NYPD Shooting Incident Data Report

B. Knight

12/2/2024

This data report consists of every shooting incident in New York from 2006 to 2023.

This data is manually updated quarterly and reviewed by the Office of Management Analysis and Planning before being published on the NYPD website. Each entry represents a shooting incident, including details about the event, its location, and the time it occurred. Additionally, demographic information about the suspects and victims is included. The public can use this data to analyze trends in shooting and criminal activity. For more details, refer to the NYPD Shooting Incident Data (Historic) on CKAN.

Step 0: Import Library

```
# install.packages("tidyverse")
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(readr)
library(ggplot2)
```

Step 1: Load Data

```
# URL of the raw CSV file in Github
github_raw_url <- "https://raw.githubusercontent.com/BKnightHD/MS-Data-Science/refs/heads/main/Vital%20

# Read the CSV file
shooting_data <- read_csv(github_raw_url)
```

```r
# Display the first few rows
head(data)
```

```
##
## 1 function (..., list = character(), package = NULL, lib.loc = NULL,
## 2     verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
## 3 {
## 4     fileExt <- function(x) {
## 5         db <- grepl("\\\\.[^.]+\\\\.(gz|bz2|xz)$", x)
## 6         ans <- sub(".*\\\\.", "", x)
```

Step 2: Data Cleaning / Transform

Step 1.a: Check for Missing Values

```r
# Summarize missing values for each column
missing_summary <- colSums(is.na(shooting_data))
missing_summary
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO       LOC_OF_OCCUR_DESC                PRECINCT
##                       0                   25596                       0
##       JURISDICTION_CODE      LOC_CLASSFCTN_DESC           LOCATION_DESC
##                       2                   25596                   14977
## 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
##                      59                      59                      59
```

Step 1.b: Rename Columns for Consistency

```r
# Rename the "BORO" column to "BOROUGH" before making all headers consistent
shooting_data <- shooting_data %>%
  rename(BOROUGH = BORO) %>%
  rename_all(~str_replace_all(., " ", "_") %>% tolower())

# Display the new column names
colnames(shooting_data)
```

```
##  [1] "incident_key"            "occur_date"
##  [3] "occur_time"              "borough"
##  [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"
```

Step 1.c: Remove Duplicates

```
# Check for and remove duplicate rows
shooting_data <- shooting_data %>%
  distinct()

# Confirm the number of rows after removing duplicates
nrow(shooting_data)
```

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

Step 1.d: Handle Missing Data

```
# Replace missing values in selected columns with "Unknown" or similar placeholders
shooting_data <- shooting_data %>%
  mutate(across(c(perp_race, perp_sex, vic_race, vic_sex), ~replace_na(., "Unknown")))

# Verify changes
summary(shooting_data)
```

```
##   incident_key        occur_date         occur_time          borough
##  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
```

Step 1.e: Convert Dates to Proper Format

```r
# Convert date columns to Date format
shooting_data <- shooting_data %>%
  mutate(occur_date = as.Date(occur_date, format = "%m/%d/%Y"))

# Verify the date conversion
summary(shooting_data$occur_date)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2006-01-01" "2009-09-04" "2013-09-20" "2014-06-07" "2019-09-29" "2023-12-29"
```

Step 1.f: Filter and Select Relevant Columns

```r
# Keep only relevant columns for analysis
shooting_data <- shooting_data %>%
  select(occur_date, occur_time, borough, precinct, perp_race, vic_race, perp_age_group, vic_age_group,

# Display the structure of the cleaned dataset
glimpse(shooting_data)
```

```
## Rows: 28,562
## Columns: 9
## $ occur_date              <date> 2022-05-05, 2022-07-04, 2012-05-27, 2019-09-2~
## $ occur_time              <time> 00:10:00, 22:20:00, 19:35:00, 21:00:00, 21:00~
## $ borough                 <chr> "MANHATTAN", "BRONX", "QUEENS", "BRONX", "BROO~
## $ precinct                <dbl> 14, 48, 103, 42, 83, 23, 113, 77, 48, 49, 73, ~
## $ perp_race               <chr> "BLACK", "(null)", "Unknown", "UNKNOWN", "BLAC~
## $ vic_race                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ perp_age_group          <chr> "25-44", "(null)", NA, "25-44", "25-44", NA, N~
## $ vic_age_group           <chr> "25-44", "18-24", "18-24", "25-44", "25-44", "~
## $ statistical_murder_flag <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, ~
```

