NYPD Shooting Incident Data Analysis

Cristian

Introduction

This report analyzes the NYPD Shooting Incident Data Historic. The dataset includes information about shooting incidents across various boroughs and over time. We will clean the data, explore it through visualizations, and build a logistic regression model to predict the likelihood of shooting incidents.

Data Import

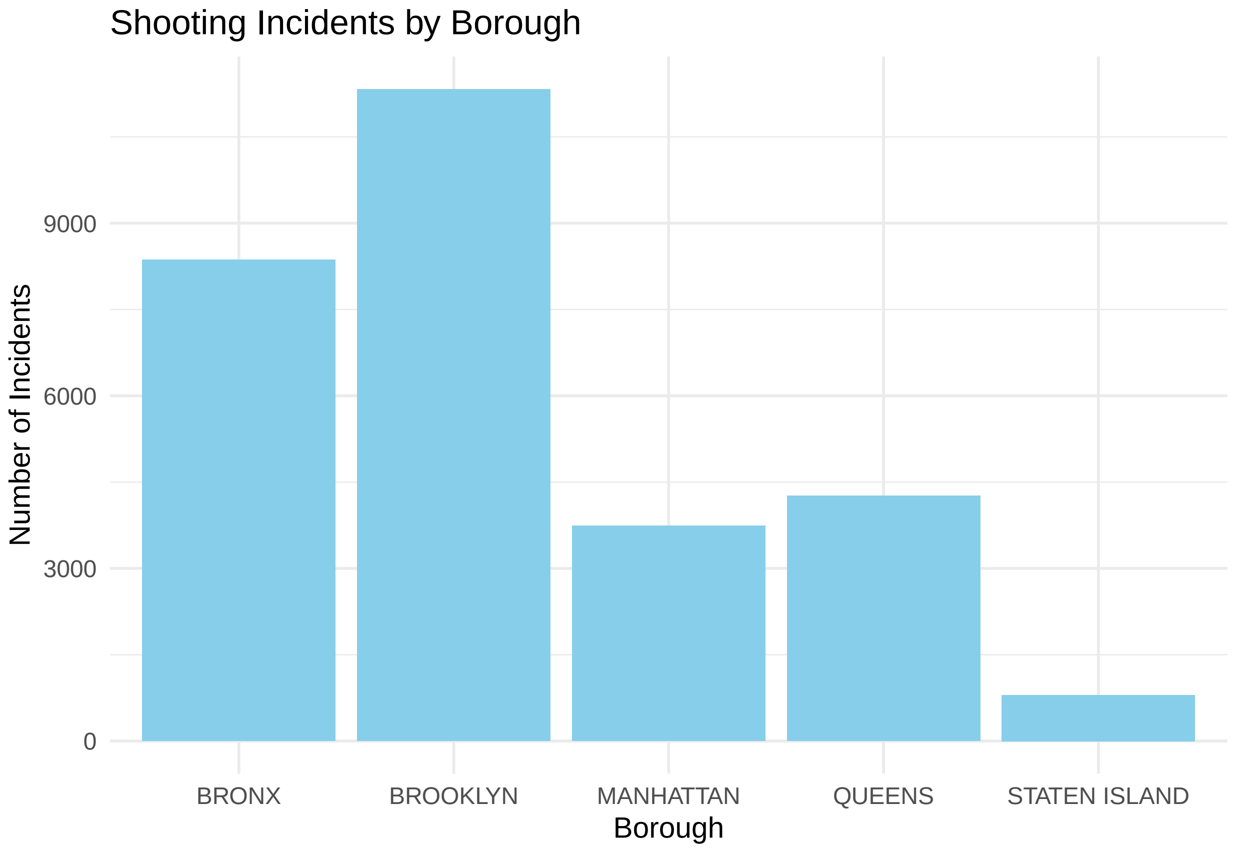
```
setwd("C:/Users/Cristian/Downloads")
nypd_data <- read.csv("NYPD_Shooting_Incident_Data__Historic_ (1).csv")
# Load the necessary libraries
library(tidyverse)
```

```
## Warning: package ’tidyverse’ was built under R version 4.4.3

## Warning: package ’ggplot2’ was built under R version 4.4.3

## Warning: package ’tibble’ was built under R version 4.4.3

## Warning: package ’tidyr’ was built under R version 4.4.3

## Warning: package ’readr’ was built under R version 4.4.3

## Warning: package ’purrr’ was built under R version 4.4.3

## Warning: package ’dplyr’ was built under R version 4.4.3

## Warning: package ’stringr’ was built under R version 4.4.3

## Warning: package ’forcats’ was built under R version 4.4.3

## Warning: package ’lubridate’ was built under R version 4.4.3

## -- 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 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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 error
```

