NYPD Shooting Incident Data

Vatsal

2024-08-10

This R Markdown report is about the NYPD shooting incident dataset from data.gov.It has data about shooting incidents in New York split according to various categories like boroughs, time of day, location, race etc. This is the data source: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv

Importing data:

Import the data from the URL

```
data_source_url <- ("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv")
nypd_shooting_incidents_historical <- read.csv(data_source_url)
```

Tidying the imported data:

We will remove the attributes for lat, long and x and y coordinates as we don't need it for now

```
tidy_nypd_shooting_data <- nypd_shooting_incidents_historical %>%
  select(-X_COORD_CD, -Y_COORD_CD, -Latitude, -Longitude, -Lon_Lat)
```

Transforming the imported data:

We will first check the structure of the cleaned data and then do the following transformations:

1. According to this the date and time columns are "chr" which we will need to transform to date and time respectively.
2. Loc_of_occur_desc and statistical_murder_flag can be a "factor" type but we can change the name to be more intuitive

```
str(tidy_nypd_shooting_data)
```

```
## 'data.frame':    28562 obs. of  16 variables:
##  $ INCIDENT_KEY           : int  244608249 247542571 84967535 202853370 27078636 230311078 229224142
##  $ OCCUR_DATE             : chr  "05/05/2022" "07/04/2022" "05/27/2012" "09/24/2019" ...
##  $ OCCUR_TIME             : chr  "00:10:00" "22:20:00" "19:35:00" "21:00:00" ...
##  $ BORO                   : chr  "MANHATTAN" "BRONX" "QUEENS" "BRONX" ...
##  $ LOC_OF_OCCUR_DESC      : chr  "INSIDE" "OUTSIDE" "" "" ...
##  $ PRECINCT               : int  14 48 103 42 83 23 113 77 48 49 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 2 0 0 0 0 ...
```

```
##  $ LOC_CLASSFCTN_DESC     : chr  "COMMERCIAL" "STREET" "" "" ...
##  $ LOCATION_DESC          : chr  "VIDEO STORE" "(null)" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "true" "true" "false" "false" ...
##  $ PERP_AGE_GROUP         : chr  "25-44" "(null)" "" "25-44" ...
##  $ PERP_SEX               : chr  "M" "(null)" "" "M" ...
##  $ PERP_RACE              : chr  "BLACK" "(null)" "" "UNKNOWN" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "18-24" "18-24" "25-44" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
```

```
tidy_nypd_shooting_data <- tidy_nypd_shooting_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE), OCCUR_TIME = hms(OCCUR_TIME))
tidy_nypd_shooting_data <- tidy_nypd_shooting_data %>%
  rename(
    Location_Description = LOC_OF_OCCUR_DESC,
    Is_Murder = STATISTICAL_MURDER_FLAG
  ) %>%
  mutate(
    Location_Description = factor(Location_Description),
    Is_Murder = factor(Is_Murder, levels = c("true", "false"))
  )
```

Descriptive analysis:

