NY Shooting Crimes

2022-11-04

A study of shooting crimes in New york

This government data shows shooting incidents in New York City from 2006 through 2021 with a number of classifiers, including both victim and perpetrator demographics, shooting location and whether it was a fatality.

```
file = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
mydata <- read.csv(file)
```

```
summary(mydata)
```

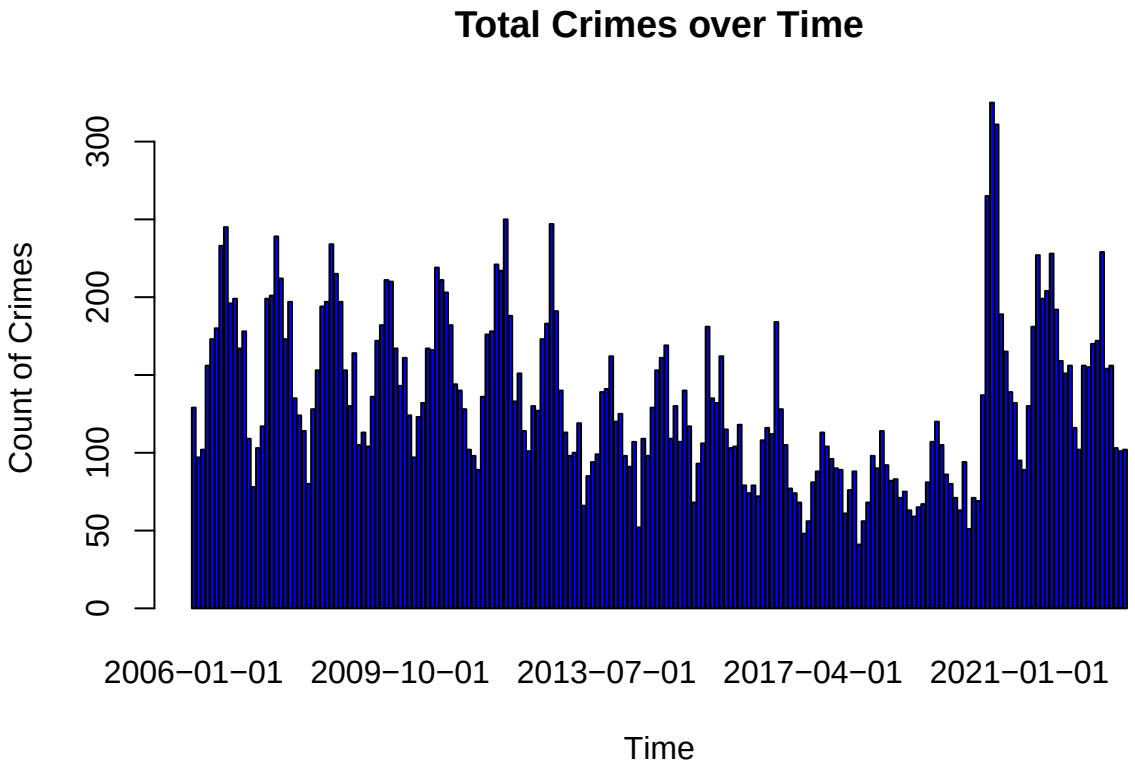
```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character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

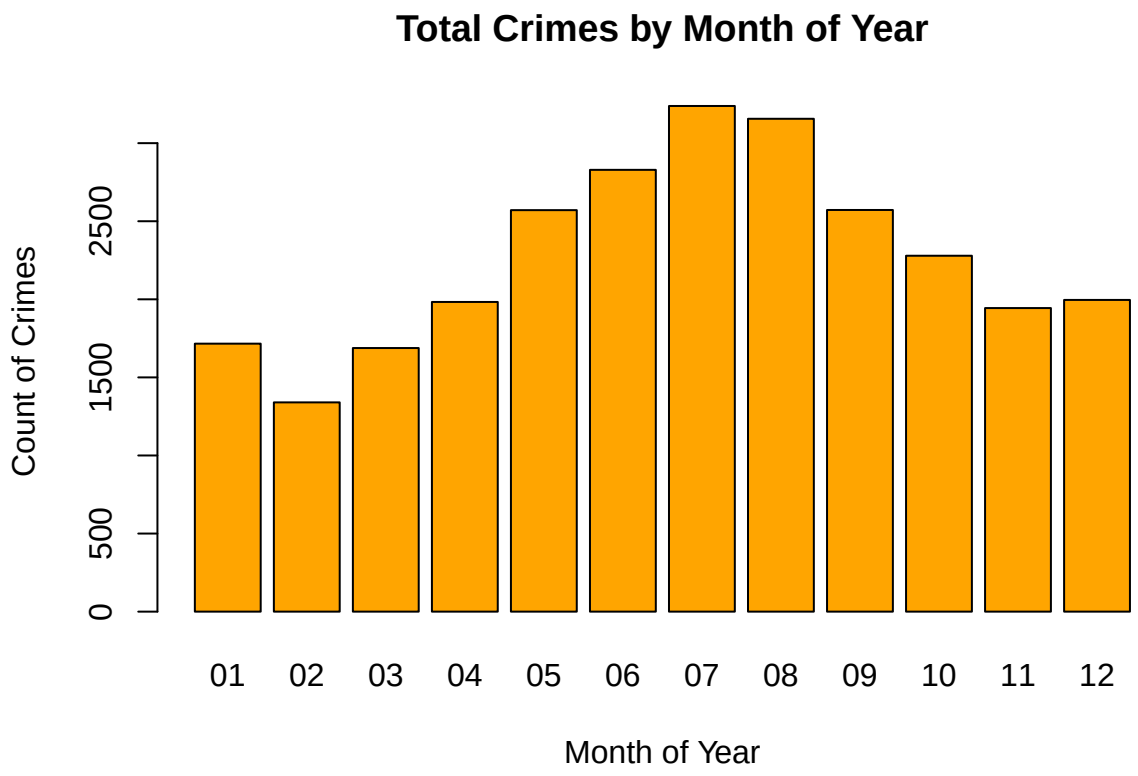
Here we wrangle the data

