nypd_report.Rmd

Valentina Anthonio

2023-06-22

Data Science as Field NYPD SHOOTING INCIDENT DATA project

This report presents the findings from the nypd shooting data which covers shooting incidences of five different boroughs spanning from January 01 2006 to December 31 2022. The report focuses on the the distribution of cases by the hour, location and gender distribution.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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(lubridate)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(dplyr)
library(knitr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(ggtext)
library(ggpie)
```

Data Sources

Import data

The data used for this analysis was extracted from DATA.GOV

```
url_in <-"https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
df_ny <-read_csv(url_in)
```

```
## Rows: 27312 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.
```

Data Cleaning and Exploration

```
head(df_ny,5)
```

```
## # A tibble: 5 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO   LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>  <chr>                <dbl>
## 1    228798151 05/27/2021 21:30      QUEENS <NA>                   105
## 2    137471050 06/27/2014 17:40      BRONX  <NA>                    40
## 3    147998800 11/21/2015 03:56      QUEENS <NA>                   108
## 4    146837977 10/09/2015 18:30      BRONX  <NA>                    44
## 5     58921844 02/19/2009 22:58      BRONX  <NA>                    47
## # 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>
```

```
#drop some columns from the data
df_ny <- df_ny %>% select(-c(INCIDENT_KEY,PRECINCT,JURISDICTION_CODE,LOC_OF_OCCUR_DESC,LOC_CLASSFCTN_DES
#replace n/a values with unknown
df_ny <- df_ny %>% replace_na(list(PERP_AGE_GROUP="Unknown",PERP_SEX="Unknown", PERP_RACE="Unknown", VIC
```

```
#merge n/a, null and unknown values into one.
df_ny$PERP_AGE_GROUP = recode(df_ny$PERP_AGE_GROUP, UNKNOWN ="Unknown")
df_ny$PERP_AGE_GROUP = recode(df_ny$PERP_AGE_GROUP,"(null)" = "Unknown")
df_ny$PERP_AGE_GROUP = recode(df_ny$PERP_AGE_GROUP, "N/A" = "Unknown")
df_ny$VIC_AGE_GROUP = recode(df_ny$VIC_AGE_GROUP, UNKNOWN = "UnKnown")
df_ny$VIC_AGE_GROUP = as.factor(df_ny$VIC_AGE_GROUP)
df_ny$PERP_SEX = recode(df_ny$PERP_SEX,"(null)" = "Unknown")
df_ny$VIC_RACE = recode(df_ny$VIC_RACE,UNKNOWN ="Unknown")
df_ny$LOCATION_DESC = recode(df_ny$LOCATION_DESC,"(null)" = "Unknown")

# convert TRUE/FALSE to 1 and 0
df_ny$STATISTICAL_MURDER_FLAG[df_ny$STATISTICAL_MURDER_FLAG=="TRUE"] <-1
df_ny$STATISTICAL_MURDER_FLAG[df_ny$STATISTICAL_MURDER_FLAG =="FALSE"]<- 0

#convert date column of date formate
df_ny$OCCUR_DATE <- mdy(df_ny$OCCUR_DATE)
df_ny$OCCUR_TIME <- hour(hms(as.character(df_ny$OCCUR_TIME)))

#rename columns
df_ny <- df_ny %>% rename(DATE = OCCUR_DATE,
                      TIME = OCCUR_TIME)
head(df_ny)
```

```
## # A tibble: 6 x 11
##   DATE        TIME BORO     LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##   <date>     <dbl> <chr>    <chr>                           <dbl> <chr>
## 1 2021-05-27    21 QUEENS   Unknown                             0 Unknown
## 2 2014-06-27    17 BRONX    Unknown                             0 Unknown
## 3 2015-11-21     3 QUEENS   Unknown                             1 Unknown
## 4 2015-10-09    18 BRONX    Unknown                             0 Unknown
## 5 2009-02-19    22 BRONX    Unknown                             1 25-44
## 6 2020-10-21    21 BROOKLYN Unknown                             1 Unknown
## # i 5 more variables: PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <fct>,
## #   VIC_SEX <chr>, VIC_RACE <chr>
```

```
# get the age groupings of victims and perpetrators
table(df_ny$VIC_AGE_GROUP)
```

```
##
##     <18     1022    18-24   25-44   45-64      65+ UnKnown
##    2839        1    10086   12281    1863      181      61
```

```
table(df_ny$PERP_AGE_GROUP)
```

```
##
##     <18     1020    18-24      224    25-44    45-64      65+      940 Unknown
##    1591        1     6222        1     5687      617       60        1   13132
```

```
#drop the outliers in the age groups
df_ny <- df_ny[!(df_ny$PERP_AGE_GROUP ==1020 ),]
df_ny <- df_ny[!(df_ny$PERP_AGE_GROUP ==940 ),]
df_ny <- df_ny[!(df_ny$PERP_AGE_GROUP ==224 ),]
df_ny <- df_ny[!(df_ny$VIC_AGE_GROUP ==1022),]


