Assignment 2

B. Nunez

2022-11-03

R Markdown

NOTE that tidyverse and e1071 packages must be installed.

Collecting the data:

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

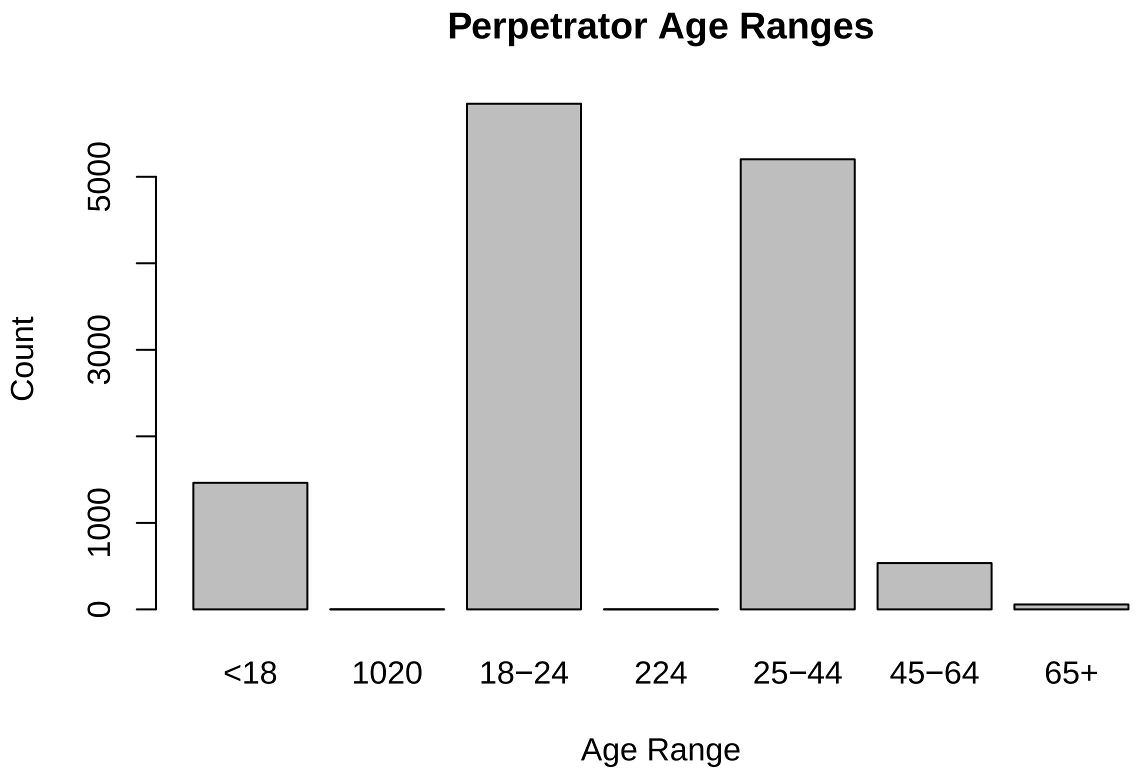
Cleaning up and formatting the data: remove rows where the age and race values are empty, then summarize the data.