```
mydata$OCCUR_DATE <- as.Date(mydata$OCCUR_DATE, format="%m/%d/%Y")
mydata$STATISTICAL_MURDER_FLAG <- as.logical(mydata$STATISTICAL_MURDER_FLAG)
mydata$PERP_SEX <- factor(mydata$PERP_SEX)
mydata$VIC_SEX <- factor(mydata$VIC_SEX)
mydata$PERP_AGE_GROUP <- factor(mydata$PERP_AGE_GROUP)
mydata$VIC_AGE_GROUP <- factor(mydata$VIC_AGE_GROUP)
mydata$PERP_RACE <- factor(mydata$PERP_RACE)
mydata$VIC_RACE <- factor(mydata$VIC_RACE)
df_grp_date = mydata %>% group_by(month = lubridate::floor_date(OCCUR_DATE, 'month')) %>%
  summarise(total_crime = n(),
            .groups = 'drop')

mydata$PERP_AGE_GROUP[mydata$PERP_AGE_GROUP == "1020"] <- "UNKNOWN"
mydata$PERP_AGE_GROUP[mydata$PERP_AGE_GROUP == "940"] <- "UNKNOWN"
mydata$PERP_AGE_GROUP[mydata$PERP_AGE_GROUP == "224"] <- "UNKNOWN"
mydata$PERP_AGE_GROUP[mydata$PERP_AGE_GROUP == ""] <- "UNKNOWN"

mydata$PERP_AGE_GROUP <- factor(mydata$PERP_AGE_GROUP)

mydata$PERP_SEX[mydata$PERP_SEX == ""] <- "U"

mydata$PERP_SEX <- factor(mydata$PERP_SEX)

mydata$PERP_RACE[mydata$PERP_RACE == ""] <- "UNKNOWN"
mydata$PERP_RACE <- factor(mydata$PERP_RACE)

