5301 NYPD Shooting Assignment

Tim McCracken

2024-02-19

NYPD Shooting Incident Data Analysis

This document will load in every shooting incident that occurred in NYC from 2006 to the end of the previous calendar year and then provide analysis using visuals and models.

Question: Where do the majority of shooting incidents and murders occur in NYC? What are the defining characteristics of shooting perpetrators and shooting victims in NYC?

Per the City of New York, 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."

Source - https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Load necessary libaries

```
# install.packages("tidyverse")
# install.packages("lubridate")
library(tidyverse)
library(lubridate)
```

Import Data

```
shooting_incidents = read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DO
```

Tidy and Transform Data

Here I look at the data and select which columns that I will need for this particular analysis. Then I clean up any missing/incomplete data so the data can be interpreted, modeled, and analyzed.

```
summary(shooting_incidents)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms         Class :character
##  Median : 90372218   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :120860536                      Mode  :numeric
```

```
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  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:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

```
tidied_incidents = select(shooting_incidents, OCCUR_DATE, BORO,STATISTICAL_MURDER_FLAG,
                          PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE)
summary(tidied_incidents)
```

```
##   OCCUR_DATE            BORO           STATISTICAL_MURDER_FLAG
##  Length:27312       Length:27312       Mode :logical
##  Class :character   Class :character   FALSE:22046
##  Mode  :character   Mode  :character   TRUE :5266
##  PERP_AGE_GROUP       PERP_SEX           PERP_RACE         VIC_AGE_GROUP
##  Length:27312       Length:27312       Length:27312       Length:27312
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    VIC_SEX            VIC_RACE
##  Length:27312       Length:27312
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
colSums(is.na(tidied_incidents))
```

```
##              OCCUR_DATE                    BORO STATISTICAL_MURDER_FLAG
##                       0                       0                       0
##          PERP_AGE_GROUP                PERP_SEX               PERP_RACE
##                    9344                    9310                    9310
##           VIC_AGE_GROUP                 VIC_SEX                VIC_RACE
##                       0                       0                       0
```

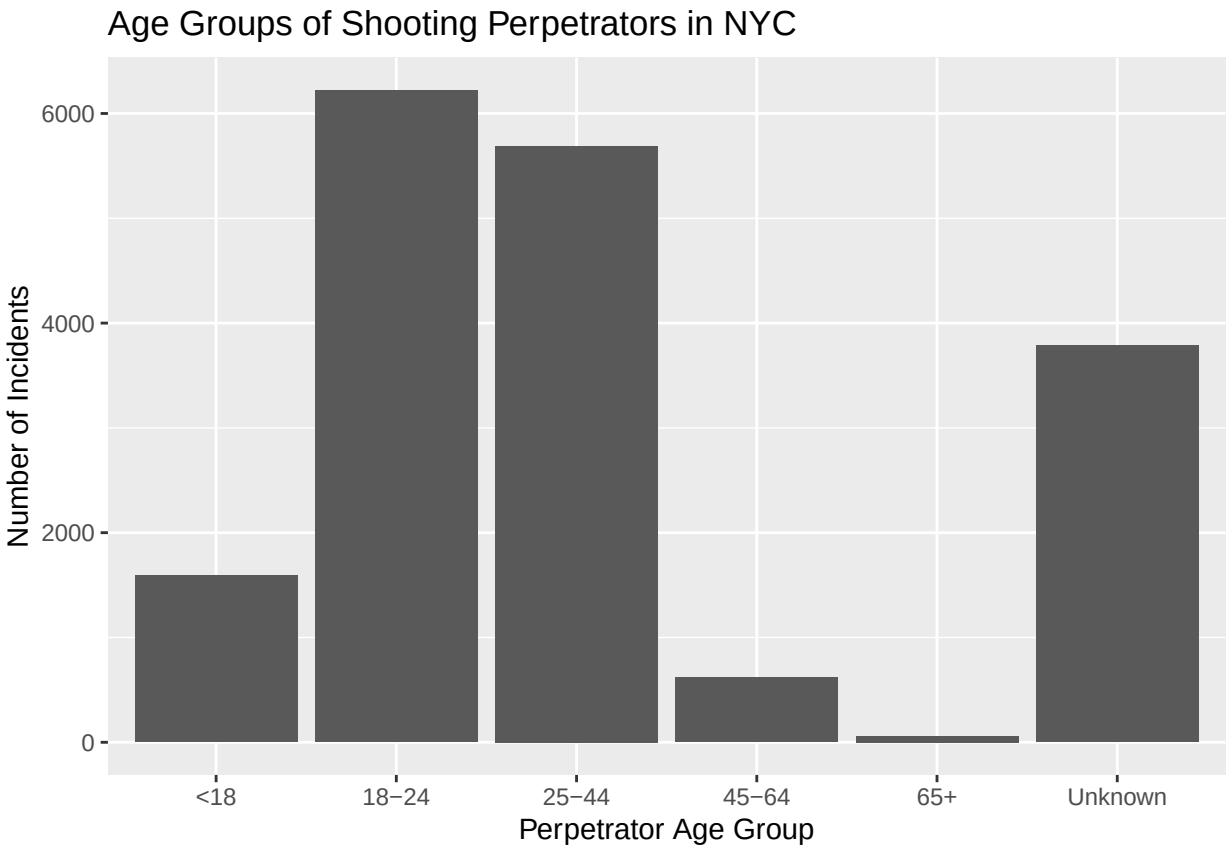
```
tidied_incidents = subset(tidied_incidents, PERP_AGE_GROUP != "224" & PERP_AGE_GROUP != "940"
                          & PERP_AGE_GROUP != "1020" & VIC_AGE_GROUP != "1022")