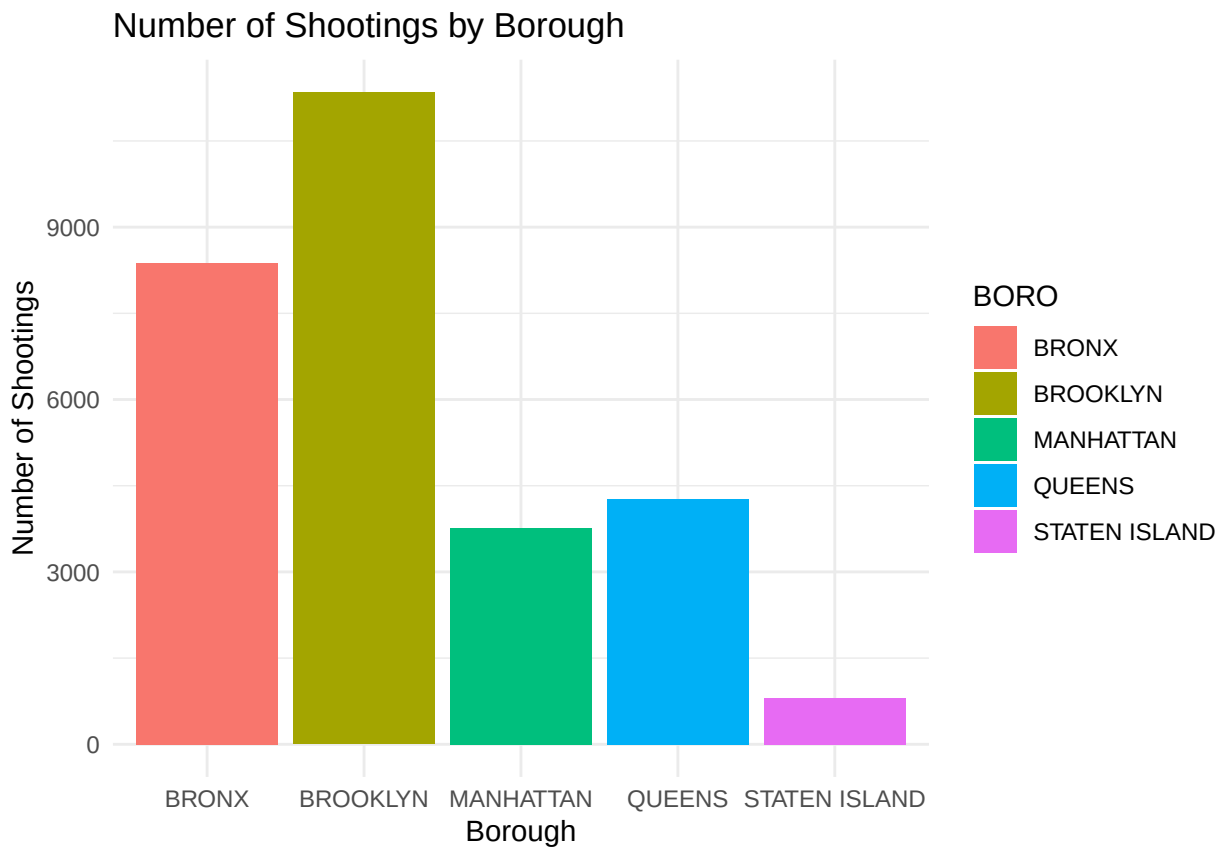
1. Frequency of Shootings by Borough: This shows the number of shootings per borough to identify which areas have higher incidents of violence. This can help in understanding geographical patterns and possibly infer reasons for these patterns.

```
shootings_by_borough <- tidy_nypd_shooting_data %>%
  count(BORO, sort = TRUE)
shootings_by_borough
```

```
##            BORO     n
## 1      BROOKLYN 11346
## 2         BRONX  8376
## 3        QUEENS  4271
## 4     MANHATTAN  3762
## 5 STATEN ISLAND   807
```

According to this Brooklyn has the highest incidents of shooting while Staten Island has the lowest. Here's a visual representation:

```
ggplot(data = shootings_by_borough, aes(x = BORO, y = n, fill = BORO)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of Shootings by Borough", x = "Borough", y = "Number of Shootings") +
  theme_minimal()
```