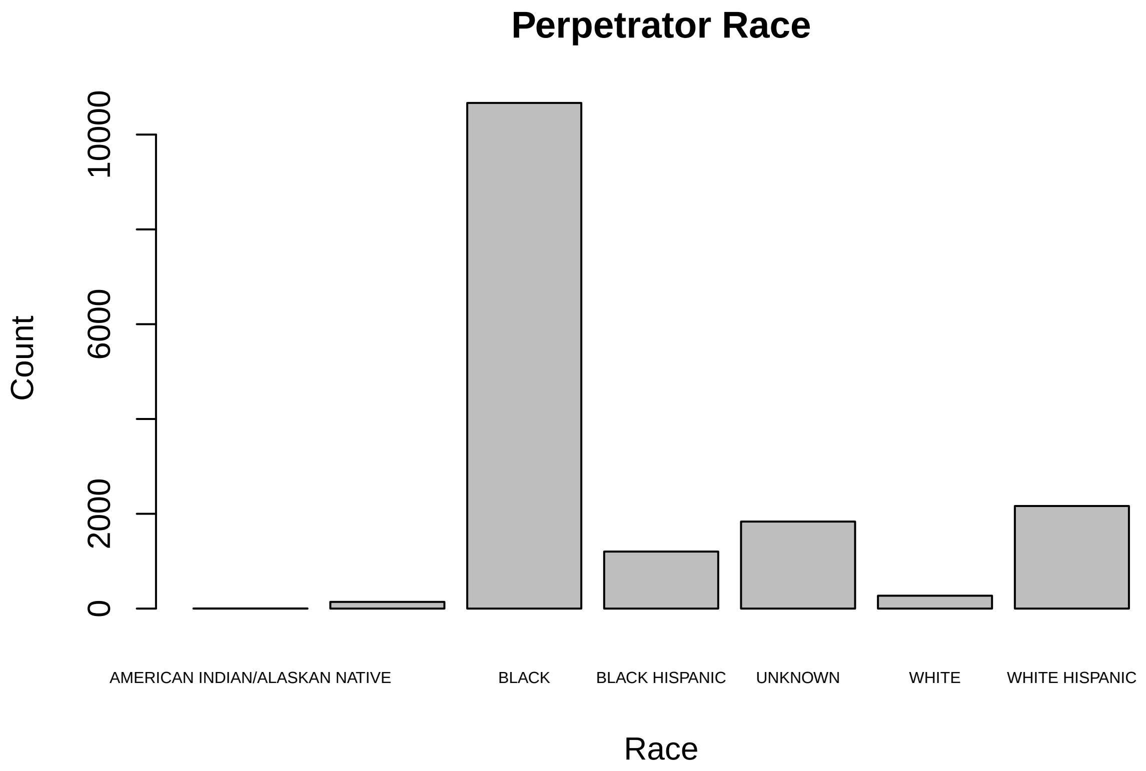
```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA’s   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##    VIC_SEX            VIC_RACE            X_COORD_CD         Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                         Mean   :1009455   Mean   :207894
##                                         3rd Qu.:1016838   3rd Qu.:239429
##                                         Max.   :1066815   Max.   :271128
##
##     Latitude        Longitude         Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Including Plots

Visualizations:





Note that erroneous or unknown value data points were removed in the visualizations. We can see that the perpetrator in the crime is disproportionately inclined to one race and age.

We create a model to predict victim race based on perpetrator race:

```
##
## pred                             AMERICAN INDIAN/ALASKAN NATIVE
##   AMERICAN INDIAN/ALASKAN NATIVE                              0
##   ASIAN / PACIFIC ISLANDER                                    0
##   BLACK                                                       2
##   BLACK HISPANIC                                              0
##   UNKNOWN                                                     0
##   WHITE                                                       0
##   WHITE HISPANIC                                              0
##
## pred                             ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     0              0
##   ASIAN / PACIFIC ISLANDER                              0     0              0
##   BLACK                                                72  3642            515
##   BLACK HISPANIC                                        0     0              0
##   UNKNOWN                                               0     0              0
##   WHITE                                                 0     0              0
##   WHITE HISPANIC                                        0     0              0
##
## pred                             UNKNOWN WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0     0              0
##   BLACK                                   14   129             745
##   BLACK HISPANIC                           0     0               0
##   UNKNOWN                                  0     0               0
##   WHITE                                    0     0               0
##   WHITE HISPANIC                           0     0               0
```

```
## [1] 0.7114671
```

This shows that the model predicts the victim race based on the perpetrator race with about 74% accuracy.

Sources of bias: Possibly the significant omission of unknown and erroneous data points that may have changed the model in some way, were their true values to be known. Also, there too many black perpetrator / victim data points, so that the model is biased in such a way that it is difficult to tell if there is a statistically significant relationship between races in general, versus only black perpetrator / victim data points.