mydata$VIC_RACE[mydata$VIC_RACE == ""] <- "UNKNOWN"
mydata$VIC_RACE <- factor(mydata$VIC_RACE)
```

Crime had been steadily declining until 2020 when see see a significant spike. This is likely the result of the aftermath of Covid.



We see that crimes are more frequent in the summer months

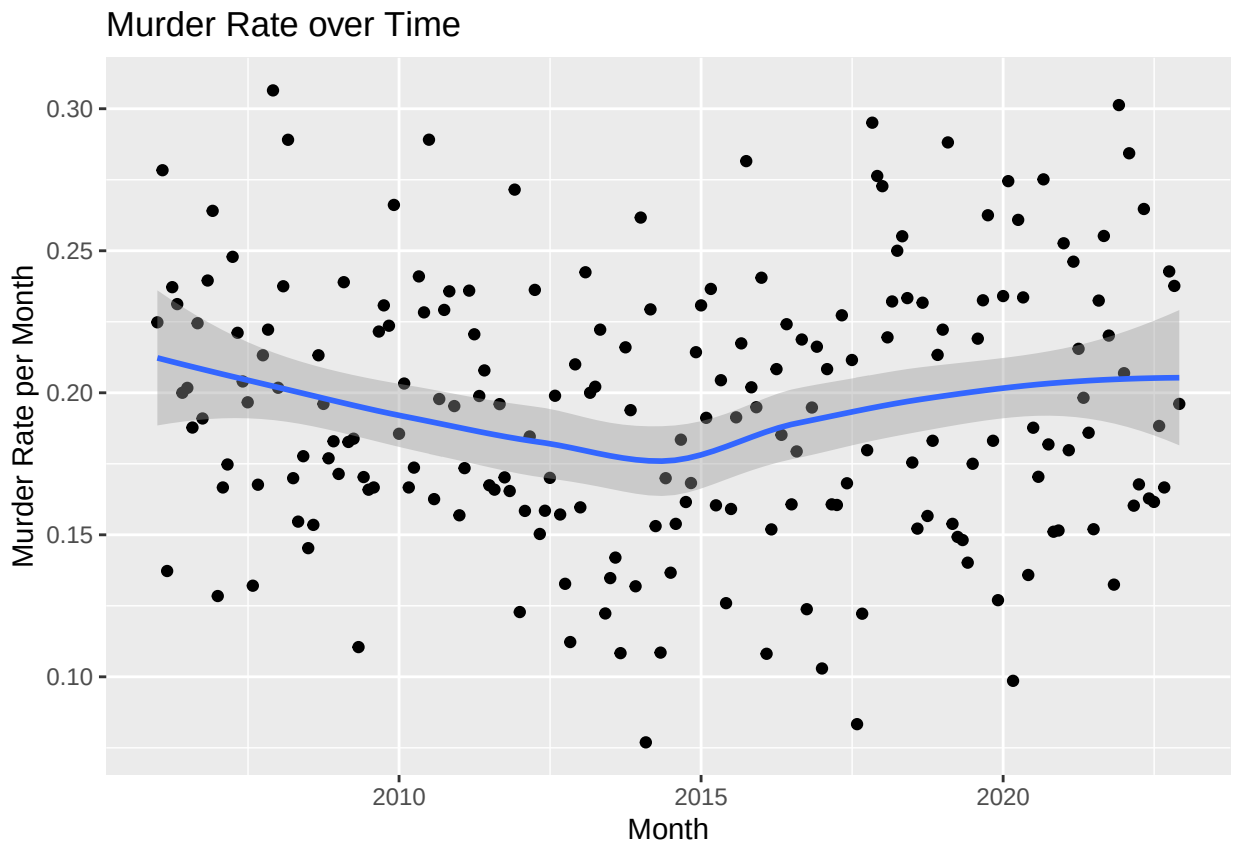


Murders had been declining through 2015, although we see a steady rise since then.

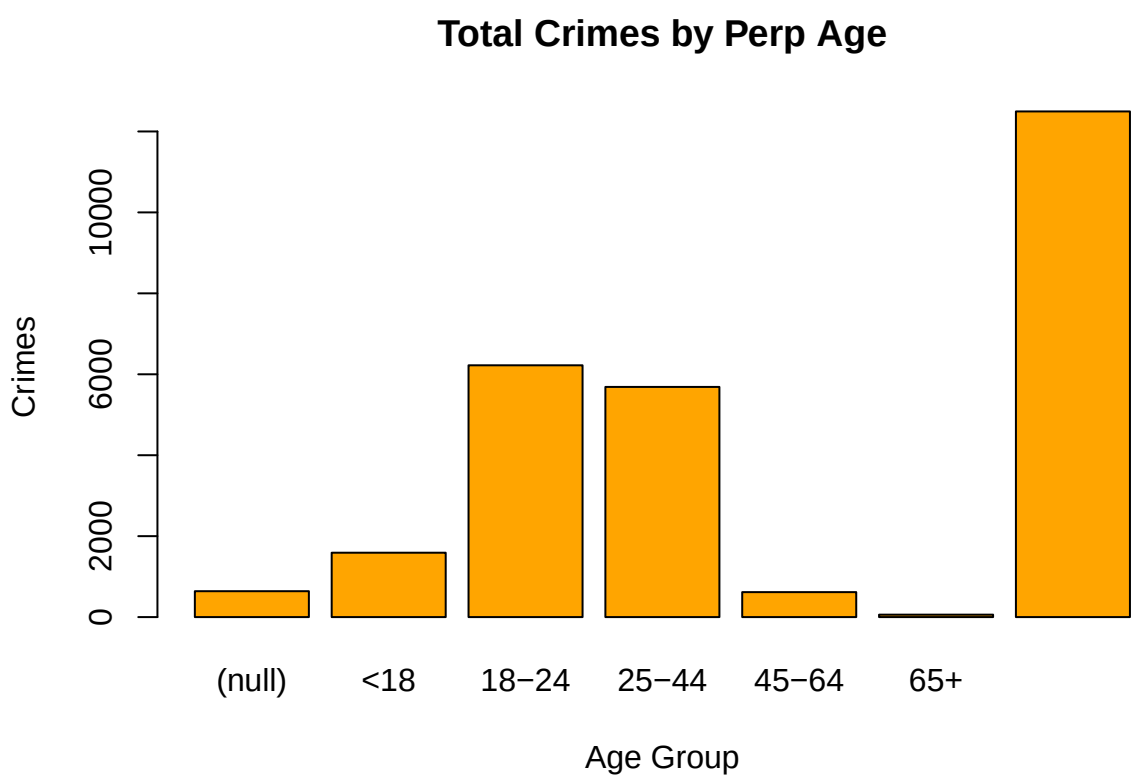
```
# Core wrapping function
wrap.it <- function(x, len)
{
  sapply(x, function(y) paste(strwrap(y, len),
                              collapse = "\n"),
         USE.NAMES = FALSE)
}


