Assignment 2 - Sample Solutions

Liam Tay Kearney

2022-02-01

```
library(tidyverse)
```

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on Monday, January 31st.

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_20-21.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. 
##   16.00   30.00   40.00   40.77   52.00   64.00
```

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

```
##    Length     Class      Mode 
##     16876 character character
```

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

```
##    Length     Class      Mode 
##     16876 character character
```

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

```
##    Length     Class      Mode 
##     16876 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':    16876 obs. of  15 variables:
##  $ year    : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ 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  44 47 45 29 28 59 25 24 56 42 ...
##  $ sex     : chr  "Female" "Female" "Male" "Female" ...
##  $ race    : chr  "Black/Negro" "Black/Negro" "Black/Negro" "Black/Negro" ...
##  $ college : chr  "No college degree" "No college degree" "No college degree" "No college degree" ..
##  $ earnweek: num  750 1093 760 510 800 ...
##  $ hrsworkt: int  40 40 45 40 50 40 40 40 40 40 ...
##  $ hispanic: chr  "Not Hispanic" "Not Hispanic" "Not Hispanic" "Not Hispanic" ...
##  $ ind     : int  6290 8090 6390 1370 4280 7860 6170 6970 7690 6380 ...
##  $ 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  5 7 7 36 39 48 55 59 69 100 ...
##  $ sex.fac : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 1 1 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:885] 51 64 68 71 91 148 149 152 160 161 ...
##   ..- attr(*, "names")= chr [1:885] "51" "64" "68" "71" ...
```

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
## 1 2020     6        1  44 Female Black/Negro No college degree    750.0
## 2 2020     6        1  47 Female Black/Negro No college degree   1092.6
## 3 2020     6        1  45   Male Black/Negro No college degree    760.0
## 4 2020     6        1  29 Female Black/Negro No college degree    510.0
## 5 2020     6        1  28   Male Black/Negro No college degree    800.0
##   hrsworkt     hispanic        hhid    personid sex.fac    race.fac
## 1       40 Not Hispanic 2.02003e+13 2.02003e+13  Female Black/Negro
## 2       40 Not Hispanic 2.01903e+13 2.01903e+13  Female Black/Negro
## 3       45 Not Hispanic 2.01903e+13 2.01903e+13    Male Black/Negro
## 4       40 Not Hispanic 2.02003e+13 2.02003e+13  Female Black/Negro
## 5       50 Not Hispanic 2.01903e+13 2.01903e+13    Male Black/Negro
##         college.fac
## 1 No college degree
## 2 No college degree
## 3 No college degree
## 4 No college degree
## 5 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
Female	Black/Negro	No college degree	750.0
Female	Black/Negro	No college degree	1092.6
Male	Black/Negro	No college degree	760.0
Female	Black/Negro	No college degree	510.0
Male	Black/Negro	No college degree	800.0

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-Hawaiian/Pacific Islander"
##  [3] "American Indian/Aleut/Eskimo"
##  [4] "Asian-Hawaiian/Pacific Islander"
##  [5] "Asian only"
##  [6] "Black-American Indian"
##  [7] "Black-American Indian-Asian"
##  [8] "Black-Asian"
##  [9] "Black/Negro"
## [10] "Four or five races, unspecified"
## [11] "Hawaiian/Pacific Islander only"
## [12] "White"
## [13] "White-American Indian"
## [14] "White-American Indian-Asian"
## [15] "White-Asian"
## [16] "White-Asian-Hawaiian/Pacific Islander"
## [17] "White-Black"
## [18] "White-Black--Hawaiian/Pacific Islander"
## [19] "White-Black-American Indian"
## [20] "White-Black-American Indian-Asian"
## [21] "White-Black-Asian"
## [22] "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 2020

- store as a new object cps_2020,

- print the first 5 observations,

```
cps_2020 <- cps.temp1 %>%
  filter(year == 2020)

head(cps_2020, n = 5)
```

```
##   year month statefip age    sex        race           college earnweek
## 1 2020     6        1  44 Female Black/Negro No college degree    750.0
## 2 2020     6        1  47 Female Black/Negro No college degree   1092.6
## 3 2020     6        1  45   Male Black/Negro No college degree    760.0
## 4 2020     6        1  29 Female Black/Negro No college degree    510.0
## 5 2020     6        1  28   Male Black/Negro No college degree    800.0
##   hrsworkt     hispanic        hhid    personid sex.fac    race.fac
## 1       40 Not Hispanic 2.02003e+13 2.02003e+13  Female Black/Negro
## 2       40 Not Hispanic 2.01903e+13 2.01903e+13  Female Black/Negro
## 3       45 Not Hispanic 2.01903e+13 2.01903e+13    Male Black/Negro
## 4       40 Not Hispanic 2.02003e+13 2.02003e+13  Female Black/Negro
## 5       50 Not Hispanic 2.01903e+13 2.01903e+13    Male Black/Negro
##         college.fac
## 1 No college degree
## 2 No college degree
## 3 No college degree
## 4 No college degree
## 5 No college degree
```

- confirm your data only includes observations for 2020

```
summary(cps_2020$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2020    2020    2020    2020    2020    2020
```

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

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

2 Describe the cps_2020 data frame

a) What is the unit of observation?

