NYPD Shooting Data Analysis

2024-08-19

NYPD Dataset Analysis

Importing the data

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

Data Cleanup

In this section we will be organizing the data by removing un-necessary columns that are not directly relevant to the analysis. We focus on key variables such as `OCCUR_DATE`, `OCCUR_TIME`, `BORO`, `VIC_AGE_GROUP`, and `VIC_SEX`, which are most relevant to our analysis.

```
cleaned_data <- shooting_data %>%
  select(-INCIDENT_KEY,-LOC_OF_OCCUR_DESC,-PRECINCT,-JURISDICTION_CODE,
         -LOC_CLASSFCTN_DESC,-LOCATION_DESC,-X_COORD_CD,-Y_COORD_CD,-Latitude,
         -Longitude,-Lon_Lat,-PERP_AGE_GROUP,-PERP_SEX,-PERP_RACE,-STATISTICAL_MURDER_FLAG)
```

Summarizing the cleaned data

```
summary(cleaned_data)
```

```
##   OCCUR_DATE         OCCUR_TIME           BORO           VIC_AGE_GROUP
##  Length:28562       Length:28562      Length:28562       Length:28562
##  Class :character   Class1:hms        Class :character   Class :character
##  Mode  :character   Class2:difftime   Mode  :character   Mode  :character
##                     Mode  :numeric
##    VIC_SEX            VIC_RACE
##  Length:28562       Length:28562
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
```

Data Analysis

Converting Data and Time Columns

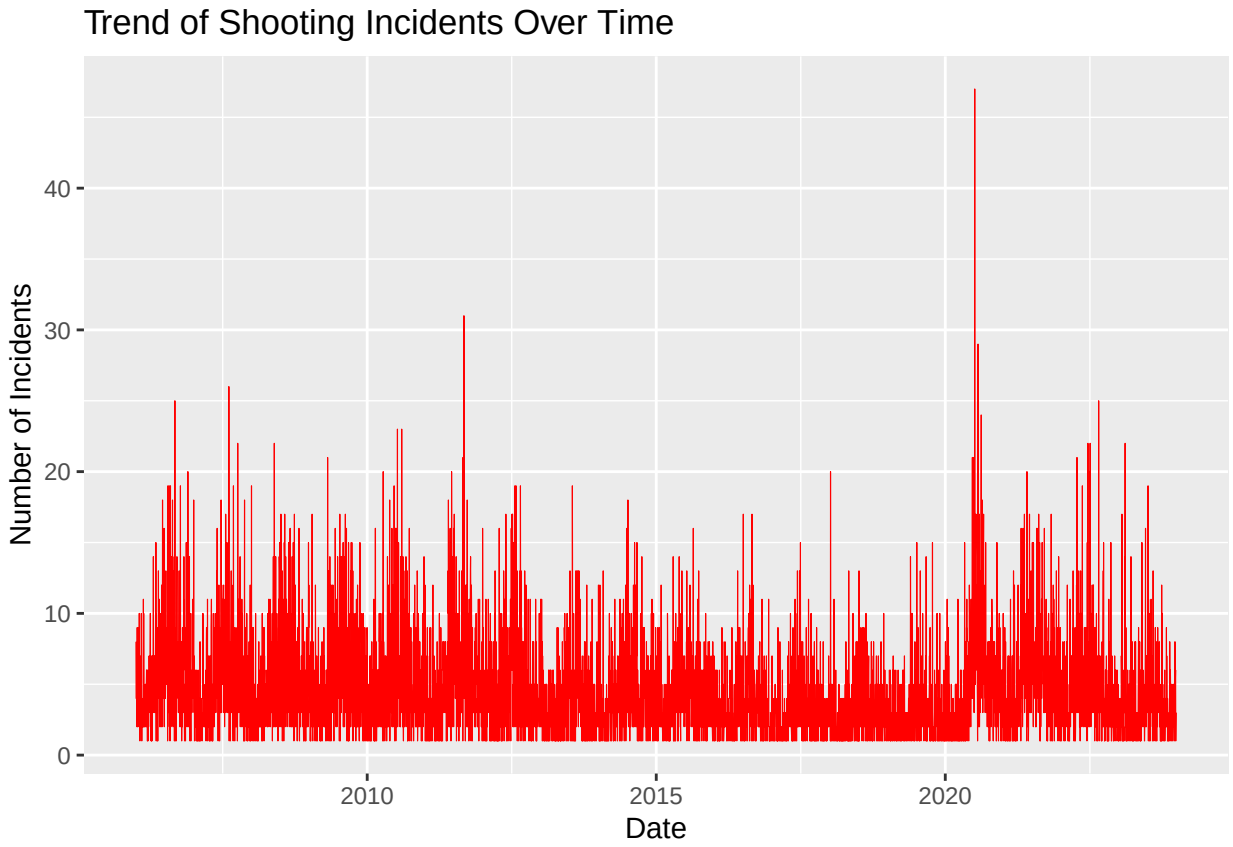
Before performing any analysis, it is important to ensure that the date and time columns are in a format suitable for time-series analysis. We convert the `OCCUR_DATE` into a date format and `OCCUR_TIME` into a time format, allowing us to properly group and visualize incidents over time.

```
cleaned_data <- cleaned_data %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"),
         OCCUR_TIME = hms::as_hms(OCCUR_TIME))
```