# Call this function with a list or vector
wrap.labels <- function(x, len)
{
  if (is.list(x))
  {
    lapply(x, wrap.it, len)
  } else {
    wrap.it(x, len)
  }
}
```

```
ggplot(data=df_grp_date, aes(x = month, y = Murder_Rate)) + geom_point() +
  geom_smooth() + ggtitle("Murder Rate over Time") + xlab("Month") + ylab("Murder Rate per Month")
```

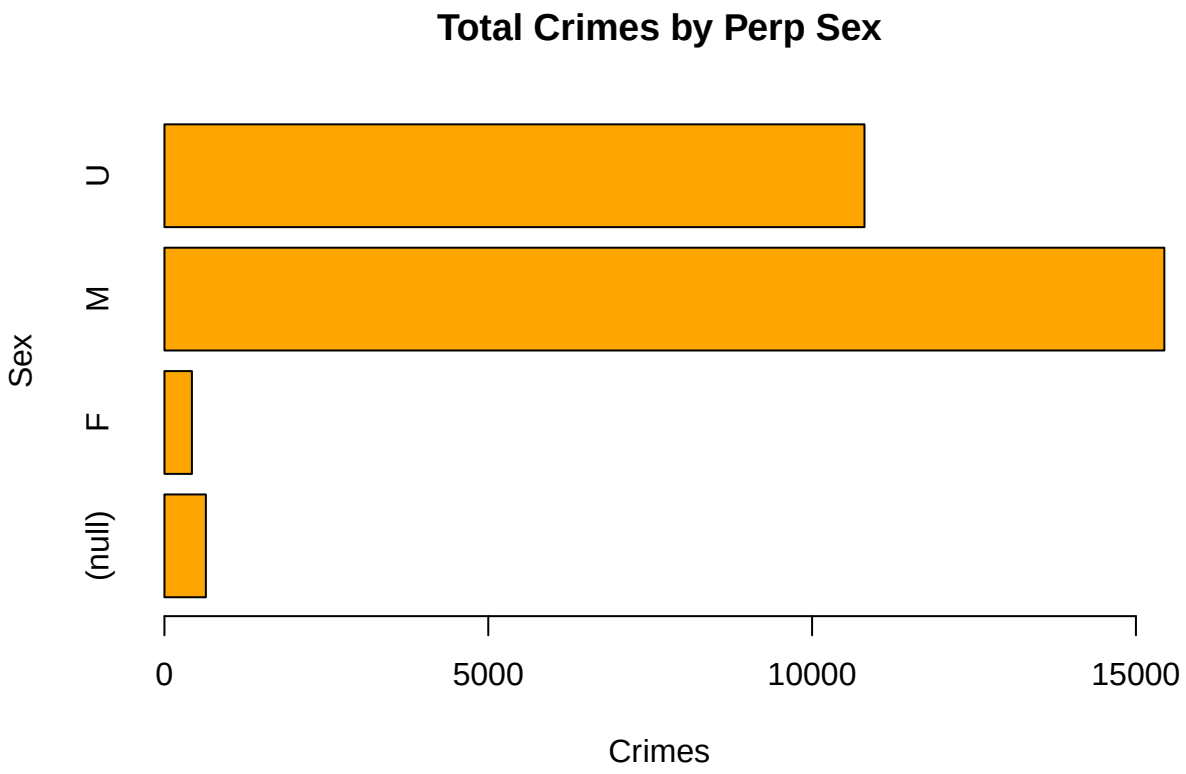
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



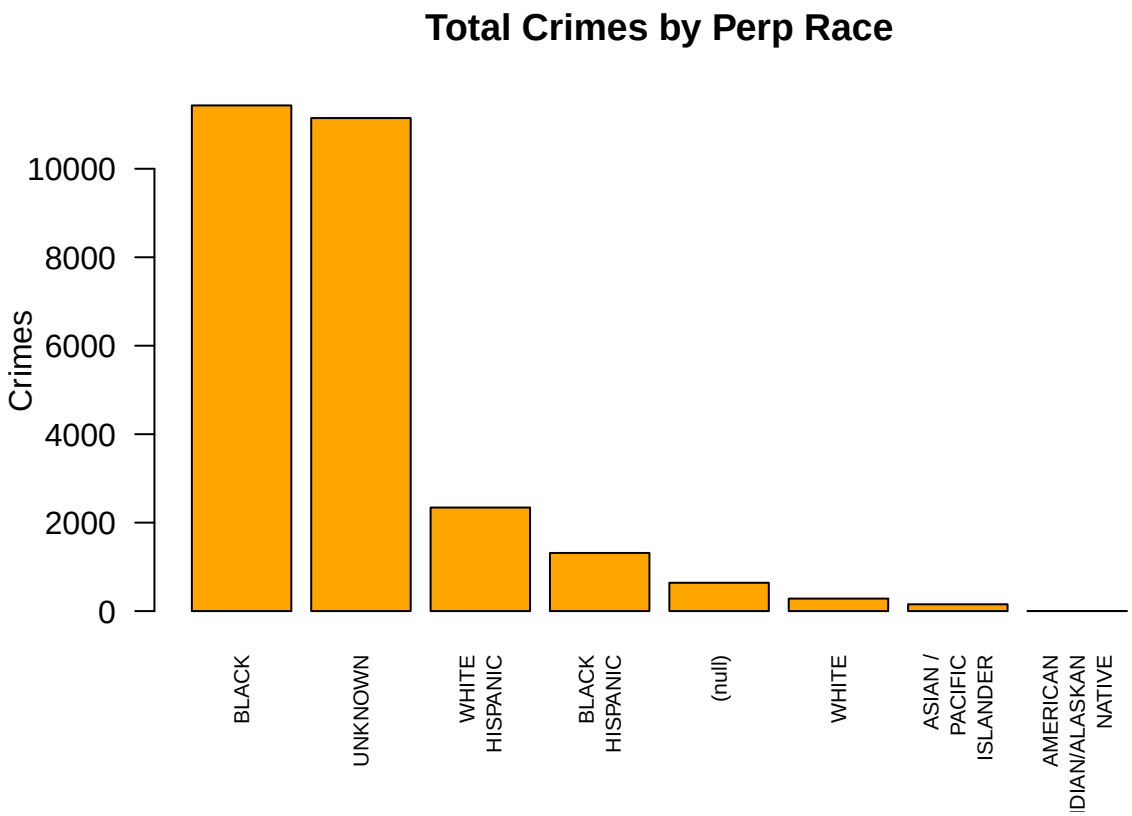

