Project1: NYPD Shooting Incident Data

F.Mogharabin

2024-02-03

Loading Libraries

Loading tidyverse, lubridate and ggplot2 libraries:

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

Importing Data

Reading in the data from https://data.gov/ and loading it to our variable

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

```
nypd_shooting_cvs <- read_csv(url_nypd)
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

Cleaning Data

Checking summary of the imported data

```r
summary(nypd_shooting_cvs)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA’s   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD       Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA’s   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA’s   :10
```

Removing the columns that are not significant to our study from the data and converting the date and time to time objects.

```
nypd_c <- nypd_shooting_cvs %>%
    select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, INCIDENT_KEY,Lon_Lat,LOC_CLASSFCTN_DESC,LOCA
    mutate(OCCUR_DATE = mdy(OCCUR_DATE),
           OCCUR_TIME = hms(OCCUR_TIME))
```

Checking the first few row of the data

```
head(nypd_c)
```

```
## # A tibble: 6 x 11
##   OCCUR_DATE OCCUR_TIME BORO     PRECINCT STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##   <date>     <Period>   <chr>       <dbl> <lgl>                   <chr>
## 1 2021-05-27 21H 30M 0S QUEENS        105 FALSE                   <NA>
## 2 2014-06-27 17H 40M 0S BRONX          40 FALSE                   <NA>
## 3 2015-11-21 3H 56M 0S  QUEENS        108 TRUE                    <NA>
## 4 2015-10-09 18H 30M 0S BRONX          44 FALSE                   <NA>
## 5 2009-02-19 22H 58M 0S BRONX          47 TRUE                    25-44
## 6 2020-10-21 21H 36M 0S BROOKLYN       81 TRUE                    <NA>
## # i 5 more variables: PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>
```

By looking at the first few rows, it seems that some values in certain fields are missing. Using the sapply function, we apply the is.na function to each column and then sum the results to obtain the count of NAs for each column.

```
sapply(nypd_c, function(x) sum(is.na(x)))
```

```
##              OCCUR_DATE              OCCUR_TIME                    BORO
##                       0                       0                       0
##                PRECINCT STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                       0                       0                    9344
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                    9310                    9310                       0
##                 VIC_SEX                VIC_RACE
##                       0                       0
```

The information regarding the perpetrator appears to be incomplete, possibly due to the cases being unsolved or still under investigation. For now, we will ignore the missing data and leave the information as it is.

Checking the format of data in VIC_AGE_GROUP:

```
table(nypd_c$VIC_AGE_GROUP)
```

```
##
##     <18    1022   18-24   25-44   45-64     65+ UNKNOWN
##    2839       1   10086   12281    1863     181      61
```

It appears that we have a value that does not match the expected format. We will filter it out.

```r
nypd_c_filtered <- nypd_c %>%
filter(VIC_AGE_GROUP != "1022")
```

We will also update the binary value of STATISTICAL_MURDER_FLAG to ‘Fatal’ and ‘Non-Fatal’ to make it easier for the viewer to understand.