Visualize the Shooting Incidents over time

The plot shows a significant spike in shooting incidents in early to mid-2020.

```
# Group by date and count incidents
daily_incidents <- cleaned_data %>%
  group_by(OCCUR_DATE) %>%
  summarise(count = n())

# Plot the trend
ggplot(daily_incidents, aes(x = OCCUR_DATE, y = count)) +
  geom_line(color = "red", linewidth = 0.2) +
  labs(title = "Trend of Shooting Incidents Over Time", x = "Date", y = "Number of Incidents")
```

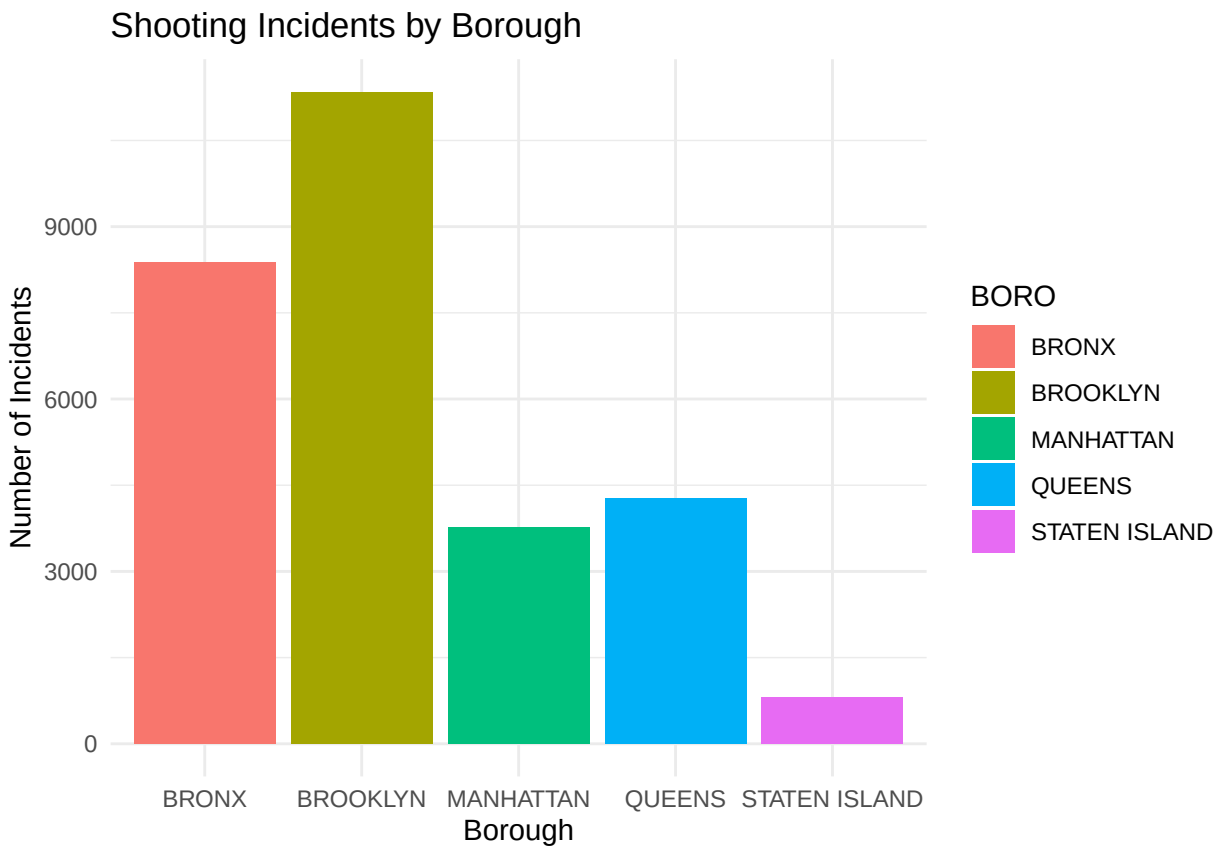


Visualize Incidents by Borough

The distribution of incidents across boroughs reveals that Brooklyn has the highest number of shooting incidents. Possible factors contributing to this concentration could include socioeconomic conditions, population density, or historical crime patterns in those neighborhoods.

```
# Group by borough and count incidents
borough_incidents <- cleaned_data %>%
  group_by(BORO) %>%
  summarise(count = n())

# Plot the distribution
ggplot(borough_incidents, aes(x = BORO, y = count, fill = BORO)) +
  geom_bar(stat = "identity") +
  labs(title = "Shooting Incidents by Borough", x = "Borough", y = "Number of Incidents") +
```

```
  theme_minimal()
```

