W5_NYPD analysis

(Assignment)

2023/11/11

0) load the library

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
```

1) Read the CSV file (NYDP Shooting Incident)

```
NYPD <- read.csv("https://data.cityofnewyork.us/api/views/5ucz-vwe8/rows.csv?accessType=DOWNLOAD")
```

2) Overview the data

```
summary(NYPD)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :261194183   Length:991         Length:991         Length:991
##  1st Qu.:265297120   Class :character   Class :character   Class :character
##  Median :268973603   Mode  :character   Mode  :character   Mode  :character
##  Mean   :268591547
##  3rd Qu.:271818282
##  Max.   :275218028
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:991         Min.   :  5.00   Min.   :0.0000    Length:991
##  Class :character   1st Qu.: 43.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 49.00   Median :0.0000    Mode  :character
##                     Mean   : 61.55   Mean   :0.2119
##                     3rd Qu.: 78.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:991         Length:991              Length:991
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:991         Length:991         Length:991         Length:991
##  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:991         Min.   : 929510   Min.   :127539   Min.   :40.52
##  Class :character   1st Qu.:1000930   1st Qu.:185903   1st Qu.:40.68
##  Mode  :character   Median :1008697   Median :221195   Median :40.77
##                     Mean   :1008871   Mean   :215468   Mean   :40.76
##                     3rd Qu.:1016100   3rd Qu.:244385   3rd Qu.:40.84
##                     Max.   :1057854   Max.   :268868   Max.   :40.90
##                                                        NA’s   :41
##    Longitude      New.Georeferenced.Column
##  Min.   :-74.20   Length:991
##  1st Qu.:-73.94   Class :character
##  Median :-73.91   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.73
##  NA’s   :41
```

```
head(NYPD)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME      BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    265303128 03/18/2023   03:45:00    QUEENS           OUTSIDE      102
## 2    264075661 02/22/2023   16:55:00     BRONX           OUTSIDE       44
## 3    270760379 07/03/2023   21:25:00  BROOKLYN           OUTSIDE       75
## 4    265124475 03/14/2023   09:49:00 MANHATTAN           OUTSIDE       20
## 5    266761946 04/15/2023   15:46:00 MANHATTAN            INSIDE       32
## 6    273520496 08/25/2023   22:45:00     BRONX           OUTSIDE       46
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC           LOCATION_DESC
## 1                 0              OTHER                HOSPITAL
## 2                 0             STREET                  (null)
## 3                 0             STREET                  (null)
## 4                 0             STREET         COMMERCIAL BLDG
## 5                 0           DWELLING MULTI DWELL - APT BUILD
## 6                 0             STREET                   (null)
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX      PERP_RACE VIC_AGE_GROUP
## 1                       N          25-44        M          BLACK         25-44
## 2                       N          25-44        M WHITE HISPANIC         25-44
## 3                       N          18-24        M          BLACK         25-44
## 4                       N          18-24        M          BLACK           <18
## 5                       N         (null)   (null)         (null)         25-44
## 6                       Y          25-44        M BLACK HISPANIC         25-44
##   VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1       M          BLACK    1030953     194101       NA        NA
## 2       M WHITE HISPANIC    1004343     243407 40.83475 -73.92739
## 3       M          BLACK    1008769     177614 40.65415 -73.91163
## 4       M          BLACK     988755     221899 40.77574 -73.98373
## 5       M WHITE HISPANIC    1000980     239318 40.82353 -73.93955
## 6       M BLACK HISPANIC    1009333     247239 40.84525 -73.90934
##                      New.Georeferenced.Column
## 1
## 2                 POINT (-73.927388 40.834751)
## 3                 POINT (-73.911632 40.654153)
## 4                 POINT (-73.983734 40.775738)
## 5                 POINT (-73.939551 40.823533)
## 6 POINT (-73.90934182293881 40.84525339546618)
```