```r
STATISTICAL_MURDER_FLAG_factor <- as.factor(nypd_c_filtered$STATISTICAL_MURDER_FLAG)
```

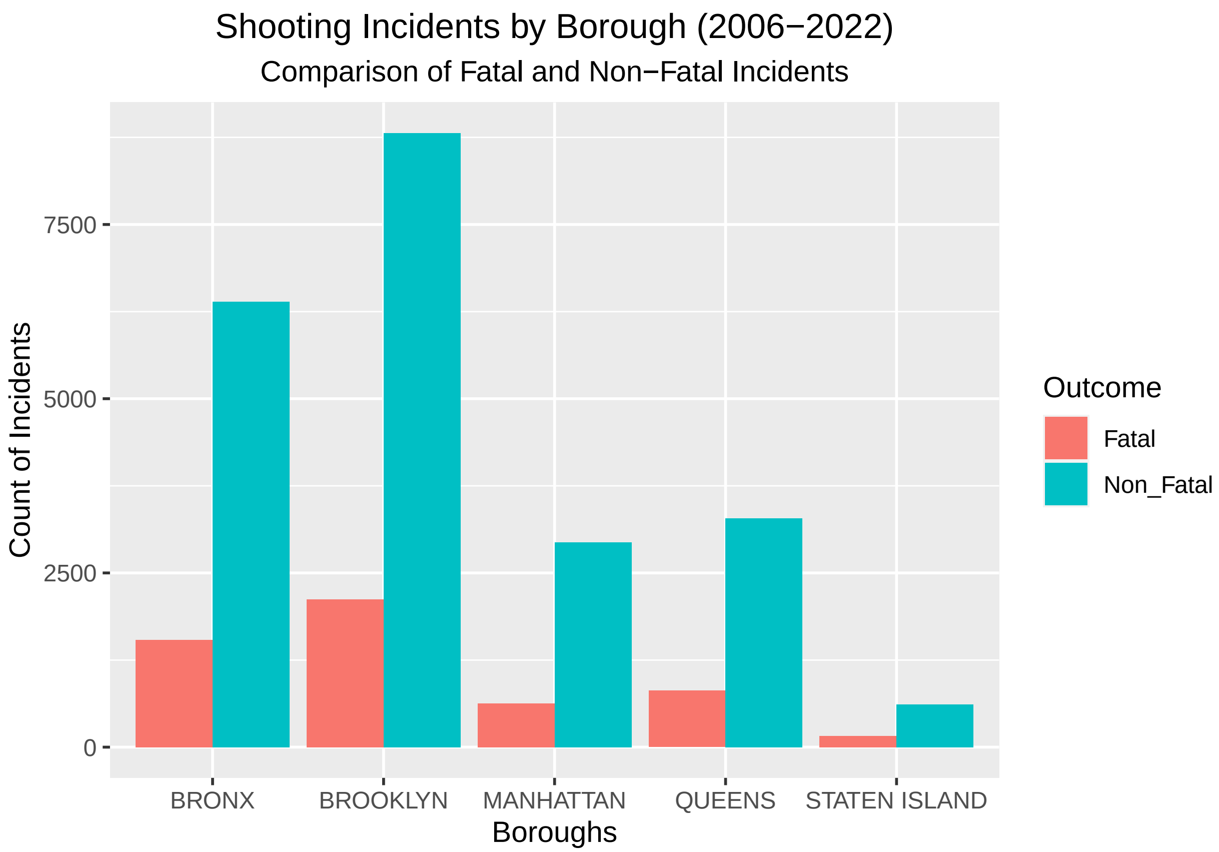
```r
nypd_c_filtered <- nypd_c_filtered %>%
  mutate(Outcome = ifelse(STATISTICAL_MURDER_FLAG, "Fatal", "Non_Fatal"))
```

Parsing the OCCUR_DATE into Year, Month, and Weekday.

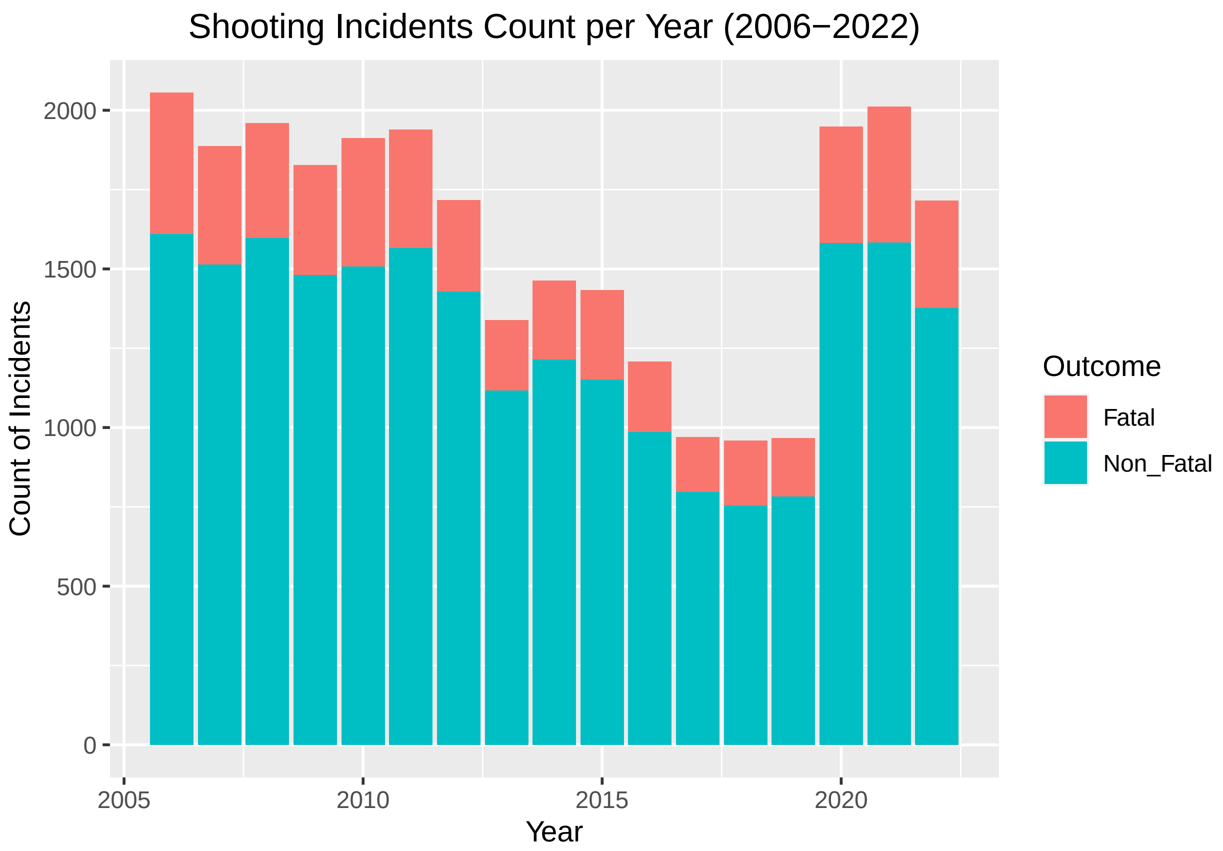
```r
nypd_c_filtered <- nypd_c_filtered %>%
  mutate(OCCUR_DATE = parse_date_time(as.character(OCCUR_DATE), orders = c("mdy", "my", "ymd")),
