Analysis of NYPD Historic Shooting Incidents using R

Marco F

2025-06-14

Table of Contents

[Introduction 1](#_Toc200830884)

[1. Importing Data 2](#_Toc200830885)

[2. Data Cleaning 6](#_Toc200830886)

[3. Visualizing Data 13](#_Toc200830887)

[Visualization 1: Shooting Incidents Over Years 13](#_Toc200830888)

[Visualization 2: Shooting Incidents by Borough 14](#_Toc200830889)

[Visualization 3: Shooting Incidents by Borough Per 100K Residents 16](#_Toc200830890)

[Visualization 4: Incidents by Hour of Day 17](#_Toc200830891)

[4. Analyzing and Modeling 19](#_Toc200830892)

[Logistic Regression Analysis of Fatal Shooting Likelihood 19](#_Toc200830893)

[Generalized Linear Model (GLM) Output Explanation 20](#_Toc200830894)

[Key Findings 20](#_Toc200830895)

[Model Fit 21](#_Toc200830896)

[Notes 21](#_Toc200830897)

[5 Personal Bias and Mitigation 21](#_Toc200830898)

[6 Conclusion 22](#_Toc200830899)

## Introduction

This report conducts an analysis of the NYPD Shooting Incident Data (Historic). The primary goal is to explore patterns, trends, and characteristics of shooting incidents in New York City. We will follow a standard data analysis workflow:

1. **Importing Data**: Loading the dataset into R.
2. **Data Cleaning**: Preparing the data for analysis by handling missing values, correcting data types, and transforming variables.
3. **Visualizing Data**: Creating visualizations to understand distributions and relationships.
4. **Analyzing and Modeling**: Performing statistical analysis and building a predictive model.
5. **Bias Discussion**: Identifying potential biases in the data and analysis.
6. **Conclusion**: Summarizing key findings and limitations.

The dataset is publicly available from NYC OpenData.

## 1. Importing Data

First, we load the necessary R packages and import the dataset. We’ll use tidyverse for general data manipulation and visualization, lubridate for date-time operations, skimr for quick summaries, rpart and rpart.plot for decision tree modeling.

# Install packages if not already installed  
# install.packages(c("tidyverse", "lubridate", "skimr", "rpart", "rpart.plot", "kableExtra"))  
  
library(tidyverse)  
library(lubridate)  
library(skimr)  
library(rpart)  
library(rpart.plot)  
library(kableExtra)  
library(broom)  
  
# URL for the NYPD Shooting Incident Data (Historic) - Same as https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD  
data\_url <- "https://raw.githubusercontent.com/marcofanti/Final-Project-1-Data-Science-NYPD-Shooting-Incident-Historic/refs/heads/main/Data/NYPD\_Shooting\_Incident\_Data\_\_Historic\_.csv"  
  
# Load the dataset  
nypd\_shootings\_raw <- read\_csv(data\_url, show\_col\_types = FALSE)  
  
cat("Data loaded successfully.\n")

## Data loaded successfully.

cat(paste("Dataset dimensions:", paste(dim(nypd\_shootings\_raw), collapse = " x "), "\n\n"))

## Dataset dimensions: 29744 x 21

cat("First few rows of raw data:\n")

## First few rows of raw data:

kable(head(nypd\_shootings\_raw))

