Problem Set 01 Solution

Professor Bradley Warner

June, 2022

Documentation:

We used all the resources available to instructors from the authors of Modern Dive.

Setup

```
library(dplyr)
library(ggplot2)
library(readr)
```

```
nc <- read_csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vTm2WZwNBoQdZhMgot7urbtu8eG7tzAq-60ZJsQ
```

```
glimpse(nc)
```

Exercise 1

Looking at the output of the `glimpse()` function, has does `R` classify the variable `mature`? Same question for the variable `gained`. (Answer with text)

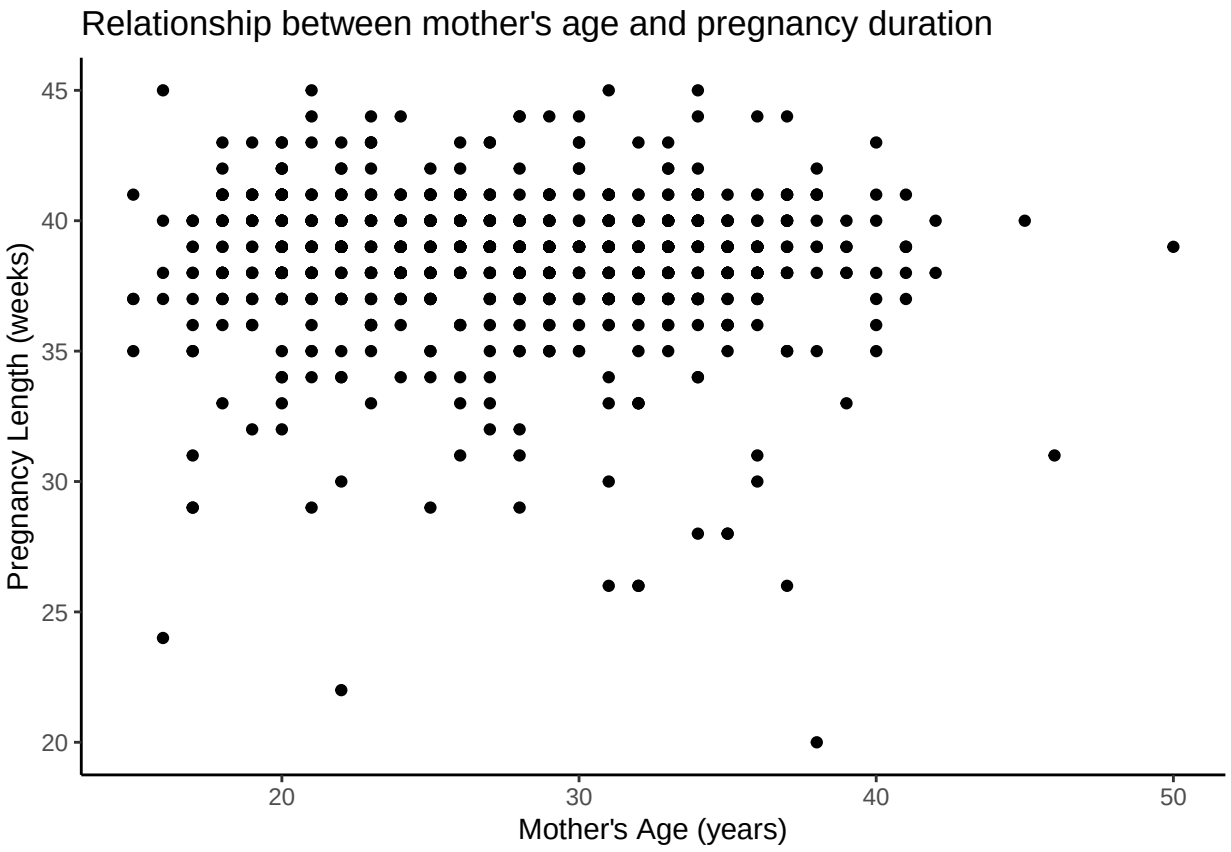
Answer: mature is listed as (character), and visits is an double precision