table(df_ny$VIC_AGE_GROUP)
```

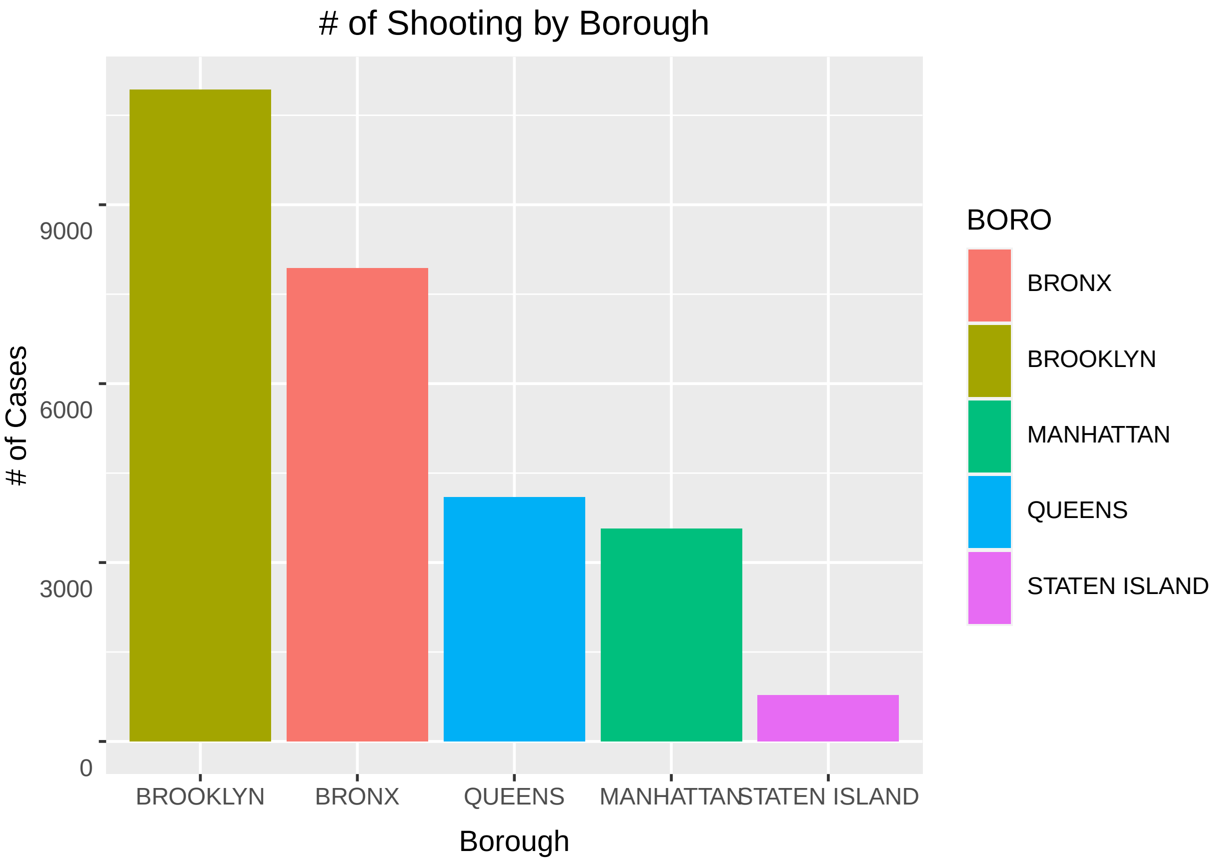
```
##
##     <18    1022   18-24   25-44   45-64     65+ UnKnown
##    2839       0   10085   12279    1863     181      61
```

```
table(df_ny$PERP_AGE_GROUP)
```

```
##
##     <18   18-24   25-44   45-64     65+ Unknown
##    1591    6221    5687     617      60   13132
```

Data Exploration

```
#plot data points
#get cases group by sex
shooting <- df_ny %>% group_by(BORO) %>% summarize(incidents = n())

ggplot(shooting,aes(reorder(BORO,-incidents), y= incidents, fill=BORO)) +geom_bar(stat = "identity",posi
  ggtitle("# of Shooting by Borough") +theme(plot.title = element_text(hjust = 0.5), axis.text.y = eleme
```



From the graph above, its evident that most of the shootings took place in the Brooklyn followed by the Bronx with Staten Island being the least.

