NYC Shooting Analysis

Nate Smik

2024-06-12

Introduction

The data analyzed in this report comes from the City of New York and represents historical shooting incidents across New York City from 2006 through the last calendar year. The purpose of the project is to clean and analyze the data and then create a model.

Based on an initial view of the data, my objective was to see if there was a predictor that would help determine if a shooting incident resulted in death.

Descriptions of Columns in Data Used:

- OCCUR_DATE: Exact date of the shooting incident
- OCCUR_TIME: Exact time of the shooting incident
- BORO: Borough where the shooting incident occurred
- STATISTICAL_MURDER_FLAG: Shooting resulted in the victim's death which would be counted as a murder
- VIC_AGE_GROUP: Victim's age withint a category
- VIC_SEX: Victim's sex description
- VIC_RACE: Victim's race description

Import Libraries and Data

First the necessary R libraries along with the NYPD dataset are imported.

```r
library(tidyverse)
library(lubridate)
library(ggplot2)
```

```r
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

NYC_data <- read_csv(url)
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Inspect and Clean Data

The initial data is checked for missing data and other potential anomalies to help determine how to properly clean the data.

```
head(NYC_data)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

Analyze and Account for Missing Data

Missing data is taken as a percentage of total data to get a better idea of any columns that lack sufficient data for analysis. Based on the amount of missing data some categories were excluded from further analysis such as data relating to the perp.

```
missing_counts <- colSums(is.na(NYC_data))
total_rows <- nrow(NYC_data)
missing_percentages <- round((missing_counts / total_rows) * 100, 1)
missing_percentages
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                     0.0                     0.0                     0.0
##                    BORO       LOC_OF_OCCUR_DESC                PRECINCT
##                     0.0                    89.6                     0.0
##       JURISDICTION_CODE      LOC_CLASSFCTN_DESC           LOCATION_DESC
##                     0.0                    89.6                    52.4
## STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP                PERP_SEX
##                     0.0                    32.7                    32.6
##               PERP_RACE           VIC_AGE_GROUP                 VIC_SEX
##                    32.6                     0.0                     0.0
##                VIC_RACE              X_COORD_CD              Y_COORD_CD
##                     0.0                     0.0                     0.0
##                Latitude               Longitude                 Lon_Lat
##                     0.2                     0.2                     0.2
```

Change Data Types and Remove Columns

The OCCUR_DATE column data is in the wrong format and must be changed to the date format. Columns that won't be used in the analysis are removed.

```
NYC_data <- NYC_data %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>% select(-c(LOC_OF_OCCUR_DESC, LOCATION_
```

Remove Bad Date Entries

There was one entry in age group "1022" that seemed to be an error in the data entry and was removed.

```
NYC_data <- NYC_data[NYC_data$VIC_AGE_GROUP != "1022", ]
```

Factorize Categorical Variables

Categorical variables are factorized to help in analysis and summarization of the data.