| INCIDENT\_KEY | OCCUR\_DATE | OCCUR\_TIME | BORO | LOC\_OF\_OCCUR\_DESC | PRECINCT | JURISDICTION\_CODE | LOC\_CLASSFCTN\_DESC | LOCATION\_DESC | STATISTICAL\_MURDER\_FLAG | PERP\_AGE\_GROUP | PERP\_SEX | PERP\_RACE | VIC\_AGE\_GROUP | VIC\_SEX | VIC\_RACE | X\_COORD\_CD | Y\_COORD\_CD | Latitude | Longitude | Lon\_Lat |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 231974218 | 08/09/2021 | 01:06:00 | BRONX | NA | 40 | 0 | NA | NA | FALSE | NA | NA | NA | 18-24 | M | BLACK | 1006343.0 | 234270.0 | 40.80967 | -73.92019 | POINT (-73.92019278899994 40.80967347200004) |
| 177934247 | 04/07/2018 | 19:48:00 | BROOKLYN | NA | 79 | 0 | NA | NA | TRUE | 25-44 | M | WHITE HISPANIC | 25-44 | M | BLACK | 1000082.9 | 189064.7 | 40.68561 | -73.94291 | POINT (-73.94291302299996 40.685609672000055) |
| 255028563 | 12/02/2022 | 22:57:00 | BRONX | OUTSIDE | 47 | 0 | STREET | GROCERY/BODEGA | FALSE | (null) | (null) | (null) | 25-44 | M | BLACK | 1020691.0 | 257125.0 | 40.87235 | -73.86823 | POINT (-73.868233 40.872349) |
| 25384540 | 11/19/2006 | 01:50:00 | BROOKLYN | NA | 66 | 0 | NA | PVT HOUSE | TRUE | UNKNOWN | U | UNKNOWN | 18-24 | M | BLACK | 985107.3 | 173349.8 | 40.64249 | -73.99691 | POINT (-73.99691224999998 40.642489932000046) |
| 72616285 | 05/09/2010 | 01:58:00 | BRONX | NA | 46 | 0 | NA | MULTI DWELL - APT BUILD | TRUE | 25-44 | M | BLACK | <18 | F | BLACK | 1009853.5 | 247502.6 | 40.84598 | -73.90746 | POINT (-73.90746098599993 40.84598358900007) |
| 85875439 | 07/22/2012 | 21:35:00 | BRONX | NA | 42 | 2 | NA | MULTI DWELL - PUBLIC HOUS | FALSE | 18-24 | M | BLACK | 18-24 | M | BLACK | 1011046.7 | 239814.2 | 40.82488 | -73.90318 | POINT (-73.90317908399999 40.82487781900005) |

cat("\n\nSummary of raw data:\n")

##   
##   
## Summary of raw data:

skim(nypd\_shootings\_raw) %>%   
 select(skim\_type, skim\_variable, n\_missing, complete\_rate, character.empty, numeric.mean)

Data summary

|  |  |
| --- | --- |
| Name | nypd\_shootings\_raw |
| Number of rows | 29744 |
| Number of columns | 21 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 12 |
| difftime | 1 |
| logical | 1 |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | empty |
| --- | --- | --- | --- |
| OCCUR\_DATE | 0 | 1.00 | 0 |
| BORO | 0 | 1.00 | 0 |
| LOC\_OF\_OCCUR\_DESC | 25596 | 0.14 | 0 |
| LOC\_CLASSFCTN\_DESC | 25596 | 0.14 | 0 |
| LOCATION\_DESC | 14977 | 0.50 | 0 |
| PERP\_AGE\_GROUP | 9344 | 0.69 | 0 |
| PERP\_SEX | 9310 | 0.69 | 0 |
| PERP\_RACE | 9310 | 0.69 | 0 |
| VIC\_AGE\_GROUP | 0 | 1.00 | 0 |
| VIC\_SEX | 0 | 1.00 | 0 |
| VIC\_RACE | 0 | 1.00 | 0 |
| Lon\_Lat | 97 | 1.00 | 0 |

**Variable type: difftime**

| skim\_variable | n\_missing | complete\_rate |
| --- | --- | --- |
| OCCUR\_TIME | 0 | 1 |

**Variable type: logical**

| skim\_variable | n\_missing | complete\_rate |
| --- | --- | --- |
| STATISTICAL\_MURDER\_FLAG | 0 | 1 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean |
| --- | --- | --- | --- |
| INCIDENT\_KEY | 0 | 1 | 133850951.08 |
| PRECINCT | 0 | 1 | 65.23 |
| JURISDICTION\_CODE | 2 | 1 | 0.32 |
| X\_COORD\_CD | 0 | 1 | 1009442.08 |
| Y\_COORD\_CD | 0 | 1 | 208721.99 |
| Latitude | 97 | 1 | 40.74 |
| Longitude | 97 | 1 | -73.91 |

## 2. Data Cleaning

Data cleaning is crucial for reliable analysis. This section involves:

Standardizing column names.

Converting data types (e.g., dates, times).

Handling missing values.

