NYPD Shooting Incident Data Analysis

A

2024-06-22

NYPD Shooting Incident Data (Historic)

This dataset consists of a detailed summary of all shooting incidents in NYC from 2006 to the end of the previous calendar year. The data is manually extracted quarterly and reviewed by the Office of Management Analysis and Planning before being published on the NYPD website. Each record represents a shooting incident and includes information about the event, location, and time of occurrence, as well as suspect and victim demographics.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(ggplot2)
library(readr)
```

Read The NYPD Shootings Data

```
# NYPD Shootings URL
NYPD_url <- "https://data.cityofnewyork.us/resource/833y-fsy8.csv"

# Read the CSV file
NYPD <- read_csv(NYPD_url)
```

```
## Rows: 1000 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (11): boro, loc_of_occur_desc, loc_classfctn_desc, location_desc, perp_...
## dbl   (7): incident_key, precinct, jurisdiction_code, x_coord_cd, y_coord_cd...
```

```
## lgl   (1): statistical_murder_flag
## dttm  (1): occur_date
## 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 few rows of the dataset
head(NYPD)
```

```
## # A tibble: 6 x 21
##   incident_key occur_date          occur_time boro    loc_of_occur_desc precinct
##          <dbl> <dttm>              <time>     <chr>   <chr>                <dbl>
## 1    279683077 2023-12-29 00:00:00 03:43      QUEENS  INSIDE                 113
## 2    279709792 2023-12-29 00:00:00 21:22      BROOKL~ OUTSIDE                 75
## 3    279758069 2023-12-29 00:00:00 18:40      BRONX   OUTSIDE                 40
## 4    279609499 2023-12-27 00:00:00 19:47      BRONX   OUTSIDE                 42
## 5    279547332 2023-12-26 00:00:00 23:31      BRONX   OUTSIDE                 46
## 6    279547333 2023-12-26 00:00:00 23:43      QUEENS  OUTSIDE                106
## # 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>, geocoded_column <chr>
```

Inspect the Data

```
# Check the structure of the data
tail(NYPD)
```

```
## # A tibble: 6 x 21
##   incident_key occur_date          occur_time boro    loc_of_occur_desc precinct
##          <dbl> <dttm>              <time>     <chr>   <chr>                <dbl>
## 1    265354835 2023-03-19 00:00:00 23:48      BRONX   INSIDE                  47
## 2    265327272 2023-03-18 00:00:00 12:15      QUEENS  INSIDE                 102
## 3    265303128 2023-03-18 00:00:00 03:45      QUEENS  OUTSIDE                102
## 4    265339107 2023-03-18 00:00:00 15:45      MANHAT~ OUTSIDE                 32
## 5    265303127 2023-03-18 00:00:00 02:40      BROOKL~ INSIDE                  75
## 6    265304929 2023-03-18 00:00:00 05:25      QUEENS  OUTSIDE                102
## # 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>, geocoded_column <chr>
```