tidied_incidents$BORO = factor(tidied_incidents$BORO)

tidied_incidents <- tidied_incidents %>%
replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))
tidied_incidents$PERP_AGE_GROUP = recode(tidied_incidents$PERP_AGE_GROUP, "UNKNOWN" = "Unknown")
tidied_incidents$PERP_AGE_GROUP = recode(tidied_incidents$PERP_AGE_GROUP, "(null)" = "Unknown")
tidied_incidents$PERP_AGE_GROUP = factor(tidied_incidents$PERP_AGE_GROUP)

tidied_incidents$PERP_SEX = recode(tidied_incidents$PERP_SEX, "U" = "Unknown")
tidied_incidents$PERP_SEX = recode(tidied_incidents$PERP_SEX, "(null)" = "Unknown")
tidied_incidents$PERP_SEX = factor(tidied_incidents$PERP_SEX)

tidied_incidents$PERP_RACE = recode(tidied_incidents$PERP_RACE, "UNKNOWN" = "Unknown")
tidied_incidents$PERP_RACE = recode(tidied_incidents$PERP_RACE, "(null)" = "Unknown")
tidied_incidents$PERP_RACE = factor(tidied_incidents$PERP_RACE)

tidied_incidents$VIC_AGE_GROUP = recode(tidied_incidents$VIC_AGE_GROUP, "UNKNOWN" = "Unknown")
tidied_incidents$VIC_AGE_GROUP = factor(tidied_incidents$VIC_AGE_GROUP)

tidied_incidents$VIC_SEX = recode(tidied_incidents$VIC_SEX, "U" = "Unknown")
tidied_incidents$VIC_SEX= factor(tidied_incidents$VIC_SEX)

tidied_incidents$VIC_RACE = recode(tidied_incidents$VIC_RACE, "UNKNOWN" = "Unknown")
tidied_incidents$VIC_RACE = factor(tidied_incidents$VIC_RACE)


summary(tidied_incidents)
```

```
##   OCCUR_DATE                   BORO      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:17964       BRONX        :5423   Mode :logical           <18    :1591
##  Class :character   BROOKLYN     :6641   FALSE:14404             18-24  :6221
##  Mode  :character   MANHATTAN    :2541   TRUE :3560              25-44  :5687
##                     QUEENS       :2728                           45-64  : 617
##                     STATEN ISLAND: 631                           65+    :  60
```

```
##                                                                  Unknown:3788
##
##     PERP_SEX                                PERP_RACE     VIC_AGE_GROUP
##  F       :  424   AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    :2027
##  M       :15435   ASIAN / PACIFIC ISLANDER      :  154   18-24  :6517
##  Unknown: 2105   BLACK                         :11430   25-44  :7937
##                  BLACK HISPANIC                : 1314   45-64  :1290
##                  Unknown                       : 2442   65+    : 137
##                  WHITE                         :  283   Unknown:  56
##                  WHITE HISPANIC                : 2339
##     VIC_SEX                                VIC_RACE
##  F       : 1922   AMERICAN INDIAN/ALASKAN NATIVE:    8
##  M       :16034   ASIAN / PACIFIC ISLANDER      :  307
##  Unknown:    8   BLACK                         :12250
##                  BLACK HISPANIC                : 1800
##                  Unknown                       :   48
##                  WHITE                         :  552
##                  WHITE HISPANIC                : 2999
```

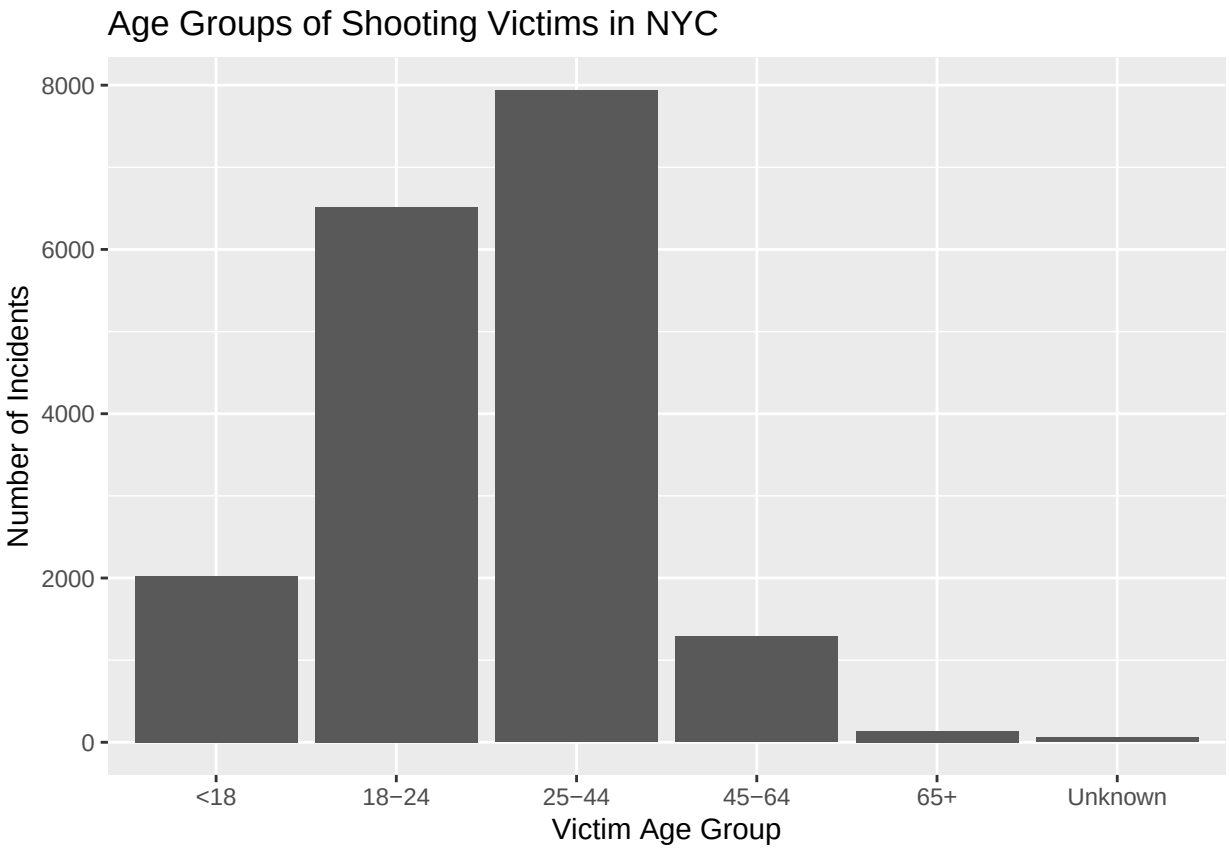
Visualize and Analyze the Data