Feature engineering (e.g., extracting year, month, hour).

nypd\_shootings\_clean <- nypd\_shootings\_raw %>%  
 # Standardize column names (lowercase and replace spaces with underscores)  
 rename\_with(tolower) %>%  
 rename\_with(~gsub(" ", "\_", .x))   
  
# Convert OCCUR\_DATE to Date object  
nypd\_shootings\_clean <- nypd\_shootings\_clean %>%  
 mutate(occur\_date = mdy(occur\_date)) # Assumes MM/DD/YYYY format  
  
# Extract year, month, day\_of\_week  
nypd\_shootings\_clean <- nypd\_shootings\_clean %>%  
 mutate(  
 year = year(occur\_date),  
 month = month(occur\_date, label = TRUE, abbr = FALSE),  
 day\_of\_week = wday(occur\_date, label = TRUE, abbr = FALSE)  
 )  
  
# Handle OCCUR\_TIME - convert to hour  
nypd\_shootings\_clean <- nypd\_shootings\_clean %>%  
 mutate(  
 occur\_hour = case\_when(  
 !is.na(occur\_time) ~ hour(hms(occur\_time, quiet = TRUE)), # Try to parse HH:MM:SS  
 TRUE ~ NA\_integer\_ # If parsing fails or is NA, keep as NA  
 )  
 )  
  
# Impute missing occur\_hour with the median hour  
median\_hour <- median(nypd\_shootings\_clean$occur\_hour, na.rm = TRUE)  
nypd\_shootings\_clean <- nypd\_shootings\_clean %>%  
 mutate(occur\_hour = ifelse(is.na(occur\_hour), median\_hour, occur\_hour))  
  
# Handle missing values for key categorical features by replacing with 'UNKNOWN'  
cols\_to\_fill\_unknown <- c(  
 "boro", "perp\_age\_group", "perp\_sex", "perp\_race",  
 "vic\_age\_group", "vic\_sex", "vic\_race",   
 "loc\_of\_occur\_desc", "loc\_classfctn\_desc", "location\_desc"   
)  
  
# Ensure these columns exist before trying to mutate them  
existing\_cols\_to\_fill <- intersect(cols\_to\_fill\_unknown, names(nypd\_shootings\_clean))  
  
nypd\_shootings\_clean <- nypd\_shootings\_clean %>%  
 mutate(across(all\_of(existing\_cols\_to\_fill), ~replace\_na(as.character(.x), "UNKNOWN"))) %>%  
 mutate(across(all\_of(existing\_cols\_to\_fill), ~ifelse(.x %in% c("", "(null)"), "UNKNOWN", .x)))  
  
  
# Convert STATISTICAL\_MURDER\_FLAG to a factor (0 for No, 1 for Yes)  
# The column might be logical (TRUE/FALSE) or character ("true"/"false")  
nypd\_shootings\_clean <- nypd\_shootings\_clean %>%  
 mutate(  
 statistical\_murder\_flag = case\_when(  
 is.logical(statistical\_murder\_flag) ~ factor(statistical\_murder\_flag, levels = c(FALSE, TRUE), labels = c("No", "Yes")),  
 is.character(statistical\_murder\_flag) ~ factor(tolower(statistical\_murder\_flag), levels = c("false", "true"), labels = c("No", "Yes")),  
 TRUE ~ factor(NA, levels = c("No", "Yes")) # Handle other cases or if column doesn't exist as expected  
 ),  
 # If it was already T/F, ensure NAs are handled (e.g. replace with "No" or a specific category)  
 statistical\_murder\_flag = replace\_na(statistical\_murder\_flag, "No")   
 )  
  
  
# Drop columns not immediately useful for this analysis or with too many unique values for simple modeling  
# For example, specific coordinates if lat/long are used, or high-cardinality IDs.  
# `incident\_key` is an ID, `lon\_lat` is a point geometry.  
cols\_to\_drop <- c("x\_coord\_cd", "y\_coord\_cd", "lon\_lat", "jurisdiction\_code", "incident\_key", "patch\_code", "zip\_code", "census\_tract") # Added a few more common ones  
# Ensure columns exist before trying to drop  
existing\_cols\_to\_drop <- intersect(cols\_to\_drop, names(nypd\_shootings\_clean))  
if (length(existing\_cols\_to\_drop) > 0) {  
 nypd\_shootings\_clean <- nypd\_shootings\_clean %>%  
 select(-all\_of(existing\_cols\_to\_drop))  
}  
  
cat("\nCleaned data summary:\n")

##   
## Cleaned data summary:

skim(nypd\_shootings\_clean) %>%   
 select(skim\_type, skim\_variable, n\_missing, complete\_rate, character.empty, numeric.mean, factor.ordered, factor.n\_unique)