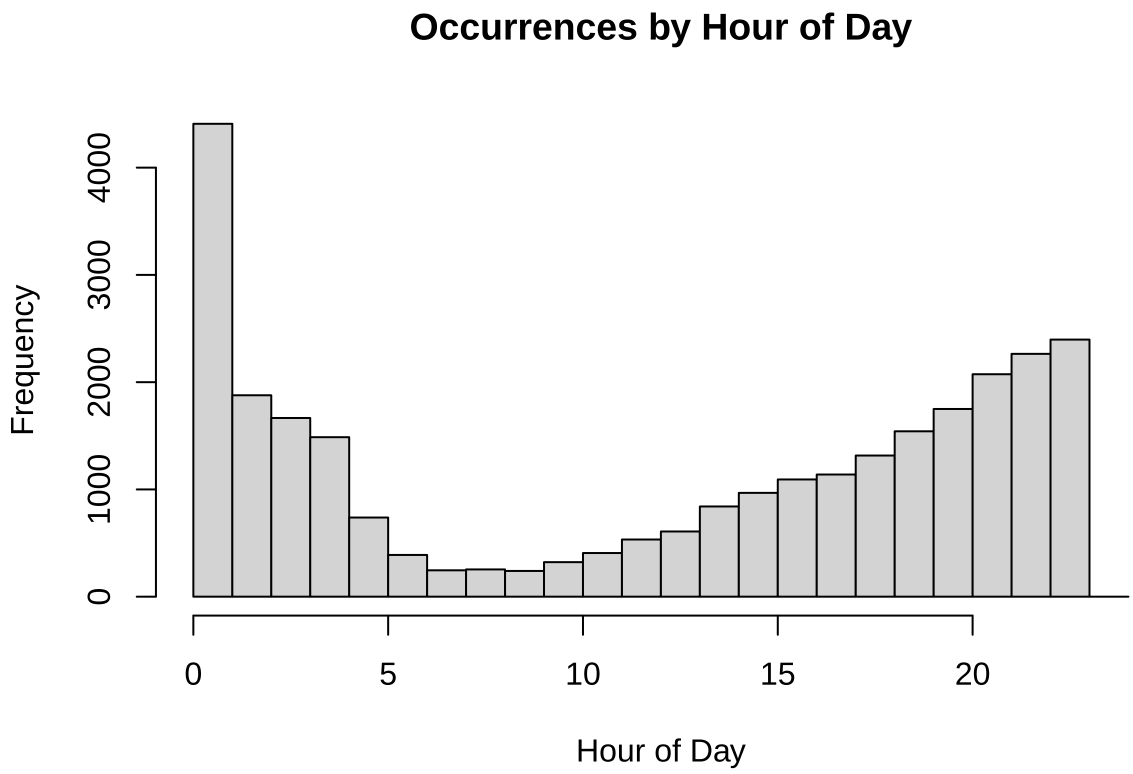
```
NYC_data$BORO = as.factor(NYC_data$BORO)
NYC_data$VIC_AGE_GROUP = as.factor(NYC_data$VIC_AGE_GROUP)
NYC_data$VIC_SEX = as.factor(NYC_data$VIC_SEX)
NYC_data$VIC_RACE = as.factor(NYC_data$VIC_RACE)

summary(NYC_data)
```

```
##    OCCUR_DATE          OCCUR_TIME                  BORO
##  Min.   :2006-01-01   Length:28561      BRONX        : 8376
##  1st Qu.:2009-09-04   Class1:hms        BROOKLYN     :11346
##  Median :2013-09-20   Class2:difftime   MANHATTAN    : 3761
##  Mean   :2014-06-07   Mode  :numeric    QUEENS       : 4271
##  3rd Qu.:2019-09-29                     STATEN ISLAND:  807
##  Max.   :2023-12-29
##
##  STATISTICAL_MURDER_FLAG VIC_AGE_GROUP   VIC_SEX
##  Mode :logical           <18     : 2954   F: 2760
##  FALSE:23035             18-24   :10384   M:25789
##  TRUE :5526              25-44   :12973   U:   12
##                          45-64   : 1981
##                          65+     :  205
##                          UNKNOWN:   64
##
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   11
##  ASIAN / PACIFIC ISLANDER      :  440
##  BLACK                         :20234
##  BLACK HISPANIC                : 2795
##  UNKNOWN                       :   70
##  WHITE                         :  728
##  WHITE HISPANIC                : 4283
```

Initial Analysis and Visualizations

Incidents are plotted on a histogram based on time of day to see if there is a pattern on when these incidents occur.

```
NYC_data$hour <- hour(NYC_data$OCCUR_TIME)
hist(NYC_data$hour, breaks = seq(0, 24, by = 1), main = "Occurrences by Hour of Day", xlab = "Hour of Da
```

