NYPD Shooting Data Project

Angela Wilson

2024-11-05

Project Step 1: Start an Rmd Document

My analysis finds the the majority of Black victims of NYPD shooting incidents faced Black perpetrators over 2006 through 2023.

About the Dataset The NYPD Shooting Incident Data (Historic) dataset was taken from https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic on 10/30/2024. It is a non-federal dataset that lists every shooting incident that occurred in New York City from 2006 through 2023. Data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record in the dataset contains information about each shooting incident including time, location, and demographics of victims and suspects.

```
# Load libraries
library(tidyverse)
```

Import Data

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.1

## Warning: package 'lubridate' was built under R version 4.3.1
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     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 1.9.2     v tidyr     1.3.1
## v purrr     1.0.1
## -- 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
```

```
# Read in Data
url = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
raw_data = read_csv(url)
```

```
## Rows: 28562 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.
```

Project Step 2: Tidy and Transform Your Data

Initial Summary of Data Next, we called `summary(data)` to get an overview of our dataset.

```
# Initial Summary of raw data
summary(raw_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  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
```

We clicked the landing page (https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8/about_data) of the dataset to find descriptions for each of those 21 variables.

Column (Variable) Name	Description
INCIDENT_KEY	Randomly generated persistent ID for each arrest
OCCUR_DATE	Exact date of the shooting incident
OCCUR_TIME	Exact time of the shooting incident
BORO	Borough where the shooting incident occurred
LOC_OF_OCCUR_DESC	(no description available)
PRECINCT	Precinct where the shooting incident occurred
JURISDICTION_CODE	Jurisdiction where the shooting incident occurred. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions
LOC_CLASSFCTN_DESC	(no description available)
LOCATION_DESC	Location of the shooting incident
STATISTICAL_MURDER_FLAG	Shooting resulted in the victim's death which would be counted as a murder
PERP_AGE_GROUP	Perpetrator's age within a category
PERP_SEX	Perpetrator's sex description
PERP_RACE	Perpetrator's race description

Column (Variable) Name	Description
VIC_AGE_GROUP	Victim's age within a category
VIC_SEX	Victim's sex description
VIC_RACE	Victim's race description
X_COORD_CD	Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Lon_Lat	Longitude and Latitude Coordinates for mapping

Clean and Transform Data We perform the following steps to properly clean and transform our data:

- X_COORD_CD, Y_COORD_CD, Latitude, Longitude, and Lon_Lat are not needed for our analysis, so they were removed from our dataset.
- LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC LOCATION_DESC all have a significant number of missing values, so they were removed as well.
- JURISDICTION_CODE, PERP_AGE_GROUP, PERP_SEX, and PERP_RACE were kept in our dataset, but rows with missing values for these columns were removed.

```
# Remove variables: X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat,
#     LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC
data = raw_data %>%
    select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat,
              LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC))

# Remove rows with missing data
data = data[complete.cases(data), ]
```

Further examination of the data showed that missing values were also entered in as strings such as "(null)", "U" or "UNKNOWN". We removed those rows as well.

```
# Remove rows containing strings as missing values
data = data %>%
  filter(!grepl("null", PERP_AGE_GROUP)) %>%
  filter(!grepl("UNKNOWN", PERP_AGE_GROUP)) %>%
  filter(!grepl("1020", PERP_AGE_GROUP)) %>%
  filter(!grepl("940", PERP_AGE_GROUP)) %>%
  filter(!grepl("224", PERP_AGE_GROUP)) %>%
  filter(!grepl("1028", PERP_AGE_GROUP)) %>%
  filter(!grepl("U", PERP_SEX)) %>%
  filter(!grepl("U", PERP_SEX)) %>%
  filter(!grepl("(null)", PERP_RACE)) %>%
  filter(!grepl("UNKNOWN", PERP_RACE)) %>%
  filter(!grepl("UNKNOWN", VIC_AGE_GROUP)) %>%
  filter(!grepl("1022", VIC_AGE_GROUP)) %>%
```

```
  filter(!grepl("U", VIC_SEX)) %>%
  filter(!grepl("UNKNOWN", VIC_RACE))
```

When we further examine our data, we see that out of 14680 observations, we only have 2 incidents where the PERP_RACE was "AMERICAN INDIAN/ALASKAN NATIVE". There are also only 4 incidents where the VIC_RACE was "AMERICAN INDIAN/ALASKAN NATIVE". Because our counts in these categories are negligible, these rows were removed.

```
# Remove rows containing "AMERICAN INDIAN/ALASKA NATIVE" in `PERP_RACE` or `VIC_RACE`
data = data %>%
  filter(!grepl("AMERICAN INDIAN/ALASKAN NATIVE", PERP_RACE)) %>%
  filter(!grepl("AMERICAN INDIAN/ALASKAN NATIVE", VIC_RACE))
```

Change the OCCUR_DATE column from character to date.

