Week 3 Assingment - NYPD Shooting Incident Data

```
library(tidyverse)
```

Importing tidyverse and lubridate libraries.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Step 1: Importing Data

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

Importing csv file from data.cityofnewyork.us

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO PRECINCT JURISDICTION_CODE
## 1     24050482 08/27/2006   05:35:00    BRONX       52                 0
## 2     77673979 03/11/2011   12:03:00   QUEENS      106                 0
## 3    203350417 10/06/2019   01:09:00 BROOKLYN       77                 0
## 4     80584527 09/04/2011   03:35:00    BRONX       40                 0
## 5     90843766 05/27/2013   21:16:00   QUEENS      100                 0
```

```
## 6     92393427 09/01/2013   04:17:00 BROOKLYN       67                 0
##   LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE
## 1                                  true
## 2                                 false
## 3                                 false
## 4                                 false
## 5                                 false
## 6                                 false
##   VIC_AGE_GROUP VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1         25-44       F BLACK HISPANIC    1017542   255918.9 40.86906 -73.87963
## 2           65+       M          WHITE    1027543   186095.0 40.67737 -73.84392
## 3         18-24       F          BLACK     995325   185155.0 40.67489 -73.96008
## 4           <18       M          BLACK    1007453   233952.0 40.80880 -73.91618
## 5         18-24       M          BLACK    1041267   157133.5 40.59780 -73.79469
## 6           <18       M          BLACK    1001694   170112.9 40.63359 -73.93715
##                                        Lon_Lat
## 1  POINT (-73.87963173099996 40.86905819000003)
## 2 POINT (-73.84392019199998 40.677366895000034)
## 3 POINT (-73.96007501899999 40.674885741000026)
## 4  POINT (-73.91618413199996 40.80879780500004)
## 5 POINT (-73.79468553799995 40.597796249000055)
## 6  POINT (-73.93715330699996 40.63358818100005)
```

Step 2: Tidy and Transform Data

```
df_new <- df %>% select(OCCUR_DATE, OCCUR_TIME, BORO, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_SEX
head(df_new)
```

Since I don't think a lot of the columns are relevant, I will only choose "OCCUR_DATE", "OCCUR_TIME", "BORO", "STATISTICAL_MURDER_FLAG", "PERP_AGE_GROUP", "PERP_SEX", "PERP_RACE", "VIC_AGE_GROUP", "VIC_SEX",and "VIC_RACE".

```
##   OCCUR_DATE OCCUR_TIME     BORO STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1 08/27/2006   05:35:00    BRONX                    true
## 2 03/11/2011   12:03:00   QUEENS                   false
## 3 10/06/2019   01:09:00 BROOKLYN                   false
## 4 09/04/2011   03:35:00    BRONX                   false
## 5 05/27/2013   21:16:00   QUEENS                   false
## 6 09/01/2013   04:17:00 BROOKLYN                   false
##   PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                            25-44       F BLACK HISPANIC
## 2                              65+       M          WHITE
## 3                            18-24       F          BLACK
## 4                              <18       M          BLACK
## 5                            18-24       M          BLACK
## 6                              <18       M          BLACK
```

There are numbers of missing perpetrator information in this dataset, maybe due to the fact that the perpetrator was never caught, or it's under active investigation, so I will change the

blank spaces to unknown to match rest of the dataset. I will also update time and dates to a more readable format for easy analysis.

