NYPD Shooting

Ayush Jain

2025-06-18

NYPD Shooting Incident Data Report

This report analyzes NYPD Shooting Incident data from 2006 through 2024. The dataset is manually extracted and reviewed quarterly by the Office of Management Analysis and Planning, then published by the NYPD. Each record represents a shooting incident in New York City, including event, location, and demographic information about suspects and victims. The dataset offers valuable public insight into the patterns and risk factors associated with gun violence in New York City.

Data Source - https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Library Imports

```r
library(readr)
library(tidyverse)
library(party)
library(caret)
library(e1071)
```

Importing the data

The data is read from a CSV file which is present in the same folder.

```r
file_path = 'NYPD_Shooting_Incident_Data__Historic_.csv'
data <- read_csv(file_path)
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): 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.
```

Display the first 10 rows:

```r
head(data)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    231974218 08-09-2021 01:06      BRONX    <NA>                    40
## 2    177934247 04-07-2018 19:48      BROOKLYN <NA>                    79
## 3    255028563 12-02-2022 22:57      BRONX    OUTSIDE                 47
## 4     25384540 11/19/2006 01:50      BROOKLYN <NA>                    66
## 5     72616285 05-09-2010 01:58      BRONX    <NA>                    46
## 6     85875439 07/22/2012 21:35      BRONX    <NA>                    42
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, 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>
```

Data Cleaning and Transformation

The following steps remove unused columns, ensure proper data types, and create new variables such as crime year, weekday, and hour.

```r
data = data %>%
  select(-c(X_COORD_CD, Y_COORD_CD))

data$BORO = as.factor(data$BORO)
data$PERP_AGE_GROUP = as.factor(data$PERP_AGE_GROUP)
data$PERP_SEX = as.factor(data$PERP_SEX)
data$PERP_RACE = as.factor(data$PERP_RACE)
data$VIC_AGE_GROUP = as.factor(data$VIC_AGE_GROUP)
data$VIC_RACE = as.factor(data$VIC_RACE)
data$VIC_SEX = as.factor(data$VIC_SEX)
data$LOC_CLASSFCTN_DESC = as.factor(data$LOC_CLASSFCTN_DESC)

data = data %>%
  mutate(OCCUR_HOUR=as.integer(format(strptime(OCCUR_TIME,"%H:%M:%S"),'%H')))

data = data %>%
  mutate(OCCUR_DATE=as.Date(gsub('-', '/', OCCUR_DATE), format="%m/%d/%Y")) %>%
  mutate(OCCUR_YEAR=as.integer(format(OCCUR_DATE,"%Y")))

data$OCCUR_WEEKDAY = wday(data$OCCUR_DATE)
```

Summarize all columns:

```r
summary(data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :00:00:00.000000
##  1st Qu.: 67321140   1st Qu.:2009-10-29   1st Qu.:03:30:45.000000
```

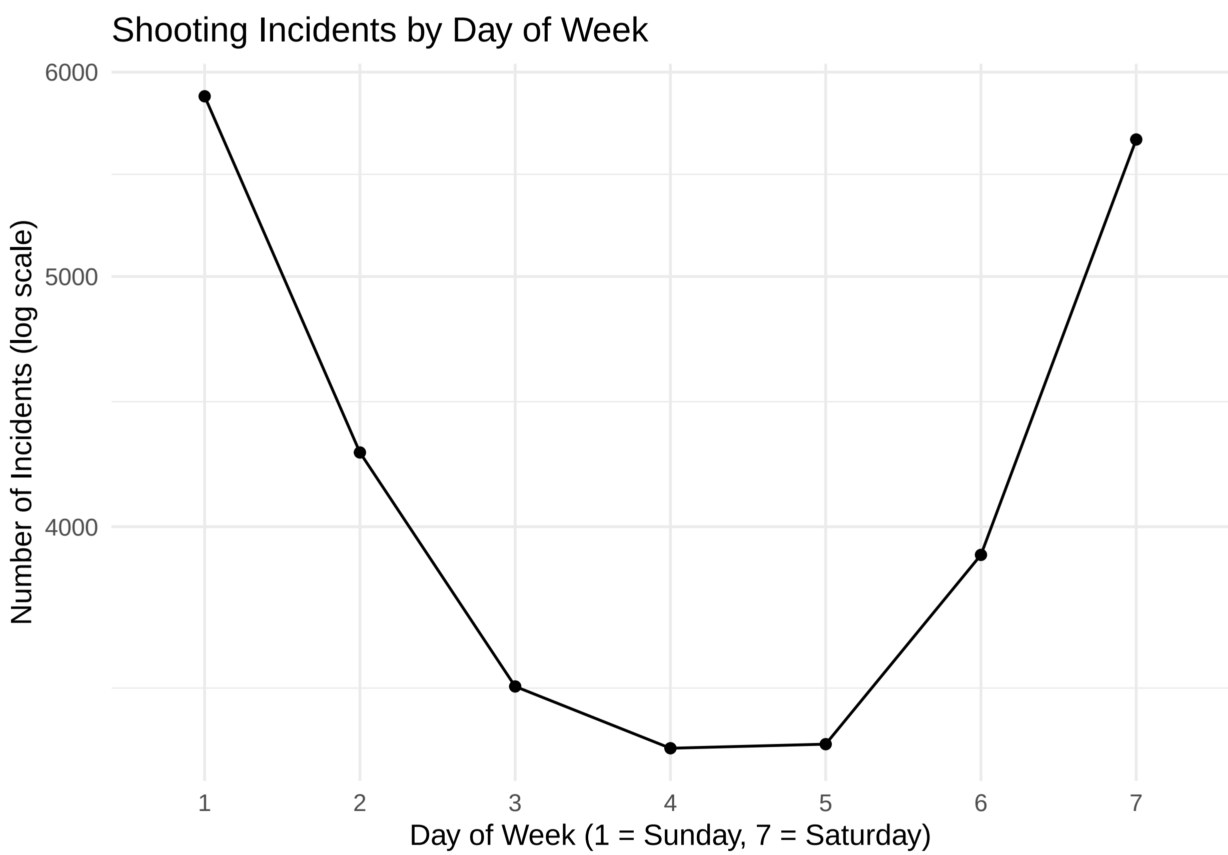
```
##  Median :109291972   Median :2014-03-25   Median :15:15:00.000000
##  Mean   :133850951   Mean   :2014-10-31   Mean   :12:46:10.874798
##  3rd Qu.:214741917   3rd Qu.:2020-06-29   3rd Qu.:20:44:00.000000
##  Max.   :299462478   Max.   :2024-12-31   Max.   :23:59:00.000000
##
##             BORO       LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  BRONX        : 8834   Length:29744       Min.   :  1.00   Min.   :0.0000
##  BROOKLYN     :11685   Class :character   1st Qu.: 44.00   1st Qu.:0.0000
##  MANHATTAN    : 3977   Mode  :character   Median : 67.00   Median :0.0000
##  QUEENS       : 4426                      Mean   : 65.23   Mean   :0.3181
##  STATEN ISLAND:  822                      3rd Qu.: 81.00   3rd Qu.:0.0000
##                                           Max.   :123.00   Max.   :2.0000
##                                                            NA's   :2
##   LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  STREET    : 2639    Length:29744       Mode :logical           18-24  :6630
##  HOUSING   :  643    Class :character   FALSE:23979             25-44  :6342
##  DWELLING  :  341    Mode  :character   TRUE :5765              UNKNOWN:3148
##  COMMERCIAL:  276                                               <18    :1805
##  OTHER     :   74                                               (null) :1628
##  (Other)   :  175                                               (Other): 847
##  NA's      :25596                                               NA's   :9344
##    PERP_SEX              PERP_RACE      VIC_AGE_GROUP   VIC_SEX
##  (null): 1628   BLACK         :12323   <18    : 3081   F: 2891
##  F     :  461   WHITE HISPANIC: 2667   1022   :    1   M:26841
##  M     :16845   UNKNOWN       : 1838   18-24  :10677   U:   12
##  U     : 1500   (null)        : 1628   25-44  :13563
##  NA's  : 9310   BLACK HISPANIC: 1487   45-64  : 2118
##                 (Other)       :  491   65+    :  236
##                 NA's          : 9310   UNKNOWN:   68
##                            VIC_RACE        Latitude       Longitude
##  AMERICAN INDIAN/ALASKAN NATIVE:   13   Min.   :40.51   Min.   :-74.25
##  ASIAN / PACIFIC ISLANDER      :  478   1st Qu.:40.67   1st Qu.:-73.94
##  BLACK                         :20999   Median :40.70   Median :-73.91
##  BLACK HISPANIC                : 2930   Mean   :40.74   Mean   :-73.91
##  UNKNOWN                       :   72   3rd Qu.:40.83   3rd Qu.:-73.88
##  WHITE                         :  741   Max.   :40.91   Max.   :-73.70
##  WHITE HISPANIC                : 4511   NA's   :97      NA's   :97
##    Lon_Lat           OCCUR_HOUR     OCCUR_YEAR   OCCUR_WEEKDAY
##  Length:29744       Min.   : 0.0   Min.   :2006   Min.   :1.000
##  Class :character   1st Qu.: 3.0   1st Qu.:2009   1st Qu.:2.000
##  Mode  :character   Median :15.0   Median :2014   Median :4.000
##                     Mean   :12.3   Mean   :2014   Mean   :3.947
##                     3rd Qu.:20.0   3rd Qu.:2020   3rd Qu.:6.000
##                     Max.   :23.0   Max.   :2024   Max.   :7.000
##
```