```
# Change OCCUR_DATE to date object
data$OCCUR_DATE = as.Date(data$OCCUR_DATE, "%m/%d/%Y")
```

```
# Give a summary of a newly cleaned and transformed data.
summary(data)
```

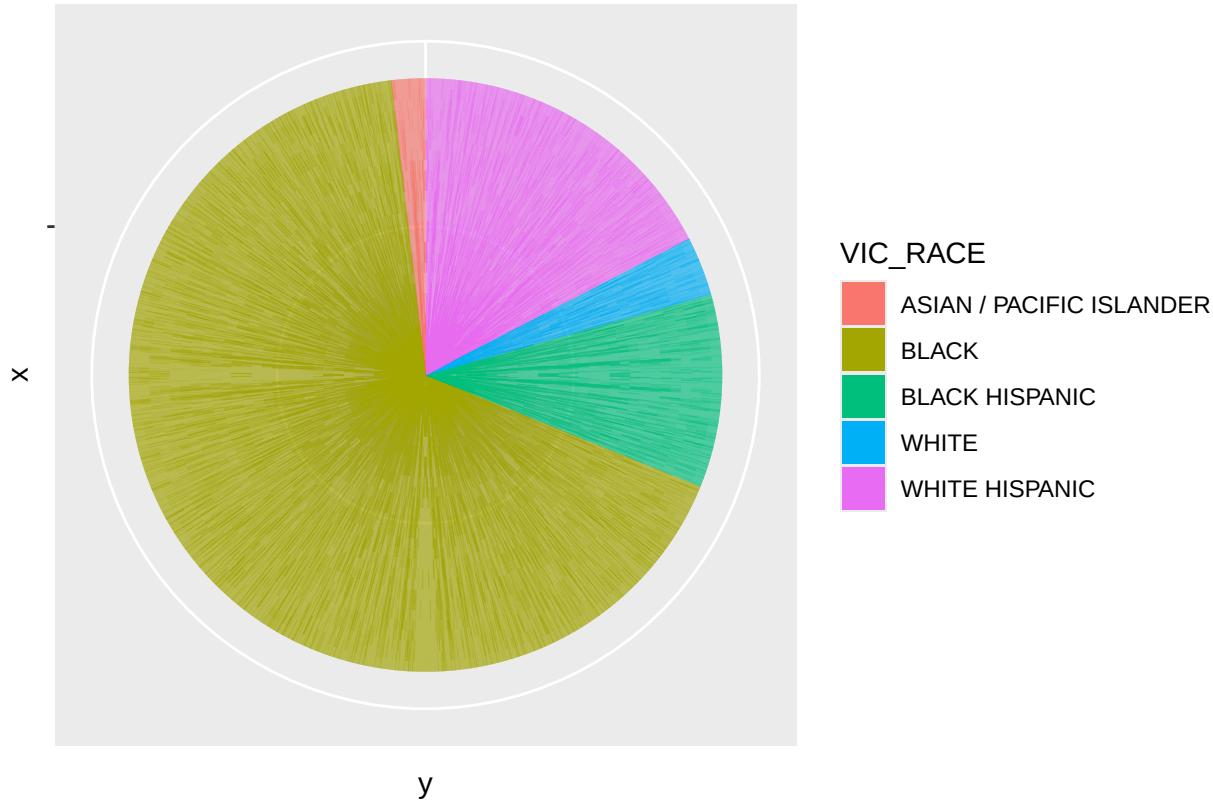
Final Summary of Cleaned and Transformed Data

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:14674      Length:14674
##  1st Qu.: 63506963   1st Qu.:2009-07-07   Class1:hms        Class :character
##  Median : 94282722   Median :2014-01-01   Class2:difftime   Mode  :character
##  Mean   :129032578   Mean   :2014-07-11   Mode  :numeric
##  3rd Qu.:206311808   3rd Qu.:2019-12-09
##  Max.   :279709792   Max.   :2023-12-29
##     PRECINCT     JURISDICTION_CODE STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Min.   :  1.0   Min.   :0.0000    Mode :logical           Length:14674
##  1st Qu.: 43.0   1st Qu.:0.0000    FALSE:11159             Class :character
##  Median : 67.0   Median :0.0000    TRUE :3515              Mode  :character
##  Mean   : 64.6   Mean   :0.3146
##  3rd Qu.: 81.0   3rd Qu.:0.0000
##  Max.   :123.0   Max.   :2.0000
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:14674       Length:14674       Length:14674       Length:14674
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    VIC_RACE
##  Length:14674
##  Class :character
##  Mode  :character
##
##
##
```

Project Step 3: Add Visualizations and Analysis

Visualization 1 Create a pie chart showing the proportion of victims by race.