Step 1.g: Save the Cleaned Dataset

```r
# Save the cleaned data to a new CSV file
write_csv(shooting_data, "Cleaned_NYPD_Shooting_Incident_Data.csv")

# Confirm the file creation
list.files(pattern = "Cleaned_NYPD_Shooting_Incident_Data.csv")
```

```
## [1] "Cleaned_NYPD_Shooting_Incident_Data.csv"
```

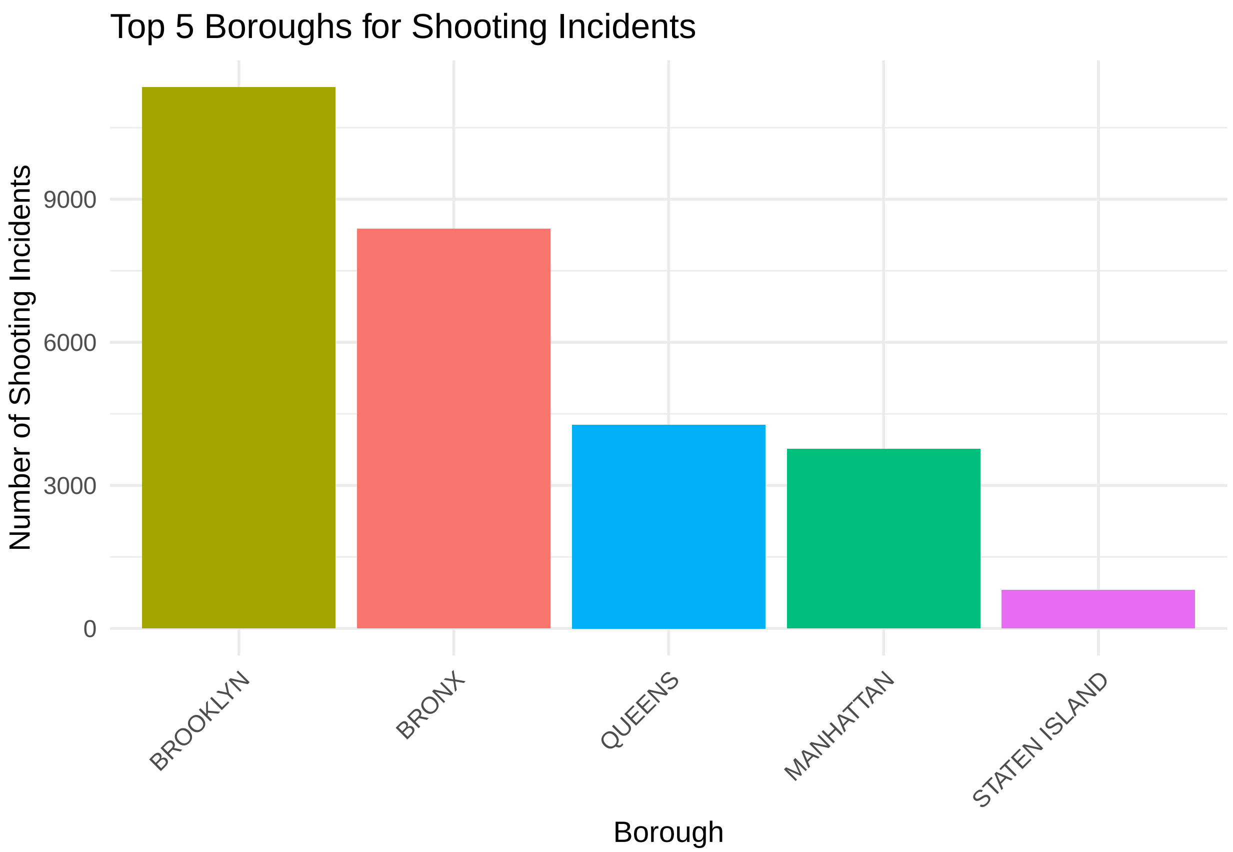
Step 2: Visualization

The next visualizations will help answer the following two questions:

