DTSA 5301 - Project_1_NYPD

DTSA 5301 MSDS Student, Oct 2024

2024-10-08

Data Analysis Using NYPD Shooting Historic Data.

Loading NYPD Shooting Historic Data and Read.

The primary source of data for this project is publicly available from NYPD website.

A link for the same is here

Description of the data

From the description available on the website, data set contains list of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year, 2023. NYPD claims the data to be manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website.

Lets read the data and inspect it ourselves.

```
#URL for the NYPD shooting data available in CSV
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

#Read in the csv
nypd_data <- read_csv(url_in, show_col_types = FALSE)
```

Inspect the data

```
#inspect the data to understand it better
dim(nypd_data)
```

```
## [1] 28562    21
```

```
colnames(nypd_data)
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "LOC_OF_OCCUR_DESC"       "PRECINCT"
##  [7] "JURISDICTION_CODE"       "LOC_CLASSFCTN_DESC"
##  [9] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
## [11] "PERP_AGE_GROUP"          "PERP_SEX"
## [13] "PERP_RACE"               "VIC_AGE_GROUP"
```

```
## [15] "VIC_SEX"                "VIC_RACE"
## [17] "X_COORD_CD"             "Y_COORD_CD"
## [19] "Latitude"               "Longitude"
## [21] "Lon_Lat"
```

```
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms         Class :character
##  Median : 92711254   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  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:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
```

```
##  NA's   :59
```

```
#more inspection with results not included in report
str(nypd_data)
```

Priliminary Observation of the data We can see data has 21 columns and 28562 rows of records that representing shooting incident including information about the event, location (longitude, latitude), date and time of occurrence along with information related to perpetrator's and victim's demographics, age, etc along with information about the precint, jurisdiction, borough/counties.

Clean the data

Lets clean the data for analysis.

```
#from summary, we also see there are few NA's, esp for latitude and Longitude.
#Lets drop those rows from the data set and also, drop 2 rows where jurisdiction code is NA.
nypd_data_cleaned <- nypd_data %>%
  filter(!is.na(Latitude) & !is.na(Longitude) & !is.na(JURISDICTION_CODE))

#there are (null) values I see, lets get a count
null_val_counts <- sapply(nypd_data_cleaned, function(x) sum(x == "(null)", na.rm = TRUE))
null_val_counts
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                       0                        0                        0
##                    BORO      LOC_OF_OCCUR_DESC                 PRECINCT
##                       0                        0                        0
##       JURISDICTION_CODE     LOC_CLASSFCTN_DESC            LOCATION_DESC
##                       0                        2                     1668
## STATISTICAL_MURDER_FLAG         PERP_AGE_GROUP                 PERP_SEX
##                       0                     1115                     1115
##               PERP_RACE          VIC_AGE_GROUP                  VIC_SEX
##                    1115                        0                        0
##                VIC_RACE             X_COORD_CD               Y_COORD_CD
##                       0                        0                        0
##                Latitude              Longitude                  Lon_Lat
##                       0                        0                        0
```

```
#there are NA values I see, lets get a count
na_val_counts <- sapply(nypd_data_cleaned, function(x) sum(is.na(x)))
na_val_counts
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                       0                        0                        0
##                    BORO      LOC_OF_OCCUR_DESC                 PRECINCT
##                       0                    25594                        0
##       JURISDICTION_CODE     LOC_CLASSFCTN_DESC            LOCATION_DESC
##                       0                    25594                    14976
## STATISTICAL_MURDER_FLAG         PERP_AGE_GROUP                 PERP_SEX
##                       0                     9344                     9310
##               PERP_RACE          VIC_AGE_GROUP                  VIC_SEX
```

```
##                    9310                       0                       0
##                VIC_RACE             X_COORD_CD             Y_COORD_CD
##                       0                       0                       0
##                Latitude              Longitude                 Lon_Lat
##                       0                       0                       0
```

```
#there are a lot of NA values and NULL values esp for perp-age-group(1115),
#perp-race(1115) and perp-sex(1115). Set the values to "UNKNOWN" to use that
#category if needed, instead of dropping the rows as it significantly
#affects/introduces bias.

nypd_data_cleaned$PERP_AGE_GROUP[is.na(nypd_data_cleaned$PERP_AGE_GROUP)] <- "Unknown"
nypd_data_cleaned$PERP_RACE[is.na(nypd_data_cleaned$PERP_RACE)] <- "Unknown"
nypd_data_cleaned$PERP_SEX[is.na(nypd_data_cleaned$PERP_SEX)] <- "Unknown"

#examine each columns as table
table(nypd_data_cleaned$PERP_AGE_GROUP)
```

```
##
##  (null)     <18    1020   18-24     224   25-44   45-64     65+     940 Unknown
##    1115    1673       1    6425       1    6032     697      65       1    9344
## UNKNOWN
##    3147
```

```
table(nypd_data_cleaned$PERP_RACE)
```

```
##
##                          (null) AMERICAN INDIAN/ALASKAN NATIVE
##                            1115                               2
##       ASIAN / PACIFIC ISLANDER                           BLACK
##                             169                           11880
##                  BLACK HISPANIC                         Unknown
##                            1388                            9310
##                         UNKNOWN                           WHITE
##                            1837                             298
##                  WHITE HISPANIC
##                            2502
```

```
table(nypd_data_cleaned$PERP_SEX)
```

```
##
##  (null)       F       M       U Unknown
##    1115     443   16134    1499    9310
```

```
#replace (null) values with "Unknown" and merge "UNKNOWN" with "Unknown"
nypd_data_cleaned$PERP_AGE_GROUP[nypd_data_cleaned$PERP_AGE_GROUP == "(null)"] <- "Unknown"
nypd_data_cleaned$PERP_RACE[nypd_data_cleaned$PERP_RACE == "(null)"] <- "Unknown"
nypd_data_cleaned$PERP_SEX[nypd_data_cleaned$PERP_SEX == "(null)"] <- "Unknown"
nypd_data_cleaned$PERP_AGE_GROUP[nypd_data_cleaned$PERP_AGE_GROUP == "UNKNOWN"] <- "Unknown"
nypd_data_cleaned$PERP_RACE[nypd_data_cleaned$PERP_RACE == "UNKNOWN"] <- "Unknown"
nypd_data_cleaned$PERP_SEX[nypd_data_cleaned$PERP_SEX == "UNKNOWN"] <- "Unknown"
```

```r
#replace incorrect values in PERP_AGE_GROUP
nypd_data_cleaned$PERP_AGE_GROUP[nypd_data_cleaned$PERP_AGE_GROUP %in%
                                   c("1020", "224", "940")] <- "Unknown"

#replace incorrect values in VIC_AGE_GROUP
nypd_data_cleaned$VIC_AGE_GROUP[nypd_data_cleaned$VIC_AGE_GROUP %in% c("1022")] <- "Unknown"

#re-examine each columns as table
table(nypd_data_cleaned$PERP_AGE_GROUP)
```

```
##
##     <18   18-24   25-44   45-64     65+ Unknown
##    1673    6425    6032     697      65   13609
```

```r
table(nypd_data_cleaned$PERP_RACE)
```

```
##
## AMERICAN INDIAN/ALASKAN NATIVE       ASIAN / PACIFIC ISLANDER
##                              2                            169
##                          BLACK                 BLACK HISPANIC
##                          11880                           1388
##                        Unknown                          WHITE
##                          12262                            298
##                WHITE HISPANIC
##                           2502
```

```r
table(nypd_data_cleaned$PERP_SEX)
```

```
##
##       F       M       U Unknown
##     443   16134    1499   10425
```

```r
table(nypd_data_cleaned$VIC_AGE_GROUP)
```

```
##
##     <18   18-24   25-44   45-64     65+ Unknown UNKNOWN
##    2946   10362   12945    1978     205       1      64
```