```
ggplot(tidied_incidents, aes(x = PERP_AGE_GROUP)) +
  geom_bar() +
  labs(x = "Perpetrator Age Group", y = "Number of Incidents",
       title = "Age Groups of Shooting Perpetrators in NYC")
```



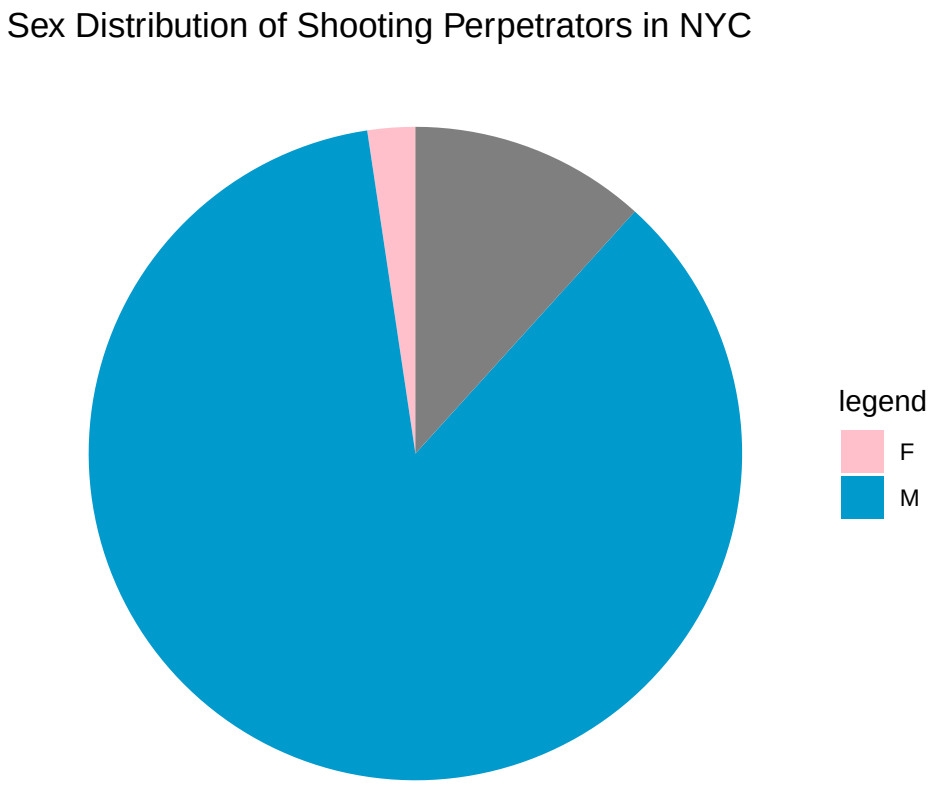
This graph shows the distribution of shooting perpetrators in NYC. It is clear that two age groups dominate the amount of shooting incidents in NYC, 18-24 and 25-44. There is also a substantial amount of incidents from the <18 category which is surprising to me. However, there is almost 4000 incidents where the age of the perpetrator is unknown. My assumption would be that in many shooting incidents it may be difficult to identify how old the shooter is.