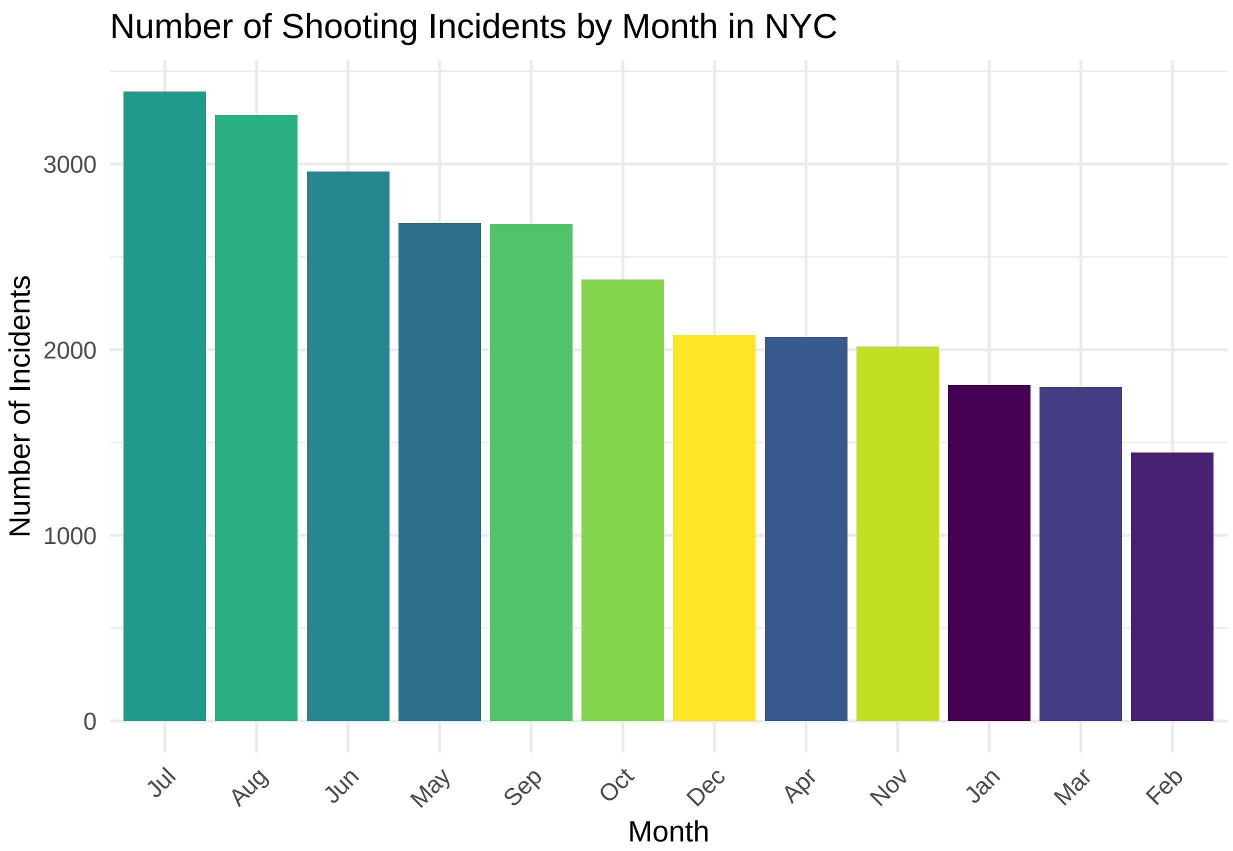
- Which borough has the highest number of shooting incidents?
- What is the most dangerous month to in New York in terms of shooting incidents?