Visualization

Shooting Incidents by Day of Week

The plot below shows the number of shooting incidents according to the day of the week. Incidents tend to be more frequent on weekends. Accordingly, Police Department can increase the patrolling on these days.

```r
data %>%
  group_by(OCCUR_WEEKDAY) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = as.factor(OCCUR_WEEKDAY), y = Total_Incidents, group = 1)) +
  geom_line() +
  geom_point() +
  scale_y_log10() +
  labs(
    title = "Shooting Incidents by Day of Week",
    x = "Day of Week (1 = Sunday, 7 = Saturday)",
    y = "Number of Incidents (log scale)"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



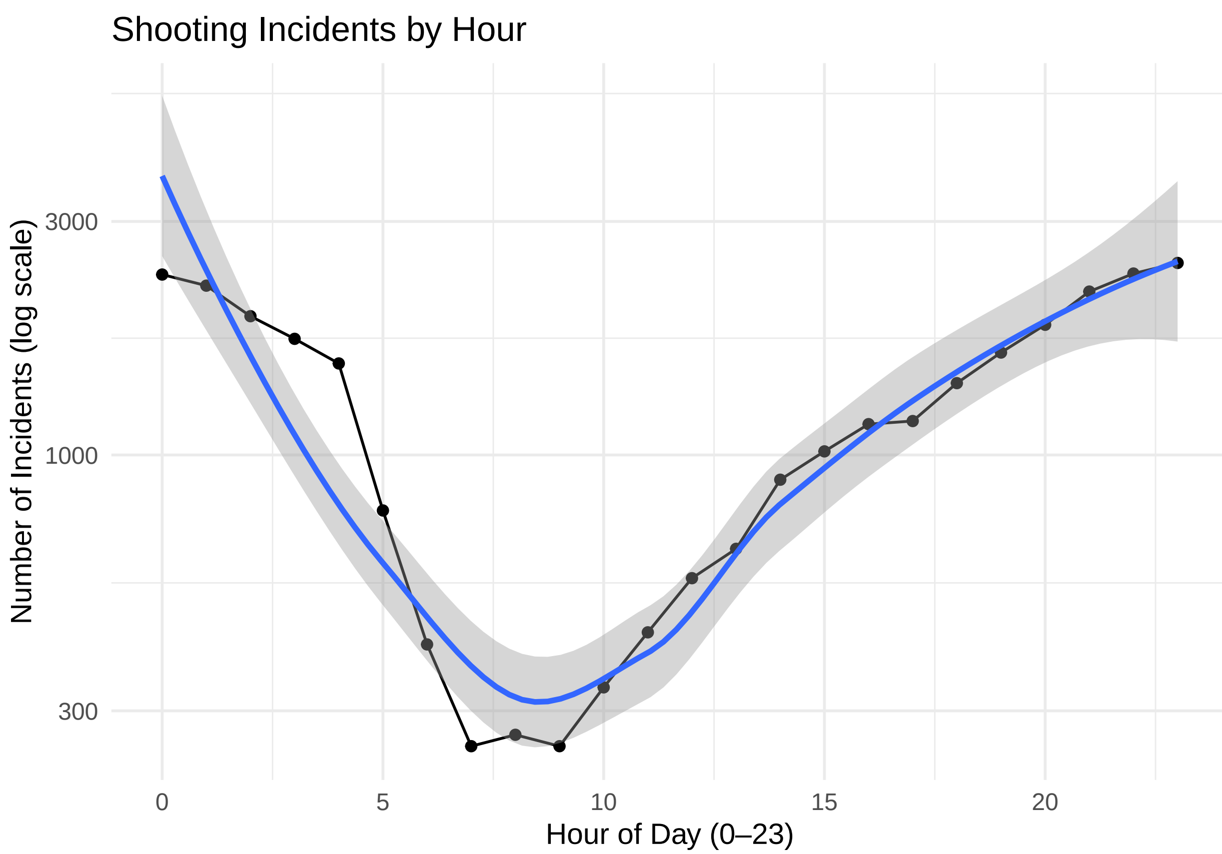
Shooting Incidents by Hour

The next plot shows incident frequency by hour of the day. Shootings are more common at night. Accordingly, Police Department can increase the patrolling on these time.

```r
data %>%
  group_by(OCCUR_HOUR) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = OCCUR_HOUR, y = Total_Incidents)) +
```

```r
  geom_line() +
  geom_point() +
  geom_smooth() +
  scale_y_log10() +
  labs(
    title = "Shooting Incidents by Hour",
    x = "Hour of Day (0-23)",
    y = "Number of Incidents (log scale)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