```r
str(NYPD)
```

```
## ’data.frame':    991 obs. of  21 variables:
##  $ INCIDENT_KEY            : int  265303128 264075661 270760379 265124475 266761946 273520496 275218
##  $ OCCUR_DATE              : chr  "03/18/2023" "02/22/2023" "07/03/2023" "03/14/2023" ...
##  $ OCCUR_TIME              : chr  "03:45:00" "16:55:00" "21:25:00" "09:49:00" ...
##  $ BORO                    : chr  "QUEENS" "BRONX" "BROOKLYN" "MANHATTAN" ...
##  $ LOC_OF_OCCUR_DESC       : chr  "OUTSIDE" "OUTSIDE" "OUTSIDE" "OUTSIDE" ...
##  $ PRECINCT                : int  102 44 75 20 32 46 79 42 69 46 ...
##  $ JURISDICTION_CODE       : int  0 0 0 0 0 0 2 2 0 0 ...
##  $ LOC_CLASSFCTN_DESC      : chr  "OTHER" "STREET" "STREET" "STREET" ...
##  $ LOCATION_DESC           : chr  "HOSPITAL" "(null)" "(null)" "COMMERCIAL BLDG" ...
##  $ STATISTICAL_MURDER_FLAG : chr  "N" "N" "N" "N" ...
##  $ PERP_AGE_GROUP          : chr  "25-44" "25-44" "18-24" "18-24" ...
##  $ PERP_SEX                : chr  "M" "M" "M" "M" ...
##  $ PERP_RACE               : chr  "BLACK" "WHITE HISPANIC" "BLACK" "BLACK" ...
##  $ VIC_AGE_GROUP           : chr  "25-44" "25-44" "25-44" "<18" ...
##  $ VIC_SEX                 : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE                : chr  "BLACK" "WHITE HISPANIC" "BLACK" "BLACK" ...
##  $ X_COORD_CD              : int  1030953 1004343 1008769 988755 1000980 1009333 1000301 1010158 101
##  $ Y_COORD_CD              : int  194101 243407 177614 221899 239318 247239 192923 242490 175595 249
##  $ Latitude                : num  NA 40.8 40.7 40.8 40.8 ...
##  $ Longitude               : num  NA -73.9 -73.9 -74 -73.9 ...
##  $ New.Georeferenced.Column: chr  "" "POINT (-73.927388 40.834751)" "POINT (-73.911632 40.654153)" "
```

```r
na_columns <- NYPD %>%
  select_if(function(x) any(is.na(x))) %>%
  names()
na_columns # identify which columns contains NA values
```

```
## [1] "Latitude"  "Longitude"
```

Some rows lack spacial data, but this time we are to focus on "age" and "sex" of the victim. So there seems to be no such a big problem even if we the whole data.

We have no specific theory to test at the present… Let's just explore **"VIC_AGE_GROUP"**, *and* **"VIC_SEX"**!

3-A) Plot; univariate

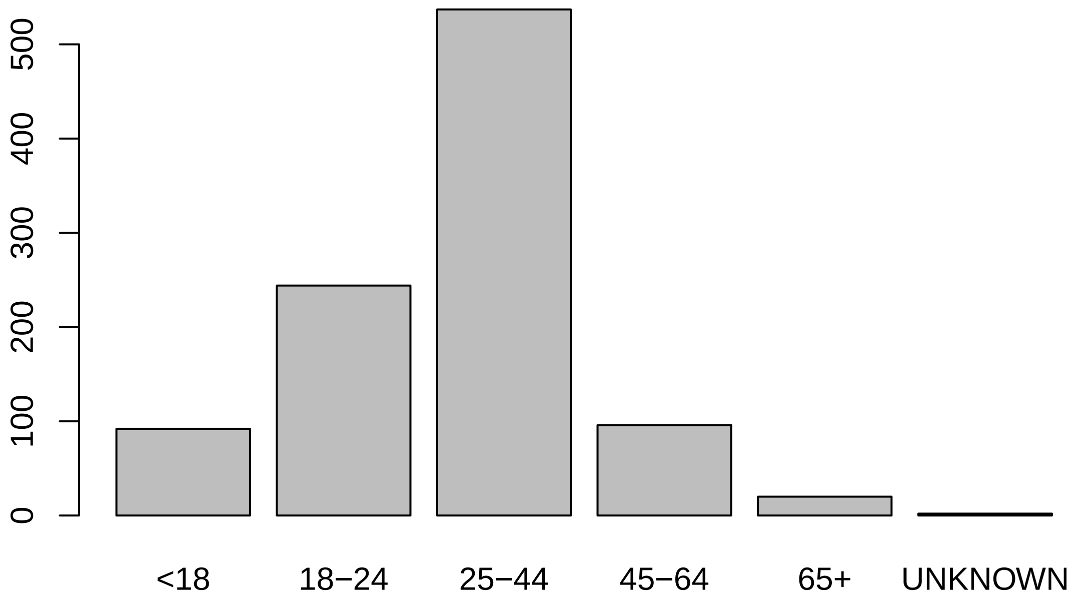
```
NYPD.2v = NYPD[,c("VIC_AGE_GROUP","VIC_SEX")]
NYPD.2v[] <- lapply(NYPD.2v, factor)
str(NYPD.2v)
```

