NYPD Historic Shooting Incidents

Anonymous Student

2025-04-09

R Markdown

This is an R Markdown document on the NYPD Shooting Incident (Historic) found on the DATA.GOV website.

Getting the NYDP shooting data by loading the url

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Load the data
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nyc_data <- read.csv(url)
```

We will now have to look at the data structure and decide what we can do to clean it up if necessary.

```
# Preview structure
str(nyc_data)
```

```
## 'data.frame':    28562 obs. of  21 variables:
##  $ INCIDENT_KEY           : int  231974218 177934247 255028563 25384540 72616285 85875439 79780323 85
##  $ OCCUR_DATE             : chr  "08/09/2021" "04/07/2018" "12/02/2022" "11/19/2006" ...
##  $ OCCUR_TIME             : chr  "01:06:00" "19:48:00" "22:57:00" "01:50:00" ...
##  $ BORO                   : chr  "BRONX" "BROOKLYN" "BRONX" "BROOKLYN" ...
##  $ LOC_OF_OCCUR_DESC      : chr  "" "" "OUTSIDE" "" ...
```

```
##  $ PRECINCT               : int  40 79 47 66 46 42 71 69 75 69 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 2 0 2 0 0 ...
##  $ LOC_CLASSFCTN_DESC     : chr  "" "" "STREET" "" ...
##  $ LOCATION_DESC          : chr  "" "" "GROCERY/BODEGA" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "true" "false" "true" ...
##  $ PERP_AGE_GROUP         : chr  "" "25-44" "(null)" "UNKNOWN" ...
##  $ PERP_SEX               : chr  "" "M" "(null)" "U" ...
##  $ PERP_RACE              : chr  "" "WHITE HISPANIC" "(null)" "UNKNOWN" ...
##  $ VIC_AGE_GROUP          : chr  "18-24" "25-44" "25-44" "18-24" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  1006343 1000083 1020691 985107 1009854 ...
##  $ Y_COORD_CD             : num  234270 189065 257125 173350 247503 ...
##  $ Latitude               : num  40.8 40.7 40.9 40.6 40.8 ...
##  $ Longitude              : num  -73.9 -73.9 -73.9 -74 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.92019278899994 40.80967347200004)" "POINT (-73.94291302
```

```r
# Preview the data
summary(nyc_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class :character   Class :character
##  Median : 92711254   Mode  :character   Mode  :character   Mode  :character
##  Mean   :127405824
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Length:28562            Length:28562
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
```