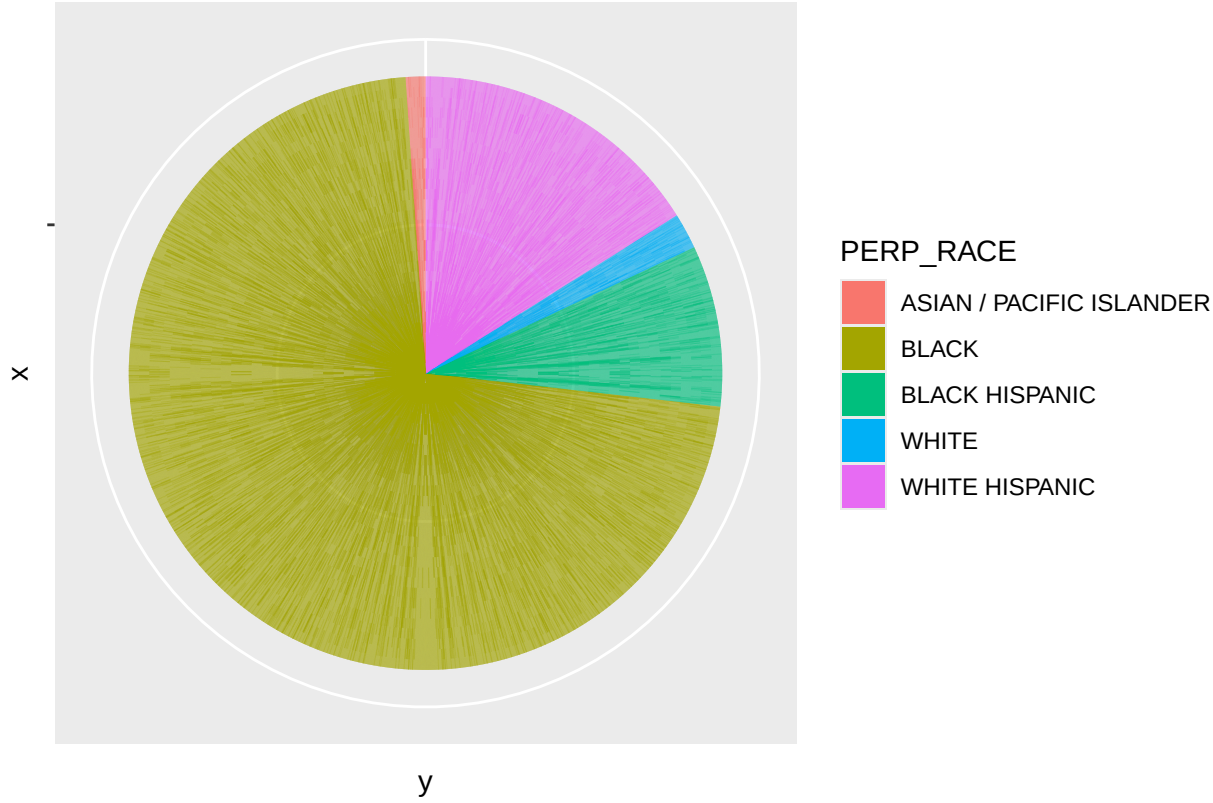
```
# Get proportions of victims by race
VIC_RACE_prop_table = as.data.frame(prop.table(table(data$VIC_RACE)))
VIC_RACE_prop_table = (VIC_RACE_prop_table[order(VIC_RACE_prop_table[,2],
                                                 decreasing = TRUE), ])

VIC_RACE_list = VIC_RACE_prop_table[, 1]
VIC_RACE_proportions = VIC_RACE_prop_table[, 2]

# Pie Chart of victims by race
ggplot(data, aes(x = "", y = "", fill = VIC_RACE)) +
    geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start = 0)
```



Visualization 2 Create a pie chart showing the proportion of perpetrators by race.

```
# Get proportions of perpetrators by race
PERP_RACE_prop_table = as.data.frame(prop.table(table(data$PERP_RACE)))
PERP_RACE_prop_table = (
```

```
    PERP_RACE_prop_table[order(PERP_RACE_prop_table[, 2], decreasing = TRUE), ])

PERP_RACE_list = PERP_RACE_prop_table[, 1]
PERP_RACE_proportions = PERP_RACE_prop_table[, 2]

# Pie Chart of perpetrators by race
ggplot(data, aes(x = "", y = "", fill = PERP_RACE)) +
    geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start = 0)
```



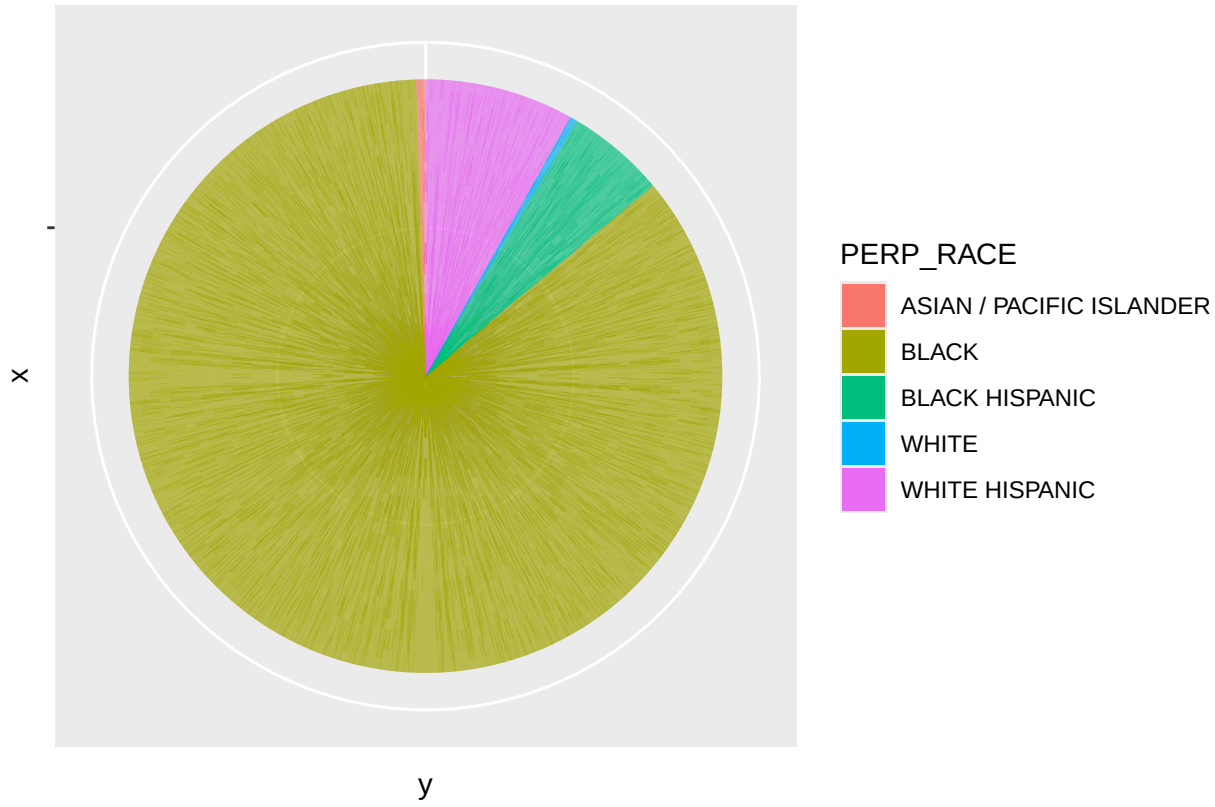
Visualization 3 Both previous visualizations showed that the majority of victims and perpetrators are both Black. Let's look at our subset of Black victims and see the proportion of their perpetrators by race.

```
# Subset of Black victims
data_VIC_RACE_BLACK = data %>%
    filter(VIC_RACE == "BLACK") %>%
    group_by(VIC_RACE, PERP_RACE, OCCUR_DATE) %>%
    select(VIC_RACE, PERP_RACE, OCCUR_DATE, everything())

# Get proportions of perpetrators by race
data_VIC_RACE_BLACK_prop_table = as.data.frame(
    prop.table(table(data_VIC_RACE_BLACK$PERP_RACE)))
data_VIC_RACE_BLACK_prop_table = (
    data_VIC_RACE_BLACK_prop_table[order(data_VIC_RACE_BLACK_prop_table[,2],
```

```
                                          decreasing = TRUE), ])
print(data_VIC_RACE_BLACK_prop_table)
```

```
##                       Var1        Freq
## 2                    BLACK 0.855908074
## 5           WHITE HISPANIC 0.081147041
## 3           BLACK HISPANIC 0.053792963
## 1 ASIAN / PACIFIC ISLANDER 0.005084401
## 4                    WHITE 0.004067521
```

```
# Pie Chart of perpetrators by race for Black victims
ggplot(data_VIC_RACE_BLACK,
       aes(x = "", y = "", fill = PERP_RACE)) +
    geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start = 0)
```