```
# Check the column names
colnames(NYPD)
```

```
##  [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] "geocoded_column"
```

Data Cleaning and Transformation

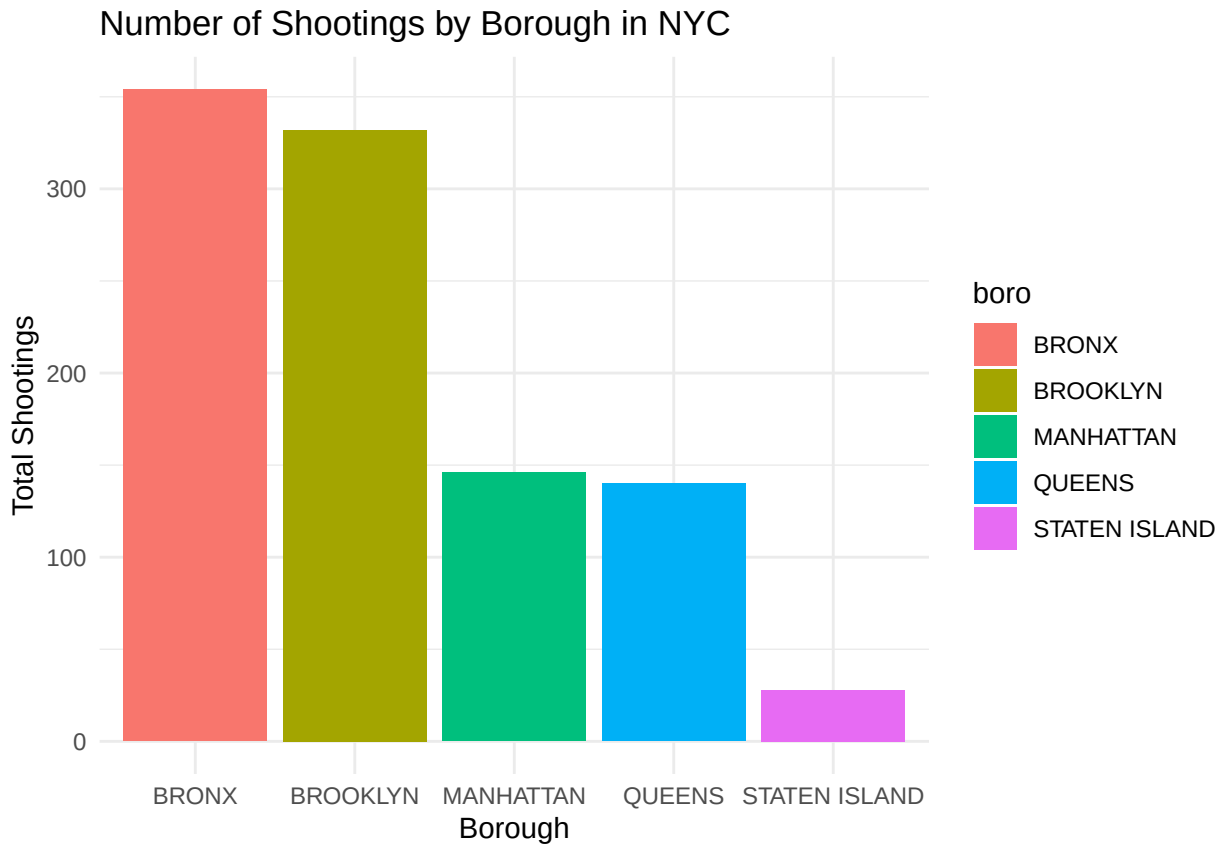
```
# Clean and tidy the data
NYPD_clean <- NYPD %>%
  select(c("occur_date", "occur_time", "boro", "precinct", "statistical_murder_flag", "vic_age_group",
  mutate(
    occur_date = ymd(occur_date),  # Adjusted date parsing function
    occur_time = hms(occur_time),
    statistical_murder_flag = as.logical(statistical_murder_flag),
    shootings = 1,
    year = year(occur_date),
    day_of_week = wday(occur_date, label = TRUE)
  )

