NYPD Shooting Incident

2025-06-30

R Packages Utilized in the Analysis

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Downloading the NYPD Shooting Incident Dataset

First, you'll need to assign the URL and read in the CSV file. I located the dataset at the following link: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic/resource/c564b578-fd8a-4005-8365-34150d306cc4.

However, the direct URL to the CSV file is: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

The code below demonstrates how to read the CSV file into R. Please note that I use the `head()` function to display the first few rows of the uncleaned data set.

```
# Reads the CSV directly from the URL
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read.csv(url)

# Head function checks the first few rows
head(nypd_data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    231974218 08/09/2021   01:06:00    BRONX                          40
## 2    177934247 04/07/2018   19:48:00 BROOKLYN                          79
## 3    255028563 12/02/2022   22:57:00    BRONX           OUTSIDE       47
```

```
## 4     25384540 11/19/2006   01:50:00 BROOKLYN                          66
## 5     72616285 05/09/2010   01:58:00    BRONX                          46
## 6     85875439 07/22/2012   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
## 1       M    BLACK                 1006343                  234270 40.80967
## 2       M    BLACK 1000082.937500000000000 189064.671875000000000 40.68561
## 3       M    BLACK                 1020691                  257125 40.87235
## 4       M    BLACK  985107.312500000000000 173349.796875000000000 40.64249
## 5       F    BLACK 1009853.500000000000000 247502.562500000000000 40.84598
## 6       M    BLACK 1011046.687500000000000 239814.234375000000000 40.82488
##   Longitude                                      Lon_Lat
## 1 -73.92019  POINT (-73.92019278899994 40.80967347200004)
## 2 -73.94291 POINT (-73.94291302299996 40.685609672000055)
## 3 -73.86823                  POINT (-73.868233 40.872349)
## 4 -73.99691 POINT (-73.99691224999998 40.642489932000046)
## 5 -73.90746  POINT (-73.90746098599993 40.84598358900007)
## 6 -73.90318  POINT (-73.90317908399999 40.82487781900005)
```

Selecting Relevant Columns for Analysis

We do not wish to work with all of the columns within the data set for analysis. Therefore, we will be dropping the following columns: INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat

```
selected_data <- nypd_data %>%
  select(-INCIDENT_KEY, -PRECINCT, -JURISDICTION_CODE,
         -X_COORD_CD, -Y_COORD_CD, -Latitude, -Longitude, -Lon_Lat)

head(selected_data)
```

```
##   OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC LOC_CLASSFCTN_DESC
## 1 08/09/2021   01:06:00    BRONX
## 2 04/07/2018   19:48:00 BROOKLYN
## 3 12/02/2022   22:57:00    BRONX           OUTSIDE             STREET
## 4 11/19/2006   01:50:00 BROOKLYN
## 5 05/09/2010   01:58:00    BRONX
## 6 07/22/2012   21:35:00    BRONX
##               LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
## 1                                             false
```

```
## 2                                          true          25-44        M
## 3            GROCERY/BODEGA                false         (null)   (null)
## 4                 PVT HOUSE                true        UNKNOWN        U
## 5   MULTI DWELL - APT BUILD                true          25-44        M
## 6 MULTI DWELL - PUBLIC HOUS                false         18-24        M
##        PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE
## 1                        18-24       M    BLACK
## 2 WHITE HISPANIC         25-44       M    BLACK
## 3         (null)         25-44       M    BLACK
## 4        UNKNOWN         18-24       M    BLACK
## 5          BLACK           <18       F    BLACK
## 6          BLACK         18-24       M    BLACK
```