```
## 'data.frame':    991 obs. of  2 variables:
##  $ VIC_AGE_GROUP: Factor w/ 6 levels "<18","18-24",..: 3 3 3 1 3 3 3 3 3 2 ...
##  $ VIC_SEX      : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
```

```
table(NYPD.2v$VIC_AGE_GROUP)
```

```
## 
##     <18   18-24   25-44   45-64     65+ UNKNOWN 
##      92     244     537      96      20       2
```

```
barplot(table(NYPD.2v$VIC_AGE_GROUP))
```

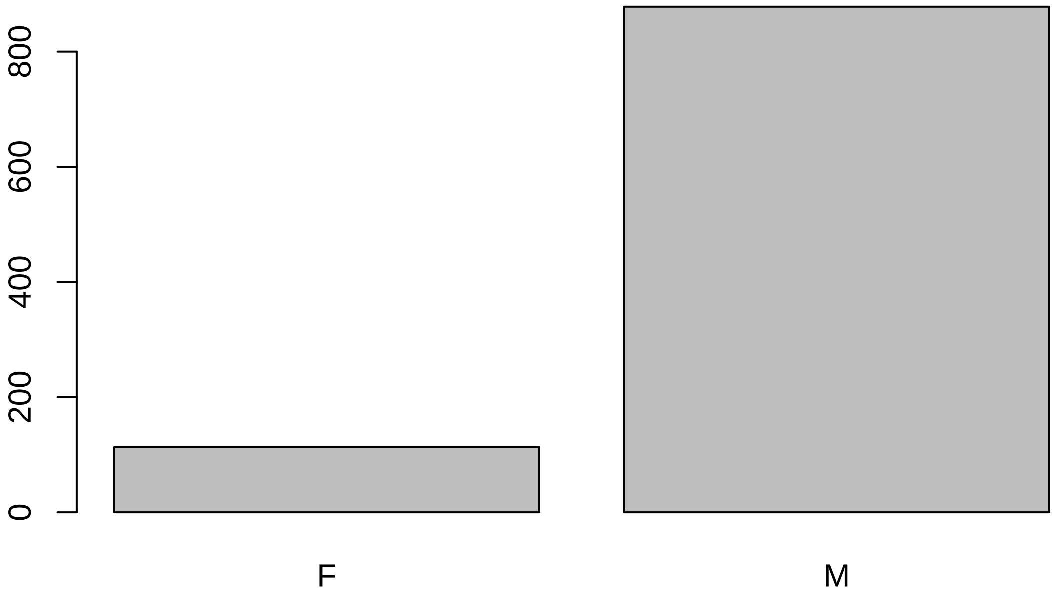


Note that the ranges of each age group are not consistent, which may lead to issues when interpreting the results.