# Summary of the cleaned data
summary(NYPD_clean)
```

```
##    occur_date           occur_time                          boro
##  Min.   :2023-03-18   Min.   :0S                        Length:1000
##  1st Qu.:2023-05-31   1st Qu.:5H 7M 30S                 Class :character
##  Median :2023-07-24   Median :16H 2M 0S                 Mode  :character
##  Mean   :2023-07-31   Mean   :13H 34M 28.6200000000026S
##  3rd Qu.:2023-10-01   3rd Qu.:20H 45M 30S
##  Max.   :2023-12-29   Max.   :23H 59M 0S
##
##     precinct      statistical_murder_flag vic_age_group       vic_sex
##  Min.   :  5.00   Mode :logical           Length:1000        Length:1000
##  1st Qu.: 43.00   FALSE:803               Class :character   Class :character
##  Median : 56.00   TRUE :197               Mode  :character   Mode  :character
##  Mean   : 62.92
##  3rd Qu.: 79.00
##  Max.   :123.00
##
##    vic_race           shootings      year      day_of_week
##  Length:1000        Min.   :1   Min.   :2023   Sun:181
##  Class :character   1st Qu.:1   1st Qu.:2023   Mon:152
##  Mode  :character   Median :1   Median :2023   Tue:124
##                     Mean   :1   Mean   :2023   Wed:112
##                     3rd Qu.:1   3rd Qu.:2023   Thu:115
##                     Max.   :1   Max.   :2023   Fri:134
##                                                Sat:182
```

Analysis and Visualizations

Number of Incidents by Borough

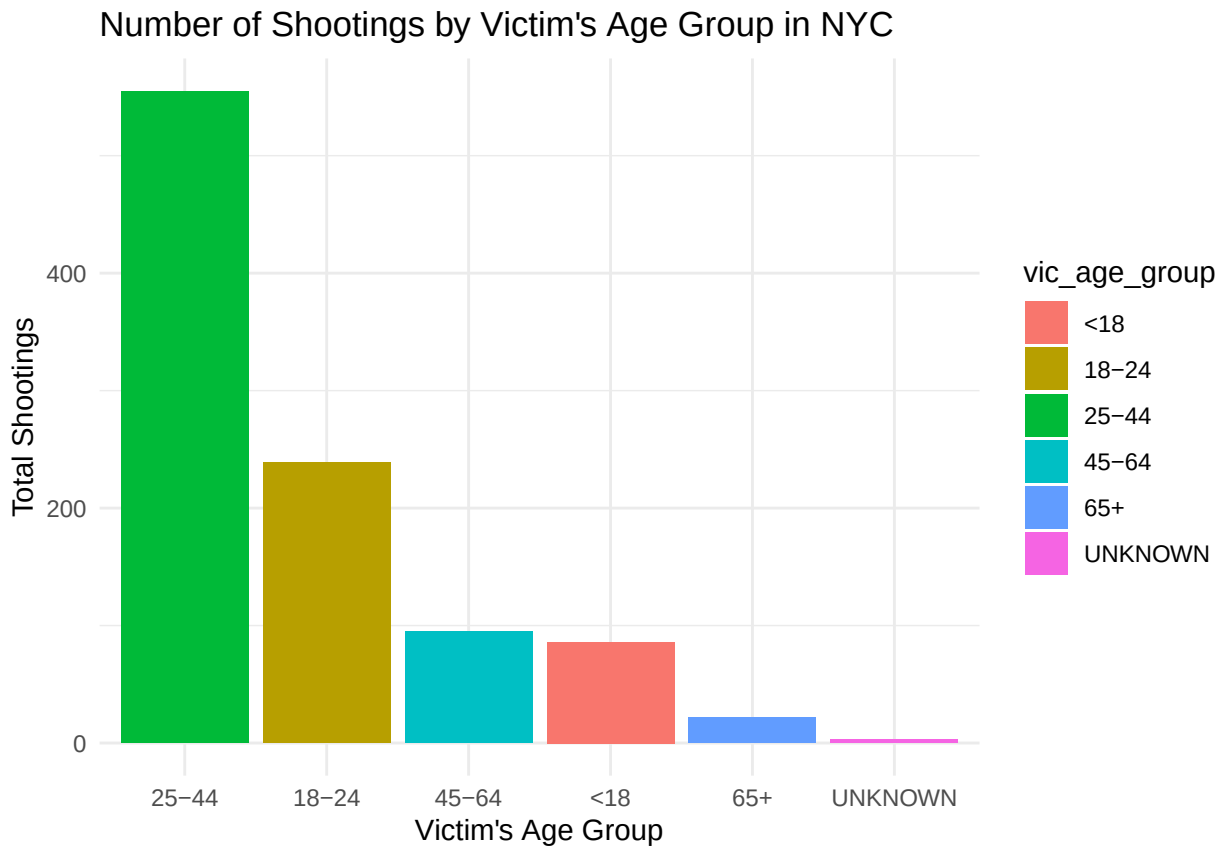
```
# Number of shootings by borough
shootings_by_boro <- NYPD_clean %>%
  group_by(boro) %>%
  summarize(total_shootings = sum(shootings))