```
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

```r
# Check column names
colnames(nyc_data)
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "LOC_OF_OCCUR_DESC"       "PRECINCT"
##  [7] "JURISDICTION_CODE"       "LOC_CLASSFCTN_DESC"
##  [9] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
## [11] "PERP_AGE_GROUP"          "PERP_SEX"
## [13] "PERP_RACE"               "VIC_AGE_GROUP"
## [15] "VIC_SEX"                 "VIC_RACE"
## [17] "X_COORD_CD"              "Y_COORD_CD"
## [19] "Latitude"                "Longitude"
## [21] "Lon_Lat"
```

```r
# Convert date columns
nyc_data$`OCCUR_DATE` <- as.Date(nyc_data$`OCCUR_DATE`, format = "%m/%d/%Y")

# Convert categorical columns to factor
nyc_data$`BORO` <- as.factor(nyc_data$`BORO`)
nyc_data$`LOC_OF_OCCUR_DESC` <- as.factor(nyc_data$`LOC_OF_OCCUR_DESC`)
nyc_data$`PERP_SEX` <- as.factor(nyc_data$`PERP_SEX`)
nyc_data$`PERP_RACE` <- as.factor(nyc_data$`PERP_RACE`)
nyc_data$`VIC_SEX` <- as.factor(nyc_data$`VIC_SEX`)
nyc_data$`VIC_RACE` <- as.factor(nyc_data$`VIC_RACE`)
nyc_data$`STATISTICAL_MURDER_FLAG` <- as.factor(nyc_data$`STATISTICAL_MURDER_FLAG`)
```

```r
# Remove unnecessary columns
nyc_clean <- nyc_data %>%
  select(-c(`Latitude`, `Longitude`, `Lon_Lat`, `X_COORD_CD`, `Y_COORD_CD`))
# Check cleaned data
str(nyc_clean)
```

```
## 'data.frame':    28562 obs. of  16 variables:
##  $ INCIDENT_KEY          : int  231974218 177934247 255028563 25384540 72616285 85875439 79780323 85
##  $ OCCUR_DATE            : Date, format: "2021-08-09" "2018-04-07" ...
##  $ OCCUR_TIME            : chr  "01:06:00" "19:48:00" "22:57:00" "01:50:00" ...
```

```
##  $ BORO                    : Factor w/ 5 levels "BRONX","BROOKLYN",..: 1 2 1 2 1 1 2 2 2 2 ...
##  $ LOC_OF_OCCUR_DESC       : Factor w/ 3 levels "","INSIDE","OUTSIDE": 1 1 3 1 1 1 1 1 1 1 ...
##  $ PRECINCT                : int  40 79 47 66 46 42 71 69 75 69 ...
##  $ JURISDICTION_CODE       : int  0 0 0 0 0 2 0 2 0 0 ...
##  $ LOC_CLASSFCTN_DESC      : chr  "" "" "STREET" "" ...
##  $ LOCATION_DESC           : chr  "" "" "GROCERY/BODEGA" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: Factor w/ 2 levels "false","true": 1 2 1 2 2 1 2 1 1 1 ...
##  $ PERP_AGE_GROUP          : chr  "" "25-44" "(null)" "UNKNOWN" ...
##  $ PERP_SEX                : Factor w/ 5 levels "","(null)","F",..: 1 4 2 5 4 4 1 1 4 4 ...
##  $ PERP_RACE               : Factor w/ 9 levels "","(null)","AMERICAN INDIAN/ALASKAN NATIVE",..: 1 9 1
##  $ VIC_AGE_GROUP           : chr  "18-24" "25-44" "25-44" "18-24" ...
##  $ VIC_SEX                 : Factor w/ 3 levels "F","M","U": 2 2 2 2 1 2 2 2 2 2 ...
##  $ VIC_RACE                : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 3 3 3 3 3 3 3 7 3
```

```
# Summary of cleaned data
summary(nyc_clean)
```

```
##   INCIDENT_KEY          OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class :character
##  Median : 92711254   Median :2013-09-20   Mode  :character
##  Mean   :127405824   Mean   :2014-06-07
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##
##             BORO       LOC_OF_OCCUR_DESC    PRECINCT     JURISDICTION_CODE
##  BRONX        : 8376          :25596     Min.   :  1.0   Min.   :0.0000
##  BROOKLYN     :11346   INSIDE :  460     1st Qu.: 44.0   1st Qu.:0.0000
##  MANHATTAN    : 3762   OUTSIDE: 2506     Median : 67.0   Median :0.0000
##  QUEENS       : 4271                     Mean   : 65.5   Mean   :0.3219
##  STATEN ISLAND:  807                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                                          Max.   :123.0   Max.   :2.0000
##                                                          NA's   :2
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:28562       Length:28562       false:23036
##  Class :character   Class :character   true : 5526
##  Mode  :character   Mode  :character
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX              PERP_RACE     VIC_AGE_GROUP
##  Length:28562             : 9310   BLACK         :11903   Length:28562
##  Class :character   (null): 1141                 : 9310   Class :character
##  Mode  :character   F     :  444   WHITE HISPANIC: 2510   Mode  :character
##                     M     :16168   UNKNOWN       : 1837
##                     U     : 1499   BLACK HISPANIC: 1392
##                                    (null)        : 1141
##                                    (Other)       :  469
##  VIC_SEX                              VIC_RACE
##  F: 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  M:25790   ASIAN / PACIFIC ISLANDER      :  440
##  U:   12   BLACK                         :20235
##            BLACK HISPANIC                : 2795
```

```
##             UNKNOWN                    :   70
##             WHITE                      :  728
##             WHITE HISPANIC             : 4283
```