Exercise 2

Make a graph showing a mother's age `mage` on the x axis and the variable `weeks` on the y axis. Include axis labels with measurement units, and a title. (`R` code and output)

Answer:

```
ggplot(data = nc, aes(x = mage, y = weeks))+
  geom_point() +
  labs(x = "Mother's Age (years)", y = "Pregnancy Length (weeks)",
       title = "Relationship between mother's age and pregnancy duration") +
  theme_classic()
```



Exercise 3

Study the code below, and the resulting graphical output. Note that we added a new argument of `color = premie` and `shape = marital` **inside** the `aes`thetic mapping. The variable `premie` indicates whether a birth was early (premie) or went full term and `marital` represents the marital status of the mother. Please answer with text:

A. What did adding the argument `alpha = 0.3` accomplish?

Answer: It makes the data points more transparent so we can see overlapping points.

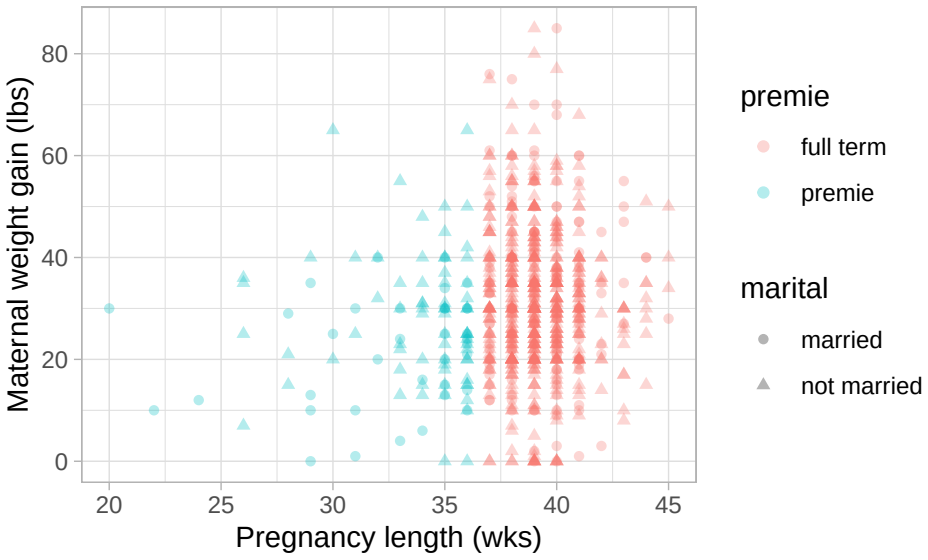
B. How many **variables** are now displayed on this plot?

Answer: 4 variables are shown

C. What appears to (roughly) be the pregnancy length cutoff for classifying a newborn as a "premie"" versus a "full term".

Answer: anywhere between 36 and 38 seems reasonable

```
ggplot(data = nc, aes(x = weeks, y = gained, color = premie, shape = marital))+
  geom_point(alpha=0.3) +
  labs(x = "Pregnancy length (wks)", y = "Maternal weight gain (lbs)") +
  theme_light()
```