Extracting Murder-Related Records

Since we are primarily focused on murders, we will filter the data to include only rows where `filter(toupper(STATISTICAL_MURDER_FLAG) == "TRUE")`, ensuring that we are analyzing murder cases exclusively. Note that we use the `toupper()` function because this column is not a logical indicator. In case the data isn't standardized, converting all values to uppercase ensures we capture all variations of the string "true." We will call the resulting data set murder_data_set.

```
murder_data_set <- selected_data %>%
  filter(toupper(STATISTICAL_MURDER_FLAG) == "TRUE")

head(murder_data_set)
```

```
##   OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC LOC_CLASSFCTN_DESC
## 1 04/07/2018   19:48:00 BROOKLYN
## 2 11/19/2006   01:50:00 BROOKLYN
## 3 05/09/2010   01:58:00    BRONX
## 4 07/12/2011   22:26:00 BROOKLYN
## 5 06/24/2011   04:36:00    BRONX
## 6 09/17/2018   16:48:00 BROOKLYN
##               LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
## 1                                              true          25-44        M
## 2                 PVT HOUSE                    true        UNKNOWN        U
## 3   MULTI DWELL - APT BUILD                    true          25-44        M
## 4                                              true
## 5                                              true          18-24        M
## 6 MULTI DWELL - PUBLIC HOUS                    true
##        PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE
## 1 WHITE HISPANIC         25-44       M    BLACK
## 2        UNKNOWN         18-24       M    BLACK
## 3          BLACK           <18       F    BLACK
## 4                        25-44       M    BLACK
## 5          BLACK         25-44       M    BLACK
## 6                        25-44       M    BLACK
```

Filling in all Null and Blank Data

To address the null or blank values in our data set, we will replace them with the label "NOT DOCUMENTED", indicating that the NYPD did not record the information. We intentionally avoid using the

term "UNKNOWN", as it is already a distinct category within the data. There could be multiple reasons why certain information was not documented, and analyzing these cases may help reveal potential patterns or biases in how the NYPD records data. We will want to see if documentation practices vary based on victim characteristics or by borough. Therefore, identifying and tracking these documentation gaps will be an important aspect of our analysis.

As a first step, we will convert any blank strings (" ") or values that contain the string"null" to NA. In R, these are not treated as missing values by default, and standardizing them as NA allows us to handle and replace them more efficiently across the dataset.

```
murder_data_set$LOC_OF_OCCUR_DESC[murder_data_set$LOC_OF_OCCUR_DESC == "" | murder_data_set$LOC_OF_OCCU
murder_data_set$LOC_CLASSFCTN_DESC[murder_data_set$LOC_CLASSFCTN_DESC == "" | murder_data_set$LOC_CLASS
murder_data_set$LOCATION_DESC[murder_data_set$LOCATION_DESC == "" | murder_data_set$LOCATION_DESC == "(
murder_data_set$PERP_AGE_GROUP[murder_data_set$PERP_AGE_GROUP == "" | murder_data_set$PERP_AGE_GROUP ==
murder_data_set$PERP_SEX[murder_data_set$PERP_SEX == "" | murder_data_set$PERP_SEX == "(null)"] <- NA
murder_data_set$PERP_RACE[murder_data_set$PERP_RACE == "" | murder_data_set$PERP_RACE == "(null)"] <- NA

head(murder_data_set)
```

```
##   OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC LOC_CLASSFCTN_DESC
## 1 04/07/2018   19:48:00 BROOKLYN              <NA>               <NA>
## 2 11/19/2006   01:50:00 BROOKLYN              <NA>               <NA>
## 3 05/09/2010   01:58:00    BRONX              <NA>               <NA>
## 4 07/12/2011   22:26:00 BROOKLYN              <NA>               <NA>
## 5 06/24/2011   04:36:00    BRONX              <NA>               <NA>
## 6 09/17/2018   16:48:00 BROOKLYN              <NA>               <NA>
##               LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
## 1                      <NA>                    true          25-44        M
## 2                 PVT HOUSE                    true        UNKNOWN        U
## 3   MULTI DWELL - APT BUILD                    true          25-44        M
## 4                      <NA>                    true           <NA>     <NA>
## 5                      <NA>                    true          18-24        M
## 6 MULTI DWELL - PUBLIC HOUS                    true           <NA>     <NA>
##        PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE
## 1 WHITE HISPANIC         25-44       M    BLACK
## 2        UNKNOWN         18-24       M    BLACK
## 3          BLACK           <18       F    BLACK
## 4           <NA>         25-44       M    BLACK
## 5          BLACK         25-44       M    BLACK
## 6           <NA>         25-44       M    BLACK
```