```
crime_smry_perp_age <- table(mydata$PERP_AGE_GROUP)
crime_smry_perp_sex <- table(mydata$PERP_SEX)
crime_smry_perp_race <- table(mydata$PERP_RACE)
crime_smry_vic_age <- table(mydata$VIC_AGE_GROUP)
crime_smry_vic_sex <- table(mydata$VIC_SEX)
crime_smry_vic_race <- table(mydata$VIC_RACE)

crime_smry_perp_race <- crime_smry_perp_race[order(crime_smry_perp_race,decreasing=TRUE)]
crime_smry_vic_race <- crime_smry_vic_race[order(crime_smry_vic_race,decreasing=TRUE)]

crime_smry_perp_race2 <- wrap.labels(names(crime_smry_perp_race), 10)
crime_smry_vic_race2 <- wrap.labels(names(crime_smry_vic_race), 10)


barplot(crime_smry_perp_age,
        main = "Total Crimes by Perp Age",
        xlab = "Age Group",
        ylab = "Crimes",
        col = "Orange",
        horiz = FALSE)
```



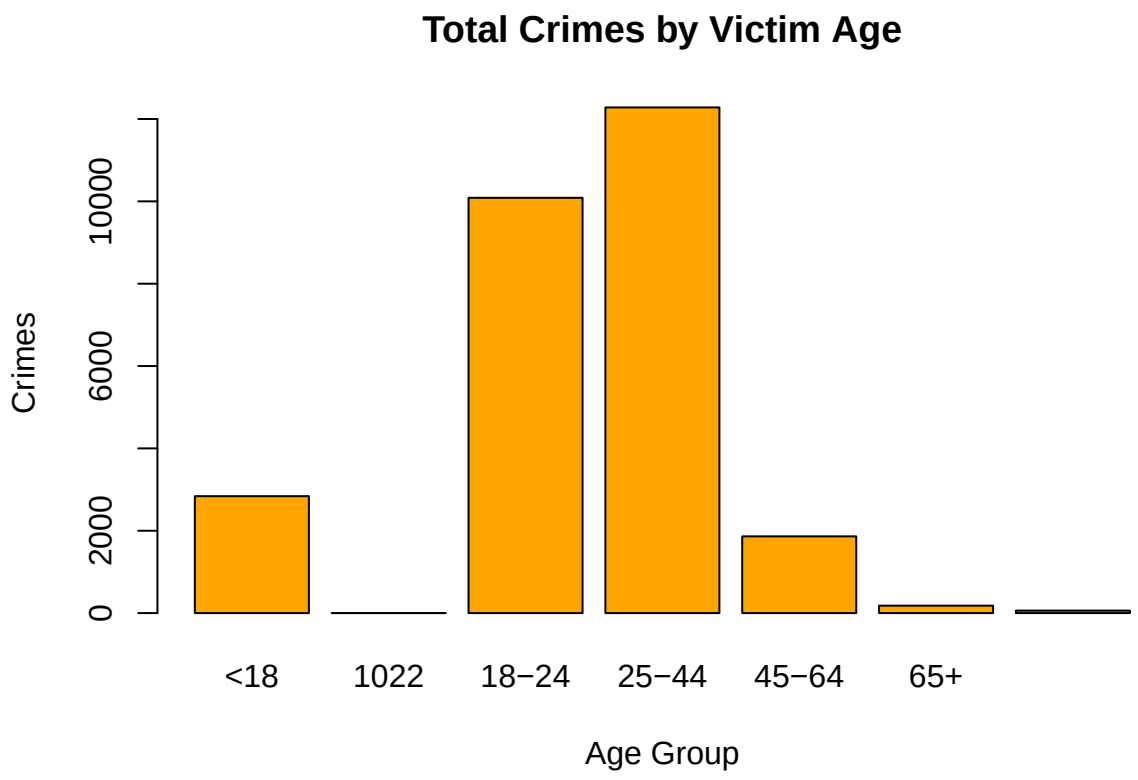