```
# get the number of shooting incidents by location
table(df_ny$LOCATION_DESC)
```

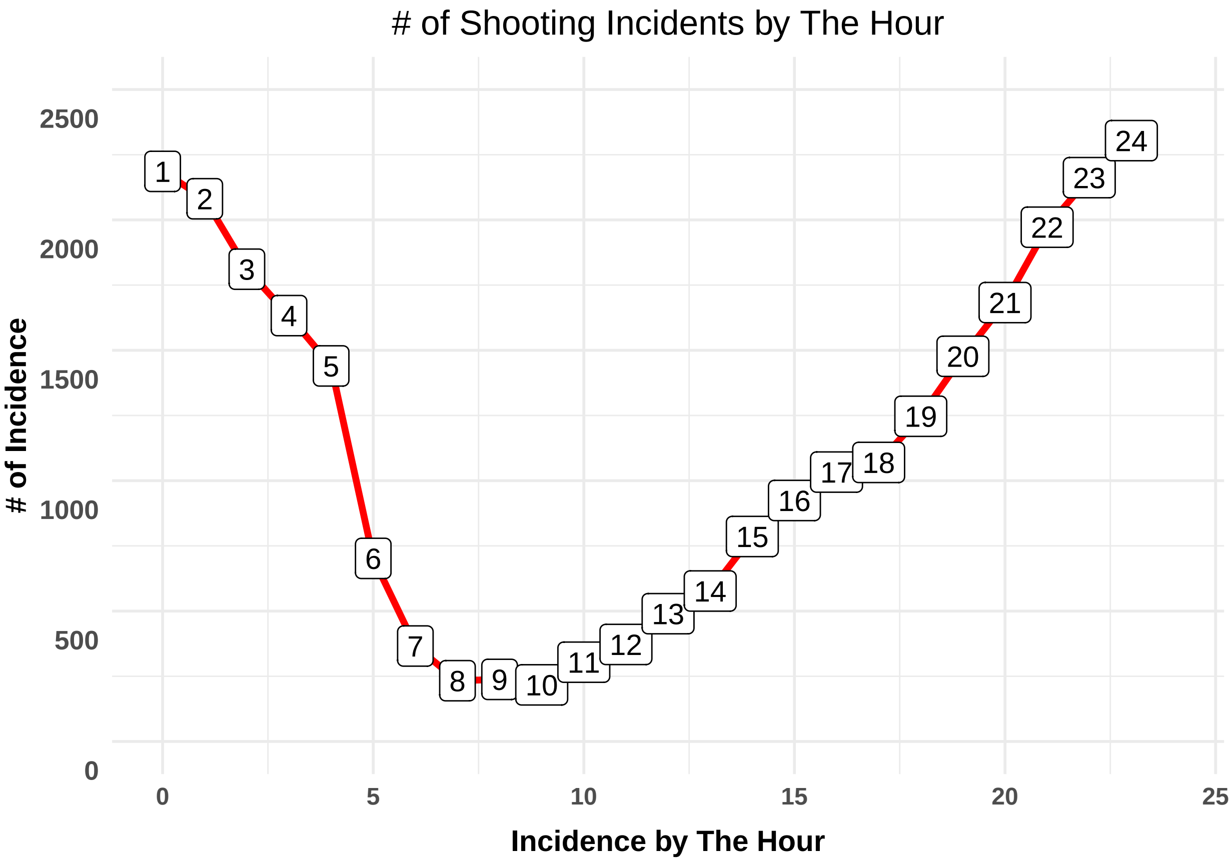
```
##
##                       ATM                      BANK            BAR/NIGHT CLUB
##                         1                         3                       627
##         BEAUTY/NAIL SALON               CANDY STORE               CHAIN STORE
##                       112                         7                         5
##                CHECK CASH          CLOTHING BOUTIQUE           COMMERCIAL BLDG
##                         1                        14                       292
##                DEPT STORE           DOCTOR/DENTIST                DRUG STORE
##                         9                         1                        14
##       DRY CLEANER/LAUNDRY         FACTORY/WAREHOUSE                 FAST FOOD
##                        31                         8                       104
##               GAS STATION            GROCERY/BODEGA     GYM/FITNESS FACILITY
##                        70                       694                         3
##                  HOSPITAL               HOTEL/MOTEL             JEWELRY STORE
##                        65                        35                        12
##              LIQUOR STORE              LOAN COMPANY  MULTI DWELL - APT BUILD
##                        41                         1                      2835
## MULTI DWELL - PUBLIC HOUS                      NONE         PHOTO/COPY STORE
##                      4831                       175                         1
##                 PVT HOUSE         RESTAURANT/DINER                    SCHOOL
##                       951                       204                         1
##                SHOE STORE            SMALL MERCHANT SOCIAL CLUB/POLICY LOCATI
##                        10                        37                        72
##          STORAGE FACILITY        STORE UNCLASSIFIED               SUPERMARKET
##                         1                        36                        21
##           TELECOMM. STORE                   Unknown             VARIETY STORE
##                        11                     15953                        11
##               VIDEO STORE
##                         8
```

```
shooting<- df_ny %>%
  group_by(LOCATION_DESC) %>%
  reframe(total_shootings =n(), paste(round(total_shootings/(count(df_ny))*100,2), "%")) %>%
  arrange(desc(total_shootings))

shooting<- shooting %>% rename(percentage = 3)

shooting
```

```
## # A tibble: 40 x 3
##    LOCATION_DESC             total_shootings percentage
##    <chr>                               <int> <chr>
##  1 Unknown                             15953 58.42 %
##  2 MULTI DWELL - PUBLIC HOUS            4831 17.69 %
##  3 MULTI DWELL - APT BUILD              2835 10.38 %
##  4 PVT HOUSE                             951 3.48 %
##  5 GROCERY/BODEGA                        694 2.54 %
##  6 BAR/NIGHT CLUB                        627 2.3 %
##  7 COMMERCIAL BLDG                       292 1.07 %
##  8 RESTAURANT/DINER                      204 0.75 %
##  9 NONE                                  175 0.64 %
## 10 BEAUTY/NAIL SALON                     112 0.41 %
## # i 30 more rows
```

58.4% of most shootings had no specific location information but multi dwelling public places had **17.69%** following by multi dwelling apartment of **10.38%**.

```
#check which day of the week has most cases
defaultW <-getOption("warn")
options(warning=-1)

# get shooting data by the hour
df_hour <- df_ny %>%  group_by(TIME) %>% count()

ggplot(df_hour, aes(x=TIME, y=n)) + geom_line(color="red", linewidth=1.2) + labs(x= "Incidence by The H
```



The number of cases witnessed a significant decline between 2015 and 2020 in the early late hours of the day to midnight as per the graph above. It would be interesting to investigate the reasons behind this decline to assist in future planning for preventing such incidents.