Now we can fill the NA values we created using the `replace_na()` function from the tidyverse. This function allows us to replace missing values with specified content, making the data more complete and easier to analyze.

```
murder_data_set <- murder_data_set %>%
  replace_na(list(LOC_OF_OCCUR_DESC = "NOT DOCUMENTED", LOC_CLASSFCTN_DESC = "NOT DOCUMENTED",
  LOCATION_DESC = "NOT DOCUMENTED", PERP_AGE_GROUP = "NOT DOCUMENTED",
  PERP_SEX = "NOT DOCUMENTED", PERP_RACE = "NOT DOCUMENTED"))

head(murder_data_set)
```

```
##   OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC LOC_CLASSFCTN_DESC
```

```
## 1 04/07/2018   19:48:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
## 2 11/19/2006   01:50:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
## 3 05/09/2010   01:58:00    BRONX    NOT DOCUMENTED     NOT DOCUMENTED
## 4 07/12/2011   22:26:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
## 5 06/24/2011   04:36:00    BRONX    NOT DOCUMENTED     NOT DOCUMENTED
## 6 09/17/2018   16:48:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
##               LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1            NOT DOCUMENTED                    true          25-44
## 2                PVT HOUSE                     true        UNKNOWN
## 3   MULTI DWELL - APT BUILD                    true          25-44
## 4            NOT DOCUMENTED                    true NOT DOCUMENTED
## 5            NOT DOCUMENTED                    true          18-24
## 6 MULTI DWELL - PUBLIC HOUS                    true NOT DOCUMENTED
##         PERP_SEX      PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE
## 1              M WHITE HISPANIC         25-44       M    BLACK
## 2              U        UNKNOWN         18-24       M    BLACK
## 3              M          BLACK           <18       F    BLACK
## 4 NOT DOCUMENTED NOT DOCUMENTED         25-44       M    BLACK
## 5              M          BLACK         25-44       M    BLACK
## 6 NOT DOCUMENTED NOT DOCUMENTED         25-44       M    BLACK
```

Question One: Has the Murder Rate Gone Down Over Time among the Borough Locations?

For our first analysis, we aim to determine whether the murder rate has been increasing or decreasing over time, and whether this trend is consistent across all boroughs. To explore this, we'll use the previously created murder_data_set and generate a line graph using ggplot2, with each borough represented by a distinct color to easily compare their trends.

The first step is to convert the OCCUR_Date field into a proper date format and ensure it is recognized as a Date object rather than a string. We'll then adjust it to display only the month and year by setting the date to the first day of each month. This prevents overly granular daily data and ensures accurate grouping by date and borough in our analysis.

```r
# Extract the OCCUR_DATE column (character dates in MM/DD/YYYY format)
date_obj <- murder_data_set$OCCUR_DATE

# Convert to Date class
formatted_date <-  as.Date(date_obj, format = "%m/%d/%Y")

#Trim the date to get rid of the actual day
Date_Char <- format(formatted_date, "%m/%Y")

#paste0 concatenates strings without spaces. It adds "/01" to the end of "MM/YYYY", making it "MM/YYYY/
first_date <- paste0(Date_Char, "/01")

# Convert back to Date object with full date
Month_year <-as.Date(first_date, format = "%m/%Y/%d")

# Add Month_year as a new column to the data frame
murder_data_set$MONTH_YEAR <- Month_year

head(murder_data_set)
```

```
##   OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC LOC_CLASSFCTN_DESC
## 1 04/07/2018   19:48:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
## 2 11/19/2006   01:50:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
## 3 05/09/2010   01:58:00    BRONX    NOT DOCUMENTED     NOT DOCUMENTED
## 4 07/12/2011   22:26:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
## 5 06/24/2011   04:36:00    BRONX    NOT DOCUMENTED     NOT DOCUMENTED
## 6 09/17/2018   16:48:00 BROOKLYN    NOT DOCUMENTED     NOT DOCUMENTED
##               LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1            NOT DOCUMENTED                    true          25-44
## 2                PVT HOUSE                     true        UNKNOWN
## 3   MULTI DWELL - APT BUILD                    true          25-44
## 4            NOT DOCUMENTED                    true NOT DOCUMENTED
## 5            NOT DOCUMENTED                    true          18-24
## 6 MULTI DWELL - PUBLIC HOUS                    true NOT DOCUMENTED
##         PERP_SEX      PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE MONTH_YEAR
## 1              M WHITE HISPANIC         25-44       M    BLACK 2018-04-01
## 2              U        UNKNOWN         18-24       M    BLACK 2006-11-01
## 3              M          BLACK           <18       F    BLACK 2010-05-01
## 4 NOT DOCUMENTED NOT DOCUMENTED         25-44       M    BLACK 2011-07-01
## 5              M          BLACK         25-44       M    BLACK 2011-06-01
## 6 NOT DOCUMENTED NOT DOCUMENTED         25-44       M    BLACK 2018-09-01
```