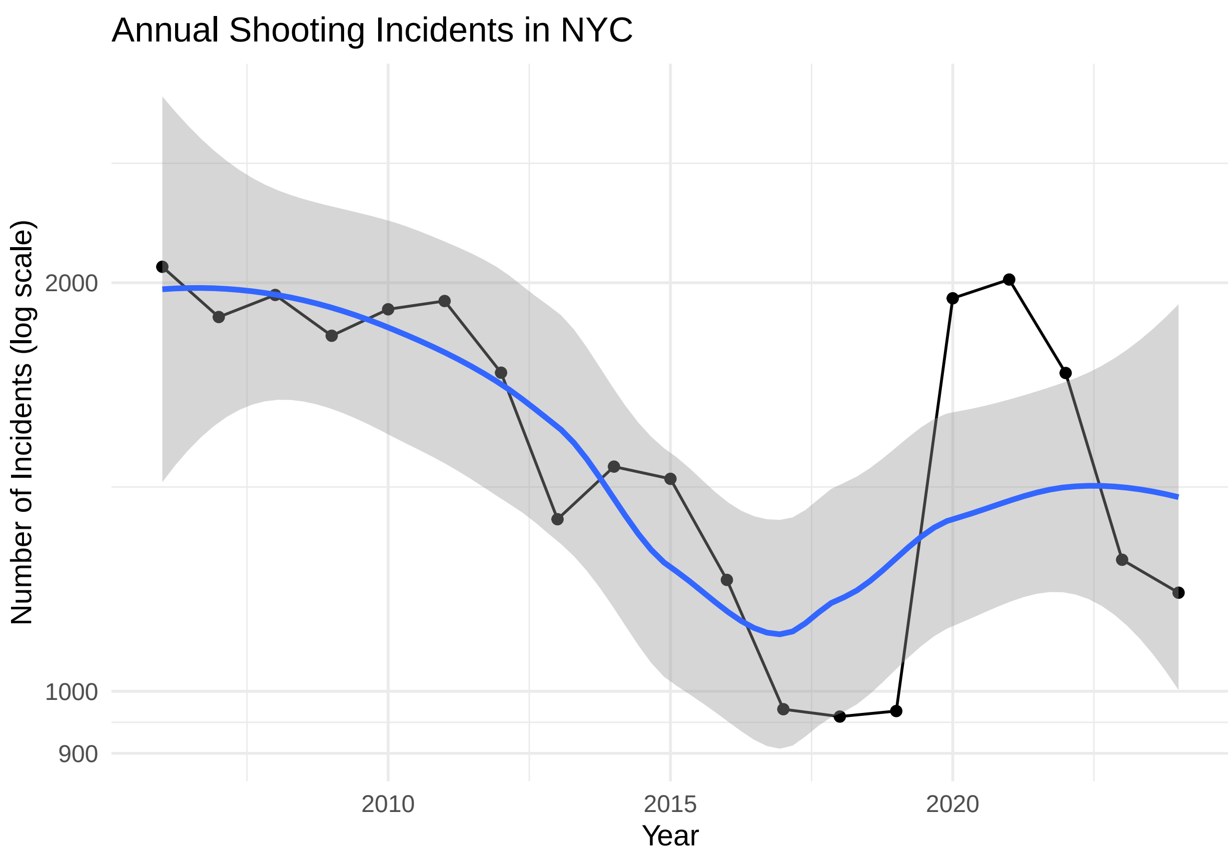
Shooting Incidents by Year

This time series shows the number of incidents each year, highlighting long-term trends.

```r
data %>%
  group_by(OCCUR_YEAR) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = OCCUR_YEAR, y = Total_Incidents)) +
  geom_line() +
  geom_point() +
  geom_smooth() +
```

```r
  scale_y_log10() +
  labs(
    title = "Annual Shooting Incidents in NYC",
    x = "Year",
    y = "Number of Incidents (log scale)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Victim Age Group and Sex

This bar plot shows the distribution of shooting victims by age group and sex. Males in the 25–44 age group are the most frequent victims. It is understandable as this age group usually spend most of their time outside at late hours and weekends.

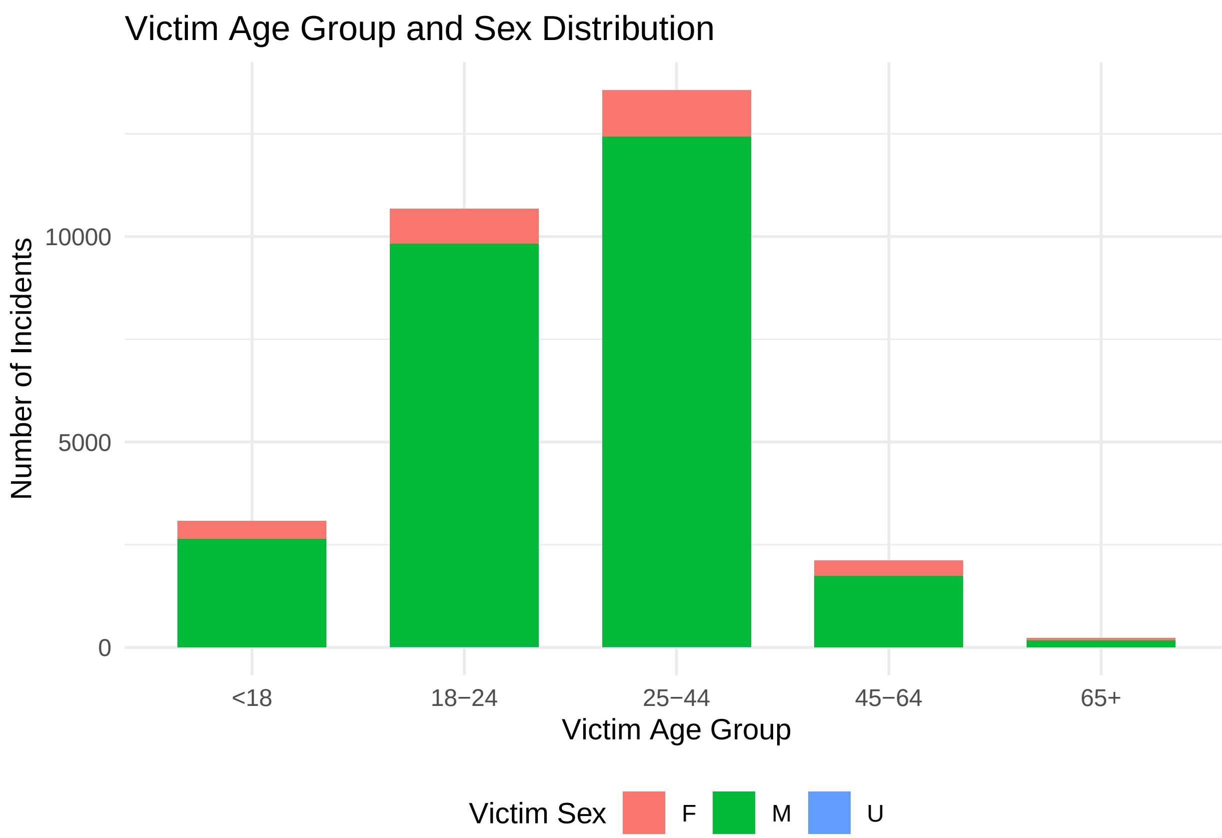
```r
data %>%
  group_by(VIC_AGE_GROUP, VIC_SEX) %>%
  summarise(total_cases = length(VIC_AGE_GROUP))
```

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

```
## # A tibble: 16 x 3
## # Groups:   VIC_AGE_GROUP [7]
##    VIC_AGE_GROUP VIC_SEX total_cases
##    <fct>         <fct>         <int>
##  1 <18           F               441
##  2 <18           M              2640
##  3 1022          M                 1
##  4 18-24         F               858
##  5 18-24         M              9815
##  6 18-24         U                 4
##  7 25-44         F              1132
##  8 25-44         M             12429
##  9 25-44         U                 2
## 10 45-64         F               385
## 11 45-64         M              1733
## 12 65+           F                70
## 13 65+           M               166
## 14 UNKNOWN       F                 5
## 15 UNKNOWN       M                57
## 16 UNKNOWN       U                 6
```

```r
data %>%
  filter(!VIC_AGE_GROUP %in% c('1022', 'UNKNOWN', '(null)'),
         !is.na(VIC_AGE_GROUP),
         !is.null(VIC_SEX)) %>%
  group_by(VIC_AGE_GROUP, VIC_SEX) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = Total_Incidents, fill = VIC_SEX)) +
  geom_col(width = 0.7) +
  scale_fill_hue(c = 100, name = "Victim Sex") +
  labs(
    title = "Victim Age Group and Sex Distribution",
    x = "Victim Age Group",
    y = "Number of Incidents"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

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



Victim Race

This plot shows the count of victims by race. The distribution reflects broader social factors and demographic patterns.