```
ggplot(tidied_incidents, aes(x = VIC_AGE_GROUP)) +
  geom_bar() +
  labs(x = "Victim Age Group", y = "Number of Incidents",
       title = "Age Groups of Shooting Victims in NYC")
```



The victims of shooting incidents in NYC have a similar distribution to the perpetrators except the unknown amount reduces almost to zero. This is probably due to the fact that in shooting cases the victim can confirm their age while the perpetrator may not be caught and may not be able to be identified.

```
ggplot(tidied_incidents,aes(x="",fill=PERP_SEX)) +
  geom_bar() +
  labs(x = "Perpetrator Sex", y = "Number of Incidents",
       title = "Sex Distribution of Shooting Perpetrators in NYC") +
  scale_fill_manual("legend",values=c("M" = "deepskyblue3", F = "pink")) +
  coord_polar("y", start=0) +
  theme_void()
```




```
round(sum(tidied_incidents$PERP_SEX == "M")/length(tidied_incidents$PERP_SEX)*100,1)
```

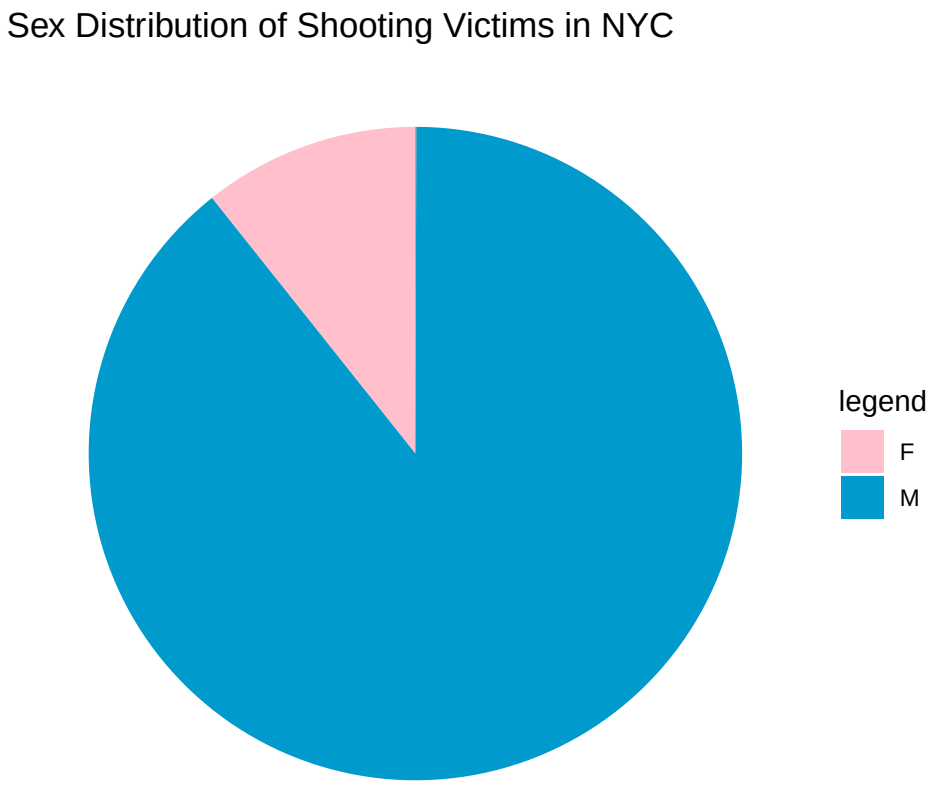
```
## [1] 85.9
```

```
round(sum(tidied_incidents$PERP_SEX == "F")/length(tidied_incidents$PERP_SEX)*100,1)
```

```
## [1] 2.4
```

This is the sex distribution of perpetrators of shooting incidents in NYC. Males dominate this category and are responsible for almost 86% of reported shooting incidents in NYC. Females are only responsible for 2.4% of shooting incidents. Similar to the age group distribution, there is quite a substantial amount of incidents where the sex of the perpetrator is unknown (11.7%),

```
ggplot(tidied_incidents,aes(x="",fill=VIC_SEX)) +
  geom_bar() +
  labs(x = "Victim Sex", y = "Number of Incidents",
       title = "Sex Distribution of Shooting Victims in NYC") +
  scale_fill_manual("legend",values=c("M" = "deepskyblue3", F = "pink")) +
  coord_polar("y", start=0) +
  theme_void()
```




