Data Analysis Report on NYPD Shooting Incident

Umair Shaikh

2023-12-10

Introduction

This study looks at NYPD's data on shooting incidents to understand who's most affected by gun violence in New York City. I explore details like where these shootings often happen, if they're deadly, and who the victims and perpetrators usually are, in terms of their age.

My aim is to figure out who is at greater risk of being involved in these incidents, find out where in the city gun violence is most common, and see at what times these shootings typically occur.

By understanding these patterns, I hope to help direct police, community help, and medical aid to the people and places that need them most.

Data Source

Data is in csv format from this site: https://catalog.data.gov/dataset?q=NYPD+Shooting+Incident+Data+%28Historic%29&sort=views_recent+desc&ext_location=&ext_bbox=&ext_prev_extent=

Required Libraries

```
# Load required libraries
library(tidyr)
library(dplyr)
library(ggplot2)
library(viridis)
library(lubridate)
library(tidymodels)
```

Load and Summarize Data

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

# Display the first few rows
head(nypd_data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                       105
```

```
## 2    137471050 06/27/2014   17:40:00    BRONX                          40
## 3    147998800 11/21/2015   03:56:00   QUEENS                         108
## 4    146837977 10/09/2015   18:30:00    BRONX                          44
## 5     58921844 02/19/2009   22:58:00    BRONX                          47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                          81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                  false
## 2                 0                                                  false
## 3                 0                                                   true
## 4                 0                                                  false
## 5                 0                                                   true
## 6                 0                                                   true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                    Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

```
#Generate Summary Statistics of the data set
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
```

```
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  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_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Brief Overview of Structure:

- INCIDENT_KEY: A unique identifier for each incident.
- OCCUR_DATE: The date of the incident.
- OCCUR_TIME: The time of the incident.
- BORO: Borough where the incident occurred.
- LOC_OF_OCCUR_DESC: Description of the location of occurrence (if available).
- PRECINCT: NYPD precinct where the incident occurred.
- JURISDICTION_CODE: Jurisdiction code.
- LOC_CLASSFCTN_DESC: Classification description of the location (if available).
- LOCATION_DESC: Detailed description of the location (if available).
- STATISTICAL_MURDER_FLAG: Indicates if the incident was a statistical murder.
- PERP_AGE_GROUP: Age group of the perpetrator (if known).
- PERP_SEX: Sex of the perpetrator (if known).
- PERP_RACE: Race of the perpetrator (if known).
- VIC_AGE_GROUP: Age group of the victim.
- VIC_SEX: Sex of the victim.
- VIC_RACE: Race of the victim.
- X_COORD_CD, Y_COORD_CD: X and Y coordinates of the incident location.
- Latitude, Longitude: Geographical coordinates of the incident.
- Lon_Lat: Combined longitude and latitude in a point format.

```r
# Convert 'OCCUR_DATE' and 'OCCUR_TIME' columns
nypd_data <- nypd_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME))

# Drop irrelevant columns
columns_to_drop <- c("LOC_OF_OCCUR_DESC", "LOC_CLASSFCTN_DESC", "LOCATION_DESC", "Lon_Lat")
nypd_data <- nypd_data %>%
  select(-all_of(columns_to_drop))

# Replace missing values with 'Unknown' in specific columns
nypd_data <- nypd_data %>%
  mutate(PERP_AGE_GROUP = replace_na(PERP_AGE_GROUP, "Unknown"),
         PERP_SEX = replace_na(PERP_SEX, "Unknown"),
         PERP_RACE = replace_na(PERP_RACE, "Unknown"))

# Remove rows with missing values in 'JURISDICTION_CODE', 'Latitude', and 'Longitude'
nypd_data <- nypd_data %>%
  drop_na(JURISDICTION_CODE, Latitude, Longitude)
head(nypd_data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO PRECINCT JURISDICTION_CODE
## 1    228798151 2021-05-27 21H 30M 0S   QUEENS      105                 0
## 2    137471050 2014-06-27 17H 40M 0S    BRONX       40                 0
## 3    147998800 2015-11-21  3H 56M 0S   QUEENS      108                 0
## 4    146837977 2015-10-09 18H 30M 0S    BRONX       44                 0
## 5     58921844 2009-02-19 22H 58M 0S    BRONX       47                 0
## 6    219559682 2020-10-21 21H 36M 0S BROOKLYN       81                 0
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                   false                                          18-24
## 2                   false                                          18-24
## 3                    true                                          25-44
## 4                   false                                            <18
## 5                    true          25-44        M     BLACK         45-64
## 6                    true                                          25-44
##   VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1       M          BLACK    1058925   180924.0 40.66296 -73.73084
## 2       M          BLACK    1005028   234516.0 40.81035 -73.92494
## 3       M          WHITE    1007668   209836.5 40.74261 -73.91549
## 4       M WHITE HISPANIC    1006537   244511.1 40.83778 -73.91946
## 5       M          BLACK    1024922   262189.4 40.88624 -73.85291
## 6       M          BLACK    1004234   186461.7 40.67846 -73.92795
```

