NYShootings

David

2025-01-28

```
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     1.0.2
## -- 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
```

```
library(dplyr)
library(ggplot2)
```

NYPD Shootings Data

Step 1: Importing and Describing Data

Data is sourced from here: https://catalog.data.gov/dataset

The dataset can be found here: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

This dataset covers every shooting in NYC from 2006-2023 and includes location, time, and other relevant details surrounding the event.

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

```
## 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.
```

Step 2: Data Summary and Clean Up

Removing columns that won't be used.

```
nypd_shooting_data <- nypd_shooting_data %>%
  select(-c('INCIDENT_KEY','JURISDICTION_CODE','X_COORD_CD',
            'Y_COORD_CD',Latitude,Longitude,Lon_Lat, OCCUR_TIME))
```

Changing the formatting of the date column.

```
nypd_shooting_data$Year <- year(mdy(nypd_shooting_data$OCCUR_DATE))
```

Cleaning up bad data

```
nypd_shooting_data <- nypd_shooting_data %>%

  #changing unknowns in sex columns to 'not stated' for later factoring
  mutate(
    VIC_SEX = recode(VIC_SEX, 'F'='F', 'M'='M', .default = 'Not Stated'),
    PERP_SEX = recode(PERP_SEX, 'F'='F', 'M'='M', .default = 'Not Stated')
        ) %>%

  #replacing unknowns and other bad values with NA
  mutate(across(-c(OCCUR_DATE, Year, PRECINCT, STATISTICAL_MURDER_FLAG,
                   VIC_SEX, PERP_SEX),
                ~ na_if(.,'(null)'))) %>%
  mutate(across(-c(OCCUR_DATE, Year, PRECINCT, STATISTICAL_MURDER_FLAG,
                   VIC_SEX, PERP_SEX),
                ~ na_if(.,'UNKNOWN'))) %>%
  mutate(across(-c(OCCUR_DATE,Year,, PRECINCT, STATISTICAL_MURDER_FLAG,
                   VIC_SEX, PERP_SEX),
                ~ na_if(.,'U')))

#replacing specific bad values that I found with NA
nypd_shooting_data$PERP_AGE_GROUP[nypd_shooting_data$PERP_AGE_GROUP %in% c('1020','1028','224','940')] =

nypd_shooting_data$VIC_AGE_GROUP[nypd_shooting_data$VIC_AGE_GROUP %in%
                                   c('1022')] = NA
```

Changing NA values in PERP_SEX to 'Not Stated'

```
nypd_shooting_data$PERP_SEX <- nypd_shooting_data$PERP_SEX %>%
#replacing NA in PERP_SEX with 'Not Stated'
   replace_na('Not Stated')
```

Factoring appropriate columns.

```
nypd_shooting_data <- nypd_shooting_data %>%
  mutate(across(c(BORO,PRECINCT, PERP_AGE_GROUP, PERP_SEX, PERP_RACE,
                  VIC_AGE_GROUP, VIC_SEX, VIC_RACE,LOC_CLASSFCTN_DESC,
                  LOCATION_DESC, STATISTICAL_MURDER_FLAG), factor))
nypd_shooting_data = droplevels(nypd_shooting_data) #drop unused levels
```

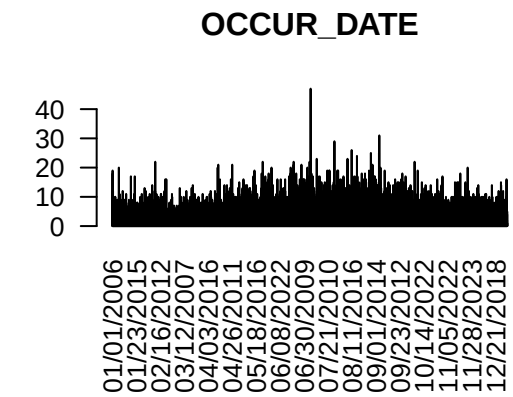
Summary Time!