Analysis Right now, our data over the years 2006 - 2023 is all lumped together. Let's break up our data by year so we can see how it changes over time.

```
seq(as.Date("2006-01-01"), to = as.Date("2023-12-29"), by = "1 year")
```

```
##  [1] "2006-01-01" "2007-01-01" "2008-01-01" "2009-01-01" "2010-01-01"
##  [6] "2011-01-01" "2012-01-01" "2013-01-01" "2014-01-01" "2015-01-01"
```

```
## [11] "2016-01-01" "2017-01-01" "2018-01-01" "2019-01-01" "2020-01-01"
## [16] "2021-01-01" "2022-01-01" "2023-01-01"
```

We created separate code chunks to get the proportions of perpetrators by race for black victims for each year within our dataset. We showed the code for extracting 2006 data. Years were changed accordingly to extract data for years 2007-2023, but codes were not shown for brevity.

```
# 2006 perpetrators by race for Black victims
data_VIC_RACE_BLACK_06 = data_VIC_RACE_BLACK %>%
    filter(between(OCCUR_DATE, as.Date("2006-01-01"), as.Date("2006-12-31")))

data_VIC_RACE_BLACK_06_prop_table = as.data.frame(
    prop.table(table(data_VIC_RACE_BLACK_06$PERP_RACE)))
data_VIC_RACE_BLACK_06_prop_table = (data_VIC_RACE_BLACK_06_prop_table[order(
    data_VIC_RACE_BLACK_06_prop_table[, 2], decreasing = TRUE), ])
```