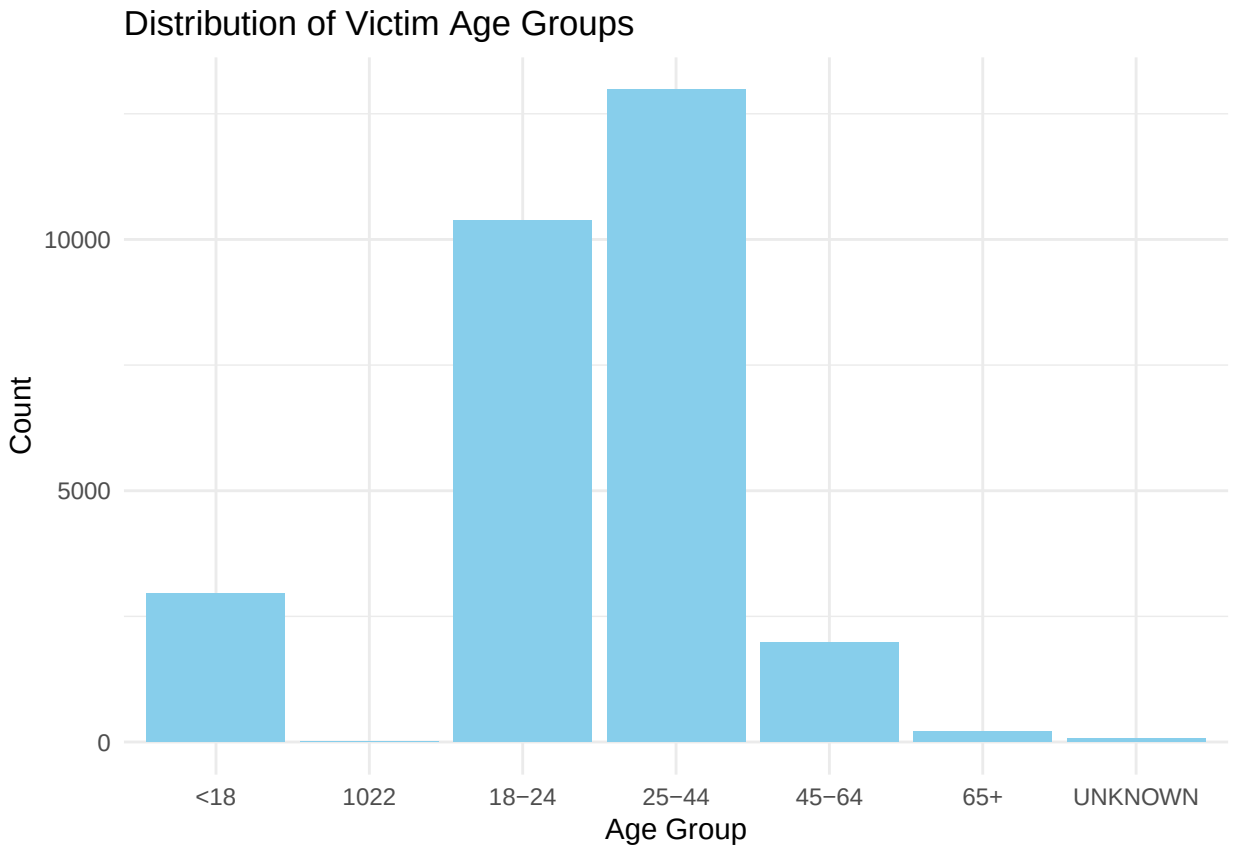


Visualize Victim Demographics

By Age Group

- The distribution of victim age groups indicates that certain age ranges are more frequently affected by shootings. This insight could be useful for community programs aimed at high-risk age groups.

