Hw1

Lauren Done

9/1/2021

4. PUMS-NY data

Dataset at a Glimpse

```
load("acs2017_ny_data.RData")
#glimpse(acs2017_ny) try this later
acs2017_ny[1:10,1:7]
```

```
##    AGE female educ_nohs educ_hs educ_somecoll educ_college educ_advdeg
## 1   72      1         0       0             0            0           1
## 2   72      0         0       0             0            0           1
## 3   31      0         0       0             0            1           0
## 4   28      1         0       0             0            1           0
## 5   54      0         0       0             0            0           1
## 6   45      1         0       1             0            0           0
## 7   84      1         0       0             1            0           0
## 8   71      0         0       0             0            1           0
## 9   68      1         0       0             1            0           0
## 10  37      1         1       0             0            0           0
```

```
attach(acs2017_ny)
```

Summary of and Number of People In Dataset

```
summary(acs2017_ny)
```

```
##       AGE            female         educ_nohs        educ_hs
##  Min.   : 0.00   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:22.00   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :42.00   Median :1.0000   Median :0.000   Median :0.0000
##  Mean   :41.57   Mean   :0.5156   Mean   :0.271   Mean   :0.2804
##  3rd Qu.:60.00   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :95.00   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##
##  educ_somecoll    educ_college     educ_advdeg                  SCHOOL
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000   N/A              :  5569
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000   No, not in school:144968
##  Median :0.000   Median :0.0000   Median :0.000   Yes, in school   : 46048
##  Mean   :0.173   Mean   :0.1567   Mean   :0.119   Missing          :     0
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.000
##
##                         EDUC
```

```
##  Grade 12                 :55119
##  4 years of college       :30802
##  5+ years of college      :23385
##  1 year of college        :19947
##  Nursery school to grade 4:14240
##  2 years of college       :14065
##  (Other)                  :39027
##                                          EDUCD
##  Regular high school diploma                :35689
##  Bachelor's degree                          :30802
##  1 or more years of college credit, no degree:19947
##  Master's degree                            :17010
##  Associate's degree, type not specified     :14065
##  Some college, but less than 1 year         : 9086
##  (Other)                                    :69986
##                                      DEGFIELD
##  N/A                                    :142398
##  Business                               :  9802
##  Education Administration and Teaching  :  6708
##  Social Sciences                        :  4836
##  Medical and Health Sciences and Services:  3919
##  Fine Arts                              :  3491
##  (Other)                                : 25431
##                                    DEGFIELDD
##  N/A                                   :142398
##  Psychology                            :  2926
##  Business Management and Administration:  2501
##  Accounting                            :  2284
##  General Education                     :  2238
##  English Language and Literature       :  2202
##  (Other)                               : 42036
##                                   DEGFIELD2
##  N/A                                  :190425
##  Business                             :   972
##  Social Sciences                      :   853
##  Education Administration and Teaching:   611
##  Fine Arts                            :   465
##  Communications                       :   352
##  (Other)                              :  2907
##                                                                  