Homework 2: Beginning to Learn

A. Short answer

B. Empirical exercise

Submission

Start Over

B. Empirical exercise

In this exercise, we will extend the gender pay gap analysis we carried out using the same March 2009 CPS extract. Before we get to the analysis, we will do a little data documentation work.

✓ Data Documentation and Structure

Use the March 2009 extract description posted on eLC to answer the next few questions.

Question 1:

The extract was filtered to include individuals who had worked at least _____ hours per week and _____ weeks during the past year.

36, 48

Correct!

Question 2:

The filtered extract contains _____ observations on _____ variables.

50742, 12

Correct!

Question 3:

The variable `age` is top-coded at _____.

85

Correct!

Question 4:

There are _____ categories in the variable `race` with individuals identifying as Asian only assigned the value _____.

21, 4

Correct!

Continue

✓ Gender earnings gap review

The analysis we carried out in class did not place any age restrictions on the data. Let's summarize what we found.

Question 5:

Using the 2009 CPS data we find, men are almost _____ percentage points more likely to earn at least $100,000.

9

Correct!

Question 6:

On average, we found that men earned roughly $_____ (round to the nearest thousand) more than women, which translates into about a _____ percentage (round to the nearest integer) earnings gap

19000, 43

Correct!

Continue

Gender earnings gap among younger workers

Now, let's see what happens when we focus on workers in the early stages of their careers. First, we need to load the data. Because the March 2009 extract is contained in an `.xlsx` file, we will the `read_xlsx` function provided by the `readxl` package. Load the package and read the file, note the file name is `cps09mar.xlsx`. (You need to get this right to move on, so we will check your code here.)

R Code | Start Over | Run Code | Submit Answer

```
library(readxl)
cps09mar <- read_xlsx("./data/cps09mar.xlsx")

```

You should be proud. Correct!

Next, let's filter down to workers who are between 25-34 years old. This will allow us to focus on individuals who (in all likelihood) have completed their educations but are still in the early stage of their careers. To do this we will need the `filter` function from the `dplyr` package. To parallel the analysis covered in the class slide deck, we will create a `gender` indicator using the `mutate` function from `dplyr`. We could easily do both things in one chunk but we will split them up.

The filtering action will take the original data set, `cps09mar`, filter on the age restriction, and create a new data set containing only the younger workers. The syntax goes like this:

```
newdataset <- olddataset %>%
  filter(age <= _____,
         age >= _____)
```

Note that we have used the `%>%`, the "pipe", to pass the data to the `filter` function. The pipe is very useful because it allows you to code a sequence of operations that is easy to read and interpret.

Now, use this sample code to complete the filtering operation to create a new data set called `cps09mar_2534` containing only the younger workers.

R Code | Start Over | Run Code | Submit Answer

```
cps09mar_2534 <- cps09mar %>%
  filter(age <= 34,
         age >= 25)
```

I couldn't have done it better myself. Correct!

The next step is, strictly speaking, not necessary because `cps09mar` already contains a `female` indicator, but as we said, we want to parallel the analysis in the deck. Here is one way to create the new variable `gender` indicator using the `mutate` and `case_when` functions from `dplyr`. Then we will add the new variable to `cps09mar_2534`.

```
current_data_set <- current_data_set %>%
  mutate(newvariable = case_when(_____ == 1 ~ 'Female',
                                 _____ == 0 ~ 'Male'))
```

Use this sample code to complete the mutating operation that will create a new `gender` variable that takes on the values `Female` and `Male` and adds it to `cps09mar_2534`.

R Code | Start Over | Run Code | Submit Answer

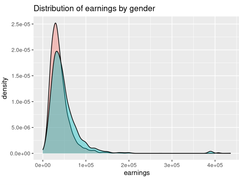
```
cps09mar_2534 <- cps09mar_2534 %>%
  mutate(gender = case_when(female == 1 ~ "Female",
                            female == 0 ~ "Male"))
```

Bravo! Correct!

Ok, now we are ready to do some work. First, let's replicate the earnings distributions shown in Figure 4 of the deck for 25-34 year-olds. Here is all the code you need. Just fill in the blank with the earnings distribution object name and see what you get.

R Code | Start Over | Run Code | Submit Answer

```
earnings_dist_fvm <- ggplot(cps09mar_2534,
                            aes(x=earnings, group = gender, fill = gender)
  geom_density(adjust=1.5, alpha = 0.4) +
  labs(title="Distribution of earnings by gender")
earnings_dist_fvm
```



Nice job! Correct!

Question 7:

Compared with the earnings distributions in Figure 4, these show (more/less) _____ overlap.

more

Correct!

Next, let's compute the gap in average earnings in dollar and percentage terms. Use code on slide 49 to carry out this calculation.

R Code | Start Over | Run Code

```
earnings_bar <- cps09mar_2534%>%
  group_by(gender) %>%
  summarise (avg_earnings = mean(earnings))
earnings_bar
```

gender <chr>	avg_earnings <dbl>
female	39050.74
male	49145.72

2 rows

Question 8:

The average gender earnings gap among 25-34 year-olds is $_____ (round to the nearest hundred), which is about $_____ smaller than the gap for all workers. (Use your first answer to Part B Q6.)

10100, 8900

Correct!

Question 9:

The dollar gap among 25-34 year-olds translates into a _____ percent (round to the nearest integer), which is about _____ percentage points smaller than the gap for all workers.

28, 17

Correct!

We'll finish this exercise by calculating the gender gap in the likelihood of earning six figures among 25-34 year-olds. Use the code on slide 38 to complete this code chunk. This calculation may take a minute to run, but reach out to Abbi if you have issues getting the answer to finish calculating.

R Code | Start Over | Run Code

```
six_figs_fvm <- _____ %>%
  _____(_____) %>%
  _____(six_figs_shrs = mean(earnings >= _____))
print(_____)
```

Question 10:

For 25-34 year-olds, the gender gap in the likelihood of earning at least $100,000 is only _____ percentage points. Answer to 2 decimal places, an example answer would read "1.12" percentage points.

3.54

Correct!

Continue

Previous Topic | Next Topic