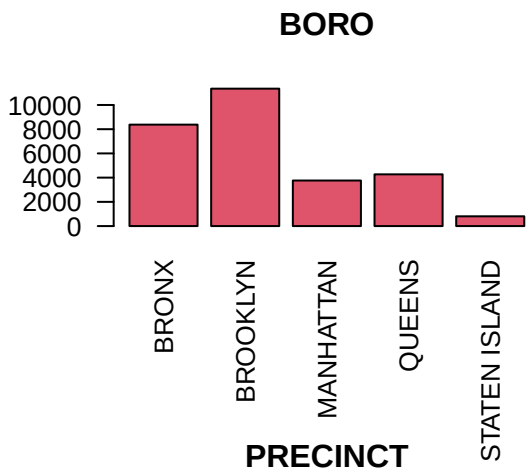
```
summary(nypd_shooting_data)
```

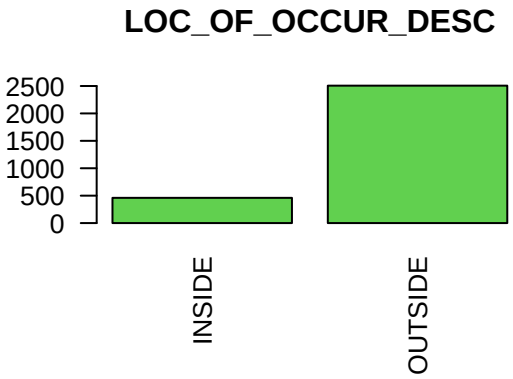
```
##   OCCUR_DATE                   BORO       LOC_OF_OCCUR_DESC     PRECINCT
##  Length:28562       BRONX        : 8376   Length:28562       75     : 1628
##  Class :character   BROOKLYN     :11346   Class :character   73     : 1500
##  Mode  :character   MANHATTAN    : 3762   Mode  :character   67     : 1259
##                     QUEENS       : 4271                      44     : 1076
##                     STATEN ISLAND:  807                      79     : 1045
##                                                              47     : 1006
##                                                              (Other):21048
##   LOC_CLASSFCTN_DESC                   LOCATION_DESC   STATISTICAL_MURDER_FLAG
##  STREET    : 1886    MULTI DWELL - PUBLIC HOUS: 5007   FALSE:23036
##  HOUSING   :  460    MULTI DWELL - APT BUILD  : 2964   TRUE : 5526
##  DWELLING  :  243    PVT HOUSE                :  983
##  COMMERCIAL:  208    GROCERY/BODEGA           :  750
##  OTHER     :   59    BAR/NIGHT CLUB           :  668
##  (Other)   :  108    (Other)                  : 1502
##  NA's      :25598    NA's                     :16688
##  PERP_AGE_GROUP       PERP_SEX                              PERP_RACE
##  <18  : 1682    F          :  444   AMERICAN INDIAN/ALASKAN NATIVE:    2
##  18-24: 6438    M          :16168   ASIAN / PACIFIC ISLANDER      :  169
##  25-44: 6041    Not Stated:11950   BLACK                         :11903
##  45-64:  699                       BLACK HISPANIC                : 1392
##  65+  :   65                       WHITE                         :  298
##  NA's :13637                       WHITE HISPANIC                : 2510
##                                    NA's                          :12288
##  VIC_AGE_GROUP       VIC_SEX                              VIC_RACE
##  <18  : 2954   F          : 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  18-24:10384   M          :25790   ASIAN / PACIFIC ISLANDER      :  440
##  25-44:12973   Not Stated:   12   BLACK                         :20235
##  45-64: 1981                       BLACK HISPANIC                : 2795
##  65+  :  205                       WHITE                         :  728
##  NA's :   65                       WHITE HISPANIC                : 4283
##                                    NA's                          :   70
##       Year
##  Min.   :2006
##  1st Qu.:2009
##  Median :2013
##  Mean   :2014
##  3rd Qu.:2019
##  Max.   :2023
##
```

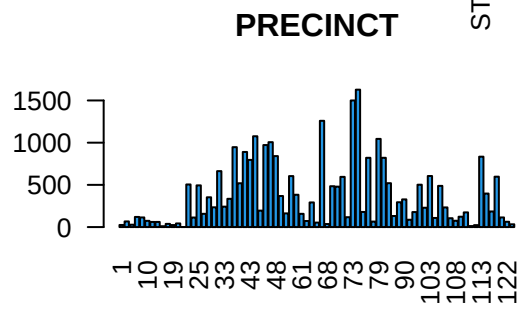
Step 3: Data Analysis and Visualization

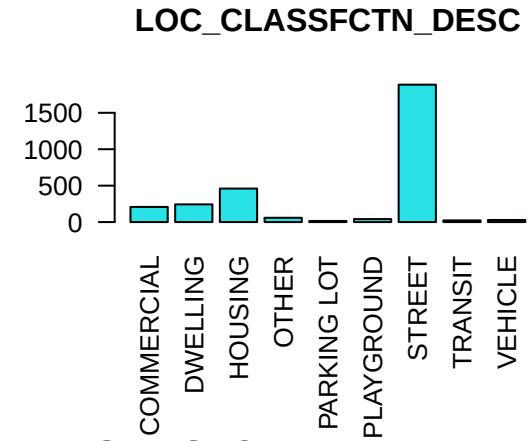
```
#just getting a look at the data (graphs are messy but that's okay)
par(mfrow = c(2,2))
for (i in 1:13){
      barplot(table(nypd_shooting_data[i]), col = i, las = 2, main = colnames(nypd_shooting_data)[i])
}
```