Setting goals for analysis of the data

After getting a sense of the data from preliminary inspection and knowing what kind of data I have, I found that there are about 13606 incidents where perpetrator's age group information is not available, about 12262 incidents where perpetrators race information is not available and about 10425 incidents where perpetrators gender is not known. Such a large missing information will affect some type of analysis I was interested in doing, but will continue with that as bias from the data over any analysis.

What am I interested in with this data set ?

I am interested to explore the data and analyze it for the 4 of the below questions.

My questions, finding answers which will be the goal of this project for, are:

1. What is the profile/demographic relationship between perpetrators and victims in nyc shooting incidents ?
2. Could there be any seasonal trends/months of the year when shootings are more frequent ?
3. Which date of the year, every year since 2003, saw maximum shootings ?
4. What is the ratio/proportion of shootings that involve: a. female perpetrators and male victims. b. female perpetrators and female victims. c. male perpetrators and male victims. d. male perpetrators and female victims.

Lets analyze !

Analysis - Case 1:

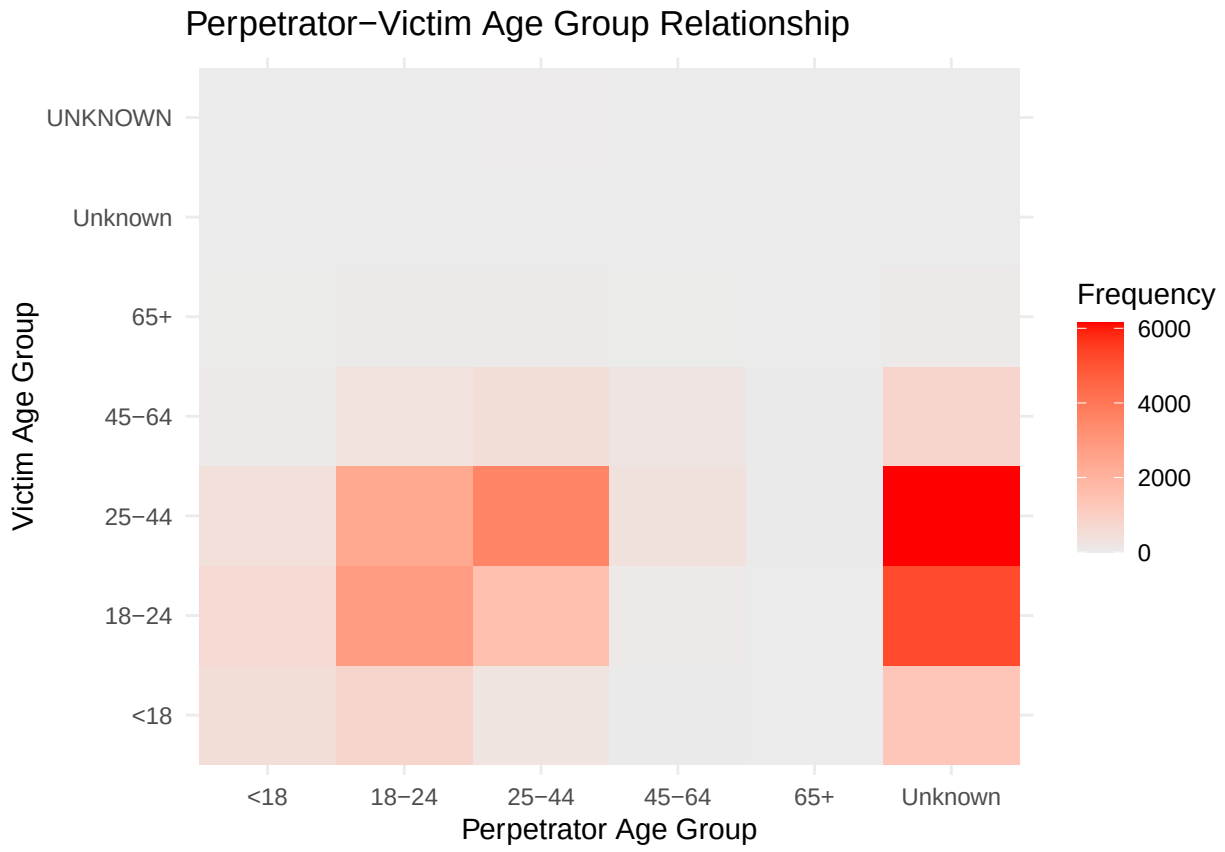
What is the profile/demographic relationships between perpetrators and victims in nyc shooting incidents ?

```
# make a table for perp-victim using age group
age_group_relationship <- table(nypd_data_cleaned$PERP_AGE_GROUP,
                                nypd_data_cleaned$VIC_AGE_GROUP)

# view the table
age_group_relationship
```