```r
# Summary of the cleaned dataset
summary(nypd_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 63859933   1st Qu.:2009-07-18   1st Qu.:3H 27M 0S
##  Median : 90340495   Median :2013-04-27   Median :15H 11M 30S
##  Mean   :120812778   Mean   :2014-01-05   Mean   :12H 41M 34.298901098904S
##  3rd Qu.:188587325   3rd Qu.:2018-10-08   3rd Qu.:20H 45M 0S
##  Max.   :261190187   Max.   :2022-12-31   Max.   :23H 59M 0S
```

```
##      BORO              PRECINCT      JURISDICTION_CODE STATISTICAL_MURDER_FLAG
##  Length:27300       Min.   :  1.00   Min.   :0.000     Length:27300
##  Class :character   1st Qu.: 44.00   1st Qu.:0.000     Class :character
##  Mode  :character   Median : 68.00   Median :0.000     Mode  :character
##                     Mean   : 65.64   Mean   :0.327
##                     3rd Qu.: 81.00   3rd Qu.:0.000
##                     Max.   :123.00   Max.   :2.000
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:27300       Length:27300       Length:27300       Length:27300
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:27300       Length:27300       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000033   1st Qu.:182832
##  Mode  :character   Mode  :character   Median :1007742   Median :194478
##                                        Mean   :1009451   Mean   :208128
##                                        3rd Qu.:1016838   3rd Qu.:239518
##                                        Max.   :1066815   Max.   :271128
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.92
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
```

Analyze and Visualize Data

Insightful visualizations I considered:

1) Perpetrator and Victim Demographics: This involves creating bar charts comparing the age of perpetrators and victims. This visualization can provide insights into the demographics of those involved in shooting incidents

2) Time of Day Incidents: Analyzing the time of day when shootings occur most frequently can provide insights into when these incidents are more likely to happen. This involves a histogram showing the number of incidents by hour of the day.

3) Number of Shooting Incidents Over Time (Yearly): This bar chart shows the annual number of shooting incidents. It provides a clear view of how the frequency of incidents has changed over the years, highlighting any trends, such as increases or decreases in shooting incidents.

4) Distribution of Incidents Across Boroughs: This bar chart displays the distribution of shooting incidents across different boroughs. It helps in understanding which boroughs have higher incidences of shootings, potentially indicating areas with higher crime rates.

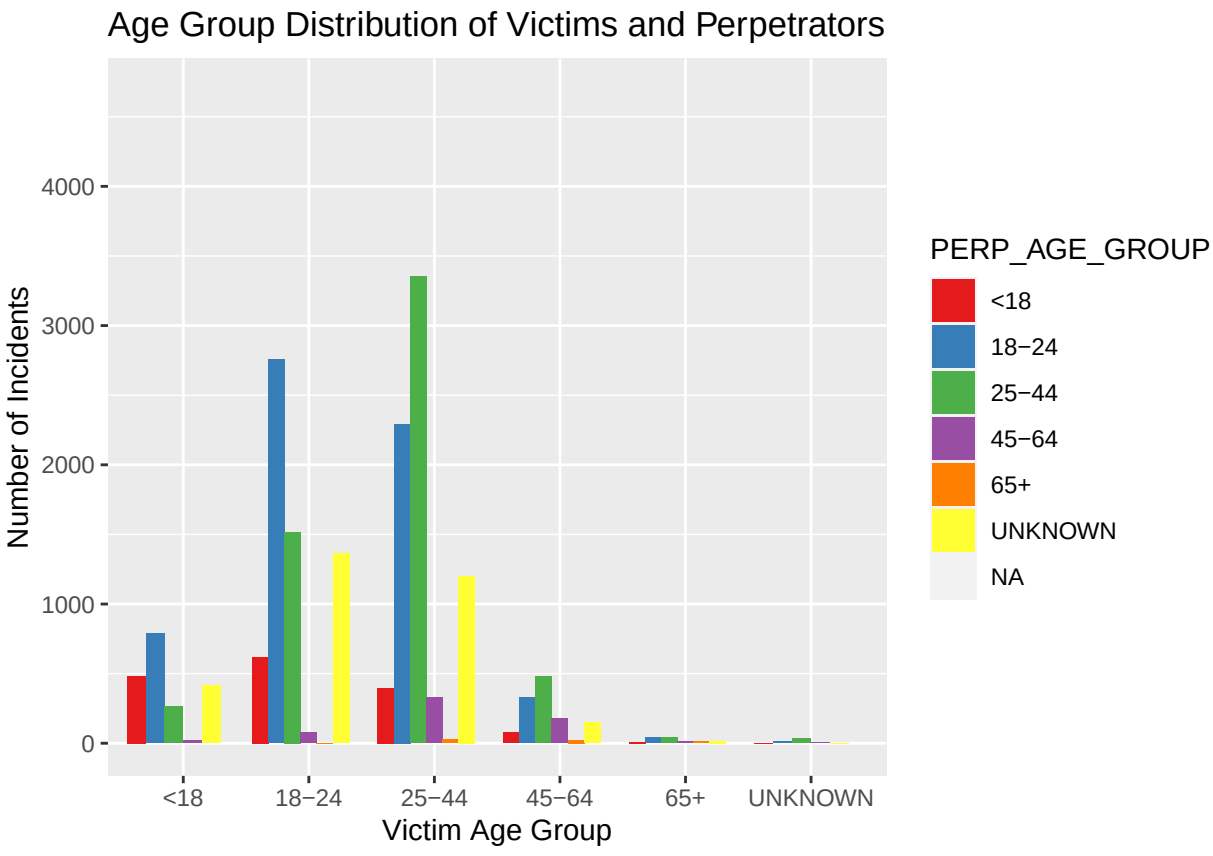
```r
# Age Group Distribution