```
# perpetrator age distribution
options(repr.plot.width = 25, repr.plot.height=25)

perp_b<- ggplot(data=df_ny)  +  geom_bar(mapping = aes(x=PERP_AGE_GROUP, fill=PERP_AGE_GROUP), show.lege

perp_flip <- perp_b + coord_flip()+
  theme(axis.text.x = element_blank())  + geom_text(aes(x=PERP_AGE_GROUP, label = ..count..), size=3, f
```

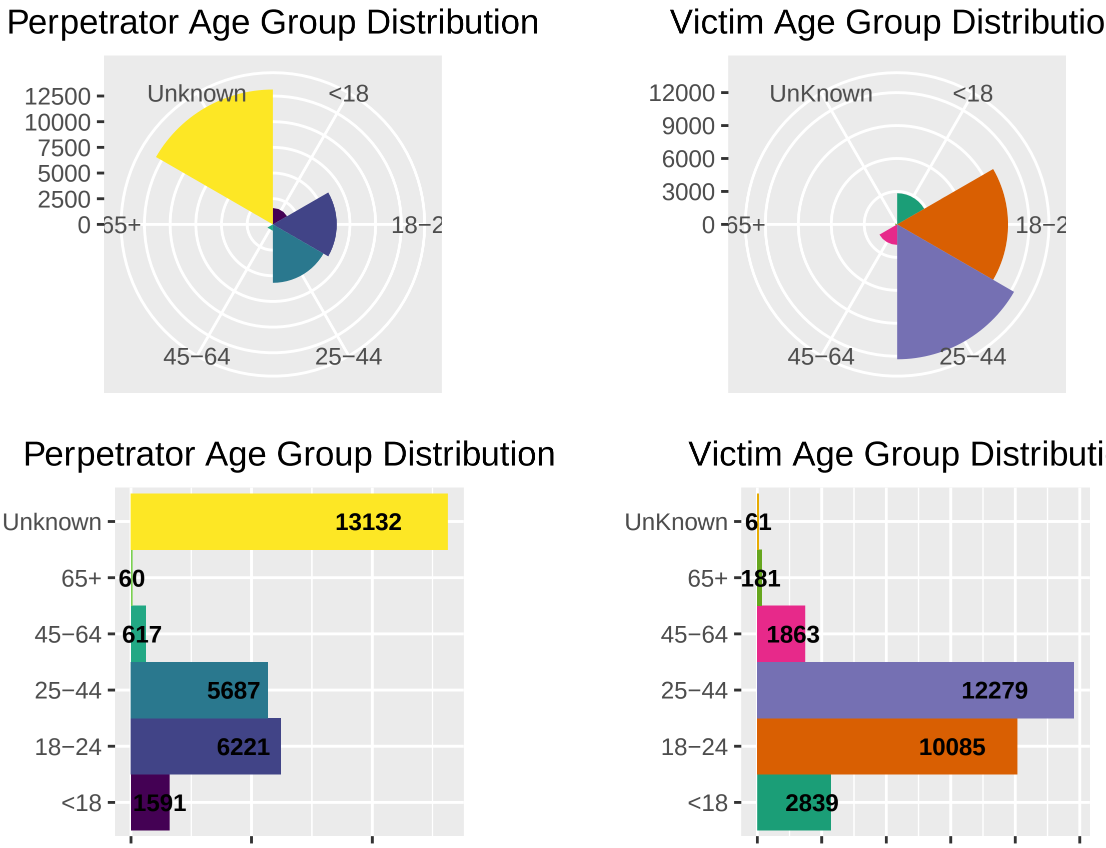
```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.
```

```
options(repr.plot.width = 25, repr.plot.height=25)

vict_b <- ggplot(data=df_ny) + geom_bar(mapping = aes(x=VIC_AGE_GROUP, fill=VIC_AGE_GROUP),show.legend

vict_b_flip <- vict_b + coord_flip()+
  theme(axis.text.x = element_blank())  + geom_text(aes(x=VIC_AGE_GROUP, label = ..count..), size=3, fo
```