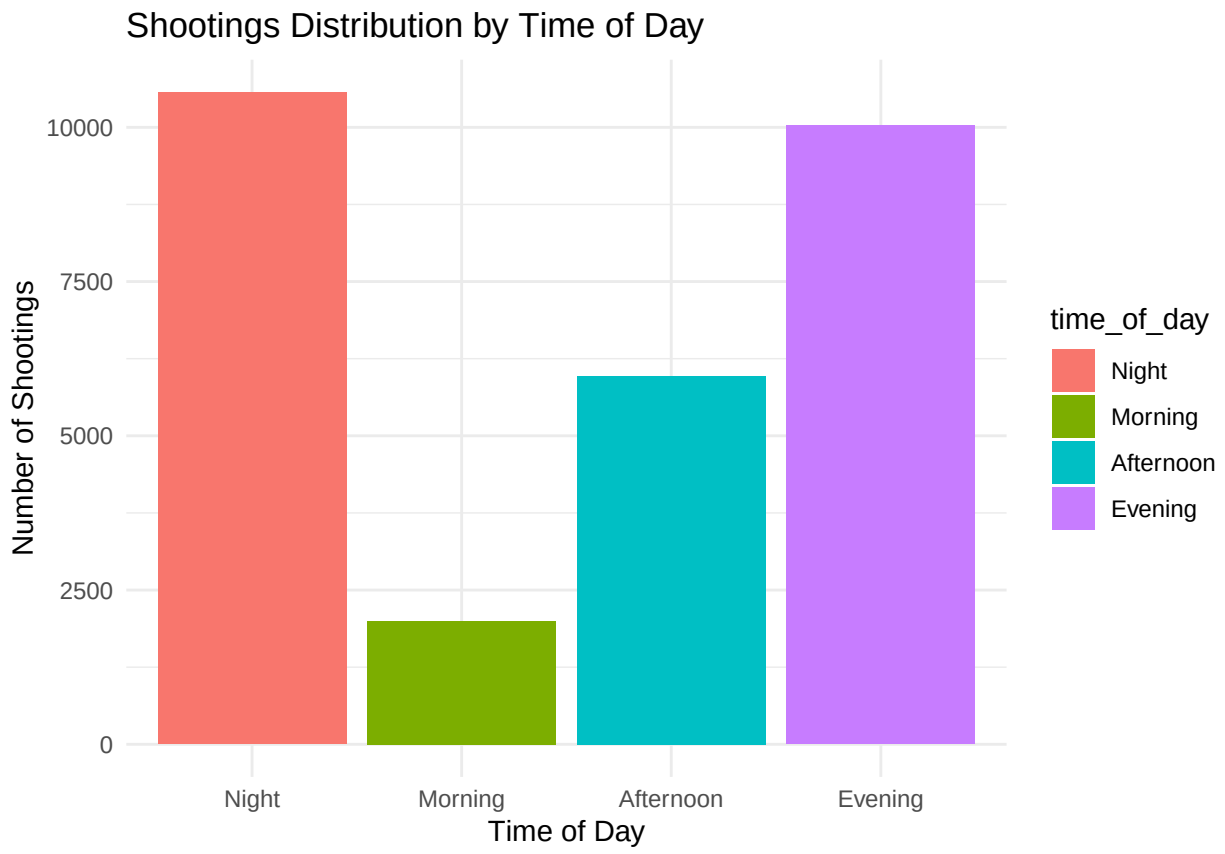
2. Analysis of Shootings by Time of Day: This shows shootings by different time blocks (morning, afternoon, evening, and night) to see if there's a specific time when shootings are more likely to occur. This could be useful for law enforcement to allocate resources more effectively.

```
tidy_nypd_shooting_data$hour <- hour(tidy_nypd_shooting_data$OCCUR_TIME)
shootings_by_time <- tidy_nypd_shooting_data %>%
  mutate(time_of_day = cut(hour, breaks = c(0, 6, 12, 18, 24),
                           labels = c("Night", "Morning", "Afternoon", "Evening"),
                           include.lowest = TRUE)) %>%
  count(time_of_day, sort = TRUE)
shootings_by_time
```

```
##   time_of_day     n
## 1       Night 10567
## 2     Evening 10027
## 3   Afternoon  5966
## 4     Morning  2002
```

So it seems like most of the shootings happen post afternoon (evening and night) so maybe it is worth looking into allocating more resources during that time. Here is a visual analysis for this:

```
ggplot(data = shootings_by_time, aes(x = time_of_day, y = n, fill = time_of_day)) +
  geom_bar(stat = "identity") +
  labs(title = "Shootings Distribution by Time of Day", x = "Time of Day", y = "Number of Shootings") +
  theme_minimal()
```



3. Murder Flag Analysis: This counts how many shootings were flagged as murders (Is_Murder) to see the severity of incidents.

```
murder_analysis <- tidy_nypd_shooting_data %>%
  count(Is_Murder, sort = TRUE)
murder_analysis
```