# Check unique values of VIC_AGE_GROUP
unique(nypd_data$VIC_AGE_GROUP)
```

```
## [1] "18-24"   "25-44"   "<18"     "45-64"   "65+"     "UNKNOWN" "1022"
```

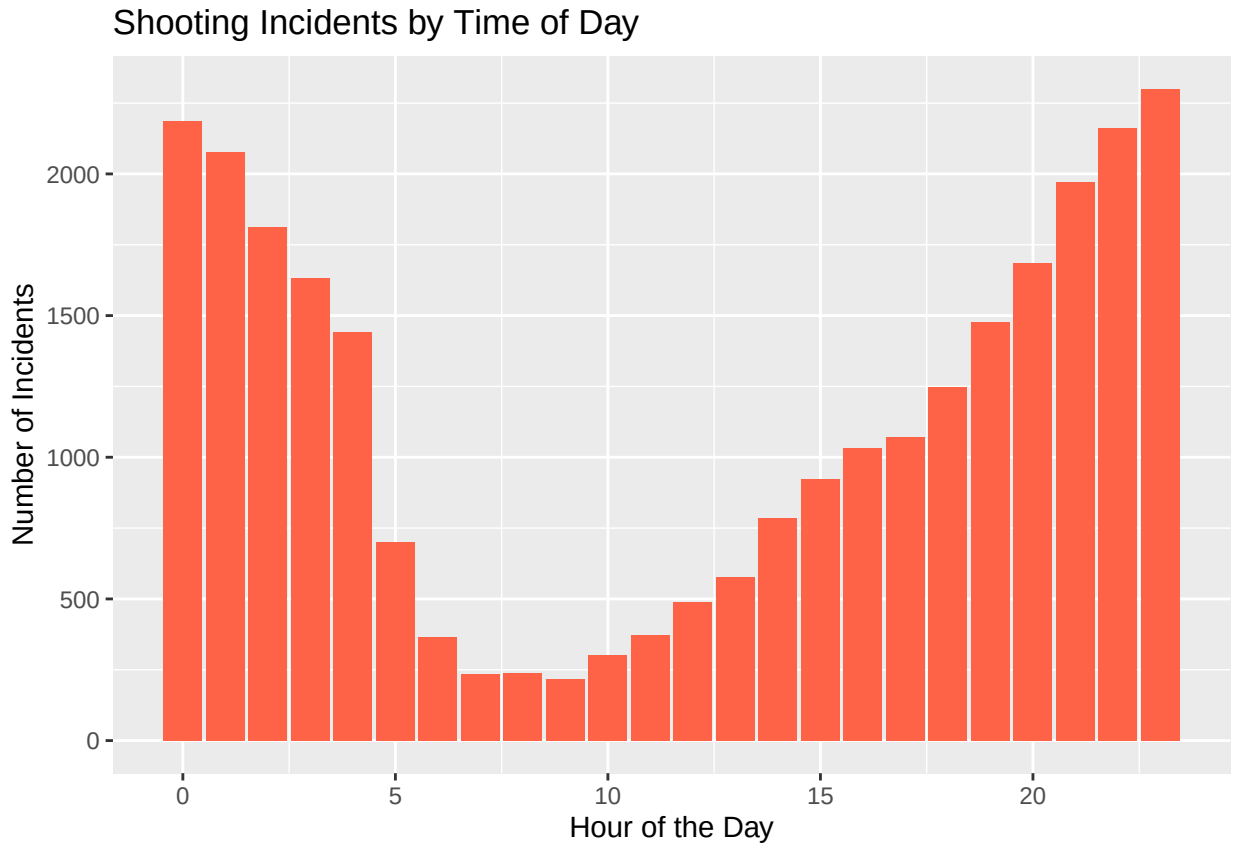
```r
# Redo the age groups (Assuming 1022 was meant to be 18-24 and 224 was meant to be 25-44)
nypd_data <- nypd_data %>%
  mutate(VIC_AGE_GROUP = case_when(
    VIC_AGE_GROUP == "1022" ~ "18-24",
    VIC_AGE_GROUP == "224"  ~ "25-44",
    TRUE                    ~ as.character(VIC_AGE_GROUP)
  ))

