U6614: Assignment 2

Philip Crane (plc2137)

2024-01-25

```
library(tidyverse)
```

1 Load and inspect CPS data:

1a) Inspect the data frame and data types for each column

- remember to remove NAs
- make sure to inspect the age, sex, race, college columns

```
cps <- read.csv("cps_june_22-23.csv")
  cps <- na.omit(cps)
```

```
summary(cps$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   30.00   41.00   42.19   54.00   85.00
```

```
summary(cps$sex)
```

```
##    Length     Class      Mode
##     20120 character character
```

```
summary(cps$race)
```

```
##    Length     Class      Mode
##     20120 character character
```

```
summary(cps$college)
```

```
##    Length     Class      Mode
##     20120 character character
```

1b) Use the mutate function to create new column for sex

- sex.fac = as.factor(sex),
- check if it worked by calling the str() function

```
mutate(cps, sex.fac = as.factor(sex))
```

```
str(mutate(cps, sex.fac = as.factor(sex)))
```

```
## 'data.frame':    20120 obs. of  15 variables:
##  $ year    : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...
##  $ month   : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ statefip: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  48 24 23 46 65 26 27 50 46 22 ...
##  $ sex     : chr  "Male" "Male" "Female" "Male" ...
##  $ race    : chr  "White" "White" "White" "Black" ...
##  $ college : chr  "College degree" "No college degree" "No college degree" "No college degree" ...
##  $ earnweek: num  2880 720 420 654 1510 600 600 1730 1460 300 ...
##  $ hrsworkt: int  40 40 40 40 24 40 40 40 40 30 ...
##  $ hispanic: chr  "Not Hispanic" "Not Hispanic" "Not Hispanic" "Not Hispanic" ...
##  $ ind     : int  2190 7680 5170 9160 8191 7480 7480 1270 6991 5080 ...
##  $ hhid    : num  2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
##  $ personid: num  2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
##  $ serial  : int  11 14 14 38 40 54 54 76 79 79 ...
##  $ sex.fac : Factor w/ 2 levels "Female","Male": 2 2 1 2 2 1 1 2 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1032] 44 108 117 144 180 200 205 232 269 312 ...
##   ..- attr(*, "names")= chr [1:1032] "44" "108" "117" "144" ...
```

1c) Include sex.fac in a new data frame called cps.temp1

- also create factors for race and college education,
- use a pipe to exclude the columns for serial, ind
- after creating cps.temp1, print the first 5 observations

```
cps.temp1 <- cps %>%
  mutate(sex.fac = as.factor(sex),
         race.fac = as.factor(race),
         college.fac = as.factor(college)) %>%
  select(-serial, -ind)

head(cps.temp1, n = 5)
```

```
##   year month statefip age    sex  race           college earnweek hrsworkt
## 1 2022     6        1  48   Male White    College degree     2880       40
## 2 2022     6        1  24   Male White No college degree      720       40
## 3 2022     6        1  23 Female White No college degree      420       40
## 4 2022     6        1  46   Male Black No college degree      654       40
## 5 2022     6        1  65   Male Black No college degree     1510       24
##       hispanic        hhid    personid sex.fac race.fac       college.fac
## 1 Not Hispanic 2.02203e+13 2.02203e+13    Male    White    College degree
## 2 Not Hispanic 2.02203e+13 2.02203e+13    Male    White No college degree
## 3 Not Hispanic 2.02203e+13 2.02203e+13  Female    White No college degree
## 4 Not Hispanic 2.02203e+13 2.02203e+13    Male    Black No college degree
## 5 Not Hispanic 2.02103e+13 2.02103e+13    Male    Black No college degree
```

1d) Inspect race.fac, sex.fac, and college.fac using the levels() function

- what package is the levels() function located in?

```r
levels(cps.temp1$sex.fac)
```

```
## [1] "Female" "Male"
```

```r
levels(cps.temp1$race.fac)
```

```
##  [1] "American Indian-Asian"
##  [2] "American Indian/Aleut/Eskimo"
##  [3] "Asian-Hawaiian/Pacific Islander"
##  [4] "Asian only"
##  [5] "Black"
##  [6] "Black-American Indian"
##  [7] "Black-Asian"
##  [8] "Black-Hawaiian/Pacific Islander"
##  [9] "Hawaiian/Pacific Islander only"
## [10] "White"
## [11] "White-American Indian"
## [12] "White-Asian"
## [13] "White-Asian-Hawaiian/Pacific Islander"
## [14] "White-Black"
## [15] "White-Black--Hawaiian/Pacific Islander"
## [16] "White-Black-American Indian"
## [17] "White-Black-American Indian-Asian"
## [18] "White-Black-Asian"
## [19] "White-Hawaiian/Pacific Islander"
```

```r
levels(cps.temp1$college.fac)
```

```
## [1] "College degree"    "No college degree"
```