```
str(cps_2020)
```

```
## 'data.frame':    7970 obs. of  15 variables:
##  $ year       : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ 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  44 47 45 29 28 59 25 24 56 42 ...
##  $ sex        : chr  "Female" "Female" "Male" "Female" ...
##  $ race       : chr  "Black/Negro" "Black/Negro" "Black/Negro" "Black/Negro" ...
##  $ college    : chr  "No college degree" "No college degree" "No college degree" "No college degree"
##  $ earnweek   : num  750 1093 760 510 800 ...
##  $ hrsworkt   : int  40 40 45 40 50 40 40 40 40 40 ...
##  $ 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 2 1 2 1 1 2 2 2 ...
##  $ race.fac   : Factor w/ 22 levels "American Indian-Asian",..: 9 9 9 9 9 12 12 12 12 9 ...
##  $ college.fac: Factor w/ 2 levels "College degree",..: 2 2 2 2 2 1 1 1 1 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:885] 51 64 68 71 91 148 149 152 160 161 ...
##   ..- attr(*, "names")= chr [1:885] "51" "64" "68" "71" ...
```

The unit of observation is the individual (individuals surveyed in June, 2020).

b) How many individuals are observed? From how many households?

```
summarise(cps_2020, n_distinct(personid))
```

```
##   n_distinct(personid)
## 1                 7970
```

```
summarise(cps_2020, n_distinct(hhid))
```

```
##   n_distinct(hhid)
## 1             5530
```

```
#Alternative way:
cps_2020 %>%
  summarise(num_persons = n_distinct(personid),
            num_households = n_distinct(hhid))
```

There are 7970 unique individuals, and 5530 unique households.

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

```
cps_2020 %>%
  summarise(avg_age = mean(age),
            min_age = min(age),
            max_age = max(age))
```

```
##   avg_age min_age max_age
## 1 41.1803      16      64
```

Alternatively, using inline code:

```
sumstats <- cps_2020 %>%
              summarise(avg_age = mean(age),
                        min_age = min(age),
                        max_age = max(age))
```

The average age in the sample is 41.18, with individuals ranging from 16 to 64 years old.

3 Earnings per week for different groups in June 2020

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_2020 %>%
                  summarise(max_earnings = max(earnweek),)
max_earnings
```

```
##   max_earnings
## 1       2884.5
```

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

```
cps_2020 %>%
  arrange(desc(earnweek)) %>%
  select(earnweek) %>%
  head(n = 1)
```

```
##   earnweek
## 1   2884.5
```

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_2020 %>%
                  filter(earnweek == max_earnings[1,])

#Alternative way:
cps_max_earn <- cps_2020 %>%
                  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 Asian only  39 2.01903e+13 College degree   2884.5
```

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  39 Male Asian only
```

Alternatively, to make it look nicer, we can pipe the output to knitr::kable().

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

age	sex	race
39	Male	Asian only

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

```
cps_2020 %>%
  arrange(desc(earnweek)) %>%
  select(age,sex,race, earnweek) %>%
  head(n=10) %>%
  knitr::kable()
```

age	sex	race	earnweek
39	Male	Asian only	2884.5
36	Male	White	2884.0
41	Male	White	2884.0
57	Female	White	2884.0
45	Male	White	2884.0
40	Male	White	2884.0
59	Male	White	2884.0
64	Male	White	2884.0
49	Male	Asian only	2884.0
36	Male	White	2884.0

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

```
cps_2020 %>%
  filter(earnweek > 2000) %>%
  nrow()
```

```
## [1] 602
```

4 Wage gaps between males and females:

a) Use the filter function to subset observations for males

- assign to new data frame, cps_2020_male,
- sort in descending order of weekly earnings
- check if it worked

```
cps_2020_male <- cps_2020 %>%
                    filter(sex.fac == "Male") %>%
                    arrange(desc(earnweek))

#Check
cps_2020_male %>%
  select(sex.fac,earnweek) %>%
  head(n = 3) %>%
  knitr::kable()
```

sex.fac	earnweek
Male	2884.5
Male	2884.0
Male	2884.0

b) Repeat part a for females and create a new data frame, cps_2020_female

```
cps_2020_female <- cps_2020 %>%
                    filter(sex.fac == "Female") %>%
                    arrange(desc(earnweek))

#Check
cps_2020_female %>%
  select(sex.fac,earnweek) %>%
  head(n = 3) %>%
  knitr::kable()
```

sex.fac	earnweek
Female	2884
Female	2884
Female	2884

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_2020_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
1101.879	0.01	2884.5

```
cps_2020_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
920.7292	0.23	2884

- what is the gender gap in mean weekly earnings?

The gender gap in weekly earnings is $181.15.

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

```
cps_2020_wh_male <- cps_2020_male %>%
                    filter(race.fac == "White")
```

```
cps_2020_bl_female <- cps_2020_female %>%
                      filter(race.fac == "Black/Negro")
```

The weekly earnings gap between white males and Black females is $263.03.

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

```
cps_2020_wh_male_college <- cps_2020_male %>%
                                filter(college.fac == "College degree" &
                                         race.fac == "White")
```

```
cps_2020_bl_female_college <- cps_2020_female %>%
                                filter(college.fac == "College degree" &
                                         race.fac == "Black/Negro")
```

The weekly earnings gap between white college-educated males and Black college-educated females is $345.83.

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

End of assignment.