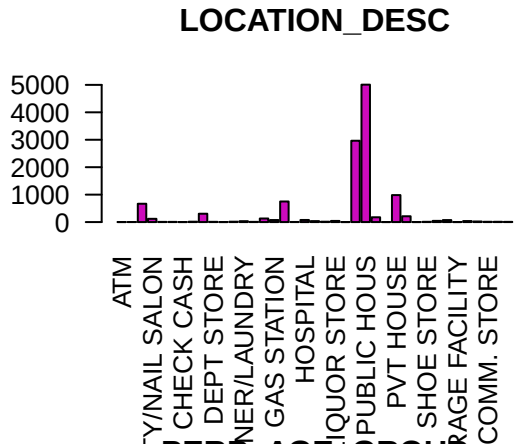

OCCUR_DATE
40
30
20
10
0
01/01/2006
01/23/2015
02/16/2012
03/12/2007
04/03/2016
04/26/2011
05/18/2016
06/08/2022
06/30/2009
07/21/2010
08/11/2016
09/01/2014
09/23/2012
10/14/2022
11/05/2022
11/28/2023
12/21/2018

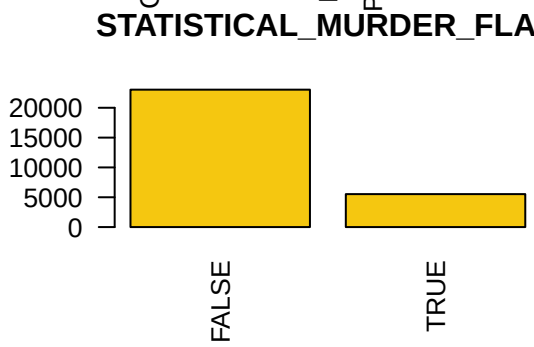

BORO
10000
8000
6000
4000
2000
0
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND

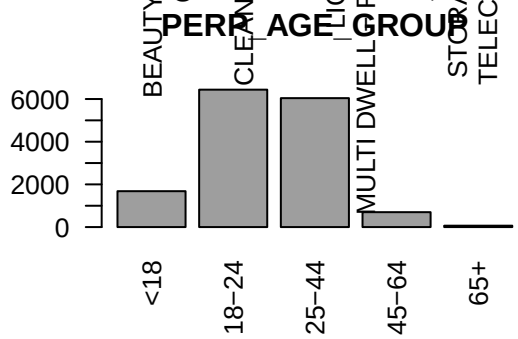

LOC_OF_OCCUR_DESC
2500
2000
1500
1000
500
0
INSIDE
OUTSIDE

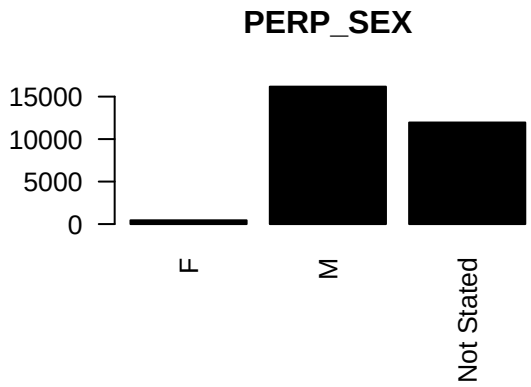

PRECINCT
1500
1000
500
0
1
10
19
25
33
43
48
61
68
73
79
90
103
108
113
122

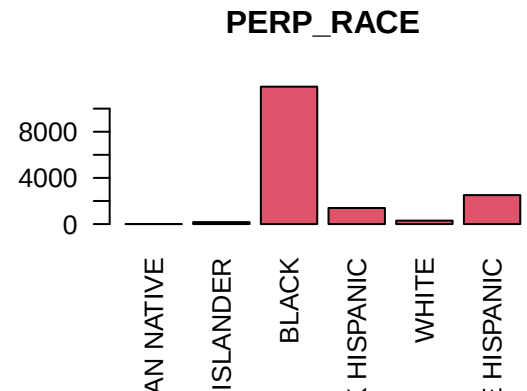

LOC_CLASSFCTN_DESC
1500
1000
500
0
COMMERCIAL
DWELLING
HOUSING
OTHER
PARKING LOT
PLAYGROUND
STREET
TRANSIT
VEHICLE

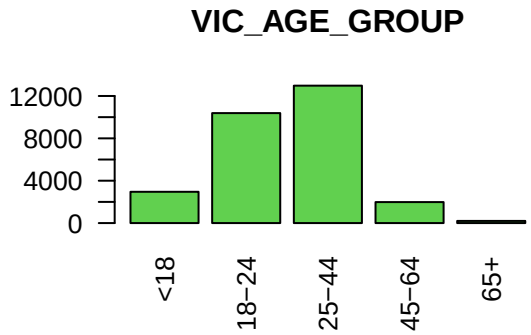

LOCATION_DESC
5000
4000
3000
2000
1000
0
ATM
BEAUTY/NAIL SALON
CHECK CASH
DEPT STORE
CLEANER/LAUNDRY
GAS STATION
HOSPITAL
LIQUOR STORE
PUBLIC HOUS
PVT HOUSE
SHOE STORE
STORAGE FACILITY
TELECOMM. STORE

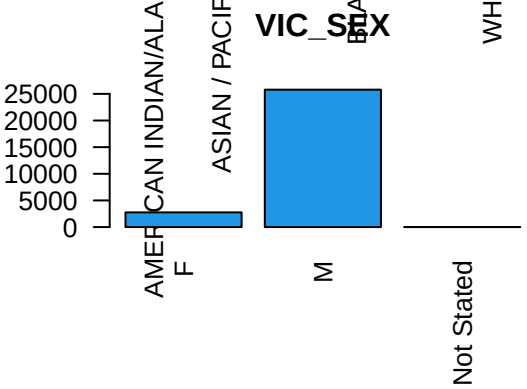

STATISTICAL_MURDER_FLAG
20000
15000
10000
5000
0
FALSE
TRUE

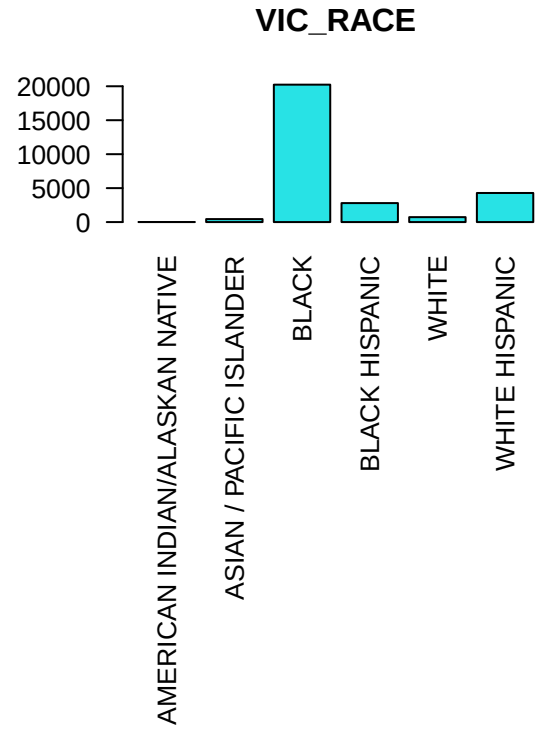

PERP_AGE_GROUP
6000
4000
2000
0
<18
18−24
25−44
45−64
65+


PERP_SEX
15000
10000
5000
0
F
M
Not Stated


PERP_RACE
8000
4000
0
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN / PACIFIC ISLANDER
BLACK
BLACK HISPANIC
WHITE
WHITE HISPANIC