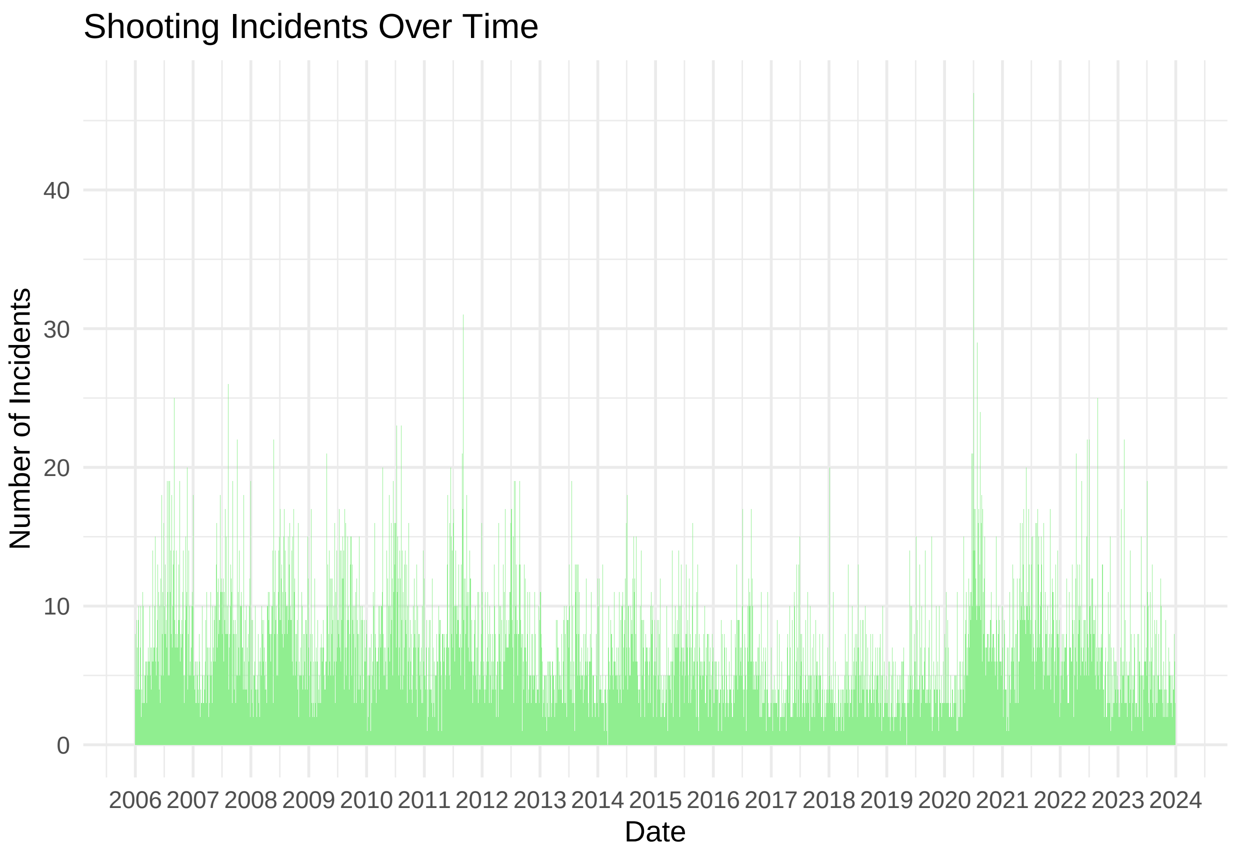
```r
library(lubridate)

# Clean the data
nypd_data_clean <- nypd_data %>%
  drop_na() %>%           # Remove rows with missing values
  distinct() %>%          # Remove duplicates
  mutate(
    OCCUR_DATE = mdy(OCCUR_DATE),   # Convert OCCUR_DATE to Date format
    BORO = factor(BORO)  # Ensure BORO is a factor for categorical analysis
  )

# Check the cleaned data
head(nypd_data_clean)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    231974218 2021-08-09   01:06:00    BRONX                         40
## 2    177934247 2018-04-07   19:48:00 BROOKLYN                         79
## 3    255028563 2022-12-02   22:57:00    BRONX           OUTSIDE       47
## 4     25384540 2006-11-19   01:50:00 BROOKLYN                         66
## 5     72616285 2010-05-09   01:58:00    BRONX                         46
## 6     85875439 2012-07-22   21:35:00    BRONX                         42
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC             LOCATION_DESC
## 1                 0
## 2                 0
## 3                 0             STREET            GROCERY/BODEGA
## 4                 0                                    PVT HOUSE
## 5                 0                      MULTI DWELL - APT BUILD
## 6                 2                    MULTI DWELL - PUBLIC HOUS
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX      PERP_RACE VIC_AGE_GROUP