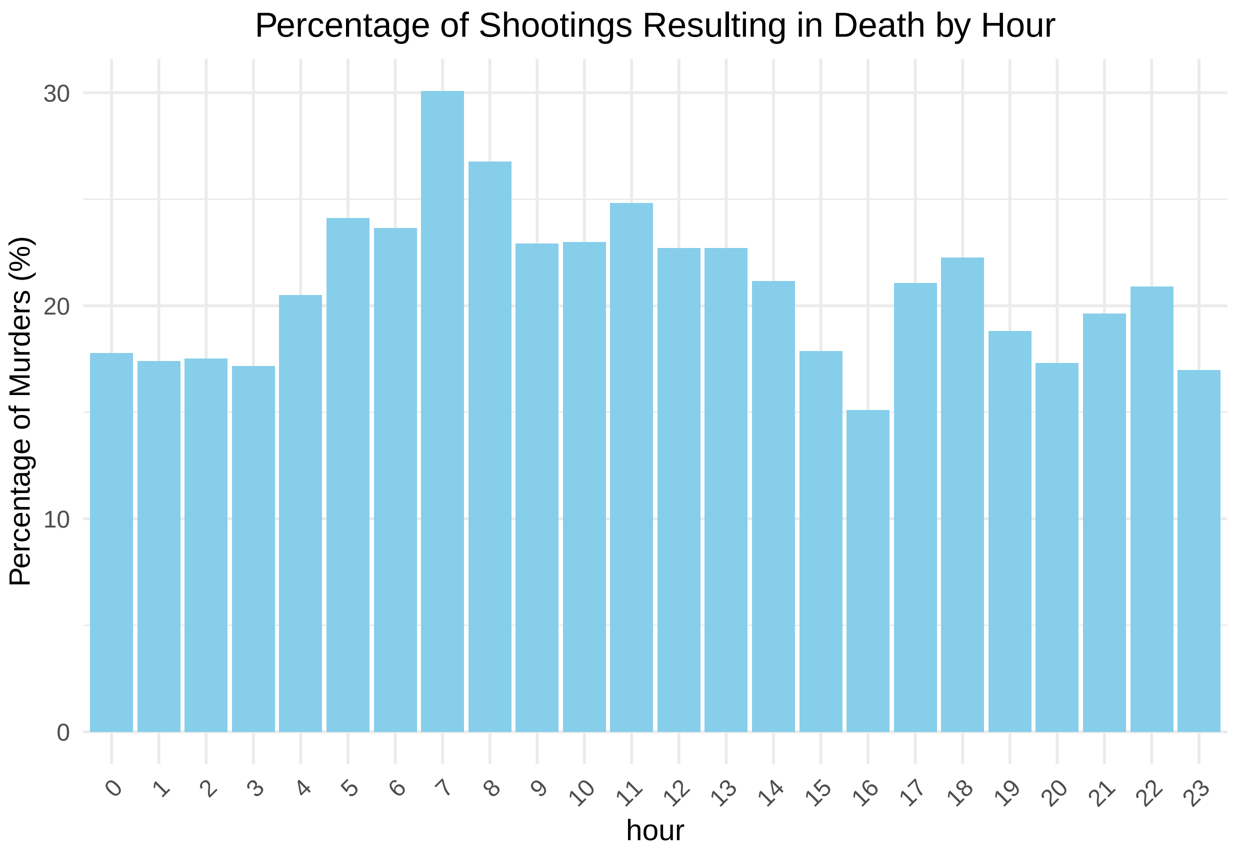


Most of the incidents appear to occur late night and early morning.

Percentage of Shootings Resulting in Death

Incidents are now transformed to see a percentage of shootings that end up as a murder and plotted against time to see how the distribution looks.

```
incidents_by_hour <- table(NYC_data$hour)

# Count murders by borough
murders_data <- subset(NYC_data, STATISTICAL_MURDER_FLAG == TRUE)
murders_by_hour <- table(murders_data$hour)

# Calculate percentage of murders relative to total incidents for each borough
percentage_murders_by_hour <- (murders_by_hour / incidents_by_hour) * 100

# Convert to data frame
percentage_murders_by_hour_df <- as.data.frame(percentage_murders_by_hour)
names(percentage_murders_by_hour_df) <- c("hour", "percentage")

# Plotting
ggplot(percentage_murders_by_hour_df, aes(x = hour, y = percentage)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Percentage of Shootings Resulting in Death by Hour",
       x = "hour",
       y = "Percentage of Murders (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))
```