Perpetrator-Victim Age Group Relationship

```
##
##             <18 18-24 25-44 45-64  65+ Unknown UNKNOWN
##    <18      517   648   412    79   15       0       2
##    18-24    808  2834  2388   335   47       1      12
##    25-44    270  1558  3594   523   49       0      38
##    45-64     21    85   371   202   13       0       5
##    65+        0     2    27    24   12       0       0
##    Unknown 1330  5235  6153   815   69       0       7
```

```
# plot the relationship using a heatmap
library(ggplot2)
ggplot(as.data.frame(age_group_relationship), aes(Var1, Var2, fill = Freq)) +
  geom_tile() +
  #scale_fill_viridis_c(option = "plasma") +
  scale_fill_gradient(low = "grey92", high = "red") +
  labs(title = "Perpetrator-Victim Age Group Relationship",
       x = "Perpetrator Age Group",
       y = "Victim Age Group",
       fill = "Frequency") +
  theme_minimal()
```



Visual Observation/Inference:

- From the heat map, we can see majority of victims are in the age-group of 25-44 and majority of the perpetrator's (excluding those whose age-group is unknown) are also in the age group of 25-44 !

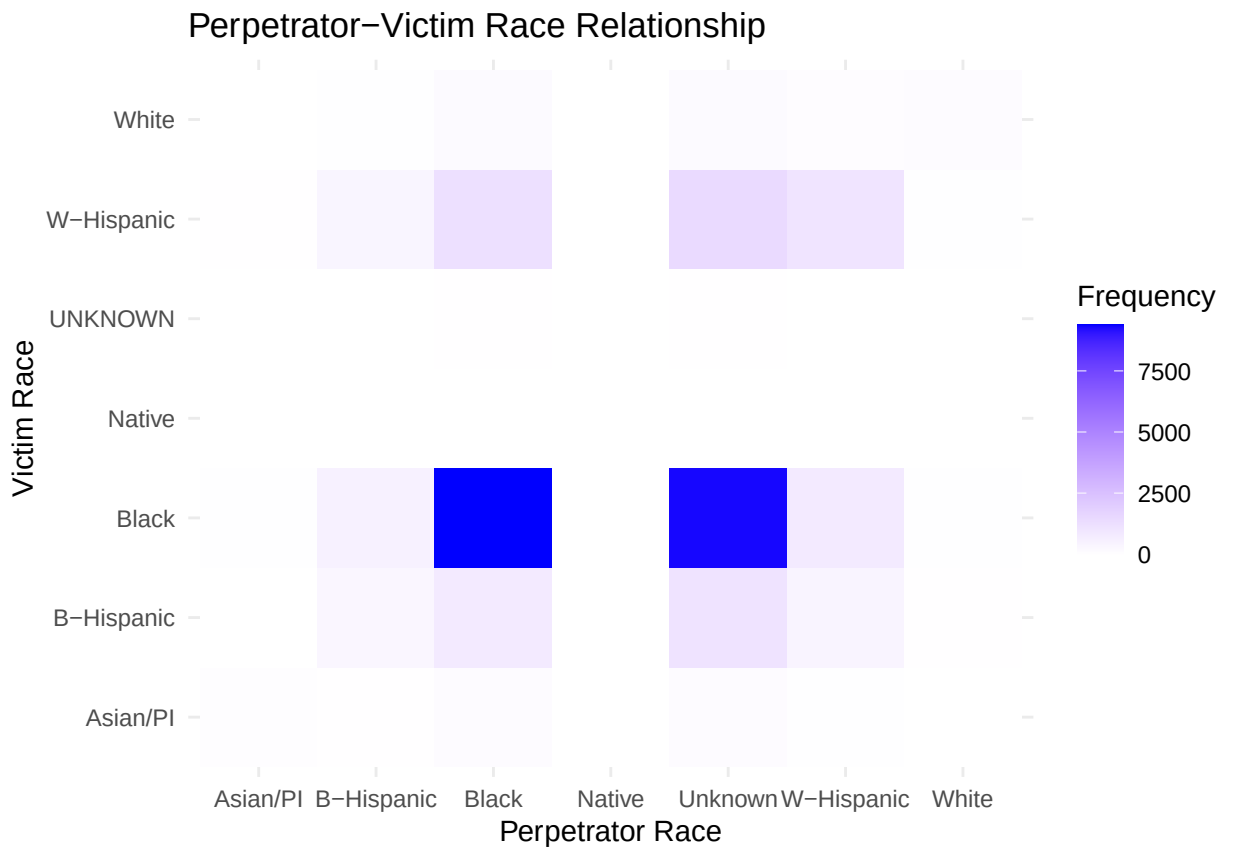
```
#clean/shorten long race names to see them better on the map
nypd_data_cleaned$PERP_RACE <- recode(nypd_data_cleaned$PERP_RACE,
    "AMERICAN INDIAN/ALASKAN NATIVE" = "Native",
    "ASIAN / PACIFIC ISLANDER" = "Asian/PI",
    "BLACK" = "Black",
    "BLACK HISPANIC" = "B-Hispanic",
    "WHITE" = "White",
    "WHITE HISPANIC" = "W-Hispanic",
    "Unknown" = "Unknown"
)

# Shorten the race names in VIC_RACE
nypd_data_cleaned$VIC_RACE <- recode(nypd_data_cleaned$VIC_RACE,
    "AMERICAN INDIAN/ALASKAN NATIVE" = "Native",
    "ASIAN / PACIFIC ISLANDER" = "Asian/PI",
    "BLACK" = "Black",
    "BLACK HISPANIC" = "B-Hispanic",
    "WHITE" = "White",
```

```
    "WHITE HISPANIC" = "W-Hispanic",
    "Unknown" = "Unknown"
)

#make a table for perpetrator-victim race
race_relationship <- table(nypd_data_cleaned$PERP_RACE,
                           nypd_data_cleaned$VIC_RACE)

# view thetable
race_relationship
```

Perpetrator-Victim Race Relationship

```
##
##              Asian/PI B-Hispanic Black Native UNKNOWN W-Hispanic White
##   Asian/PI         61         14    56      0       0         26    12
##   B-Hispanic       20        365   561      0       6        400    36
##   Black           164        836  9396      4      25       1250   205
##   Native            0          0     2      0       0          0     0
##   Unknown         140       1109  9300      6      26       1474   207
##   W-Hispanic       42        440   844      1      12       1060   103
##   White            13         23    42      0       1         54   165
```

```
# Visualize the relationship using a heatmap
ggplot(as.data.frame(race_relationship), aes(Var1, Var2, fill = Freq)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "blue") +
  labs(title = "Perpetrator-Victim Race Relationship",
       x = "Perpetrator Race",
       y = "Victim Race",
       fill = "Frequency") +
  theme_minimal()
```



Visual Observation/Inference:

- From the heat map, (excluding Unknown Race) we see a majority of perpetrators as well as majority of victims are both from the race identified as Black.

