Project Overview and Question of Interest

The data for this project is at https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic. It lists shooting incidents that occurred in NYC from 2006 through the end of the last year. Per the site:

> "This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes >for additional information about this dataset."

The first thing I'm interested in are exploring the overall counts and percentages for boroughs and precincts as tables. And that of victim's sex and race. Turning to visualization, I'm curious about how to best to present shootings by time of day; opting for a heat map. I then turn my attention to visualizing murder versus non-murder counts per year, expressed as a stacked bar chart. Finally, the model explores the probability (if shot) of getting murdered by borough.

1. Import libraries, data, and prepare.

```
library(tidyverse) # A collection of data science packages.
library(lubridate) # A package that makes it easier to work with dates and times.
library(knitr) # For tables in PDF/HTML.
```

```
# Import the data, reading the csv in as a data frame.
data0 = read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

```
# Choose the columns I think are relevant to my analysis.
data = data0 %>%
select (
   OCCUR_DATE,
   OCCUR_TIME,
   BORO,
   PRECINCT,
   STATISTICAL_MURDER_FLAG,
   PERP_AGE_GROUP,
   PERP_SEX,
   PERP_RACE,
   VIC_AGE_GROUP,
   VIC_SEX,
   VIC_RACE
)
```

```
# Replace empty values in the data set. Unsurprisingly, the perpetrator data is spotty.
data = data %>%
replace_na(list(PERP_AGE_GROUP = "UNKNOWN", PERP_SEX = "UNKNOWN", PERP_RACE = "UNKNOWN"))
```

2. Analyze

```
# Counts and percentages of shooting incidents by borough.
data %>%
count(BORO) %>%
mutate(Percent = round(n / sum(n) * 100, 1)) %>%
arrange(desc(n)) %>%
kable(caption = "Counts and Percentages by Borough")
```

Table 1: Counts and Percentages by Borough

BORO	n	Percent
BROOKLYN	11685	39.3
BRONX	8834	29.7
QUEENS	4426	14.9
MANHATTAN	3977	13.4
STATEN ISLAND	822	2.8

```
# Top 10 precincts with the most shootings. Sorted descending.
data %>%
count(PRECINCT, sort = TRUE) %>%
head(10) %>%
kable(caption = "Top 10 Precincts with the Most Shootings")
```

Table 2: Top 10 Precincts with the Most Shootings

PRECINCT	n
75	1680
73	1561
67	1288
44	1159
79	1073
47	1048
46	1044
40	1002
42	936
48	879

