NYPD Data

Packages Needed

- Tidyverse
- Lubridate

```
library(tidyverse)
library(lubridate)
```

Importing the data

First, I'll import the data from https://catalog.data.gov/dataset. This data represents information about every shooting incident in New York City since 2006.

```
url_nypd <- paste0("https://data.cityofnewyork.us/api/views/833y-fsy8/",
              "rows.csv?accessType=DOWNLOAD")

nypd_shootings <- read_csv(url_nypd)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Tidying Data

Looking at the column details, I can see some columns are not the correct variable types. Therefore, I will make the following changes

- *Occur_Date* is listed as a string/character type
  - This needs to change to a date column using the **lubridate** package
- The following variables will need to be changed to a factor type because they are categorical
  - *BORO*
  - *JURISDICTION_CODE*
  - *PERP_AGE_GROUP*
  - *PERP_SEX*
  - *PERP_RACE*
  - *VIC_AGE_GROUP*
  - *VIC_SEX*
  - *VIC_RACE*
- I'm also removing a few variables that I don't feel have as much impact to the analysis. INCIDENT_KEY would be important if we were joining multiple datasets. In this case, we aren't; therefore, I am removing it along with the geographical data. **LOCATION_DESC** can be very useful; however, at first glance it seems as if there is a lot of missing data. First we'll take a look at the missing amount.

```
sum(is.na(nypd_shootings$LOCATION_DESC)) / nrow(nypd_shootings)
```

```
## [1] 0.5762475
```

Because over half of the data is missing, we will remove **LOCATION_DESC** as well.

```
factor_cols <- c("BORO", "JURISDICTION_CODE", "PERP_AGE_GROUP", "PERP_SEX",
                 "PERP_RACE", "VIC_AGE_GROUP", "VIC_SEX", "VIC_RACE")

nypd_shootings <- nypd_shootings %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE),
                                           across(.cols = all_of(factor_cols),
                                                  as.factor)) %>%
  select(-c(INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat,
            LOCATION_DESC))
```

Viewing the summary, we can see that about of a third of the PERP_AGE_GROUP, PERP_SEX, AND PER_RACE are missing. Thus, I will drop all rows that are missing data in these columns.If we had access to more data, I could probably fill the missing data using various methods. Also, JURISDICTION_CODE only has two observations where the data is missing, I will fill them with a random number between 0 and 2.

```
summary(nypd_shootings)
```

```
##    OCCUR_DATE          OCCUR_TIME                    BORO         PRECINCT
##  Min.   :2006-01-01   Length:23568      BRONX        :6700   Min.   :  1.00
##  1st Qu.:2008-12-30   Class1:hms        BROOKLYN     :9722   1st Qu.: 44.00
##  Median :2012-02-26   Class2:difftime   MANHATTAN    :2921   Median : 69.00
##  Mean   :2012-10-03   Mode  :numeric    QUEENS       :3527   Mean   : 66.21
##  3rd Qu.:2016-02-28                     STATEN ISLAND: 698   3rd Qu.: 81.00
##  Max.   :2020-12-31                                          Max.   :123.00
##
##  JURISDICTION_CODE STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  0   :19624        Mode :logical           18-24  :5448   F   :  334
##  1   :   54        FALSE:19080             25-44  :4613   M   :13305
##  2   : 3888        TRUE :4488              UNKNOWN:3156   U   : 1504
##  NA's:    2                                <18    :1354   NA's: 8425
##                                            45-64  : 481
##                                            (Other):  57
##                                            NA's   :8459
##          PERP_RACE    VIC_AGE_GROUP    VIC_SEX
##  BLACK         :9855   <18    : 2525   F: 2195
##  WHITE HISPANIC:1961   18-24  : 9000   M:21353
##  UNKNOWN       :1869   25-44  :10287   U:   20
##  BLACK HISPANIC:1081   45-64  : 1536
##  WHITE         : 255   65+    :  155
##  (Other)       : 122   UNKNOWN:   65
##  NA's          :8425
##                               VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  320
##  BLACK                         :16846
##  BLACK HISPANIC                : 2244
##  UNKNOWN                       :  102
##  WHITE                         :  615
##  WHITE HISPANIC                : 3432
```

```
nypd_shootings <- nypd_shootings %>%
  mutate(JURISDICTION_CODE = replace(JURISDICTION_CODE, is.na(JURISDICTION_CODE)
                                     , sample(0:2, 1))) %>%
  drop_na(PERP_AGE_GROUP, PERP_SEX, PERP_RACE)
sprintf("The number of missing values is: %i", sum(is.na(nypd_shootings)))
```

```
## [1] "The number of missing values is: 0"
```

```
summary(nypd_shootings)
```

```
##    OCCUR_DATE          OCCUR_TIME                  BORO         PRECINCT
##  Min.   :2006-01-01   Length:15109      BRONX        :4497   Min.   :  1.00
##  1st Qu.:2008-04-02   Class1:hms        BROOKLYN     :5744   1st Qu.: 44.00
##  Median :2010-07-10   Class2:difftime   MANHATTAN    :1994   Median : 69.00
##  Mean   :2011-09-26   Mode  :numeric    QUEENS       :2308   Mean   : 65.93
##  3rd Qu.:2015-01-04                     STATEN ISLAND: 566   3rd Qu.: 81.00
##  Max.   :2020-12-29                                          Max.   :123.00
##
##  JURISDICTION_CODE STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  0:12680           Mode :logical           18-24  :5448   F:  334
##  1:   43           FALSE:12233             25-44  :4613   M:13305
##  2: 2386           TRUE :2876              UNKNOWN:3156   U: 1470
##                                            <18    :1354
##                                            45-64  : 481
##                                            65+    :  54
##                                            (Other):   3
##                            PERP_RACE    VIC_AGE_GROUP  VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:   2   <18    :1788   F: 1576
##  ASIAN / PACIFIC ISLANDER      : 120   18-24  :5714   M:13521
##  BLACK                         :9855   25-44  :6400   U:   12
##  BLACK HISPANIC                :1081   45-64  :1033
##  UNKNOWN                       :1835   65+    : 117
##  WHITE                         : 255   UNKNOWN:  57
##  WHITE HISPANIC                :1961
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    7
##  ASIAN / PACIFIC ISLANDER      :  235
##  BLACK                         :10325
##  BLACK HISPANIC                : 1490
##  UNKNOWN                       :   68
##  WHITE                         :  477
##  WHITE HISPANIC                : 2507
```

From the summarize table, we can see that there are three 'Other' variables. As we can see below, these age groups seem as if they're typos. Therefore, we will change the values to unknown.

```
nypd_shootings %>% filter(nypd_shootings$PERP_AGE_GROUP != "18-24" &
                          nypd_shootings$PERP_AGE_GROUP != "25-44" &
                          nypd_shootings$PERP_AGE_GROUP != "UNKNOWN" &
                          nypd_shootings$PERP_AGE_GROUP != "<18" &
                          nypd_shootings$PERP_AGE_GROUP != "45-64" &
                          nypd_shootings$PERP_AGE_GROUP != "65+"
)
```

```
## # A tibble: 3 x 12
##   OCCUR_DATE OCCUR_TIME BORO     PRECINCT JURISDICTION_CO~ STATISTICAL_MURDER_F~
##   <date>     <time>     <fct>       <dbl> <fct>            <lgl>
## 1 2015-04-19 02:05      BRONX          47 2                FALSE
## 2 2013-03-12 20:28      BROOKLYN       90 0                FALSE
## 3 2010-03-06 04:14      BRONX          41 0                FALSE
## # ... with 6 more variables: PERP_AGE_GROUP <fct>, PERP_SEX <fct>,
## #   PERP_RACE <fct>, VIC_AGE_GROUP <fct>, VIC_SEX <fct>, VIC_RACE <fct>
```

```
nypd_shootings["PERP_AGE_GROUP"][nypd_shootings["PERP_AGE_GROUP"] == "1020" |
                              nypd_shootings["PERP_AGE_GROUP"] == "940" |
                              nypd_shootings["PERP_AGE_GROUP"] == "224"] <-
  "UNKNOWN"
```

Visualization and Analyzation

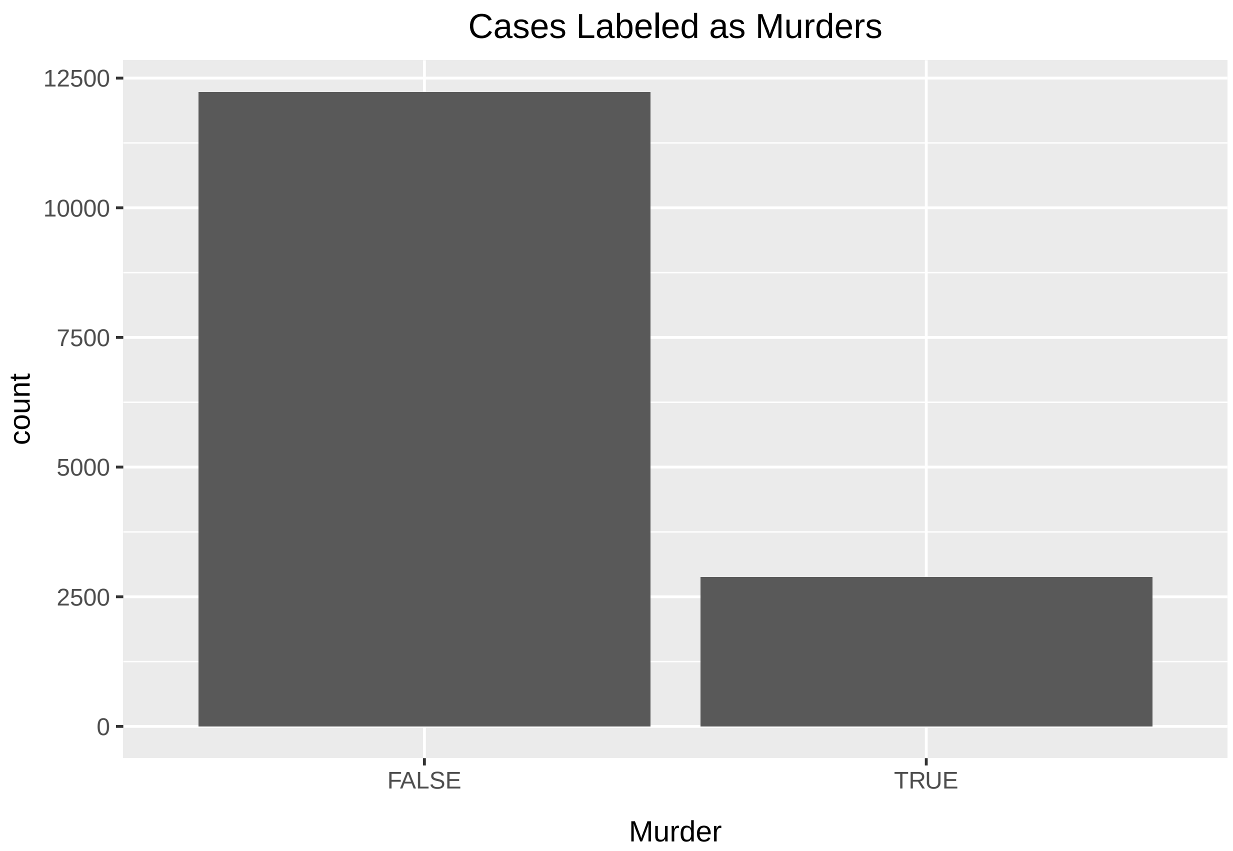
First, lets view the summary of the data