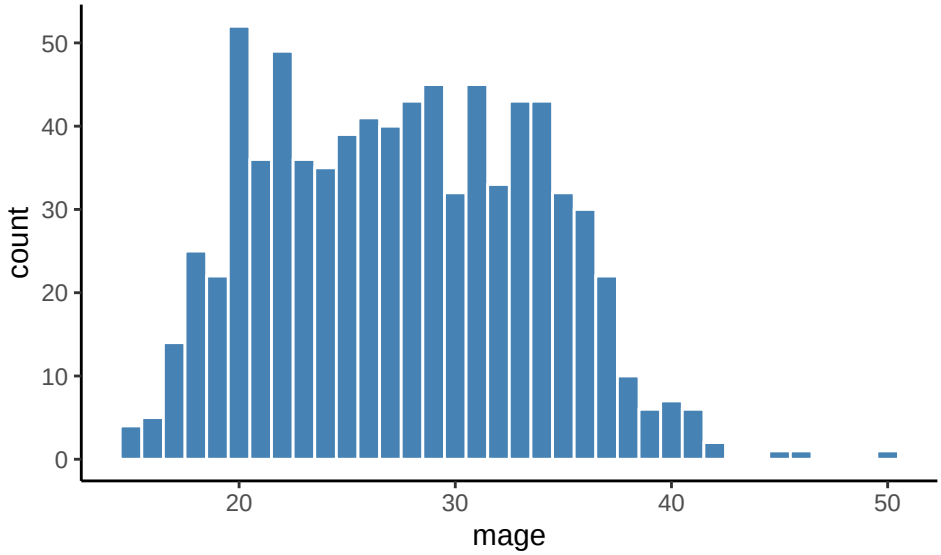
Exercise 4

Make a new scatterplot that shows a mothers age on the x axis (variable called `mage`) and pregnancy length on the y axis (`weeks`). Color the points on the plot based on the gender of the resulting baby (variable called `gender`). Change the shape based on the marital status of the mother (variable called `marital`). Use `alpha` value of 0.4 and use the classic theme. There should not appear to be any strong relationship between a mother's age and the pregnancy length. (`R` code and output)

Answer:

Exercise 5

```
ggplot(data = nc, aes(x = mage))+
  geom_histogram(binwidth = 1, color = "white", fill = "steelblue") +
  theme_classic()
```



Inspect the histogram of the `mage` variable. Answer each of the following with **text**.

A. The y axis is labeled **count**. What is specifically being counted in this case? Hint: think about what each case is in this data set.

Answer: the number of mothers whose age fall into each bin specified on the histogram

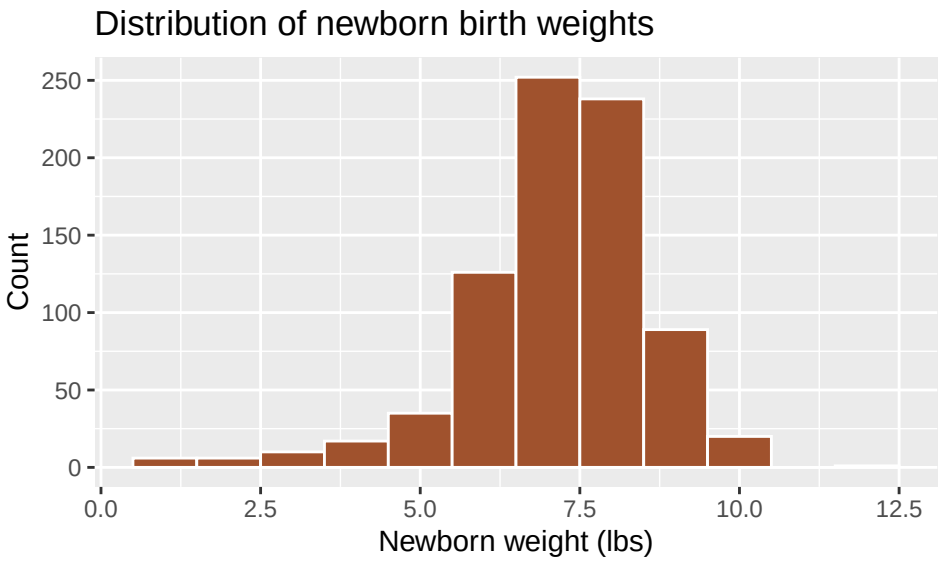
B. What appears to be roughly the average mother's age in years?