VIC_AGE_GROUP
12000
8000
4000
0
<18
18−24
25−44
45−64
65+

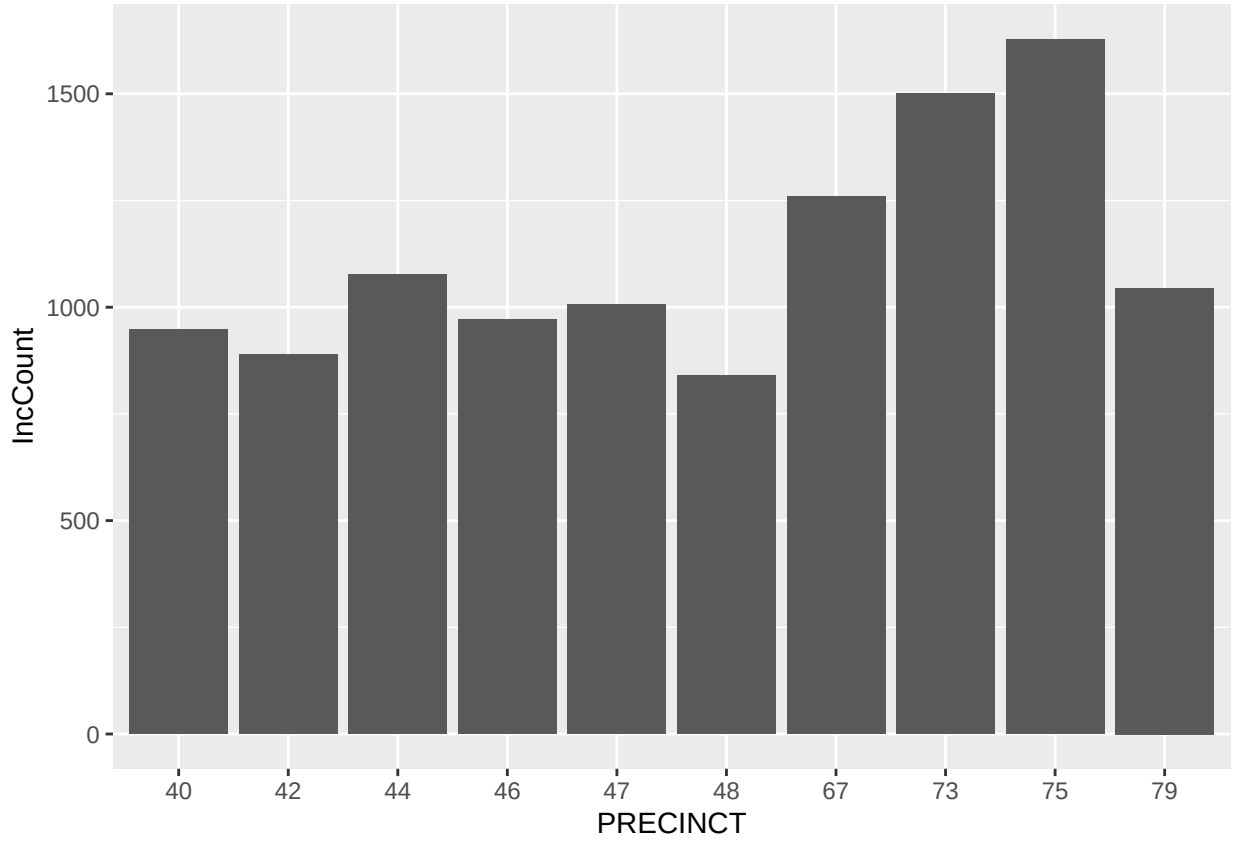

VIC_SEX
25000
20000
15000
10000
5000
0
F
M
Not Stated



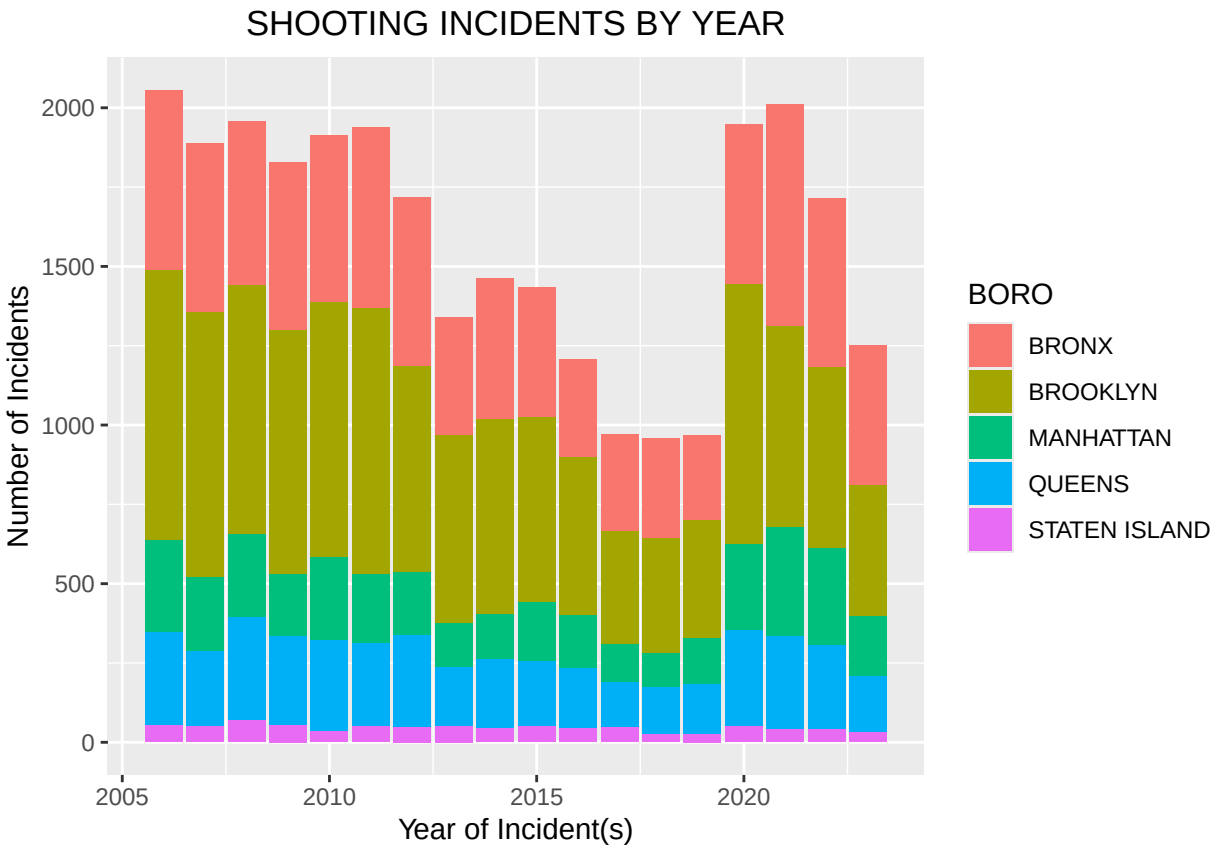

```
#looking at percentage of incidents in precincts
nypd_shooting_data <- nypd_shooting_data %>%
  add_count(BORO, PRECINCT, name = 'IncCount')

incidentsPrec <- distinct(nypd_shooting_data, PRECINCT,.keep_all = TRUE)
incidentsPrec <- incidentsPrec %>%
  mutate(
    percentage= ((IncCount)/nrow(nypd_shooting_data) *100)) %>%
  select(c(BORO ,PRECINCT, IncCount, percentage))
#isolating top 10 highest inccount precincts
top10precincts <- head(incidentsPrec[order(-incidentsPrec$IncCount),],n=10)
top10precincts
```

```
## # A tibble: 10 x 4
##    BORO     PRECINCT IncCount percentage
##    <fct>    <fct>       <int>      <dbl>
##  1 BROOKLYN 75           1628       5.70
##  2 BROOKLYN 73           1500       5.25
##  3 BROOKLYN 67           1259       4.41
##  4 BRONX    44           1076       3.77
##  5 BROOKLYN 79           1045       3.66
##  6 BRONX    47           1006       3.52
##  7 BRONX    46            972       3.40
##  8 BRONX    40            947       3.32
##  9 BRONX    42            890       3.12
## 10 BRONX    48            841       2.94
```

```r
#plot of highest incident precincts
PrecinctPlot <- ggplot(top10precincts,aes(PRECINCT,IncCount)) + geom_col()
print(PrecinctPlot)
```




```r
#Incidents per Year by Borough
IncidentsPerYear <- ggplot(nypd_shooting_data, aes(unclass(Year), fill = BORO)) +
  geom_bar() +
  ggtitle('SHOOTING INCIDENTS BY YEAR') +
  ylab('Number of Incidents') +
  xlab('Year of Incident(s)') +
  theme(plot.title = element_text(hjust = 0.5))
print(IncidentsPerYear)
```



Step 4: Conclusion and Bias Sources

From the data we find that the Bronx and Brooklyn, together, have all of the top 10 highest incident count precincts. This is further shown in the Shooting Incidents By Year graph. This graph also shows an interesting trend where shooting incidents rocket back up in 2020 after continually declining since the mid 2000s. Sources of bias in this analysis come include the data collectors (NYPD). Is it possible that Manhattan, Queens, and Staten Island have less police on staff therefore lowering shooting incidents reported? This is just one of many questions that could be raised surrounding the bias of this data.