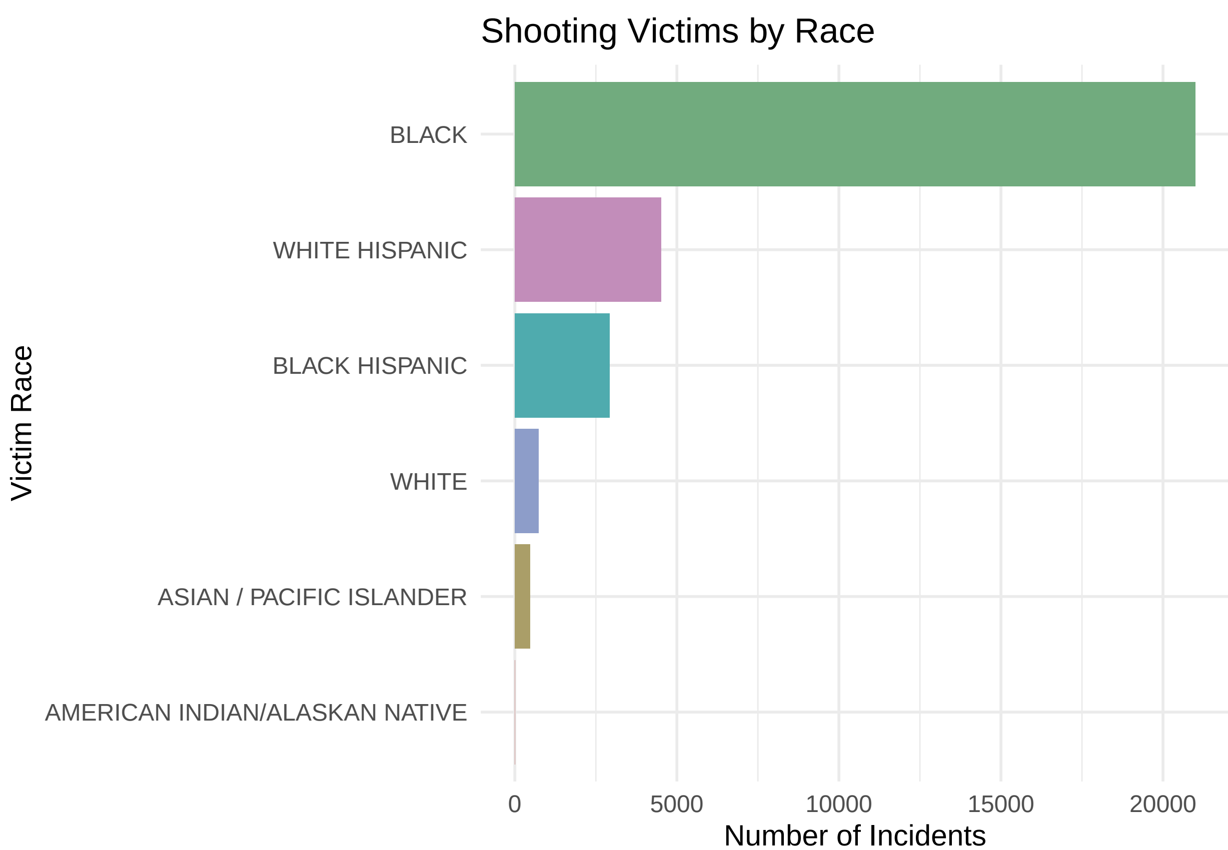
```
data %>%
  group_by(VIC_RACE) %>%
  summarise(total_cases = length(VIC_RACE))
```

```
## # A tibble: 7 x 2
##   VIC_RACE                       total_cases
##   <fct>                                <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE          13
## 2 ASIAN / PACIFIC ISLANDER               478
## 3 BLACK                                20999
## 4 BLACK HISPANIC                        2930
## 5 UNKNOWN                                 72
## 6 WHITE                                  741
## 7 WHITE HISPANIC                        4511
```

```
data %>%
  filter(!VIC_RACE %in% c('UNKNOWN', '(null)'), !is.na(VIC_RACE)) %>%
  group_by(VIC_RACE) %>%
  summarise(Total_Incidents = n()) %>%
```

```r
ggplot(aes(x = reorder(VIC_RACE, Total_Incidents), y = Total_Incidents, fill = VIC_RACE)) +
  geom_col() +
  coord_flip() +
  scale_fill_hue(c = 40, guide = FALSE) +
  labs(
    title = "Shooting Victims by Race",
    x = "Victim Race",
    y = "Number of Incidents"
  ) +
  theme_minimal()
```

```
## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
## ggplot2 3.3.4.
## i Please use "none" instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



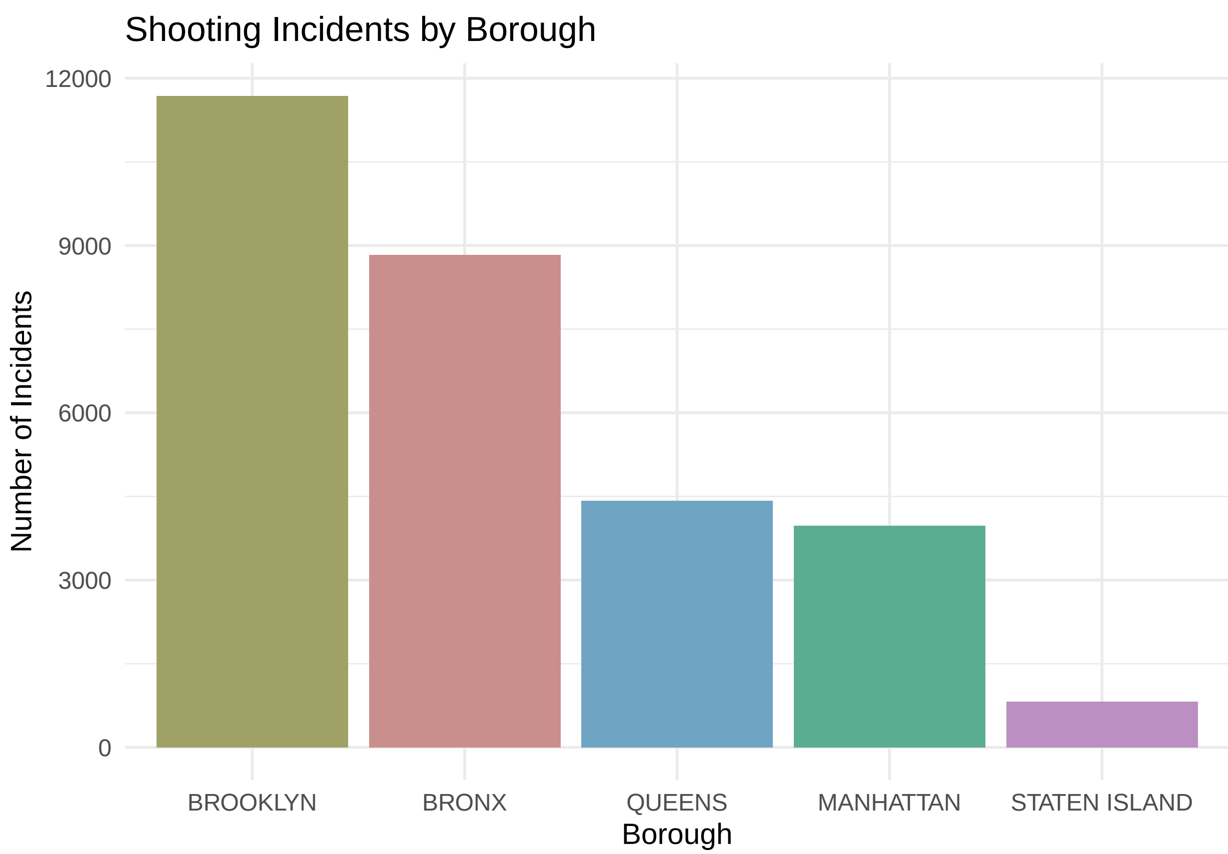
Boroughs

This bar plot shows the number of incidents by borough. Brooklyn has the highest count, while Staten Island has the lowest. The difference can be caused because of the average income in these areas.