```
summary(nypd_shootings)
```

```
##    OCCUR_DATE           OCCUR_TIME                    BORO          PRECINCT
##  Min.   :2006-01-01   Length:15109      BRONX        :4497   Min.   :  1.00
##  1st Qu.:2008-04-02   Class1:hms        BROOKLYN     :5744   1st Qu.: 44.00
##  Median :2010-07-10   Class2:difftime   MANHATTAN    :1994   Median : 69.00
##  Mean   :2011-09-26   Mode  :numeric    QUEENS       :2308   Mean   : 65.93
##  3rd Qu.:2015-01-04                     STATEN ISLAND: 566   3rd Qu.: 81.00
##  Max.   :2020-12-29                                          Max.   :123.00
##
##  JURISDICTION_CODE STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  0:12680           Mode :logical           18-24  :5448   F:  334
##  1:   43           FALSE:12233             25-44  :4613   M:13305
##  2: 2386           TRUE :2876              UNKNOWN:3159   U: 1470
##                                            <18    :1354
##                                            45-64  : 481
##                                            65+    :  54
##                                            (Other):   0
##                              PERP_RACE    VIC_AGE_GROUP  VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:   2   <18    :1788   F: 1576
##  ASIAN / PACIFIC ISLANDER      : 120   18-24  :5714   M:13521
##  BLACK                         :9855   25-44  :6400   U:   12
##  BLACK HISPANIC                :1081   45-64  :1033
##  UNKNOWN                       :1835   65+    : 117
##  WHITE                         : 255   UNKNOWN:  57
##  WHITE HISPANIC                :1961
##                               VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    7
##  ASIAN / PACIFIC ISLANDER      :  235
##  BLACK                         :10325
##  BLACK HISPANIC                : 1490
##  UNKNOWN                       :   68
##  WHITE                         :  477
##  WHITE HISPANIC                : 2507
```

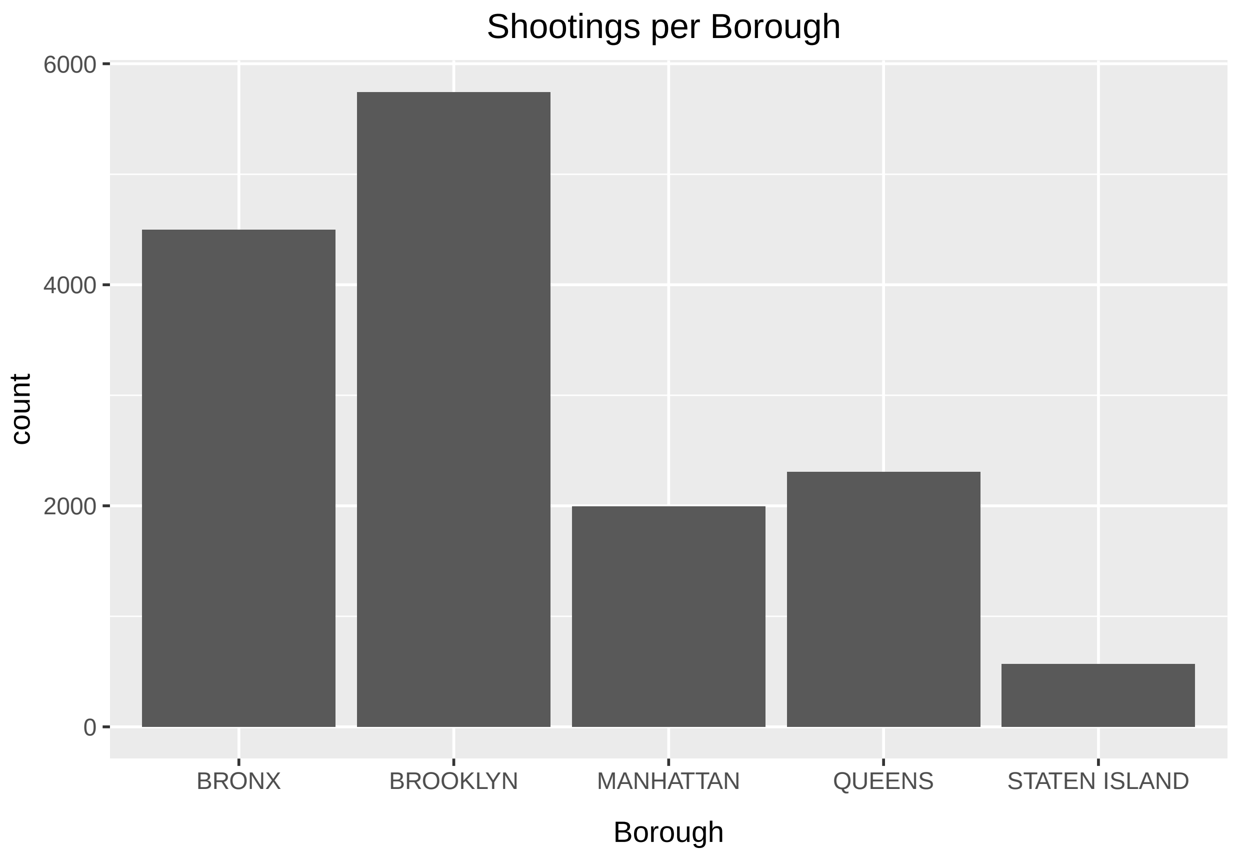
First, we can see that about 19% of the all of the shootings were labeled as murders.

```
ggplot(nypd_shootings, aes(x=STATISTICAL_MURDER_FLAG)) +
  geom_bar() +
  labs(title = "Cases Labeled as Murders", x = "Murder") +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```



Maybe the amount of shooting incidents differ between different boroughs? It seems as if there are more shootings between the Bronx and Brooklyn boroughs compared to others. However, it seems the percentage of these shootings that are labeled as murders is consistent across all boroughs.

```
#Done
ggplot(nypd_shootings, aes(x=BORO)) +
  geom_bar() +
  labs(title = "Shootings per Borough ", x = "Borough") +
  theme(axis.title.x = element_text(margin =
                                    margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

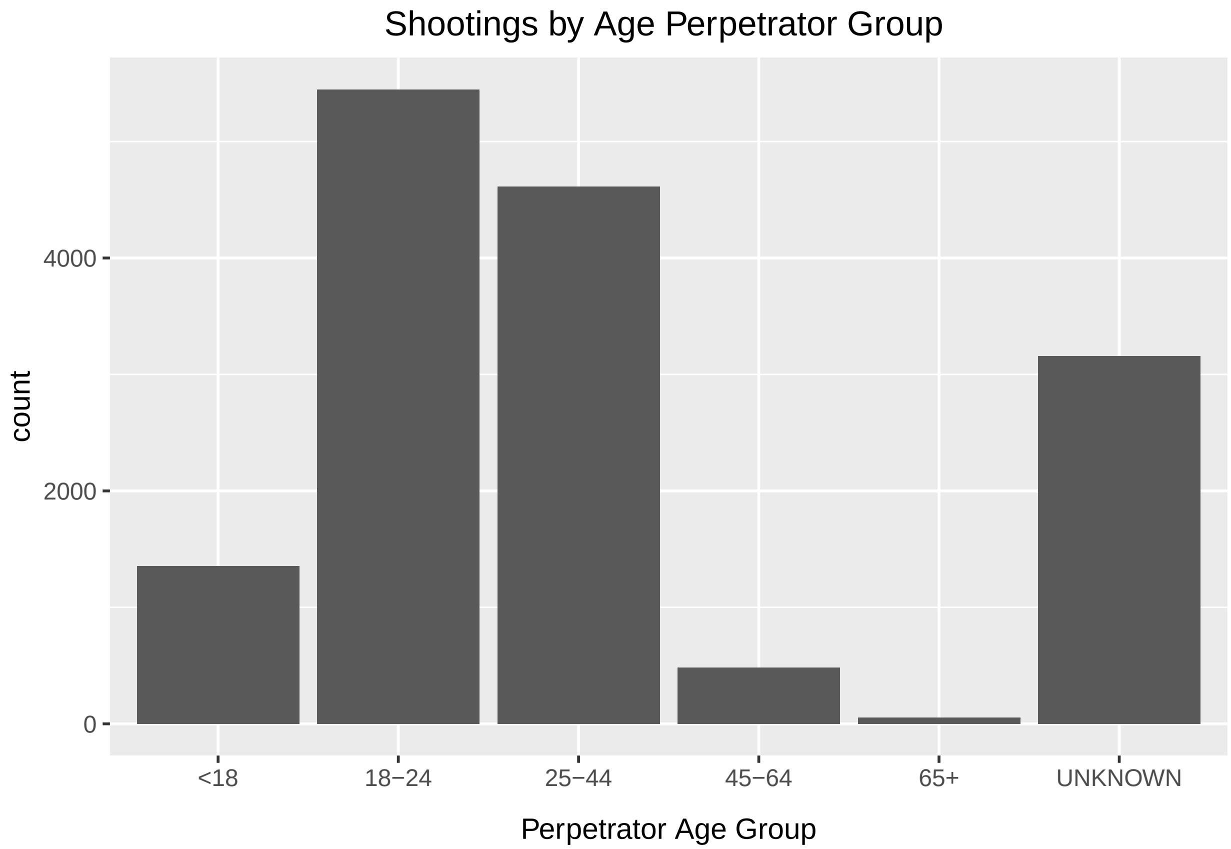


```
#Done
nypd_shootings %>% group_by(BORO) %>% summarise(
  total_shootings = n(),
  statistical_murder = sum(STATISTICAL_MURDER_FLAG == TRUE),
  percentage = statistical_murder / total_shootings) %>%
  arrange(desc(percentage)) %>%
  rename("Cases" = total_shootings, "Murder Label" = statistical_murder,
         "%" = percentage)
```

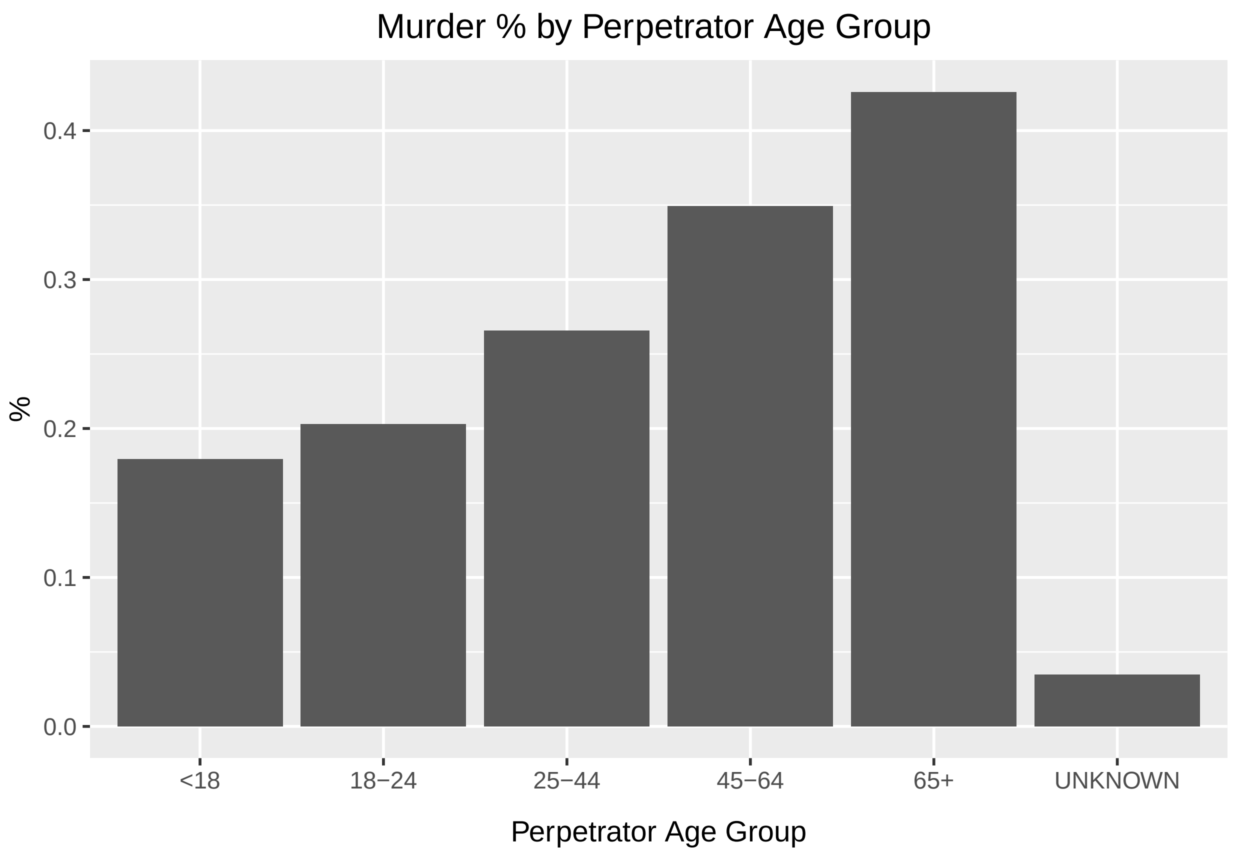
```
## # A tibble: 5 x 4
##   BORO          Cases `Murder Label`   `%`
##   <fct>         <int>          <int> <dbl>
## 1 STATEN ISLAND   566            116 0.205
## 2 BRONX          4497            906 0.201
## 3 QUEENS         2308            449 0.195
## 4 MANHATTAN      1994            367 0.184
## 5 BROOKLYN       5744           1038 0.181
```

Next, differing age groups may have different experiences within the city. Therefore, there may be different reasons for shootings. We can tell by the graphs below there are more shooting incidents between perpetrators of 18-44 years; however, perpetrators aged 45 years or older had a higher proportion of cases being labeled as a murder.