# Plotting the number of shootings by borough
ggplot(shootings_by_boro, aes(x = reorder(boro, -total_shootings), y = total_shootings, fill = boro)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of Shootings by Borough in NYC", x = "Borough", y = "Total Shootings") +
  theme_minimal()
```



Number of Shootings by Victim's Age Group

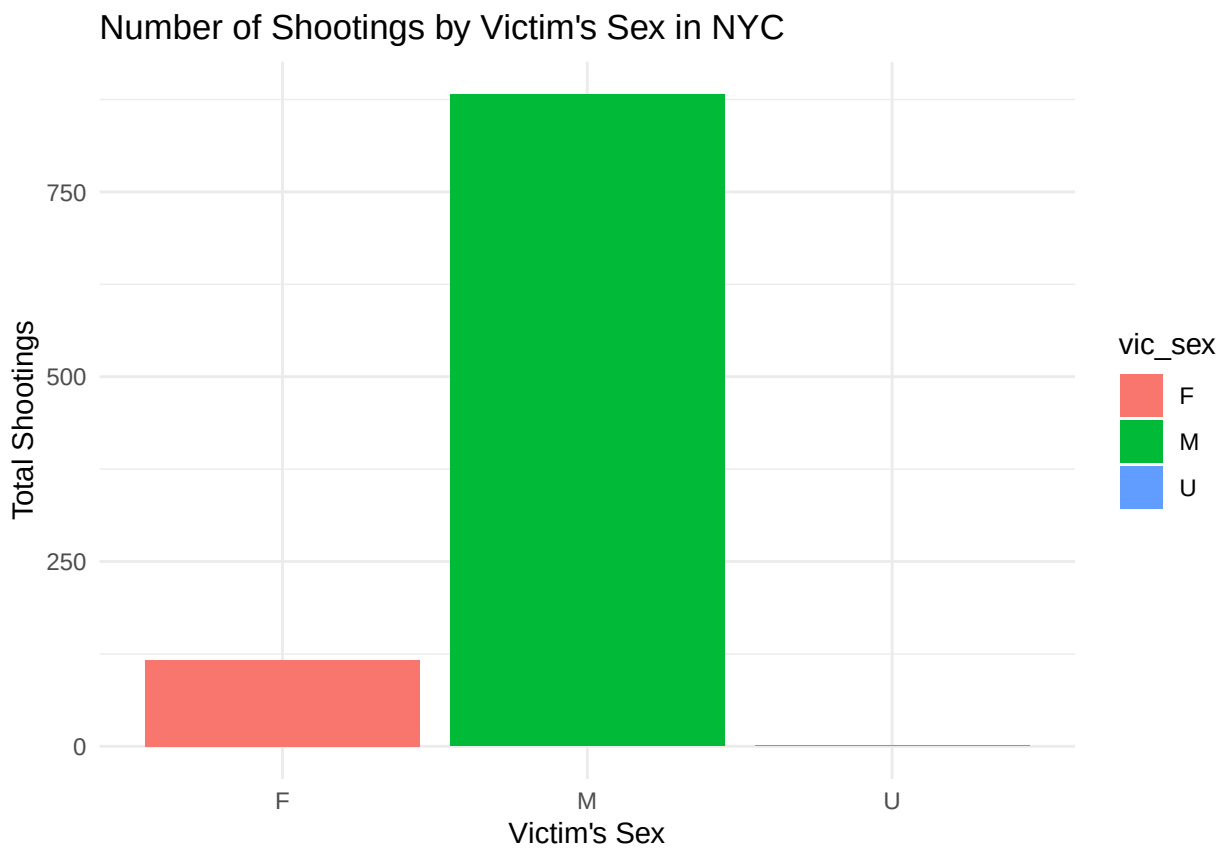
```
# Number of shootings by victim's age group
shootings_by_age_group <- NYPD_clean %>%
  group_by(vic_age_group) %>%
  summarize(total_shootings = sum(shootings))