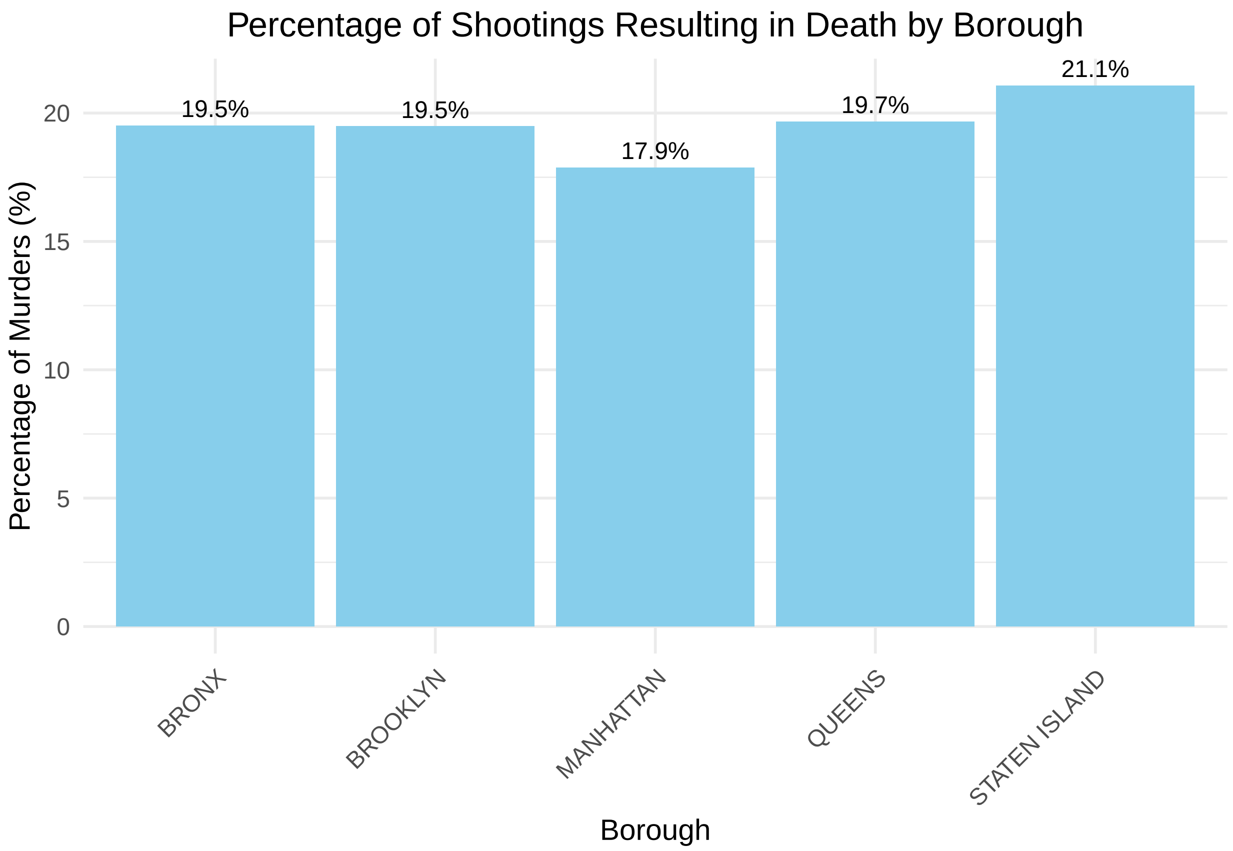


Based on the graph, it doesn't appear that time of day has a major impact on whether or not a shooting results in death.

The percentage of shootings resulting in death is now analyzed against the Borough, victim age, and victim sex for any potential relationships.