The levels function is located in the base R package

1e) Use filter() to only include rows only for June 2022

- store as a new object cps_2022,
- print the first 5 observations,
- confirm your data only includes observations for 2022

```r
cps_2022 <- cps.temp1 %>%
  filter(year == 2022)

head(cps_2022, n = 5)
```

```
##   year month statefip age    sex  race           college earnweek hrsworkt
## 1 2022     6        1  48   Male White    College degree     2880       40
## 2 2022     6        1  24   Male White No college degree      720       40
## 3 2022     6        1  23 Female White No college degree      420       40
## 4 2022     6        1  46   Male Black No college degree      654       40
## 5 2022     6        1  65   Male Black No college degree     1510       24
##       hispanic        hhid    personid sex.fac race.fac       college.fac
## 1 Not Hispanic 2.02203e+13 2.02203e+13    Male    White    College degree
## 2 Not Hispanic 2.02203e+13 2.02203e+13    Male    White No college degree
## 3 Not Hispanic 2.02203e+13 2.02203e+13  Female    White No college degree
## 4 Not Hispanic 2.02203e+13 2.02203e+13    Male    Black No college degree
## 5 Not Hispanic 2.02103e+13 2.02103e+13    Male    Black No college degree
```

1f) Remove the cps.temp1 object from memory using the rm() function

```
rm(cps.temp1)
```

2 Describe the cps_2022 data frame

2a) What is the unit of observation?

The unit of observation is the individual survey respondent.

2b) How many individuals are observed? from how many households?

```
summarise(cps_2022, n_distinct(personid))
```

```
##   n_distinct(personid)
## 1                10239
```

```
summarise(cps_2022, n_distinct(hhid))
```

```
##   n_distinct(hhid)
## 1             6729
```

There are 10239 individuals and 6729 households

2c) What is the average age of individuals in the sample? Youngest and oldest person?

```
sumstats <- cps_2022 %>%
  summarise(avg_age = mean(age),
            min_age = min(age),
            max_age = max(age))
```

The average age is 42.08, the oldest person is 85 and the youngest is 15.

3 Earnings per week for different groups in June 2022

3a) Find the observation for the top weekly earnings using the summarise() function

- assign this to a new object called max_earnings

```
max_earnings <- cps_2022 %>%
  summarise(max_earning = max(earnweek))
```

3b) Find max weekly earnings using the arrange function instead of summarise

```
cps_2022 %>%
    arrange(desc(earnweek)) %>%
    select(earnweek) %>%
    head(n=1)
```

```
##   earnweek
## 1  2884.61
```

3c) Use the filter function to subset for the observation with max weekly earnings

- don't hardcode the max earnings to filter on, refer to the max_earnings object from a),
- store in new data frame cps_max_earn,
- confirm it worked

```
cps_max_earn <- cps_2022 %>%
  filter(earnweek == max_earnings[1,])

summary(cps_max_earn$earnweek)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2885    2885    2885    2885    2885    2885
```

3d) What is the age, sex, and race of the top weekly earner in the sample?

```
cps_max_earn[1,4:6]
```

```
##   age  sex  race
## 1  38 Male Black
```

3e) List the age, sex, and race of the top 10 weekly earners in the sample

```
cps_2022 %>%
  arrange(desc(earnweek)) %>%
  select(age, sex, race) %>%
  head(n=10)
```

```
##    age    sex                  race
## 1   38   Male                 Black
## 2   33 Female                 White
## 3   49 Female Black-American Indian
## 4   38   Male                 White
## 5   66 Female                 White
## 6   38   Male                 White
## 7   54 Female                 White
## 8   63   Male                 White
## 9   30   Male                 White
## 10  29   Male                 White
```

3f) How many individuals earned more than $2000 in weekly earnings?

```
cps_2022 %>%
  filter(earnweek > 2000) %>%
  nrow()
```

```
## [1] 1501
```