         Year = year(OCCUR_DATE),
         Month = month(OCCUR_DATE, label = TRUE),
         Weekday = format(OCCUR_DATE, "%A"))
```

Visualizations

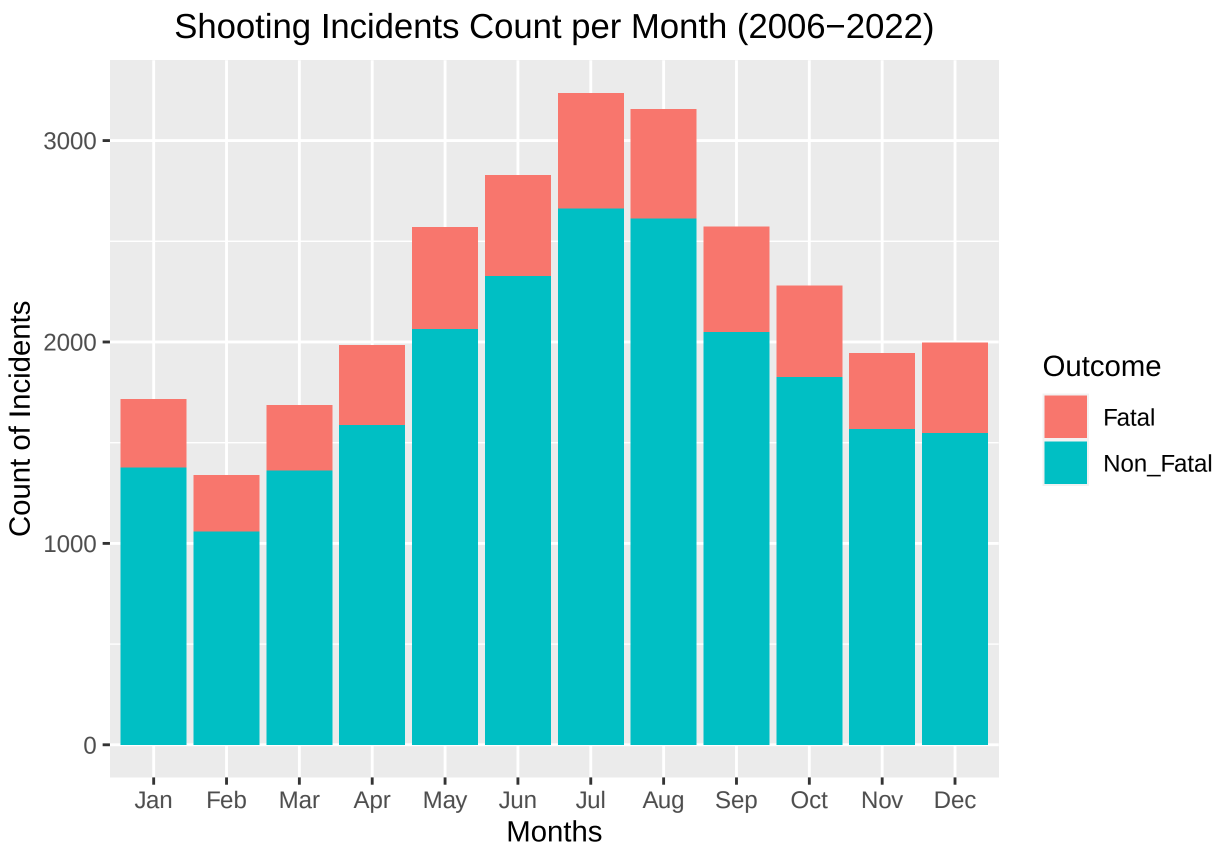
```r
ggplot(nypd_c_filtered, aes(x = BORO,fill = Outcome)) +
    geom_bar(position = "dodge") +
    labs(x = "Boroughs", y = "Count of Incidents", title = "Shooting Incidents by Borough (2006-2022)",
    theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```