DEGFIELD2D
##  N/A                                                                :190425
##  Psychology                                                         :   284
##  Economics                                                          :   260
##  Political Science and Government                                   :   243
##  Business Management and Administration                             :   217
##  French, German, Latin and Other Common Foreign Language Studies:   205
##  (Other)                                                            :  4951
##       PUMA            GQ          OWNERSHP       OWNERSHPD        MORTGAGE
##  Min.   : 100   Min.   :1.000   Min.   :0.000   Min.   : 0.00   Min.   :0.000
##  1st Qu.:1500   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:12.00   1st Qu.:0.000
##  Median :3201   Median :1.000   Median :1.000   Median :13.00   Median :1.000
##  Mean   :2713   Mean   :1.148   Mean   :1.266   Mean   :14.95   Mean   :1.453
##  3rd Qu.:3902   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:22.00   3rd Qu.:3.000
##  Max.   :4114   Max.   :5.000   Max.   :2.000   Max.   :22.00   Max.   :4.000
```

```
##
##     OWNCOST           RENT         COSTELEC       COSTGAS        COSTWATR
##  Min.   :    0   Min.   :   0   Min.   :   0   Min.   :   0   Min.   :   0
##  1st Qu.: 1208   1st Qu.:   0   1st Qu.: 960   1st Qu.: 840   1st Qu.: 320
##  Median : 2891   Median :   0   Median :1560   Median :2400   Median :1400
##  Mean   :38582   Mean   : 393   Mean   :2311   Mean   :5032   Mean   :4836
##  3rd Qu.:99999   3rd Qu.: 630   3rd Qu.:2520   3rd Qu.:9993   3rd Qu.:9993
##  Max.   :99999   Max.   :3800   Max.   :9997   Max.   :9997   Max.   :9997
##
##     COSTFUEL       HHINCOME          FOODSTMP        LINGISOL
##  Min.   :   0   Min.   : -11800   Min.   :1.000   Min.   :0.000
##  1st Qu.:9993   1st Qu.:  41600   1st Qu.:1.000   1st Qu.:1.000
##  Median :9993   Median :  81700   Median :1.000   Median :1.000
##  Mean   :7935   Mean   : 114902   Mean   :1.147   Mean   :1.002
##  3rd Qu.:9993   3rd Qu.: 140900   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :9997   Max.   :2030000   Max.   :2.000   Max.   :2.000
##                 NA's   :10630
##      ROOMS           BUILTYR2         UNITSSTR        FUELHEAT
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.00   Min.   :0.000
##  1st Qu.: 4.000   1st Qu.: 1.000   1st Qu.: 3.00   1st Qu.:2.000
##  Median : 6.000   Median : 3.000   Median : 3.00   Median :2.000
##  Mean   : 5.887   Mean   : 3.711   Mean   : 4.39   Mean   :2.959
##  3rd Qu.: 8.000   3rd Qu.: 5.000   3rd Qu.: 6.00   3rd Qu.:4.000
##  Max.   :16.000   Max.   :22.000   Max.   :10.00   Max.   :9.000
##
##       SSMC            FAMSIZE          NCHILD          NCHLT5
##  Min.   :0.00000   Min.   : 1.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.: 2.000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000   Median : 3.000   Median :0.0000   Median :0.00000
##  Mean   :0.01102   Mean   : 3.087   Mean   :0.5009   Mean   :0.08441
##  3rd Qu.:0.00000   3rd Qu.: 4.000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :2.00000   Max.   :19.000   Max.   :9.0000   Max.   :5.00000
##
##      RELATE          RELATED           MARST            RACE          RACED
##  Min.   : 1.000   Min.   : 101.0   Min.   :1.000   Min.   :1.00   Min.   :100
##  1st Qu.: 1.000   1st Qu.: 101.0   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:100
##  Median : 2.000   Median : 201.0   Median :5.000   Median :1.00   Median :100
##  Mean   : 3.307   Mean   : 335.6   Mean   :3.742   Mean   :2.03   Mean   :205
##  3rd Qu.: 3.000   3rd Qu.: 301.0   3rd Qu.:6.000   3rd Qu.:2.00   3rd Qu.:200
##  Max.   :13.000   Max.   :1301.0   Max.   :6.000   Max.   :9.00   Max.   :990
##
##      HISPAN          HISPAND                  BPL
##  Min.   :0.0000   Min.   :  0.00   New York     :128517
##  1st Qu.:0.0000   1st Qu.:  0.00   West Indies  :  8481
##  Median :0.0000   Median :  0.00   China        :  4964
##  Mean   :0.4153   Mean   : 44.75   SOUTH AMERICA:  4957
##  3rd Qu.:0.0000   3rd Qu.:  0.00   India        :  3476
##  Max.   :4.0000   Max.   :498.00   Pennsylvania :  3303
##                                    (Other)      : 42887
##                  BPLD                            ANCESTR1
##  New York          :128517   Not Reported             :32021
##  China             :  4116   Italian                  :20577
##  Dominican Republic:  3517   Irish, various subheads,:16388
##  Pennsylvania      :  3303   German                   :12781
```

```
##  New Jersey        :  3127   African-American        : 9559
##  Puerto Rico       :  2272   United States           : 8209
##  (Other)           : 51733   (Other)                 :97050
##                                    ANCESTR1D             ANCESTR2
##  Not Reported                          :32021   Not Reported:141487
##  Italian (1990-2000, ACS, PRCS)        :20577   German      :  9476
##  Irish                                 :15651   Irish       :  9238
##  German (1990-2000, ACS/PRCS)          :12605   English     :  4895
##  African-American (1990-2000, ACS, PRCS): 9559   Italian     :  4531
##  United States                         : 8209   Polish      :  3113
##  (Other)                               :97963   (Other)     : 23845
##                            ANCESTR2D         CITIZEN          YRSUSA1
##  Not Reported                 :141487   Min.   :0.0000   Min.   : 0.000
##  German (1990-2000, ACS, PRCS) :  9441   1st Qu.:0.0000   1st Qu.: 0.000
##  Irish                        :  8809   Median :0.0000   Median : 0.000
##  English                      :  4895   Mean   :0.4793   Mean   : 5.377
##  Italian (1990-2000, ACS, PRCS):  4531   3rd Qu.:0.0000   3rd Qu.: 0.000
##  Polish                       :  3113   Max.   :3.0000   Max.   :92.000
##  (Other)                      : 24309
##     HCOVANY         HCOVPRIV         SEX            EMPSTAT
##  Min.   :1.000   Min.   :1.000   Male  : 95222   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:1.000   Female:101363   1st Qu.:1.000
##  Median :2.000   Median :2.000                   Median :1.000
##  Mean   :1.951   Mean   :1.691                   Mean   :1.514
##  3rd Qu.:2.000   3rd Qu.:2.000                   3rd Qu.:3.000
##  Max.   :2.000   Max.   :2.000                   Max.   :3.000
##
##     EMPSTATD        LABFORCE          OCC              IND
##  Min.   : 0.00   Min.   :0.000   0      : 79987   0      :79987
##  1st Qu.:10.00   1st Qu.:1.000   2310   :  3494   7860   : 9025
##  Median :10.00   Median :2.000   5700   :  3235   8680   : 6354
##  Mean   :15.16   Mean   :1.331   430    :  3025   770    : 6279
##  3rd Qu.:30.00   3rd Qu.:2.000   4720   :  2666   8190   : 5873
##  Max.   :30.00   Max.   :2.000   4760   :  2563   7870   : 4041
##                                  (Other):101615   (Other):85026
##     CLASSWKR       CLASSWKRD        WKSWORK2        UHRSWORK
##  Min.   :0.000   Min.   : 0.00   Min.   :0.000   Min.   : 0.00
##  1st Qu.:0.000   1st Qu.: 0.00   1st Qu.:0.000   1st Qu.: 0.00
##  Median :2.000   Median :22.00   Median :1.000   Median :12.00
##  Mean   :1.116   Mean   :13.03   Mean   :2.701   Mean   :19.77
##  3rd Qu.:2.000   3rd Qu.:22.00   3rd Qu.:6.000   3rd Qu.:40.00
##  Max.   :2.000   Max.   :29.00   Max.   :6.000   Max.   :99.00
##
##      INCTOT           FTOTINC           INCWAGE          POVERTY
##  Min.   :  -7300   Min.   : -11800   Min.   :     0   Min.   :  0.0
##  1st Qu.:   8000   1st Qu.:  35550   1st Qu.:     0   1st Qu.:159.0
##  Median :  25000   Median :  74000   Median : 10000   Median :351.0
##  Mean   :  45245   Mean   : 107110   Mean   : 33796   Mean   :318.7
##  3rd Qu.:  56500   3rd Qu.: 132438   3rd Qu.: 47000   3rd Qu.:501.0
##  Max.   :1563000   Max.   :2030000   Max.   :638000   Max.   :501.0
##  NA's   :31129     NA's   :10817     NA's   :33427
##     MIGRATE1       MIGRATE1D        MIGPLAC1         MIGCOUNTY1
##  Min.   :0.000   Min.   : 0.00   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:1.000   1st Qu.:10.00   1st Qu.:  0.000   1st Qu.:  0.000
```

