MyfirstRMD

2024-02-28

```
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org"
options(repos = r)
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   3.5.0     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 error
```

```
library(ggplot2)
```

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_cases <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## 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.
```

```
nypd_cases
```

```
## # A tibble: 27,312 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
##  1    228798151 05/27/2021 21:30      QUEENS   <NA>                   105
##  2    137471050 06/27/2014 17:40      BRONX    <NA>                    40
##  3    147998800 11/21/2015 03:56      QUEENS   <NA>                   108
##  4    146837977 10/09/2015 18:30      BRONX    <NA>                    44
##  5     58921844 02/19/2009 22:58      BRONX    <NA>                    47
## 6    219559682 10/21/2020 21:36      BROOKLYN <NA>                        81
## 7     85295722 06/17/2012 22:47      QUEENS   <NA>                       114
## 8     71662474 03/08/2010 19:41      BROOKLYN <NA>                        81
## 9     83002139 02/05/2012 05:45      QUEENS   <NA>                       105
## 10    86437261 08/26/2012 01:10      QUEENS   <NA>                       101
## # i 27,302 more rows
## # 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>, Lon_Lat <chr>
```

```
selected_data <- nypd_cases[, c("OCCUR_DATE", "BORO", "PERP_RACE", "VIC_RACE", "VIC_AGE_GROUP")]
View(selected_data)
selected_data <- nypd_cases[, c("OCCUR_DATE", "BORO", "PERP_RACE", "VIC_RACE", "VIC_AGE_GROUP", "VIC_SE
```