Brooklyn looks to be the most dangerous while July seems to be the most dangerous month.

```r
# Count the number of shooting incidents per borough
borough_counts <- shooting_data %>%
  count(borough, sort = TRUE) %>%
  slice_max(n, n = 5) # Get the top 5 boroughs

# Create a bar plot for the top 5 boroughs
ggplot(borough_counts, aes(x = reorder(borough, -n), y = n, fill = borough)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(
    title = "Top 5 Boroughs for Shooting Incidents",
    x = "Borough",
    y = "Number of Shooting Incidents"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
# Extract the month from the occur_date column
shooting_data <- shooting_data %>%
  mutate(month = lubridate::month(occur_date, label = TRUE, abbr = TRUE))

# Count incidents by month
monthly_counts <- shooting_data %>%
  count(month, sort = TRUE)

# Create a bar plot for shooting incidents by month
ggplot(monthly_counts, aes(x = reorder(month, -n), y = n, fill = month)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(
    title = "Number of Shooting Incidents by Month in NYC",
    x = "Month",
    y = "Number of Incidents"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Step 3: Modeling shooting data for statistical analysis

Logistic Regression Model for Analyzing Shooting Data

In this section, we build a logistic regression model to explore the factors associated with shooting incidents being classified as murders. This type of analysis helps to identify patterns and relationships within the data, potentially aiding in better understanding and prevention efforts.

Model Overview:

- **Outcome Variable:** `murder_flag` (1 if the incident was classified as a murder, 0 otherwise).
- **Predictors:**
  - **Perpetrator's Race:** To examine demographic trends.
  - **Borough:** To explore geographical variations in incidents.
  - **Perpetrator's Age Group:** To investigate how age demographics correlate with murder classification.

The logistic regression model outputs the estimated relationship between each predictor and the likelihood of an incident being classified as a murder. These results help highlight significant factors that may warrant further investigation or policy consideration.

```r
# Prepare data for modeling
model_data <- shooting_data %>%
  filter(!is.na(statistical_murder_flag) & !is.na(perp_race) & !is.na(borough)) %>% # Remove rows with
  mutate(
    murder_flag = as.numeric(statistical_murder_flag == "TRUE"), # Convert murder flag to numeric
    perp_race = as.factor(perp_race),                           # Convert to factor
    borough = as.factor(borough)                                # Convert to factor
  )

# Fit logistic regression model
logistic_model <- glm(
  murder_flag ~ perp_race + borough + perp_age_group,
  data = model_data,
  family = binomial(link = "logit")
)