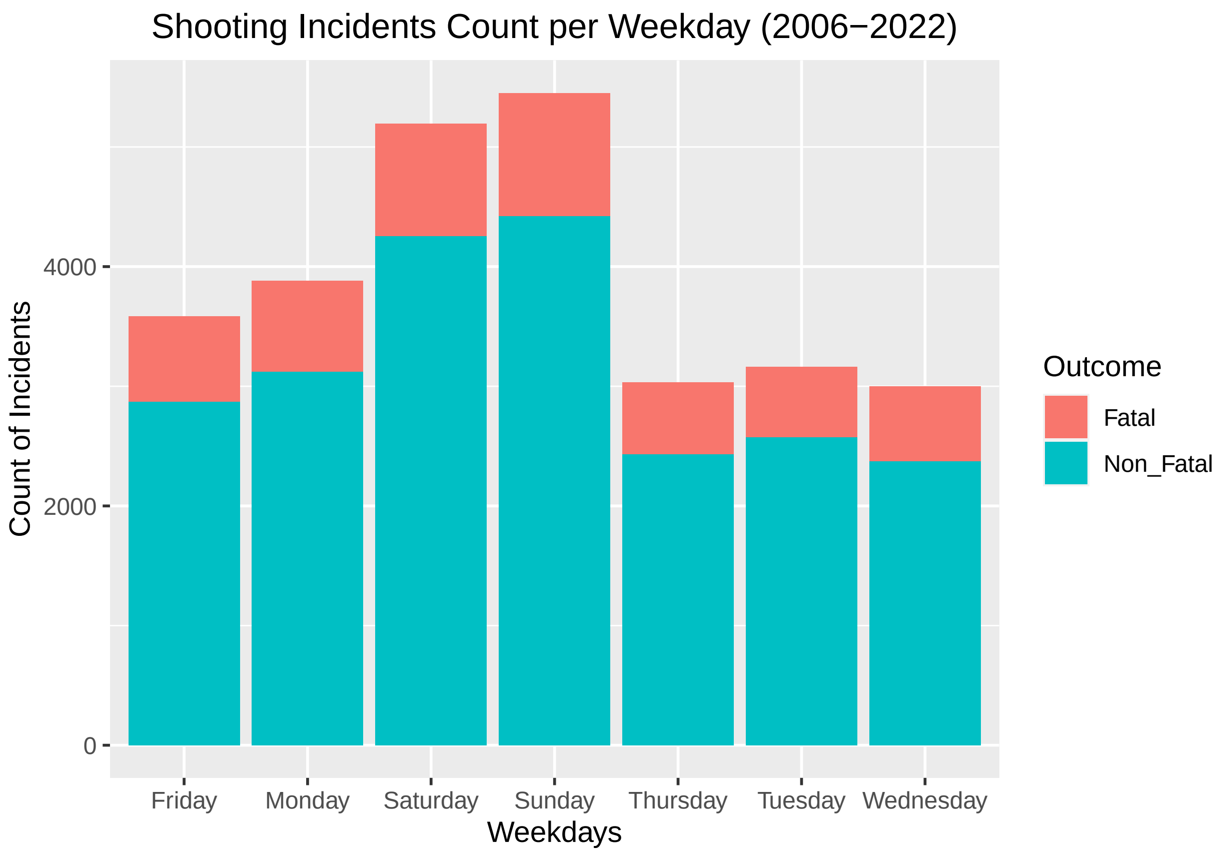
```
ggplot(nypd_c_filtered, aes(x = Year,fill = Outcome)) +
  geom_bar() +
  labs(title = "Shooting Incidents Count per Year (2006-2022)",
       x = "Year",
       y = "Count of Incidents")+
      theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```



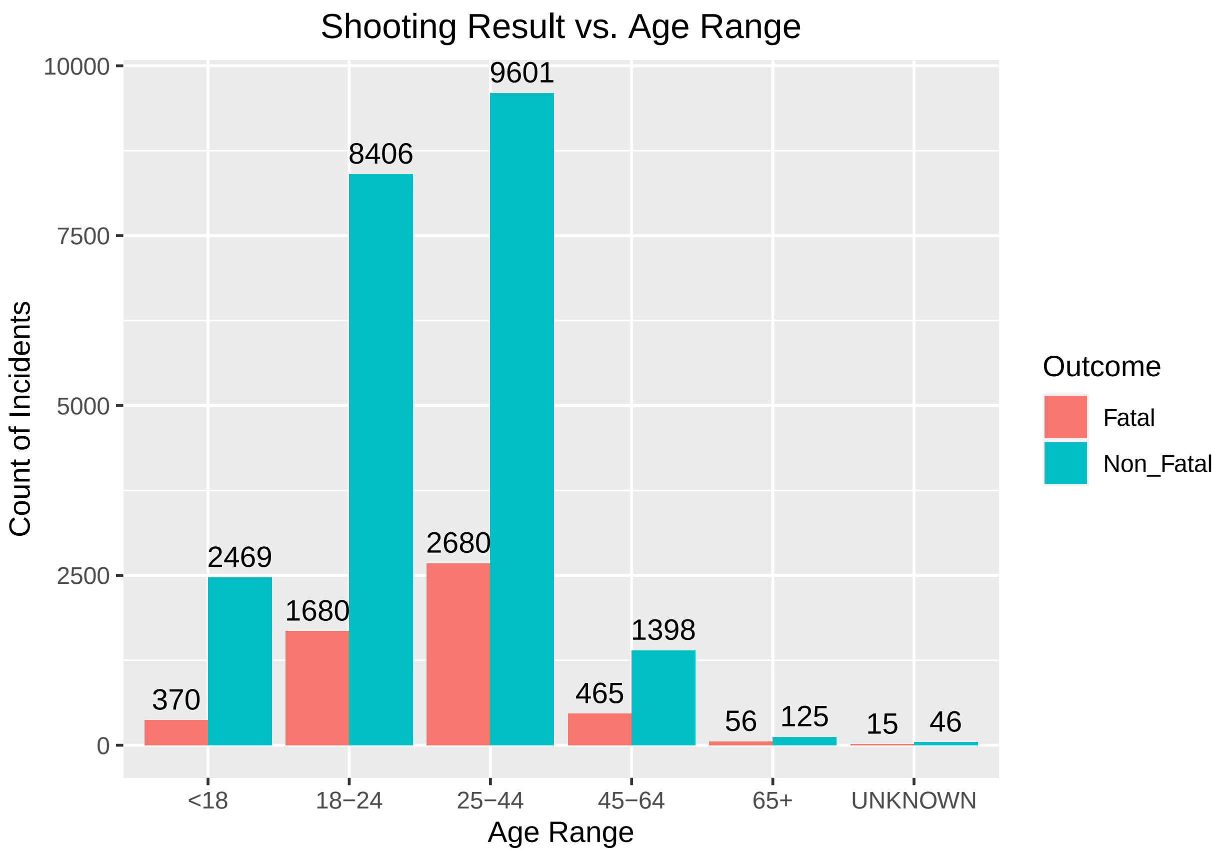
```
ggplot(nypd_c_filtered, aes(x = Month,fill = Outcome)) +
  geom_bar() +
  labs(title = "Shooting Incidents Count per Month (2006-2022)",
       x = "Months",
       y = "Count of Incidents")+
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```