```
selected_data = selected_data %>%
rename(DATE = OCCUR_DATE)
selected_data = selected_data %>%
rename(Location = BORO, Perp_race = PERP_RACE)
selected_data = selected_data %>%
rename(Date = DATE, Vic_race = VIC_RACE, Vic_age = VIC_AGE_GROUP, Vic_sex = VIC_SEX, Perp_sex = PERP_SE
summary(selected_data)
```

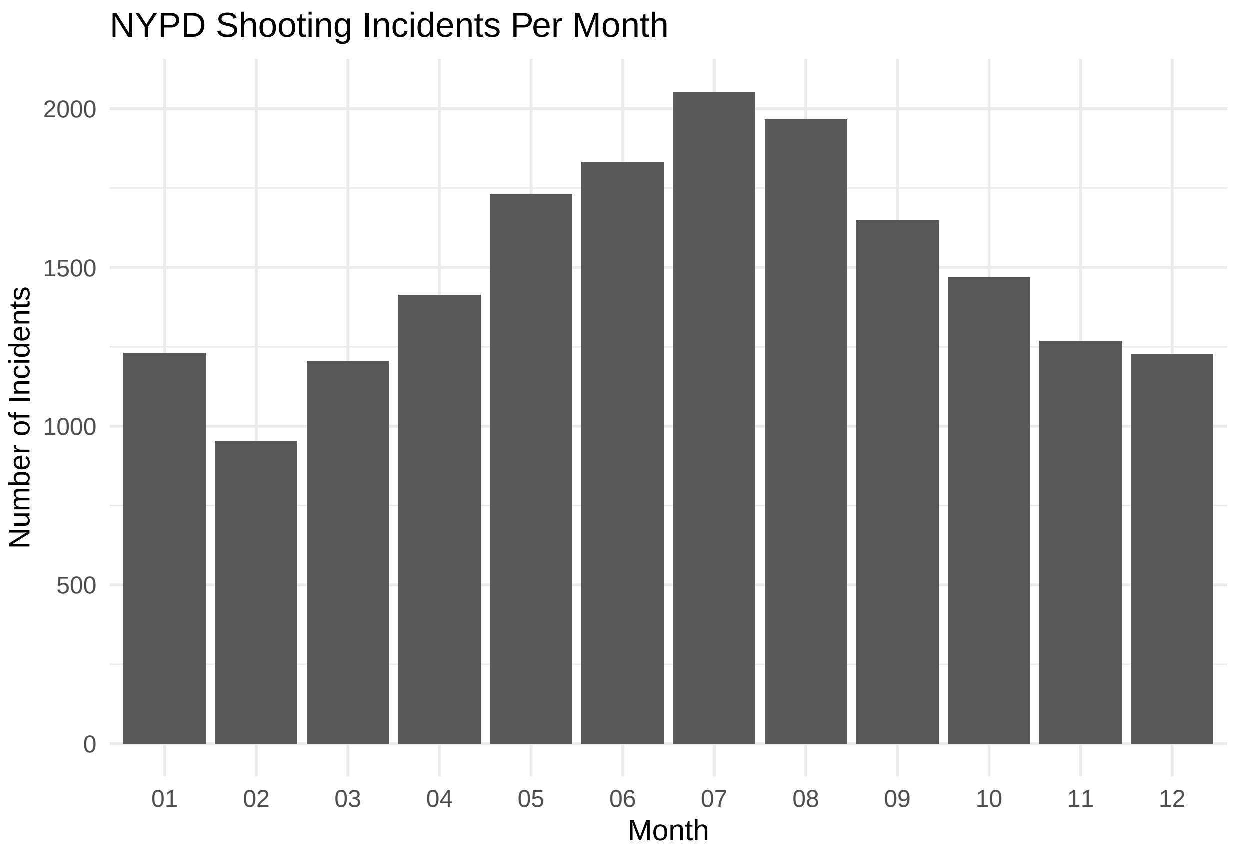
```
##      Date             Location          Perp_race           Vic_race
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    Vic_age            Vic_sex            Perp_sex
##  Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
selected_data = na.omit(selected_data)
```

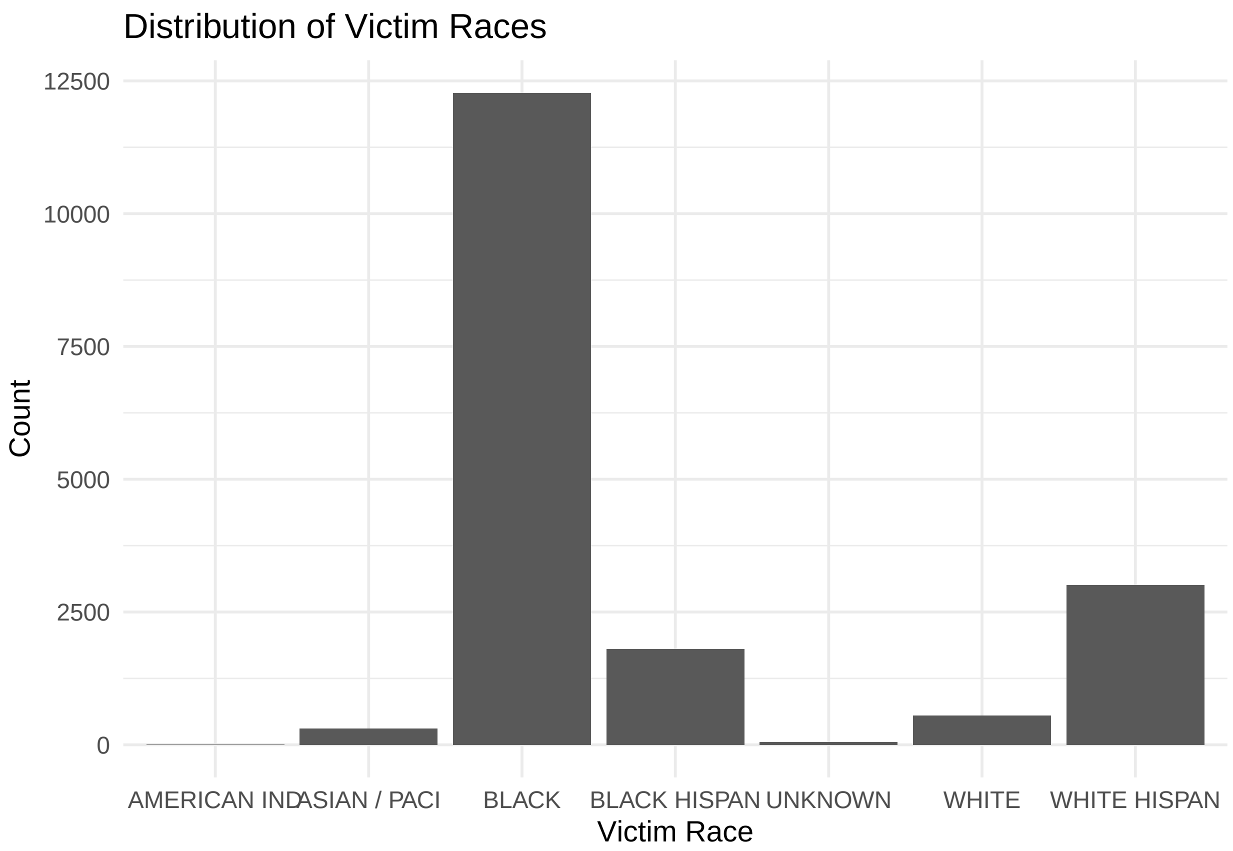
```
selected_data
```