```
#Done
ggplot(nypd_shootings, aes(x=PERP_AGE_GROUP)) +
  geom_bar() +
  labs(title = "Shootings by Age Perpetrator Group ",
       x = "Perpetrator Age Group") +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

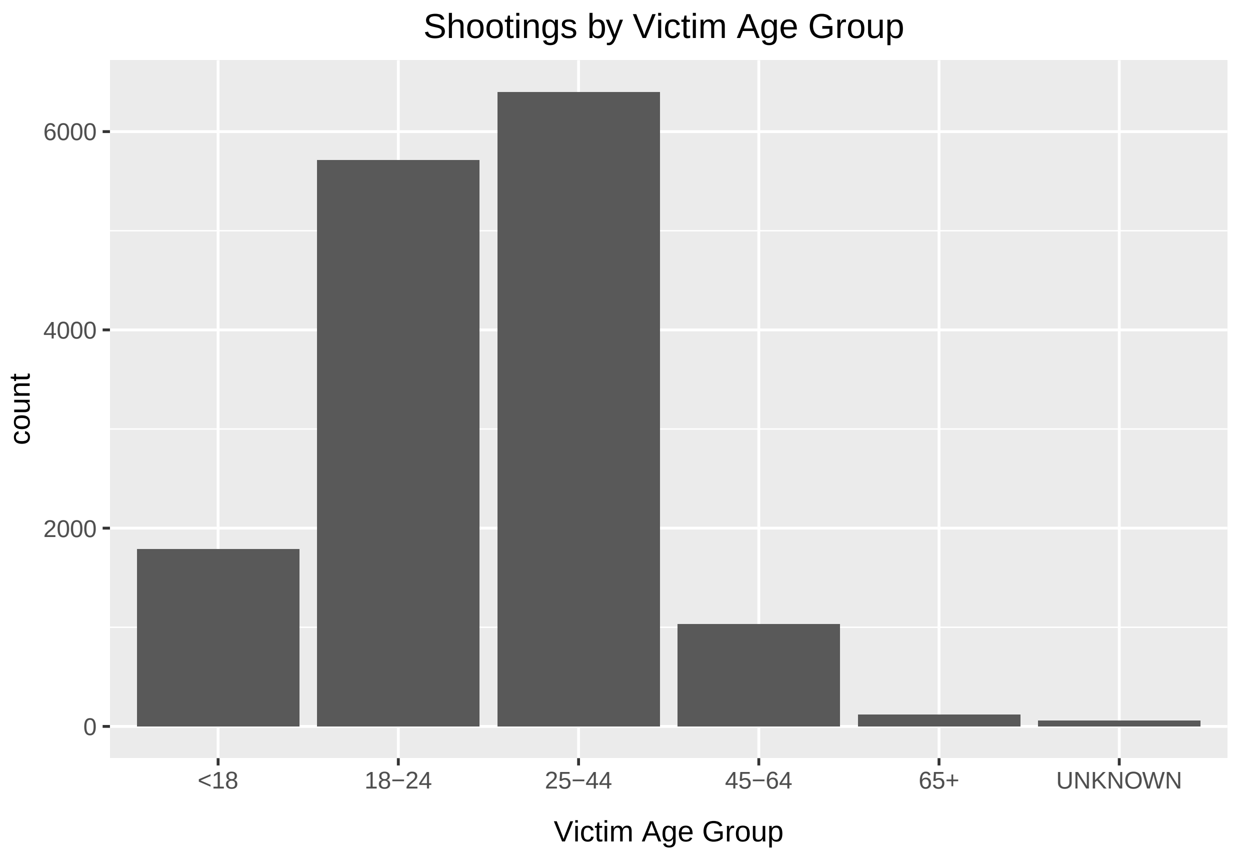


```
#Keep
nypd_shootings %>% group_by(PERP_AGE_GROUP) %>% summarise(
  total_shootings = n(),
  statistical_murder = sum(STATISTICAL_MURDER_FLAG == TRUE),
  percentage = statistical_murder / total_shootings) %>%
  ggplot(aes(x = PERP_AGE_GROUP, y = percentage)) +
  geom_col() +
  labs(title = "Murder % by Perpetrator Age Group ",
       x = "Perpetrator Age Group", y = "%") +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

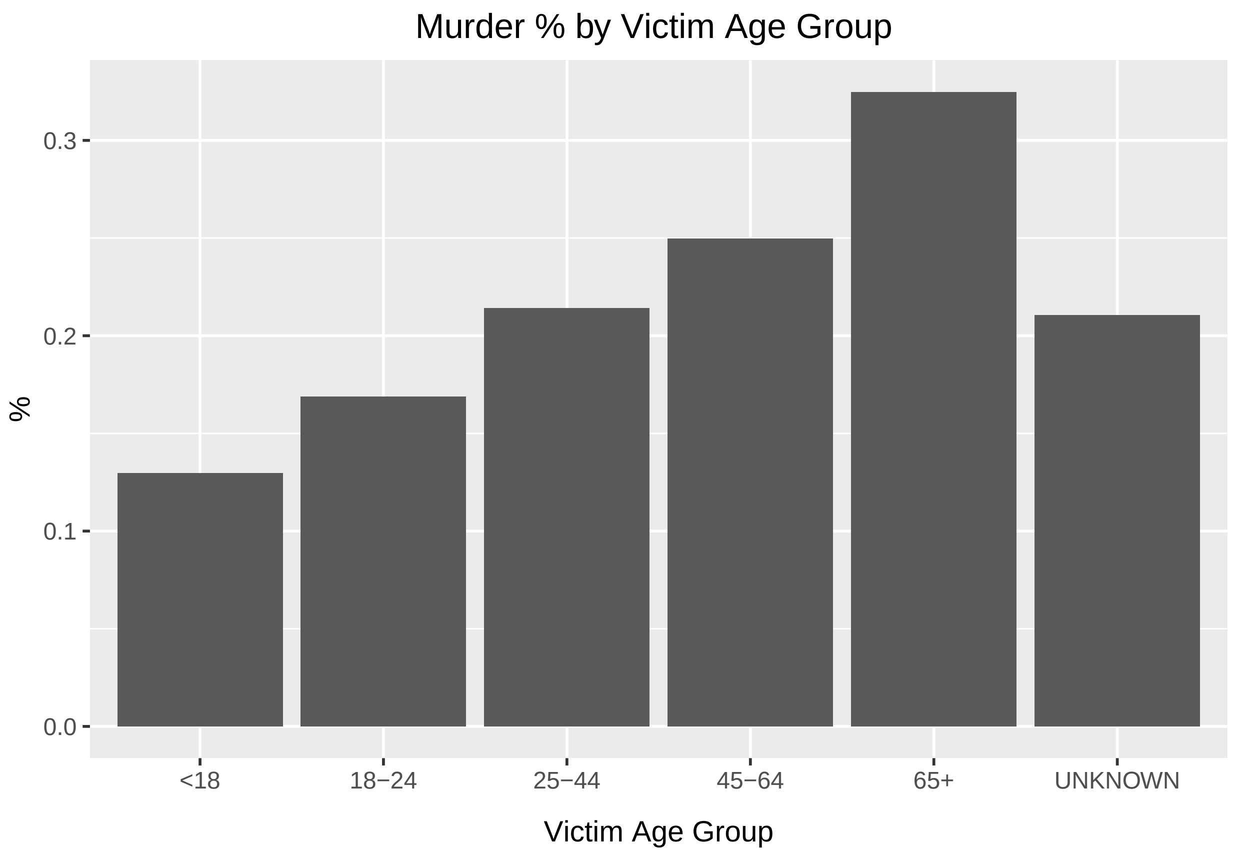


As stated before, different age groups have different experiences within New York City. Similar to the perpetrator, there are a higher number of cases in which the victim was aged between 18-44 while victims aged 45+ years had a higher proportion of their cases labeled as a murder.

```
#Done
ggplot(nypd_shootings, aes(x=VIC_AGE_GROUP)) +
  geom_bar() +
  labs(title = "Shootings by Victim Age Group ", x = "Victim Age Group") +
  theme(axis.title.x = element_text(margin =
                                    margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