```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.
```

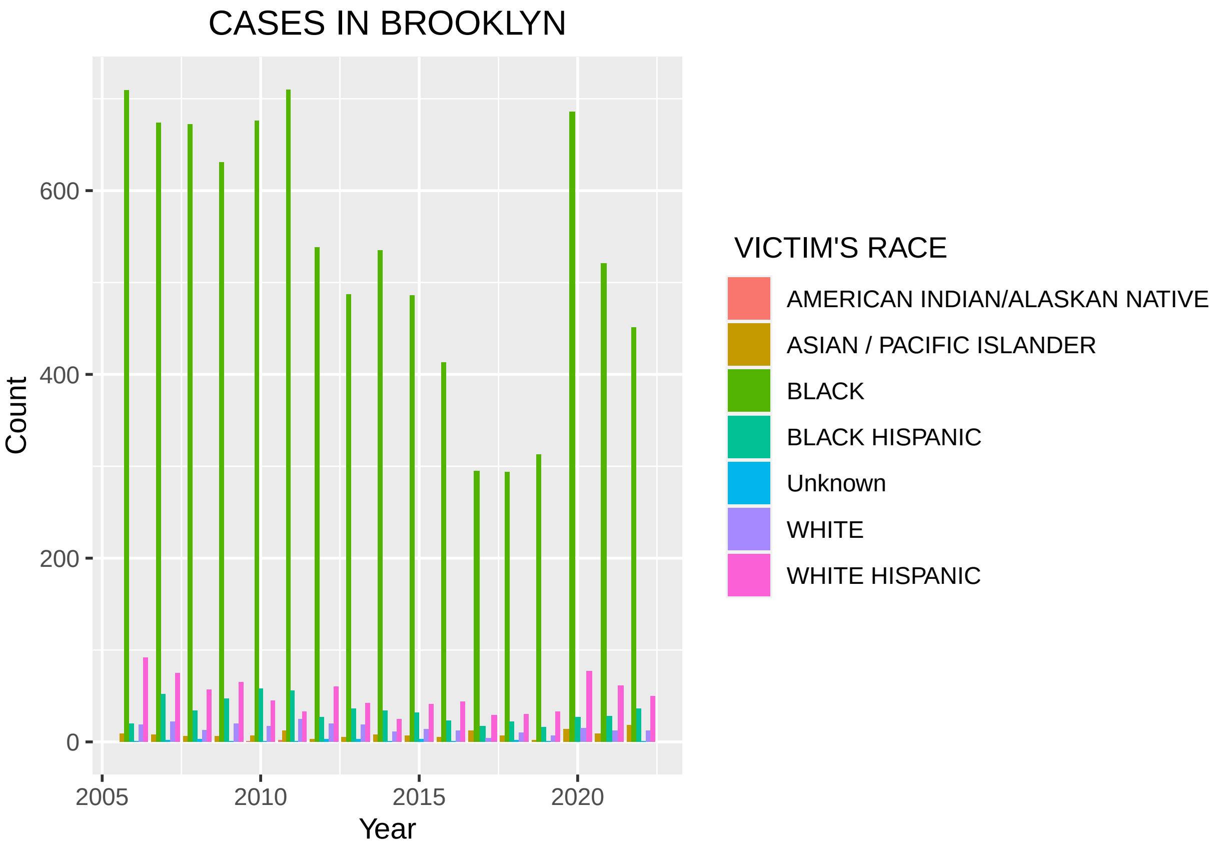
```r
grid.arrange(perp_b, vict_b, perp_flip, vict_b_flip, ncol=2)
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

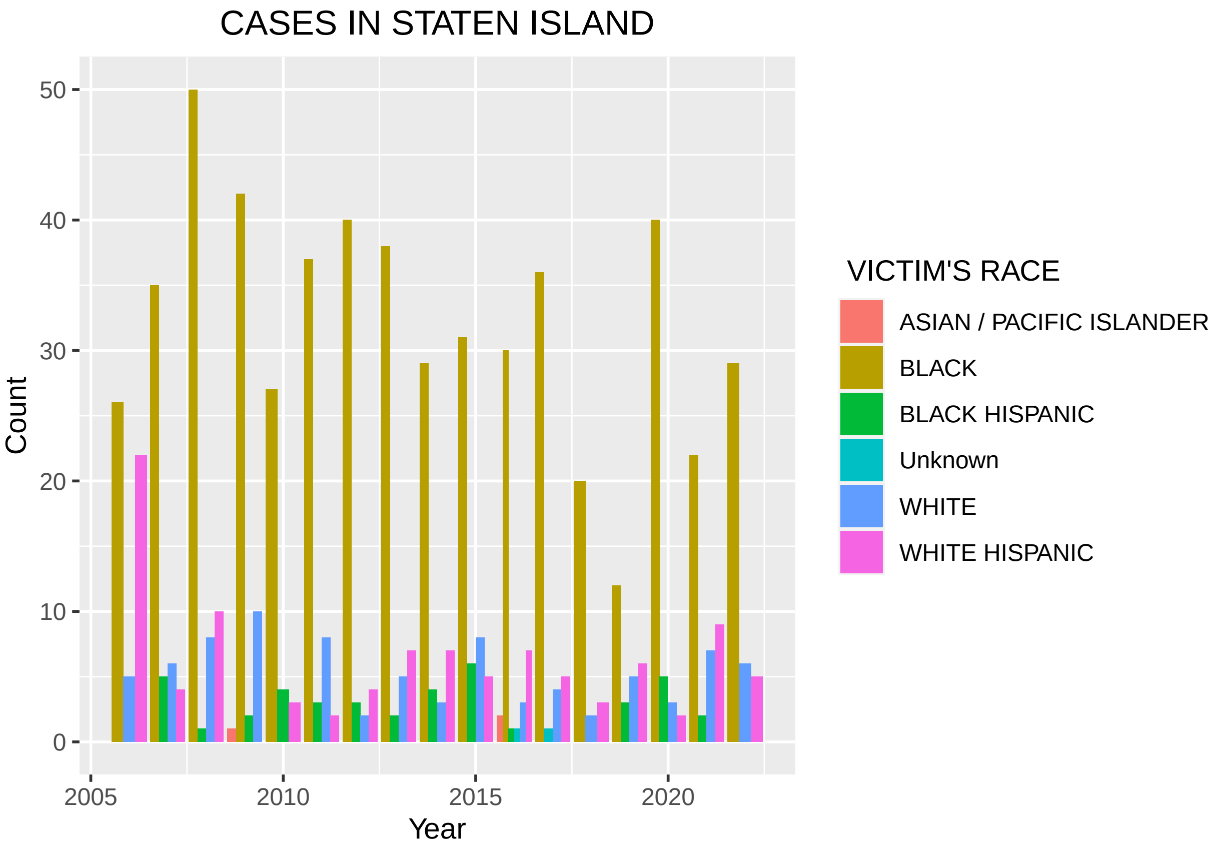


With the age group of most perpetrators unknown, significant grouped after this were between the ages of 18-24 and 25-44 with the reverse being the case of victims where most victims were between 25-44 followed by 18-24 with about 61 cases classified as unknown.

```r
df_ny %>% filter(BORO =='BROOKLYN') %>% ggplot(aes(x = year(DATE), fill = VIC_RACE)) +
  geom_bar(position = 'dodge') +
  labs(x = "Year", y = "Count", fill = " VICTIM'S RACE",title= "CASES IN BROOKLYN") + theme(plot.title
```