```r
# Victim sex distribution.
data %>%
count(VIC_SEX) %>%
mutate(Percent = round(n / sum(n) * 100, 1)) %>%
kable(caption = "Victim Sex Distribution")
```

Table 3: Victim Sex Distribution

VIC_SEX	n	Percent
F	2891	9.7
M	26841	90.2
U	12	0.0

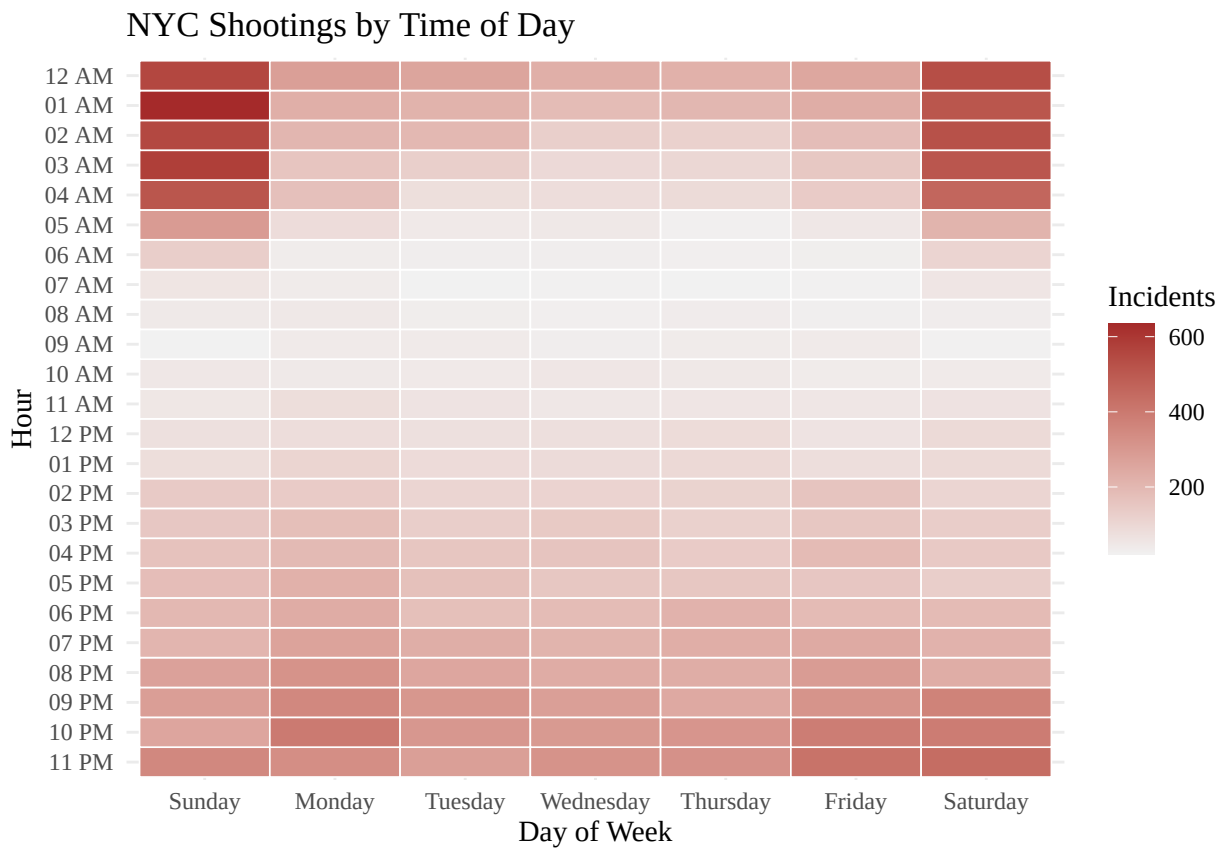
```r
# Victim race distribution.
data %>%
count(VIC_RACE) %>%