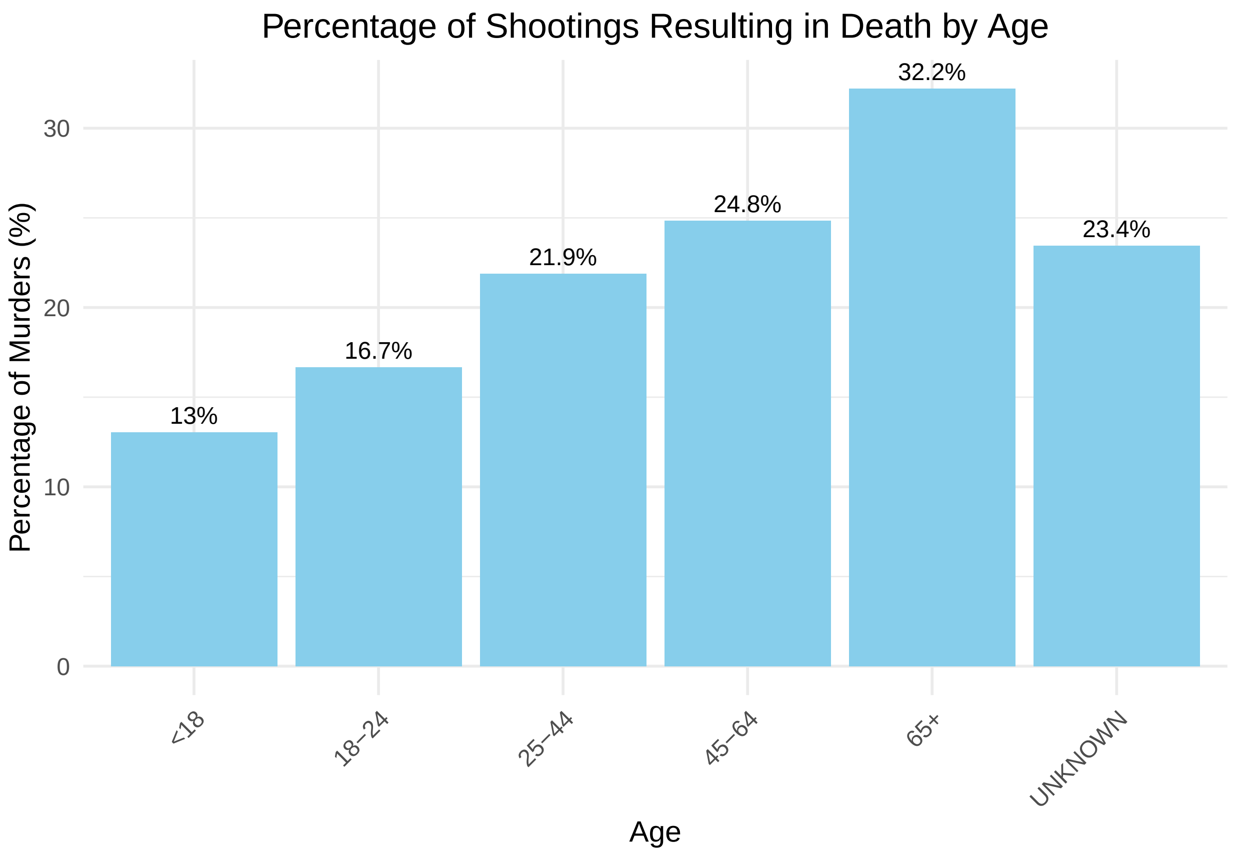
```
incidents_by_boro <- table(NYC_data$BORO)

# Count murders by borough
murders_data <- subset(NYC_data, STATISTICAL_MURDER_FLAG == TRUE)
murders_by_boro <- table(murders_data$BORO)

# Calculate percentage of murders relative to total incidents for each borough
percentage_murders_by_boro <- (murders_by_boro / incidents_by_boro) * 100

# Convert to data frame
percentage_murders_by_boro_df <- as.data.frame(percentage_murders_by_boro)
names(percentage_murders_by_boro_df) <- c("BORO", "percentage")
```

```
# Plotting
ggplot(percentage_murders_by_boro_df, aes(x = BORO, y = percentage)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5, size = 3) +
  labs(title = "Percentage of Shootings Resulting in Death by Borough",
       x = "Borough",
       y = "Percentage of Murders (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))
```