```
## # A tibble: 18,002 x 7
##    Date       Location  Perp_race Vic_race       Vic_age Vic_sex Perp_sex
##    <chr>      <chr>     <chr>     <chr>          <chr>   <chr>   <chr>
##  1 02/19/2009 BRONX     BLACK     BLACK          45-64   M       M
##  2 08/26/2012 QUEENS    BLACK     BLACK          25-44   M       M
##  3 08/29/2010 BROOKLYN  BLACK     BLACK          25-44   M       M
##  4 05/25/2011 BRONX     UNKNOWN   WHITE          18-24   M       U
##  5 11/09/2008 BROOKLYN  UNKNOWN   BLACK HISPANIC 25-44   M       U
##  6 07/05/2007 BRONX     UNKNOWN   BLACK          18-24   M       M
##  7 07/27/2010 MANHATTAN BLACK     BLACK          25-44   M       M
##  8 03/07/2021 BROOKLYN  BLACK     WHITE          25-44   M       M
##  9 02/01/2015 MANHATTAN BLACK     BLACK          18-24   F       M
## 10 03/03/2006 BROOKLYN  BLACK     WHITE          45-64   M       M
## # i 17,992 more rows
```

```
selected_data$Date <- as.Date(selected_data$Date, format = "%m/%d/%Y")

selected_data$Month <- format(selected_data$Date, "%m")

ggplot(selected_data, aes(x = Month)) +
     geom_bar() +
     labs(title = "NYPD Shooting Incidents Per Month",
          x = "Month",
          y = "Number of Incidents") +
     theme_minimal()
```



```
library(stringr)

ggplot(selected_data, aes(x = str_sub(Vic_race, 1, 12))) +
  geom_bar() +
  labs(title = "Distribution of Victim Races",
       x = "Victim Race",
       y = "Count") +
  theme_minimal()
```



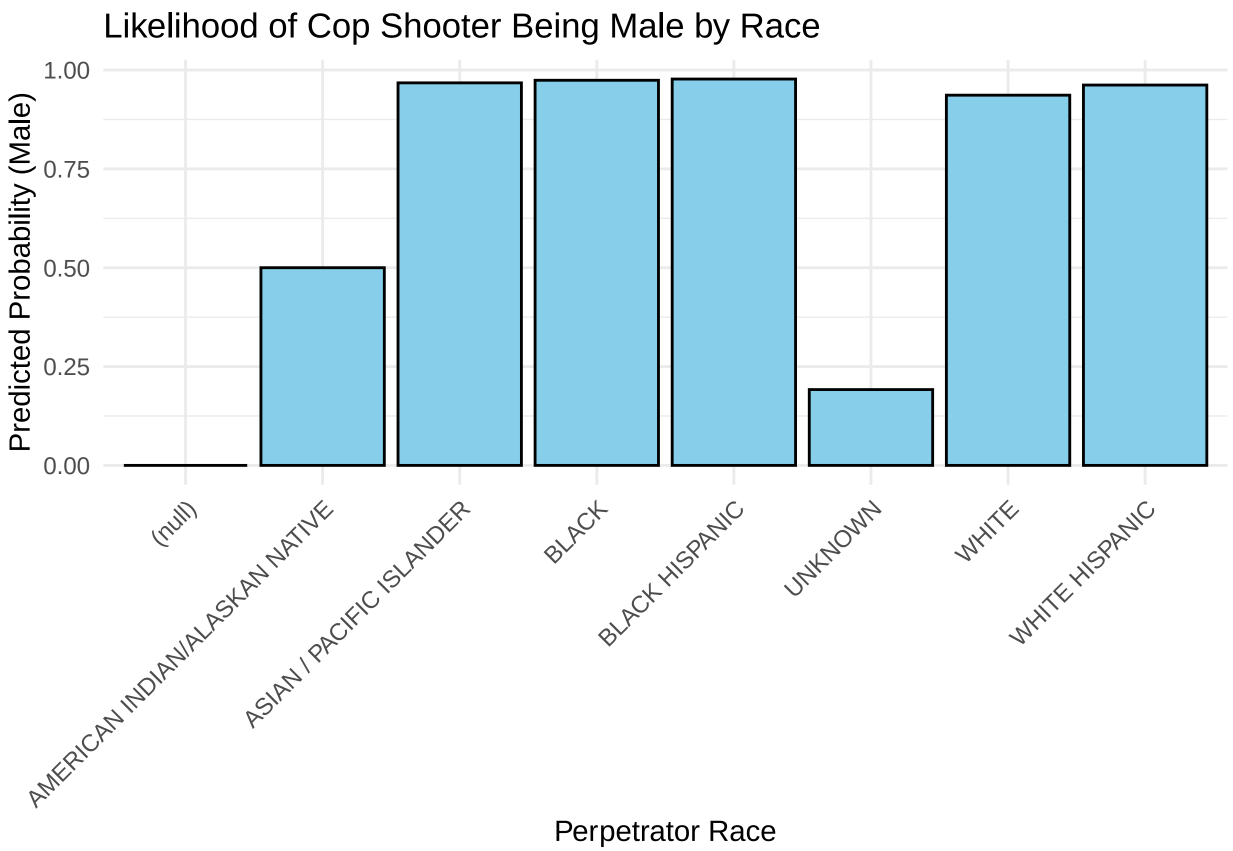
#Here i only include the first 12 characters of the race in order to avoid overlap as some of the names are very long