```
##  Median :1.000   Median :10.00   Median :  0.000   Median :  0.000
##  Mean   :1.122   Mean   :11.51   Mean   :  6.184   Mean   :  4.117
##  3rd Qu.:1.000   3rd Qu.:10.00   3rd Qu.:  0.000   3rd Qu.:  0.000
##  Max.   :4.000   Max.   :40.00   Max.   :900.000   Max.   :810.000
##
##     MIGPUMA1        VETSTAT          VETSTATD         PWPUMA00
##  Min.   :    0   Min.   :0.0000   Min.   : 0.000   Min.   :    0
##  1st Qu.:    0   1st Qu.:1.0000   1st Qu.:11.000   1st Qu.:    0
##  Median :    0   Median :1.0000   Median :11.000   Median :    0
##  Mean   :  277   Mean   :0.8621   Mean   : 9.412   Mean   : 1255
##  3rd Qu.:    0   3rd Qu.:1.0000   3rd Qu.:11.000   3rd Qu.: 3100
##  Max.   :70100   Max.   :2.0000   Max.   :20.000   Max.   :59300
##
##     TRANWORK         TRANTIME         DEPARTS           in_NYC
##  Min.   : 0.000   Min.   :  0.00   Min.   :   0.0   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:   0.0   1st Qu.:0.0000
##  Median : 0.000   Median :  0.00   Median :   0.0   Median :0.0000
##  Mean   : 9.725   Mean   : 14.75   Mean   : 373.3   Mean   :0.3615
##  3rd Qu.:10.000   3rd Qu.: 20.00   3rd Qu.: 732.0   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :138.00   Max.   :2345.0   Max.   :1.0000
##
##     in_Bronx       in_Manhattan       in_StatenI       in_Brooklyn
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.000
##  Mean   :0.0538   Mean   :0.04981   Mean   :0.02084   Mean   :0.126
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000
##
##    in_Queens      in_Westchester      in_Nassau          Hispanic
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.1111   Mean   :0.04413   Mean   :0.07032   Mean   :0.1387
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##     Hisp_Mex          Hisp_PR          Hisp_Cuban         Hisp_DomR
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.000000   Median :0.00000
##  Mean   :0.01626   Mean   :0.0436   Mean   :0.003403   Mean   :0.02827
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.000000   Max.   :1.00000
##
##      white             AfAm           Amindian            Asian
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :1.0000   Median :0.000   Median :0.000000   Median :0.00000
##  Mean   :0.6997   Mean   :0.125   Mean   :0.003779   Mean   :0.08656
##  3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.000000   Max.   :1.00000
##
##     race_oth        unmarried       veteran        has_AnyHealthIns
```

```
##  Min.   :0.0000   Min.   :0.00   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :0.0000   Median :0.00   Median :0.00000   Median :1.0000
##  Mean   :0.1324   Mean   :0.45   Mean   :0.04443   Mean   :0.9513
##  3rd Qu.:0.0000   3rd Qu.:1.00   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.00   Max.   :1.00000   Max.   :1.0000
##
##  has_PvtHealthIns  Commute_car      Commute_bus      Commute_subway
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :1.0000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.6906   Mean   :0.2997   Mean   :0.02162   Mean   :0.07468
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##
##   Commute_rail     Commute_other     below_povertyline below_150poverty
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000     Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000     1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.000     Median :0.0000
##  Mean   :0.01332   Mean   :0.05506   Mean   :0.122     Mean   :0.1965
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000     3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.000     Max.   :1.0000
##
##  below_200poverty   foodstamps
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.2676   Mean   :0.1465
##  3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