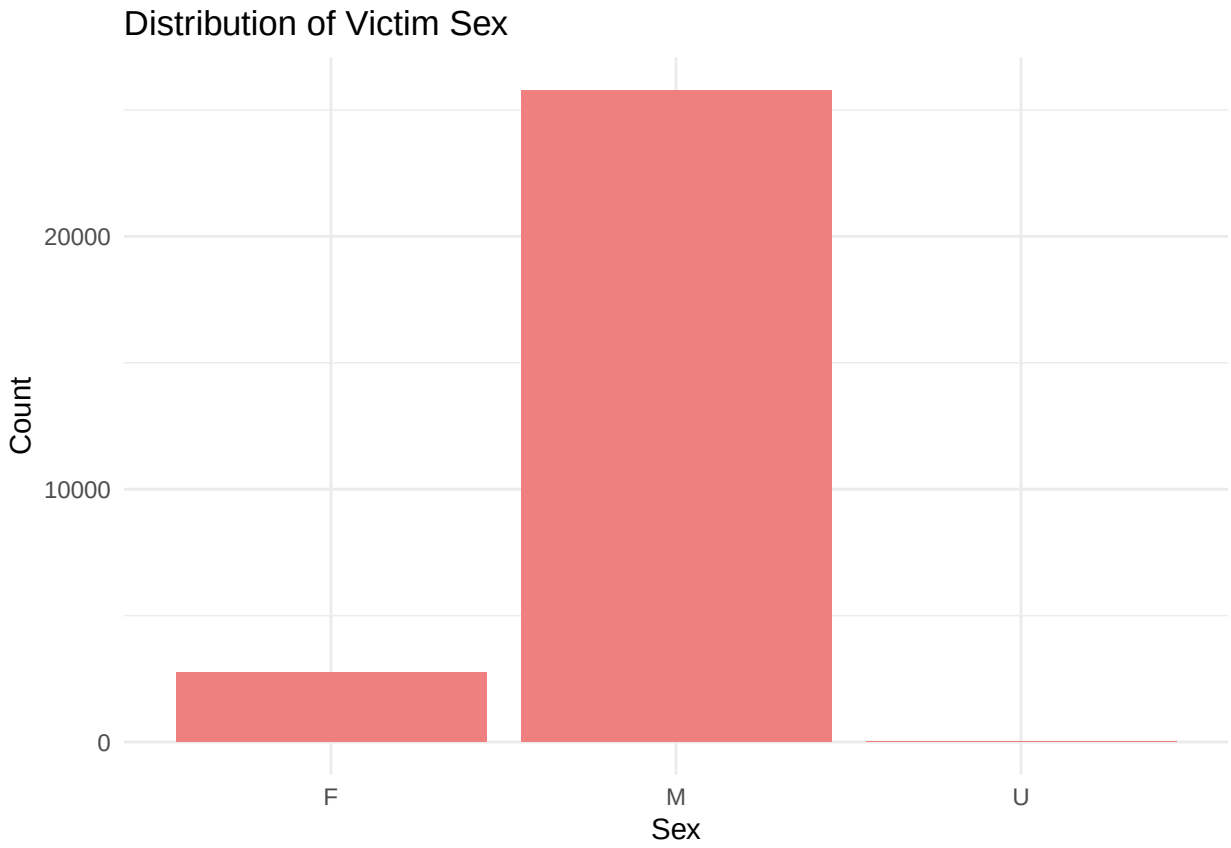
```
# Age group distribution
ggplot(cleaned_data, aes(x = VIC_AGE_GROUP)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of Victim Age Groups", x = "Age Group", y = "Count") +
  theme_minimal()
```