mutate(Percent = round(n / sum(n) * 100, 1)) %>%
arrange(desc(n)) %>%
kable(caption = "Victim Race Distribution")
```

Table 4: Victim Race Distribution

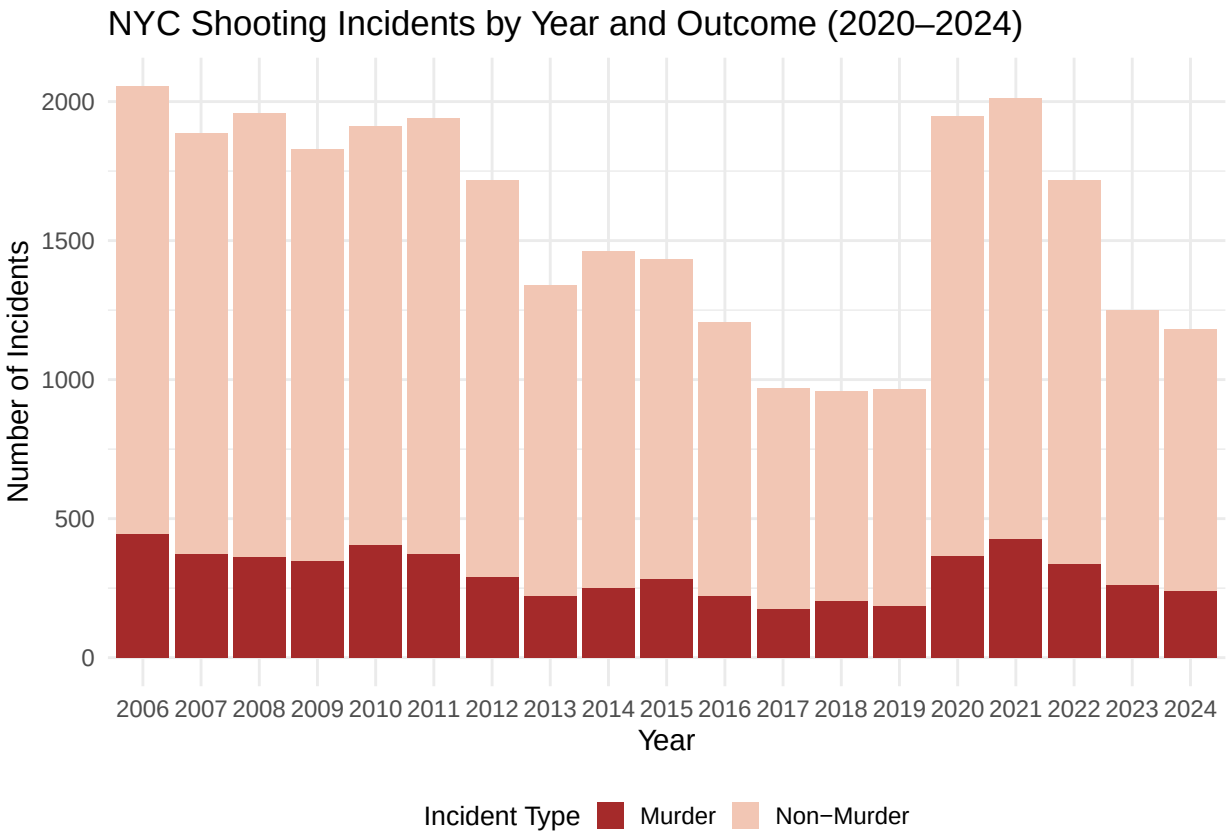
VIC_RACE	n	Percent
BLACK	20999	70.6
WHITE HISPANIC	4511	15.2
BLACK HISPANIC	2930	9.9
WHITE	741	2.5
ASIAN / PACIFIC ISLANDER	478	1.6
UNKNOWN	72	0.2
AMERICAN INDIAN/ALASKAN NATIVE	13	0.0

3. Visualize