4 Wage gaps between males and females:

4a) Use the filter function to subset observations for males

- assign to new data frame, cps_2022_male,
- sort in descending order of weekly earnings
- check if it worked

```
cps_2022_male <- cps_2022 %>%
  filter(sex == 'Male') %>%
  arrange(desc(earnweek))

str(cps_2022_male)
```

```
## ’data.frame':    5384 obs. of  15 variables:
##  $ year       : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...
##  $ month      : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ statefip   : int  1 1 1 1 2 2 4 4 4 4 ...
##  $ age        : int  38 38 38 63 30 29 42 41 31 52 ...
##  $ sex        : chr  "Male" "Male" "Male" "Male" ...
##  $ race       : chr  "Black" "White" "White" "White" ...
##  $ college    : chr  "College degree" "College degree" "College degree" "No college degree" ...
##  $ earnweek   : num  2885 2885 2885 2885 2885 ...
##  $ hrsworkt   : int  40 55 50 50 80 60 40 48 40 60 ...
##  $ hispanic   : chr  "Not Hispanic" "Not Hispanic" "Not Hispanic" "Not Hispanic" ...
##  $ hhid       : num  2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
##  $ personid   : num  2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
##  $ sex.fac    : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ race.fac   : Factor w/ 19 levels "American Indian-Asian",..: 5 10 10 10 10 10 10 10 10 10 ...
##  $ college.fac: Factor w/ 2 levels "College degree",..: 1 1 1 2 1 1 1 1 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1032] 44 108 117 144 180 200 205 232 269 312 ...
##   ..- attr(*, "names")= chr [1:1032] "44" "108" "117" "144" ...
```

4b) Repeat part a for females and create a new data frame, cps_2022_female

```r
cps_2022_female <- cps_2022 %>%
  filter(sex == 'Female') %>%
  arrange(desc(earnweek))

str(cps_2022_female)
```

```
## 'data.frame':    4855 obs. of  15 variables:
##  $ year       : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...
##  $ month      : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ statefip   : int  1 1 1 1 4 6 6 6 6 6 ...
##  $ age        : int  33 49 66 54 52 50 36 37 39 46 ...
##  $ sex        : chr  "Female" "Female" "Female" "Female" ...
##  $ race       : chr  "White" "Black-American Indian" "White" "White" ...
##  $ college    : chr  "College degree" "College degree" "College degree" "College degree" ...
##  $ earnweek   : num  2885 2885 2885 2885 2885 ...
##  $ hrsworkt   : int  40 40 60 25 60 70 40 40 32 50 ...
##  $ hispanic   : chr  "Not Hispanic" "Not Hispanic" "Not Hispanic" "Not Hispanic" ...
##  $ hhid       : num  2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
##  $ personid   : num  2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
##  $ sex.fac    : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ race.fac   : Factor w/ 19 levels "American Indian-Asian",..: 10 6 10 10 4 10 10 10 4 10 ...
##  $ college.fac: Factor w/ 2 levels "College degree",..: 1 1 1 1 1 2 1 1 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1032] 44 108 117 144 180 200 205 232 269 312 ...
##   ..- attr(*, "names")= chr [1:1032] "44" "108" "117" "144" ...
```

4c) Use summarise to find mean, min & max for males and females, separately

- name each statistic appropriately (i.e. name each column in the 1-row table of stats)
- what is the gender gap in mean weekly earnings?

```r
summarise(cps_2022_male, avg_earn = mean(earnweek),
                         min_earn = min(earnweek),
                         max_earn = max(earnweek))
```

```
##   avg_earn min_earn max_earn
## 1 1268.948        4  2884.61
```

```r
summarise(cps_2022_female, avg_earn = mean(earnweek),
                           min_earn = min(earnweek),
                           max_earn = max(earnweek))
```

```
##   avg_earn min_earn max_earn
## 1 1014.649        4  2884.61
```

The average gender gap in mean weekly earnings is 254.3

4d) What is the wage gap in weekly earnings between white males and Black females?

```r
cps_2022_male_white <- cps_2022_male %>%
  filter(race == 'White')

cps_2022_female_black <- cps_2022_female %>%
  filter(race == 'Black')
```

The average wage gap between white males and black females is 395.53

4e) What is the wage gap between college educated white males and college educated Black females?

```r
cps_2022_male_white_edu <- cps_2022_male_white %>%
  filter(college == 'College degree')

cps_2022_female_black_edu <- cps_2022_female_black %>%
  filter(college == 'College degree')
```

The average wage gap between college educated white males and college educated black females is 436.12