```
round(sum(tidied_incidents$VIC_SEX == "M")/length(tidied_incidents$VIC_SEX)*100,1)
```

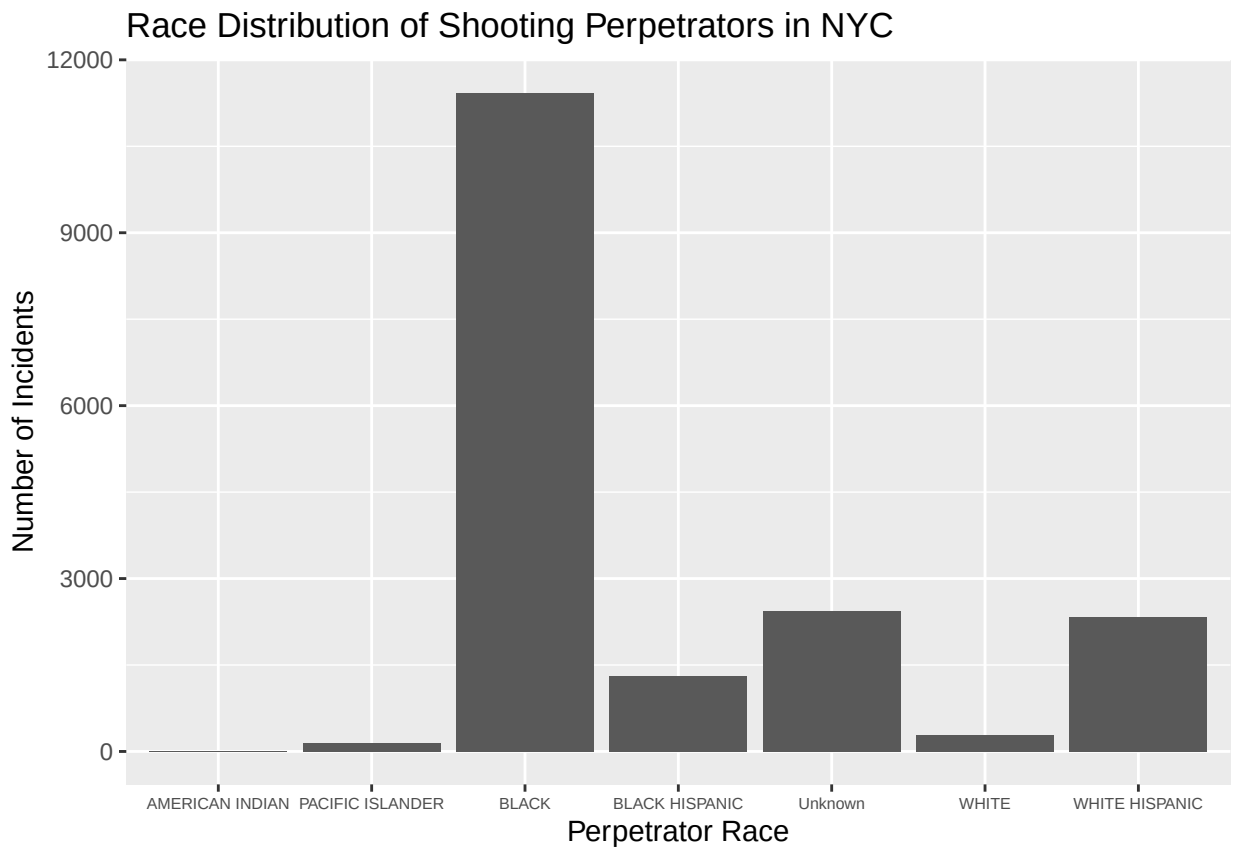
```
## [1] 89.3
```

```
round(sum(tidied_incidents$VIC_SEX == "F")/length(tidied_incidents$VIC_SEX)*100,1)
```

```
## [1] 10.7
```

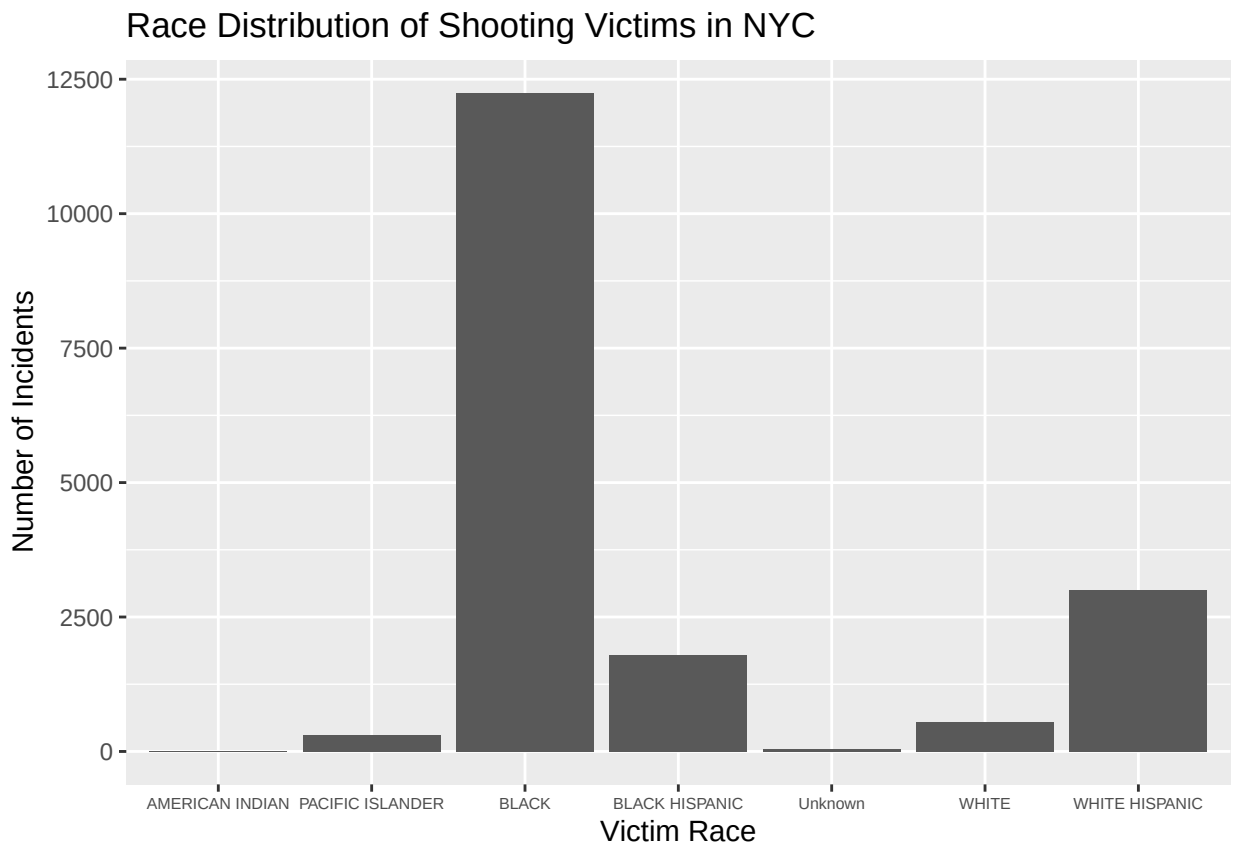
Again, males dominate the shooting victim category at 89.3% of all incidents. Females are victims in 10.7% of shooting incidents. And similar to the age group distribution, the amount of unknown incidents reduces to almost zero for victims.