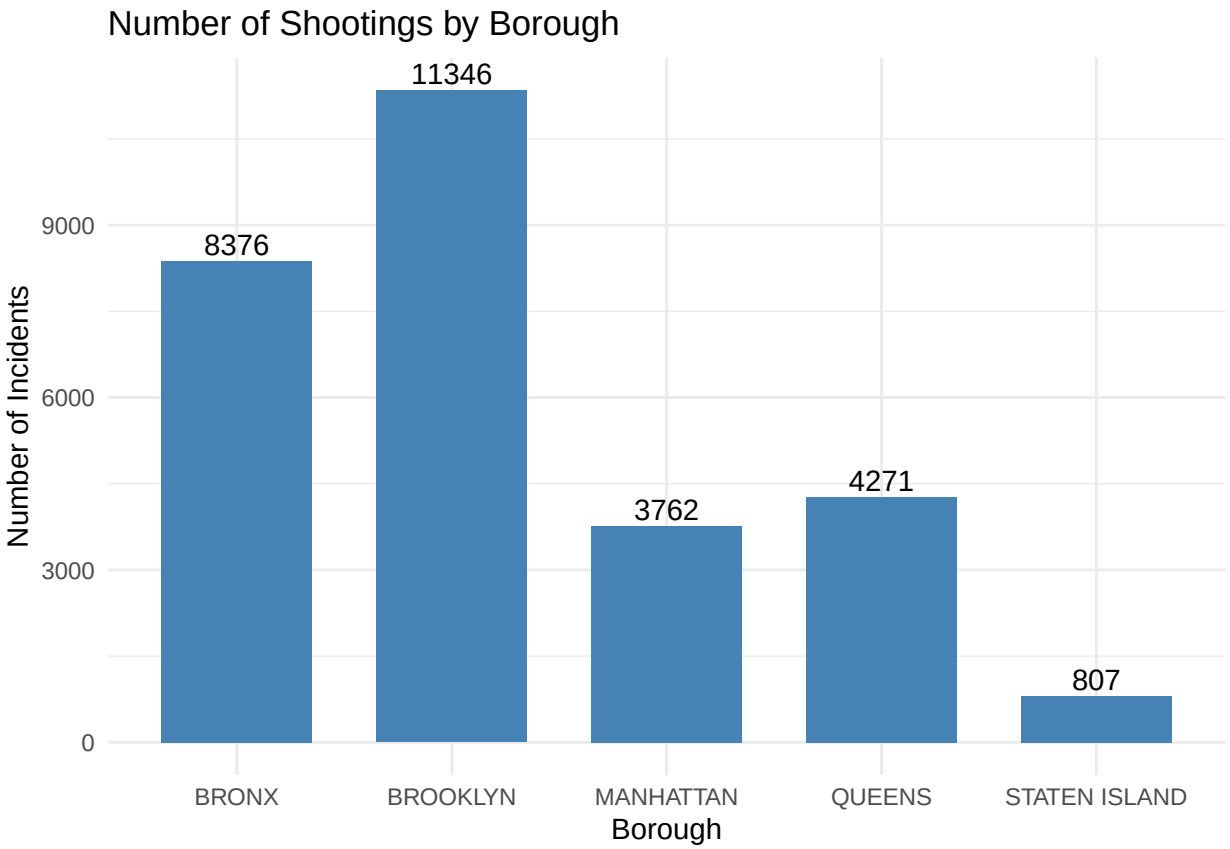
Ways to handle missing data

By observing our data, we can see where unknown or missing values are located. Common ways to handle situations like this are:

1. We can remove rows with missing values, this may cause a loss of a lot of data if many rows have missing values.
2. We can impute missing values, this will only work if we have sensible replacements and I wouldn't advise if you are not fully familiar with the database.
3. Leave the missing values in, some functions can handle it smoothly and may even exclude it automatically.