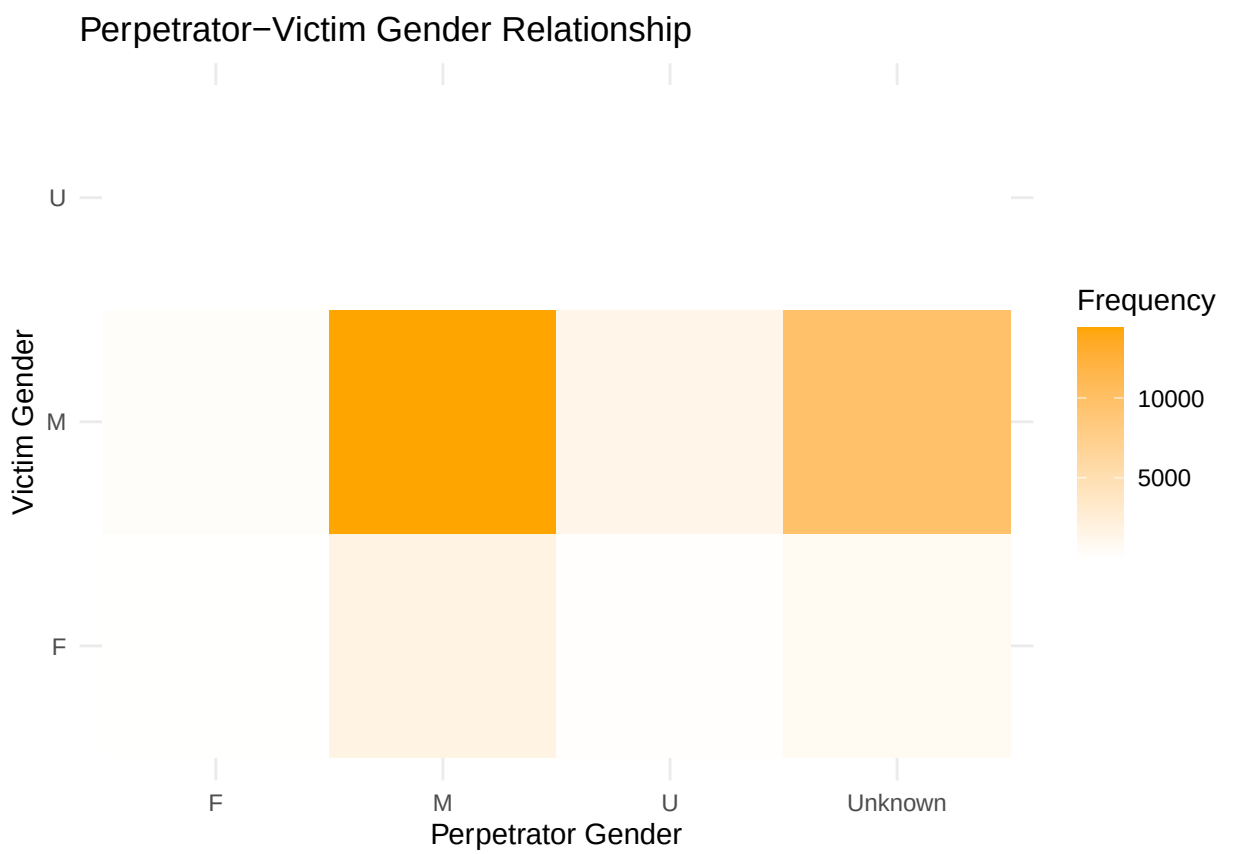
```
# make a table using Perpetrator-Victim identified sex/gender
gender_relationship <- table(nypd_data_cleaned$PERP_SEX, nypd_data_cleaned$VIC_SEX)

# View the table
gender_relationship
```

Perpetrator-Victim Gender Relationship

```
##
##                F     M     U
##   F           77   365     1
##   M         1752 14375     7
##   U          112  1386     1
##   Unknown    812  9610     3
```

```
# Visualize the relationship using a heatmap
ggplot(as.data.frame(gender_relationship), aes(Var1, Var2, fill = Freq)) +
  geom_tile() +
```

```
  scale_fill_gradient(low = "white", high = "orange") +
  labs(title = "Perpetrator-Victim Gender Relationship",
       x = "Perpetrator Gender",
       y = "Victim Gender",
       fill = "Frequency") +
  theme_minimal()
```



Visual Observation/Inference:

- From the heat map, (excluding Unknown sex/gender) we see a majority of perpetrators as well as majority of victims are both Males.

Also, because from the data set we don't have sufficient data identified as female perpetrators or female victims, my 4th question - i.e, analysis for finding the ratio across combination of genders will not be meaningful at all to do on this data set.

Analysis - Case 2:

Could there be any seasonal trends/months of the year when shootings are more frequent ?

```
#check date format
nypd_data_cleaned$OCCUR_DATE <- as.Date(nypd_data_cleaned$OCCUR_DATE, format = "%m/%d/%Y")

#extract month and add it as a column to the data set. Using %B to get full month name.
```

```r
nypd_data_cleaned$Month <- format(nypd_data_cleaned$OCCUR_DATE, "%B")

#create an additional 4-seasons column and add it to the data set
nypd_data_cleaned$Season <- cut(as.numeric(format(nypd_data_cleaned$OCCUR_DATE, "%m")),
                                breaks = c(0, 3, 6, 9, 12),
                                labels = c("Winter", "Spring", "Summer", "Fall"),
                                include.lowest = TRUE)

#check to see if we have new colums added
table(nypd_data_cleaned$Month)
```