The following vectors show the frequency rates per year of perpetrators by race for Black victims of NYPD shooting incidents.

```
# Frequency rate of Black perpetrators
FREQ_BLACK = c(
    data_VIC_RACE_BLACK_06_prop_table[1, 2],
    data_VIC_RACE_BLACK_07_prop_table[1, 2],
    data_VIC_RACE_BLACK_08_prop_table[1, 2],
    data_VIC_RACE_BLACK_09_prop_table[1, 2],
    data_VIC_RACE_BLACK_10_prop_table[1, 2],
    data_VIC_RACE_BLACK_11_prop_table[1, 2],
    data_VIC_RACE_BLACK_12_prop_table[1, 2],
    data_VIC_RACE_BLACK_13_prop_table[1, 2],
    data_VIC_RACE_BLACK_14_prop_table[1, 2],
    data_VIC_RACE_BLACK_15_prop_table[1, 2],
    data_VIC_RACE_BLACK_16_prop_table[1, 2],
    data_VIC_RACE_BLACK_17_prop_table[1, 2],
    data_VIC_RACE_BLACK_18_prop_table[1, 2],
    data_VIC_RACE_BLACK_19_prop_table[1, 2],
    data_VIC_RACE_BLACK_20_prop_table[1, 2],
    data_VIC_RACE_BLACK_21_prop_table[1, 2],
    data_VIC_RACE_BLACK_22_prop_table[1, 2],
    data_VIC_RACE_BLACK_23_prop_table[1, 2])
```

```
# Frequency rate of White-Hispanic perpetrators
FREQ_WHITE_HISP = c(
    data_VIC_RACE_BLACK_06_prop_table[2, 2],
    data_VIC_RACE_BLACK_07_prop_table[2, 2],
    data_VIC_RACE_BLACK_08_prop_table[2, 2],
    data_VIC_RACE_BLACK_09_prop_table[2, 2],
    data_VIC_RACE_BLACK_10_prop_table[2, 2],
    data_VIC_RACE_BLACK_11_prop_table[2, 2],
    data_VIC_RACE_BLACK_12_prop_table[2, 2],
    data_VIC_RACE_BLACK_13_prop_table[2, 2],
    data_VIC_RACE_BLACK_14_prop_table[3, 2],
    data_VIC_RACE_BLACK_15_prop_table[2, 2],
    data_VIC_RACE_BLACK_16_prop_table[2, 2],
```

```
    data_VIC_RACE_BLACK_17_prop_table[2, 2],
    data_VIC_RACE_BLACK_18_prop_table[2, 2],
    data_VIC_RACE_BLACK_19_prop_table[2, 2],
    data_VIC_RACE_BLACK_20_prop_table[2, 2],
    data_VIC_RACE_BLACK_21_prop_table[2, 2],
    data_VIC_RACE_BLACK_22_prop_table[2, 2],
    data_VIC_RACE_BLACK_23_prop_table[2, 2])
```

```
# Frequency rate of Black-Hispanic perpetrators

FREQ_BLACK_HISP = c(
    data_VIC_RACE_BLACK_06_prop_table[3, 2],
    data_VIC_RACE_BLACK_07_prop_table[3, 2],
    data_VIC_RACE_BLACK_08_prop_table[3, 2],
    data_VIC_RACE_BLACK_09_prop_table[3, 2],
    data_VIC_RACE_BLACK_10_prop_table[3, 2],
    data_VIC_RACE_BLACK_11_prop_table[3, 2],
    data_VIC_RACE_BLACK_12_prop_table[3, 2],
    data_VIC_RACE_BLACK_13_prop_table[3, 2],
    data_VIC_RACE_BLACK_14_prop_table[2, 2],
    data_VIC_RACE_BLACK_15_prop_table[3, 2],
    data_VIC_RACE_BLACK_16_prop_table[3, 2],
    data_VIC_RACE_BLACK_17_prop_table[3, 2],
    data_VIC_RACE_BLACK_18_prop_table[3, 2],
    data_VIC_RACE_BLACK_19_prop_table[3, 2],
    data_VIC_RACE_BLACK_20_prop_table[3, 2],
    data_VIC_RACE_BLACK_21_prop_table[3, 2],
    data_VIC_RACE_BLACK_22_prop_table[3, 2],
    data_VIC_RACE_BLACK_23_prop_table[3, 2])
```

```
# Frequency rate of White perpetrators
FREQ_WHITE = c(
    data_VIC_RACE_BLACK_06_prop_table[5, 2],
    data_VIC_RACE_BLACK_07_prop_table[4, 2],
    data_VIC_RACE_BLACK_08_prop_table[5, 2],
    data_VIC_RACE_BLACK_09_prop_table[5, 2],
    NA,
    data_VIC_RACE_BLACK_11_prop_table[4, 2],
    data_VIC_RACE_BLACK_12_prop_table[5, 2],
    data_VIC_RACE_BLACK_13_prop_table[5, 2],
    data_VIC_RACE_BLACK_14_prop_table[5, 2],
    data_VIC_RACE_BLACK_15_prop_table[5, 2],
    data_VIC_RACE_BLACK_16_prop_table[5, 2],
    data_VIC_RACE_BLACK_17_prop_table[4, 2],
    data_VIC_RACE_BLACK_18_prop_table[4, 2],
    data_VIC_RACE_BLACK_19_prop_table[4, 2],
    data_VIC_RACE_BLACK_20_prop_table[5, 2],
    data_VIC_RACE_BLACK_21_prop_table[5, 2],
    data_VIC_RACE_BLACK_22_prop_table[4, 2],
    data_VIC_RACE_BLACK_23_prop_table[4, 2])
```

```r
# Frequency rate of Asian perpetrators
FREQ_ASIAN = c(
    data_VIC_RACE_BLACK_06_prop_table[4, 2],
    NA,
    data_VIC_RACE_BLACK_08_prop_table[4, 2],
    data_VIC_RACE_BLACK_09_prop_table[4, 2],
    NA,
    data_VIC_RACE_BLACK_11_prop_table[5, 2],
    data_VIC_RACE_BLACK_12_prop_table[4, 2],
    data_VIC_RACE_BLACK_13_prop_table[4, 2],
    data_VIC_RACE_BLACK_14_prop_table[4, 2],
    data_VIC_RACE_BLACK_15_prop_table[4, 2],
    data_VIC_RACE_BLACK_16_prop_table[4, 2],
    NA,
    NA,
    data_VIC_RACE_BLACK_19_prop_table[5, 2],
    data_VIC_RACE_BLACK_20_prop_table[4, 2],
    data_VIC_RACE_BLACK_21_prop_table[4, 2],
    data_VIC_RACE_BLACK_22_prop_table[5, 2],
    data_VIC_RACE_BLACK_23_prop_table[5, 2])
```