## 1                   false                                                18-24
## 2                    true          25-44        M WHITE HISPANIC         25-44
## 3                   false         (null)   (null)         (null)         25-44
## 4                    true        UNKNOWN        U        UNKNOWN         18-24
## 5                    true          25-44        M          BLACK           <18
## 6                   false          18-24        M          BLACK         18-24
##   VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1       M    BLACK  1006343.0   234270.0 40.80967 -73.92019
## 2       M    BLACK  1000082.9   189064.7 40.68561 -73.94291
## 3       M    BLACK  1020691.0   257125.0 40.87235 -73.86823
## 4       M    BLACK   985107.3   173349.8 40.64249 -73.99691
## 5       F    BLACK  1009853.5   247502.6 40.84598 -73.90746
## 6       M    BLACK  1011046.7   239814.2 40.82488 -73.90318
##                                          Lon_Lat
## 1  POINT (-73.92019278899994 40.80967347200004)
## 2 POINT (-73.94291302299996 40.685609672000055)
## 3                  POINT (-73.868233 40.872349)
## 4 POINT (-73.99691224999998 40.642489932000046)
## 5  POINT (-73.90746098599993 40.84598358900007)
## 6  POINT (-73.90317908399999 40.82487781900005)
```

```r
# Bar plot of incidents by borough
ggplot(nypd_data_clean, aes(x = BORO)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Shooting Incidents by Borough",
       x = "Borough",
       y = "Number of Incidents") +
  theme_minimal()
```



```
# Time series plot of incidents over time
ggplot(nypd_data_clean, aes(x = OCCUR_DATE)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Shooting Incidents Over Time",
       x = "Date",
       y = "Number of Incidents") +
  theme_minimal() +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y")
```