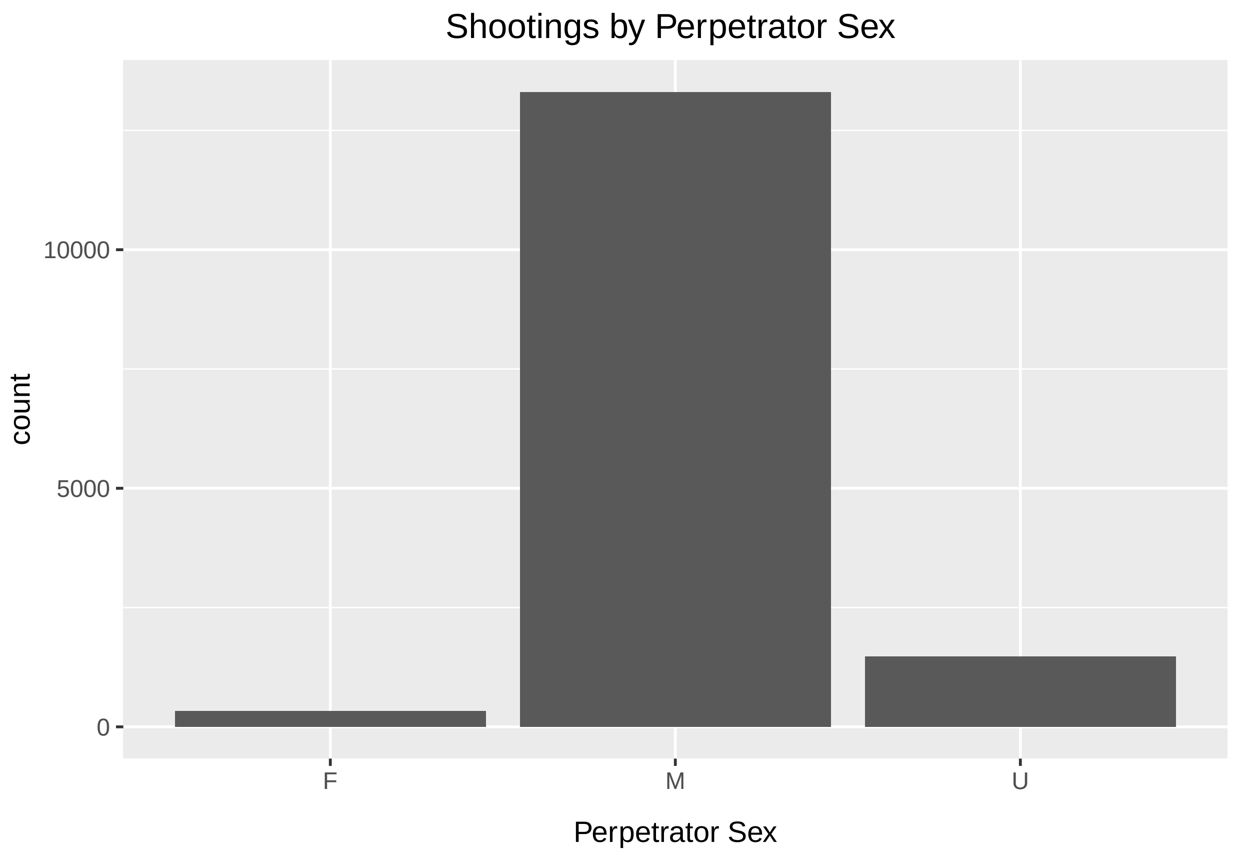


```
#Done
nypd_shootings %>% group_by(VIC_AGE_GROUP) %>% summarise(
  total_shootings = n(),
  statistical_murder = sum(STATISTICAL_MURDER_FLAG == TRUE),
  percentage = statistical_murder / total_shootings) %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = percentage)) +
  geom_col() +
  labs(title = "Murder % by Victim Age Group ", x = "Victim Age Group", y = "%") +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

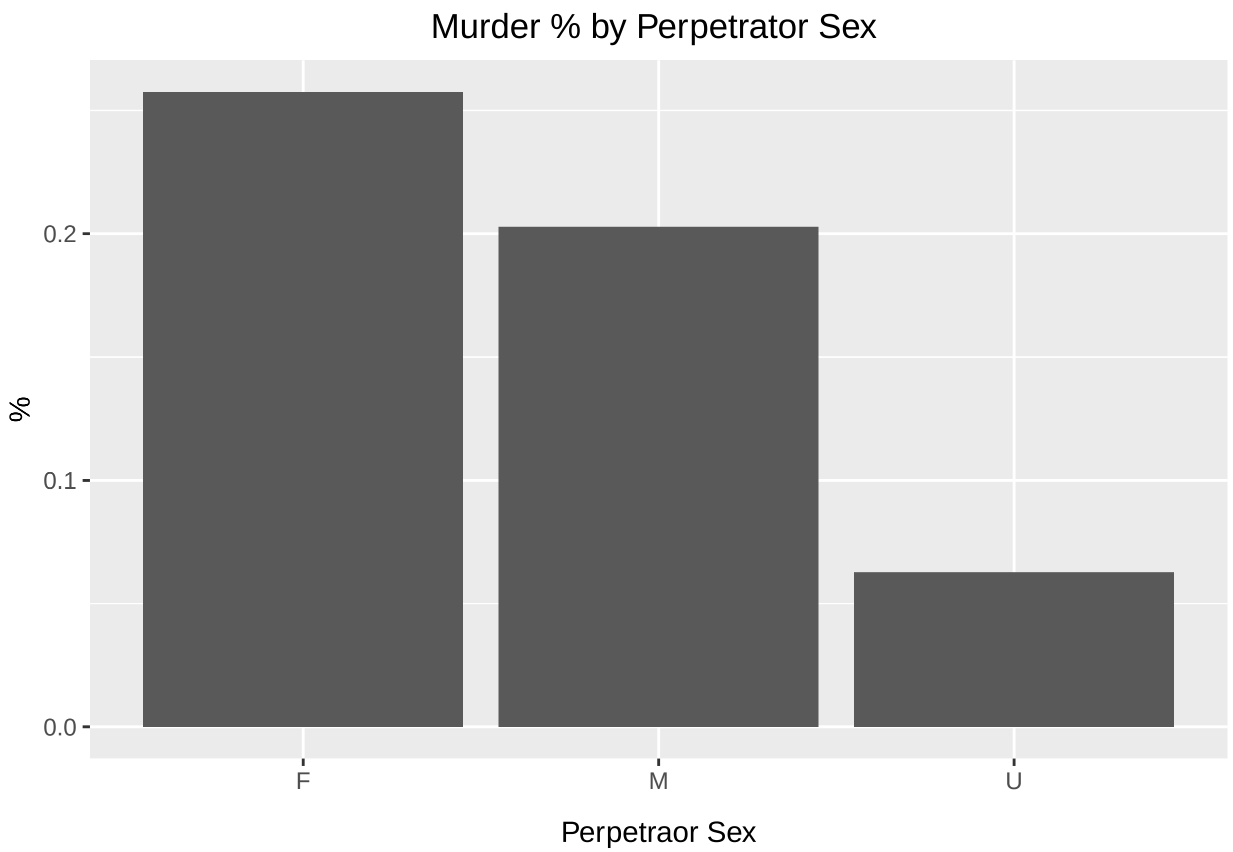


Next, viewing the differences between the quantity of cases among the age groups compared to the differences between murder proportion made me curious to view the differences between the perpetrator/victims sex. For both the victim and perpetrator, males were involved in more shootings compared to females; however, a higher percentage of female cases were considered murders compared to males.

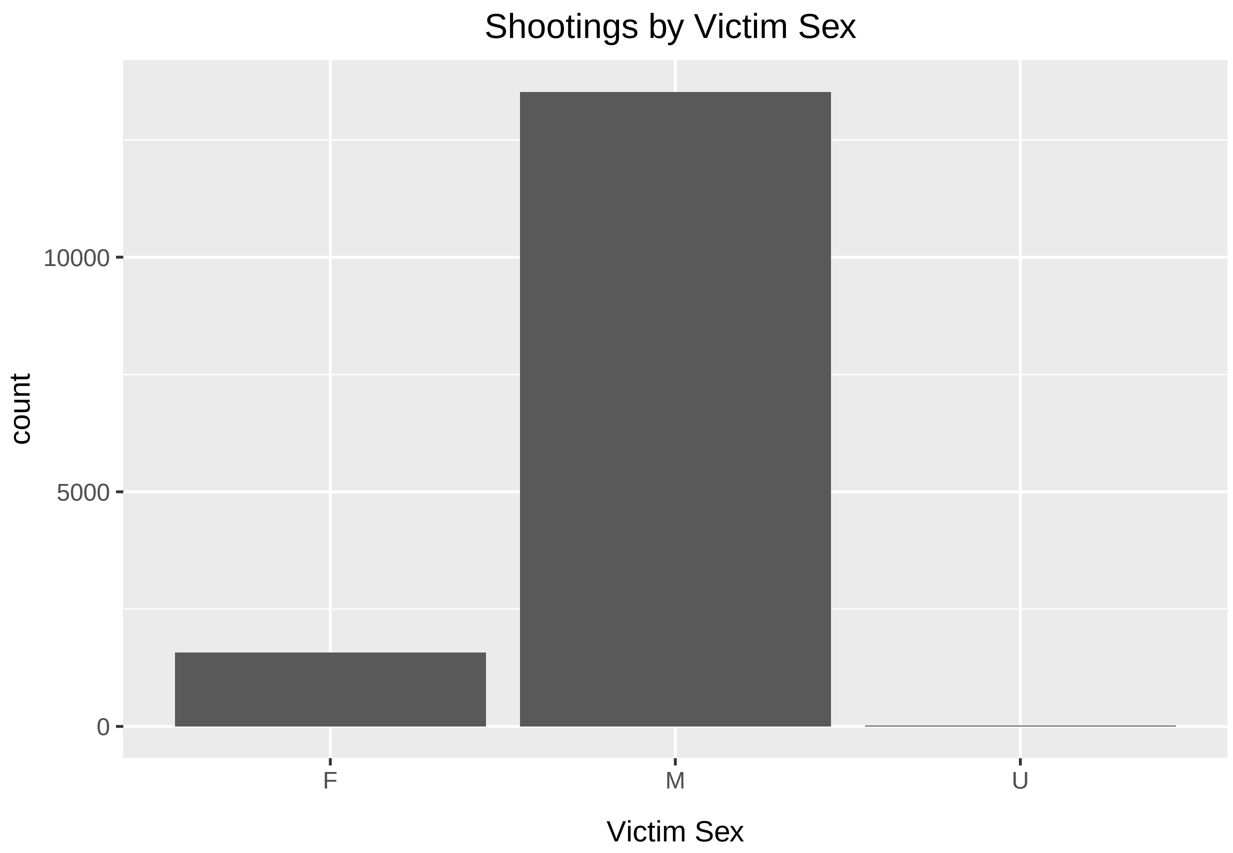
```
#Done
ggplot(nypd_shootings, aes(x=PERP_SEX)) +
  geom_bar() +
  labs(title = "Shootings by Perpetrator Sex ", x = "Perpetrator Sex") +
  theme(axis.title.x = element_text(margin =
                                    margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```



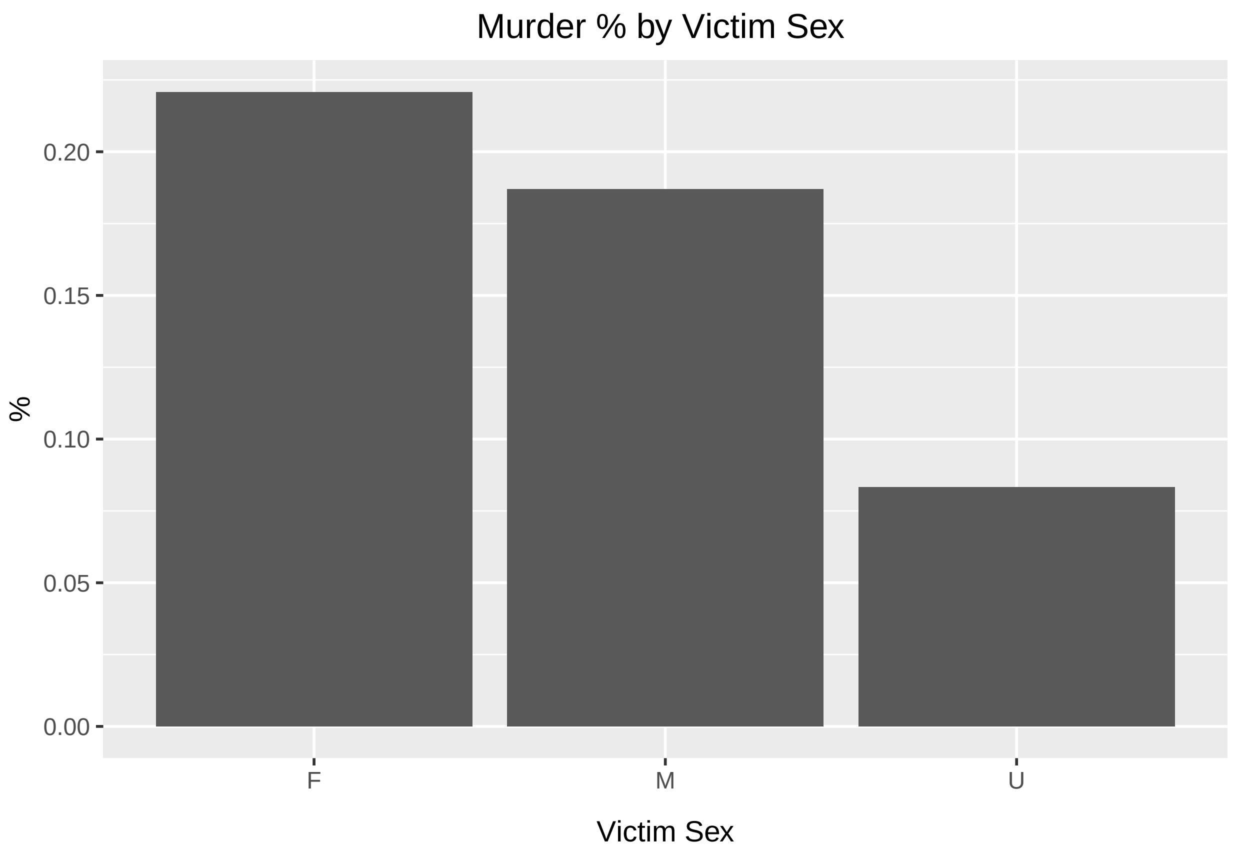
```
#Done
nypd_shootings %>% group_by(PERP_SEX) %>% summarise(
  total_shootings = n(),
  statistical_murder = sum(STATISTICAL_MURDER_FLAG == TRUE),
  percentage = statistical_murder / total_shootings) %>%
  ggplot(aes(x = PERP_SEX, y = percentage)) +
  geom_col() +
  labs(title = "Murder % by Perpetrator Sex ", x = "Perpetraor Sex", y = "%") +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```



```
#Done
ggplot(nypd_shootings, aes(x=VIC_SEX)) +
  geom_bar() +
  labs(title = "Shootings by Victim Sex ", x = "Victim Sex") +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