```r
data %>%
  group_by(BORO) %>%
  summarise(total_cases = length(BORO))
```

```
## # A tibble: 5 x 2
##   BORO          total_cases
##   <fct>               <int>
## 1 BRONX                8834
## 2 BROOKLYN            11685
## 3 MANHATTAN            3977
## 4 QUEENS               4426
## 5 STATEN ISLAND         822
```

```r
data %>%
  group_by(BORO) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = reorder(BORO, -Total_Incidents), y = Total_Incidents, fill = BORO)) +
  geom_col() +
  scale_fill_hue(c = 40, guide = FALSE) +
  labs(
    title = "Shooting Incidents by Borough",
    x = "Borough",
    y = "Number of Incidents"
  ) +
  theme_minimal()
```



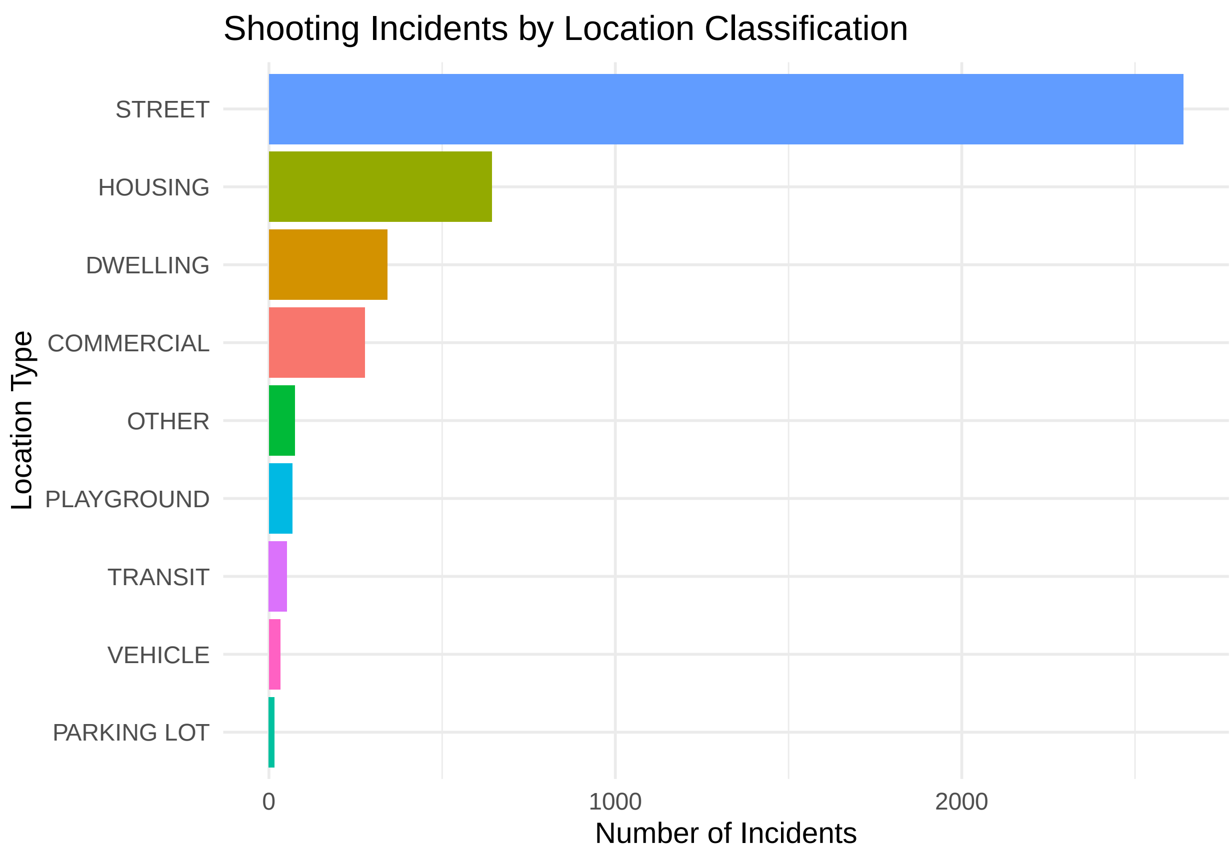
Location Classification for Incidents

This plot displays the location types with the frequency of shootings. Most number of cases are reported in street because of high population and minimum cameras. Small issues can go out of hand easily causing some gunfires. Parking lots report minimum number of issues as they are highly equipped with camera which can be risky for the perpetrator

```r
data %>%
  group_by(LOC_CLASSFCTN_DESC) %>%
  summarise(total_cases = length(LOC_CLASSFCTN_DESC))
```

```
## # A tibble: 11 x 2
##    LOC_CLASSFCTN_DESC total_cases
##    <fct>                    <int>
##  1 (null)                       7
##  2 COMMERCIAL                 276
##  3 DWELLING                   341
##  4 HOUSING                    643
##  5 OTHER                       74
##  6 PARKING LOT                 16
##  7 PLAYGROUND                  67
##  8 STREET                    2639
##  9 TRANSIT                     52
## 10 VEHICLE                     33
## 11 <NA>                     25596
```

```r
data %>%
  filter(!LOC_CLASSFCTN_DESC %in% c('NA', '(null)'), !is.na(LOC_CLASSFCTN_DESC)) %>%
  group_by(LOC_CLASSFCTN_DESC) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = reorder(LOC_CLASSFCTN_DESC, Total_Incidents), y = Total_Incidents, fill = LOC_CLASSFCTI
  geom_col() +
  coord_flip() +
  scale_fill_hue(c = 100, guide = FALSE) +
  labs(
    title = "Shooting Incidents by Location Classification",
    x = "Location Type",
    y = "Number of Incidents"
  ) +
  theme_minimal()
```



Perpetrator Age Group and Sex

This bar plot explores suspect age and sex breakdown.The graph explains that males are the top perpetrator in NY and the age-group which does this the most are from 18-24. This age group is young and filled with a lot emotions which can sometimes become violent and can lead to shootings.