Next, we'll count the number of murders for each borough by month, which is essentially calculating the monthly murder totals per borough.

```
murder_counts <- murder_data_set %>%
  #group_by(Month_year, BORO_NM): groups the data by month and borough
  group_by(MONTH_YEAR, BORO) %>%
  #summarize(Count = n()): counts the number of rows in each group
  summarize(Count = n(), .groups = "drop")

head(murder_counts)
```

```
## # A tibble: 6 x 3
##   MONTH_YEAR BORO      Count
##   <date>     <chr>     <int>
## 1 2006-01-01 BRONX         9
## 2 2006-01-01 BROOKLYN     12
## 3 2006-01-01 MANHATTAN     4
## 4 2006-01-01 QUEENS        4
## 5 2006-02-01 BRONX         7
## 6 2006-02-01 BROOKLYN      8
```

Finally, we will show the visualization for murder counts per borough per month.

```
#
## Question One: Has the Murder Rate Gone Down Over Time among the Borough Locations? Initializes the p
ggplot(murder_counts, aes(x = MONTH_YEAR, y = Count, color = BORO)) +
  geom_line(size = 1) +
  #Splits the plot into separate panels, one for each borough. Free_y allows each panel to have its own
  facet_wrap(~ BORO, scales = "free_y", ncol = 1) +
  #labs function sets the plot title and axis labels
  labs(title = "Monthly Murder Counts by Borough",
```

```
      x = "Month-Year", y = "Murder Count") +
  #theme_minimal function applies a clean, minimalist theme with no background grid lines.
  #base_size = 8: assigns a font size.
  theme_minimal(base_size = 8) +
  #axis.text.x = element_text() function rotates the x-axis labels by 45 degrees so that long or dense
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

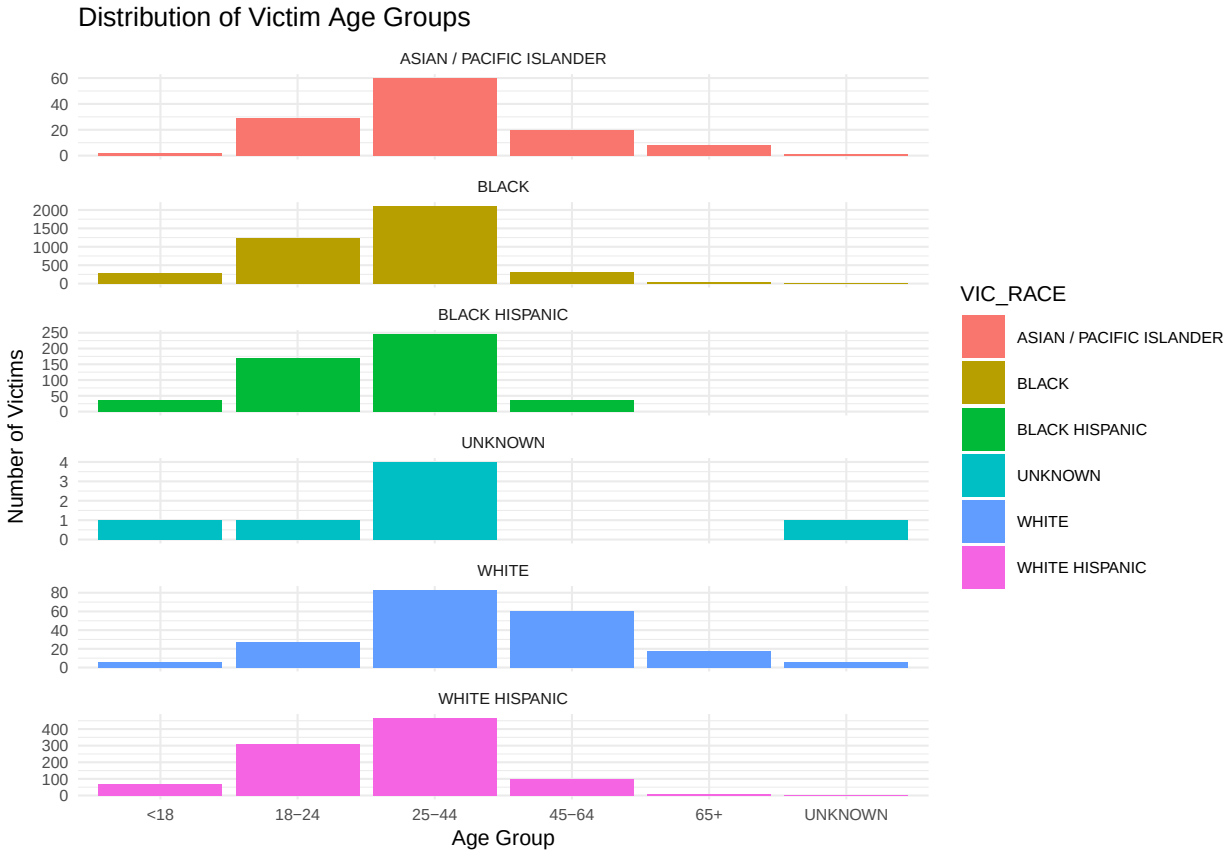
Question One Answer: Has the murder rate decreased over time across the different boroughs?

The overall pattern is no clear citywide increase or decrease in the murder rate. Most of the boroughs show irregular fluctuations or possible seasonal fluctuations rather than a steady downward or upward trend.

Question Two: What is the typical age of murder victims, and do age patterns differ across racial groups?

We will use a histogram to explore the distribution of victim ages and apply the `facet_wrap()` function to visualize how this distribution varies across different racial groups.