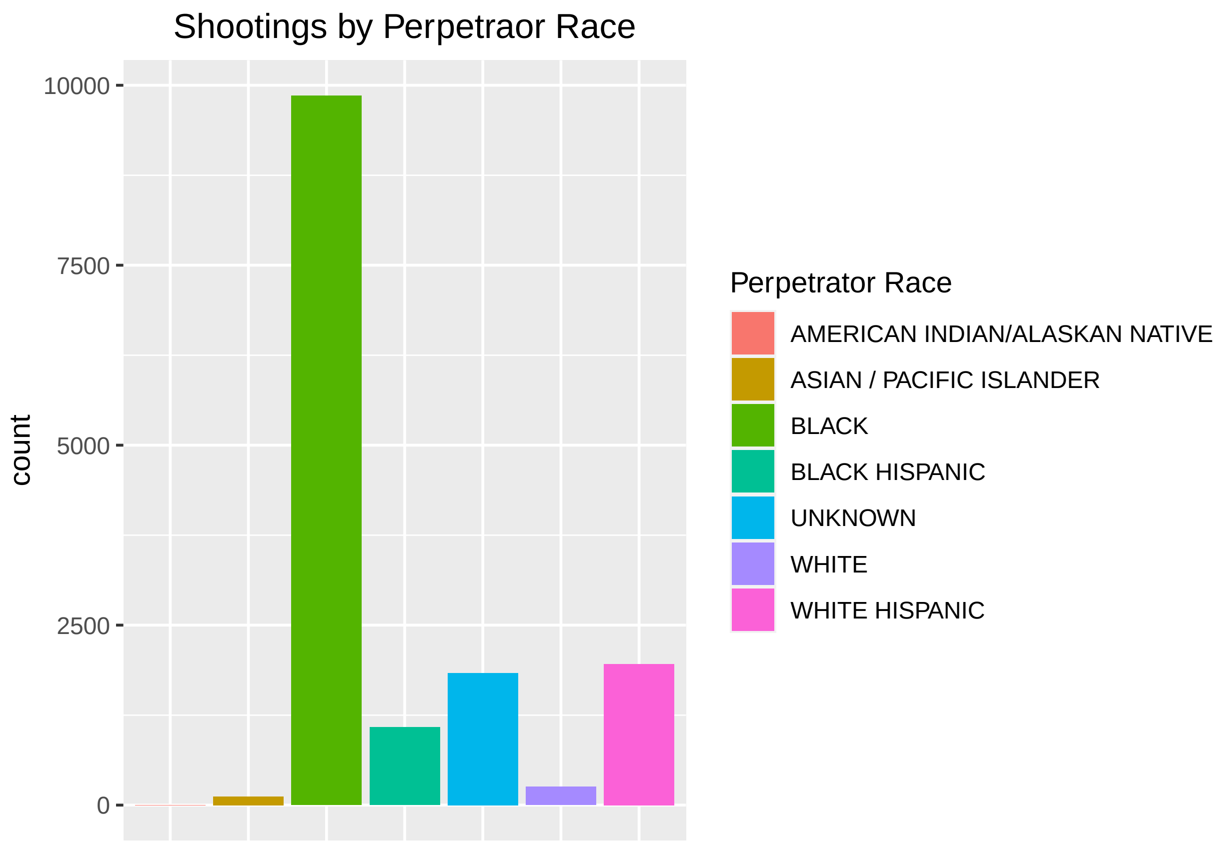


```
#Done
nypd_shootings %>% group_by(VIC_SEX) %>% summarise(
  total_shootings = n(),
  statistical_murder = sum(STATISTICAL_MURDER_FLAG == TRUE),
  percentage = statistical_murder / total_shootings) %>%
  ggplot(aes(x = VIC_SEX, y = percentage)) +
  geom_col() +
  labs(title = "Murder % by Victim Sex ", x = "Victim Sex", y = "%") +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

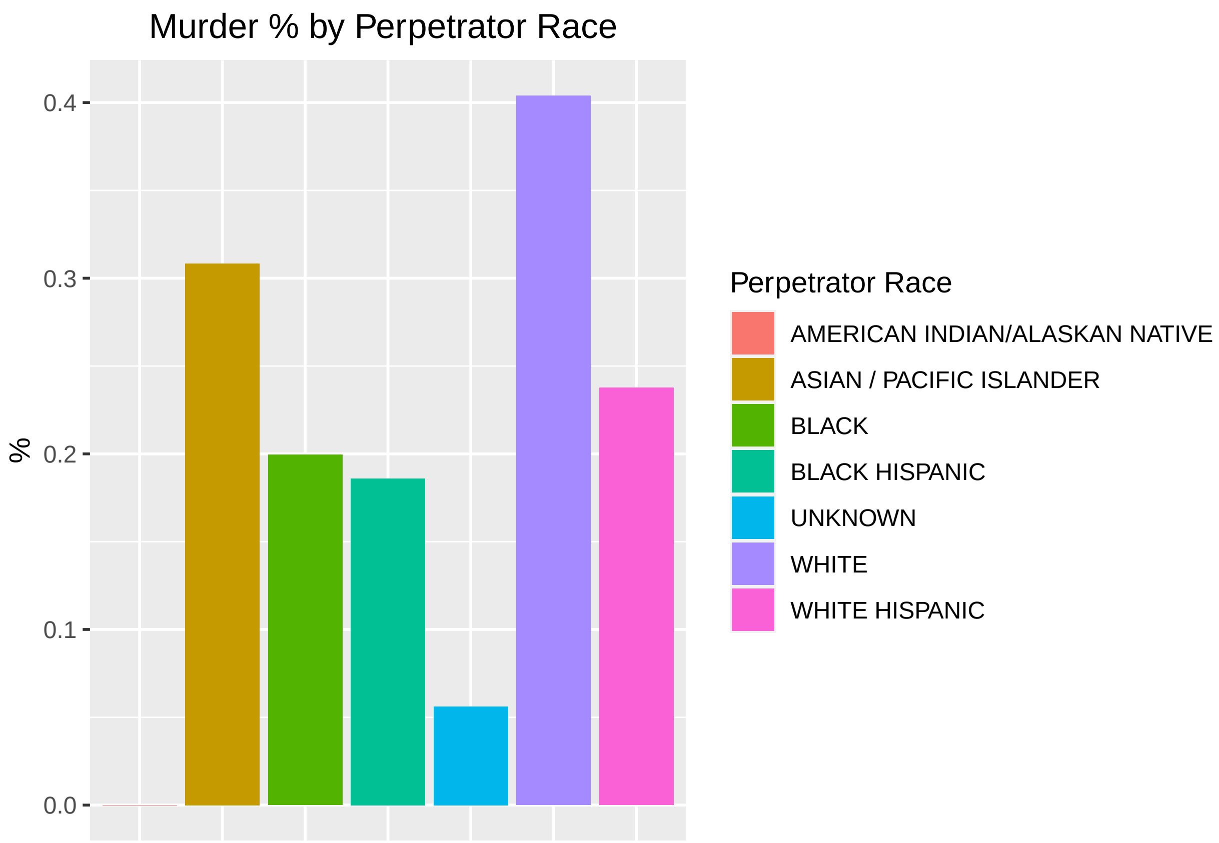


Lastly, many of the shootings involved an african american perpetrator and/or victim. However, a higher percentage of cases involving white americans were labeled as murders.

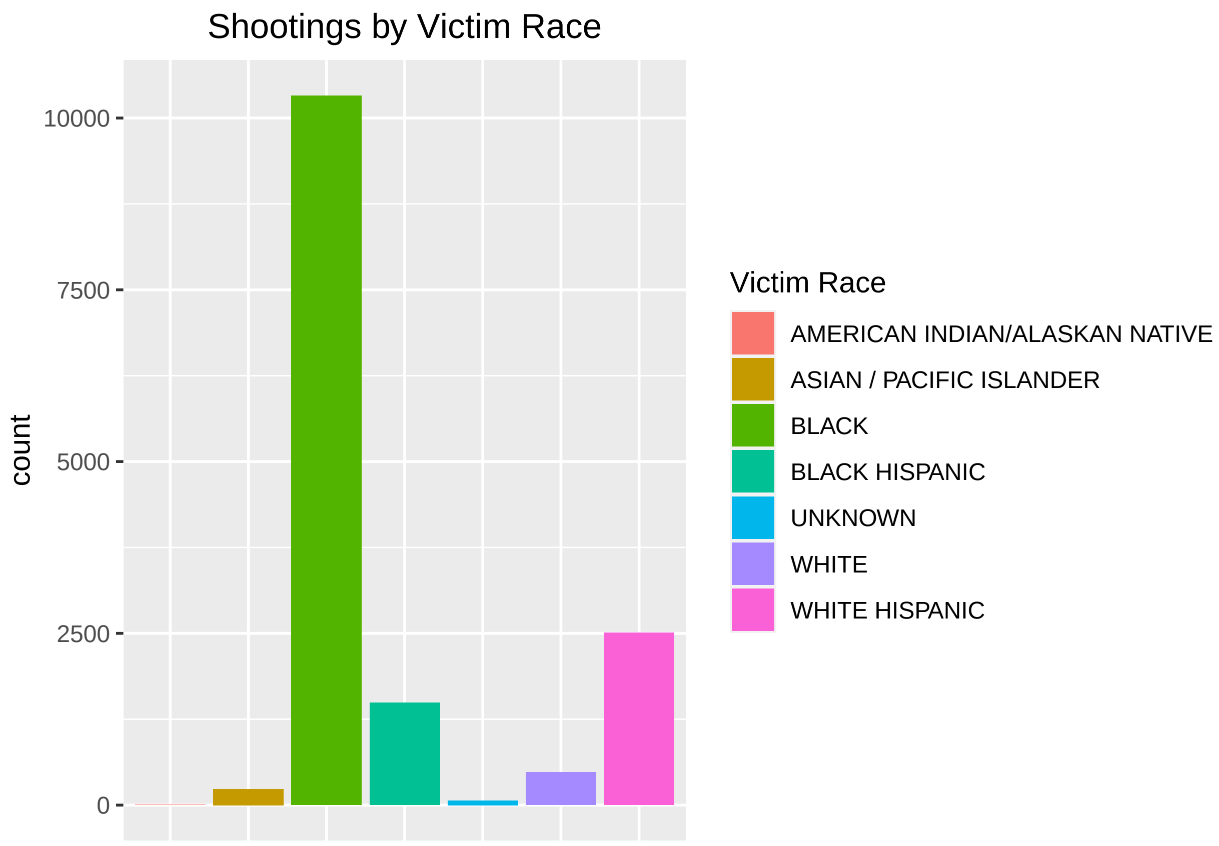
```
#Done
ggplot(nypd_shootings, aes(x=PERP_RACE, fill = PERP_RACE)) +
  geom_bar() +
  labs(title = "Shootings by Perpetraor Race", x = "Perpetrator Race") +
  theme(axis.title.x = element_text(margin =
                                    margin(t = 10)),
        plot.title = element_text(hjust = 0.5)) +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  guides(fill=guide_legend(title="Perpetrator Race"))
```



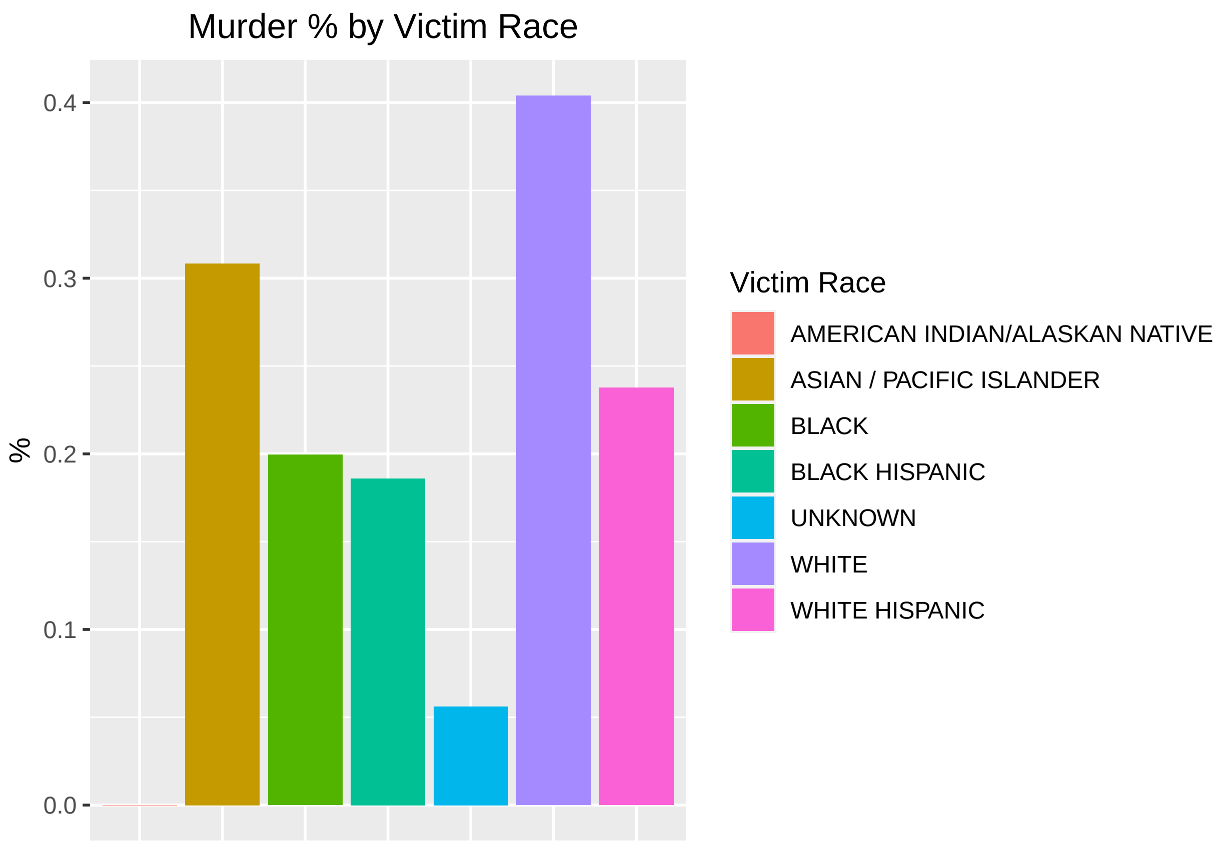
```
#Done
nypd_shootings %>% group_by(PERP_RACE) %>% summarise(
  total_shootings = n(),
  statistical_murder = sum(STATISTICAL_MURDER_FLAG == TRUE),
  percentage = statistical_murder / total_shootings) %>%
  ggplot(aes(x = PERP_RACE, y = percentage, fill = PERP_RACE)) +
  geom_col() +
  labs(title = "Murder % by Perpetrator Race ", x = "Perpetraor Race", y = "%") +
  theme(axis.title.x = element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="Perpetrator Race"))
```