# Summarize the model
summary(logistic_model)
```

```
##
## Call:
## glm(formula = murder_flag ~ perp_race + borough + perp_age_group,
##     family = binomial(link = "logit"), data = model_data)
##
## Coefficients: (1 not defined because of singularities)
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -1.68132    0.08830 -19.042  < 2e-16
## perp_raceAMERICAN INDIAN/ALASKAN NATIVE -13.08521  229.36276  -0.057   0.9545
## perp_raceASIAN / PACIFIC ISLANDER        -1.28243    0.22726  -5.643 1.67e-08
## perp_raceBLACK                           -1.69296    0.15089 -11.220  < 2e-16
## perp_raceBLACK HISPANIC                  -1.80407    0.16471 -10.953  < 2e-16
## perp_raceUNKNOWN                         -1.46636    0.14086 -10.410  < 2e-16
## perp_raceWHITE                           -1.13550    0.19533  -5.813 6.13e-09
## perp_raceWHITE HISPANIC                  -1.57646    0.15748 -10.011  < 2e-16
## boroughBROOKLYN                          -0.10165    0.04639  -2.191   0.0284
## boroughMANHATTAN                         -0.14975    0.05944  -2.519   0.0118
## boroughQUEENS                            -0.13234    0.05878  -2.252   0.0243
## boroughSTATEN ISLAND                     -0.15218    0.10275  -1.481   0.1386
## perp_age_group<18                         1.91691    0.13962  13.729  < 2e-16
## perp_age_group1020                       -9.19178  324.74372  -0.028   0.9774
## perp_age_group1028                       -9.03960  324.74373  -0.028   0.9778
## perp_age_group18-24                       2.11238    0.12801  16.502  < 2e-16
## perp_age_group224                        -9.30828  324.74372  -0.029   0.9771
## perp_age_group25-44                       2.41804    0.12767  18.939  < 2e-16
## perp_age_group45-64                       2.77928    0.14808  18.769  < 2e-16
## perp_age_group65+                         2.83093    0.28747   9.848  < 2e-16
## perp_age_group940                        -9.20663  324.74372  -0.028   0.9774
## perp_age_groupUNKNOWN                          NA         NA      NA       NA
##
## (Intercept)                             ***
## perp_raceAMERICAN INDIAN/ALASKAN NATIVE
## perp_raceASIAN / PACIFIC ISLANDER       ***
## perp_raceBLACK                          ***
## perp_raceBLACK HISPANIC                 ***
## perp_raceUNKNOWN                        ***
## perp_raceWHITE                          ***
## perp_raceWHITE HISPANIC                 ***
## boroughBROOKLYN                         *
## boroughMANHATTAN                        *
## boroughQUEENS                           *
## boroughSTATEN ISLAND
## perp_age_group<18                       ***
## perp_age_group1020
## perp_age_group1028
## perp_age_group18-24                     ***
## perp_age_group224
## perp_age_group25-44                     ***
## perp_age_group45-64                     ***
## perp_age_group65+                       ***
## perp_age_group940
## perp_age_groupUNKNOWN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19168  on 19217  degrees of freedom
## Residual deviance: 18071  on 19197  degrees of freedom
##   (9344 observations deleted due to missingness)
## AIC: 18113
##
## Number of Fisher Scoring iterations: 11
```

Explanation:

1. **Data Preparation:**
   - Filter out rows with missing values in key columns (`statistical_murder_flag`, `perp_race`, and `borough`).
   - Convert `statistical_murder_flag` to a binary numeric variable (`murder_flag`) for modeling.
   - Convert categorical variables (`perp_race` and `borough`) to factors.
2. **Model Selection:**
   - Logistic regression predicts whether an incident is a murder based on predictors like `perp_race`, `borough`, and `perp_age_group`.
3. **Model Summary:**
   - The `summary(logistic_model)` function provides insights into the relationships between predictors and the likelihood of an incident being classified as a murder.

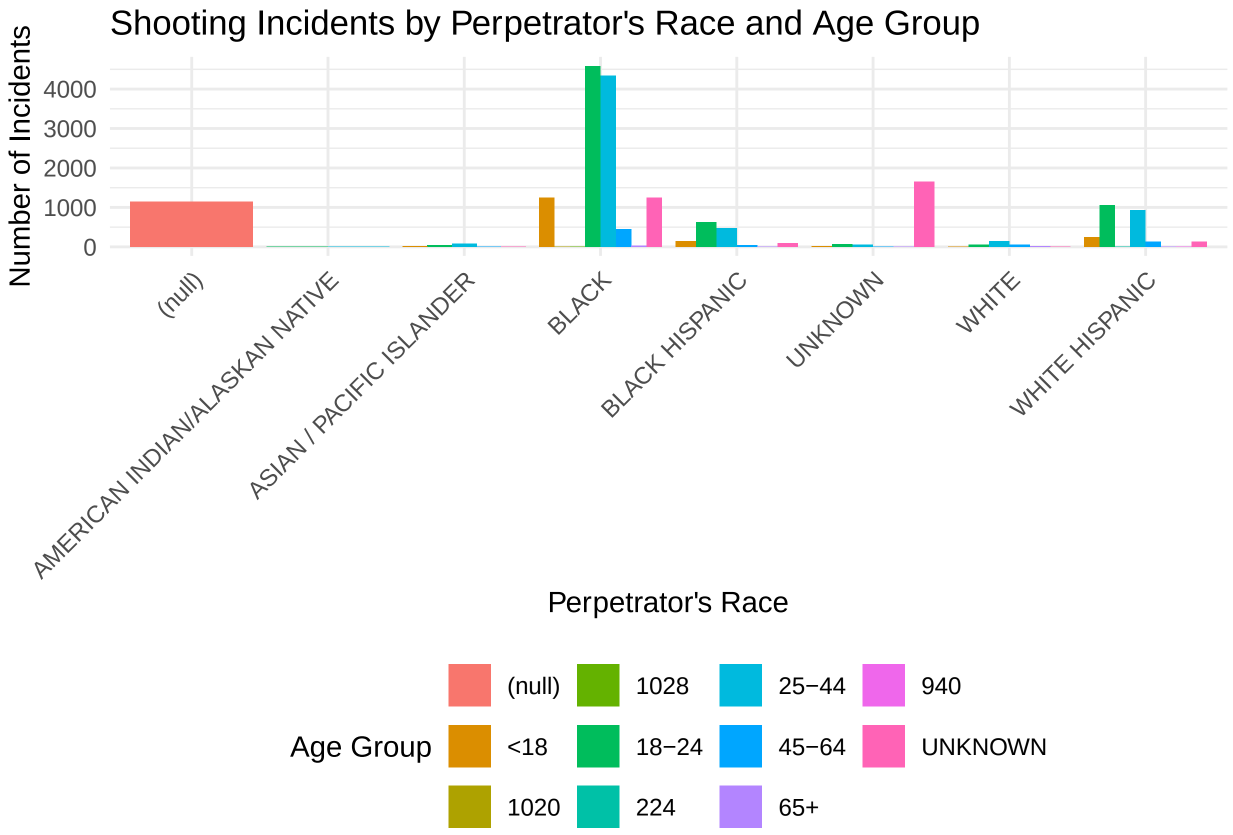
Step 4: Analysis of Perpetrators' Race and Age Group

Understanding the demographics of individuals involved in shooting incidents can provide insights into patterns and potential areas of intervention. The visualization below highlights the distribution of perpetrators' race and age group based on the reported data. Each bar represents the number of incidents attributed to a specific race, with colors differentiating the age groups.

This chart helps identify: - Which racial groups have higher reported incidents. - How age groups are distributed within each racial group.

By examining these trends, we can better understand demographic factors related to shooting incidents and design targeted prevention strategies.