```
barplot(crime_smry_perp_sex,
        main = "Total Crimes by Perp Sex",
        xlab = "Crimes",
        ylab = "Sex",
        col = "Orange",
        horiz = TRUE)
```



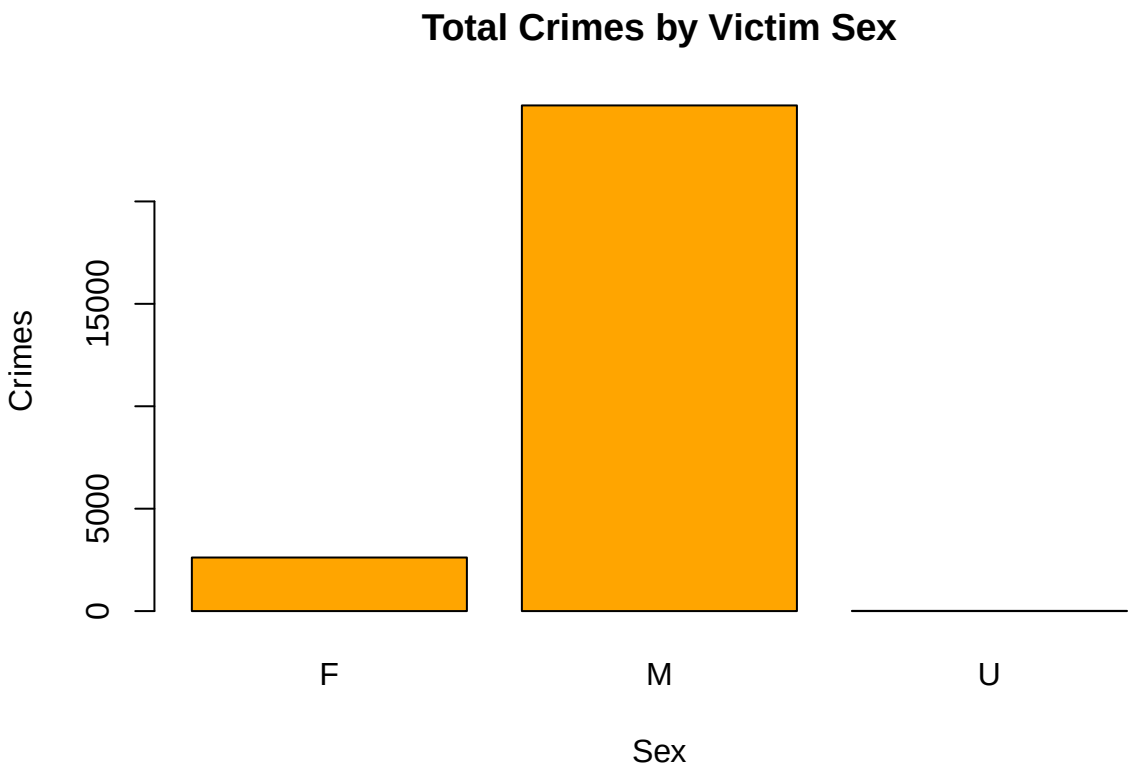

