Assignment 2

Sample Solutions

2023-09-18

```
library(tidyverse)
```

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on the due date.

Before knitting your rmd file as a pdf, you will need to install TinyTex for Latex distribution by running the following code:

`tinytex::install_tinytex()`

Please visit this link for more information on TinyTex installation.

1 Load and inspect CPS data:

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

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

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

```
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
```

b) 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)) #output suppressed
```

```
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" ...
```

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

- also create new factor columns for race and college education,
- in the same pipe, exclude the columns for serial and 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
```

```
#A neater way to present (key data only, other cols omitted)
head(cps.temp1, n = 5) %>%
  select(sex.fac, race.fac, college.fac, earnweek) %>%
  knitr::kable()
```

sex.fac	race.fac	college.fac	earnweek
Male	White	College degree	2880
Male	White	No college degree	720
Female	White	No college degree	420
Male	Black	No college degree	654
Male	Black	No college degree	1510

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

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

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

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

```
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"
```

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

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

```
#?levels #from the documentation, the levels function is located in base R.
```

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

- store as a new object cps_2022,
- print the first 5 observations,

```
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
```

- confirm your data only includes observations for 2022

```
summary(cps_2022$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2022    2022    2022    2022    2022    2022
```

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

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

2 Describe the cps_2022 data frame

a) What is the unit of observation?

```
str(cps_2022)
```

```
## 'data.frame':    10239 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" ...
##  $ 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 1 2 2 1 1 2 1 1 ...
##  $ race.fac   : Factor w/ 19 levels "American Indian-Asian",..: 10 10 10 5 5 10 10 5 5 5 ...
##  $ college.fac: Factor w/ 2 levels "College degree",..: 1 2 2 2 2 2 2 2 1 2 ...
##  - 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" ...
```

The unit of observation is the individual (individuals surveyed in June, 2022).

b) 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
```

```
#Alternative way:
cps_2022 %>%
  summarise(num_persons = n_distinct(personid),
            num_households = n_distinct(hhid))
```

There are 10239 unique individuals, and 6729 unique households.

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

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

```
##    avg_age min_age max_age
## 1 42.08409      15      85
```

Alternatively, using inline code:

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

The average age in the sample is 42.08, with individuals ranging from 15 to 85 years old.

3 Earnings per week for different groups in June 2022

a) 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_earnings = max(earnweek),)
max_earnings
```

```
##   max_earnings
## 1      2884.61
```

b) 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
```

c) 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,

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

#Alternative way:
cps_max_earn <- cps_2022 %>%
                  arrange(desc(earnweek)) %>%
                  head(n = 1)
```

- confirm it worked

```
cps_max_earn %>%
  select(sex, race, age, personid, college, earnweek)
```

```
##    sex  race age    personid        college earnweek
## 1 Male Black  38 2.02203e+13 College degree  2884.61
```

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

```
cps_max_earn %>%
  select(age,sex,race) %>%
  head(n = 1)
```

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

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

```
cps_2022 %>%
  arrange(desc(earnweek)) %>%
```

```
  select(age,sex,race, earnweek) %>%
  head(n=10) %>%
  knitr::kable()
```

age	sex	race	earnweek
38	Male	Black	2884.61
33	Female	White	2884.61
49	Female	Black-American Indian	2884.61
38	Male	White	2884.61
66	Female	White	2884.61
38	Male	White	2884.61
54	Female	White	2884.61
63	Male	White	2884.61
30	Male	White	2884.61
29	Male	White	2884.61

f) 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:

a) 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.fac == "Male") %>%
                    arrange(desc(earnweek))

#Check
cps_2022_male %>%
  select(sex.fac,earnweek) %>%
  head(n = 3) %>%
  knitr::kable()
```

sex.fac	earnweek
Male	2884.61
Male	2884.61
Male	2884.61

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

```
cps_2022_female <- cps_2022 %>%
                    filter(sex.fac == "Female") %>%
                    arrange(desc(earnweek))

#Check
cps_2022_female %>%
  select(sex.fac,earnweek) %>%
  head(n = 3) %>%
  knitr::kable()
```

sex.fac	earnweek
Female	2884.61
Female	2884.61
Female	2884.61

c) 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)

```
cps_2022_male %>%
  summarise(mean_earnings_male = mean(earnweek),
            min_earnings_male = min(earnweek),
            max_earnings_male = max(earnweek)) %>%
  knitr::kable()
```

mean_earnings_male	min_earnings_male	max_earnings_male
1268.948	4	2884.61

```
cps_2022_female %>%
  summarise(mean_earnings_female = mean(earnweek),
            min_earnings_female = min(earnweek),
            max_earnings_female = max(earnweek)) %>%
  knitr::kable()
```

mean_earnings_female	min_earnings_female	max_earnings_female
1014.649	4	2884.61

- what is the gender gap in mean weekly earnings?

The gender gap in weekly earnings is 254.3 .

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

```
cps_2022_wh_male <- cps_2022_male %>%
                      filter(race.fac == "White")
```

```
cps_2022_bl_female <- cps_2022_female %>%
                        filter(race.fac == "Black")
```

The weekly earnings gap between white males and Black females is $395.53.

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

```
cps_2022_wh_male_college <- cps_2022_wh_male %>%
                              filter(college.fac == "College degree" &
                                       race.fac == "White")
```

```
cps_2022_bl_female_college <- cps_2022_bl_female %>%
                              filter(college.fac == "College degree" &
                                       race.fac == "Black")
```

The weekly earnings gap between white college-educated males and Black college-educated females is $436.12.

NOTE: the exercises above are done using weekly earnings, but can easily be converted to hourly wages

End of assignment.