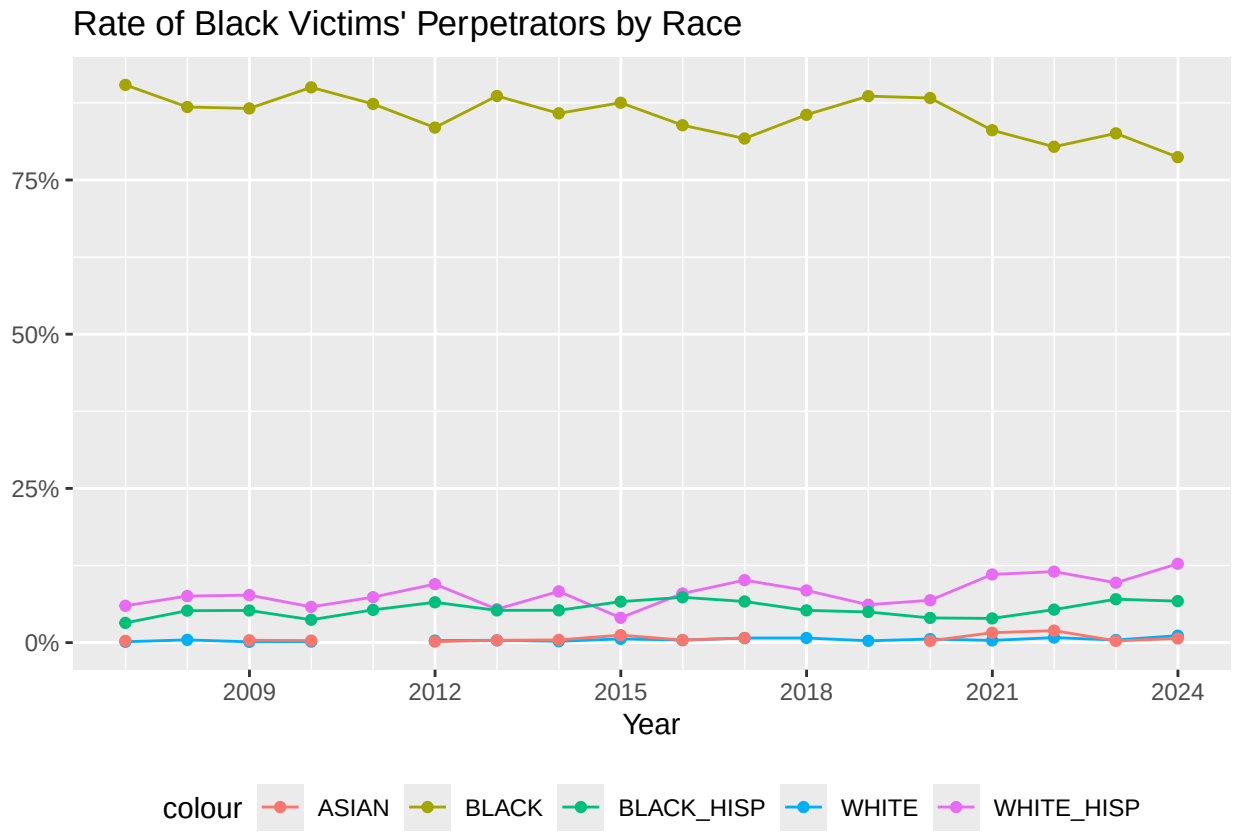
```r
# Set vector for plot x-axis
x = c(as.Date("2006-12-31"),
      as.Date("2007-12-31"),
      as.Date("2008-12-31"),
      as.Date("2009-12-31"),
      as.Date("2010-12-31"),
      as.Date("2011-12-31"),
      as.Date("2012-12-31"),
      as.Date("2013-12-31"),
      as.Date("2014-12-31"),
      as.Date("2015-12-31"),
      as.Date("2016-12-31"),
      as.Date("2017-12-31"),
      as.Date("2018-12-31"),
      as.Date("2019-12-31"),
      as.Date("2020-12-31"),
      as.Date("2021-12-31"),
      as.Date("2022-12-31"),
      as.Date("2023-12-31"))

# Create plot for "Rate of Black Victims' Perpetrators by Race"
ggplot() +
    geom_point(aes(x, y = FREQ_BLACK, color = "BLACK")) +
    geom_line(aes(x, y = FREQ_BLACK, color = "BLACK")) +
    geom_point(aes(x, y = FREQ_WHITE_HISP, color = "WHITE_HISP")) +
    geom_line(aes(x, y = FREQ_WHITE_HISP, color = "WHITE_HISP")) +
    geom_point(aes(x, y = FREQ_BLACK_HISP, color = "BLACK_HISP")) +
    geom_line(aes(x, y = FREQ_BLACK_HISP, color = "BLACK_HISP")) +
    geom_point(aes(x, y = FREQ_WHITE, color = "WHITE")) +
    geom_line(aes(x, y = FREQ_WHITE, color = "WHITE")) +
    geom_point(aes(x, y = FREQ_ASIAN, color = "ASIAN")) +
    geom_line(aes(x, y = FREQ_ASIAN, color = "ASIAN")) +
```

```
    scale_x_date(breaks = "3 years", minor_breaks = "1 year", date_labels = "%Y",
                 name = "Year") +
    scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
    theme(legend.position = "bottom", axis.text.test_date = element_text(angle = 60)) +
    labs(title = "Rate of Black Victims' Perpetrators by Race", y = NULL)
```

```
## Warning in plot_theme(plot): The ‘axis.text.test_date‘ theme element is not defined in the element
## hierarchy.
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



Scroll back up to the Visualization 3 to see the original proportions we found for Black victims and their perpetrators by race. We'll use these as our predicted values for our data if we were to project values through the end of 2024 going into 2025.