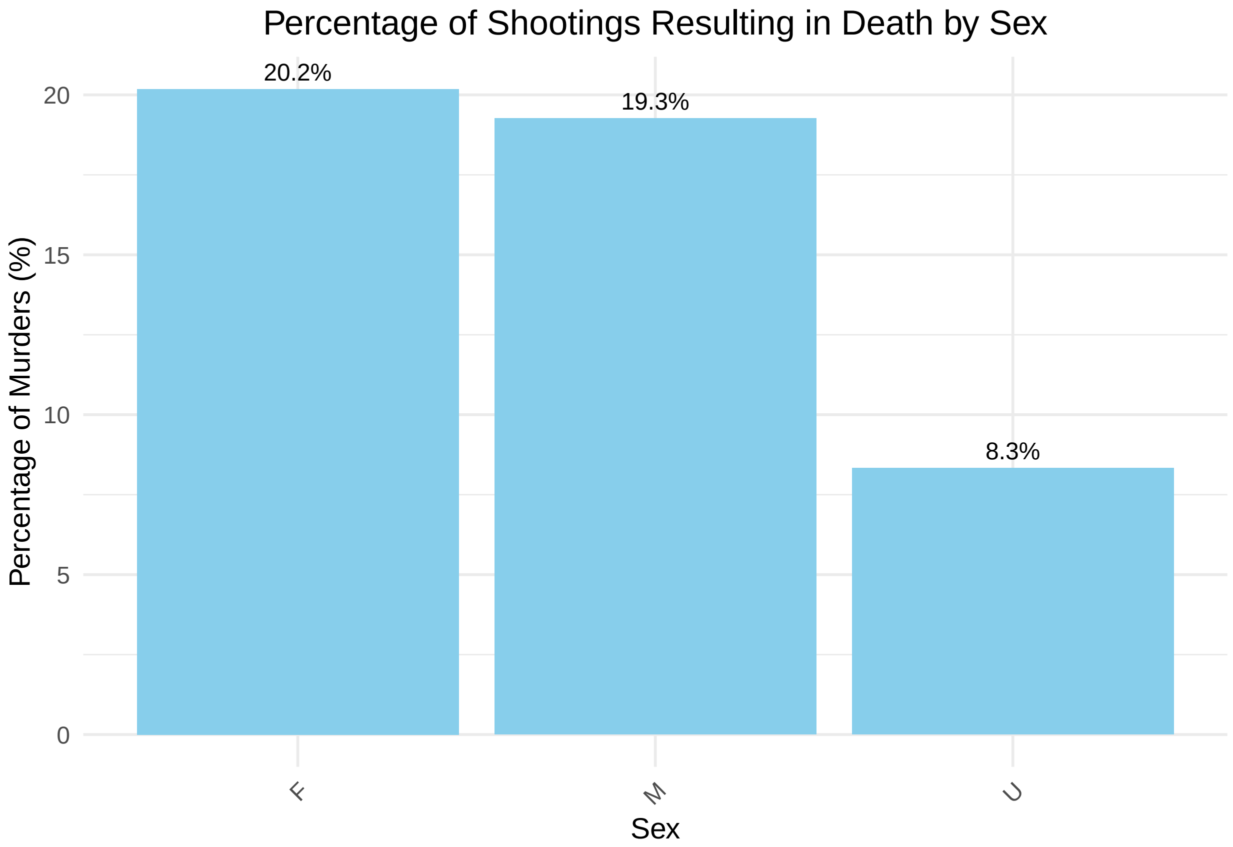


```
incidents_by_age <- table(NYC_data$VIC_AGE_GROUP)

# Count murders by borough
murders_data <- subset(NYC_data, STATISTICAL_MURDER_FLAG == TRUE)
murders_by_age <- table(murders_data$VIC_AGE_GROUP)

# Calculate percentage of murders relative to total incidents for each borough
percentage_murders_by_age <- (murders_by_age / incidents_by_age) * 100

# Convert to data frame
percentage_murders_by_age_df <- as.data.frame(percentage_murders_by_age)
names(percentage_murders_by_age_df) <- c("AGE", "percentage")

# Plotting
```

```
ggplot(percentage_murders_by_age_df, aes(x = AGE, y = percentage)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5, size = 3) +
  labs(title = "Percentage of Shootings Resulting in Death by Age",
       x = "Age",
       y = "Percentage of Murders (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))
```



```
incidents_by_sex <- table(NYC_data$VIC_SEX)

# Count murders by borough
murders_data <- subset(NYC_data, STATISTICAL_MURDER_FLAG == TRUE)
murders_by_sex <- table(murders_data$VIC_SEX)

# Calculate percentage of murders relative to total incidents for each borough
percentage_murders_by_sex <- (murders_by_sex / incidents_by_sex) * 100

# Convert to data frame
percentage_murders_by_sex_df <- as.data.frame(percentage_murders_by_sex)
names(percentage_murders_by_sex_df) <- c("Sex", "percentage")

# Plotting
ggplot(percentage_murders_by_sex_df, aes(x = Sex, y = percentage)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5, size = 3) +
  labs(title = "Percentage of Shootings Resulting in Death by Sex",
       x = "Sex",
       y = "Percentage of Murders (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the the visualizations it appears that age has the greatest relationship with incidents resulting in death.

Create a Model of the Data

A model of the data is created using a logistic regression analysis to see the potential relationship between variables (victim age, victim sex, victim race, borough, and time) and likelihood of murder based on the statistical murder flag.