Data summary

|  |  |
| --- | --- |
| Name | nypd\_shootings\_clean |
| Number of rows | 29744 |
| Number of columns | 20 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 10 |
| Date | 1 |
| difftime | 1 |
| factor | 3 |
| numeric | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | empty |
| --- | --- | --- | --- |
| boro | 0 | 1 | 0 |
| loc\_of\_occur\_desc | 0 | 1 | 0 |
| loc\_classfctn\_desc | 0 | 1 | 0 |
| location\_desc | 0 | 1 | 0 |
| perp\_age\_group | 0 | 1 | 0 |
| perp\_sex | 0 | 1 | 0 |
| perp\_race | 0 | 1 | 0 |
| vic\_age\_group | 0 | 1 | 0 |
| vic\_sex | 0 | 1 | 0 |
| vic\_race | 0 | 1 | 0 |

**Variable type: Date**

| skim\_variable | n\_missing | complete\_rate |
| --- | --- | --- |
| occur\_date | 0 | 1 |

**Variable type: difftime**

| skim\_variable | n\_missing | complete\_rate |
| --- | --- | --- |
| occur\_time | 0 | 1 |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique |
| --- | --- | --- | --- | --- |
| statistical\_murder\_flag | 0 | 1 | FALSE | 2 |
| month | 0 | 1 | TRUE | 12 |
| day\_of\_week | 0 | 1 | TRUE | 7 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean |
| --- | --- | --- | --- |
| precinct | 0 | 1 | 65.23 |
| latitude | 97 | 1 | 40.74 |
| longitude | 97 | 1 | -73.91 |
| year | 0 | 1 | 2014.31 |
| occur\_hour | 0 | 1 | 12.30 |

cat("\nFirst few rows of cleaned data:\n")

##   
## First few rows of cleaned data:

kable(head(nypd\_shootings\_clean))

| occur\_date | occur\_time | boro | loc\_of\_occur\_desc | precinct | loc\_classfctn\_desc | location\_desc | statistical\_murder\_flag | perp\_age\_group | perp\_sex | perp\_race | vic\_age\_group | vic\_sex | vic\_race | latitude | longitude | year | month | day\_of\_week | occur\_hour |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2021-08-09 | 01:06:00 | BRONX | UNKNOWN | 40 | UNKNOWN | UNKNOWN | No | UNKNOWN | UNKNOWN | UNKNOWN | 18-24 | M | BLACK | 40.80967 | -73.92019 | 2021 | August | Monday | 1 |
| 2018-04-07 | 19:48:00 | BROOKLYN | UNKNOWN | 79 | UNKNOWN | UNKNOWN | Yes | 25-44 | M | WHITE HISPANIC | 25-44 | M | BLACK | 40.68561 | -73.94291 | 2018 | April | Saturday | 19 |
| 2022-12-02 | 22:57:00 | BRONX | OUTSIDE | 47 | STREET | GROCERY/BODEGA | No | UNKNOWN | UNKNOWN | UNKNOWN | 25-44 | M | BLACK | 40.87235 | -73.86823 | 2022 | December | Friday | 22 |
| 2006-11-19 | 01:50:00 | BROOKLYN | UNKNOWN | 66 | UNKNOWN | PVT HOUSE | Yes | UNKNOWN | U | UNKNOWN | 18-24 | M | BLACK | 40.64249 | -73.99691 | 2006 | November | Sunday | 1 |
| 2010-05-09 | 01:58:00 | BRONX | UNKNOWN | 46 | UNKNOWN | MULTI DWELL - APT BUILD | Yes | 25-44 | M | BLACK | <18 | F | BLACK | 40.84598 | -73.90746 | 2010 | May | Sunday | 1 |
| 2012-07-22 | 21:35:00 | BRONX | UNKNOWN | 42 | UNKNOWN | MULTI DWELL - PUBLIC HOUS | No | 18-24 | M | BLACK | 18-24 | M | BLACK | 40.82488 | -73.90318 | 2012 | July | Sunday | 21 |

cat("\nMissing values count per column after cleaning:\n")

##   
## Missing values count per column after cleaning:

sapply(nypd\_shootings\_clean, function(x) sum(is.na(x))) %>%   
 data.frame(missing\_count = .) %>%   
 rownames\_to\_column("column\_name") %>%  
 filter(missing\_count > 0) %>%  
 kable()

| column\_name | missing\_count |
| --- | --- |
| latitude | 97 |
| longitude | 97 |

The following variables should be treated as factors:

* boro
* perp\_age\_group
* perp\_sex
* perp\_race
* vic\_age\_group
* vic\_sex
* vic\_race
* statistical\_murder\_flag

nypd\_shootings\_clean$boro <- factor(nypd\_shootings\_clean$boro)  
nypd\_shootings\_clean$perp\_age\_group <- factor(nypd\_shootings\_clean$perp\_age\_group)  
nypd\_shootings\_clean$perp\_sex <- factor(nypd\_shootings\_clean$perp\_sex)  
nypd\_shootings\_clean$perp\_race <- factor(nypd\_shootings\_clean$perp\_race)  
nypd\_shootings\_clean$vic\_age\_group <- factor(nypd\_shootings\_clean$vic\_age\_group)  
nypd\_shootings\_clean$vic\_sex <- factor(nypd\_shootings\_clean$vic\_sex)  
nypd\_shootings\_clean$vic\_race <- factor(nypd\_shootings\_clean$vic\_race)  
nypd\_shootings\_clean$statistical\_murder\_flag <- factor(nypd\_shootings\_clean$statistical\_murder\_flag)

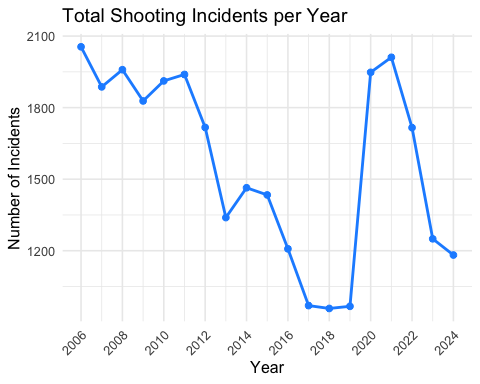
## 3. Visualizing Data

Visualizations help in understanding trends, distributions, and relationships.

### Visualization 1: Shooting Incidents Over Years

This plot shows the trend of shooting incidents annually.

if (nrow(nypd\_shootings\_clean) > 0 && "year" %in% names(nypd\_shootings\_clean)) {  
 incidents\_by\_year <- nypd\_shootings\_clean %>%  
 filter(!is.na(year)) %>% # Ensure year is not NA  
 group\_by(year) %>%  
 summarise(count = n(), .groups = 'drop') %>%  
 filter(year >= min(nypd\_shootings\_clean$year, na.rm=TRUE) & year <= max(nypd\_shootings\_clean$year, na.rm=TRUE)) # Filter out potential outlier years if any from parsing  
  
 ggplot(incidents\_by\_year, aes(x = year, y = count)) +  
 geom\_line(color = "dodgerblue", size = 1) +  
 geom\_point(color = "dodgerblue", size = 2) +  
 labs(  
 title = "Total Shooting Incidents per Year",  
 x = "Year",  
 y = "Number of Incidents"  
 ) +  
 theme\_minimal(base\_size = 12) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 scale\_x\_continuous(breaks = seq(min(incidents\_by\_year$year, na.rm=TRUE), max(incidents\_by\_year$year, na.rm=TRUE), by = 2)) # Adjust breaks as needed  
} else {  
 cat("Skipping yearly trend visualization as data is empty or 'year' column is missing.\n")  
}



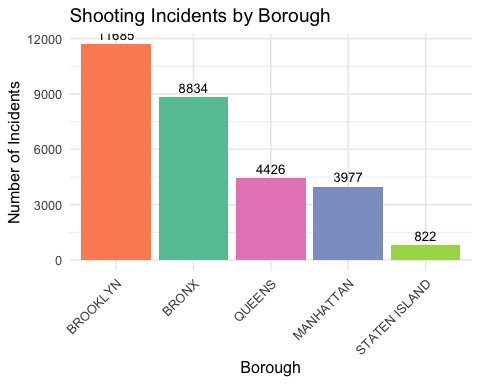
Number of Shooting Incidents per Year

Interpretation: This line chart illustrates the annual frequency of shooting incidents. It can reveal long-term trends, such as increases, decreases, or periods of stability in shooting occurrences.