# Plotting the number of shootings by victim's age group
```

```r
ggplot(shootings_by_age_group, aes(x = reorder(vic_age_group, -total_shootings), y = total_shootings, f
  geom_bar(stat = "identity") +
  labs(title = "Number of Shootings by Victim's Age Group in NYC", x = "Victim's Age Group", y = "Total
  theme_minimal()
```



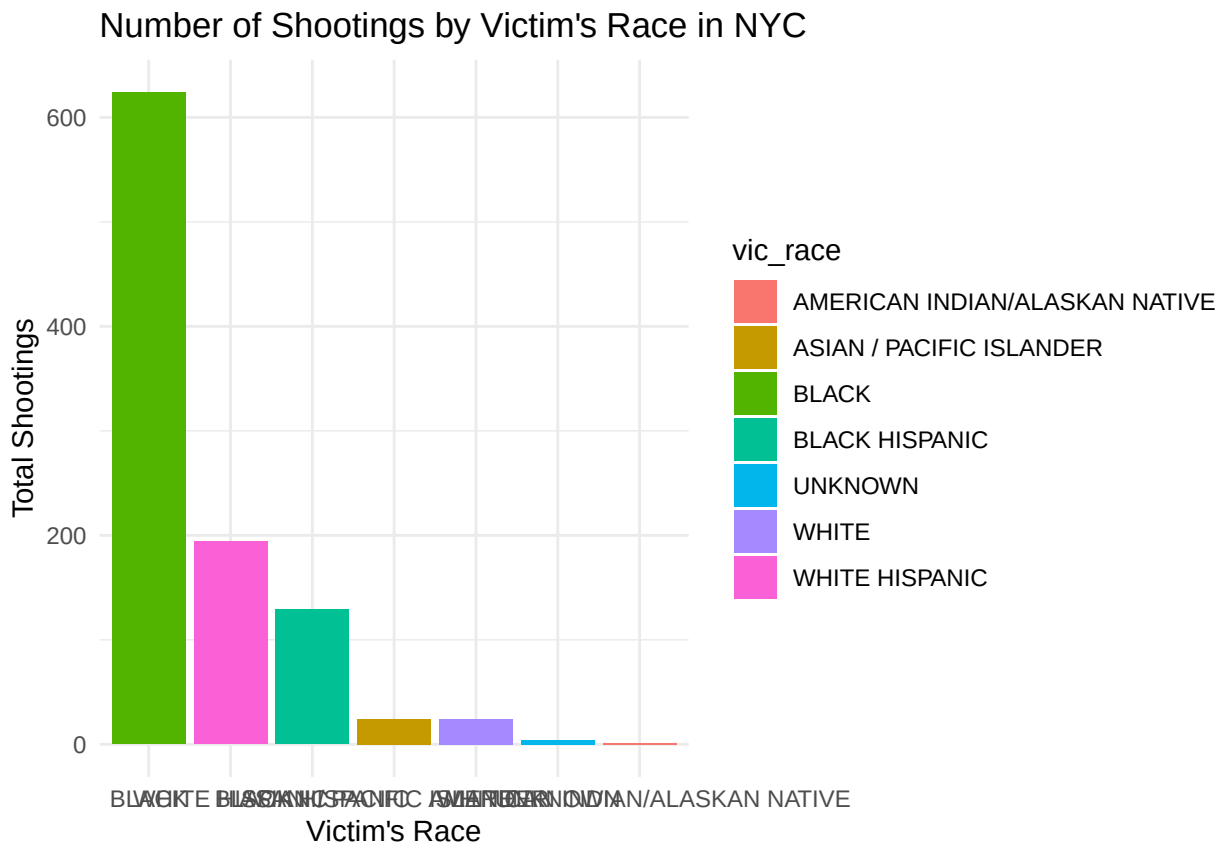
Number of Shootings by Victim's Sex

```r
# Number of shootings by victim's sex
shootings_by_sex <- NYPD_clean %>%
  group_by(vic_sex) %>%
  summarize(total_shootings = sum(shootings))