```
data %>%
  group_by(PERP_AGE_GROUP, PERP_SEX) %>%
  summarise(total_cases = length(PERP_AGE_GROUP))
```

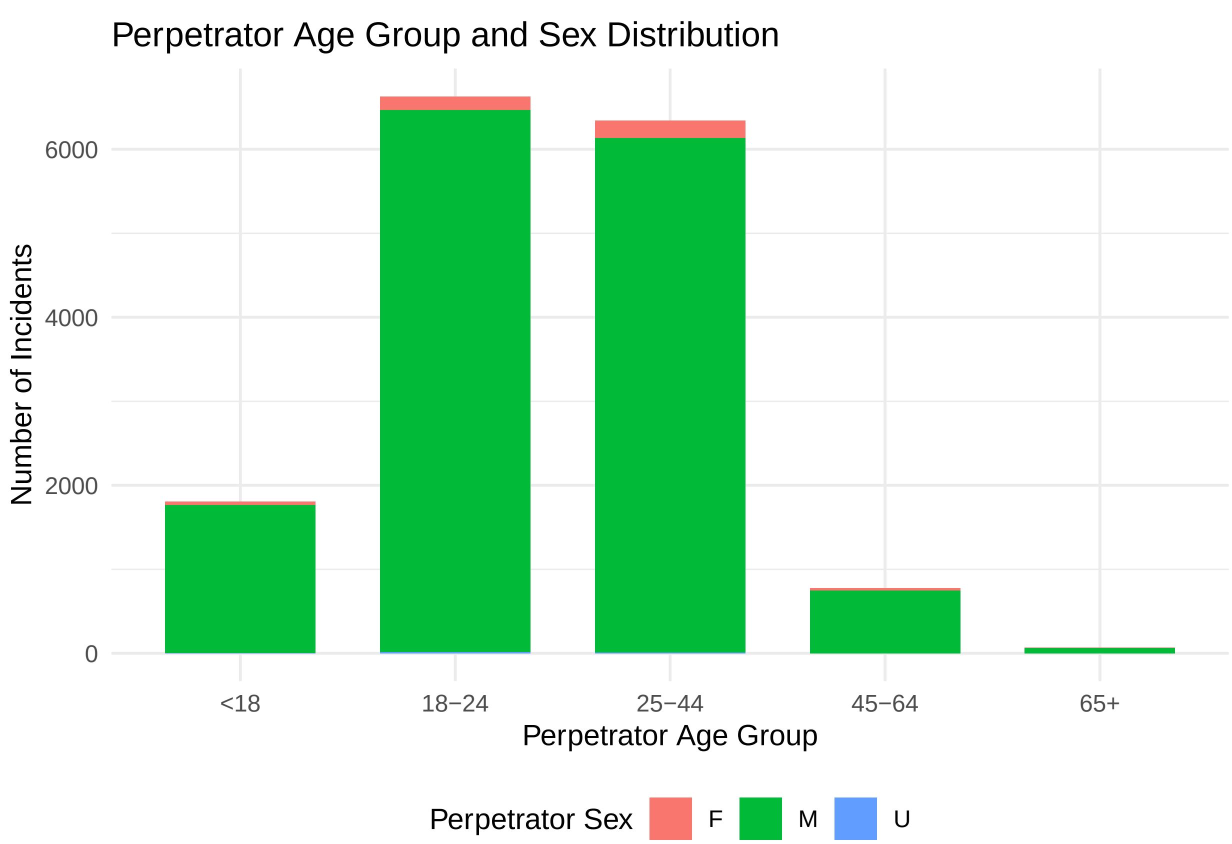
```
## `summarise()` has grouped output by 'PERP_AGE_GROUP'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 24 x 3
## # Groups:   PERP_AGE_GROUP [13]
##    PERP_AGE_GROUP PERP_SEX total_cases
##    <fct>          <fct>          <int>
##  1 (null)         (null)          1628
##  2 <18            F                 40
##  3 <18            M               1762
##  4 <18            U                  3
##  5 1020           M                  1
##  6 1028           M                  1
##  7 18-24          F                166
```

```
##   8 18-24          M                6447
##   9 18-24          U                  17
## 10 2021           M                   1
## # i 14 more rows
```

```
data %>%
  filter(!PERP_AGE_GROUP %in% c('1020', '1028', '2021', '224', '940', 'UNKNOWN', '(null)'),
         !is.na(PERP_AGE_GROUP),
         !is.null(PERP_SEX)) %>%
  group_by(PERP_AGE_GROUP, PERP_SEX) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = PERP_AGE_GROUP, y = Total_Incidents, fill = PERP_SEX)) +
  geom_col(width = 0.7) +
  scale_fill_hue(c = 100, name = "Perpetrator Sex") +
  labs(
    title = "Perpetrator Age Group and Sex Distribution",
    x = "Perpetrator Age Group",
    y = "Number of Incidents"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

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



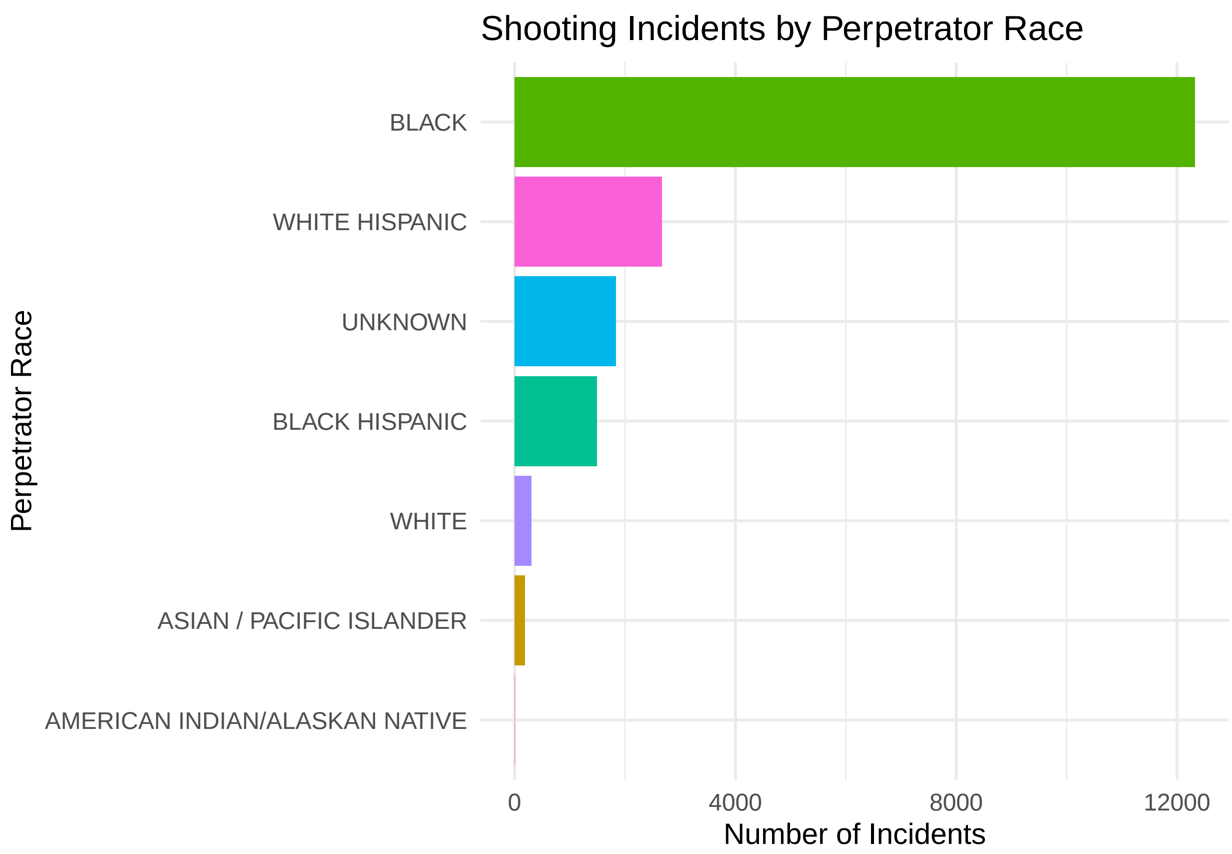
Perpetrator Race

This plot shows perpetrators by race. The graph reflects that the most number of perpetrators are People with black race. The reason for this can be to prove dominance in the society where white people mostly take the decisions.

```
data %>%
  group_by(PERP_RACE) %>%
  summarise(total_cases = length(PERP_RACE))
```

```
## # A tibble: 9 x 2
##   PERP_RACE                      total_cases
##   <fct>                                <int>
## 1 (null)                                1628
## 2 AMERICAN INDIAN/ALASKAN NATIVE           2
## 3 ASIAN / PACIFIC ISLANDER               184
## 4 BLACK                                12323
## 5 BLACK HISPANIC                        1487
## 6 UNKNOWN                               1838
## 7 WHITE                                  305
## 8 WHITE HISPANIC                        2667
## 9 <NA>                                  9310
```

```
data %>%
  filter(!PERP_RACE %in% c('NA', '(null)'), !is.na(PERP_RACE)) %>%
  group_by(PERP_RACE) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = reorder(PERP_RACE, Total_Incidents), y = Total_Incidents, fill = PERP_RACE)) +
  geom_col() +
  coord_flip() +
  scale_fill_hue(c = 100, guide = FALSE) +
  labs(
    title = "Shooting Incidents by Perpetrator Race",
    x = "Perpetrator Race",
    y = "Number of Incidents"
  ) +
  theme_minimal()
```