```
df_new[df_new == ''] <- NA
df_new <- replace_na(df_new, list(PERP_AGE_GROUP = "UNKNOWN", PERP_SEX = "U", PERP_RACE = "UNKNOWN"))
df_new$OCCUR_DATE<-mdy(df_new$OCCUR_DATE)
df_new$OCCUR_DATE<-wday(df_new$OCCUR_DATE, label=TRUE, abbr=FALSE)
df_new$OCCUR_TIME<-hour(hms(df_new$OCCUR_TIME))
df_new[df_new == 1020] <- "UNKNOWN"
df_new[df_new == 224] <- "UNKNOWN"
df_new[df_new == 940] <- "UNKNOWN"
df_new[df_new == "true"] <- "1"
df_new[df_new == "false"] <- "0"
head(df_new)
```

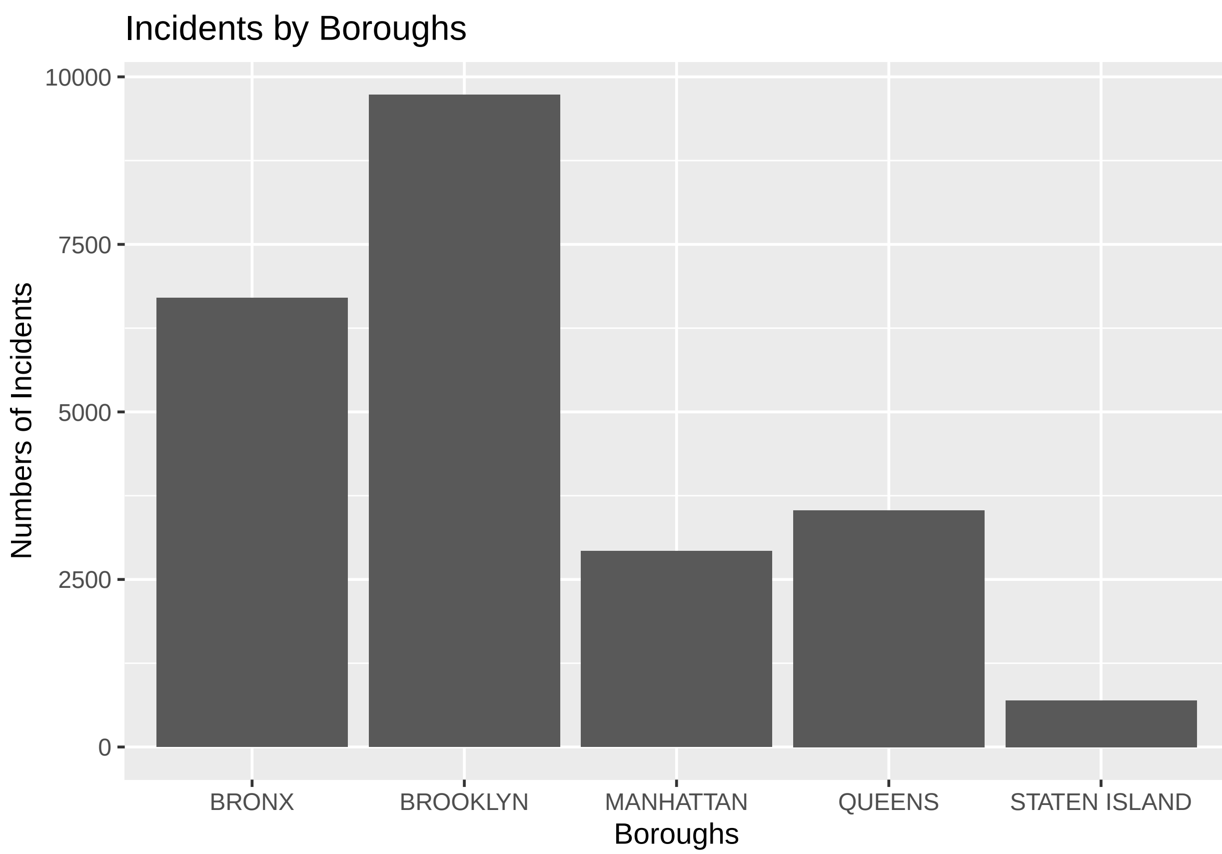
While doing analysis of this dataset, I found that there are strange numbers in the PERP_AGE column, so I will also replace these numbers with "UNKNOWN". I will also change true and false to 1 and 0 respectively to make it easier for modeling.