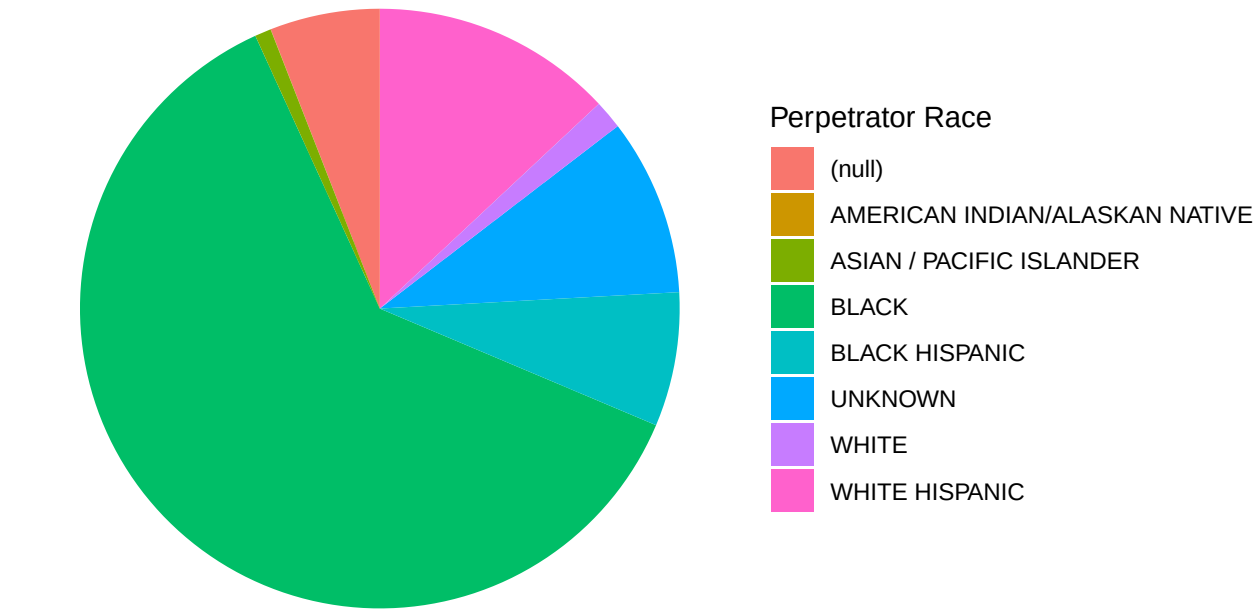
```
##   Is_Murder     n
## 1     false 23036
## 2      true  5526
```

A high number of these shootings (majority) ended in murder.