```
# Count incidents by perpetrator race and age group
perp_stats <- shooting_data %>%
  filter(!is.na(perp_race) & !is.na(perp_age_group)) %>% # Filter out missing values
  count(perp_race, perp_age_group)

# Create a grouped bar chart
ggplot(perp_stats, aes(x = perp_race, y = n, fill = perp_age_group)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Shooting Incidents by Perpetrator's Race and Age Group",
    x = "Perpetrator's Race",
    y = "Number of Incidents",
    fill = "Age Group"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom"
  )
```



Step 5: Conclusion / Inisights

The analysis of the NYPD shooting incident data has provided valuable insights into patterns and trends related to gun violence in New York City. Here's a summary of key findings:

1. **Geographical Insights:**
   - The boroughs with the highest number of shooting incidents were identified, offering a clear picture of areas most affected by gun violence.
2. **Temporal Trends:**
   - Analysis of shooting incidents by month highlighted seasonal variations, with some months consistently experiencing higher levels of violence.
   - A frequency distribution by time of day revealed patterns in the timing of incidents, suggesting potential hotspots for intervention during specific hours.
3. **Demographic Patterns:**
   - Perpetrators' race and age group distributions provided demographic insights, showing which groups were more frequently involved in reported incidents.
   - This information can guide targeted outreach and community engagement initiatives.
4. **Predictive Modeling:**
   - A logistic regression model was constructed to identify factors influencing whether a shooting incident was classified as a murder. The results pointed to significant relationships between demographics, location, and the likelihood of an incident resulting in a fatality.

Final Thoughts: This analysis sheds light on critical aspects of gun violence in NYC, offering data-driven insights for policymakers, law enforcement, and community organizations. By focusing resources on high-risk areas, times, and demographics, there is an opportunity to design more effective prevention and intervention strategies. Future research could explore additional variables, such as socioeconomic factors or repeat offenders, to further enhance the understanding of this complex issue.