Problem Set 2: Data Transformation

Due: 5pm on Wednesday, February 14.

Student identifer: 5956

- Use this .qmd template to complete the problem set
- In Canvas, you will upload the PDF produced by your .qmd file
- Put your identifier above, not your name! We want anonymous grading to be possible

This problem set draws on the following paper.

> England, Paula, Andrew Levine, and Emma Mishel. 2020. Progress toward gender equality in the United States has slowed or stalled, PNAS 117(13):6990–6997.

A note about sex and gender. As we have discussed in class, sex typically refers to categories assigned at birth (e.g., female, male). Gender is a performed construct with many possible values: man, woman, nonbinary, etc. The measure in the CPS-ASEC is "sex," coded male or female. We will use these data to study sex disparities between those identifying as male and female. The paper at times uses "gender" to refer to this construct.

1. Data analysis: Existing question

20 points. Reproduce Figure 1 from the paper.

Visit cps.ipums.org to download data from the 1962–2023 March Annual Social and Economic Supplement. Include these variables in your cart: sex, age, asecwt, empstat.

To reduce extract size, select cases to those ages 25–54. Before submitting your extract, we recommend changing the data format to "Stata (.dta)" so that you get value labels.

> Tip
>
> Look ahead: you will later study a new outcome of your own choosing. You could add it to your cart now if you want.

On your computer, analyze these data.

- filter to `asecwt > 0` (see paper footnote on p. 6995 about negative weights)
- mutate to create an `employed` variable indicating that `empstat == 10 | empstat == 12`
- mutate to convert `sex` to a factor variable using `as_factor`
- group by `sex` and `year`
- summarize the proportion employed: use `weighted.mean` to take the mean of `employed` using the weight `asecwt`

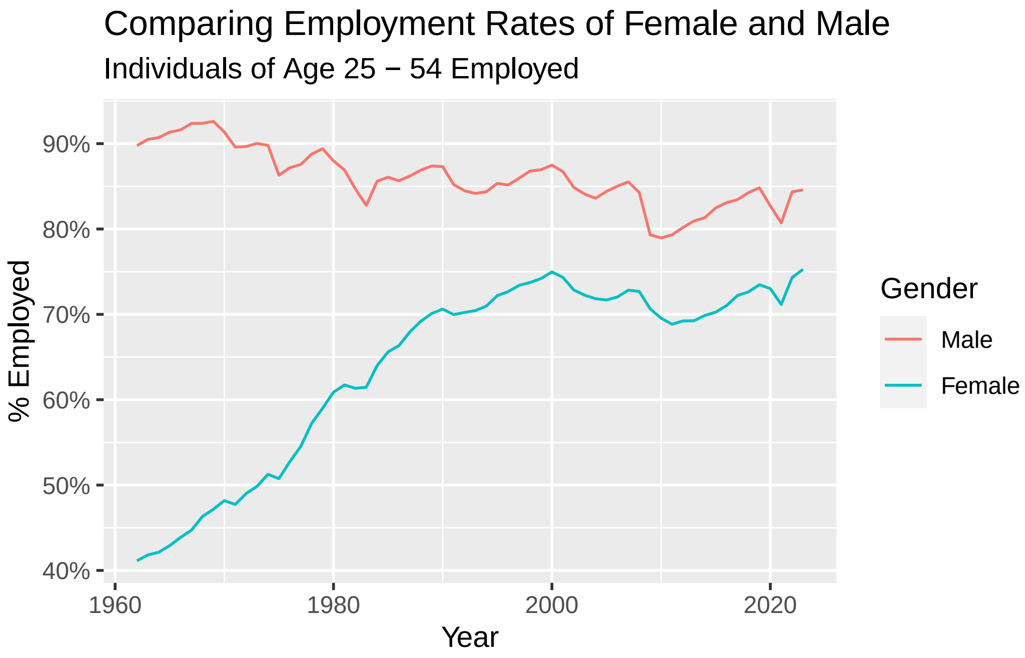
Your figure will be close but not identical to the original. Yours will include some years that the original did not. Feel free to change aesthetics of the plot, such as the words used in labels. For example, it would be more accurate to the data to label the legend "Sex" with values "Male" and "Female."

```r
library(tidyverse)
library(scales)
library(haven)

data <- read_dta("cps_00002.dta")
```

```r
# Treating the negative CPS weights as 0 to remain consistent with the paper.
data$asecwt[data$asecwt < 0] <- 0

filtered_df <- data |>
  filter(asecwt > 0) |>
  mutate(
    employed = (empstat == 10 | empstat == 12)
  ) |>
  mutate(sex = as.factor(sex)) |>
  group_by(sex, year) |>
  summarise(
    employment = weighted.mean(employed, asecwt)
  )
```

```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
```

```r
filtered_df |>
  ggplot(mapping = aes(x = year, y = employment, color = sex)) +
  labs(
    title = "Comparing Employment Rates of Female and Male",
    subtitle = "Individuals of Age 25 - 54 Employed",
    x = "Year",
  ) +
  scale_y_continuous(
    name = "% Employed",
    labels = label_percent()
  ) +
  scale_color_discrete(
    name = "Gender",
    labels = c("Male", "Female")
  ) +
  geom_line()
```