### Visualization 2: Shooting Incidents by Borough

This bar chart shows the distribution of shooting incidents across NYC boroughs.

if (nrow(nypd\_shootings\_clean) > 0 && "boro" %in% names(nypd\_shootings\_clean)) {  
 incidents\_by\_boro <- nypd\_shootings\_clean %>%  
 filter(!is.na(boro) & boro != "UNKNOWN") %>% # Exclude NA or UNKNOWN for a cleaner plot  
 group\_by(boro) %>%  
 summarise(count = n(), .groups = 'drop') %>%  
 arrange(desc(count))  
  
 ggplot(incidents\_by\_boro, aes(x = reorder(boro, -count), y = count, fill = boro)) +  
 geom\_bar(stat = "identity", show.legend = FALSE) +  
 geom\_text(aes(label = count), vjust = -0.5, size = 3.5) +  
 labs(  
 title = "Shooting Incidents by Borough",  
 x = "Borough",  
 y = "Number of Incidents"  
 ) +  
 scale\_fill\_brewer(palette = "Set2") +  
 theme\_minimal(base\_size = 12) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
} else {  
 cat("Skipping borough distribution visualization as data is empty or 'boro' column is missing.\n")  
}



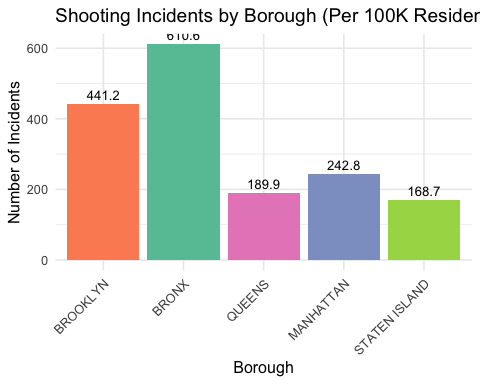
Distribution of Shooting Incidents by Borough

Interpretation: This bar chart highlights which boroughs experience the highest and lowest numbers of shooting incidents, providing insight into the geographical distribution of these events.

### Visualization 3: Shooting Incidents by Borough Per 100K Residents

This bar chart shows the distribution of shooting incidents across NYC boroughs per 100K residents.

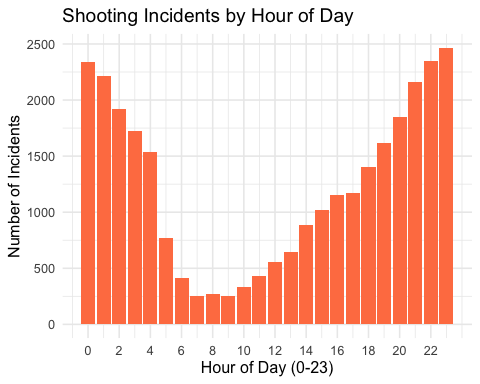
# Load population data for 2020 - Same as "https://data.cityofnewyork.us/resource/xywu-7bv9.csv")  
pop\_url <- "https://raw.githubusercontent.com/marcofanti/Final-Project-1-Data-Science-NYPD-Shooting-Incident-Historic/refs/heads/main/Data/xywu-7bv9.csv"  
  
population <- read\_csv(pop\_url, show\_col\_types = FALSE)  
  
population <- population %>%  
 # Standardize column names (lowercase and replace spaces with underscores)  
 rename\_with(tolower) %>%  
 rename\_with(~gsub(" ", "\_", .x))   
  
population\_2020 <- population %>%  
 filter(borough != "NYC Total") %>%  
 select(borough, `2020`) %>%  
 rename(boro = borough, Population = `2020`) %>%  
 mutate(boro = toupper(boro))  
  
  
if (nrow(nypd\_shootings\_clean) > 0 && "boro" %in% names(nypd\_shootings\_clean)) {  
 incidents\_by\_boro <- nypd\_shootings\_clean %>%  
 filter(!is.na(boro) & boro != "UNKNOWN") %>% # Exclude NA or UNKNOWN for a cleaner plot  
 group\_by(boro) %>%  
 summarise(count = n(), .groups = 'drop') %>%  
 arrange(desc(count))  
  
 # Join with population data  
 incidents\_population <- incidents\_by\_boro %>%  
 left\_join(population\_2020, by = "boro") %>%  
 mutate(Rate\_Per\_100k = (count / Population) \* 100000)  
  
  
 ggplot(incidents\_population, aes(x = reorder(boro, -count), y = Rate\_Per\_100k, fill = boro)) +  
 geom\_bar(stat = "identity", show.legend = FALSE) +  
 geom\_text(aes(label = round(Rate\_Per\_100k, 1)), vjust = -0.5, size = 3.5) +  
 labs(  
 title = "Shooting Incidents by Borough (Per 100K Residents)",  
 x = "Borough",  
 y = "Number of Incidents"  
 ) +  
 scale\_fill\_brewer(palette = "Set2") +  
 theme\_minimal(base\_size = 12) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
} else {  
 cat("Skipping borough distribution visualization as data is empty or 'boro' column is missing.\n")  
}

 Interpretation: This bar chart highlights which boroughs experience the highest and lowest numbers of shooting incidents per 100K residents. If we look at the incidence rate per 100,000 residents, and compare to the previous visualization, the Bronx has the highest number of shootings per capita, and Manhattan overtakes Queens in this metric.