```
ggplot(nypd_shootings, aes(x=VIC_RACE, fill = VIC_RACE)) +
  geom_bar() +
  labs(title = "Shootings by Victim Race", x = "Victim Race") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  guides(fill=guide_legend(title="Victim Race"))
```



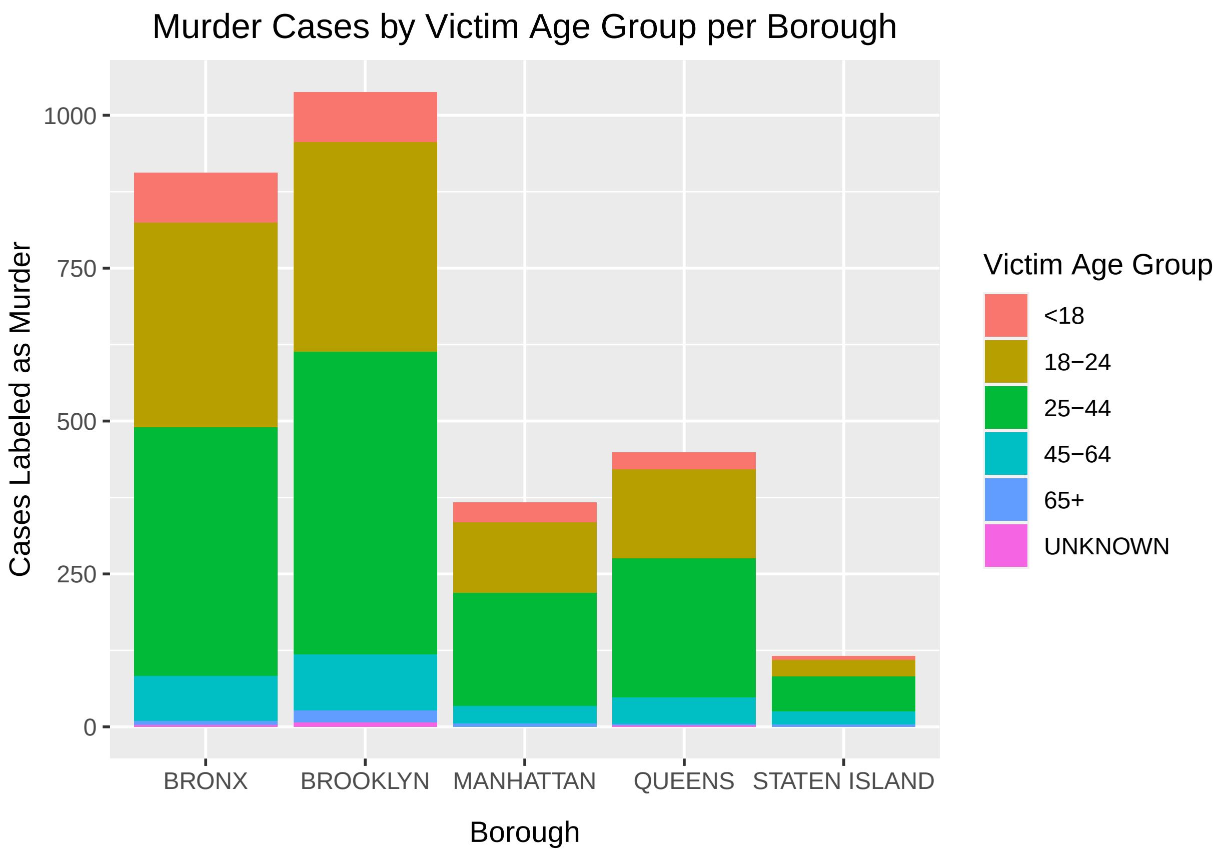
```
nypd_shootings %>% group_by(PERP_RACE) %>% summarise(
  total_shootings = n(),
  statistical_murder = sum(STATISTICAL_MURDER_FLAG == TRUE),
  percentage = statistical_murder / total_shootings) %>%
  ggplot(aes(x = PERP_RACE, y = percentage, fill = PERP_RACE)) +
  geom_col() +
  labs(title = "Murder % by Victim Race ", x = "Victim Race", y = "%") +
  theme(axis.title.x = element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="Victim Race"))
```



The difference in murder percentages compared to the counts of incidents can be due to cultural differences. for example, there are many males that love to hunt. Hunting is physically demanding; therefore, many hunters are younger to middle-aged. Accidents that occur during hunting can be considered a shooting but wouldn't be labeled as a murder. These cultural differences can cause a lower amount of shooting incidents for specific demographics which causes a higher proportion of their cases being murders.

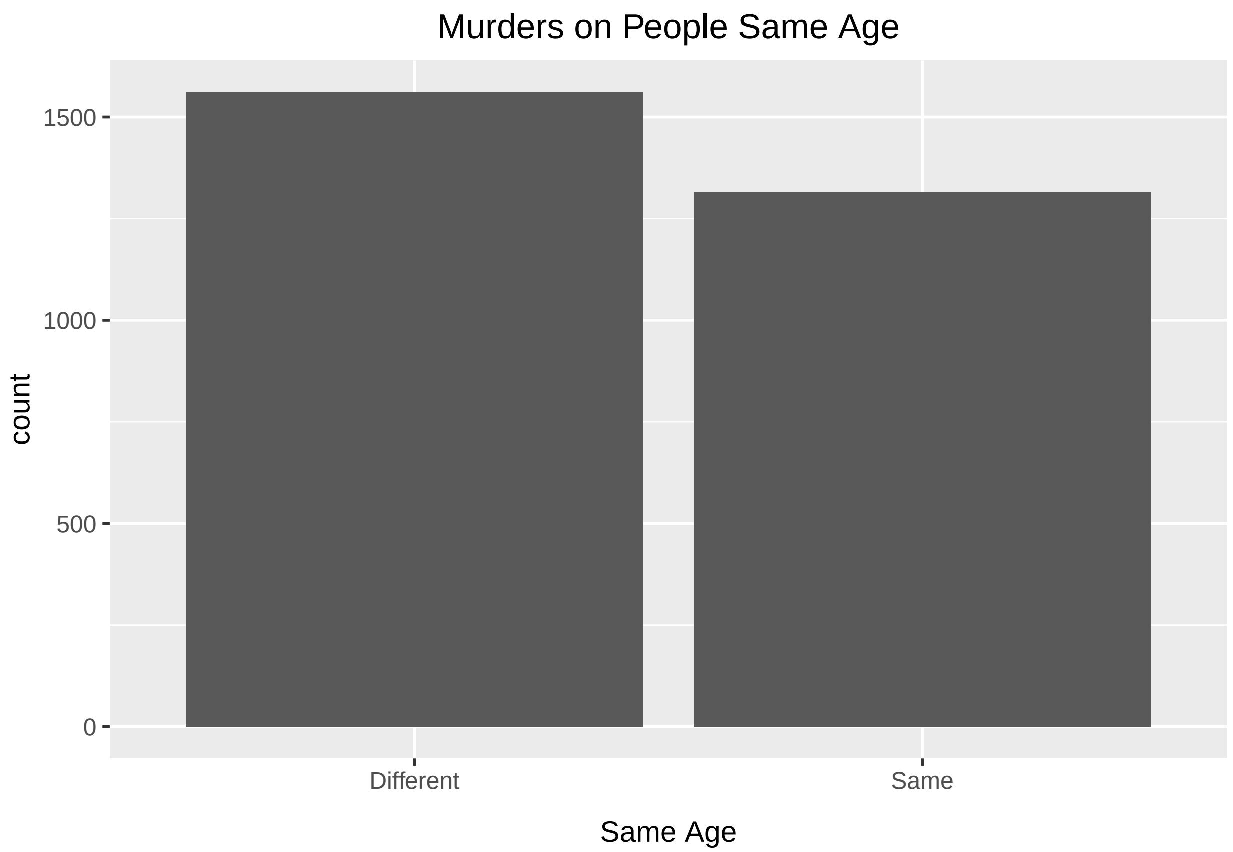
Next, Boroughs have different lifestyles due to location and differing financial situations. However, we can see that murder cases involving 18-44 year old citizens is consistently common across all boroughs.

```
#Done
murder_df <- nypd_shootings %>% filter(STATISTICAL_MURDER_FLAG == TRUE)
murder_df %>%group_by(BORO, VIC_AGE_GROUP) %>%
  ggplot(aes(x = BORO, fill = VIC_AGE_GROUP)) +
  geom_bar() +
  labs(title = "Murder Cases by Victim Age Group per Borough",
       x = "Borough", y = "Cases Labeled as Murder") +
  theme(axis.title.x = element_text(margin =
                                    margin(t = 10)),
        plot.title = element_text(hjust = 0.5))+
  guides(fill=guide_legend(title="Victim Age Group"))
```



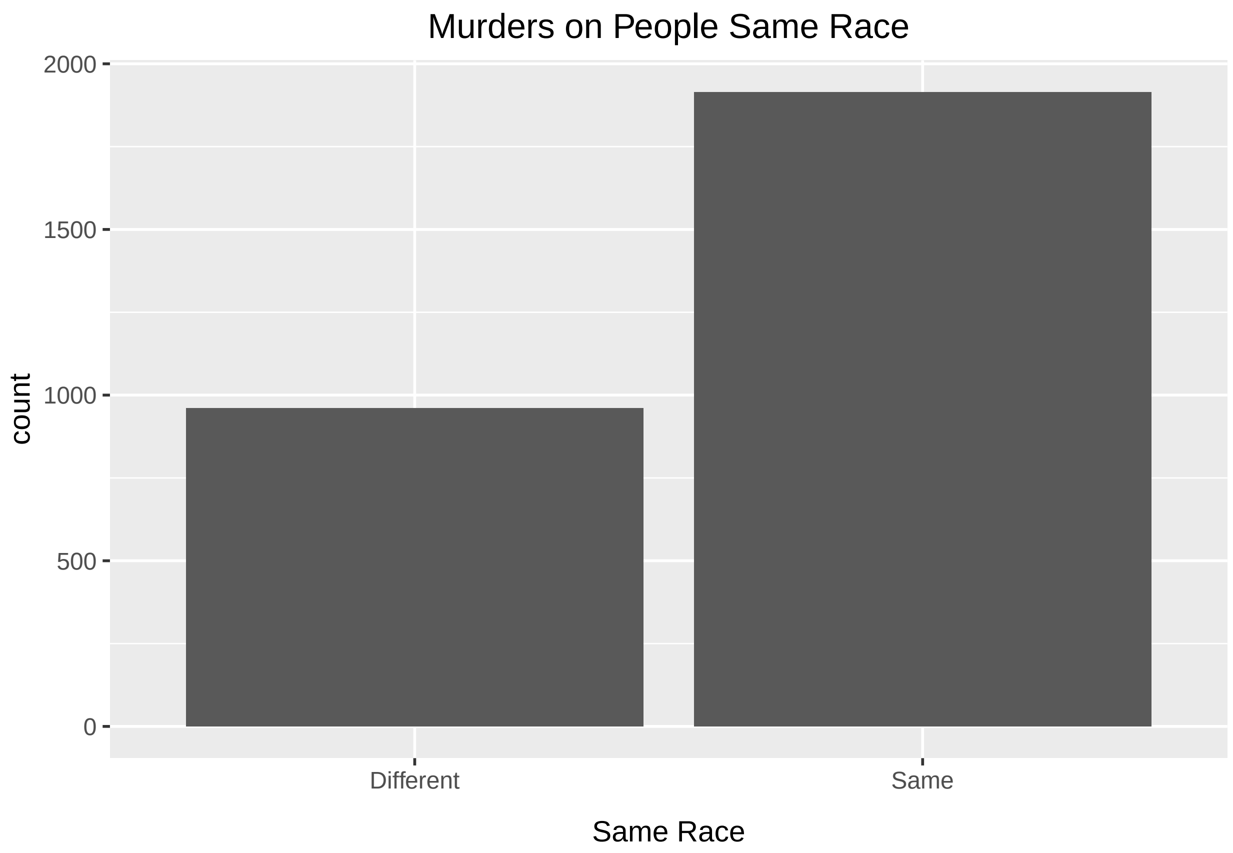
Many citizens are involved in activities and cliques with people of similar age. Does this cause murders where the perpetrator and the victim are the same age? As shown below, although it's not a staggering difference, slightly over half of the murder cases involve situations where the perpetrator and the victim are different ages.