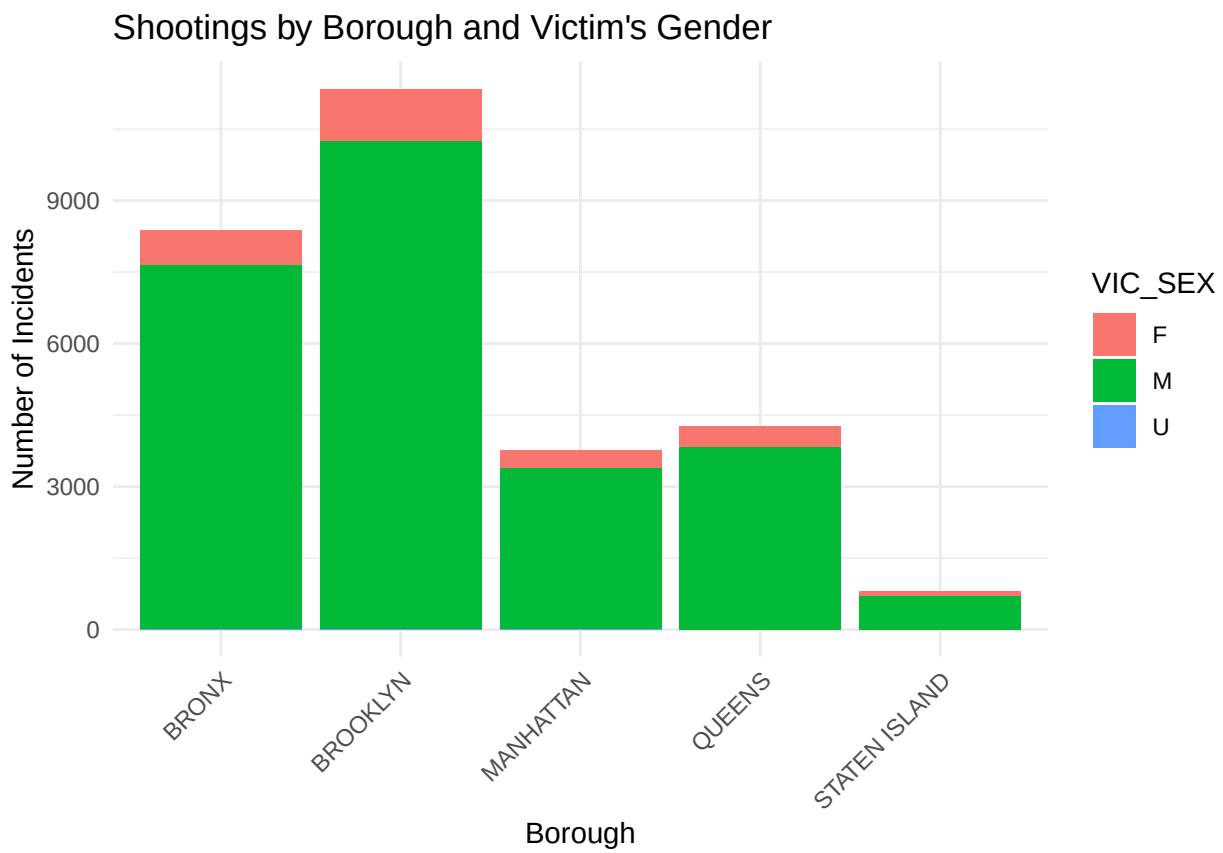
Visualization of the data

```r
library(ggplot2)

# Show how many incidents in each borough
ggplot(nyc_clean, aes(x = BORO)) +
  geom_bar(fill = "steelblue", width = 0.7) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.3) +
  theme_minimal() +
  labs(title = "Number of Shootings by Borough", x = "Borough", y = "Number of Incidents")
```