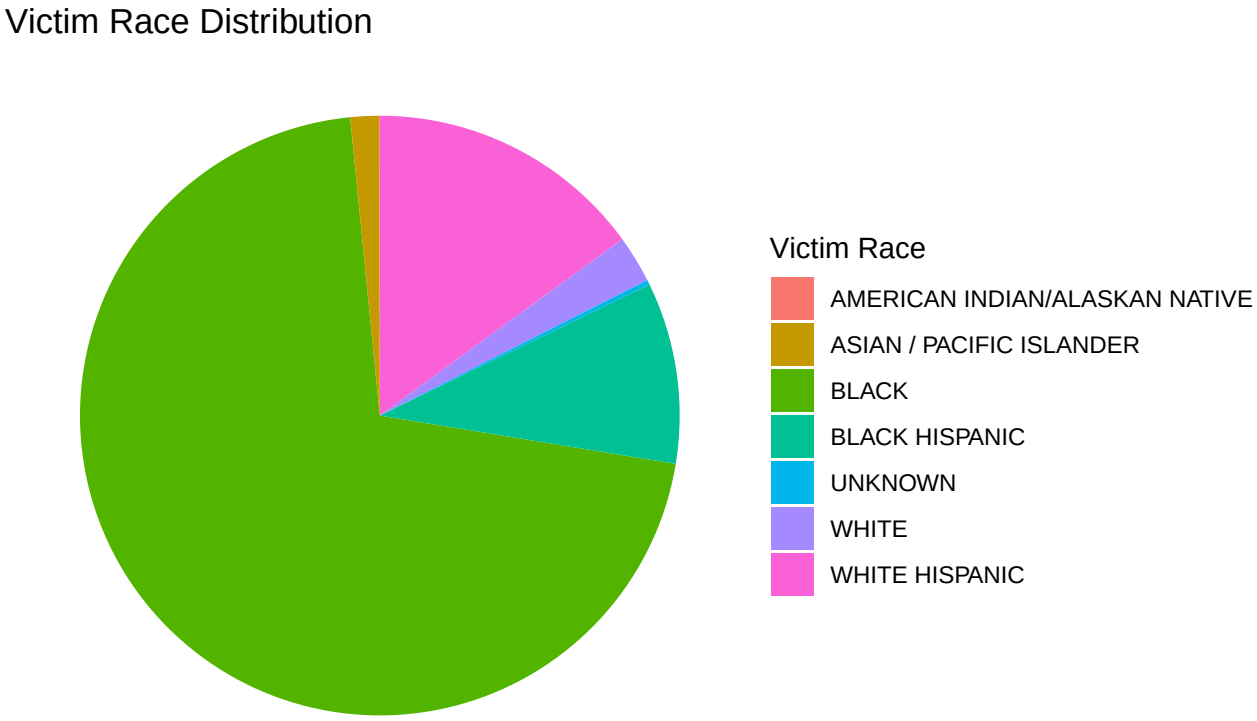
4. Pie Chart of Shooter and Victim Demographics: This shows the racial composition of perpetrators and victims in shootings, which can provide insights into demographic patterns or disparities in shooting incidents.

```
# Aggregate data for perpetrators
perp_race_distribution <- tidy_nypd_shooting_data %>%
  count(PERP_RACE, sort = TRUE) %>%
  filter(PERP_RACE != "")  # Assuming empty strings represent missing data

# Aggregate data for victims
vic_race_distribution <- tidy_nypd_shooting_data %>%
```

```
  count(VIC_RACE, sort = TRUE) %>%
  filter(VIC_RACE != "")

# Creating pie charts
# Pie chart for perpetrator race
ggplot(data = perp_race_distribution, aes(x = "", y = n, fill = PERP_RACE)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(title = "Perpetrator Race Distribution", x = NULL, y = NULL, fill = "Perpetrator Race") +
  theme_void()
```




```
# Pie chart for victim race
ggplot(data = vic_race_distribution, aes(x = "", y = n, fill = VIC_RACE)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(title = "Victim Race Distribution", x = NULL, y = NULL, fill = "Victim Race") +
  theme_void()
```



Modeling:

A straightforward model which can be useful is a logistic regression model. This model can predict the probability of an incident being flagged as a murder (Is_Murder) based on some of the available features, such as the time of the incident, the borough, and the demographic characteristics of the perpetrator or victim.