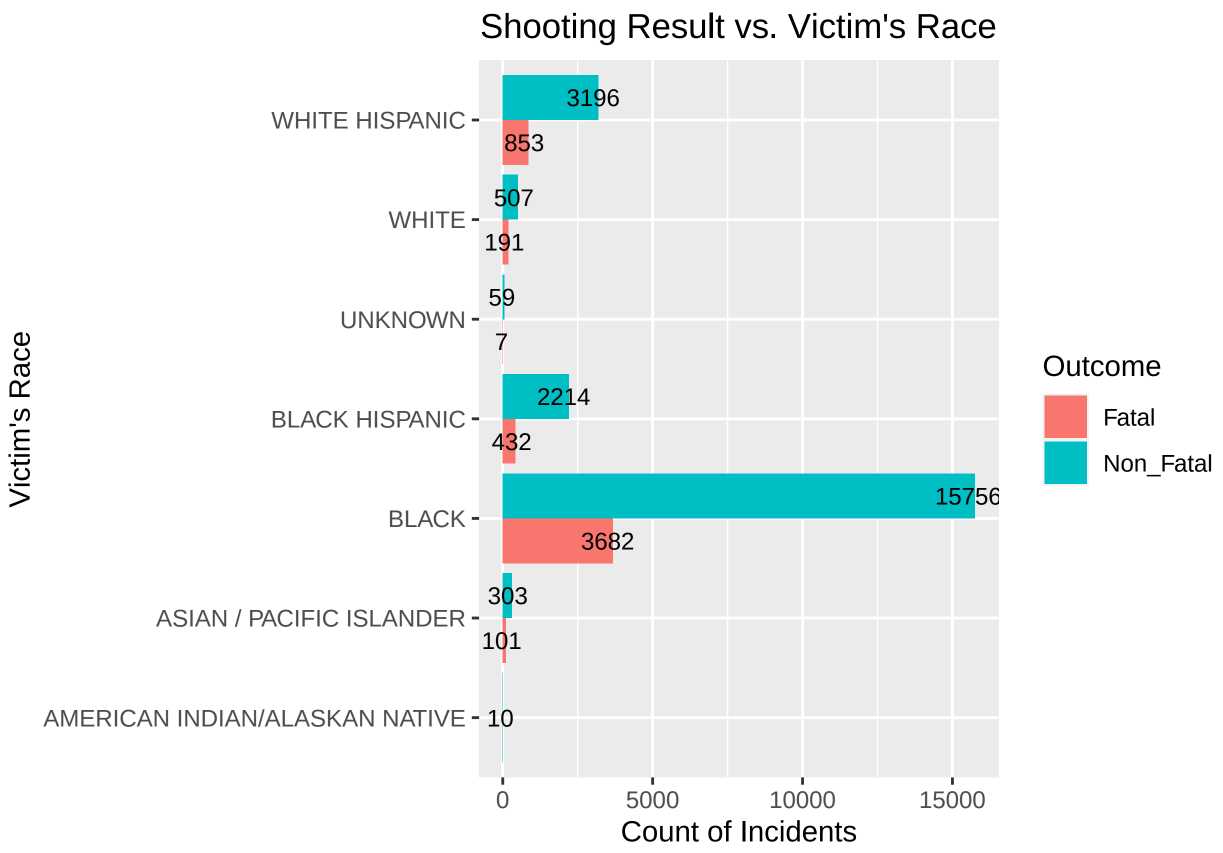
```r
ggplot(nypd_c_filtered, aes(x = Weekday,fill = Outcome)) +
  geom_bar() +
  labs(title = "Shooting Incidents Count per Weekday (2006-2022)",
       x = "Weekdays",
       y = "Count of Incidents")+
       theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```



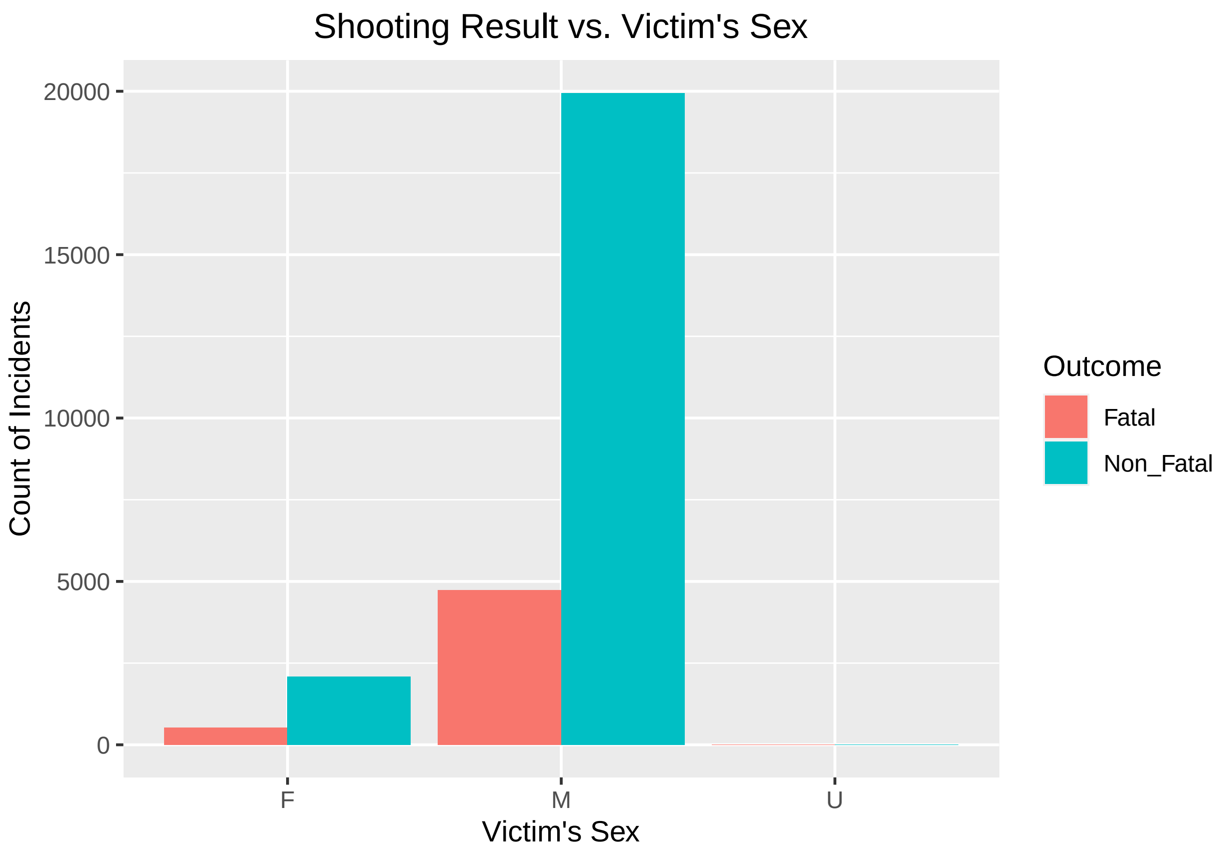
```
ggplot(nypd_c_filtered, aes(x = VIC_AGE_GROUP, fill = Outcome)) +
geom_bar(position = "dodge") +
  geom_text(stat = 'count', aes(label = after_stat(count)), position = position_dodge(width = 0.9), vju
labs(title = "Shooting Result vs. Age Range",
x = "Age Range",
y = "Count of Incidents",
fill = "Outcome") +theme(plot.title = element_text(hjust = 0.5))
```