Murder Rates

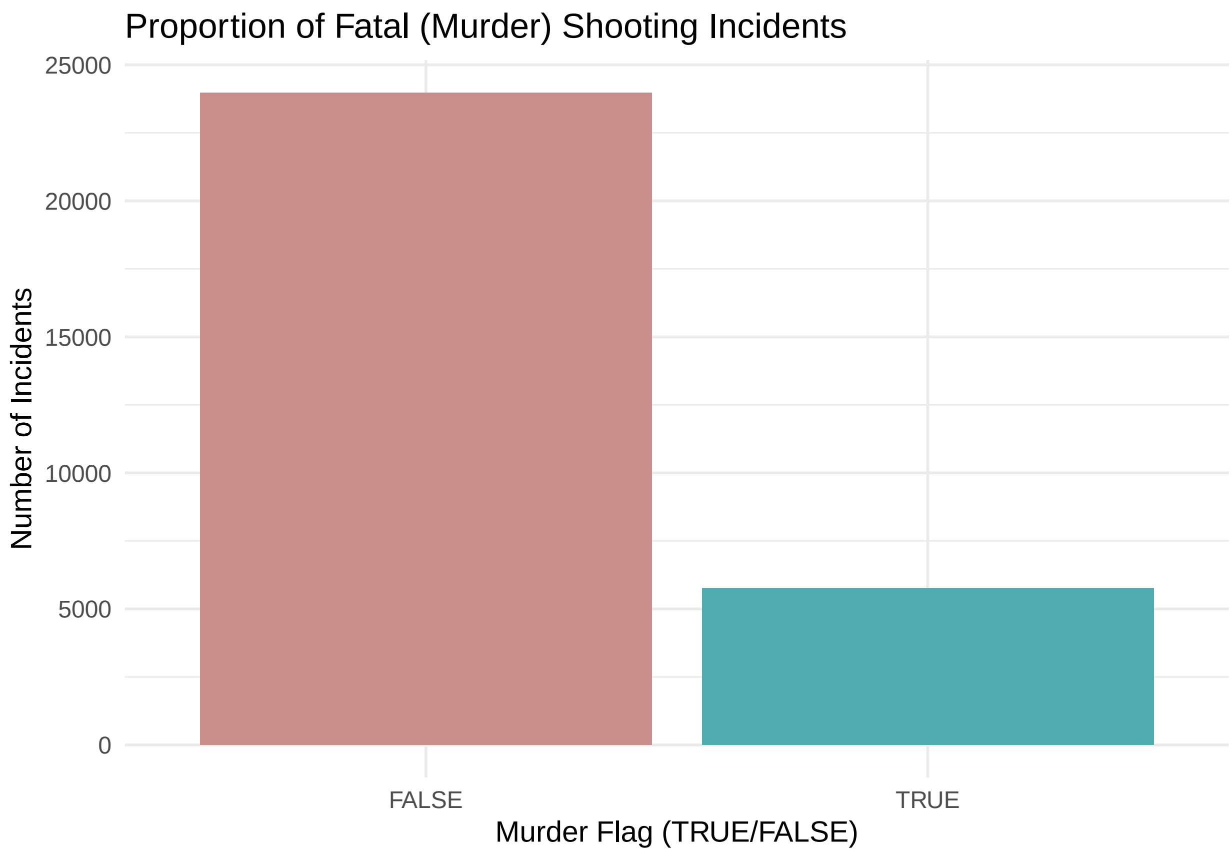
The plot below compares incidents resulting in murder with total shootings. It help us identify that most of the shooting cases doesn't lead to murder. They could have been caused to scare someone or to prove dominance and power.

```
data %>%
  group_by(STATISTICAL_MURDER_FLAG) %>%
  summarise(total_cases = length(STATISTICAL_MURDER_FLAG))
```

```
## # A tibble: 2 x 2
##   STATISTICAL_MURDER_FLAG total_cases
##   <lgl>                         <int>
## 1 FALSE                         23979
## 2 TRUE                           5765
```

```
data %>%
  group_by(STATISTICAL_MURDER_FLAG) %>%
  summarise(Total_Incidents = n()) %>%
  ggplot(aes(x = STATISTICAL_MURDER_FLAG, y = Total_Incidents, fill = as.factor(STATISTICAL_MURDER_FLAG
  geom_col() +
  scale_fill_hue(c = 40, guide = FALSE) +
  labs(
    title = "Proportion of Fatal (Murder) Shooting Incidents",
```

```
    x = "Murder Flag (TRUE/FALSE)",
    y = "Number of Incidents"
  ) +
  theme_minimal()
```



Predictive Modeling

We use to model real world situations in mathematical models to predict the future things. There are a lot of different modelling techniques in the field of data science. For this report, we want to predict that the shootout lead to a murder or not therefore coming under the tree of classification model.

Only complete cases are used. Train/test splitting is omitted for simplicity.

```
model_data = data %>%
  filter(!is.na(BORO),
         !is.na(LOC_CLASSFCTN_DESC),
         !is.na(OCCUR_HOUR),
         !is.na(OCCUR_YEAR),
         !is.na(OCCUR_WEEKDAY),
         !is.na(Latitude),
         !is.na(Longitude))
```

Logistic Regression

The model is regression technique which is used for classification problems. We identify the list of columns which can affect our decision and provide it to the model for computation. The model is created and output in predicted using the predict function.

```r
model_glm = glm(STATISTICAL_MURDER_FLAG ~ BORO + LOC_CLASSFCTN_DESC + OCCUR_HOUR + OCCUR_YEAR + OCCUR_WE
summary(model_glm)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + LOC_CLASSFCTN_DESC +
##     OCCUR_HOUR + OCCUR_YEAR + OCCUR_WEEKDAY + Latitude + Longitude,
##     family = "binomial", data = model_data)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    45.817580 140.848146   0.325   0.7450
## BOROBROOKLYN                   -0.060866   0.242306  -0.251   0.8017
## BOROMANHATTAN                   0.071894   0.145998   0.492   0.6224
## BOROQUEENS                     -0.302313   0.242108  -1.249   0.2118
## BOROSTATEN ISLAND              -0.097957   0.473394  -0.207   0.8361
## LOC_CLASSFCTN_DESCCOMMERCIAL   -0.280400   0.853249  -0.329   0.7424
## LOC_CLASSFCTN_DESCDWELLING      0.297208   0.848919   0.350   0.7263
## LOC_CLASSFCTN_DESCHOUSING      -0.884907   0.849266  -1.042   0.2974
## LOC_CLASSFCTN_DESCOTHER        -1.277733   0.931784  -1.371   0.1703
## LOC_CLASSFCTN_DESCPARKING LOT  -1.084910   1.132166  -0.958   0.3379
## LOC_CLASSFCTN_DESCPLAYGROUND   -0.063601   0.887150  -0.072   0.9428
## LOC_CLASSFCTN_DESCSTREET       -0.634995   0.843265  -0.753   0.4514
## LOC_CLASSFCTN_DESCTRANSIT       0.336088   0.891907   0.377   0.7063
## LOC_CLASSFCTN_DESCVEHICLE       1.079825   0.912215   1.184   0.2365
## OCCUR_HOUR                      0.009790   0.005100   1.920   0.0549 .
## OCCUR_YEAR                      0.005525   0.048171   0.115   0.9087
## OCCUR_WEEKDAY                   0.008846   0.018611   0.475   0.6346
## Latitude                        0.361173   1.216923   0.297   0.7666
## Longitude                       0.983288   1.188942   0.827   0.4082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4103.1  on 4050  degrees of freedom
## Residual deviance: 3993.2  on 4032  degrees of freedom
## AIC: 4031.2
##
## Number of Fisher Scoring iterations: 4
```

The confusion matrix below shows model performance. The logistic regression model achieved an accuracy of approximately 79.65%.