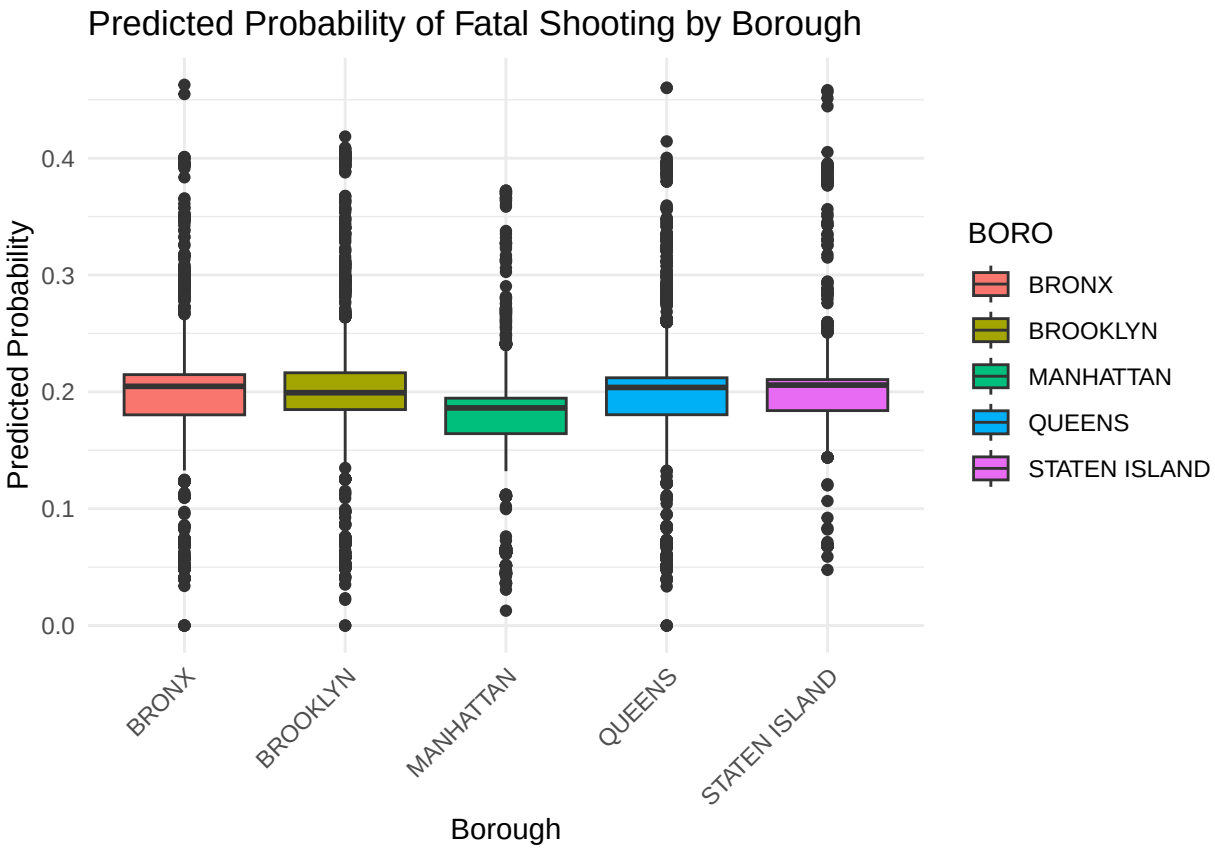
The bar graph clearly shows that historically, Brooklyn and the Bronx are where majority of shooting incidents occur. The two boroughs combined for almost 70 percent of all shooting incidents that occurred in the city.

```
library(ggplot2)

# Stacked bar plot for BORO vs VIC_SEX
ggplot(nyc_clean, aes(x = BORO, fill = VIC_SEX)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Shootings by Borough and Victim's Gender", x = "Borough", y = "Number of Incidents") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The stacked bar plot above shows that across all boroughs, the likelihood of the victim of a shooting incident being male is extremely high.25,790 of the victims are males, this is an astonishing 90 percent of all incidents. Further research into male involvement in gang affiliations or access to weapons could provide some insight for this.

```
# Load necessary library
library(caret)
# a logistic regression model
# change the murder variable to a binomial
nyc_clean$STATISTICAL_MURDER_FLAG <- as.factor(ifelse(nyc_clean$STATISTICAL_MURDER_FLAG == "true", 1, 0
fatal_model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + VIC_SEX + VIC_RACE + PERP_SEX + PERP_RACE + OCCUR_DA
                   data = nyc_clean,
                   family = "binomial")

# View the model summary
#summary(fatal_model)

# Predicted probabilities for each row in the data
nyc_clean$pred_probs <- predict(fatal_model, type = "response")

# Plot the predicted probabilities against one of the predictors (e.g., BORO)
ggplot(nyc_clean, aes(x = BORO, y = pred_probs, fill = BORO)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Predicted Probability of Fatal Shooting by Borough",
       x = "Borough", y = "Predicted Probability") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the predicted probabilities, we can observe that the chances of the incident being fatal is fairly low, with all the boroughs showing around the 20 percent fatality rate.

Conclusion

Looking at the historical data for NYC, it is observed that majority of the shooting incidents occurred in the Bronx and Brooklyn, however when looking at the chances of a fatal shooting occurring, all boroughs are approximately in the same probability range. Males are by far the gender most affected by these incidents, also observed across all boroughs. The biases from my analysis would be that I am not familiar with the demographic, political and socio-economic factors that may be affecting the city. So any inference I can make will be strictly from a numerical standpoint. The biases in the data could be in the sample (precincts reporting, neighborhoods with higher police activity), we could also have over representation of shootings involving police but a misrepresentation of shootings that occurred in lower profile areas.If a variable was available to show if persons involved were gang affiliated could give some insights as well. This being historical data may also reflect past prejudices/racial profiling that may be carried forward, especially in modeling.