```
ggplot(nypd_c_filtered, aes(x = VIC_RACE, fill = Outcome)) +
  geom_bar(position = "dodge") +
  geom_text(stat = 'count', aes(label = after_stat(count), group = Outcome),
            position = position_dodge(width = 0.9), hjust = .6, size = 3)+
  coord_flip()+
  labs(title = "Shooting Result vs. Victim's Race",
       x = "Victim's Race",
       y = "Count of Incidents",
       fill = "Outcome") + theme(plot.title = element_text(hjust = 0.5))
```



```
ggplot(nypd_c_filtered, aes(x = VIC_SEX, fill = Outcome)) +
  geom_bar(position = "dodge")+
  labs(title = "Shooting Result vs. Victim's Sex",
       x = "Victim's Sex",
       y = "Count of Incidents",
       fill = "Outcome") + theme(plot.title = element_text(hjust = 0.5))
```



Model

To create our model, we use features such as age, race, and sex as predictors to forecast the shooting outcome. We employ multivariate linear regression followed by logistic regression to examine how the results differ.

Convert "Outcome" to a binary numeric variable. The new variable will have a value of 1 if the shooting incident resulted in a fatality ("Fatal") and 0 if the outcome was non-fatal ("Non-Fatal").

```r
nypd_c_filtered$Outcome_numeric <- as.numeric(nypd_c_filtered$Outcome == "Fatal")
```

Fit a Multivariate Linear Regression

```r
model <- lm(Outcome_numeric ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX, data = nypd_c_filtered)
```

Print the summary of the model