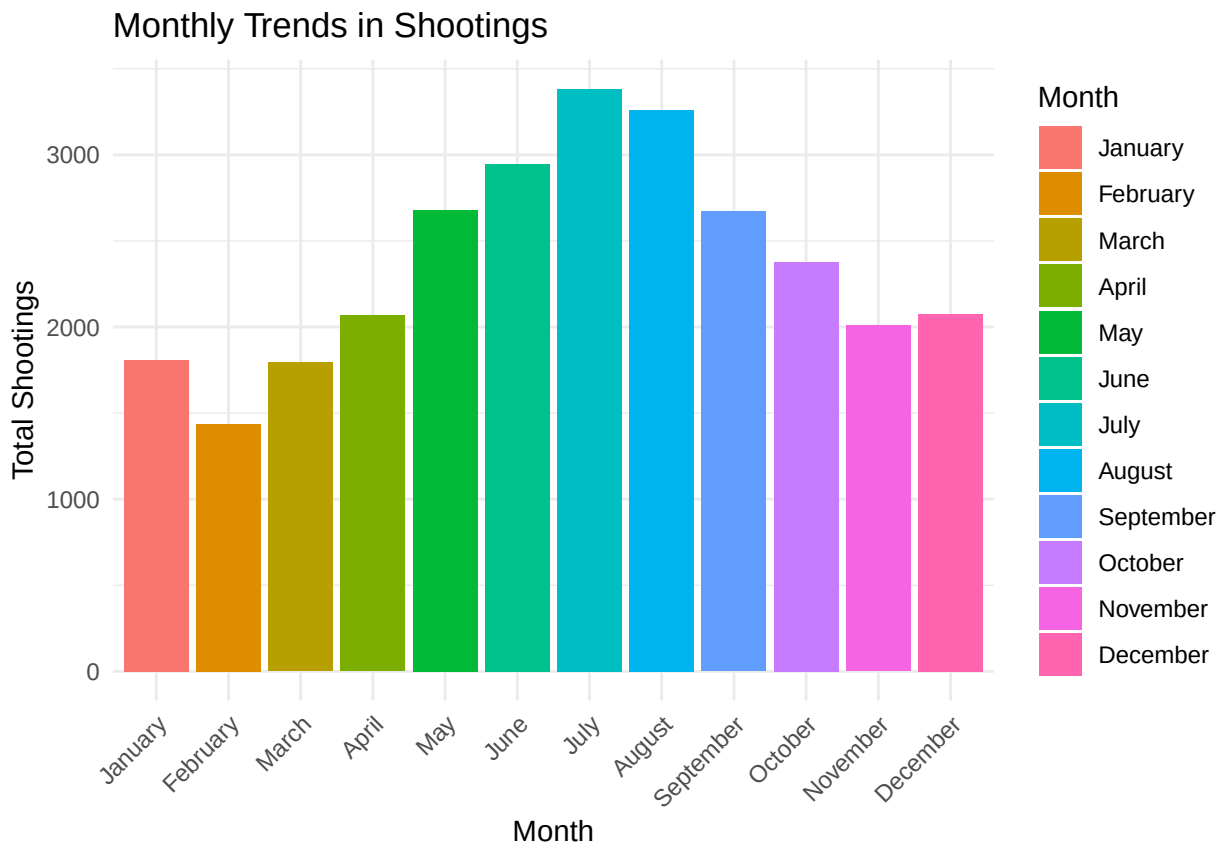
```
##
##     April    August  December  February   January      July      June     March
##      2066      3261      2075      1435      1808      3383      2947      1793
##       May  November   October September
##      2678      2009      2376      2670
```

```r
table(nypd_data_cleaned$Season)
```

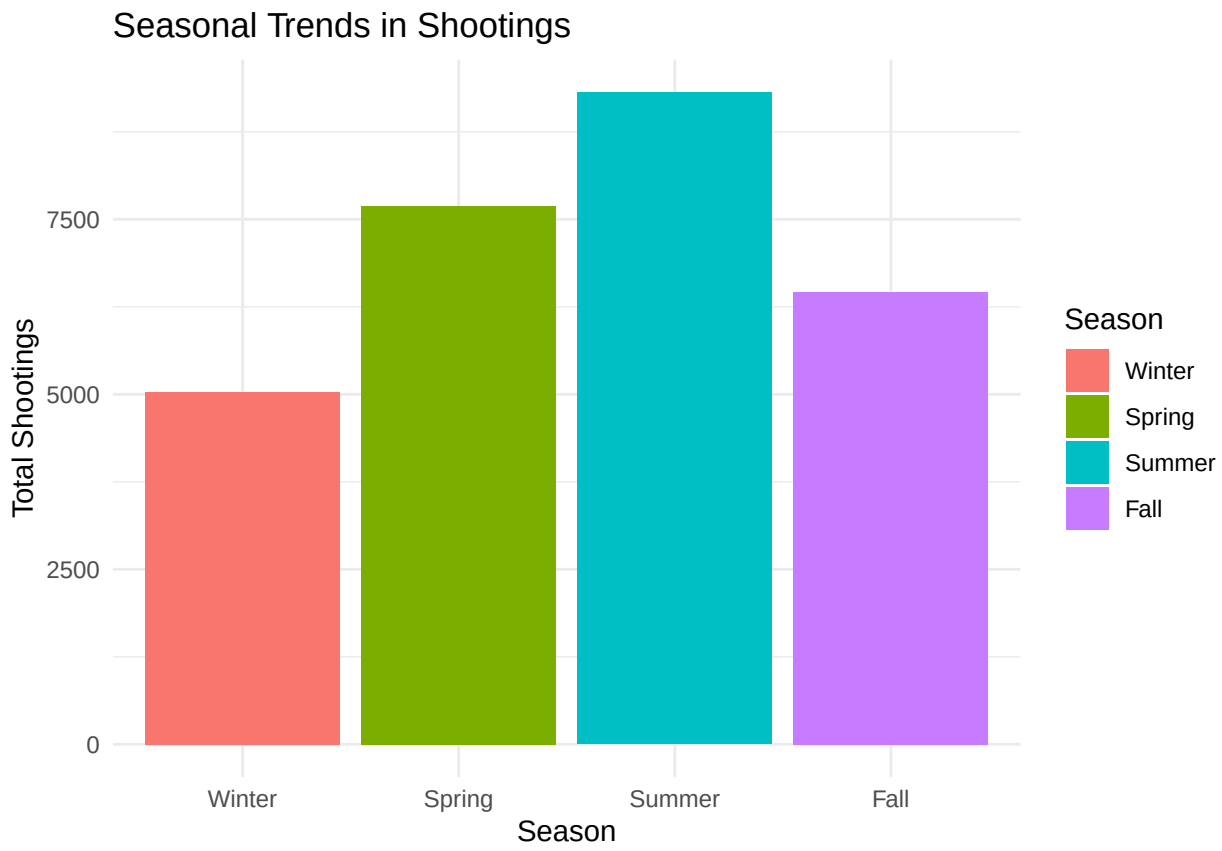
```
##
## Winter Spring Summer   Fall
##   5036   7691   9314   6460
```

```r
#group by month and count number of shootings in each month
nypd_data_cleaned$Month <- factor(nypd_data_cleaned$Month, levels = month.name)
monthly_trends <- nypd_data_cleaned %>%
  group_by(Month) %>%
  summarise(total_shootings = n()) %>%
  #arrange(match(month, month.name))  %>% #arrange by month
  ungroup()

#group by season and count number of shootings in each season
seasonal_trends <- nypd_data_cleaned %>%
  group_by(Season) %>%
  summarise(total_shootings = n()) %>%
  ungroup()
```

```r
#plot monthly trends
ggplot(monthly_trends, aes(x = Month, y = total_shootings, fill = Month)) +
  geom_bar(stat = "identity") +
  labs(title = "Monthly Trends in Shootings",
       x = "Month",
       y = "Total Shootings") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#plot seasonal trends
ggplot(seasonal_trends, aes(x = Season, y = total_shootings, fill = Season)) +
  geom_bar(stat = "identity") +
  labs(title = "Seasonal Trends in Shootings",
       x = "Season",
       y = "Total Shootings") +
  theme_minimal()
```