# Plotting the number of shootings by victim's sex
ggplot(shootings_by_sex, aes(x = vic_sex, y = total_shootings, fill = vic_sex)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of Shootings by Victim's Sex in NYC", x = "Victim's Sex", y = "Total Shootings")
  theme_minimal()
```



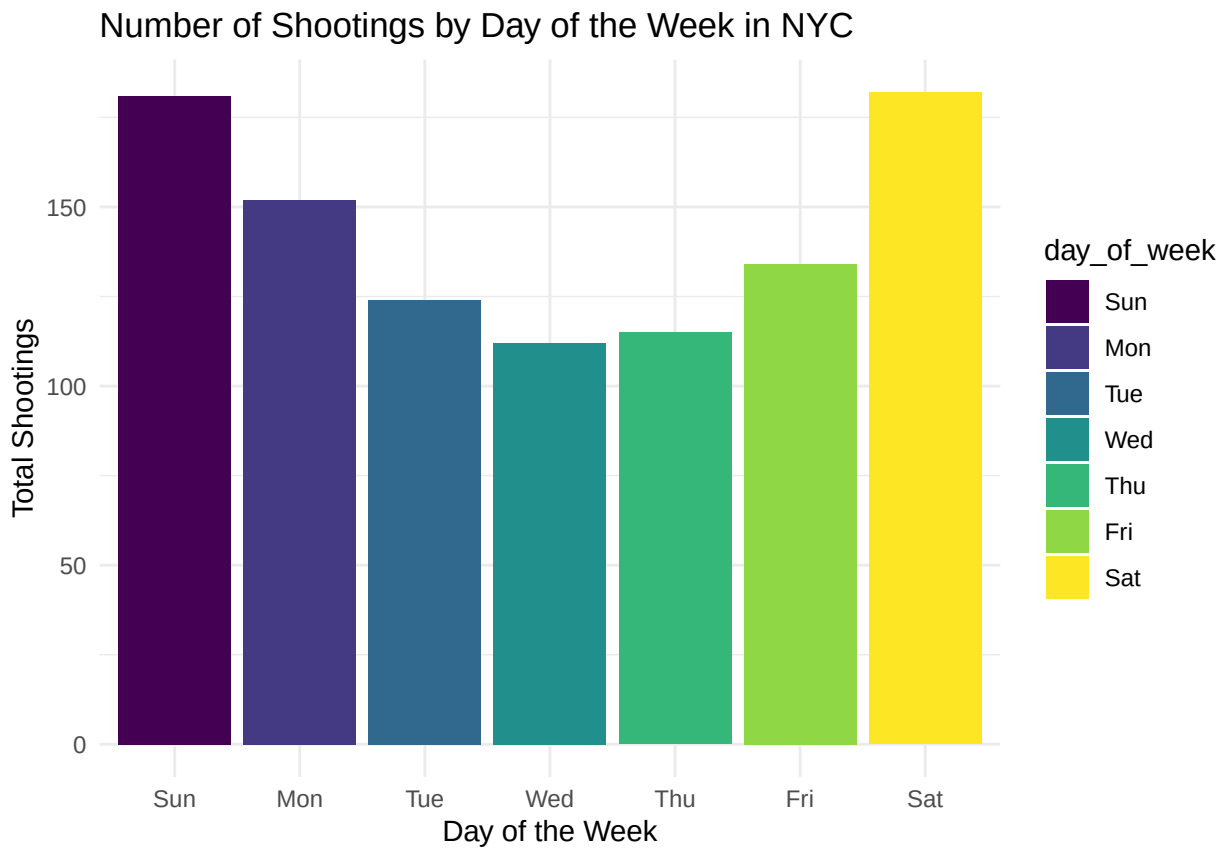
Number of Shootings by Victim's Race

```
# Number of shootings by victim's race
shootings_by_race <- NYPD_clean %>%
  group_by(vic_race) %>%
  summarize(total_shootings = sum(shootings))

# Plotting the number of shootings by victim's race
ggplot(shootings_by_race, aes(x = reorder(vic_race, -total_shootings), y = total_shootings, fill = vic_
  geom_bar(stat = "identity") +
  labs(title = "Number of Shootings by Victim's Race in NYC", x = "Victim's Race", y = "Total Shootings
  theme_minimal()
```



Number of Shootings by Day of the Week

```
# Number of shootings by day of the week
shootings_by_day_of_week <- NYPD_clean %>%
  group_by(day_of_week) %>%
  summarize(total_shootings = sum(shootings))

# Plotting the number of shootings by day of the week
ggplot(shootings_by_day_of_week, aes(x = day_of_week, y = total_shootings, fill = day_of_week)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of Shootings by Day of the Week in NYC", x = "Day of the Week", y = "Total Shoot
  theme_minimal()
```



Sources of Bias

Several potential sources of bias may affect the accuracy and interpretation of this analysis:

1. Data Quality:

The dataset relies on accurate reporting by the NYPD. Any underreporting or misclassification of incidents could skew the results.

2. Temporal Coverage:

If the data does not span multiple years or has gaps, it may not accurately represent long-term trends. ### 3. Demographic Data:

Incomplete or inaccurate demographic information can lead to incorrect conclusions about the affected populations.

4: Personal Bias :

As an analyst, my interpretations are influenced by my background and experiences. To mitigate personal bias, I have adhered strictly to the data, using objective methods for analysis and visualizations. By relying on reproducible code and transparent methodologies, I aimed to ensure that the findings are based on the data itself, rather than subjective viewpoints.