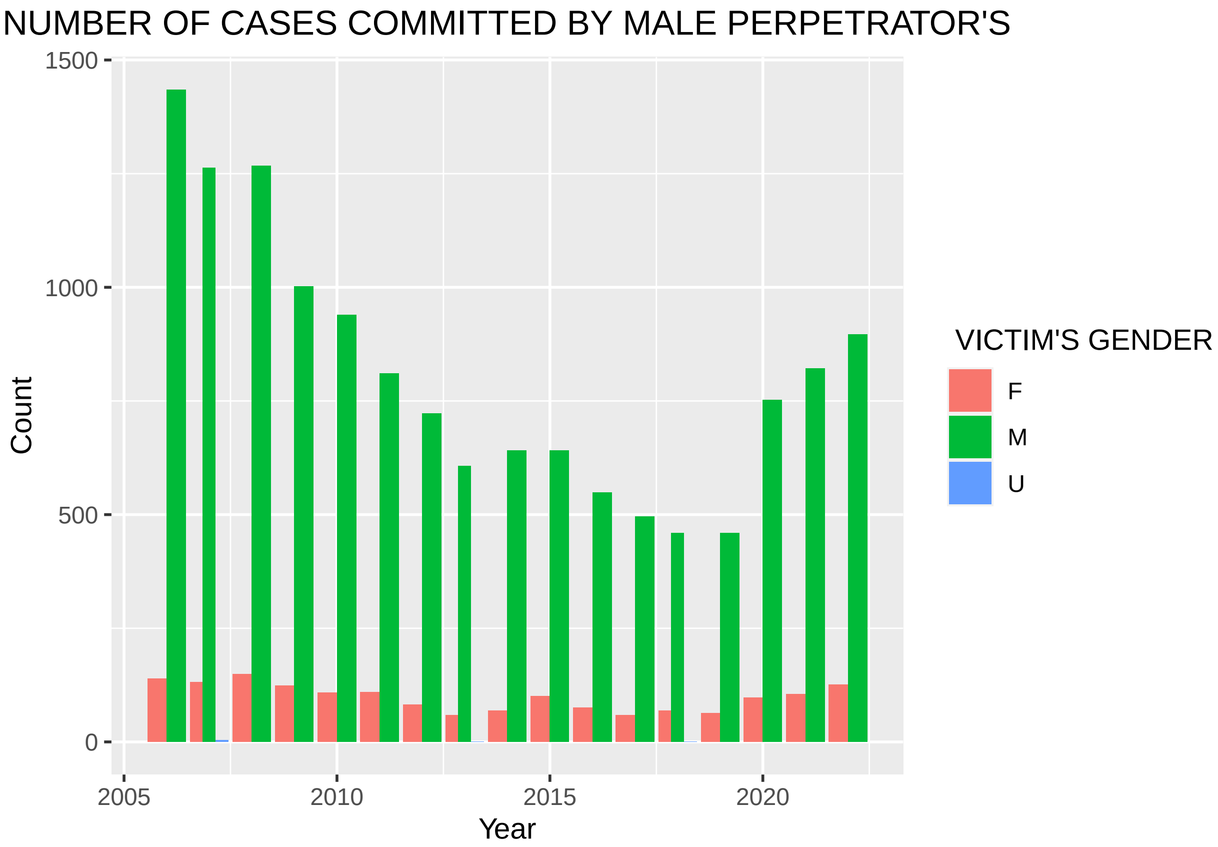


```
df_ny %>% filter(BORO =='STATEN ISLAND') %>% ggplot(aes(x = year(DATE), fill = VIC_RACE)) +
  geom_bar(position = 'dodge') +
  labs(x = "Year", y = "Count", fill = " VICTIM'S RACE",title= "CASES IN STATEN ISLAND") + theme(plot.ti
```