```
##   OCCUR_DATE OCCUR_TIME     BORO STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1     Sunday          5    BRONX                       1        UNKNOWN
## 2     Friday         12   QUEENS                       0        UNKNOWN
## 3     Sunday          1 BROOKLYN                       0        UNKNOWN
## 4     Sunday          3    BRONX                       0        UNKNOWN
## 5     Monday         21   QUEENS                       0        UNKNOWN
## 6     Sunday          4 BROOKLYN                       0        UNKNOWN
##   PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1        U   UNKNOWN         25-44       F BLACK HISPANIC
## 2        U   UNKNOWN           65+       M          WHITE
## 3        U   UNKNOWN         18-24       F          BLACK
## 4        U   UNKNOWN           <18       M          BLACK
## 5        U   UNKNOWN         18-24       M          BLACK
## 6        U   UNKNOWN           <18       M          BLACK
```

Step 3: Visulizing Data

We will now plot the amount of shootings by all the boroughs of NYC.

```
ggplot(df_new, aes(BORO))+geom_bar()+labs(title="Incidents by Boroughs", x="Boroughs",y="Numbers of Inc
```




```
table(df_new$BORO)
```

```
##
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##          6701          9734          2922          3532           696
```

For this dataset, let's find out the following:

1. **Which Borough have the highest shootings?**

1. **What race is more likely to be shooter and which race is more likely to be victim?**

2. **What age group is more likely to be part of these shootings?**

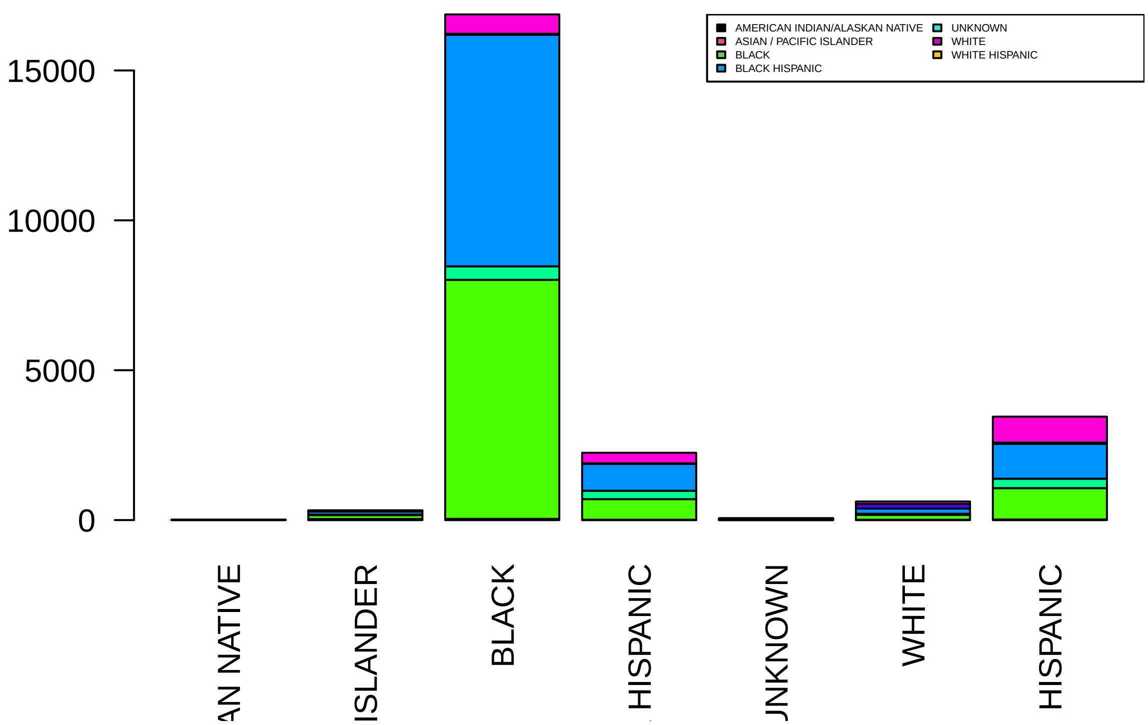
3. **Which sex is more likely to be part of these shootings?**