```
#Done
murder_df %>% mutate(same_age = ifelse(as.character(PERP_AGE_GROUP) ==
                                              as.character(VIC_AGE_GROUP),
                                            TRUE, FALSE)) %>%
  group_by(same_age) %>%
  ggplot(aes(x=same_age)) +
  geom_bar() +
  labs(title = "Murders on People Same Age", x = "Same Age") +
  scale_x_discrete(labels = c("TRUE" = "Same", "FALSE" = "Different")) +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```



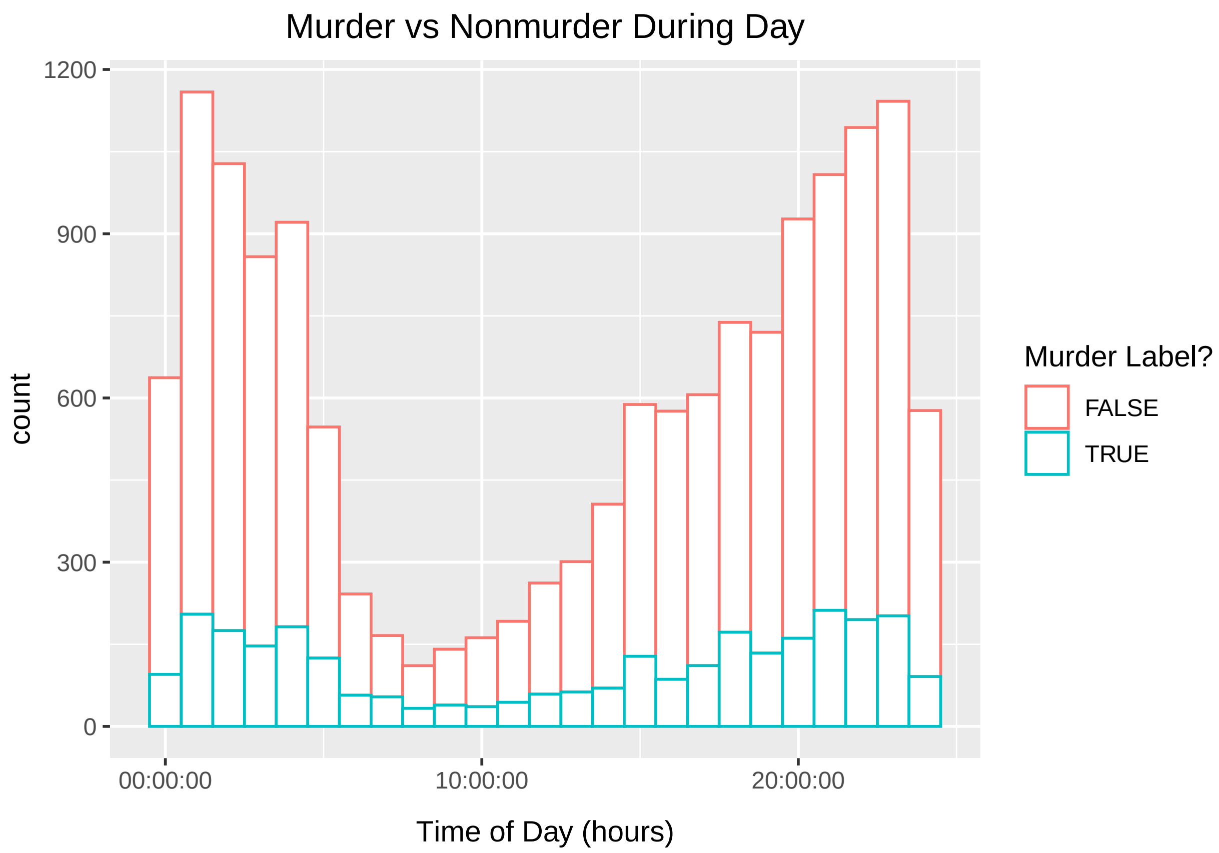
Contrary to age, murders are common among victims of the same race. This could be due to cultural similarities.

```
#Definitely keep, possible make a variable
murder_df %>% mutate(same_race = ifelse(as.character(PERP_RACE) ==
                                              as.character(VIC_RACE),
                                            TRUE, FALSE)) %>%
  group_by(same_race) %>%
  ggplot(aes(x=same_race)) +
  geom_bar() +
  labs(title = "Murders on People Same Race", x = "Same Race") +
  scale_x_discrete(labels = c("TRUE" = "Same", "FALSE" = "Different")) +
  theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```



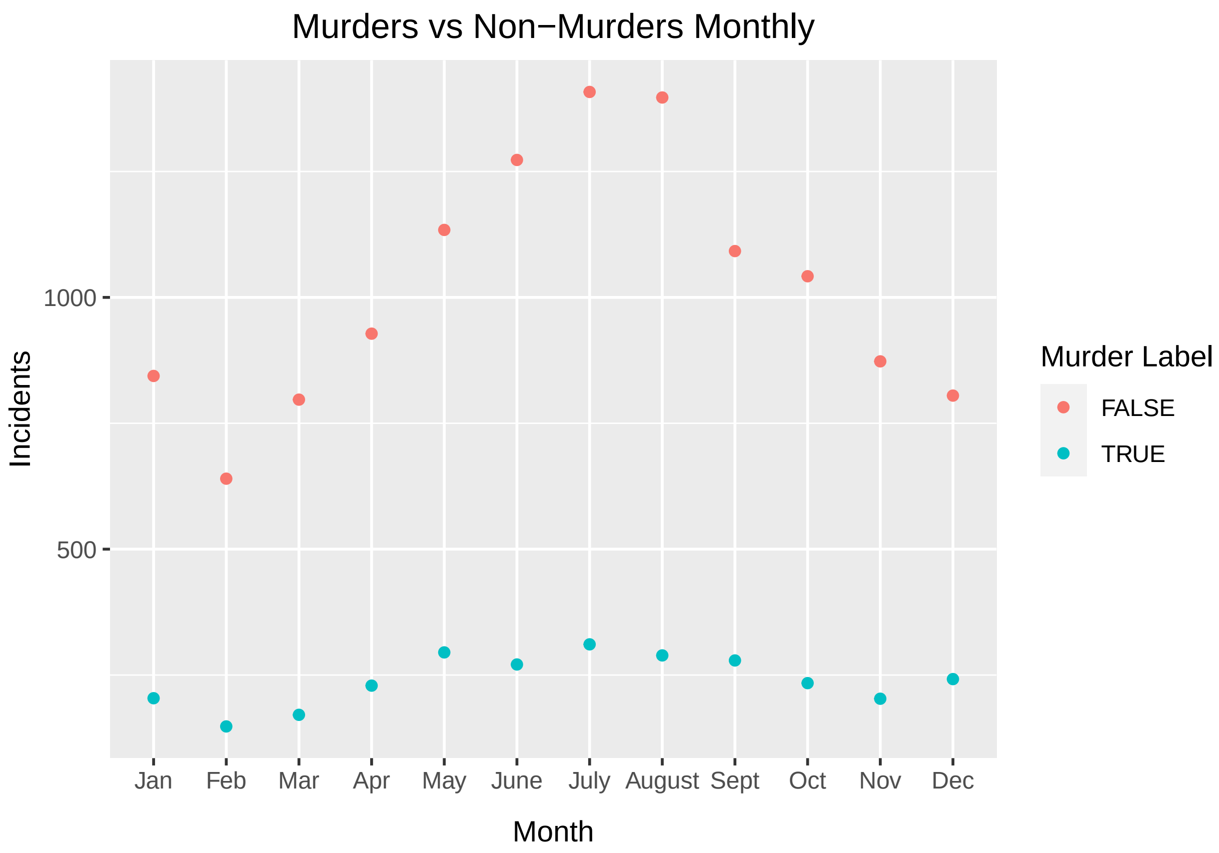
Do murder incidents occur at different times during the day? month? year? Compared to murder cases there is no disparity between a shooting incident during the time of day, month or year. Murder cases follow a similar pattern compared to other incidents. First, all cases are more frequent during the late night time into the early morning. Also, more incidents occur during the summer months compared to other seasons. This could be because of the increase in weather temperature; more people would like to go outside with friends and family causing more interpersonal contact. I doubt anyone wants to be outside during the winter; New York city winters can be quite brutal! Lastly, there seems to be no relationship between the day of the month and shooting incidents, whether they are labeled a murder or not.

```
#look at the % of murders are between the same race group
non_murder_df <- nypd_shootings %>% filter(STATISTICAL_MURDER_FLAG == FALSE)

nypd_shootings <- nypd_shootings %>% mutate(day = format(OCCUR_DATE, "%d"),
                          month = format(OCCUR_DATE, "%m"),
                          year = format(OCCUR_DATE, "%y"))

ggplot(nypd_shootings, aes(x=OCCUR_TIME, color = STATISTICAL_MURDER_FLAG)) +
  geom_histogram(binwidth = 3600, fill="white") + #Every hour
  guides(color=guide_legend(title="Murder Label?"))+
    theme(axis.title.x = element_text(margin =
                                   margin(t = 10)),
        plot.title = element_text(hjust = 0.5)) +
  labs(title = 'Murder vs Nonmurder During Day', x = 'Time of Day (hours)')
```



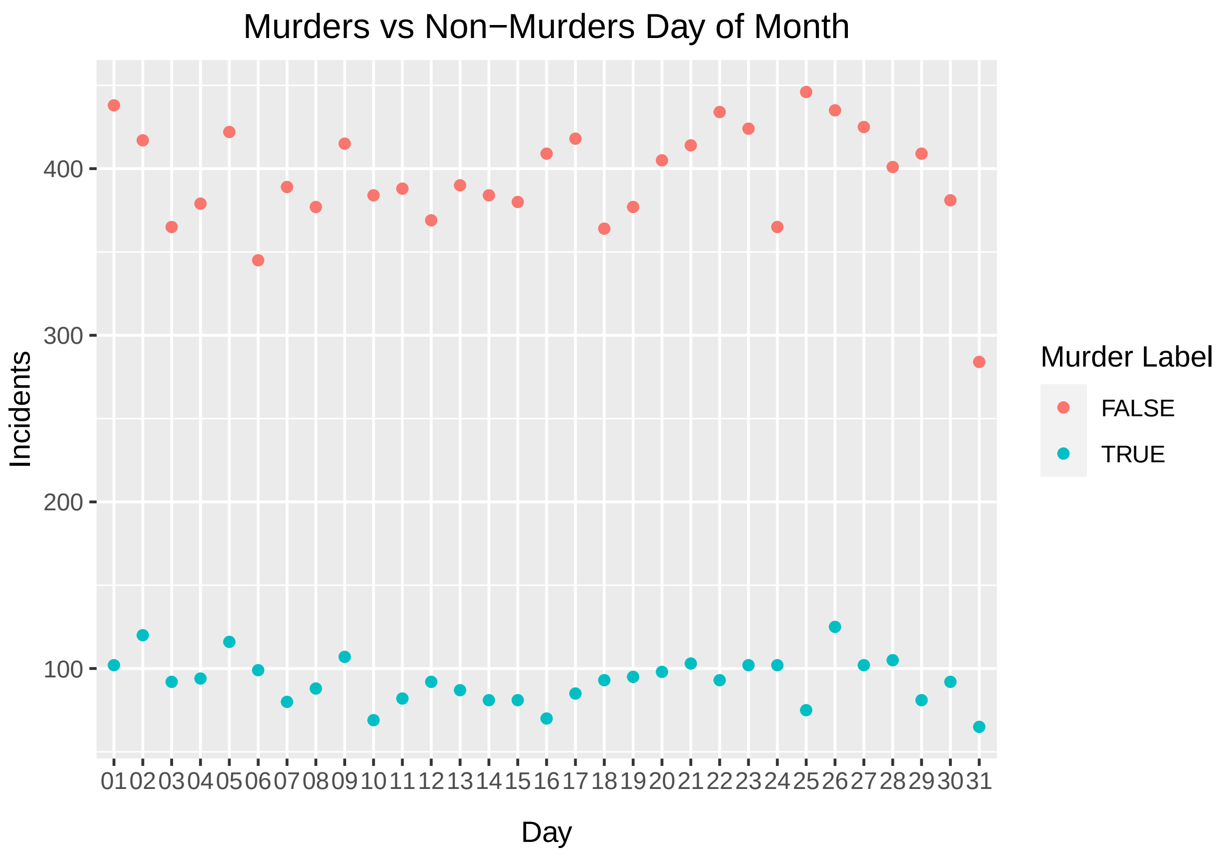
```
#look at the times of the year the murders occur by month vs non murder
nypd_shootings %>% group_by(month, STATISTICAL_MURDER_FLAG) %>%
  summarise(incidents = n()) %>%
  ggplot(aes(x=month, y = incidents)) +
  geom_point(aes(color=STATISTICAL_MURDER_FLAG)) +
  labs(title = "Murders vs Non-Murders Monthly", x= "Month", y="Incidents") +
  guides(color=guide_legend(title="Murder Label"))+
    theme(axis.title.x = element_text(margin =
                                      margin(t = 10)),
        plot.title = element_text(hjust = 0.5)) +
  scale_x_discrete(labels = c("Jan", "Feb", "Mar", "Apr", "May", "June", "July",
                              "August", "Sept", "Oct", "Nov", "Dec"))
```