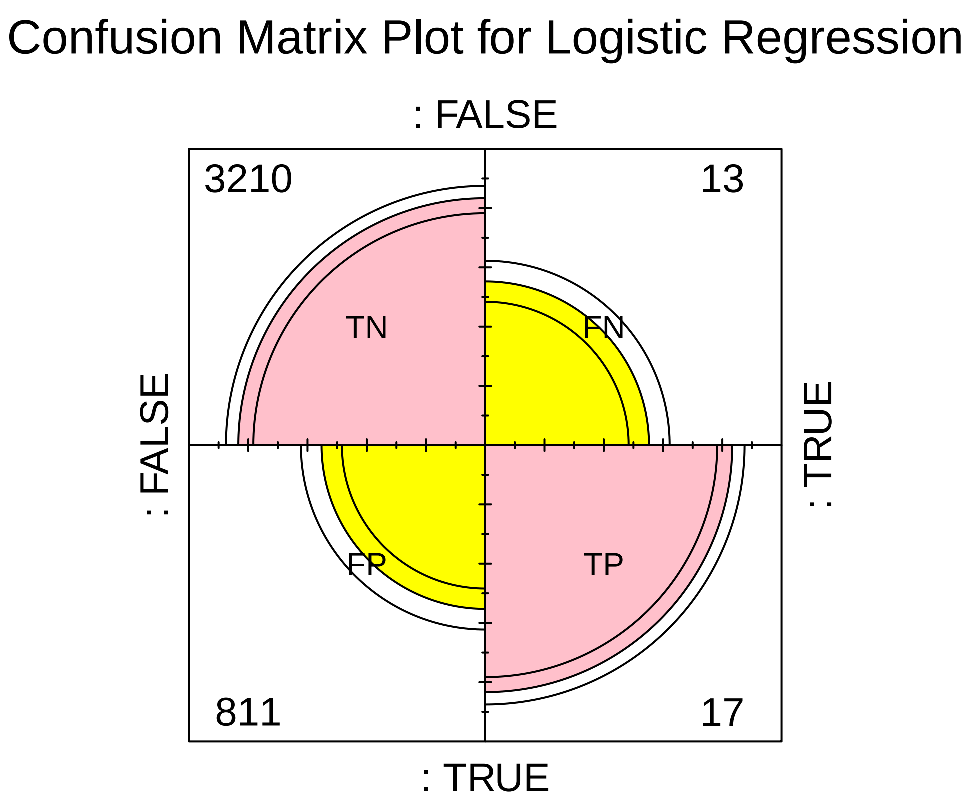
```r
predictTest = predict(model_glm, newdata = model_data, type = "response")

table(model_data$STATISTICAL_MURDER_FLAG, predictTest >= 0.5)
```

```
##
##          FALSE TRUE
##   FALSE  3210   13
##   TRUE    811   17
```

```r
(3210+17)/nrow(model_data)
```

```
## [1] 0.7965934
```

```r
fourfoldplot(table(model_data$STATISTICAL_MURDER_FLAG, predictTest >= 0.5),color=c("yellow","pink"), ma
text(-0.4,0.4, "TN", cex=1) +
text(0.4, -0.4, "TP", cex=1) +
text(0.4,0.4, "FN", cex=1) +
text(-0.4, -0.4, "FP", cex=1)
```




```
## integer(0)
```

Support Vector Machine

The model is use for classification and regression problems. It works by dividing the data points into separate classes and maximize the margin between them. We provide the list of factor columns to the model for training purpose.

```r
classifier <- svm(STATISTICAL_MURDER_FLAG ~ BORO + LOC_CLASSFCTN_DESC + OCCUR_HOUR + OCCUR_YEAR + OCCUR_
                  data = model_data,
                  type = 'C-classification',
                  kernel = 'radial',
                  gamma = 10)
```

The confusion matrix and accuracy are shown below. The SVM model achieved an accuracy of approximately 89.92%.

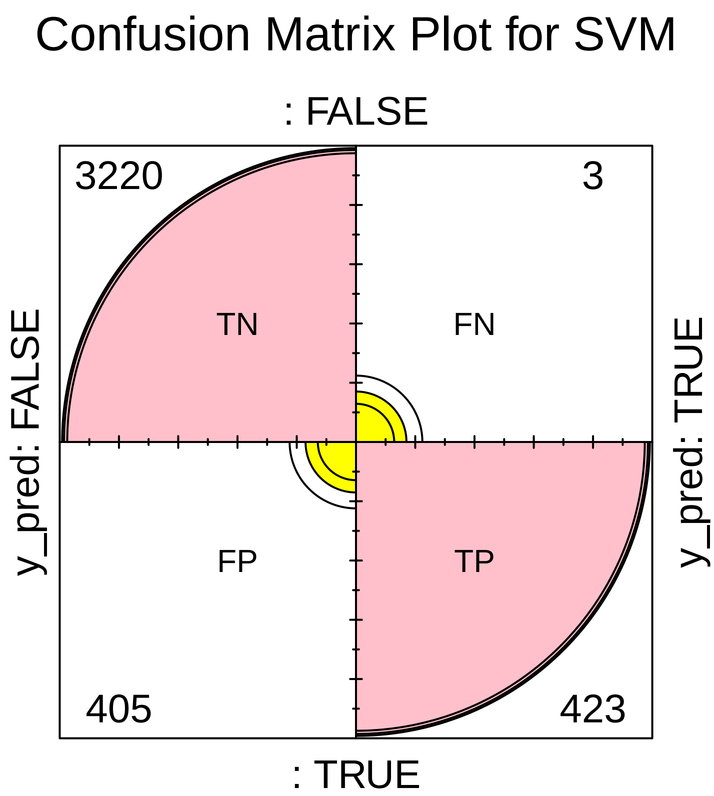
```r
y_pred <- predict(classifier, newdata = model_data)

table(model_data$STATISTICAL_MURDER_FLAG, y_pred)
```

```
##        y_pred
##         FALSE TRUE
##   FALSE  3220    3
##   TRUE    405  423
```

```r
(3220+423)/nrow(model_data)
```

```
## [1] 0.8992841
```

```r
fourfoldplot(table(model_data$STATISTICAL_MURDER_FLAG, y_pred),color=c("yellow","pink"), main = "Confus
text(-0.4,0.4, "TN", cex=1) +
text(0.4, -0.4, "TP", cex=1) +
text(0.4,0.4, "FN", cex=1) +
text(-0.4, -0.4, "FP", cex=1)
```




```
## integer(0)
```

Based on these results, the Support Vector Machine model outperforms logistic regression in this instance.

Bias

The dataset reflects higher numbers of Black suspects and victims. While this is a data-driven finding, it is important to recognize that such patterns may be influenced by a range of systemic factors, including reporting practices, policing strategies, and social determinants. Data interpretation should be undertaken with caution to avoid reinforcing societal bias or stigma. On an individual side, we might think women and old age people must be the biggest victims of this but that's not the cases from the analysis done as part of this report.

Demographic analysis indicates young males (especially ages 18-24) are both the most frequent perpetrators of shootings. It can be expected because of bias on youth people as they are filled lot of emotions. They looks for societal acceptance and always want to prove dominance. Therefore, visualizing the data in an effective and efficient manner helps in eliminating any of the human and data bias and leads the project to a proper clean direction.

Conclusion

This report provided an exploratory analysis of NYPD Shooting Incident data, including temporal, demographic, and geographic patterns. Predictive models were developed to identify risk factors for fatal shootings. The results can assist stakeholders in developing informed policy and intervention strategies, provided findings are interpreted within proper societal context and awareness of possible bias.