```
barplot( crime_smry_perp_race ,
        main = "Total Crimes by Perp Race",
        ylab = "Crimes",
        col = "Orange",
        names.arg = crime_smry_perp_race2,
        cex.names=0.7,
        horiz = FALSE,
        las = 2)
```



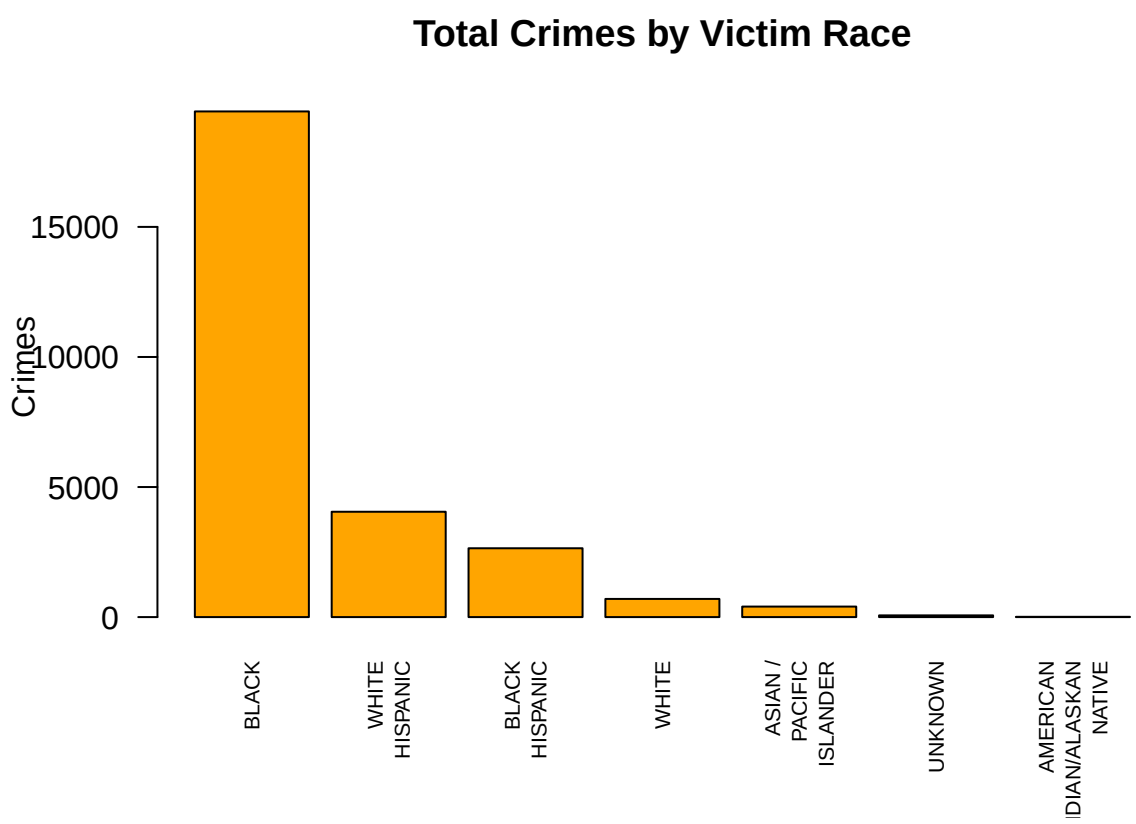

```
barplot(crime_smry_vic_age,
        main = "Total Crimes by Victim Age",
        xlab = "Age Group",
        ylab = "Crimes",
        col = "Orange",
        horiz = FALSE)
```




```
barplot(crime_smry_vic_sex,
        main = "Total Crimes by Victim Sex",
        xlab = "Sex",
        ylab = "Crimes",
        col = "Orange",
        horiz = FALSE)
```




```
barplot(crime_smry_vic_race,
        main = "Total Crimes by Victim Race",
        ylab = "Crimes",
        col = "Orange",
        names.arg = crime_smry_vic_race2,
        cex.names=0.7,
        horiz = FALSE,
        las=2)
```



A simple regression model using various demographic criteria to predict if a crime will be a murder

```
##
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ PERP_SEX + PERP_AGE_GROUP +
##     VIC_AGE_GROUP + PERP_RACE, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55087 -0.21343 -0.17978 -0.06173  1.07557
##
## Coefficients: (2 not defined because of singularities)
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.091133   0.016929   5.383 7.37e-08
## PERP_SEXF                           -0.073742   0.027262  -2.705  0.00684
## PERP_SEXM                           -0.097461   0.019849  -4.910 9.16e-07
## PERP_SEXU                            0.091871   0.027340   3.360  0.00078
## PERP_AGE_GROUP<18                    0.168425   0.013808  12.197  < 2e-16
## PERP_AGE_GROUP18-24                  0.186110   0.010878  17.108  < 2e-16
## PERP_AGE_GROUP25-44                  0.229025   0.011049  20.727  < 2e-16
## PERP_AGE_GROUP45-64                  0.295514   0.018682  15.818  < 2e-16
## PERP_AGE_GROUP65+                    0.318403   0.051776   6.150 7.87e-10
## PERP_AGE_GROUPUNKNOWN                      NA         NA      NA       NA
## VIC_AGE_GROUP1022                   -0.157570   0.388043  -0.406  0.68470
## VIC_AGE_GROUP18-24                   0.033645   0.008322   4.043 5.30e-05
```