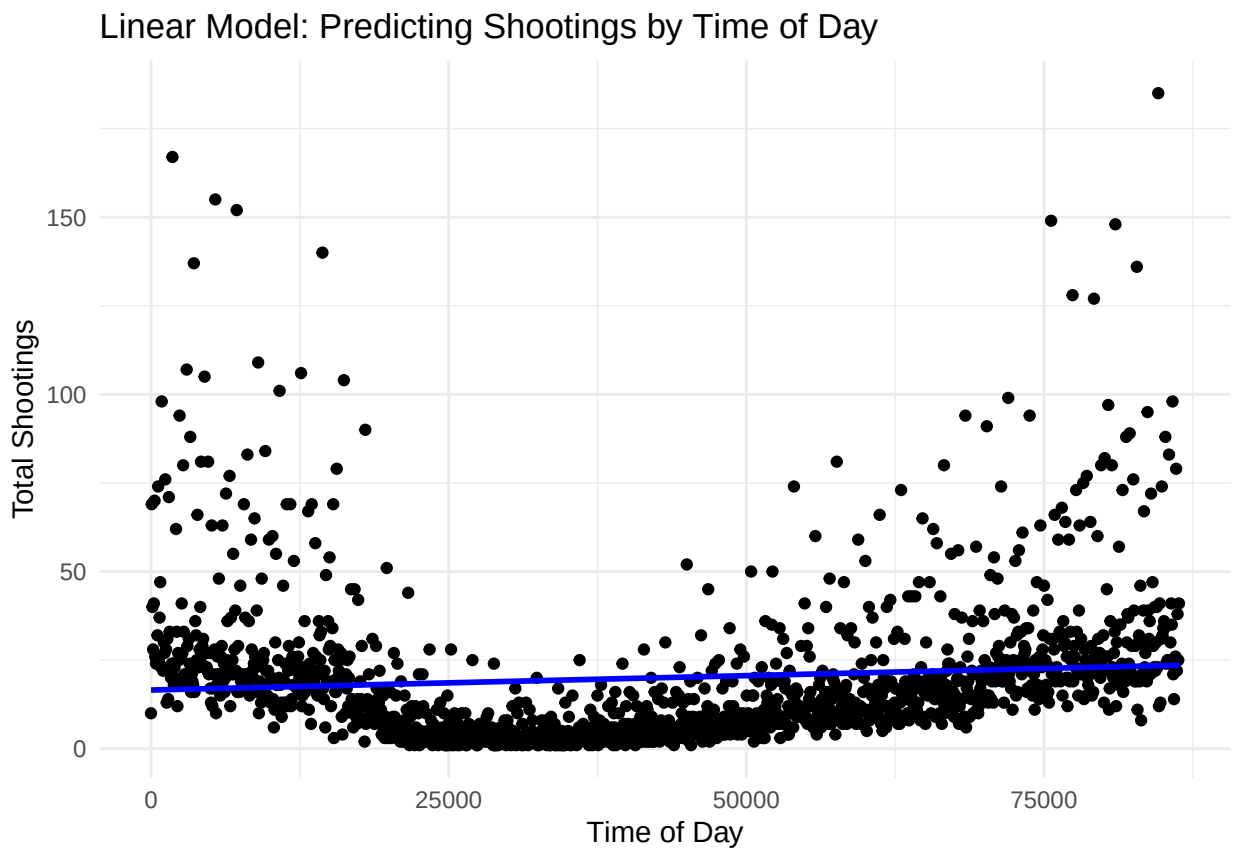
```
# Aggregate data by time of day
shootings_by_time <- tidy_nypd_shooting_data %>%
  group_by(OCCUR_TIME) %>%
  summarise(total_shootings = n())

# Fit the linear regression model
time_model <- lm(total_shootings ~ OCCUR_TIME, data = shootings_by_time)

# Summary of the model
summary(time_model)
```

```
##
## Call:
## lm(formula = total_shootings ~ OCCUR_TIME, data = shootings_by_time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.072 -14.072  -5.072   4.928 164.928
```

```
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.072      0.573   35.03   <2e-16 ***
## OCCUR_TIME        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.62 on 1422 degrees of freedom
```

```r
# Visualizing the model
ggplot(shootings_by_time, aes(x = OCCUR_TIME, y = total_shootings)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Linear Model: Predicting Shootings by Time of Day",
       x = "Time of Day", y = "Total Shootings") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Model insights:

The plot and summary indicates that the time variable is not a significant predictor, as evidenced by the "NA" values for the coefficient. The model suggests that there is no clear linear relationship between the

specific time of day and the number of shootings. This implies that other factors, such as location or socio-economic conditions, might play a more critical role in influencing the frequency of shootings, and time of day alone does not adequately capture the variability in shooting incidents. So we should explore some other models for this.

Potential bias:

While this analysis provides valuable insights, it's important to consider potential biases which could be present in the data. The dataset may be subject to reporting bias, where incidents in certain neighborhoods or involving specific demographics are either underreported or overrepresented due to varying levels of police presence, community trust, or socio-economic factors. Additionally, the data does not account for unreported shootings, which could skew the results and lead to an incomplete understanding of shooting incident patterns in New York City.

Conclusion:

This analysis of the NYPD Shooting Incident data reveals significant patterns, with Brooklyn showing the highest shooting incidents and most occurring during evening and night hours. The findings also highlight the severity of these incidents, with a notable portion resulting in murders Additionally, demographic differences among perpetrators and victims suggest the need for targeted solutions. These insights can guide policy and resource allocation to effectively address and reduce shooting incidents in New York City.