```
print(NN_obs <- length(AGE))
```

```
## [1] 196585
```

###Comparing the Age of Men and Women in the Dataset

```
summary(AGE[female == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   23.00   44.00   42.72   61.00   95.00
```

```
summary(AGE[!female])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   21.00   40.00   40.35   59.00   95.00
```

Women were found to be older than men by a couple on average.

Avg Age of Men and Women in the Data Set

```
mean(AGE[female == 1])
```

```
## [1] 42.71629
```

```
sd(AGE[female == 1])
```

```
## [1] 23.72012
```

```
mean(AGE[!female])
```

```
## [1] 40.35398
```

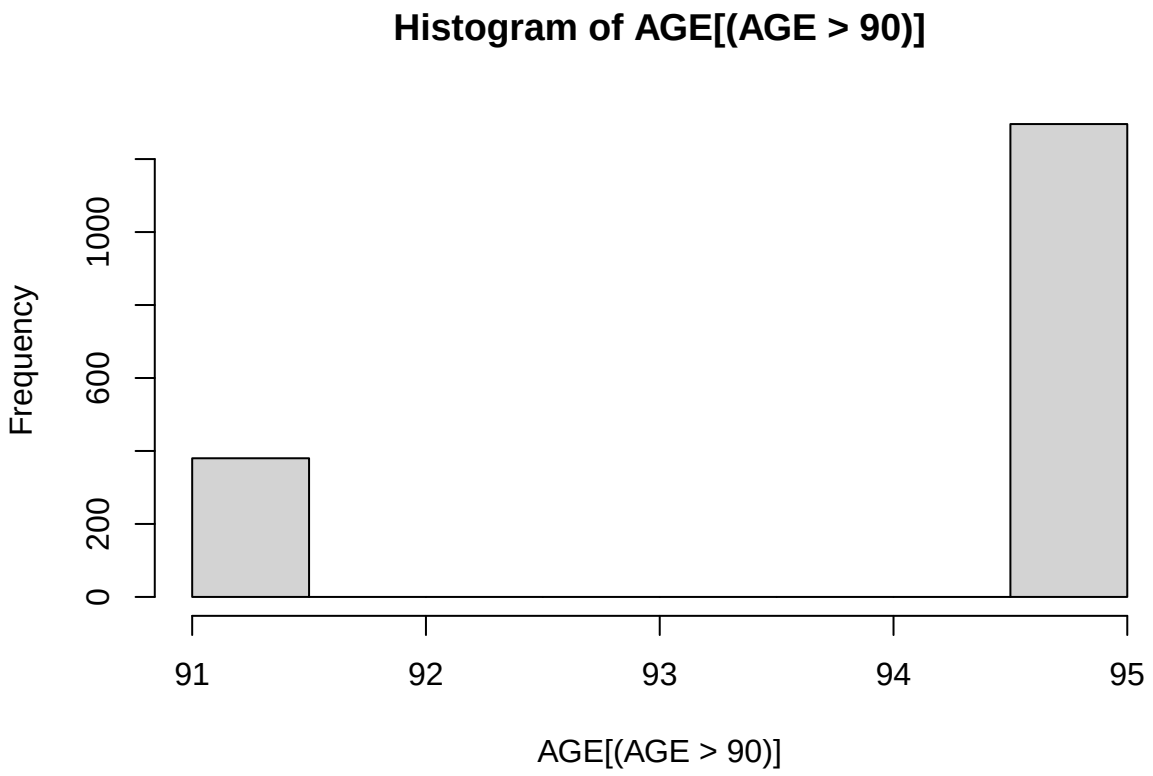
```
sd(AGE[!female])
```

```
## [1] 23.1098
```

Something interesting I learned from this dataset:

Variable Coding

```
hist(AGE[(AGE>90)])
```



"There is a bit of weirdness in the right, where it looks like there are suddenly a bunch of people who are 95 but nobody is 94 or 96. This is due to a coding choice by the Census, where really old people are just labeled as "95" (top-coding) so it actually should be interpreted as meaning "92 or older". So if you were to get finicky (and every good statistician is!) you might go back to the calculations of averages previously and modify them all like this, to select just those who are female and who are coded as having age less than 90. Many variables are topcoded! And recall that topcoding wouldn't change the median values calculated before, which is a point in favor of that statistic."

```
str(as.numeric(PUMA))
```

```
##  num [1:196585] 902 902 4002 4002 3803 ...
```

```
print(levels(female))
```

```
## NULL
```

```
levels(female) <- c("male","female")
```

```
educ_indx <- factor((educ_nohs + 2*educ_hs + 3*educ_somecoll + 4*educ_college + 5*educ_advdeg), levels=
```

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
install.packages("plyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
levels_n <- read.csv("PUMA_levels.csv")
levels_orig <- levels(PUMA)
levels_new <- join(data.frame(levels_orig),data.frame(levels_n))
```

```
## Joining by:
```

```
levels(PUMA) <- levels_new$New_Level
```