2. Reading questions

2.1 (3 points) The authors write that "change in the gender system has been deeply asymmetric." Explain this in a sentence or two to someone who hasn't read the article.

By the statement that "change in the gender system has been deeply asymmetric," the authors are making a claim that although there is an increase in women participation in the work force as full-time workers, there has not been a decrease in male employment nor an increase in male participation in family work. As women enter the work force, they are still tasked with managing family work and this prevents women from spending more time to close the gender gap in pay.

2.2 (3 points) The authors discuss cultural changes that could lead to greater equality. Propose a question that could (hypothetically) be included in the CPS-ASEC questionnaire to help answer questions about cultural changes.

A question can be: Approximately how many times have you taken time-off work for a child-related matter in one year? NA (no child), 0-5, 5-10, 10+

> Tip
>
> If you are not sure how to word a survey question, here are some examples from the American Time Use Survey, Current Population Survey, and General Social Survey.

2.3 (3 points) The authors discuss institutional changes that could lead to greater equality. Propose a question that could (hypothetically) be included in the CPS-ASEC questionnaire to help answer questions about institutional changes.

A possible question can be: When applying to jobs, most applications require the applicant to indicate their gender. To what degree do you fear your answer would impact your application? Strongly Agree, Agree, Disagree, Strongly Disagree, NA

2.4 (1 point) What was one fact presented in this paper that most surprised you?

In this paper, the authors discussed the stability in women hourly wage for those earning at the 10th, 20th, and 50th percentile while the wage for men decreased. Superficially, it seemed like equality is reached, but upon further investigation of figures 10 and 11, you can observe the lower wage that women had from the start relative to men at the same percentile. I found this surprising that even for lower-income jobs, the gender inequality is still present and even now, there is an income gap between men and women.

3. A new outcome

20 points. The CPS-ASEC has numerous variables. Pick another variable of your choosing. Add it to your cart in IPUMS, and visualize how that variable has changed over time for those identifying as male and female.

As in the previous plot, year should be on the x-axis and color should represent sex. The y-axis is up to you. You can examine something like median income, proportion holding college degrees, or the 90th percentile of usual weekly work hours. You can restrict to some subset if you want, such as those who are employed.

Your answer should include

- a written statement of what you estimated: the variable you chose, any sample restrictions you made, and how you summarized that variable
- a written interpretation of what you found
- code following style conventions
- your publication-quality visualization

Child-Related Absences Compared to Gender Analysis

In this section, I look to investigate further on the claim presented by the authors of the article that family responsibilities and childcare is usually left to females as opposed to males. Therefore, I selected the WHYABSNT variable for the group of ages 25 - 54. Among the possible responses for why the respondent was absent from the preceding week, I limited the sample to those who are employed and were absent due to childcare problems, family/personal obligations, and maternity/paternity leave. By doing so and analyzing the divide based on gender for employed, it provides insight on whether or not the larger share of family responsibilities is given to women and hinder women's ability to close the wealth gap.

```r
data3 <- read_dta("cps_00003.dta")

family_df <- data3 |>
  # The WHYABSNT variable was not tracked till the 1990s. Need to omit the period
  # when this data was not tracked.
  filter(
    (year >= 1990)
  ) |>
  # Filter to those who are employed
  filter(
    (empstat == 10 | empstat == 12)
  ) |>
  # Absent from work due to child care issues or family/personal obligation
  mutate(
    family = (whyabsnt == 07 | whyabsnt == 08 | whyabsnt == 09)
  ) |>
  mutate(sex = as.factor(sex)) |>
  group_by(sex, year) |>
  summarise(
    family_time = weighted.mean(family, asecwt)
  )
```