Visual Observation/Inference:

- From the plots, we see feburary to be the month with least number of shootings (and I dont think its because it has less days !) and summer months, July and August are when there are a lot of shootings, which is surprising !

Analysis - Case 3:

Which date of the year, every year since 2003, saw maximum shootings ?

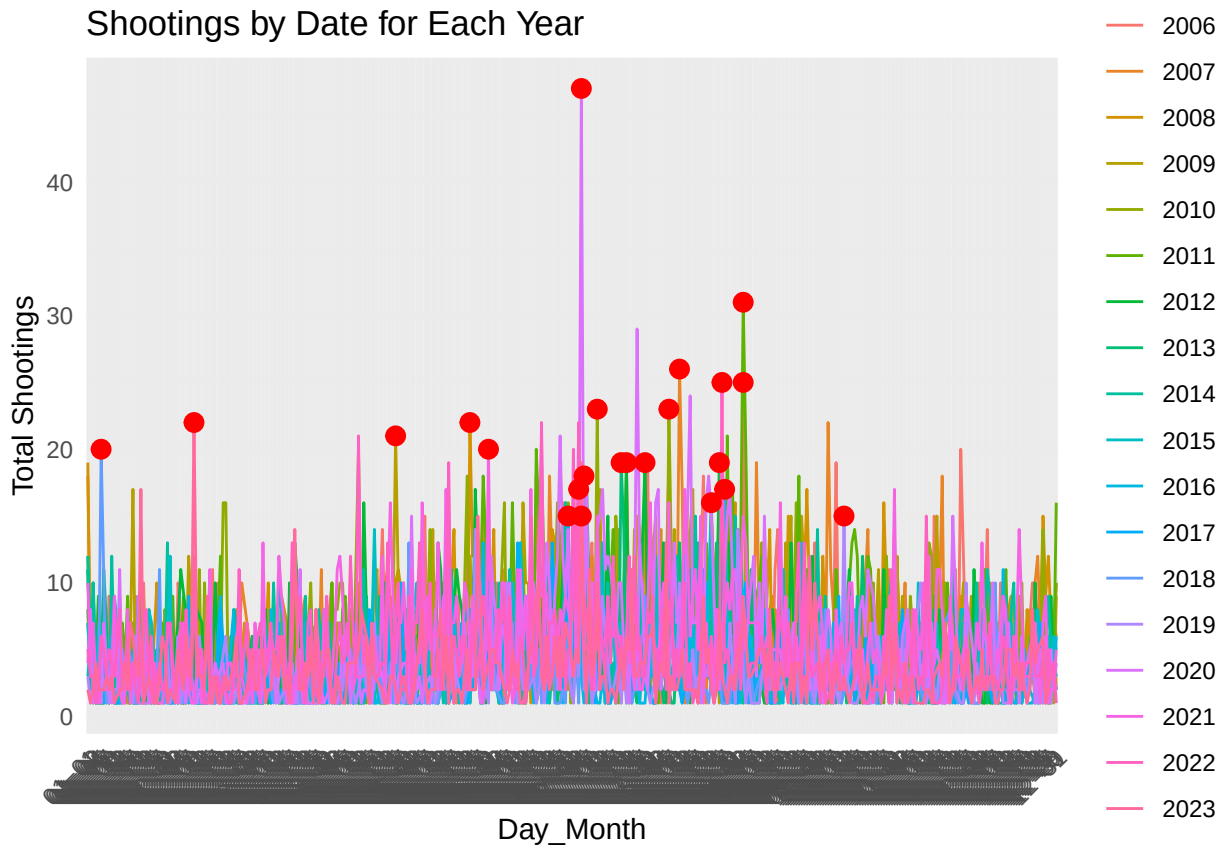
```
#extract the year and day-month (excluding the year) and add it as columns to use as group_by
nypd_data_cleaned$Year <- format(nypd_data_cleaned$OCCUR_DATE, "%Y")
nypd_data_cleaned$Day_Month <- format(nypd_data_cleaned$OCCUR_DATE, "%m-%d")

#group by year and day-month to get the number of shootings for each date
shootings_by_date <- nypd_data_cleaned %>%
  group_by(Year, Day_Month) %>%
  summarise(total_shootings = n()) %>%
  ungroup()
```