```
# Visualize incidents by day of week and time of day.
hour_levels <- format(strptime(0:23, format = "%H"), "%I %p")
hour_levels <- factor(hour_levels, levels = rev(hour_levels))  # Midnight at top

# Prepare data
heat_data <- data %>%
mutate(
OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"),
HOUR = hour(hms(OCCUR_TIME)),
HOUR_LABEL = factor(format(strptime(HOUR, format = "%H"), "%I %p"), levels = levels(hour_levels)),
DOW = wday(OCCUR_DATE, label = TRUE, abbr = FALSE)
) %>%
count(DOW, HOUR_LABEL)

# Plot heatmap
ggplot(heat_data, aes(x = DOW, y = HOUR_LABEL, fill = n)) +
geom_tile(color = "white", linewidth = 0.3) +
scale_fill_gradient(low = "#F1F1F1", high = "brown", name = "Incidents") +
labs(
title = "NYC Shootings by Time of Day",
x = "Day of Week",
y = "Hour"
) + theme_minimal(base_family = "serif")
```



```r
# Count of shooting incidents per year and outcome.

# Prepare data
stacked_data <- data %>%
  mutate(
    OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"),
    YEAR = year(OCCUR_DATE),
    MURDER_FLAG = case_when(
      STATISTICAL_MURDER_FLAG == TRUE ~ "Murder",
      STATISTICAL_MURDER_FLAG == FALSE ~ "Non-Murder",
      TRUE ~ NA_character_
    )
  ) %>%
  filter(!is.na(YEAR), YEAR >= 2006 & YEAR <= 2024, !is.na(MURDER_FLAG)) %>%
  mutate(MURDER_FLAG = factor(MURDER_FLAG, levels = c("Non-Murder", "Murder"))) %>%
  count(YEAR, MURDER_FLAG)

# Plot with murders stacked at bottom
ggplot(stacked_data, aes(x = as.factor(YEAR), y = n, fill = MURDER_FLAG)) +
  geom_bar(stat = "identity") +
  labs(
    title = "NYC Shooting Incidents by Year and Outcome (2020-2024)",
    x = "Year",
    y = "Number of Incidents",
    fill = "Incident Type"
  ) +
  scale_fill_manual(
    values = c("Murder" = "brown", "Non-Murder" = "#f2c6b4"),
    breaks = c("Murder", "Non-Murder")  # Legend order: Murder first
  ) +
  theme_minimal(base_size = 11) +
  theme(
    legend.position = "bottom",
    legend.text = element_text(size = 9),
    legend.title = element_text(size = 10),
    legend.key.size = unit(0.4, "cm")
  )
```

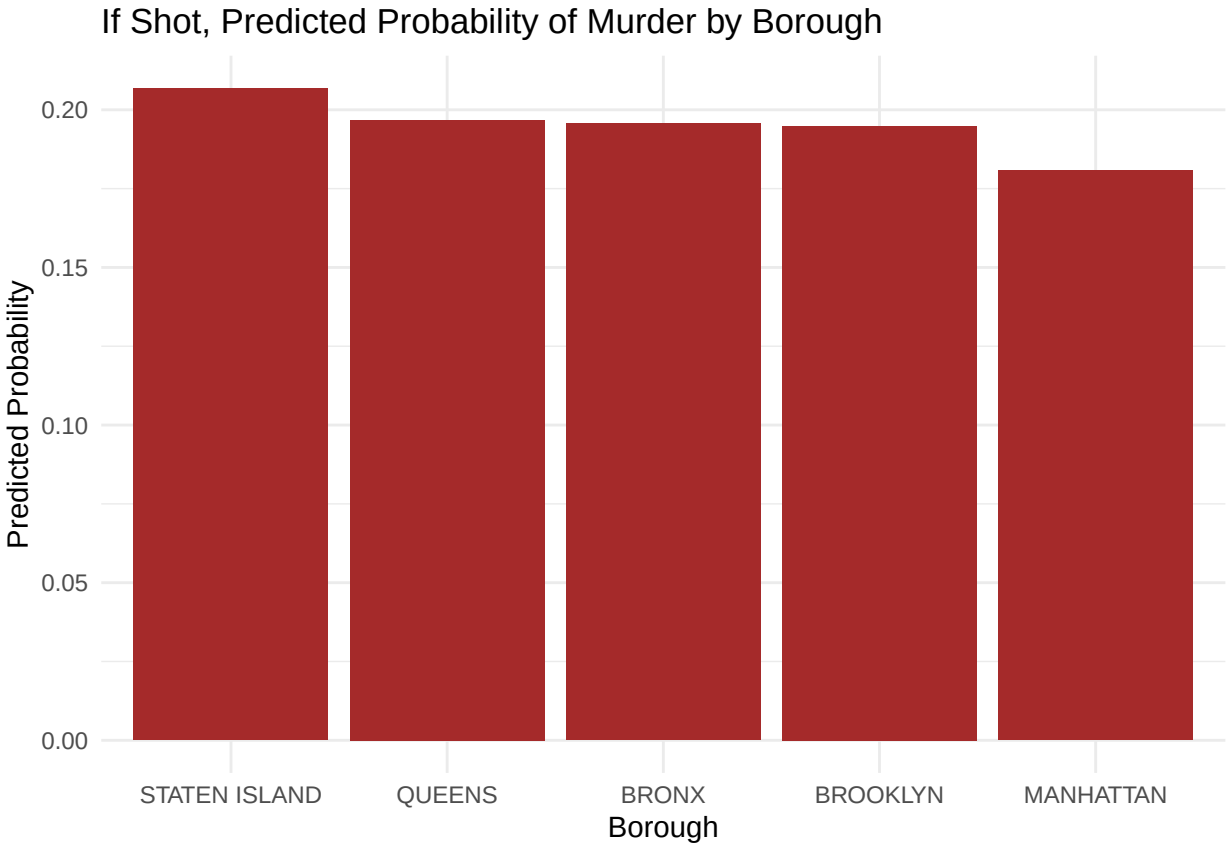

NYC Shooting Incidents by Year and Outcome (2020–2024)
Number of Incidents
2000
1500
1000
500
0
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024
Year
Incident Type
Murder
Non−Murder

4. Model