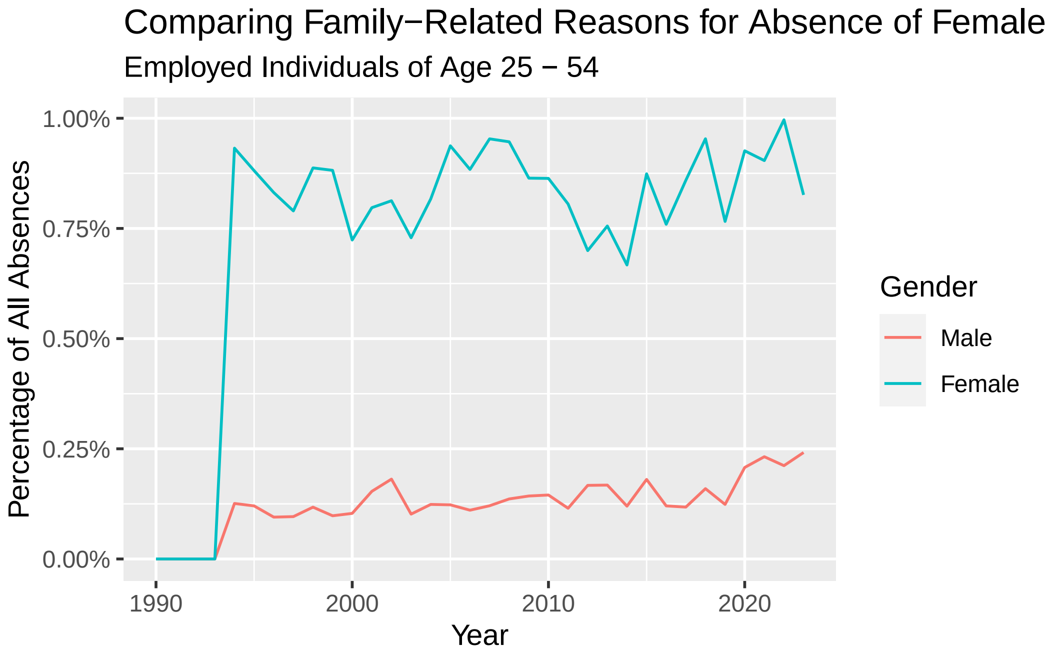
```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
```

```r
child_df <- data3 |>
  # The WHYABSNT variable was not tracked till the 1990s. Need to omit the period
  # when this data was not tracked.
  filter(
    (year >= 1990)
  ) |>
  # Filter to those who is employed
  filter(
    (empstat == 10 | empstat == 12)
  ) |>
  # Absent from work due to childcare and maternity leave
  mutate(
    family = (whyabsnt == 07 | whyabsnt == 09)
  ) |>
  mutate(sex = as.factor(sex)) |>
  group_by(sex, year) |>
  summarise(
    child_time = weighted.mean(family, asecwt)
  )
```