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

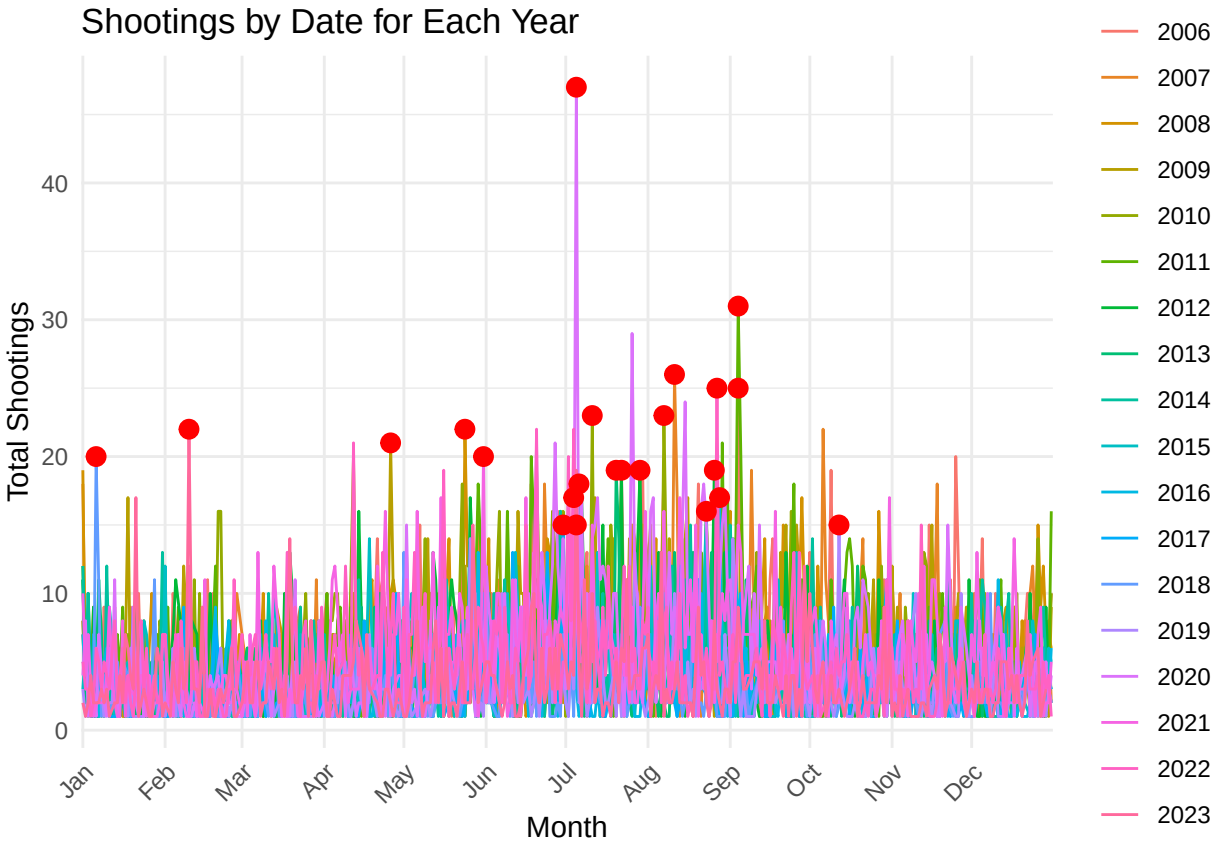
```r
#find the date with the maximum number of shootings for each year
max_shootings_by_year <- shootings_by_date %>%
  group_by(Year) %>%
  filter(total_shootings == max(total_shootings)) %>%
  ungroup()

#check what we have in max_shootings_by_year
max_shootings_by_year
```

```
## # A tibble: 23 x 3
##    Year  Day_Month total_shootings
##    <chr> <chr>               <int>
##  1 2006  09-04                  25
##  2 2007  08-11                  26
##  3 2008  05-24                  22
##  4 2009  04-26                  21
##  5 2010  07-11                  23
##  6 2010  08-07                  23
##  7 2011  09-04                  31
##  8 2012  07-22                  19
##  9 2012  07-29                  19
## 10 2012  08-26                  19
## # i 13 more rows
```

```r
#plot shootings over the year and highlight the peak date for each year
ggplot(shootings_by_date, aes(x = Day_Month, y = total_shootings, group = Year, color = Year)) +
  geom_line() +
  geom_point(data = max_shootings_by_year, aes(x = Day_Month, y = total_shootings),
             color = "red", size = 3) +
  labs(title = "Shootings by Date for Each Year",
       x = "Day_Month",
       y = "Total Shootings",
       color = "Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#create levels from 1st jan to 31st dec and factor it to add back to same Day_month column
nypd_data_cleaned$Day_Month <- factor(nypd_data_cleaned$Day_Month,
                                      levels = format(seq(as.Date("2000-01-01"),
                                                          as.Date("2000-12-31"),
                                                          by = "1 day"), "%m-%d"))

#plot it with discrete scale with mapping from %d%-%m% to actual names of the months as labels
ggplot(shootings_by_date, aes(x = Day_Month, y = total_shootings, group = Year, color = Year)) +
  geom_line() +
  geom_point(data = max_shootings_by_year, aes(x = Day_Month, y = total_shootings),
             color = "red", size = 3) +
  scale_x_discrete(breaks = c("01-01", "02-01", "03-01", "04-01", "05-01", "06-01",
                              "07-01", "08-01", "09-01", "10-01", "11-01", "12-01"),
                   labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug",
                              "Sep", "Oct", "Nov", "Dec")) +
  labs(title = "Shootings by Date for Each Year",
       x = "Month",
       y = "Total Shootings",
       color = "Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Visual Observation/Inference:

- This is a great and surprising insight ! I never thought to see, year over year (at least many years), the maximum shootings for that entire year have happened right around the 4th of July !!!

Analysis - Case 4:

What is the ratio/proportion of shootings across gender combinations of perpetrators to victims ? - female perpetrators and male victims. - female perpetrators and female victims. - male perpetrators and male victims. - male perpetrators and female victims.

Skipping this analysis:

- `From the data set we don't have sufficient data identified as female perpetrators` or female victims, hence, analysis for finding the ratio across combination of genders will not be meaningful at all to do on this data set.
- `Heat map in Analysis 1, part 3 confirms this as well.`

Modelling from the Data

Goal: The goal is to create a model applying logistic regression to predict whether the victim belongs to the 25-44 age group (a binary outcome) based on the attributes like perpetrator's age, race, gender and time of the year (month/season).

Once the model is created, then to check the model's performance, co-efficients and statistical significace of the model can be done.