```
## `summarise()` has grouped output by 'month'. You can override using the `.groups` argument.
```



```
#try to find a map representation of all the data
nypd_shootings %>% group_by(day, STATISTICAL_MURDER_FLAG) %>%
  summarise(incidents = n()) %>%
  ggplot(aes(x=day, y = incidents)) +
  geom_point(aes(color=STATISTICAL_MURDER_FLAG)) +
  labs(title = "Murders vs Non-Murders Day of Month", x= "Day", y="Incidents") +
  guides(color=guide_legend(title="Murder Label"))+
    theme(axis.title.x = element_text(margin =
                                        margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

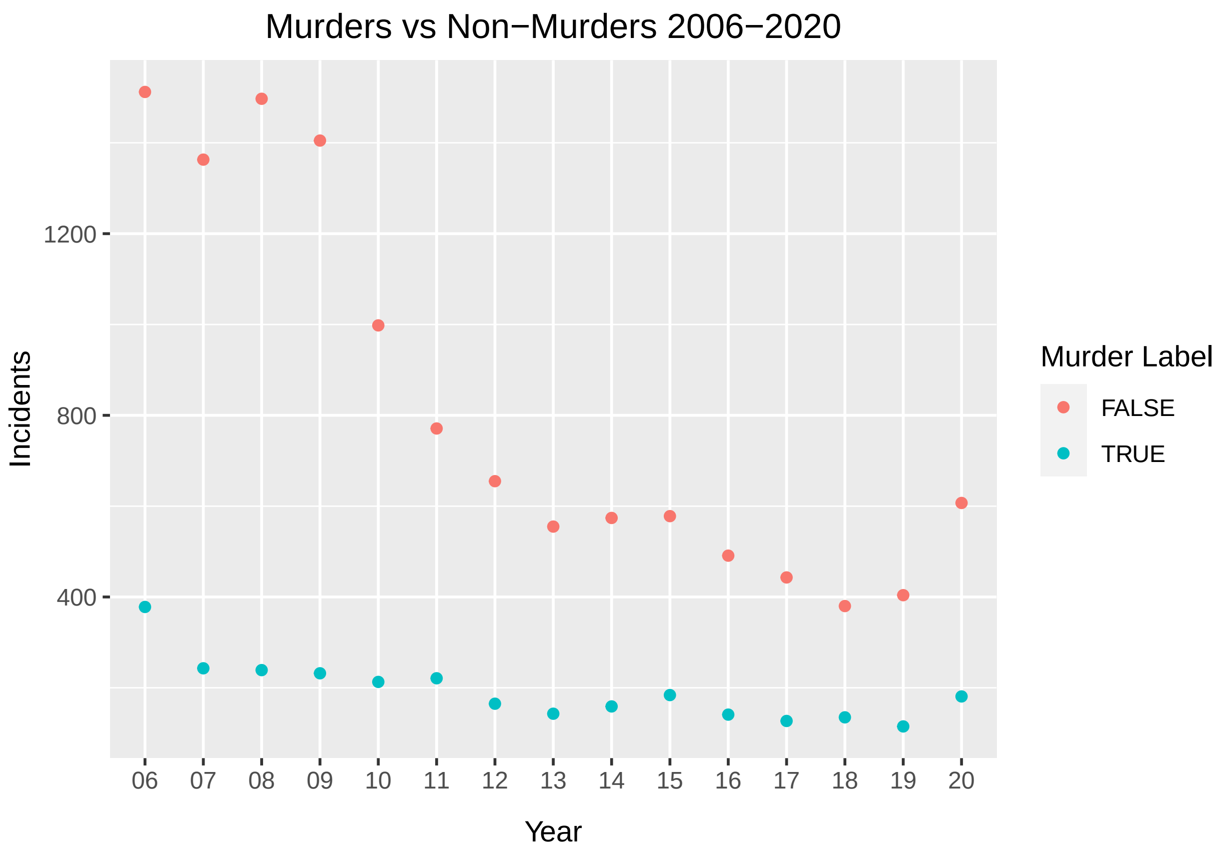
```
## 'summarise()' has grouped output by 'day'. You can override using the '.groups' argument.
```



In a perfect world, everyone would love to live in a city where there are no shooting incidents. Although that is not a reality today, New York city has significantly reduced the number of incidents and statistical murders since 2006. That being said, the policies and cultural lifestyle currently implemented is helping improve the quality of life of the city.

```
#look at the times of the year the murders occur by month vs non murder


nypd_shootings %>% group_by(year, STATISTICAL_MURDER_FLAG) %>%
  summarise(incidents = n()) %>%
  ggplot(aes(x=year, y = incidents)) +
  geom_point(aes(color=STATISTICAL_MURDER_FLAG)) +
  labs(title = "Murders vs Non-Murders 2006-2020", x= "Year", y="Incidents") +
  guides(color=guide_legend(title="Murder Label"))+
    theme(axis.title.x = element_text(margin =
                                        margin(t = 10)),
        plot.title = element_text(hjust = 0.5))
```

```
## ‘summarise()‘ has grouped output by ’year’. You can override using the ‘.groups‘ argument.
```



Model Building

After our analysis, we would like to create a model to predict whether a shooting would be considered a murder or not. First, we would like to drop a few variables that we don't believe are important or will be redundant in our model

```
nypd_shootings <- nypd_shootings %>% select(
  -c("OCCUR_DATE", "PRECINCT", "JURISDICTION_CODE", "day", "year"))
nypd_shootings$month <- as.factor(nypd_shootings$month)
```

Now, we can finally build our model!

```
logit_1 <- glm(STATISTICAL_MURDER_FLAG ~., family = binomial,
               data=nypd_shootings)

summary(logit_1)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ ., family = binomial,
##     data = nypd_shootings)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max  
## -1.5990  -0.7294  -0.6148  -0.1986   3.0804  
## 
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                       -2.353e+01  2.542e+02  -0.093 0.926244    
## OCCUR_TIME                        -8.026e-07  7.291e-07  -1.101 0.270955    
## BOROBROOKLYN                      -9.572e-02  5.466e-02  -1.751 0.079899 .  
## BOROMANHATTAN                     -1.237e-01  7.119e-02  -1.738 0.082162 .  
## BOROQUEENS                        -9.586e-02  6.826e-02  -1.404 0.160225    
## BOROSTATEN ISLAND                 -2.276e-01  1.152e-01  -1.976 0.048164 *  
## PERP_AGE_GROUP18-24                8.765e-02  7.995e-02   1.096 0.272931    
## PERP_AGE_GROUP25-44                3.616e-01  8.166e-02   4.428 9.51e-06 ***
## PERP_AGE_GROUP45-64                6.479e-01  1.248e-01   5.191 2.09e-07 ***
## PERP_AGE_GROUP65+                  8.223e-01  3.008e-01   2.733 0.006268 ** 
## PERP_AGE_GROUPUNKNOWN             -2.444e+00  1.806e-01 -13.535  < 2e-16 ***
## PERP_SEXM                         -1.364e-01  1.301e-01  -1.049 0.294318    
## PERP_SEXU                          1.718e+00  2.936e-01   5.852 4.86e-09 ***
## PERP_RACEASIAN / PACIFIC ISLANDER  1.209e+01  2.295e+02   0.053 0.957991    
## PERP_RACEBLACK                     1.172e+01  2.295e+02   0.051 0.959270    
## PERP_RACEBLACK HISPANIC            1.158e+01  2.295e+02   0.050 0.959745    
## PERP_RACEUNKNOWN                   1.091e+01  2.295e+02   0.048 0.962080    
## PERP_RACEWHITE                     1.230e+01  2.295e+02   0.054 0.957246    
## PERP_RACEWHITE HISPANIC            1.184e+01  2.295e+02   0.052 0.958862    
## VIC_AGE_GROUP18-24                 2.648e-01  8.172e-02   3.240 0.001193 ** 
## VIC_AGE_GROUP25-44                 4.185e-01  8.137e-02   5.143 2.70e-07 ***
## VIC_AGE_GROUP45-64                 4.712e-01  1.072e-01   4.396 1.10e-05 ***
## VIC_AGE_GROUP65+                   7.790e-01  2.243e-01   3.474 0.000514 ***
## VIC_AGE_GROUPUNKNOWN               1.863e-01  3.559e-01   0.524 0.600598    
## VIC_SEXM                          -6.939e-02  6.824e-02  -1.017 0.309233    
## VIC_SEXU                          -8.734e-01  1.089e+00  -0.802 0.422729    
## VIC_RACEASIAN / PACIFIC ISLANDER   1.058e+01  1.093e+02   0.097 0.922885    
## VIC_RACEBLACK                      1.036e+01  1.093e+02   0.095 0.924536    
## VIC_RACEBLACK HISPANIC             1.009e+01  1.093e+02   0.092 0.926454    
## VIC_RACEUNKNOWN                    1.017e+01  1.093e+02   0.093 0.925908    
## VIC_RACEWHITE                      1.047e+01  1.093e+02   0.096 0.923696    
## VIC_RACEWHITE HISPANIC             1.047e+01  1.093e+02   0.096 0.923689    
## month02                           -3.289e-02  1.240e-01  -0.265 0.790868    
## month03                           -1.570e-01  1.186e-01  -1.324 0.185491    
## month04                            4.638e-02  1.111e-01   0.417 0.676455    
## month05                            7.677e-02  1.053e-01   0.729 0.465837    
## month06                           -1.224e-01  1.061e-01  -1.154 0.248669    
## month07                           -5.058e-02  1.035e-01  -0.489 0.624905    
## month08                           -1.258e-01  1.046e-01  -1.203 0.228892    
## month09                            8.850e-02  1.065e-01   0.831 0.406119    
## month10                           -3.933e-02  1.100e-01  -0.358 0.720643    
## month11                           -1.274e-02  1.141e-01  -0.112 0.911067    
## month12                            2.480e-01  1.111e-01   2.232 0.025586 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 14708  on 15108  degrees of freedom
## Residual deviance: 13565  on 15066  degrees of freedom
## AIC: 13651
##
## Number of Fisher Scoring iterations: 11
```

As we can see from the model, the variables with the coefficients that are positive (all perpetrator age types except ‘unknown’, an unknown perpetrator sex, or the incident occurring in April) increase the probability that the incident is a murder case. On the other hand, a negative coefficient (the month of July or November, or a male perpetrator) decrease the likelihood of the shooting being a murder case. It’s quite surprising that the summer months negatively impact if a case was murder or not!

Conclusion

Statistical murders follow the same patterns and only account for 20% of all shootings. First, most cases involve males between the ages of 18 and 44. Although Brooklyn and the Bronx have a higher number of murders in their boroughs, the percentage of those shootings that are murders is similar to the other areas. On the other hand, cases in which the perpetrator and/or victim has an age of 65+ are more likely to be considered a murder case. The same can be said about cases involving a female and/or white perpetrator/victim. Shootings involving the same race are likely to be considered murders while ones involving different different age groups are not. Lastly, murders do not differentiate from regular shootings when discussing chronological data. Both regular shootings and murders will increase during the night time and during summer months; this could be due to warmer weather, or just a little more free time.

There is possible that there is bias contained within this report. First, we only have a few variables to look at. Income disparities, population density, etc. are important variables to consider when looking at this data. Maybe a higher population density would lead to a higher proportion of shootings being a murder? More data could help us look at other factors that contribute to the differences between shooting types. Also, the variables in this dataset have high multicollinearity. In other words, many of the variables aren’t independent of each other; therefore, increasing or decreasing one variable amount may cause another to increase or decrease unintentionally. This causes major statistical errors when attempting to make predictions. Third, I wanted to avoid looking into specific variables due to ethical issues. Race was a variable I used in the model; however, I didn’t want to investigate it much due to different issues an African American my face compared to a Caucasian or vice versa. Investigating racial differences can be tricky when working with data. Lastly, I have my own personal biases. I may have made some choices on how to look at certain parts of the data, subconsciously. For example, I chose to focus on whether a shooting was considered a murder. However, I could of focused on the disparities between different boroughs.

In conclusion, New York City has a long history of gun violence. Lets focus on controlling the violence ot make New York great for everyone.