```
ggplot(murder_data_set, aes(x = VIC_AGE_GROUP, fill = VIC_RACE)) +
  geom_bar() +  facet_wrap(~ VIC_RACE, scales = "free_y", ncol = 1) +
  labs(title = "Distribution of Victim Age Groups",
       x = "Age Group", y = "Number of Victims") +
  theme_minimal(base_size = 8)
```



Question Two Answer: What is the typical age of murder victims, and do age patterns differ across racial groups?

Based on the histogram, most murder victims are between 25 and 44 years old, followed by those aged 18 to 24. This trend is consistent across most racial groups, except for non-Hispanic White victims, where the second largest age group is 45 to 64 years old.

Question Three: Is there a statistically significant relationship between the victim's race and the likelihood that the perpetrator's race is not documented?

Before performing a logistic regression, we need to clean and prepare the data. Logistic regression is used to model the probability of binary outcomes such as yes/no, success/failure, or, in this case, whether information

is missing or not.

In this analysis, we aim to examine whether the likelihood of a perpetrator's race being undocumented is associated with the victim's race. Identifying a statistically significant relationship could suggest potential disparities in how thoroughly cases are documented, possibly indicating differences in investigative attention based on the victim's race.

```r
# Convert the perpetrator's race column into a binary variable indicating whether the race is documente
murder_data_set$PERP_RACE_MISSING <- ifelse(murder_data_set$PERP_RACE == "NOT DOCUMENTED" , 1, 0)

# Convert Victim's Race to a factor, as this is the preferred format for categorical variables in model
murder_data_set$VIC_RACE <- as.factor(murder_data_set$VIC_RACE)