```
#set a binary variable (as column) for whether the victim is in the age group 25-44
nypd_data_cleaned$VIC_AGE_25_44 <- ifelse(nypd_data_cleaned$VIC_AGE_GROUP == "25-44", 1, 0)

#get a model_data to work on and select predictors, remove rows with NA values
model_data <- nypd_data_cleaned %>%
  select(VIC_AGE_25_44, PERP_AGE_GROUP, PERP_RACE, PERP_SEX, Month, Season) %>%
  filter(!is.na(PERP_AGE_GROUP) & !is.na(PERP_RACE) & !is.na(PERP_SEX))

#apply logistic regression model using glm()
model <- glm(VIC_AGE_25_44 ~ PERP_AGE_GROUP + PERP_RACE + PERP_SEX + Month + Season,
             data = model_data, family = binomial)

#dump a summary to see the coefficients and significance levels
summary(model)
```

```
##
## Call:
## glm(formula = VIC_AGE_25_44 ~ PERP_AGE_GROUP + PERP_RACE + PERP_SEX +
##     Month + Season, family = binomial, data = model_data)
##
## Coefficients: (3 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.5316620  0.2010134  -2.645  0.00817 **
## PERP_AGE_GROUP18-24    0.5928721  0.0624014   9.501  < 2e-16 ***
## PERP_AGE_GROUP25-44    1.5025086  0.0626103  23.998  < 2e-16 ***
## PERP_AGE_GROUP45-64    1.2506929  0.0953930  13.111  < 2e-16 ***
## PERP_AGE_GROUP65+      0.8035103  0.2603000   3.087  0.00202 **
## PERP_AGE_GROUPUnknown  0.6621196  0.0765486   8.650  < 2e-16 ***
## PERP_RACEB-Hispanic   -0.3732378  0.1698128  -2.198  0.02795 *
## PERP_RACEBlack        -0.3208928  0.1614574  -1.987  0.04687 *
## PERP_RACENative       -0.3621831  1.4403137  -0.251  0.80146
## PERP_RACEUnknown      -0.2343000  0.1943898  -1.205  0.22808
## PERP_RACEW-Hispanic   -0.3632886  0.1656112  -2.194  0.02826 *
## PERP_RACEWhite        -0.3828016  0.2002479  -1.912  0.05592 .
## PERP_SEXM             -0.1923855  0.0998817  -1.926  0.05409 .
## PERP_SEXU             -0.3618266  0.1544946  -2.342  0.01918 *
## PERP_SEXUnknown        0.0649826  0.1473083   0.441  0.65912
## MonthFebruary         -0.1392321  0.0726404  -1.917  0.05527 .
## MonthMarch            -0.0614096  0.0683192  -0.899  0.36873
## MonthApril            -0.1255318  0.0660899  -1.899  0.05751 .
## MonthMay              -0.1146007  0.0624116  -1.836  0.06633 .
## MonthJune             -0.0570069  0.0611206  -0.933  0.35098
## MonthJuly             -0.0663039  0.0596640  -1.111  0.26644
## MonthAugust           -0.0477960  0.0600049  -0.797  0.42572
## MonthSeptember        -0.0674357  0.0623299  -1.082  0.27929
## MonthOctober          -0.0946224  0.0639421  -1.480  0.13892
## MonthNovember         -0.0514722  0.0663686  -0.776  0.43801
## MonthDecember         -0.0007044  0.0657759  -0.011  0.99146
## SeasonSpring                  NA         NA      NA       NA
## SeasonSummer                  NA         NA      NA       NA
## SeasonFall                    NA         NA      NA       NA
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39271  on 28500  degrees of freedom
## Residual deviance: 38172  on 28475  degrees of freedom
## AIC: 38224
##
## Number of Fisher Scoring iterations: 4
```

Final Summary

In this set of analysis of the historical NYPD shooting incident data, several important insights were uncovered.

- **Demographic Relationship Between Perpetrators and Victims**:
  - **Based on Age Group**: The majority of both perpetrators and victims was found to fall within the 25-44 age group.
  - **Based on Race**: Excluding unknown values, both perpetrators and victims are predominantly identified as Black.
  - **Based on Gender**: A significant majority of both perpetrators and victims are identified as male.
- **Seasonal Trends in NYPD Shootings**:
  - **Monthly Trends**: February consistently showed the fewest shootings, while the summer months, particularly July and August, saw the highest number of incidents. This spike in shootings during the summer months may be tied to various social or environmental factors.
- **Date With Maximum Shootings Each Year**:
  - **A surprising trend was uncovered**: Year over year, the highest number of shootings within a given year tends to occur around the 4th of July. This pattern suggests that Independence Day or the activities surrounding it may be linked to an increase in shooting incidents, making it a period of heightened concern for public safety.
- **Bias in analysis**:
  - While the analysis provides meaningful insights into patterns of shooting incidents, some of the results, especially demographic and seasonal conclusions may be affected by missing or inaccurate data.
  - Perpetrator demographics related analysis could be influenced by incomplete or missing information and as we see in the analysis, there were a lot of NA/("null") values which were considered as Unknown in our analysis and excluded in inference.
  - Thus, inference doesn't capture the full essence of analysis that would have taken into account of every shooting incident. There can as well be a selection bias at the source of data itself since we only considered data reported by NYPD.
  - Any unreported or mis-classified incidents are not included, which might cause towards under-representation, over-representation or mis-representation.

Summary of Logistic Regression Results

Using logistic regression model, I aimed to predict whether a shooting victim belongs to the 25-44 age group based on factors like the perpetrator's age, race, gender, and the time of year (month and season).

My Key findings:

- **Perpetrator Age Group**:
  - Age is the most significant predictor. Perpetrators in the 18-24, 25-44, and 45-64 age groups are strongly associated with victims also being in the 25-44 age group, with the 25-44 age group having the strongest effect.
- **Perpetrator Race**:
  - Perpetrators identified as **Black Hispanic**, **Black**, and **White Hispanic** have slightly lower odds of the victim being in the 25-44 age group, though the effect sizes are relatively small.
- **Perpetrator Gender**:
  - Gender has a marginally significant effect, with male perpetrators being associated with slightly lower odds of the victim being in the 25-44 age group.
- **Months**:
  - The month of the year shows weak effects. February has marginally lower odds of the victim being in the 25-44 age group, but other months are not significant.
- **Seasons**:
  - The season variable was excluded due to overlap with the month variable, suggesting that month alone captures most of the temporal variation.

Conclusion:

The model applied shows that the **perpetrator's age** is the strongest factor in predicting whether the victim is in the 25-44 age group, while race and gender play smaller roles. Month and season have limited influence on the outcome.

Bias in the model:

The logistic regression model is likely influenced by several biases, including selection bias, missing data bias, and omitted variable bias. These biases can lead to over or underestimation of the influence of specific predictors like age, race, and gender. Thus, biases would affect the the model's accuracy.