```r
summary(model)
```

```
##
## Call:
```

```
## lm(formula = Outcome_numeric ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX,
##     data = nypd_c_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3636 -0.2150 -0.1645 -0.1287  0.9657
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -0.037132   0.124514  -0.298   0.7655
## VIC_AGE_GROUP18-24              0.035796   0.008361   4.282 1.86e-05 ***
## VIC_AGE_GROUP25-44              0.086230   0.008201  10.515  < 2e-16 ***
## VIC_AGE_GROUP45-64              0.112839   0.011776   9.582  < 2e-16 ***
## VIC_AGE_GROUP65+                0.166065   0.030229   5.494 3.97e-08 ***
## VIC_AGE_GROUPUNKNOWN            0.131035   0.053081   2.469   0.0136 *
## VIC_RACEASIAN / PACIFIC ISLANDER 0.221248   0.125707   1.760   0.0784 .
## VIC_RACEBLACK                   0.173231   0.124200   1.395   0.1631
## VIC_RACEBLACK HISPANIC          0.147177   0.124402   1.183   0.2368
## VIC_RACEUNKNOWN                 0.078856   0.134327   0.587   0.5572
## VIC_RACEWHITE                   0.234627   0.125088   1.876   0.0607 .
## VIC_RACEWHITE HISPANIC          0.192528   0.124322   1.549   0.1215
## VIC_SEXM                       -0.007373   0.008145  -0.905   0.3654
## VIC_SEXU                       -0.073390   0.124012  -0.592   0.5540
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3926 on 27297 degrees of freedom
## Multiple R-squared:  0.009987,   Adjusted R-squared:  0.009516
## F-statistic: 21.18 on 13 and 27297 DF,  p-value: < 2.2e-16
```

Fit a Logistic Regression

```
logistic_model <- glm(Outcome_numeric ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX, data = nypd_c_filtered, fam
```

Print the summary of the model

```
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Outcome_numeric ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX,
##     family = "binomial", data = nypd_c_filtered)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -12.86411  102.16039  -0.126  0.89979
## VIC_AGE_GROUP18-24               0.28558    0.06197   4.608 4.06e-06 ***
## VIC_AGE_GROUP25-44               0.61260    0.06005  10.201  < 2e-16 ***
## VIC_AGE_GROUP45-64               0.75940    0.07781   9.759  < 2e-16 ***
## VIC_AGE_GROUP65+                 1.01923    0.17146   5.944 2.78e-09 ***
## VIC_AGE_GROUPUNKNOWN             0.87539    0.31661   2.765  0.00569 **
## VIC_RACEASIAN / PACIFIC ISLANDER  11.28112  102.16043   0.110  0.91207
## VIC_RACEBLACK                     11.00312  102.16037   0.108  0.91423
## VIC_RACEBLACK HISPANIC            10.82204  102.16038   0.106  0.91564
## VIC_RACEUNKNOWN                   10.25876  102.16123   0.100  0.92001
## VIC_RACEWHITE                     11.34231  102.16041   0.111  0.91160
## VIC_RACEWHITE HISPANIC            11.12434  102.16038   0.109  0.91329
## VIC_SEXM                          -0.04773    0.05206  -0.917  0.35928
## VIC_SEXU                          -0.58948    1.08280  -0.544  0.58616
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26780  on 27310  degrees of freedom
## Residual deviance: 26504  on 27297  degrees of freedom
## AIC: 26532
##
## Number of Fisher Scoring iterations: 11
```

Analysis

The linear regression analysis examines how age, race, and gender (predictors) are related to the outcomes of shooting events. The model suggests that age might influence the outcome of shooting incidents, whereas the impacts of gender and race are less evident. Individuals in the age groups 18-24, 25-44, 45-64, and 65+ tend to have higher average outcomes compared to those younger than 18. Being male is not strongly associated with a significant increase or decrease in average outcomes compared to being female. While some races show higher average outcomes, not all are statistically significant. The model’s overall ability to explain the outcomes is limited, as indicated by a low multiple R-squared (0.009987).

Similarly, in logistic regression, age, race, and gender (predictors) are used to relate to the outcomes of shooting events. People in the age groups 18-24, 25-44, 45-64, and 65+ have higher odds of being in a shooting incident with a fatal outcome compared to those younger than 18. Being male is associated with lower odds of being in a fatal shooting incident compared to being female. The impact of race remains unclear. The model required 11 iterations to find the best fit; however, it is still not perfect.

Conclusion

In studying NYPD shooting data, we checked how age, race, and gender relate to outcomes. Our visuals showed patterns over time. Age seemed linked to outcomes, but gender and race were less clear. Looking specifically at fatal incidents, age stood out again. Males had lower odds of fatal incidents, adding nuance to gender dynamics. However, our models couldn’t fully explain outcomes, suggesting we need more research and factors. While we found some trends, understanding these incidents is complex. Future studies could explore additional factors and consider location influences.

This analysis has biases, such as incomplete perpetrator information, potentially leading to bias. The dataset may not cover all factors affecting incidents, like social conditions, law enforcement practices, or community dynamics. My views on gun control and the current atmosphere might have influenced interpretations, but I aimed for an impartial analysis, relying on factual evidence and statistical findings rather than pre-existing assumptions.