```
tidied_incidents$PERP_RACE = recode(tidied_incidents$PERP_RACE,
                                    "AMERICAN INDIAN/ALASKAN NATIVE" = "AMERICAN INDIAN")
tidied_incidents$PERP_RACE = recode(tidied_incidents$PERP_RACE,
                                    "ASIAN / PACIFIC ISLANDER" = "PACIFIC ISLANDER")
ggplot(tidied_incidents,aes(x=PERP_RACE)) +
  geom_bar() +
  labs(x = "Perpetrator Race", y = "Number of Incidents",
       title = "Race Distribution of Shooting Perpetrators in NYC") +
  theme(axis.text.x = element_text(size=6))
```



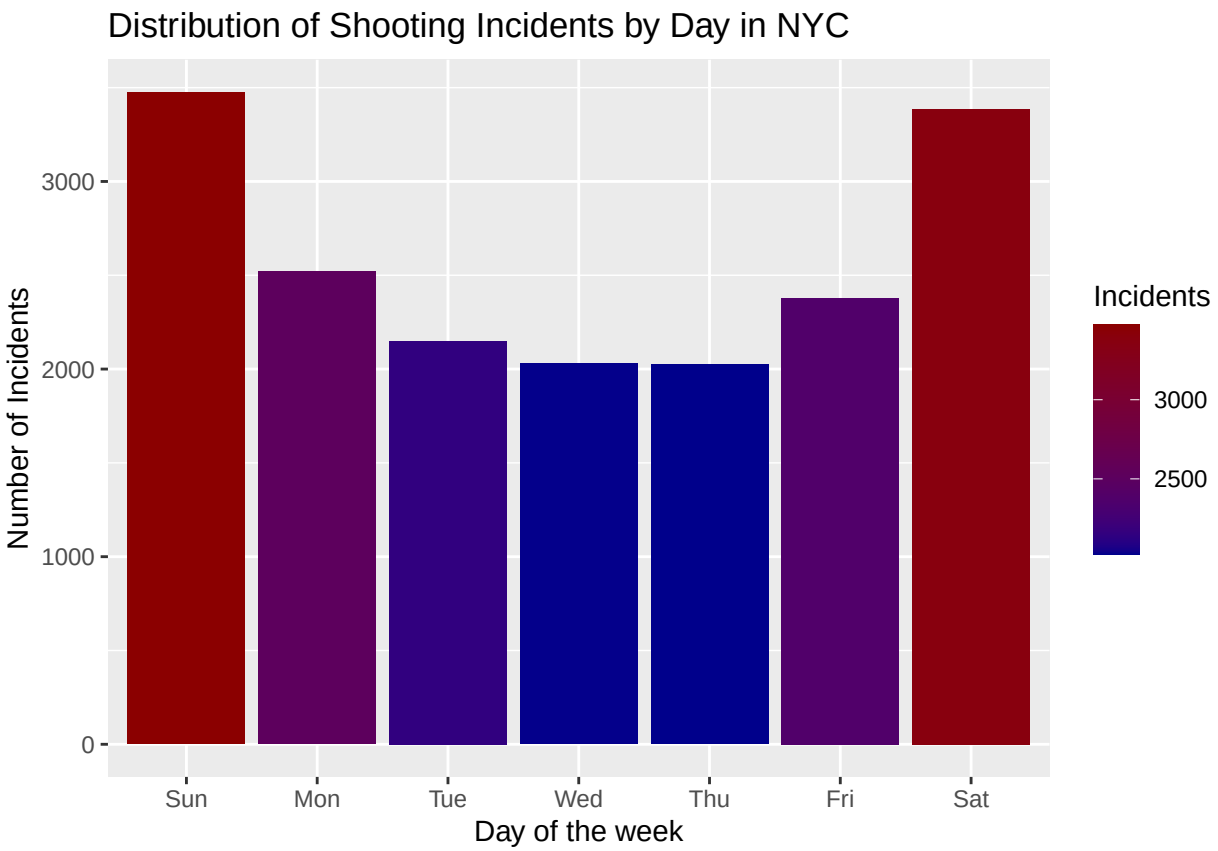
The race distribution for shooting perpetrators in NYC are largely reported to be Black with over 11,000 incidents. The next most reported shooting perpetrators are White Hispanic. Again, due to the nature of shooting cases, it may be difficult to identify characteristics of shooting perpetrators, hence the large amount cases where the race of the perpetrator is unknown.