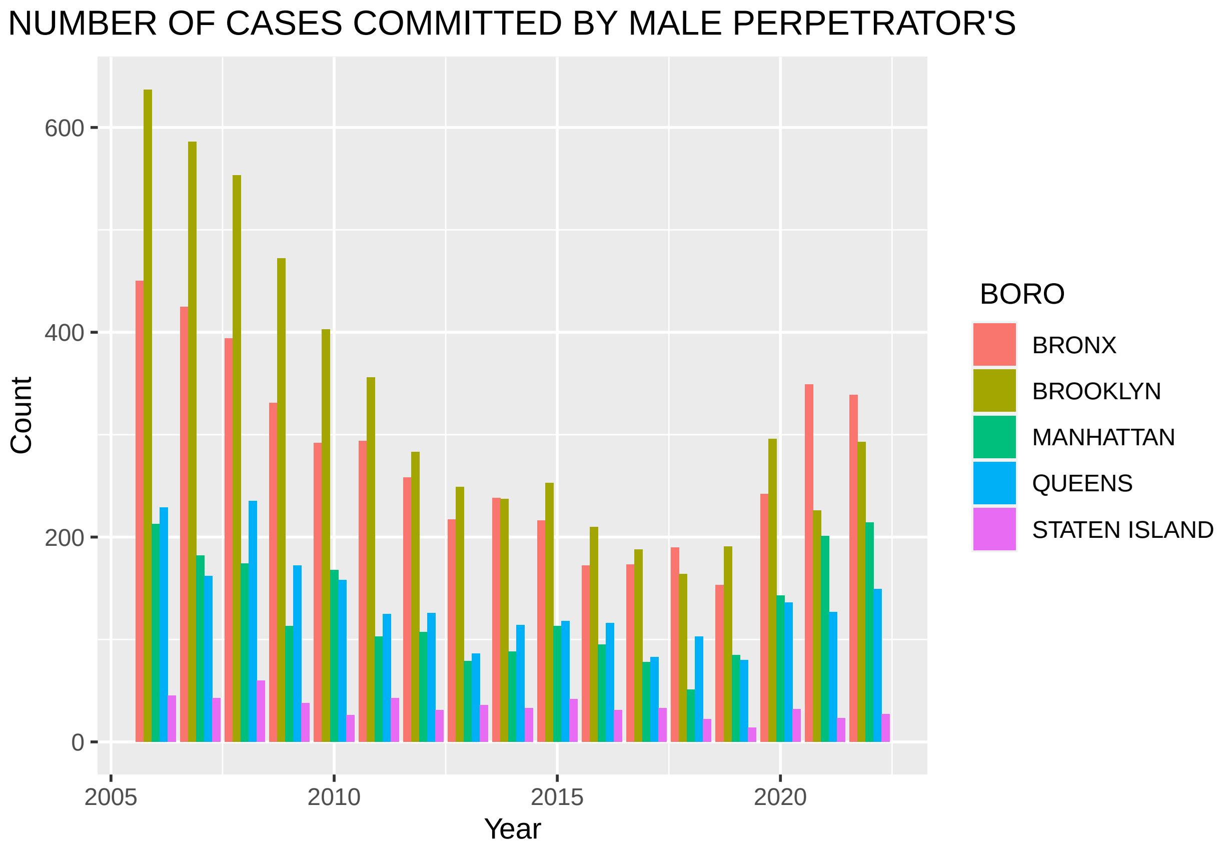


Both graphs represent the locations with the highest and lowest occurrence of cases. It is evident that the majority of victims were Black, followed by White Hispanics.

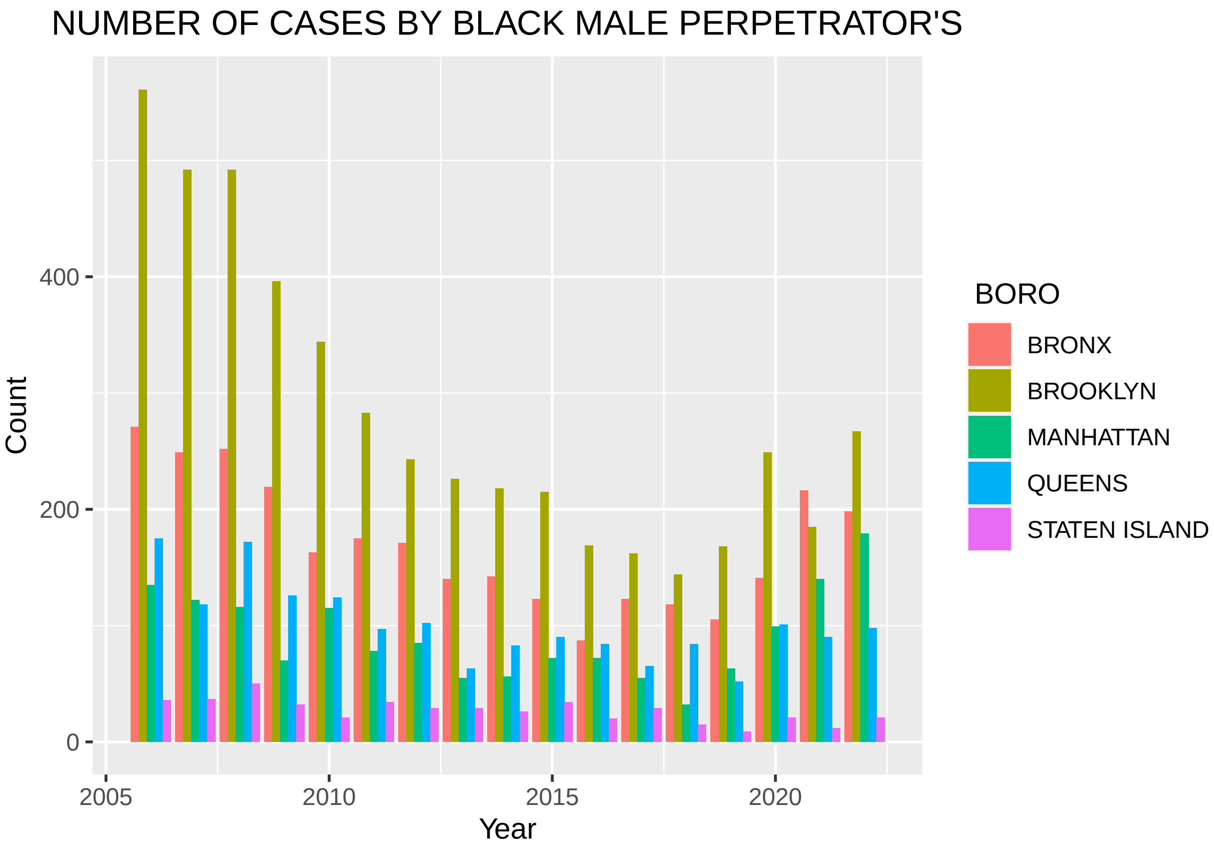
```
df_ny %>% filter(PERP_SEX =='M') %>% ggplot(aes(x = year(DATE), fill = VIC_SEX)) +
  geom_bar(position = 'dodge') +
  labs(x = "Year", y = "Count", fill = " VICTIM'S GENDER",title= "NUMBER OF CASES COMMITTED BY MALE PERP
```