```r
# Model and plot probability of murder if shot per borough

# Prepare filter and format
data_clean <- data %>%
  mutate(
    MURDER_FLAG = if_else(STATISTICAL_MURDER_FLAG == TRUE, 1, 0),
    BORO = as.factor(BORO)
  ) %>%
  filter(!is.na(MURDER_FLAG), !is.na(BORO))

# Model: Predict murder likelihood by borough
model <- glm(MURDER_FLAG ~ BORO, data = data_clean, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = MURDER_FLAG ~ BORO, family = binomial, data = data_clean)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.413975   0.026822 -52.717   <2e-16 ***
## BOROBROOKLYN      -0.004727   0.035565  -0.133   0.8943
## BOROMANHATTAN     -0.097033   0.049165  -1.974   0.0484 *
## BOROQUEENS         0.007506   0.046355   0.162   0.8714
## BOROSTATEN ISLAND  0.069729   0.090197   0.773   0.4395
## ---
## 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: 29245  on 29739  degrees of freedom
## AIC: 29255
##
## Number of Fisher Scoring iterations: 4
```

```r
# Create new data frame for each borough
newdata <- data.frame(BORO = levels(data_clean$BORO))

# Predict probabilities for each borough
newdata$predicted_prob <- predict(model, newdata = newdata, type = "response")

# Plot predicted probabilities
ggplot(newdata, aes(x = reorder(BORO, -predicted_prob), y = predicted_prob)) +
  geom_col(fill = "brown") +
  labs(
    title = "If Shot, Predicted Probability of Murder by Borough",
    x = "Borough",
    y = "Predicted Probability"
  ) + theme_minimal(base_size = 11)
```


If Shot, Predicted Probability of Murder by Borough
Predicted Probability
0.20
0.15
0.10
0.05
0.00
STATEN ISLAND
QUEENS
BRONX
BROOKLYN
MANHATTAN
Borough

5. Conclusion and Possible Bias

In conclusion... a depressing data set! Exploring it, I looked at overall differences between the boroughs and precincts by count and percentages across all years. I explored the same with victims' sex and race. I visualized shootings by time of day. I compared murder to non-murder per year with a stacked bar chart. Finally, the model I chose was the probability of getting murdered (if shot) by borough.

Possible bias in the dataset:

- There are some data points missing for some categories, especially the perpetrator columns. These can skew group analysis.
- The dataset reflects reported shooting incidents. Those that are unreported or misclassified will be ommitted.
- Victims and/or witnesses may underreport due to a lack of trust in police, or some other fear.
- Police resources may not be evenly distributed across NYC.
- Neighborhoods with higher policing may have more recorded incidents. This may overrepresent minority communities, even if similar crimes happen elsewhere.

Possible bias in my analysis:

- My analysis may reflect COVID-era anomalies.
- In my heatmap, grouping by hour/day is useful but may not consider other patterns like season. I'm considering splitting these into small multiples per year.
- In my heatmap, readers may overinterpret minor gradients of color as more meaningful than it is.
- If I don't control for population density (for example, by normalizing per capita), I may just be highlighting more populous areas.