```
df <- selected_data
df$Perp_sex_binary <- ifelse(df$Perp_sex == "M", 1, 0)
model <- glm(Perp_sex_binary ~ Perp_race, family = binomial(link = "logit"), data = df)

summary(model)
```

```
##
## Call:
## glm(formula = Perp_sex_binary ~ Perp_race, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               -17.57     156.38  -0.112    0.911
## Perp_raceAMERICAN INDIAN/ALASKAN NATIVE    17.57     156.39   0.112    0.911
## Perp_raceASIAN / PACIFIC ISLANDER          20.96     156.38   0.134    0.893
## Perp_raceBLACK                             21.19     156.38   0.136    0.892
## Perp_raceBLACK HISPANIC                    21.32     156.38   0.136    0.892
## Perp_raceUNKNOWN                           16.13     156.38   0.103    0.918
## Perp_raceWHITE                             20.26     156.38   0.130    0.897
## Perp_raceWHITE HISPANIC                    20.80     156.38   0.133    0.894
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 14734.6  on 18001  degrees of freedom
## Residual deviance:  5765.4  on 17994  degrees of freedom
## AIC: 5781.4
##
## Number of Fisher Scoring iterations: 16
```

```
race_data <- data.frame(Perp_race = unique(df$Perp_race))


race_data$predicted_prob <- predict(model, newdata = race_data, type = "response")


ggplot(race_data, aes(x = Perp_race, y = predicted_prob)) +
     geom_bar(stat = "identity", fill = "skyblue", color = "black") +
     labs(title = "Likelihood of Cop Shooter Being Male by Race",
          x = "Perpetrator Race",
          y = "Predicted Probability (Male)") +
     theme_minimal() +
     theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



#In this analysis, we used logistic regression to model the likelihood of the sex of the cop shooter being male based on the victim's race. This is a binary classification as the sex of the cop was converted to binary values for the model. As seen in the model, the likelihood of the cop shooter being a male is almost 100% for all of the races except american indian and unknown. #Regarding bias in this data, I believe there is bias present starting from the data collection and extending to the data analysis itself. Regarding the data, it is possible that some incidents were omitted and only the reported and recorded incidents are present in the data. This leaves the possibility of bias present in the data that may skew some of the analysis I did. Additionally, in the data analysis portion, one of the first things I did was remove rows that had incomplete data (N/A). This is a form of selection bias and further affects the results from the analysis I performed.