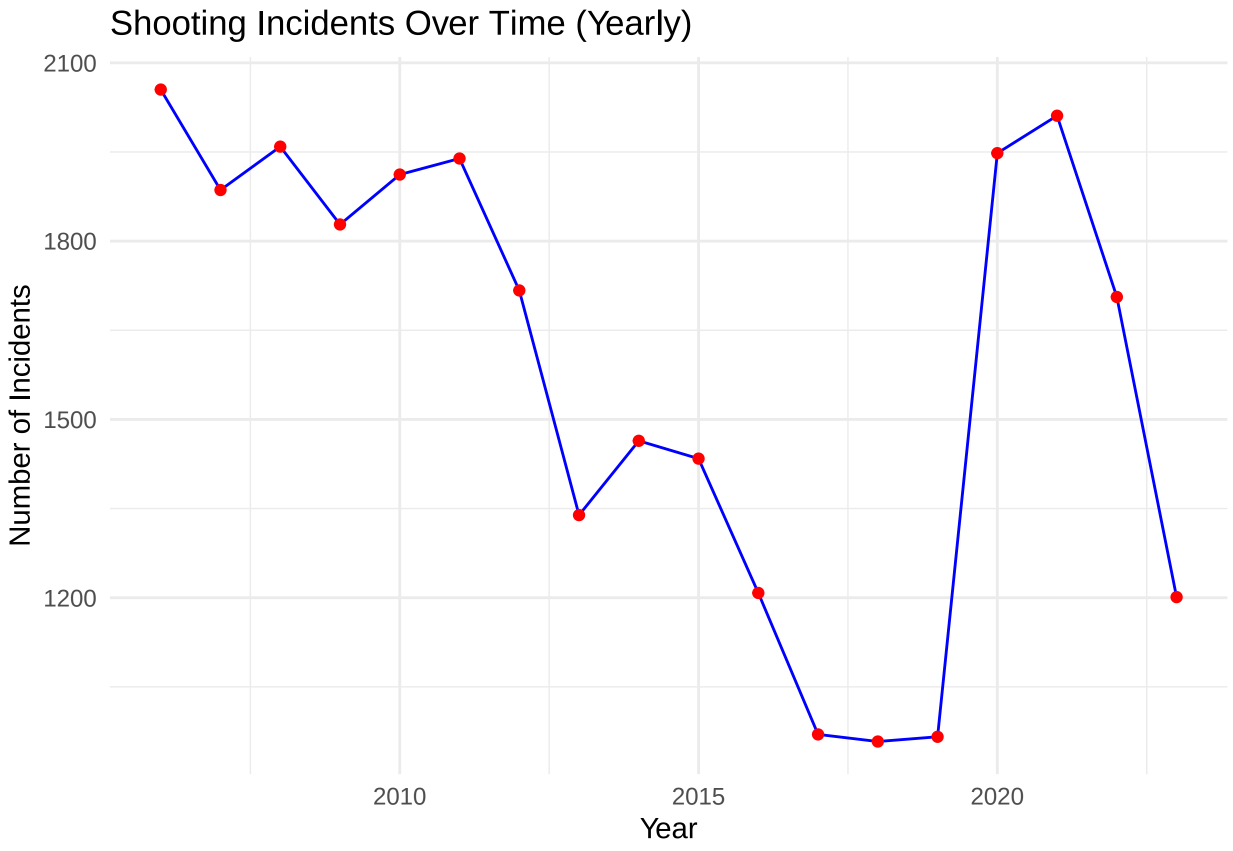


```r
# Fit a logistic regression model to predict the likelihood of a shooting incident
# For this example, we will predict if an incident was a shooting based on borough and time
nypd_data_clean$STATISTICAL_MURDER_FLAG <- as.factor(nypd_data_clean$STATISTICAL_MURDER_FLAG)

model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + OCCUR_DATE,
             data = nypd_data_clean,
             family = binomial())

summary(model)
```