# This sets WHITE NON-HISPANIC as the reference group for comparison.
# We use this group as a baseline because, historically, White individuals in the U.S. have not faced s
murder_data_set$VIC_RACE <- relevel(murder_data_set$VIC_RACE, ref = "WHITE")

model <- glm(PERP_RACE_MISSING ~ VIC_RACE, data = murder_data_set, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = PERP_RACE_MISSING ~ VIC_RACE, family = "binomial",
##     data = murder_data_set)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.4500     0.1802  -8.045 8.65e-16 ***
## VIC_RACEASIAN / PACIFIC ISLANDER   0.3514     0.2774   1.267  0.20519
## VIC_RACEBLACK                      0.9270     0.1832   5.060 4.19e-07 ***
## VIC_RACEBLACK HISPANIC             0.5676     0.2057   2.760  0.00579 **
## VIC_RACEUNKNOWN                   -0.3417     1.0951  -0.312  0.75498
## VIC_RACEWHITE HISPANIC             0.2524     0.1959   1.288  0.19763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7336.7  on 5764  degrees of freedom
## Residual deviance: 7233.8  on 5759  degrees of freedom
## AIC: 7245.8
##
## Number of Fisher Scoring iterations: 4
```

```r
# Extract summary info
model_summary <- summary(model)

# Create a tidy table
log_odds_table <- data.frame(
  Term = rownames(model_summary$coefficients),
  Log_Odds = model_summary$coefficients[, "Estimate"],
  Odds_Ratio = exp(model_summary$coefficients[, "Estimate"]),
  Std_Error = model_summary$coefficients[, "Std. Error"],
  z_value = model_summary$coefficients[, "z value"],
```

```
  p_value = model_summary$coefficients[, "Pr(>|z|)"]
)

# Extract intercept and coefficients
intercept <- log_odds_table$Log_Odds[log_odds_table$Term == "(Intercept)"]


#Applying plogis() to intercept + coefficient gives the probability for that specific category.
#It transforms log-odds from a logistic regression model into probabilities ranging from 0 to 1. The fu
log_odds_table$Predicted_Prob <- plogis(intercept + log_odds_table$Log_Odds)

#Applying plogis() to the intercept alone gives the baseline probability (~19% for White victims)
log_odds_table$Predicted_Prob[log_odds_table$Term == "(Intercept)"] <- plogis(intercept)