By Sex of the Victim

- The majority of victims are male.

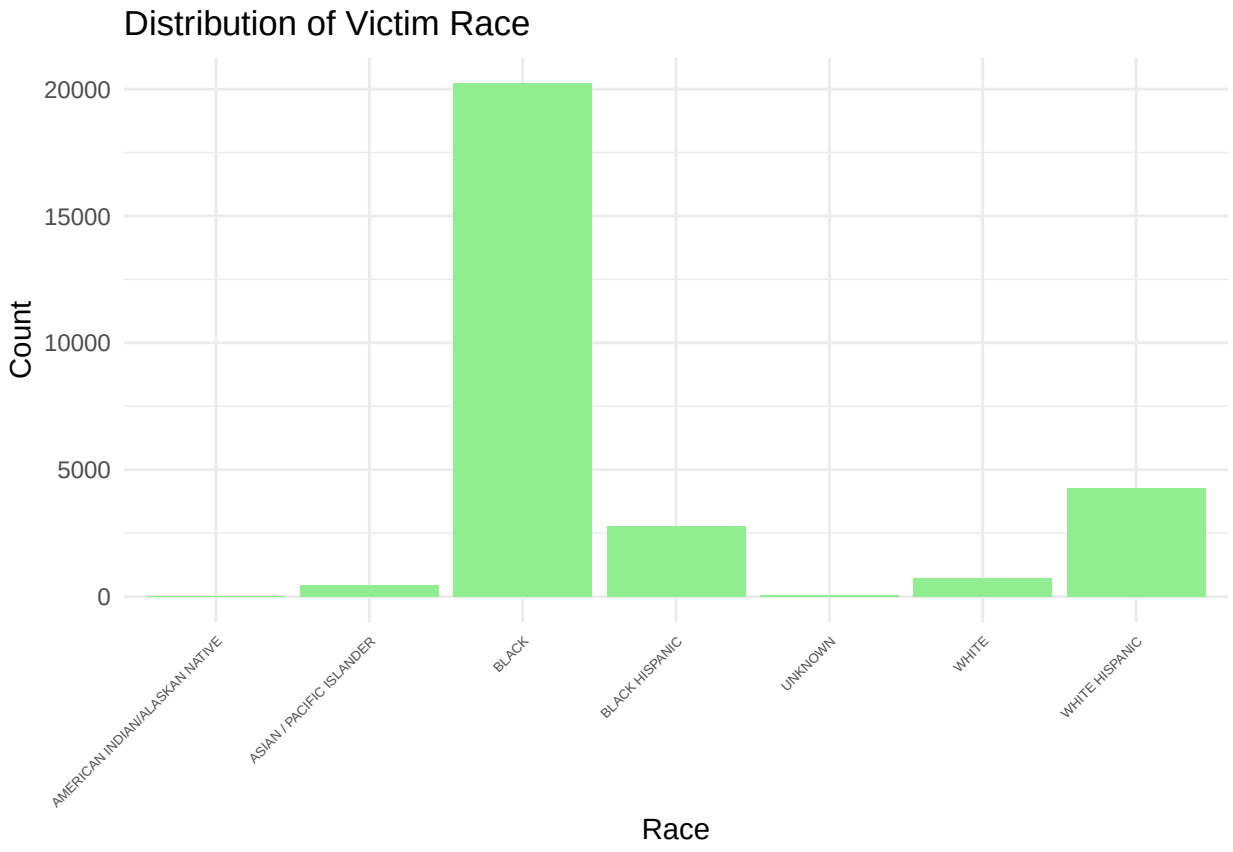
```
# Sex distribution
ggplot(cleaned_data, aes(x = VIC_SEX)) +
  geom_bar(fill = "lightcoral") +
  labs(title = "Distribution of Victim Sex", x = "Sex", y = "Count") +
  theme_minimal()
```