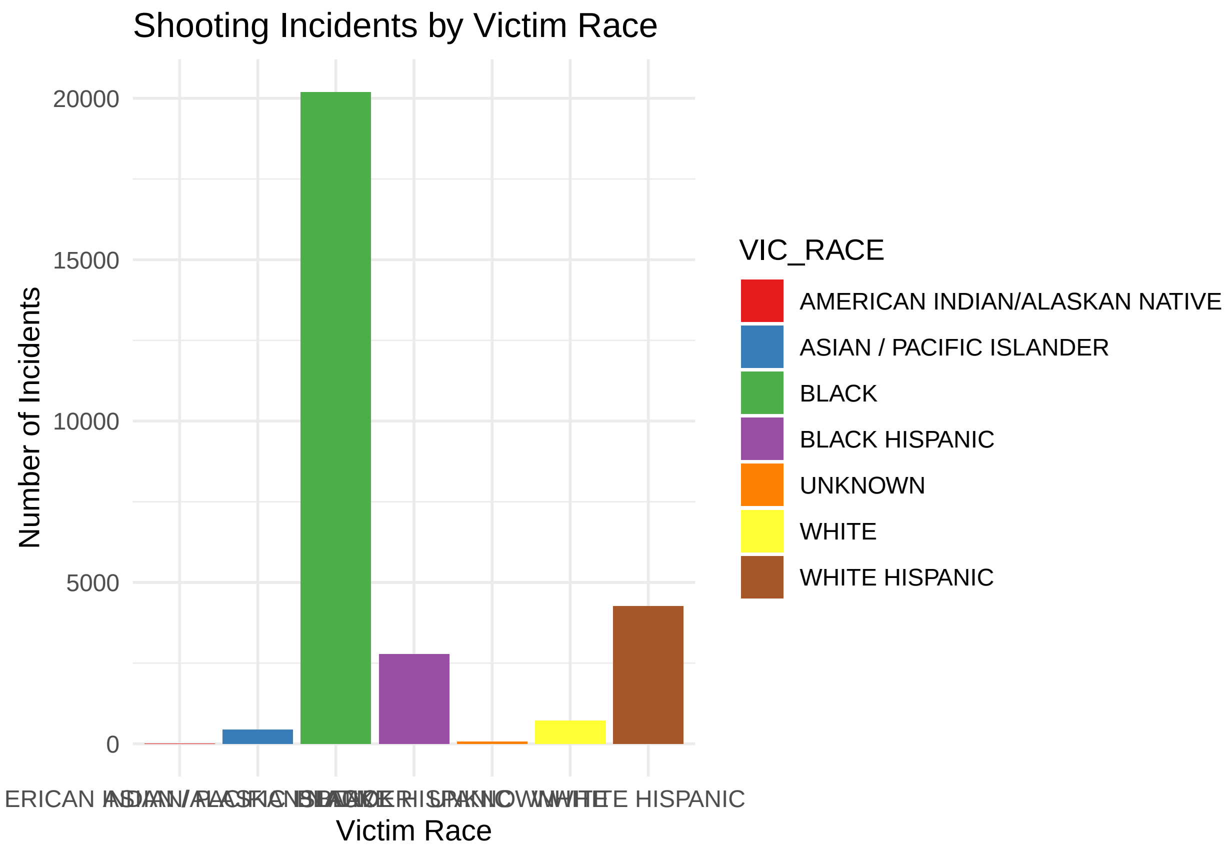
```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + OCCUR_DATE, family = binomial(),
##     data = nypd_data_clean)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.485e+00  1.278e-01 -11.624   <2e-16 ***
## BOROBROOKLYN       -2.905e-03  3.642e-02  -0.080   0.9364
## BOROMANHATTAN      -1.051e-01  5.074e-02  -2.071   0.0383 *
## BOROQUEENS          1.111e-02  4.737e-02   0.235   0.8145
## BOROSTATEN ISLAND   9.876e-02  9.091e-02   1.086   0.2773
## OCCUR_DATE          4.265e-06  7.640e-06   0.558   0.5767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28023  on 28500  degrees of freedom
## Residual deviance: 28015  on 28495  degrees of freedom
## AIC: 28027
##
## Number of Fisher Scoring iterations: 4
```

```
# Temporal bias: Check if incidents are more prevalent during specific years
yearly_count <- nypd_data_clean %>%
  mutate(year = year(OCCUR_DATE)) %>%
  group_by(year) %>%
  summarise(incident_count = n())

# Time series plot of incidents over years to check for temporal bias
ggplot(yearly_count, aes(x = year, y = incident_count)) +
  geom_line(group = 1, color = "blue") +
  geom_point(color = "red") +
  labs(title = "Shooting Incidents Over Time (Yearly)",
       x = "Year",
       y = "Number of Incidents") +
  theme_minimal()
```



```
# Analyze demographic bias by checking for distribution of victims' and perpetrators' races
victim_race_count <- nypd_data_clean %>%
  group_by(VIC_RACE) %>%
  summarise(incident_count = n())

# Plot distribution of victims' race
ggplot(victim_race_count, aes(x = VIC_RACE, y = incident_count, fill = VIC_RACE)) +
  geom_bar(stat = "identity") +
  labs(title = "Shooting Incidents by Victim Race",
       x = "Victim Race",
       y = "Number of Incidents") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")
```