```
tidied_incidents$VIC_RACE = recode(tidied_incidents$VIC_RACE,
                                   "AMERICAN INDIAN/ALASKAN NATIVE" = "AMERICAN INDIAN")
tidied_incidents$VIC_RACE = recode(tidied_incidents$VIC_RACE,
                                   "ASIAN / PACIFIC ISLANDER" = "PACIFIC ISLANDER")
ggplot(tidied_incidents,aes(x=VIC_RACE)) +
  geom_bar() +
  labs(x = "Victim Race", y = "Number of Incidents",
       title = "Race Distribution of Shooting Victims in NYC") +
  theme(axis.text.x = element_text(size=6))
```



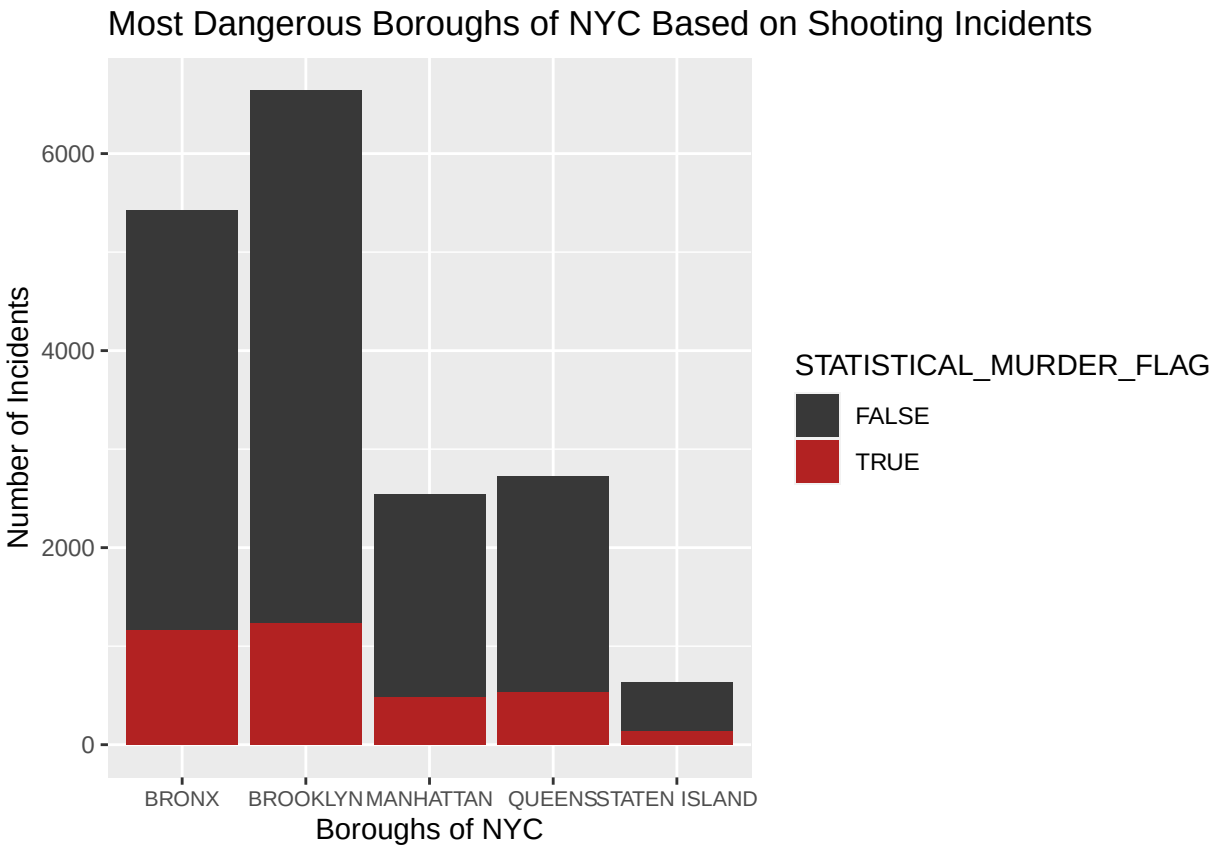
The distribution of shooting victims tell a similar story, where the most common race of shooting perpetrators is Black with next being White Hispanic. Again, we see the unknown category fall to almost zero.