```
table(NYPD.2v$VIC_SEX)
```

```
##
##   F   M
## 113 878
```

```
barplot(table(NYPD.2v$VIC_SEX))
```



The gender proportion among the victims is extremely unbalanced. While this is a fact in itself, it may introduce some sort of bias during interpretation.

3-B) Plot; bivariate

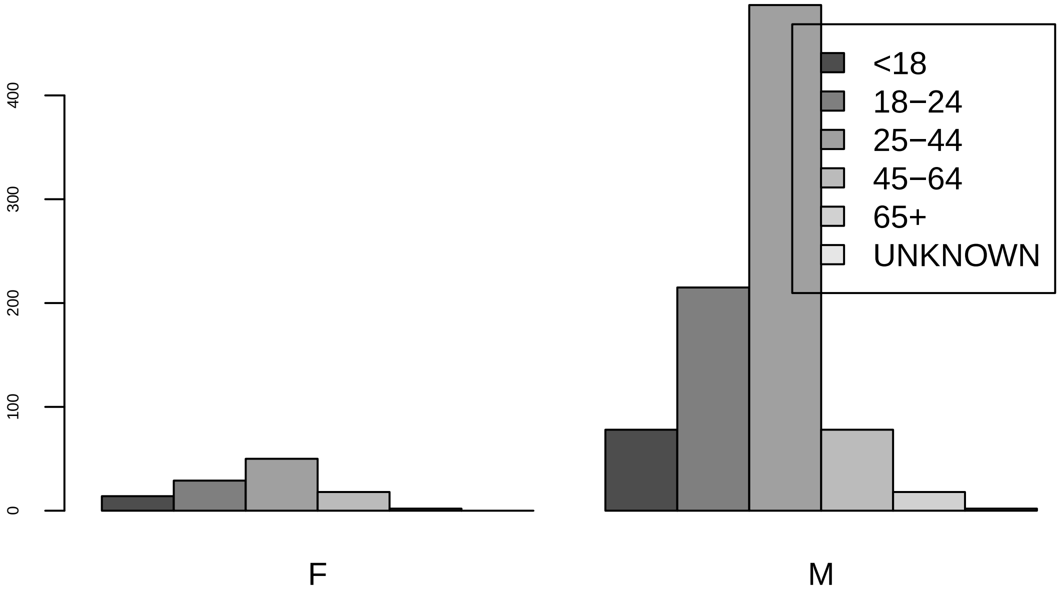
```
#tabulate
NYPD.2v <- as.data.frame(NYPD.2v)

# Create a cross table
NYPD.cross_table <- table(NYPD.2v$VIC_AGE_GROUP,NYPD.2v$VIC_SEX)

# Print the cross table
print(NYPD.cross_table)
```

```
##
##              F   M
##   <18       14  78
##   18-24     29 215
##   25-44     50 487
##   45-64     18  78
##   65+        2  18
##   UNKNOWN    0   2
```

```
barplot(NYPD.cross_table, beside=TRUE, legend = rownames(NYPD.cross_table), cex.axis = 0.5)
```



The difference in distribution doesn't seem apparent. Let's try using regression analysis to get a formal support!

4-A) Formal test

```
age_female <- NYPD.cross_table[1:5,1]#exclude the unknown data
age_male   <- NYPD.cross_table[1:5,2]
chisq.test(age_female, age_male)
```

```
## Warning in chisq.test(age_female, age_male): Chi-squared approximation may be
## incorrect
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  age_female and age_male
## X-squared = 15, df = 12, p-value = 0.2414
```

4-B) Binomial Regression Modeling: Sex~age

```
# convert factor variable into numeric (M=1, F=0)
NYPD.2v <- NYPD.2v %>% mutate(SEX_num = case_when(
  VIC_SEX == "M" ~ 1,
  VIC_SEX == "F" ~ 0))
# View the data frame to confirm changes
#head(NYPD.2v)

# simple linear modeling
mod <- glm(formula = SEX_num~VIC_AGE_GROUP, data=NYPD.2v,
           family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = SEX_num ~ VIC_AGE_GROUP, family = binomial, data = NYPD.2v)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1790   0.4421   0.4421   0.5030   0.6444
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.7177     0.2903   5.918 3.26e-09 ***
## VIC_AGE_GROUP18-24      0.2857     0.3513   0.813   0.4160
## VIC_AGE_GROUP25-44      0.5586     0.3260   1.713   0.0867 .
## VIC_AGE_GROUP45-64     -0.2513     0.3907  -0.643   0.5200
## VIC_AGE_GROUP65+        0.4796     0.7999   0.600   0.5488
## VIC_AGE_GROUPUNKNOWN  12.8484   624.1939   0.021   0.9836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 703.32  on 990  degrees of freedom
## Residual deviance: 694.66  on 985  degrees of freedom
## AIC: 706.66
##
## Number of Fisher Scoring iterations: 13
```

All age factors are statistically **insignificant** in predicting the sex of victims, according to the dataset. So, we cannot predict the victim's sex from their age information.

Discussion on Bias

- This data is from New York. However, in certain regions, age information might be meaningful for making such predictions.
- The data focuses solely on individuals who have either committed crimes or become victims. Ignoring the information about people who are not included in the data could lead to misunderstandings.