print(log_odds_table, digits = 3)
```

```
##                                                                 Term Log_Odds
## (Intercept)                                                (Intercept)   -1.450
## VIC_RACEASIAN / PACIFIC ISLANDER VIC_RACEASIAN / PACIFIC ISLANDER    0.351
## VIC_RACEBLACK                                           VIC_RACEBLACK    0.927
## VIC_RACEBLACK HISPANIC                     VIC_RACEBLACK HISPANIC    0.568
## VIC_RACEUNKNOWN                                   VIC_RACEUNKNOWN   -0.342
## VIC_RACEWHITE HISPANIC                     VIC_RACEWHITE HISPANIC    0.252
##                                  Odds_Ratio Std_Error z_value  p_value
## (Intercept)                           0.235     0.180  -8.045 8.65e-16
## VIC_RACEASIAN / PACIFIC ISLANDER      1.421     0.277   1.267 2.05e-01
## VIC_RACEBLACK                         2.527     0.183   5.060 4.19e-07
## VIC_RACEBLACK HISPANIC                1.764     0.206   2.760 5.79e-03
## VIC_RACEUNKNOWN                       0.711     1.095  -0.312 7.55e-01
## VIC_RACEWHITE HISPANIC                1.287     0.196   1.288 1.98e-01
##                                  Predicted_Prob
## (Intercept)                               0.190
## VIC_RACEASIAN / PACIFIC ISLANDER          0.250
## VIC_RACEBLACK                             0.372
## VIC_RACEBLACK HISPANIC                    0.293
## VIC_RACEUNKNOWN                           0.143
## VIC_RACEWHITE HISPANIC                    0.232
```

Plotting the Logistic Regression Results

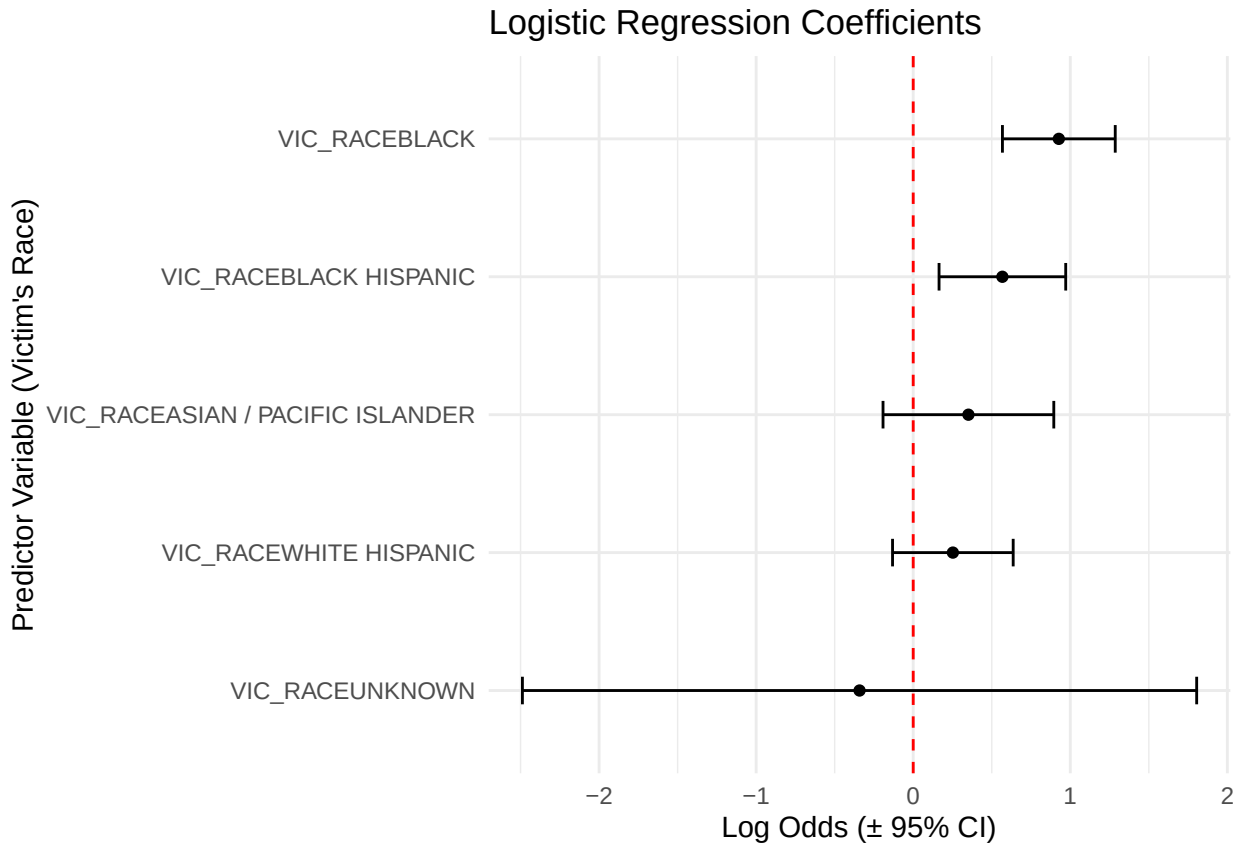
R code is creating a coefficient plot of logistic regression results, specifically the log-odds coefficients from a logistic regression, and their 95% confidence intervals. Categories where the error bar doesn't cross 0 are likely statistically significant.

```
#Creating a data frame
#Excluding the intercept
df <- log_odds_table[log_odds_table$Term != "(Intercept)", ]

#Calculating the confidence intervals
#This tells us the range within which the true log-odds are expected to fall 95% of the time.
df$CI_lower <- df$Log_Odds - 1.96 * df$Std_Error
df$CI_upper <- df$Log_Odds + 1.96 * df$Std_Error
```