```r
table(df_new$BORO, df_new$STATISTICAL_MURDER_FLAG)
```

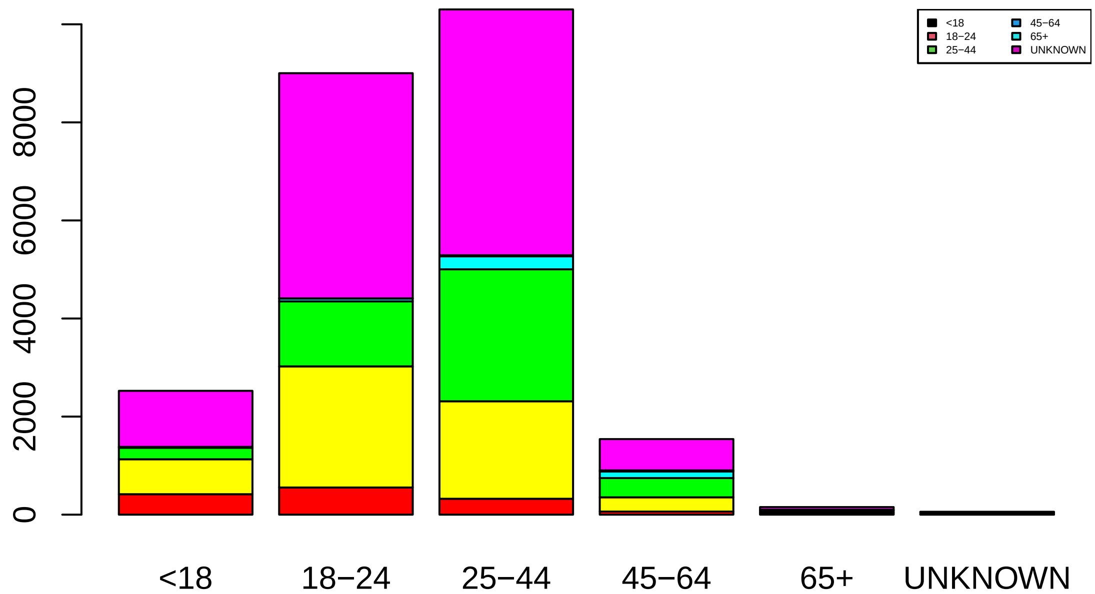
We can see that Brooklyn has the most amount of shooting incidents followed by Bronx. Based on this graph, let's find out how many of these shootings are results in murder.

```
##
##                    0    1
##   BRONX         5454 1247
##   BROOKLYN      7836 1898
##   MANHATTAN     2407  515
##   QUEENS        2835  697
##   STATEN ISLAND  553  143
```

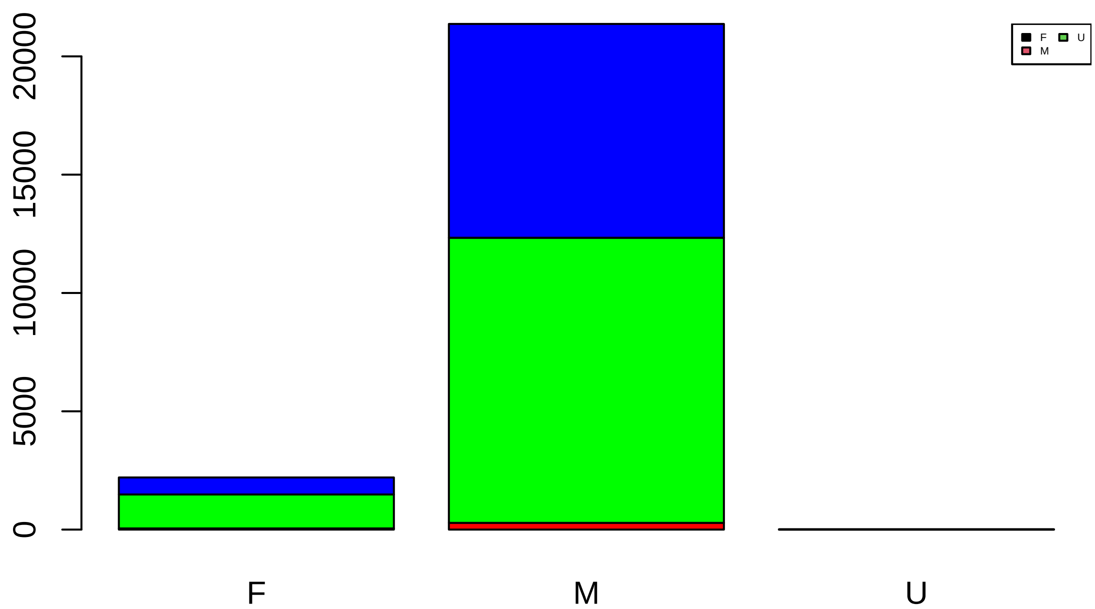
```r
r=table(df_new$PERP_RACE, df_new$VIC_RACE)
barplot(as.matrix(r), col = rainbow(7), las=2)
legend("topright",legend = rownames(r), fill = 1:7, ncol = 2,cex = 0.35)
```