```
# Set vector for x-axis with prediction
x = c(as.Date("2006-12-31"),
      as.Date("2007-12-31"),
      as.Date("2008-12-31"),
      as.Date("2009-12-31"),
```

```
      as.Date("2010-12-31"),
      as.Date("2011-12-31"),
      as.Date("2012-12-31"),
      as.Date("2013-12-31"),
      as.Date("2014-12-31"),
      as.Date("2015-12-31"),
      as.Date("2016-12-31"),
      as.Date("2017-12-31"),
      as.Date("2018-12-31"),
      as.Date("2019-12-31"),
      as.Date("2020-12-31"),
      as.Date("2021-12-31"),
      as.Date("2022-12-31"),
      as.Date("2023-12-31"),
      as.Date("2024-12-31"))

FREQ_BLACK_pred = data_VIC_RACE_BLACK_prop_table[1, 2]
FREQ_WHITE_HISP_pred = data_VIC_RACE_BLACK_prop_table[2, 2]
FREQ_BLACK_HISP_pred = data_VIC_RACE_BLACK_prop_table[3, 2]
FREQ_ASIAN_pred = data_VIC_RACE_BLACK_prop_table[4, 2]
FREQ_WHITE_pred = data_VIC_RACE_BLACK_prop_table[5, 2]
```

We created a generalized linear model for each subset of perpetrators by race for Black victims. We showed the code for Black perpetrators, but we changed `BLACK` to `WHITE_HISP`, `BLACK_HISP`, `ASIAN`, and `WHITE` to get the generalized linear models for all the other races.

```
# Black perpetrators, with prediction
df_FREQ_BLACK = data.frame(FREQ_BLACK, FREQ_BLACK_pred)
glm_FREQ_BLACK = glm(FREQ_BLACK ~ FREQ_BLACK_pred, data = df_FREQ_BLACK)
glm_FREQ_BLACK
```

```
##
## Call:  glm(formula = FREQ_BLACK ~ FREQ_BLACK_pred, data = df_FREQ_BLACK)
##
## Coefficients:
##     (Intercept)  FREQ_BLACK_pred
##          0.8552               NA
##
## Degrees of Freedom: 17 Total (i.e. Null);  17 Residual
## Null Deviance:       0.019
## Residual Deviance: 0.019     AIC: -68.29
```

```
##
## Call:  glm(formula = FREQ_WHITE_HISP ~ FREQ_WHITE_HISP_pred, data = df_FREQ_WHITE_HISP)
##
## Coefficients:
##          (Intercept)  FREQ_WHITE_HISP_pred
##               0.0812                    NA
##
## Degrees of Freedom: 17 Total (i.e. Null);  17 Residual
## Null Deviance:       0.009105
## Residual Deviance: 0.009105  AIC: -81.53
```

```
##
## Call:  glm(formula = FREQ_BLACK_HISP ~ FREQ_BLACK_HISP_pred, data = df_FREQ_BLACK_HISP)
##
## Coefficients:
##          (Intercept)  FREQ_BLACK_HISP_pred
##              0.05415                    NA
##
## Degrees of Freedom: 17 Total (i.e. Null);  17 Residual
## Null Deviance:       0.00249
## Residual Deviance: 0.00249   AIC: -104.9
```

```
##
## Call:  glm(formula = FREQ_ASIAN ~ FREQ_ASIAN_pred, data = df_FREQ_ASIAN)
##
## Coefficients:
##     (Intercept)  FREQ_ASIAN_pred
##        0.006493               NA
##
## Degrees of Freedom: 13 Total (i.e. Null);  13 Residual
##   (4 observations deleted due to missingness)
## Null Deviance:       0.0003916
## Residual Deviance: 0.0003916     AIC: -103.1
```

