom
## Multiple R-squared:  0.9044, Adjusted R-squared:  0.8853
## F-statistic: 47.29 on 1 and 5 DF,  p-value: 0.0009946
```

5. Discussion on the model

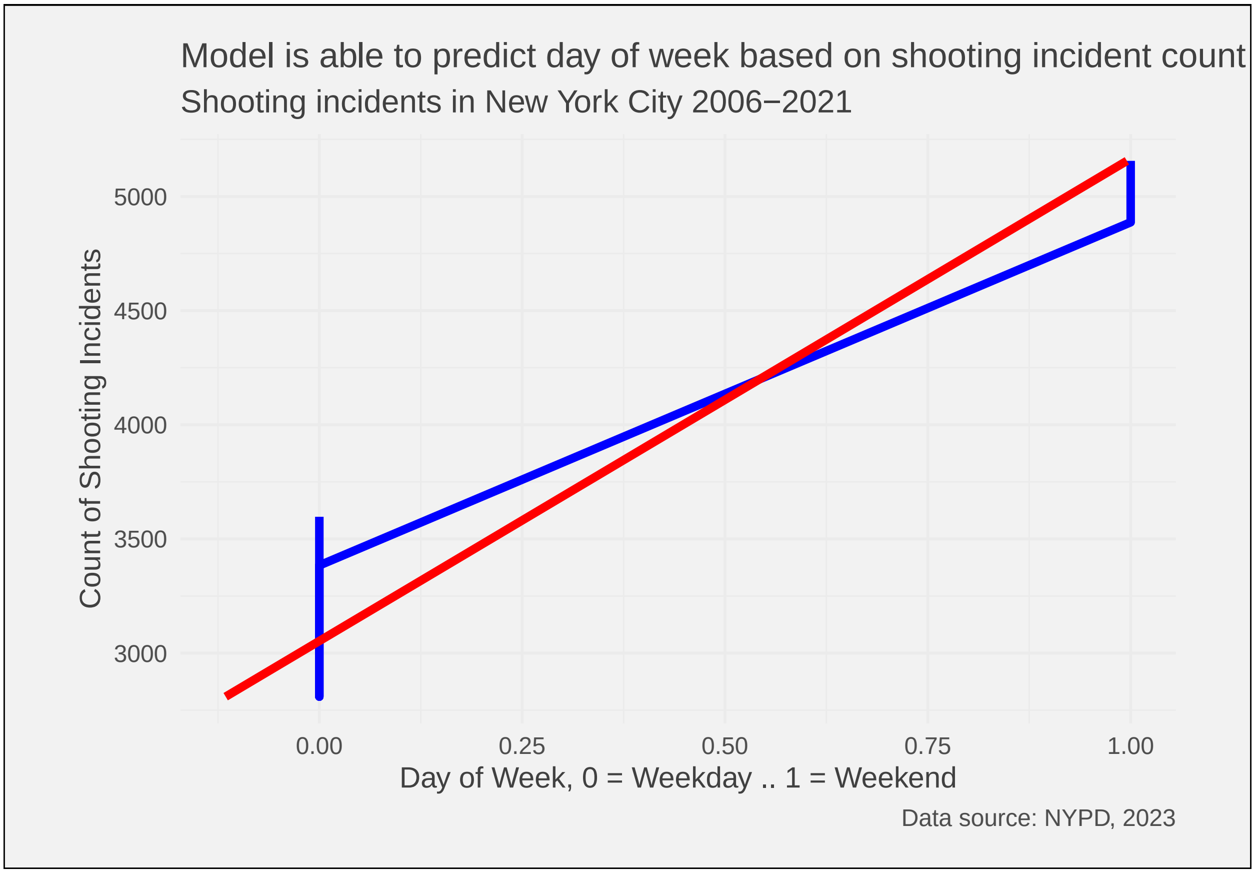
The p-value of the model is significantly less than .05 indicating that model should be able to accurately predict whether a given day is weekend or not based on the count of shooting incident. Based on above, there is a significant relationship between number of shooting incidents and day of week.

Creating a new dataset to compare linear model vs actual shooting incident

```
nypd_shooting_dow_agg <- nypd_shooting_dow_agg %>% mutate(pred = predict(nypd_shooting_model))
```

6. Model is able to closely follow the actual count occurrence on wekeday and weekend and it is statistically significant to predict a relationship between day of shooting and number of shooting incidents. ### Plotting the model performance visually

```
ggplot(data = nypd_shooting_dow_agg) +
  geom_line(color = "blue", size =1.5, aes(x = nypd_shooting_dow_agg$is_weekendd,
                                          y=nypd_shooting_dow_agg$count)) +
  geom_line(color = "red", size =1.5, aes(x = nypd_shooting_dow_agg$pred,
                                         y=nypd_shooting_dow_agg$count)) +
  labs(
    x = "Day of Week, 0 = Weekday .. 1 = Weekend",
    y = "Count of Shooting Incidents",
    title = "Model is able to predict day of week based on shooting incident count",
    subtitle = "Shooting incidents in New York City 2006-2021",
    caption = "Data source: NYPD, 2023"
  ) + theme_shooting()
```



Data bias and Limitations of the Study

1. Selection bias: If the shooting incidents in the dataset are not representative of all shooting incidents that occurred during the time period , analysis may be biased. For example, if dataset only includes shooting incidents that were reported to the police, but there were many shooting incidents that were not reported, analysis may underestimate the true number of shooting incidents and bias your results.

2. Measurement bias: If the measurement of the date and time of shooting incidents is inaccurate or incomplete, analysis may be biased. For example, if the time of a shooting incident is only recorded to the nearest hour, but the actual time of the incident was closer to a different hour, analysis may misrepresent the frequency of incidents at different times of day.

3. Temporal Bias: The dataset covers a period of several years, during which time there may have been changes in the way that shootings were reported or recorded. For example, changes in police policies or practices, changes in the political or social climate, or changes in the technology used to collect and analyze data could all affect the accuracy and completeness of the data over time.

Conclusion

1. Most shootings occur on the weekend: Based on the analysis of the shooting incidents dataset, it was found that most of the shootings occur on the weekend. This conclusion is supported by the data, which shows that the highest number of shootings occurred on Saturday and Sunday, with a smaller number occurring on weekdays. One possible explanation for this pattern is that more people are out and about on the weekend, which could lead to an increased risk of conflict or violence.

2. Most victims are male: Another conclusion drawn from the analysis of the shooting incidents dataset is that the majority of victims are male. This finding is also supported by the data, which shows that over 80% of the shooting victims in the dataset were male. One possible explanation for this pattern is that males may be more likely to be involved in violent incidents or to be targeted for violence.

3. Emergence in increase in number of shooting incidents since 2020: The analysis of the shooting incidents dataset also revealed an increase in the number of shooting incidents since 2020. This conclusion is supported by the data, which shows that there was a sharp increase in the number of shootings in 2020, which continued into 2021. One possible explanation for this pattern is that the COVID-19 pandemic and its associated economic and social disruptions may have contributed to an increase in violence in some areas.

It's important to note that these conclusions are based on the available data and the methods used to analyze it. However, there may be limitations or biases in the data that could affect the accuracy or generalization of these findings. It's also important to consider other factors that may be contributing to the patterns observed in the data, and to use caution when drawing causal inferences or making policy recommendations based on the analysis.