Answer: 27 weeks. . . 26, 28, 29, 30 or 31 also acceptable

Exercise 6

Make a histogram of the birth `weight` of newborns (which is in lbs), including a title and axis labels. (`R` code and output for answer)

```
ggplot(data = nc, aes(x =  weight)) +
  geom_histogram(binwidth = 1, color = "white", fill = "sienna") +
  labs(x = "Newborn weight (lbs)", y = "Count", title = "Distribution of newborn birth weights")
```

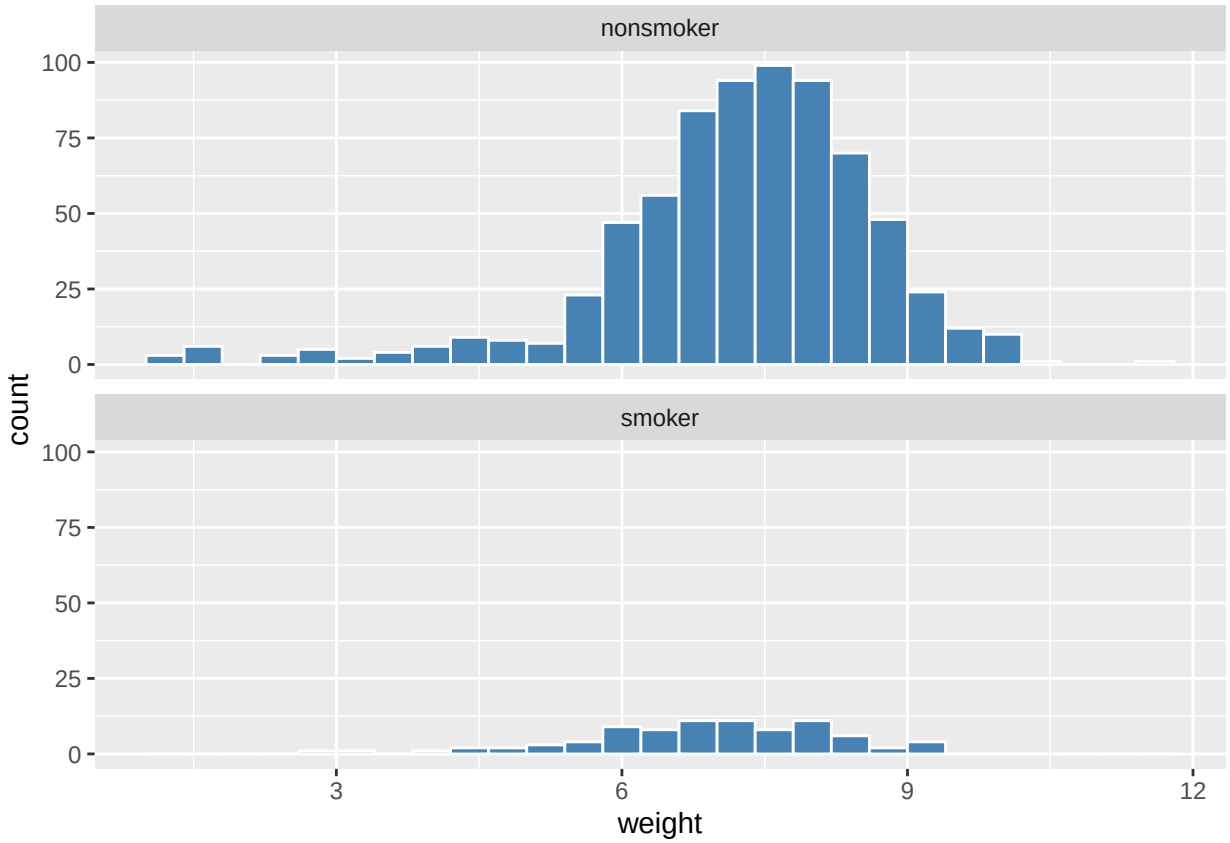


Exercise 7

Make a histogram of newborn birth `weight` split by `habit`, the smoking status of the mother. Set the binwidth to 0.5. There are many fewer mothers who smoke. Do nonsmokers appear to have babies with a slightly larger median birth weight? (Text and `R` code and output for answer)

Answer:

```
ggplot(data = nc, aes(x = weight)) +
  geom_histogram(binwidth = 0.4, color = "white", fill = "steelblue") +
  facet_wrap(~ habit, ncol = 1)
```