```
tidied_incidents$OCCUR_DATE = mdy(tidied_incidents$OCCUR_DATE)
tidied_incidents$OCCUR_DATE = wday(tidied_incidents$OCCUR_DATE,label=TRUE)
incidents_by_day <- tidied_incidents %>%
  group_by(OCCUR_DATE) %>%
  count(OCCUR_DATE)
colnames(incidents_by_day)[2] <- "Incidents"
ggplot(incidents_by_day,aes(x=OCCUR_DATE,y=Incidents)) +
  geom_col(aes(fill=Incidents)) +
  labs(x = "Day of the week", y = "Number of Incidents",
       title = "Distribution of Shooting Incidents by Day in NYC") +
  scale_fill_gradient(low = "darkblue",high = "darkred")
```



This shows the distribution of shooting incidents by day of the week. Not surprisingly, the middle of the week (Tues-Thurs) has the lowest amount of incidents while the weekend (Fri-Sun). This is probably due to the fact that more people are out on the weekend vs the weekday.

```
ggplot(tidied_incidents, aes(x=BORO,fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar() +
  labs(x = "Boroughs of NYC", y = "Number of Incidents",
       title = "Most Dangerous Boroughs of NYC Based on Shooting Incidents") +
  scale_fill_manual(values = c("grey22","firebrick")) +
  theme(axis.text.x = element_text(size=8))
```



This graph shows the distribution of shooting incidents in each borough of NYC. Brooklyn has the highest number of incidents while Staten Island has the lowest number. Even though there is a substantial difference in the amount of incidents in Brooklyn and the Bronx (almost 2000) the number of deaths as a result of shooting incidents is almost the same.

Model the Data

```
mod <- glm(family=binomial,STATISTICAL_MURDER_FLAG ~ BORO + PERP_AGE_GROUP + PERP_RACE + PERP_SEX, data
summary(mod)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + PERP_AGE_GROUP +
##     PERP_RACE + PERP_SEX, family = binomial, data = tidied_incidents)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -11.75841  139.27168  -0.084 0.932716
## BOROBROOKLYN               -0.12920    0.04813  -2.685 0.007261 **
## BOROMANHATTAN              -0.17982    0.06191  -2.904 0.003679 **
## BOROQUEENS                 -0.13239    0.06079  -2.178 0.029424 *
## BOROSTATEN ISLAND          -0.20063    0.10578  -1.897 0.057881 .
## PERP_AGE_GROUP18-24         0.18353    0.07246   2.533 0.011315 *
## PERP_AGE_GROUP25-44         0.49913    0.07201   6.931 4.18e-12 ***
```

```
## PERP_AGE_GROUP45-64         0.85588    0.10757   7.957 1.77e-15 ***
## PERP_AGE_GROUP65+           1.04197    0.27536   3.784 0.000154 ***
## PERP_AGE_GROUPUnknown      -2.32103    0.17526 -13.243  < 2e-16 ***
## PERP_RACEPACIFIC ISLANDER  10.95591  139.27177   0.079 0.937299
## PERP_RACEBLACK             10.48848  139.27166   0.075 0.939968
## PERP_RACEBLACK HISPANIC    10.35788  139.27167   0.074 0.940715
## PERP_RACEUnknown            9.88948  139.27182   0.071 0.943391
## PERP_RACEWHITE             11.10700  139.27171   0.080 0.936436
## PERP_RACEWHITE HISPANIC    10.58693  139.27166   0.076 0.939406
## PERP_SEXM                  -0.16548    0.11380  -1.454 0.145928
## PERP_SEXUnknown             1.87100    0.27576   6.785 1.16e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17887  on 17963  degrees of freedom
## Residual deviance: 16811  on 17946  degrees of freedom
## AIC: 16847
##
## Number of Fisher Scoring iterations: 10
```

I used a generalized linear model with a binomial family to model whether or not certain variables are good predictors of whether or not a shooting incident is a murder. The coefficient estimates represent the log odds ratio of each incident being a murder compared to the reference category. This model chooses reference categories based on the levels of categorical variables that the algorithm deems less relevant or less influential to the outcome. The resulting z and p values indicate the statistical significance of the coefficients. A large absolute z-value indicate the coefficient is statistically significant and a small p-value also indicates that the coefficient is highly significant.

For example, being in Queens decreases the log odds of murder by 0.13239 compared to the reference category of Manhattan, holding all else constant. If the perpetrator age is between 25-44, this increases the log odds of murder by 0.49913 compared to the reference category age group of <18.

Discuss Bias

Due to the nature of this report and selecting RACE and SEX as part of my data, this opens up the report to possible bias. For example, there may be sampling bias based on how this data is collected through the NYC police system. Police in NYC may under represent or over represent certain groups, demographics, or boroughs in NYC. The model itself may have bias in it's output due to confounding variables that it does not account for, or if the chosen predictors do not accurately capture all relevant affecting factors. There could also be bias on the reports of the age group, sex, and race of shooting perpetrators from shooting victims based on a number of factors including racial bias and geographical bias within NYC itself.

Conclusion

To answer the original questions, the majority of shooting incidents occur in the Bronx and Brooklyn. The most common shooting perpetrator profile is black, male, and 18-44 years-old. Shooting victims have a similar profile. This report provided unique insight on the characteristics of shooting perpetrators and victims in NYC, as well as the relationship between location, day of the week, and borough to shooting incidents.