```
## VIC_AGE_GROUP25-44                        0.072730   0.008279   8.785  < 2e-16
## VIC_AGE_GROUP45-64                        0.090270   0.011800   7.650 2.07e-14
## VIC_AGE_GROUP65+                          0.142340   0.029974   4.749 2.06e-06
## VIC_AGE_GROUPUNKNOWN                      0.056267   0.050321   1.118  0.26351
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -0.294599   0.274613  -1.073  0.28338
## PERP_RACEASIAN / PACIFIC ISLANDER         0.060797   0.032294   1.883  0.05976
## PERP_RACEBLACK                           -0.022211   0.008825  -2.517  0.01185
## PERP_RACEBLACK HISPANIC                  -0.038120   0.013382  -2.849  0.00439
## PERP_RACEUNKNOWN                         -0.069239   0.022342  -3.099  0.00194
## PERP_RACEWHITE                            0.096453   0.024785   3.892 9.98e-05
## PERP_RACEWHITE HISPANIC                         NA         NA      NA       NA
##
## (Intercept)                              ***
## PERP_SEXF                                **
## PERP_SEXM                                ***
## PERP_SEXU                                ***
## PERP_AGE_GROUP<18                        ***
## PERP_AGE_GROUP18-24                      ***
## PERP_AGE_GROUP25-44                      ***
## PERP_AGE_GROUP45-64                      ***
## PERP_AGE_GROUP65+                        ***
## PERP_AGE_GROUPUNKNOWN
## VIC_AGE_GROUP1022
## VIC_AGE_GROUP18-24                       ***
## VIC_AGE_GROUP25-44                       ***
## VIC_AGE_GROUP45-64                       ***
## VIC_AGE_GROUP65+                         ***
## VIC_AGE_GROUPUNKNOWN
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER        .
## PERP_RACEBLACK                           *
## PERP_RACEBLACK HISPANIC                  **
## PERP_RACEUNKNOWN                         **
## PERP_RACEWHITE                           ***
## PERP_RACEWHITE HISPANIC
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3879 on 27291 degrees of freedom
## Multiple R-squared:  0.03371,    Adjusted R-squared:  0.03301
## F-statistic: 47.61 on 20 and 27291 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ PERP_AGE_GROUP25_44 +
##     PERP_AGE_GROUP45_64 + PERP_AGE_GROUP65plus + PERP_AGE_GROUP_UNK +
##     VIC_AGE_GROUP18_24 + VIC_AGE_GROUP25_44 + VIC_AGE_GROUP45_64 +
##     VIC_AGE_GROUP65plus + PERP_SEX_U, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45536 -0.21981 -0.18261 -0.06608  1.03002
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.149940   0.009274  16.168  < 2e-16 ***
## PERP_AGE_GROUP25_44   0.059635   0.008108   7.355 1.98e-13 ***
## PERP_AGE_GROUP45_64   0.133540   0.019217   6.949 3.80e-12 ***
## PERP_AGE_GROUP65plus  0.140880   0.061431   2.293 0.021841 *
## PERP_AGE_GROUP_UNK   -0.179962   0.012282 -14.653  < 2e-16 ***
## VIC_AGE_GROUP18_24    0.036055   0.009834   3.666 0.000247 ***
## VIC_AGE_GROUP25_44    0.069868   0.009743   7.171 7.72e-13 ***
## VIC_AGE_GROUP45_64    0.096105   0.013910   6.909 5.04e-12 ***
## VIC_AGE_GROUP65plus   0.164543   0.035304   4.661 3.17e-06 ***
## PERP_SEX_U            0.142768   0.012041  11.857  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3883 on 19134 degrees of freedom
## Multiple R-squared:  0.03122,    Adjusted R-squared:  0.03077
## F-statistic: 68.51 on 9 and 19134 DF,  p-value: < 2.2e-16
```

```
## [1] 0.4553625
```

There were 3,467 murders out of 17,913 crimes, a rate of 19.3%. The model score has a maximum of 44.6%, and so the accuracy is dependent on the cut-off. With a cutoff of 30%, The model accurately predicted 79.8% of records, and captures only 145 of the 3,467 murders, or 4.2%, with a false positive rate of 33.3%. This cutoff can be reduced to capture more true positives, but the false positive rate would also increase. Overall, there is minimal information value in this data to accurately predict the probability of a crime being a murder.

```
confusionMatrix(as.factor(test_pred$default), as.factor(test_pred$pred2))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 14833   622
##          1  3404   285
##
##                Accuracy : 0.7897
##                  95% CI : (0.7839, 0.7955)
##     No Information Rate : 0.9526
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0519
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.81335
##             Specificity : 0.31422
##          Pos Pred Value : 0.95975
##          Neg Pred Value : 0.07726
##              Prevalence : 0.95262
##          Detection Rate : 0.77481
##    Detection Prevalence : 0.80730
##       Balanced Accuracy : 0.56378
##
```

```
##        'Positive' Class : 0
##
```