By Race of the Victim

- The race distribution of victims suggests that certain racial groups are disproportionately affected by shootings.

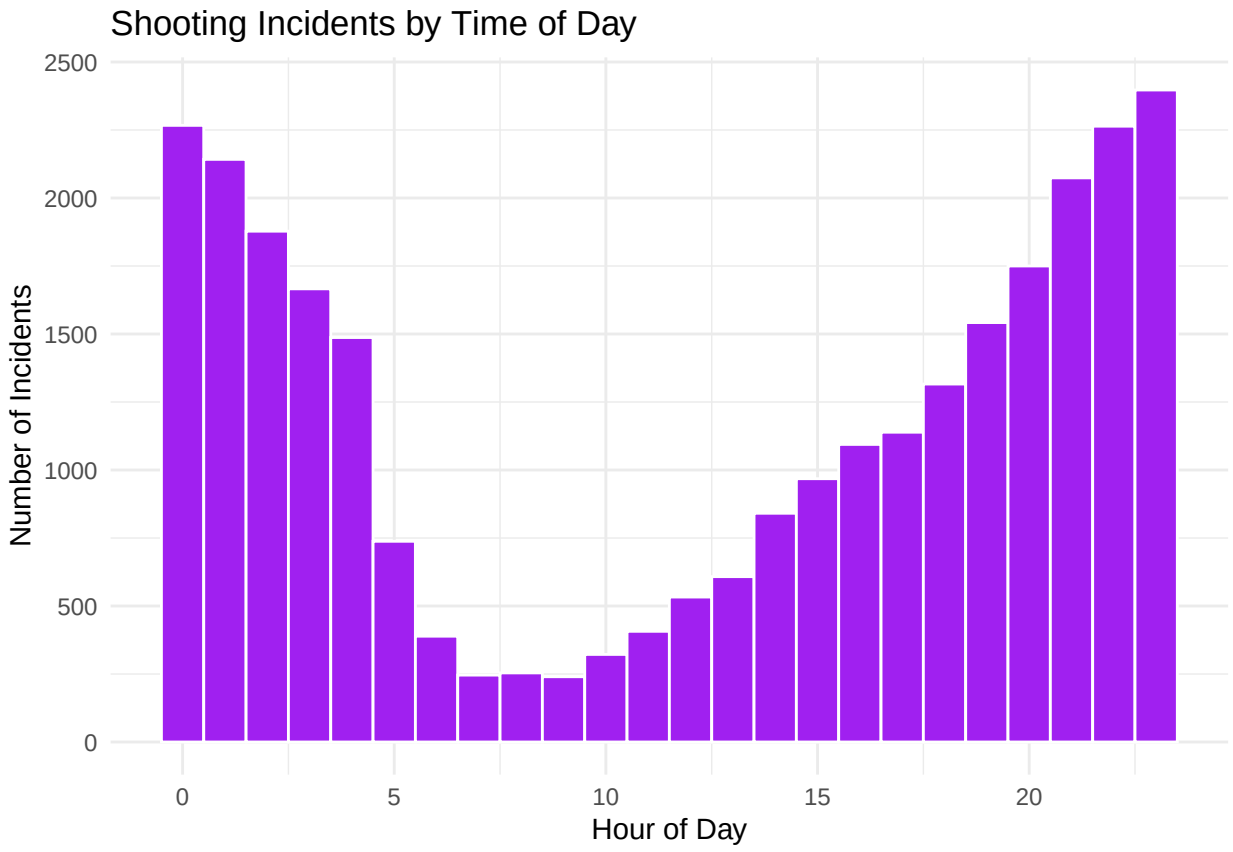
```
# Race distribution
ggplot(cleaned_data, aes(x = VIC_RACE)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribution of Victim Race", x = "Race", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1, size = 5))
```



