~~~
---
title: "NYPD Shooting"
date: "2023-02-01"
output:
  pdf_document: default
  html_document: default
---
### Load Libraries
First I will load the any necessary libraries

```{r}
library(tidyverse)
library(lubridate)
```

### Importing and Reading Data
Next, I will be importing the data through the URL and then reading the CSV file from it.

```{r}
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_shooting <- read_csv(url_in[1])
```

### View Dataset
```{r}
nypd_shooting
```

### Summary
```{r}
summary(nypd_shooting)
```

### Cleaning Data
Upon viewing the columns, it seems that many columns are not needed which include:
PRECINCT, JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, and Lon_Lat
- PRECINCT & JURISDICTION_CODE would not add much since we already have BORO and exact
latitude & longitude
- LOCATION_DESC has many missing values and we also have columns which describe location
(lat, long)
- X_COORD_CD & Y_COORD_CD are not needed since we have lat & long
- Lon_Lat is redundant because we have lat & long separately
```{r}
nypd_shooting_2 <- nypd_shooting %>%
  select(-c(PRECINCT, JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Lon_Lat))

nypd_shooting_2
```

View NA's
```{r}
apply(nypd_shooting_2, 2, function(x) any(is.na(x)))
```

After eliminating columns, it can be seen that many columns still have missing values and
it's important to handle it carefully.

Columns with missing values (N/A's) include: PERP_AGE_GROUP, PERP_SEX, PERP_RACE

I think for missing values, I will not replace them with a value that exists in the field.
I will likely replace an "N/A" with an unknown value. If for example, "PERP_AGE_GROUP" is
N/A, I will likely replace that N/A with "unknown_age_group" and use that for analysis.

### Step 2: Tidy & Transform Data
~~~

~~~
Replace NA's:
```{r}
nypd_shooting_2 <- nypd_shooting_2 %>%
  replace_na(list(PERP_AGE_GROUP = "UNKNOWN", PERP_SEX = "UNKNOWN", PERP_RACE =
"UNKNOWN"))

nypd_shooting_2
summary(nypd_shooting_2)
```

Now it can be seen that no column has missing values.

Next I want to find unique values to see if there are any outliers.
#### Find Unique values:
```{r}
unique(nypd_shooting_2$BORO)
unique(nypd_shooting_2$STATISTICAL_MURDER_FLAG)
unique(nypd_shooting_2$PERP_AGE_GROUP)
unique(nypd_shooting_2$PERP_SEX)
unique(nypd_shooting_2$PERP_RACE)
unique(nypd_shooting_2$VIC_AGE_GROUP)
unique(nypd_shooting_2$VIC_SEX)
unique(nypd_shooting_2$VIC_RACE)
```

Based on the unique() function, PERP_AGE_GROUP has some odd values which should be
excluded:
```{r}
nypd_shooting_2 <- filter(nypd_shooting_2,
                          PERP_AGE_GROUP != "1020" & PERP_AGE_GROUP != "940" &
PERP_AGE_GROUP != "224")
unique(nypd_shooting_2$PERP_AGE_GROUP)
```

Make sure that all unknowns are the same
```{r}
nypd_shooting_2$PERP_SEX[nypd_shooting_2$PERP_SEX == "U"] <- "UNKNOWN"
summary(nypd_shooting_2)
```

### Step 3: Add Visualizations and Analysis
First, I will create a bar chart to see the number of shooting incidents in each borough:
```{r}
nypd_shooting_2 %>%
  ggplot(aes(x = BORO)) +
  geom_bar(fill = "blue") +
  ggtitle("Number of Shooting Incidents by Borough") +
  xlab("Borough") +
  ylab("Number of Incidents")
```
Based on this plot, it can be seen that Brooklyn has the most number of shooting incidents
followed by Bronx. Staten Island has the least number of incidents.

#### Next, I will create a histogram with the number of shootings per victim's age group.
```{r}
ggplot(nypd_shooting_2, aes(x = VIC_AGE_GROUP)) +
  geom_bar(binwidth = 1) +
  xlab("Victim's Age Group") +
  ylab("Count") +
  ggtitle("Victim's Age Group Histogram")
```
Based on this plot, it can be seen that the most victims are between 25 to 55, with ages
18-24 being second.
~~~

#### Modeling:
For this analysis, I will be using logistic regression to to predict the STATISTICAL_MURDER_FLAG using BORO, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, Latitude, and Longitude as predictors:
```{r}
nypd_shooting_2$PERP_AGE_GROUP = as.factor(nypd_shooting_2$PERP_AGE_GROUP)
nypd_shooting_2$PERP_SEX = as.factor(nypd_shooting_2$PERP_SEX)
nypd_shooting_2$PERP_RACE = as.factor(nypd_shooting_2$PERP_RACE)
# Fit a logistic regression model
model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + PERP_AGE_GROUP + PERP_SEX + PERP_RACE +
Latitude + Longitude, data = nypd_shooting_2, family = binomial())

# Print the model summary
summary(model)
```

From the summary, we can see that Pr(>|z|) column represents the p-value associated with the value in the z value column. If the p-value is less than 0.05, it means that the predictor variable has a statistically significant relationship with the response variable. Some of these predictor variables having significant relationship with STATISTICAL_MURDER_FLAG are PERP_AGE_GROUP18-24, PERP_AGE_GROUP25-44, PERP_AGE_GROUP45-64, and PERP_AGE_GROUP65+.

### Step 4: Bias
Some possible sources of bias might include under-reporting. This means that perhaps not all incidents have been reported to he police. Furthermore, this data relies on reporters and those who have recorded such information. It is possible that there are errors in the reporting itself.

With regards to my personal bias, I think this dataset does confirm some of my beliefs. I view New York city as having a high crime rate and was not too surprised to see that Brooklyn and Bronx have the highest number of incidents. I was also not surprised to see Staten Island having lower incidents compared to Brooklyn and Bronx. Secondly, I am also not too surprised to see that the most number of victims are between the age 25 to 44. I am very surprised to see victim's under the age of 18. It's disheartening to see young people be victims of such incidents.