# Convert to a factor
nypd_data$VIC_AGE_GROUP <- factor(nypd_data$VIC_AGE_GROUP,
                                  levels = c("<18", "18-24", "25-44", "45-64", "65+", "UNKNOWN"))
nypd_data$PERP_AGE_GROUP <- factor(nypd_data$PERP_AGE_GROUP,
                                   levels = c("<18", "18-24", "25-44", "45-64", "65+", "UNKNOWN"))
# Generate the plot
ggplot(nypd_data, aes(x = VIC_AGE_GROUP, fill = PERP_AGE_GROUP)) +
  geom_bar(position = "dodge") +
  labs(title = "Age Group Distribution of Victims and Perpetrators",
       x = "Victim Age Group",
       y = "Number of Incidents") +
  scale_fill_brewer(palette = "Set1")
```

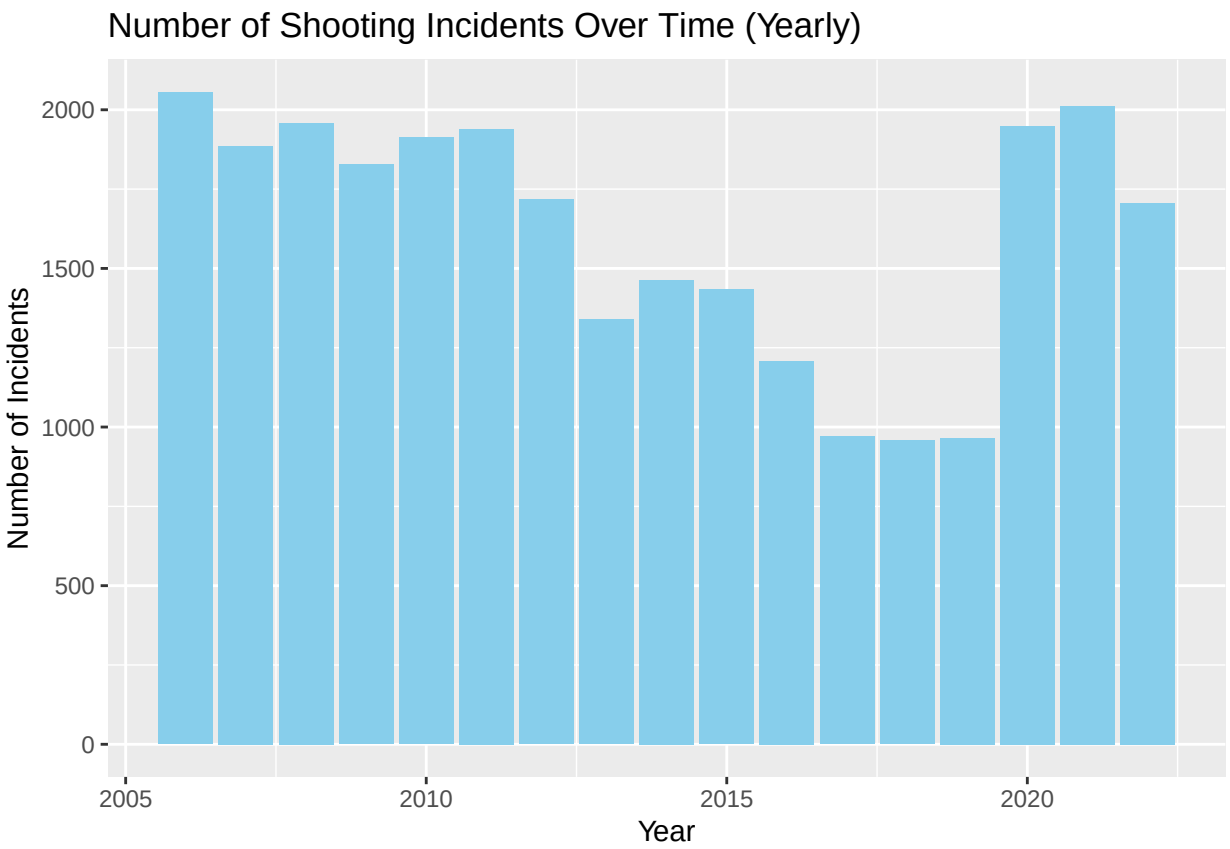