### Visualization 4: Incidents by Hour of Day

This visualization explores the temporal pattern of shootings throughout the day.

if (nrow(nypd\_shootings\_clean) > 0 && "occur\_hour" %in% names(nypd\_shootings\_clean)) {  
 incidents\_by\_hour <- nypd\_shootings\_clean %>%  
 filter(!is.na(occur\_hour)) %>%  
 group\_by(occur\_hour) %>%  
 summarise(count = n(), .groups = 'drop')  
  
 ggplot(incidents\_by\_hour, aes(x = occur\_hour, y = count)) +  
 geom\_col(fill = "coral") +  
 labs(  
 title = "Shooting Incidents by Hour of Day",  
 x = "Hour of Day (0-23)",  
 y = "Number of Incidents"  
 ) +  
 scale\_x\_continuous(breaks = seq(0, 23, by = 2)) +  
 theme\_minimal(base\_size = 12)  
} else {  
 cat("Skipping hour distribution visualization as data is empty or 'occur\_hour' column is missing.\n")  
}



Shooting Incidents by Hour of Day

Interpretation: This plot shows if there are particular times of day when shooting incidents are more frequent.

## 4. Analyzing and Modeling

## Logistic Regression Analysis of Fatal Shooting Likelihood

We model whether an incident was fatal (statistical\_murder\_flag) based on shooting incident’s borough (boro), time of occurrence (occur\_hour), victim’s age group, sex, and race.

if (nrow(nypd\_shootings\_clean) > 0 && "statistical\_murder\_flag" %in% names(nypd\_shootings\_clean)) {  
 # Select features for modeling  
 # Ensure all selected features exist in the dataframe  
 feature\_candidates <- c(  
 "boro", "occur\_hour", "vic\_age\_group", "vic\_sex", "vic\_race"  
 )  
   
 model\_features <- intersect(feature\_candidates, names(nypd\_shootings\_clean))  
   
 model\_data <- nypd\_shootings\_clean %>%  
 select(all\_of(model\_features)) %>%  
 na.omit() # Remove rows with any NAs in selected features for simplicity  
   
 # Convert character columns to factors for rpart  
 model\_data <- model\_data %>%  
 mutate(across(where(is.character), as.factor)) %>%  
 mutate(across(where(is.logical), as.factor)) # Ensure logicals are factors too  
   
 cat(paste("Dimensions of data for modeling:", paste(dim(model\_data), collapse = " x "), "\n"))  
}

## Dimensions of data for modeling: 29744 x 5

glm\_model <- glm(statistical\_murder\_flag ~ boro + occur\_hour + vic\_age\_group + vic\_sex + vic\_race,   
 data = nypd\_shootings\_clean, family = 'binomial')  
  
summary(glm\_model)

##   
## Call:  
## glm(formula = statistical\_murder\_flag ~ boro + occur\_hour + vic\_age\_group +   
## vic\_sex + vic\_race, family = "binomial", data = nypd\_shootings\_clean)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -12.907402 89.668090 -0.144 0.8855   
## boroBROOKLYN -0.022239 0.037019 -0.601 0.5480   
## boroMANHATTAN -0.117731 0.049540 -2.377 0.0175 \*   
## boroQUEENS -0.035756 0.047439 -0.754 0.4510   
## boroSTATEN ISLAND 0.020945 0.091556 0.229 0.8190   
## occur\_hour 0.001628 0.001761 0.925 0.3550   
## vic\_age\_group1022 -10.591006 324.743703 -0.033 0.9740   
## vic\_age\_group18-24 0.261897 0.059435 4.406 1.05e-05 \*\*\*  
## vic\_age\_group25-44 0.601137 0.057386 10.475 < 2e-16 \*\*\*  
## vic\_age\_group45-64 0.759364 0.073527 10.328 < 2e-16 \*\*\*  
## vic\_age\_group65+ 0.992271 0.151986 6.529 6.63e-11 \*\*\*  
## vic\_age\_groupUNKNOWN 0.772756 0.313837 2.462 0.0138 \*   
## vic\_sexM -0.030337 0.049629 -0.611 0.5410   
## vic\_sexU -0.549268 1.079340 -0.509 0.6108   
## vic\_raceASIAN / PACIFIC ISLANDER 11.326862 89.668125 0.126 0.8995   
## vic\_raceBLACK 11.054359 89.668063 0.123 0.9019   
## vic\_raceBLACK HISPANIC 10.903067 89.668075 0.122 0.9032   
## vic\_raceUNKNOWN 10.211109 89.669046 0.114 0.9093   
## vic\_raceWHITE 11.372708 89.668102 0.127 0.8991   
## vic\_raceWHITE HISPANIC 11.169376 89.668069 0.125 0.9009   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 29251 on 29743 degrees of freedom  
## Residual deviance: 28943 on 29724 degrees of freedom  
## AIC: 28983  
##   
## Number of Fisher Scoring iterations: 11