Time of Day Analysis

The analysis reveals that shooting incidents tend to occur more frequently late at night and early in the morning.

```
# Extract hour from time
cleaned_data_hour <- cleaned_data %>%
  mutate(hour = hour(OCCUR_TIME))

# Plot incidents by hour
ggplot(cleaned_data_hour, aes(x = hour)) +
  geom_histogram(binwidth = 1, fill = "purple", color = "white") +
  labs(title = "Shooting Incidents by Time of Day", x = "Hour of Day", y = "Number of Incidents") +
  theme_minimal()
```



Modeling Data

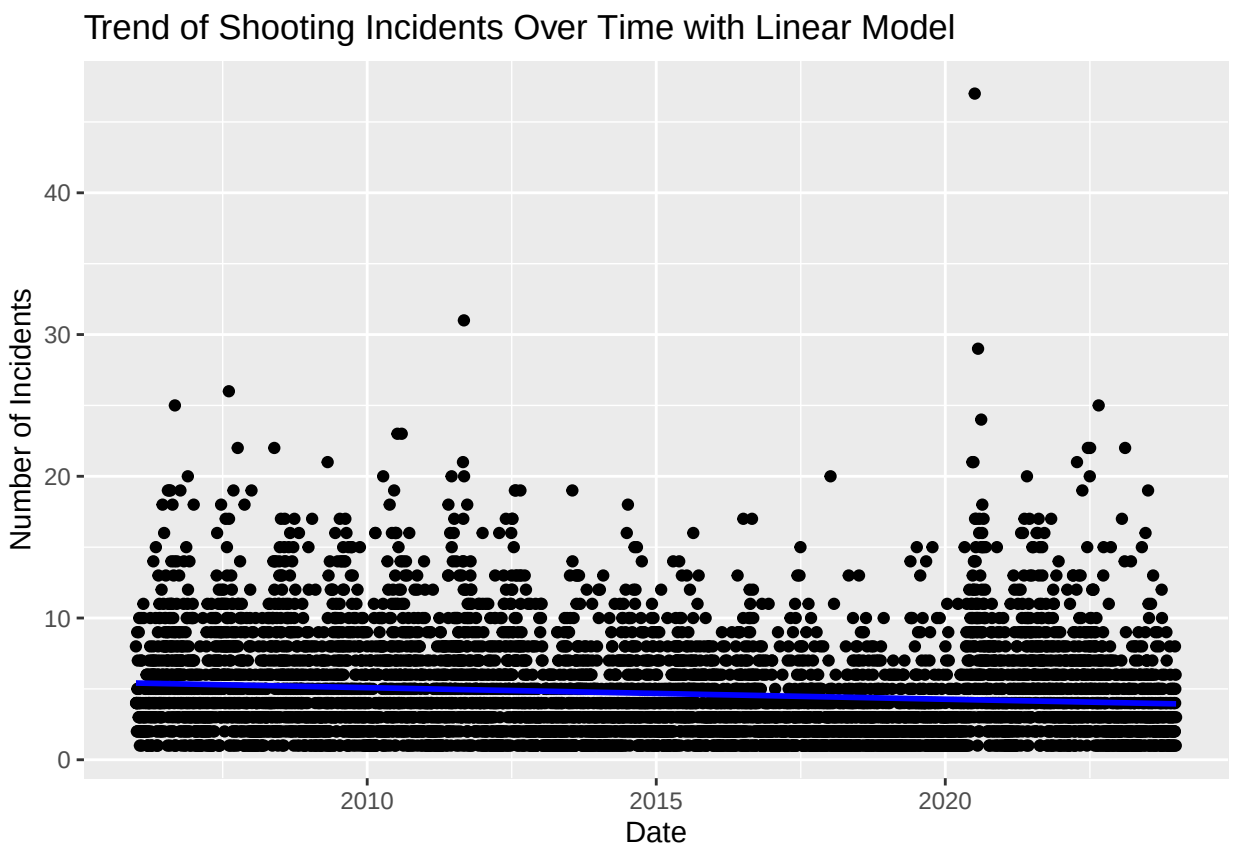
In this section, we fit a linear model to predict the number of incidents over time. The model shows a slight negative trend, indicating an slight decrease in the frequency of incidents. The regression line provides a visual representation of this trend. The p-value is extremely low, indicating that this trend is statistically significant. Even though the trend is small, it is statistically significant. This model offers a starting point for understanding how incidents have changed over time

```
# Clean and transform data
cleaned_data_hour <- cleaned_data_hour %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"),
         OCCUR_TIME = hms::as_hms(OCCUR_TIME),
         OCCUR_DATE_NUM = as.numeric(OCCUR_DATE),
         hour = hour(OCCUR_TIME))

# Aggregate daily incidents
daily_incidents <- cleaned_data_hour %>%
  group_by(OCCUR_DATE) %>%
  summarise(count = n()) %>%
  mutate(OCCUR_DATE_NUM = as.numeric(OCCUR_DATE))

# Fit a linear model for time trend
model_time <- lm(count ~ OCCUR_DATE_NUM, data = daily_incidents)
summary(model_time)
```

```
##
## Call:
```

```
## lm(formula = count ~ OCCUR_DATE_NUM, data = daily_incidents)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.408 -2.443 -0.978  1.646 42.771
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.353e+00  3.900e-01  21.419   <2e-16 ***