```r
# Shooting Incidents by Time of Day
nypd_data %>%
```

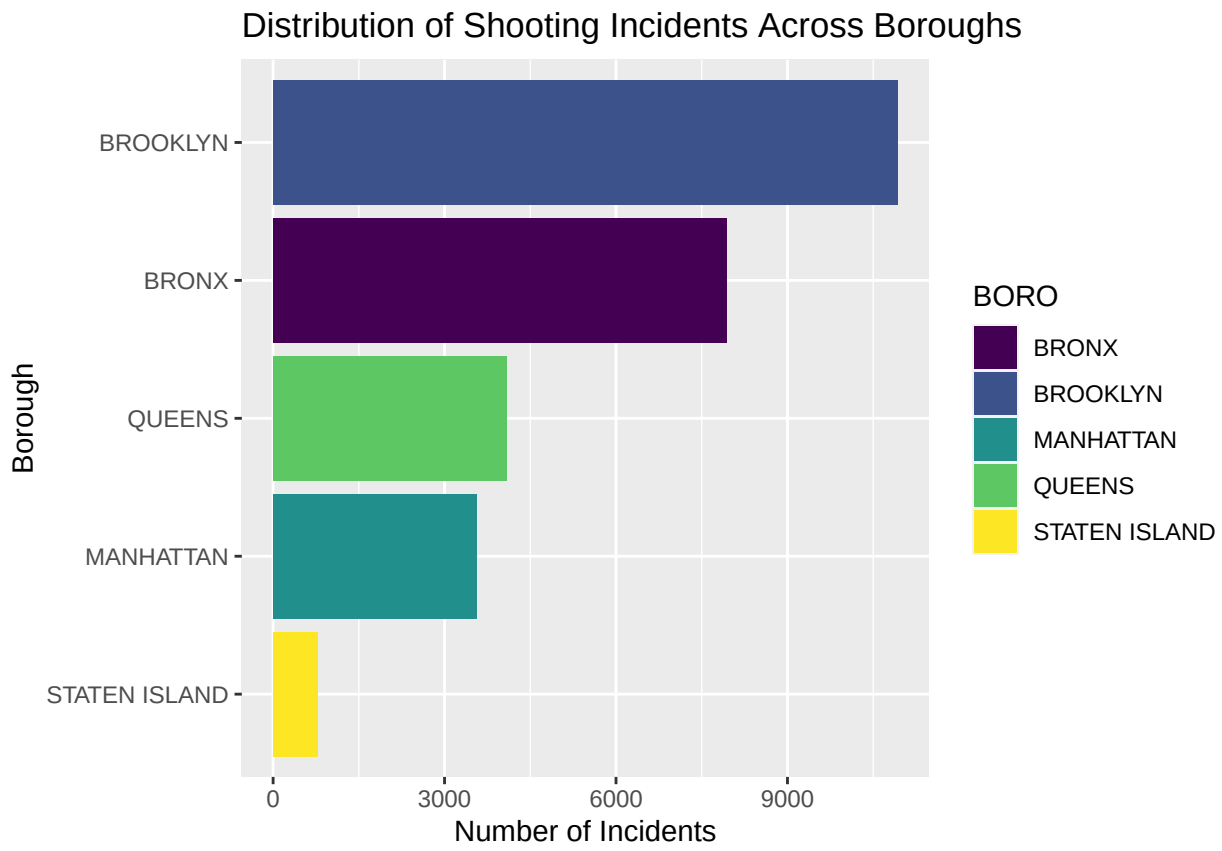
```
  mutate(hour = hour(OCCUR_TIME)) %>%
  count(hour) %>%
  ggplot(aes(x = hour, y = n)) +
  geom_bar(stat = "identity", fill = "tomato") +
  labs(title = "Shooting Incidents by Time of Day", x = "Hour of the Day", y = "Number of Incidents")
```