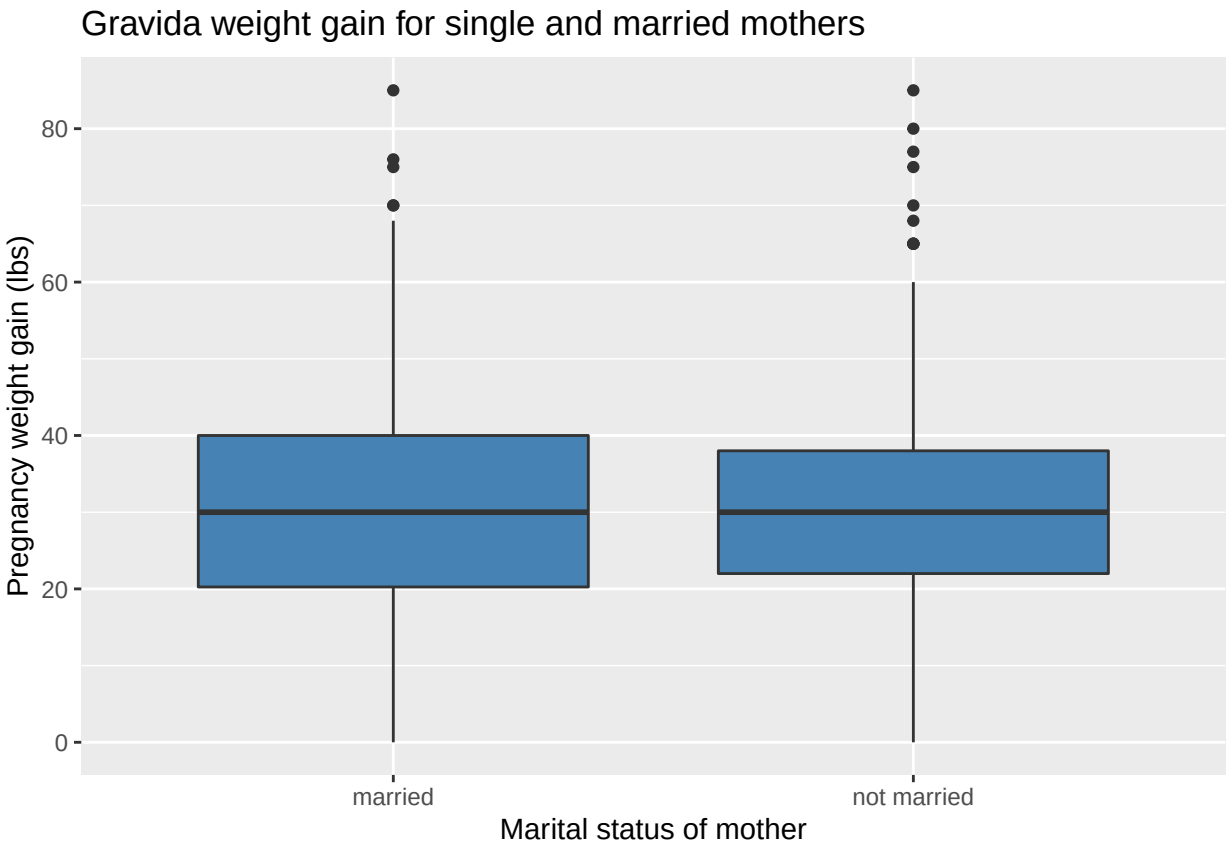
Answer: The babies of nonsmokers tend to have a slightly higher median birth weight

Exercise 8

Make a boxplot of the weight `gained` by mothers, split by the marital status of the mothers (`marital`). Include axis labels and a title on your plot. Is the variation in weight gained during pregnancy larger for married or single mothers? (Text and `R` code and output)

Answer:

```
ggplot(data = nc, aes(x = marital, y = gained)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "Marital status of mother", y = "Pregnancy weight gain (lbs)",
       title = "Gravida weight gain for single and married mothers")
```