```
##
## Call:  glm(formula = FREQ_WHITE ~ FREQ_WHITE_pred, data = df_FREQ_WHITE)
##
## Coefficients:
##     (Intercept)  FREQ_WHITE_pred
##        0.004623               NA
##
## Degrees of Freedom: 16 Total (i.e. Null);  16 Residual
##   (1 observation deleted due to missingness)
## Null Deviance:       0.0001186
## Residual Deviance: 0.0001186     AIC: -149.6
```

After running the generalized model, we obtained the following residual deviance scores:

- **Black:** 0.019
- **White Hisp:** 0.009105
- **Black Hisp:** 0.00249
- **Asian:** 0.0003916
- **White:** 0.0001186

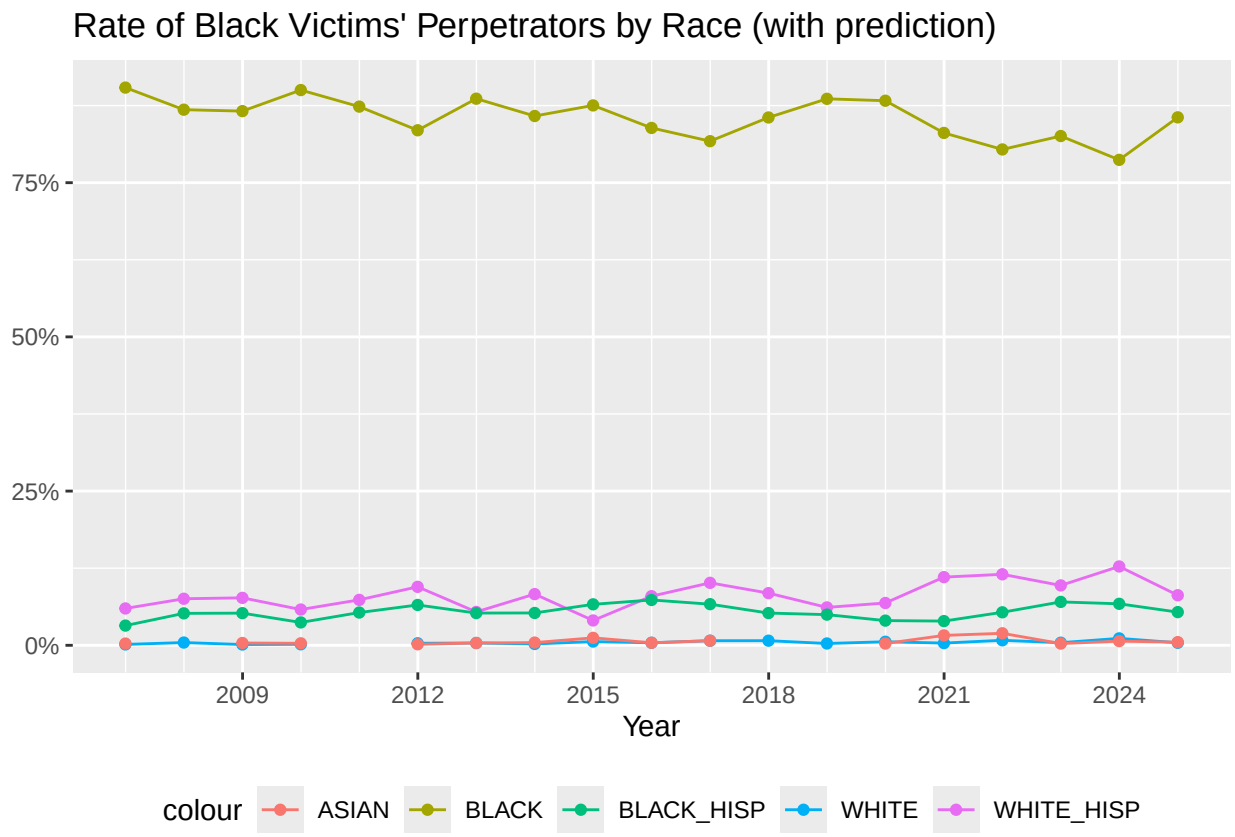
All our residual deviance scores were small, so we can use the frequency rates of perpetrators by race for Black victims as a reasonable model to predict future rates of perpetrators. We added one last visualization showing the projected rates.

```r
FREQ_BLACK_final = append(FREQ_BLACK, FREQ_BLACK_pred)
FREQ_WHITE_HISP_final = append(FREQ_WHITE_HISP, FREQ_WHITE_HISP_pred)
FREQ_BLACK_HISP_final = append(FREQ_BLACK_HISP, FREQ_BLACK_HISP_pred)
FREQ_ASIAN_final = append(FREQ_ASIAN, FREQ_ASIAN_pred)
FREQ_WHITE_final = append(FREQ_WHITE, FREQ_WHITE_pred)

ggplot() +
    geom_point(aes(x, y = FREQ_BLACK_final, color = "BLACK")) +
    geom_line(aes(x, y = FREQ_BLACK_final, color = "BLACK")) +
    geom_point(aes(x, y = FREQ_WHITE_HISP_final, color = "WHITE_HISP")) +
    geom_line(aes(x, y = FREQ_WHITE_HISP_final, color = "WHITE_HISP")) +
    geom_point(aes(x, y = FREQ_BLACK_HISP_final, color = "BLACK_HISP")) +
    geom_line(aes(x, y = FREQ_BLACK_HISP_final, color = "BLACK_HISP")) +
    geom_point(aes(x, y = FREQ_WHITE_final, color = "WHITE")) +
    geom_line(aes(x, y = FREQ_WHITE_final, color = "WHITE")) +
    geom_point(aes(x, y = FREQ_ASIAN_final, color = "ASIAN")) +
    geom_line(aes(x, y = FREQ_ASIAN_final, color = "ASIAN")) +
    scale_x_date(breaks = "3 years", minor_breaks = "1 year", date_labels = "%Y",
                 name = "Year") +
    scale_y_continuous(labels = scales::percent_format( accuracy = 1)) +
    theme(legend.position = "bottom", axis.text.test_date = element_text(angle = 60)) +
    labs(title = "Rate of Black Victims' Perpetrators by Race (with prediction)", y = NULL)
```

```
## Warning in plot_theme(plot): The ‘axis.text.test_date‘ theme element is not defined in the element
## hierarchy.
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



Additional Questions Since this analysis only covered this given dataset, we did not include information on population rates within New York City. Does this data reflect the demographic proportions of this population? Does it reflect the socioeconomic statistics of this population?

Project Step 4: Add Bias Identification

Conclusion Our analysis shows that for Black victims of NYPD shooting incidents, frequency rates of perpetrators by race have stayed relatively stable since 2006, but the majority of perpetrators has consistently been Black.

Possible Sources of Bias Based on recent events like the killing of George Floyd in 2020 and the subsequent Black Lives Matter protests, the media makes it appear that the majority of Black victims of police violence are attacked by White perpetrators. This data shows the opposite.

Further analysis should be done to see why the majority of perpetrators for Black victims of police shooting incidents were also Black.