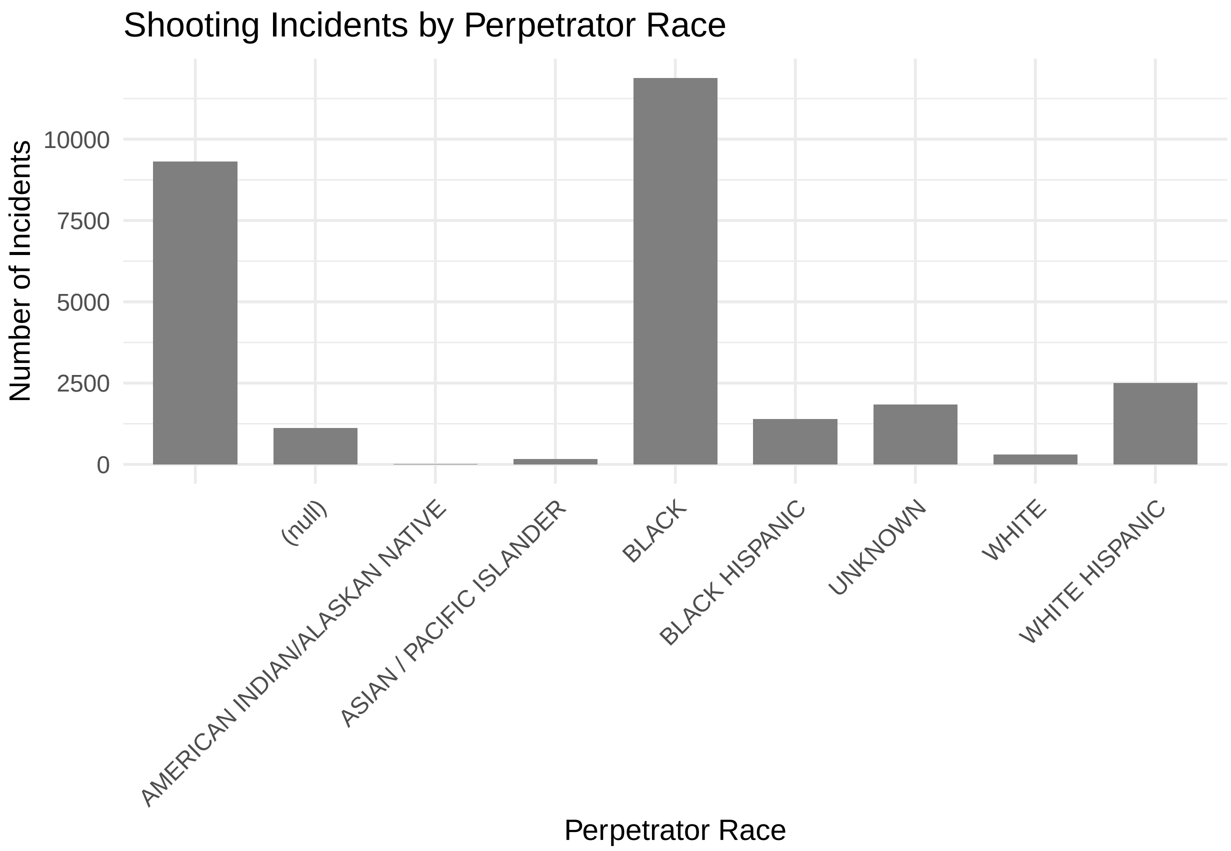


```
# Calculate incident counts for each perpetrator race
perp_race_count <- nypd_data_clean %>%
  group_by(PERP_RACE) %>%
  summarise(incident_count = n())

# Check the data to confirm it's been created correctly
head(perp_race_count)
```

```
## # A tibble: 6 x 2
##   PERP_RACE                        incident_count
##   <chr>                                     <int>
## 1 ""                                         9310
## 2 "(null)"                                   1115
## 3 "AMERICAN INDIAN/ALASKAN NATIVE"              2
## 4 "ASIAN / PACIFIC ISLANDER"                  169
## 5 "BLACK"                                   11880
## 6 "BLACK HISPANIC"                           1388
```

```r
# Plot distribution of perpetrators' race with custom colors for "White Hispanic"
ggplot(perp_race_count, aes(x = PERP_RACE, y = incident_count, fill = PERP_RACE)) +
  geom_bar(stat = "identity", width = 0.7) +  # Adjust width to space out the bars
  labs(title = "Shooting Incidents by Perpetrator Race",
       x = "Perpetrator Race",
       y = "Number of Incidents") +
  theme_minimal() +
  scale_fill_manual(values = c("White" = "#1f77b4",
                               "Black" = "#ff7f0e",
                               "Hispanic" = "#2ca02c",
                               "White Hispanic" = "#d62728",
                               "Asian" = "#9467bd",
                               "Other" = "#8c564b")) +  # Custom color for "White Hispanic"
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```

```
## Warning: No shared levels found between ‘names(values)‘ of the manual scale and the
## data’s fill values.

## Warning: No shared levels found between ‘names(values)‘ of the manual scale and the
## data’s fill values.
```


Shooting Incidents by Perpetrator Race
Number of Incidents
(null)
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN / PACIFIC ISLANDER
BLACK
BLACK HISPANIC
UNKNOWN
WHITE
WHITE HISPANIC
Perpetrator Race