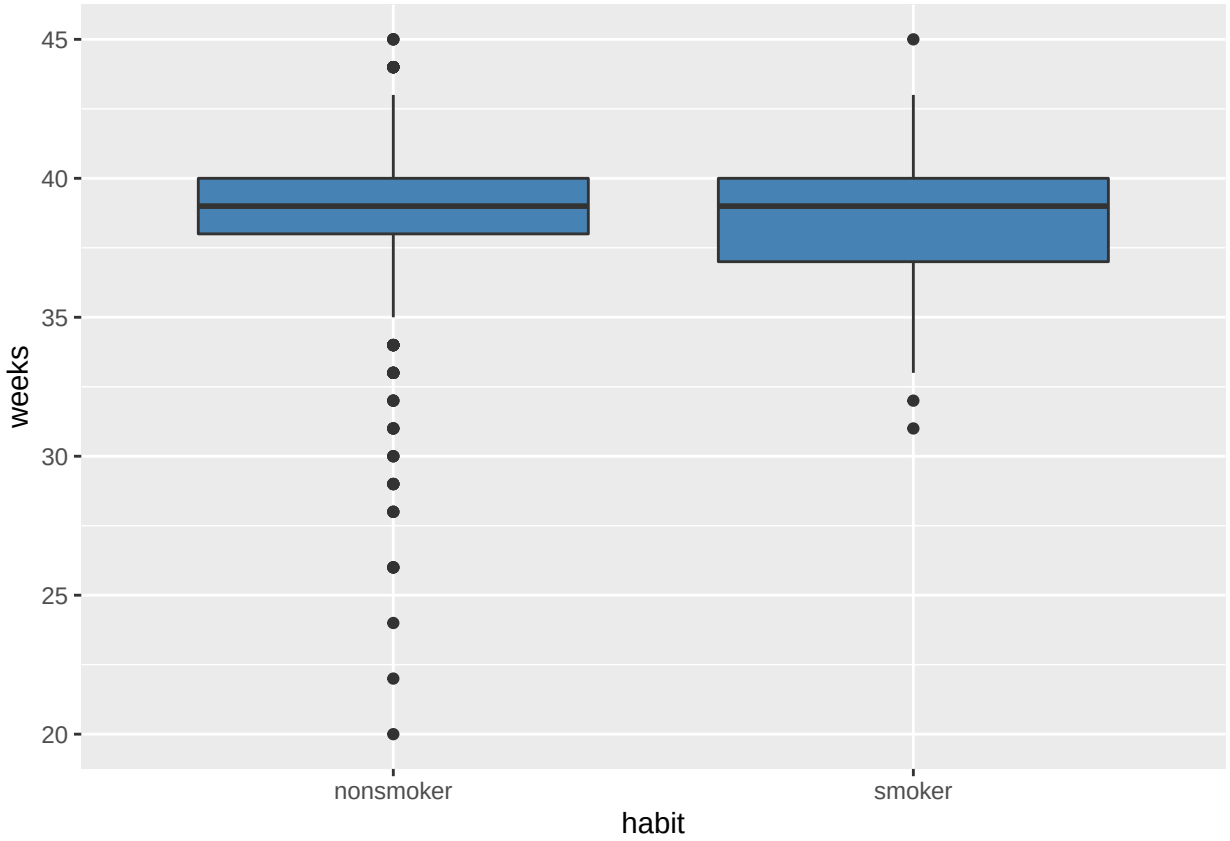
Answer: The variation is slighly greater for married mothers

Exercise 9

Make a boxplot of pregnancy duration in `weeks` by smoking `habit`. Are there more outliers in the duration of pregnancy for smokers or non-smokers? (Text and `R` code and output for answer)

Answer:

```
ggplot(data = nc, aes(x = habit, y = weeks)) +
  geom_boxplot(fill = "steelblue")
```