## OCCUR_DATE_NUM -2.235e-04  2.361e-05  -9.466   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.529 on 6093 degrees of freedom
## Multiple R-squared:  0.01449,    Adjusted R-squared:  0.01433
## F-statistic:  89.6 on 1 and 6093 DF,  p-value: < 2.2e-16
```

```r
# Plot the trend with the regression line
ggplot(daily_incidents, aes(x = OCCUR_DATE, y = count)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Trend of Shooting Incidents Over Time with Linear Model", x = "Date", y = "Number of In
```



Bias Identification

Bias in the dataset can emerge from several factors, such as under-reporting in certain communities, disproportionate targeting of specific groups, or sampling bias during data collection.
Certain groups may be over-represented or underrepresented in the data. Comparing the dataset demographics

with those of the broader population can help identify disparities. Areas with higher policing might show inflated incident counts compared to less-policed areas. Events like public protests (George Floyd protests in 2020 - 2021) or pandemics could lead to temporary spikes in incidents, skewing overall trends.

Conclusion

This analysis provides insights into the patterns and trends of NYPD shooting incidents in New York City. By examining temporal trends, geographic distribution, and victim demographics, we gain a better understanding of where and when incidents are most likely to occur, as well as which populations are most affected. However, it is essential to consider potential biases in the data, as these could impact the effectiveness of any future interventions or policies. Further research could dive deeper into other factors and explore more complex models to predict and reduce future incidents.