```r
a=table(df_new$PERP_AGE_GROUP, df_new$VIC_AGE_GROUP)
barplot(as.matrix(a), col = rainbow(6))
legend("topright",legend = rownames(a), fill = 1:6, ncol = 2,cex = 0.35)
```




```
s=table(df_new$PERP_SEX, df_new$VIC_SEX)
barplot(as.matrix(s), col = rainbow(3))
legend("topright",legend = rownames(s), fill = 1:3, ncol = 2,cex = 0.35)
```



We can see from the graphs that black men tend to be perpetrators and black/black hispanic men tend to be victims.

Step 4: Modeling Data

```
df_new$STATISTICAL_MURDER_FLAG<-as.numeric(df_new$STATISTICAL_MURDER_FLAG)
mod <- glm(STATISTICAL_MURDER_FLAG~OCCUR_DATE+PERP_AGE_GROUP+PERP_SEX+PERP_RACE, data=df_new, family="bi
summary(mod)
```

We will now build a model using logistic regression to predict if the incident will result in murder.

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ OCCUR_DATE + PERP_AGE_GROUP +
##     PERP_SEX + PERP_RACE, family = "binomial", data = df_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8992  -0.6757  -0.6149  -0.2176   2.9081
##
```

```
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -10.832962  84.415755  -0.128 0.897889
## OCCUR_DATE.L                     -0.055538   0.041519  -1.338 0.181003
## OCCUR_DATE.Q                     -0.099699   0.043684  -2.282 0.022472 *
## OCCUR_DATE.C                     -0.048915   0.044911  -1.089 0.276090
## OCCUR_DATE^4                     -0.039325   0.045668  -0.861 0.389172
## OCCUR_DATE^5                     -0.005101   0.048043  -0.106 0.915448
## OCCUR_DATE^6                     -0.040959   0.049663  -0.825 0.409520
## PERP_AGE_GROUP18-24               0.162911   0.078116   2.086 0.037023 *
## PERP_AGE_GROUP25-44               0.503244   0.077999   6.452 1.10e-10 ***
## PERP_AGE_GROUP45-64               0.827764   0.119213   6.944 3.82e-12 ***
## PERP_AGE_GROUP65+                 1.034956   0.290415   3.564 0.000366 ***
## PERP_AGE_GROUPUNKNOWN            -2.223419   0.172367 -12.899  < 2e-16 ***
## PERP_SEXM                        -0.150656   0.129302  -1.165 0.243959
## PERP_SEXU                         2.480248   0.271977   9.119  < 2e-16 ***
## PERP_RACEASIAN / PACIFIC ISLANDER   9.944287  84.415952   0.118 0.906225
## PERP_RACEBLACK                    9.443813  84.415716   0.112 0.910924
## PERP_RACEBLACK HISPANIC           9.303813  84.415749   0.110 0.912240
## PERP_RACEUNKNOWN                  8.977274  84.415968   0.106 0.915308
## PERP_RACEWHITE                   10.119528  84.415815   0.120 0.904580
## PERP_RACEWHITE HISPANIC           9.589109  84.415728   0.114 0.909560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22990  on 23584  degrees of freedom
## Residual deviance: 22127  on 23565  degrees of freedom
## AIC: 22167
##
## Number of Fisher Scoring iterations: 9
```

#Biases and Pitfalls

Biases and pitfalls would be assuming a certain borough would have more crime than others, or certain race/age group would commit more crimes than other race/age groups. Such as one would assume Brox will have more crimes and shootings due to media exposure, but Brooklyn actually has more shootings.

So maybe have an open mind and neutral mindset before starting any analysis.