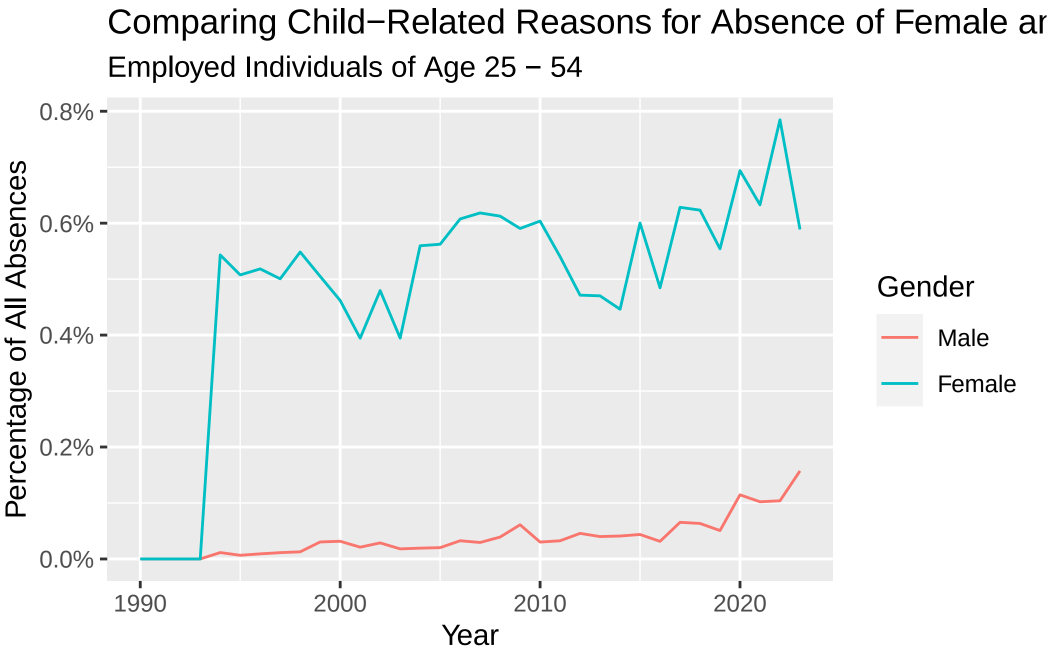
```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
```

Plotting the data

```r
family_df |>
  ggplot(mapping = aes(x = year, y = family_time, color = sex)) +
  labs(
    title = "Comparing Family-Related Reasons for Absence of Female and Male",
    subtitle = "Employed Individuals of Age 25 - 54",
    x = "Year",
  ) +
  scale_y_continuous(
    name = "Percentage of All Absences",
    labels = label_percent()
  ) +
  scale_color_discrete(
    name = "Gender",
    labels = c("Male", "Female")
  ) +
  geom_line()
```



```r
child_df |>
  ggplot(mapping = aes(x = year, y = child_time, color = sex)) +
  labs(
    title = "Comparing Child-Related Reasons for Absence of Female and Male",
    subtitle = "Employed Individuals of Age 25 - 54",
    x = "Year",
  ) +
  scale_y_continuous(
    name = "Percentage of All Absences",
    labels = label_percent()
  ) +
  scale_color_discrete(
    name = "Gender",
    labels = c("Male", "Female")
  ) +
  geom_line()
```



I created two graphs, one for overall family-related reasons and more specific child-related reasons, because in the article, the authors mentioned specifically about the child-related responsibilities assigned by societal norms to women that hinders the ability of women to close the gender gap in pay. The both graphs shows that across all years, of all the reasons of absence, family-related absences for women are more than 3 times that of men. However, there is a slow increase in family-related absences and child-related for males. This supports the authors' claims regarding the more responsibility placed on women to take care of the family. These two graphs reveal that slow progress towards equality that is still deeply symmetric.

After analyzing this variable, other focus questions that can stem from this analysis would be the same set of data but divided based on a certain income range to see if income impact the percentage of family-related absences among the two genders and the proportion of individuals who are unemployed or not participating in the work force due to child-related reasons.

Grad question: Ratio and difference

> This question is required for grad students. It is optional for undergrads, and worth no extra credit.

20 points. The figures above visualize a summary statistic for each subgroup: male and female. Another way to visualize this is with the ratio (female statistic / male statistic) or difference of the two (female statistic - male statistic).

For your own question, produce two new visualizations:

- one showing the female / male ratio over time
- one showing the female - male difference over time

Which do you find easier to interpret, and why?

> Tip
>
> It is likely that your code for the previous parts produced one column with estimates and another column `sex` containing the values `male` and `female`. One way to calculate a ratio and difference is to reshape the data so that there is one row for each year, containing both the `male` and `female` estimates. You can do this with `pivot_wider` where the names come from `sex` and the values come from the column containing your estimates. Then you can `mutate()` to create the ratio and difference.