## Generalized Linear Model (GLM) Output Explanation

This logistic regression model predicts the likelihood of a shooting incident resulting in murder (statistical\_murder\_flag) based on several predictors: borough (boro), time of occurrence (occur\_hour), victim’s age group, sex, and race.

### Key Findings

#### **Statistically Significant Predictors**

* **boroMANHATTAN**: Incidents in Manhattan are significantly less likely to result in murder compared to the baseline borough (likely the Bronx), with a small but statistically significant negative effect (p = 0.0175).
* **vic\_age\_group18-24, 25-44, 45-64, 65+, UNKNOWN**: All these age groups show statistically significant increased odds of the incident being a murder compared to the baseline (likely under 18), with the effect increasing with age.
* **vic\_age\_group18-24**: OR approximately exp(0.26) → slight increase.
* **vic\_age\_group25-44 and 45-64**: Higher estimates suggest a stronger relationship.
* **vic\_age\_group65+**: The strongest effect among the age groups (Estimate approximately 0.99).
* **vic\_age\_groupUNKNOWN**: Also significant, though likely due to data irregularities or small counts.

#### **Non-Significant Predictors**

* **Other boroughs (Brooklyn, Queens, Staten Island)**: No significant difference from the baseline.
* **occur\_hour**: No significant relationship between time of incident and likelihood of murder.
* **vic\_sex**: Neither male (M) nor unknown (U) differs significantly from the baseline (F).
* **vic\_race**: All race variables have extremely high standard errors and are not statistically significant. These inflated standard errors suggest **multicollinearity** or **complete/quasi-complete separation** in the data, leading to unreliable coefficient estimates.

### Model Fit

* **Null deviance**: 29251 → deviance of model with no predictors.
* **Residual deviance**: 28943 → deviance after including predictors.
* **AIC**: 28983 → a model selection metric; lower is better.
* **Model improvement** is minimal (small drop in deviance), suggesting limited predictive power from the included variables.

### Notes

* The very large standard errors and coefficients for race categories imply instability in the model, possibly due to sparse data or perfect prediction for certain groups.
* Only a few variables meaningfully contribute to predicting murder outcomes in the dataset, most notably certain age groups and location in Manhattan.

## 5 Personal Bias and Mitigation

**My Personal Bias**: As an analyst examining crime data, I may carry assumptions about the relationships between demographics, geography, and violence that could influence interpretation. Additionally, focusing on statistical patterns might overlook the human impact and community context of these incidents.

**Mitigation Strategies**: To address these biases, I have employed transparent methodology, avoided making causal claims about demographic relationships, focused on descriptive rather than prescriptive analysis, and explicitly acknowledged data limitations. The analysis emphasizes temporal and geographic patterns rather than making judgments about individual or group characteristics.

## 6 Conclusion

This analysis reveals several important patterns in NYC shooting incidents. The data shows a general decline in shooting incidents from 2011 to 2019, followed by a sharp increase in 2020, suggesting that broader social and economic factors significantly influence gun violence patterns. Geographic analysis indicates that Brooklyn and the Bronx experience the highest absolute numbers of incidents, while temporal analysis shows peak activity during evening and late-night hours.

The analysis is subject to limitations, primarily stemming from potential data biases as discussed. Future work could involve more advanced modeling techniques, deeper investigation into the “UNKNOWN” categories, incorporating external