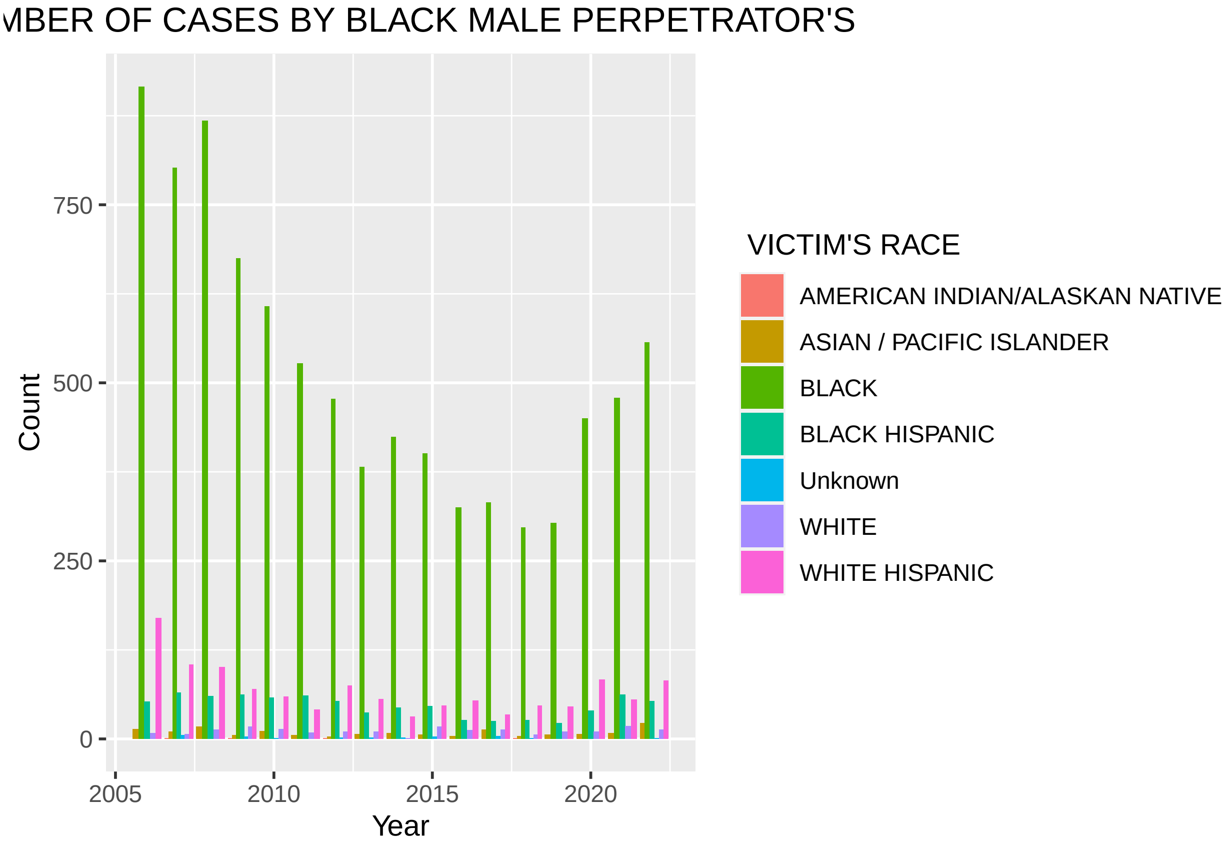
```r
df_ny %>% filter(PERP_SEX =='M') %>% ggplot(aes(x = year(DATE), fill = BORO)) +
  geom_bar(position = 'dodge') +
  labs(x = "Year", y = "Count", fill = " BORO",title= "NUMBER OF CASES COMMITTED BY MALE PERPETRATOR'S"
```



```
df_ny %>% filter(PERP_RACE=='BLACK') %>% ggplot(aes(x = year(DATE), fill = BORO)) +
  geom_bar(position = 'dodge') +
  labs(x = "Year", y = "Count", fill = " BORO",title= "NUMBER OF CASES BY BLACK MALE PERPETRATOR'S") +
```



```
df_ny %>% filter(PERP_RACE =="BLACK" & PERP_SEX =="M") %>% ggplot(aes(x = year(DATE), fill = VIC_RACE)
 geom_bar(position = 'dodge') +
 labs(x = "Year", y = "Count", fill = " VICTIM'S RACE",title= "NUMBER OF CASES BY BLACK MALE PERPETRAT
```



These plots display the distribution of cases involving male perpetrators based on gender and the neighbourhood. The distribution patterns for cases committed by Black males were similar to those for cases committed by males in general.

Conclusion

Based on the analysis of the NYPD shooting data, it is evident that most crimes were committed by individuals of Black ethnicity, with the majority of victims being Black males, followed by White Hispanics. The incidents predominantly occurred in Brooklyn and the Bronx during the early and late hours of the day. It's important to note that, any baises in relation to how the data was collected and aggregated would significantly impact this analysis as those could not at the moment be ascertained and corrected. Additionally, this dataset does not account for unlogged or unreported cases, which might impact the narrative of this report.