```
# Number of Shooting Incidents Over Time (Yearly)
nypd_data %>%
  count(Year = year(OCCUR_DATE)) %>%
  ggplot(aes(x = Year, y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Number of Shooting Incidents Over Time (Yearly)", x = "Year", y = "Number of Incidents")
```




```
# Distribution of Incidents Across Boroughs
nypd_data %>%
  count(BORO) %>%
  ggplot(aes(x = reorder(BORO, n), y = n, fill = BORO)) +
  geom_bar(stat = "identity") +
  scale_fill_viridis_d() +
  labs(title = "Distribution of Shooting Incidents Across Boroughs",
       x = "Borough",
       y = "Number of Incidents") +
  coord_flip()
```



Logistic Regression Model

I built a logistic regression model to predict the likelihood of a shooting being fatal based on various factors (e.g., borough, time of day, demographics).

```
# Select columns for the model
model_data <- nypd_data %>%
  select(
    STATISTICAL_MURDER_FLAG,
    VIC_AGE_GROUP,
    BORO,
    PERP_RACE,
    PERP_SEX
  )

# Filter out rows with missing data in any of the selected columns
model_data <- model_data %>%
  filter(
    !is.na(STATISTICAL_MURDER_FLAG),
    !is.na(VIC_AGE_GROUP),
    !is.na(PERP_SEX),
    !is.na(PERP_RACE)
  )

# Convert the outcome variable to a factor
```

```r
model_data$STATISTICAL_MURDER_FLAG <- as.factor(model_data$STATISTICAL_MURDER_FLAG)

# Split the data into training and testing sets (80% training, 20% testing)
set.seed(456)
data_split <- initial_split(model_data, prop = 0.8)
train_data <- training(data_split)
test_data <- testing(data_split)

# Logistic Regression Model
model <- logistic_reg() %>%
         set_engine("glm") %>%
         fit(STATISTICAL_MURDER_FLAG ~ BORO + VIC_AGE_GROUP + PERP_SEX + PERP_RACE, data = train_data)

# Make predictions on the test set
predictions <- predict(model, test_data, type = "class")

# Bind the predictions to the testing set
results <- bind_cols(test_data, predictions)

# Evaluate the model (e.g., using accuracy)
accuracy <- results %>%
            metrics(truth = STATISTICAL_MURDER_FLAG, estimate = .pred_class) %>%
            filter(.metric == "accuracy") %>%
            pull(.estimate) * 100

print(paste("Accuracy:", accuracy, "%"))
```

```
## [1] "Accuracy: 81.4285714285714 %"
```

Conclusions from the NYPD Shooting Incident Data

This analysis showed peaks in shooting incidents during certain times of the day, which could indicate patterns in criminal activity. For instance, higher incident rates at night might suggest a need for increased night patrols or community interventions during these hours.

By examining the victims and perpetrators age groups, we might identify a specific pattern of demographic groups that are more frequently involved in shootings. This could point towards underlying social or economic issues that need addressing, such as youth involvement in violence.

Also, certain boroughs showed higher rates of shooting incidents, and these areas could be identified as hotspots. This suggests a need for targeted interventions in these areas, such as community-based programs, increased policing, or social services.

Over time, if there are trends showing increases or decreases in shooting incidents, these could possibly be correlated with changes in law enforcement strategies, community programs, or social issues.

Identifying Possible Bias

The dataset may not capture all incidents, especially if there are cases of underreporting or misclassification.There could be biases in how incidents are reported or recorded, potentially influenced by the victim's or perpetrator's demographic characteristics.

Personal Bias

I myself don't have any personal assumption on this data set and the data analyst approach should not have any expectations ingrained about patterns. This helps to avoid biased interpretation of the data.