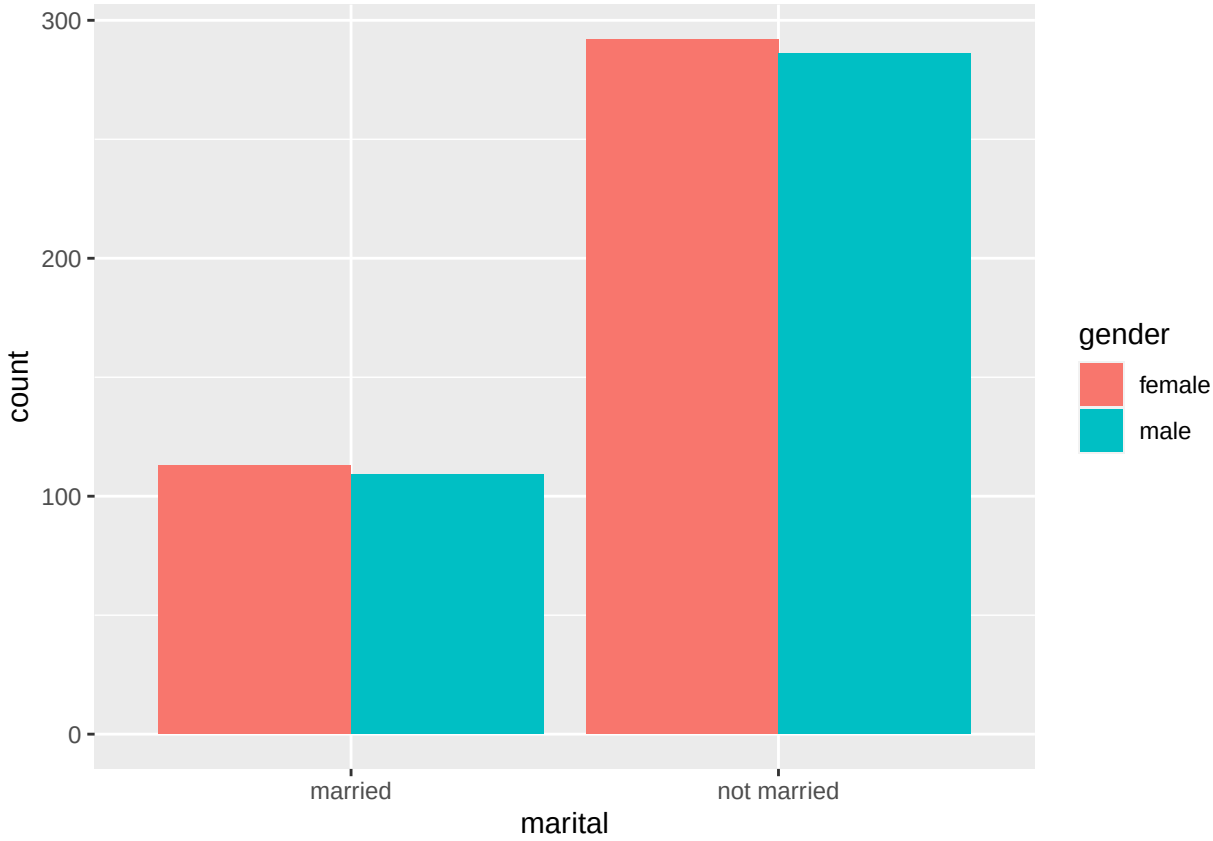
Answer: The nonsmokers have more outliers, premature babies. We are tempted to assign a cause such as the mothers are more concerned with their pregnancies and would not risk smoking. But this is only a conjecture that we have to be careful in not making.

Exercise 10

Make a side-by-side barplot of marital status `marital` and gender of the baby `gender`. Are female babies more common in both married and single mothers? (Text and `R` code and output for answer)

Answer:

```
ggplot(data = nc, aes(x = marital, fill = gender)) +
  geom_bar(position = "dodge")
```



Answer: Female babies are more common in both married and single mothers.

Documenting software

- File creation date: 2022-06-18
- R version 4.1.3 (2022-03-10)
- `dplyr` package version: 1.0.9
- `ggplot2` package version: 3.3.6
- `readr` package version: 2.1.2