```r
# Plot

#x = reorder(Term, Log_Odds): sorts terms by effect size.
#y = Log_Odds: plots the log-odds on the vertical axis.
ggplot(df, aes(x = reorder(Term, Log_Odds), y = Log_Odds)) +
  geom_point() +
  #Adds error bars to show the 95% confidence intervals.
  geom_errorbar(aes(ymin = CI_lower, ymax = CI_upper), width = 0.2) +
  #Adds a horizontal dashed red line at 0.
  #This helps visually determine which coefficients are statistically significant.
  #If the Confidence Intervals crosses zero, the effect may not be significant.
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  #Flips the x and y axes so the terms are on the y-axis and the log-odds are on the x-axis.
  #This makes the plot more readable.
  coord_flip() +
  labs(title = "Logistic Regression Coefficients",
       x = "Predictor Variable (Victim's Race)",
       y = "Log Odds (± 95% CI)") +
  theme_minimal()
```



Question 3 Answer: An Interpretation of the Logistic Regression Model

The probability that the perpetrator's race is missing for White non-Hispanic victims is 19%. We use this group as the baseline because, historically, White individuals in the U.S. have not faced systemic racial

discrimination to the same extent as other racial groups.

All other racial groups show a higher probability of missing perpetrator race data; however, only two groups show statistically significant differences (p-value < 0.05). A small p-value (less than 0.05) indicates strong evidence against the null hypothesis, suggesting that the predictor variable of victim's race has a significant effect on the outcome which is whether the perpetrator's race is missing. A large p-value suggests the data are consistent with the null hypothesis, meaning there isn't enough evidence to conclude that the victim's race influences the likelihood of missing perpetrator race data. This does not prove that there is no effect. This only proves that we lack sufficient evidence to demonstrate that certain victim categories have this effect.

In our results, the p-values for two groups are extremely low, leading us to reject the null hypothesis. This indicates that the race of the victim, specifically, being Black or Black Hispanic, is significantly associated with increased odds of the perpetrator's race being missing.

Black victims have 2.5 times the odds (Odds Ratio = 2.53) of missing perpetrator race data compared to White victims. The predicted probability of missing data for this group is approximately 37.2%.

Black Hispanic victims have 1.76 times the odds (Odds Ratio = 1.764) of missing perpetrator race data compared to White victims. The predicted probability of missing data for this group is approximately 29.3%.

What does this imply about potential bias in our dataset and in our analysis?

Our analysis found a statistically significant relationship between the victim's race and whether the perpetrator's race was documented. This was found particularly for Black and Black Hispanic victims. This suggests the possibility of systemic bias in how case details are recorded. The fact that data completeness varies meaningfully by a victim's race raises serious concerns about the fairness and consistency of investigative practices.

To determine whether this is an isolated issue or part of a broader pattern, further logistic regressions should be conducted on other descriptive characteristics of the perpetrator. This would help assess whether the same racial disparities appear in other areas of documentation. Identifying consistent gaps would strengthen the case for systemic issues in how data is collected and reported.

This is especially troubling given the context: these are murder cases, where complete and accurate descriptions of perpetrators are crucial for solving crimes and achieving justice. Gaps in documentation may hinder investigations and reduce the likelihood of holding perpetrators accountable.

Historically, non-white communities in the U.S. have faced unequal treatment by law enforcement, including under-policing and over-policing. It remains unclear whether these data gaps result from negligence by police or from mistrust of law enforcement that leads victims or witnesses to withhold information. Regardless of the cause, the result is the same: serious gaps in critical data, which call into question the integrity and reliability of the NYPD Data set.

As for potential bias in my own perspective that could have influenced how I analyzed the data, I acknowledge that I have a general lack of trust in policing institutions. This skepticism comes from being aware of documented instances of corruption and the long history of systemic oppression faced by non-white communities at the hands of law enforcement.

That said, I believe I've made a conscious effort to counterbalance this bias by not overstating what the data shows. For example, I did not claim that all non-white racial groups had statistically significant evidence of missing perpetrator race data because, in fact, many of their p-values were too high to support such a conclusion. While we did find that certain victim race categories were statistically associated with missing data, this does not definitively prove police bias. We simply don't have enough evidence to make that determination.

As I mentioned earlier, it's also possible that missing data may stem not from negligence or bias on the part of the police, but from a lack of trust in law enforcement, which may lead community members to withhold information. Either explanation is plausible, and without more data, we can't be certain which is at play.