```
logit_model <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX + VIC_RACE + BORO + hour , data =

summary(logit_model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX +
##     VIC_RACE + BORO + hour, family = binomial, data = NYC_data)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -12.911429  97.436396  -0.133  0.89458
## VIC_AGE_GROUP18-24                 0.287234   0.060926   4.714 2.42e-06 ***
## VIC_AGE_GROUP25-44                 0.620030   0.058961  10.516  < 2e-16 ***
## VIC_AGE_GROUP45-64                 0.758783   0.075960   9.989  < 2e-16 ***
## VIC_AGE_GROUP65+                   1.080636   0.160323   6.740 1.58e-11 ***
## VIC_AGE_GROUPUNKNOWN               0.862851   0.316400   2.727  0.00639 **
## VIC_SEXM                          -0.033894   0.050829  -0.667  0.50488
## VIC_SEXU                          -0.601540   1.081270  -0.556  0.57799
## VIC_RACEASIAN / PACIFIC ISLANDER  11.313954  97.436432   0.116  0.90756
## VIC_RACEBLACK                     11.044460  97.436371   0.113  0.90975
## VIC_RACEBLACK HISPANIC            10.876043  97.436383   0.112  0.91112
## VIC_RACEUNKNOWN                   10.219691  97.437277   0.105  0.91647
## VIC_RACEWHITE                     11.352670  97.436407   0.117  0.90725
## VIC_RACEWHITE HISPANIC            11.151021  97.436377   0.114  0.90889
## BOROBROOKLYN                      -0.022948   0.037838  -0.606  0.54420
## BOROMANHATTAN                     -0.125913   0.051090  -2.465  0.01372 *
## BOROQUEENS                        -0.035768   0.048435  -0.738  0.46023
## BOROSTATEN ISLAND                  0.044708   0.092011   0.486  0.62704
## hour                               0.001798   0.001793   1.003  0.31594
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28060  on 28560  degrees of freedom
## Residual deviance: 27764  on 28542  degrees of freedom
## AIC: 27802
##
## Number of Fisher Scoring iterations: 11
```

Based on the model it appears that age has the greatest impact on determining whether or not a shooting incident results in murder. The plot of age compared to percentage of incidents resulting in murder reflects this with an increase in percentage as age increases.

Possible Biases in Data

The data only includes those incidents that are reported so there may be incidents unreported for various reasons, especially incidents that don’t result in murder. It is unclear how many incidents go unreported and if it is of significance.

Although perp data was not used, there is possible bias in reporting the perp characteristics that could impact the data. Individual biases of witnesses could influence the identification of the perp.

Additional Resources

Data Descriptions: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8/about_data