Problem Set 02 Solutions

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

```
nc <- read_csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vTm2WZwNBoQdZhMgot7urbtu8eG7tzAq-6OZJsQ
```

```
glimpse(nc)
```

Exercise 1

What type of variable is R considering the variable `habit` to be? What variable type is `visits`? (answer with text)

Answer: habit is listed as (character), and visits is an integer

```
ggplot(data = nc, aes(x = weeks, y = weight)) +
  geom_point() +
  labs(x = "Length of pregnancy (in weeks)", y = "Birth weight of baby (lbs)",
       title = "Relationship between pregnancy duration and newborn weight")
```

Exercise 2

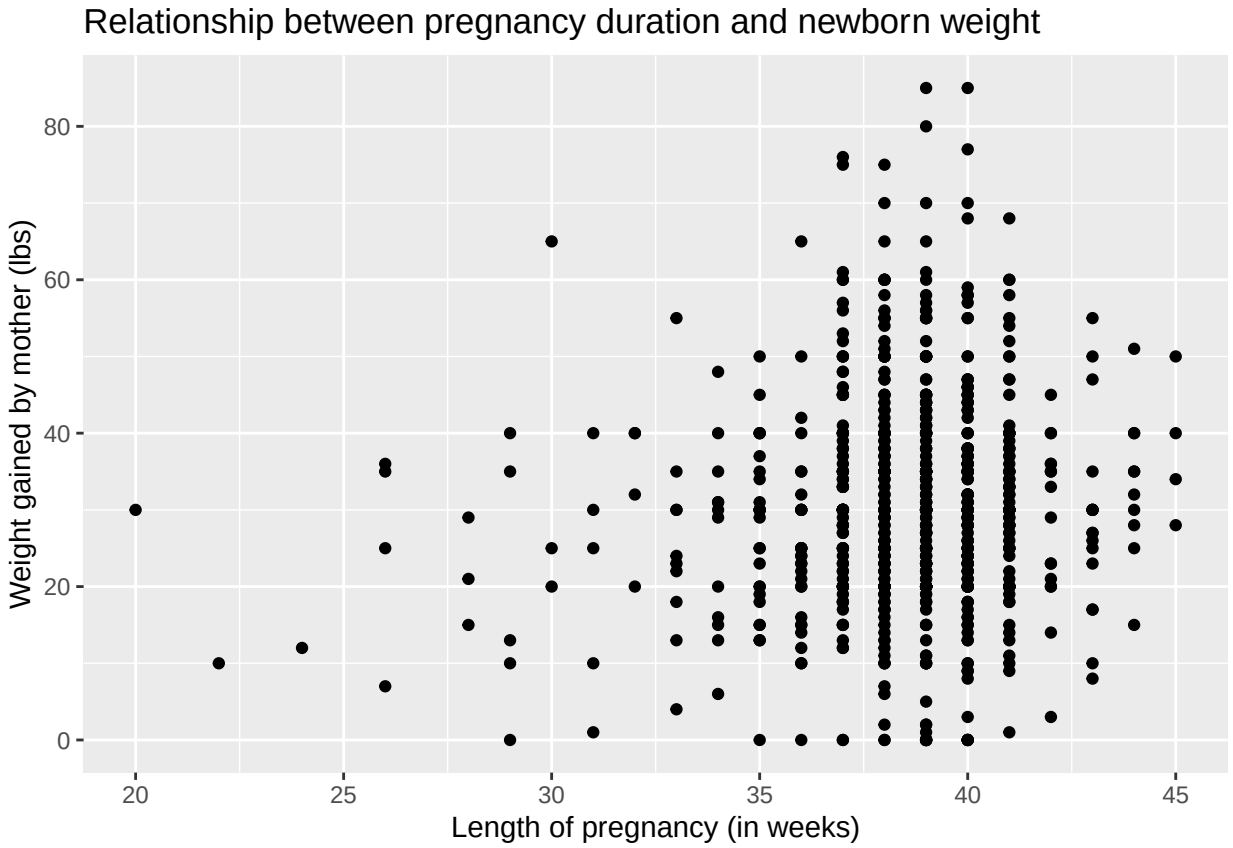
1. Is there a positive or negative relationship between these variables? (text only to answer)

Answer: positive

Exercise 3

Make a graph showing `weeks` again on the x axis and the variable `gained` on the y axis (the amount of weight a mother gained during pregnancy). Include axis labels with measurement units, and a title. (code only to answer)

```
ggplot(data = nc, aes(x = weeks, y = gained))+
  geom_point() +
  labs(x = "Length of pregnancy (in weeks)", y = "Weight gained by mother (lbs)",
       title = "Relationship between pregnancy duration and newborn weight")
```



Exercise 4

Study the code below, and the resulting graphical output.

A. What did adding the argument `color = premie` accomplish?

Answer: the color of the points corresponds to whether a birth was a premie or a full term birth

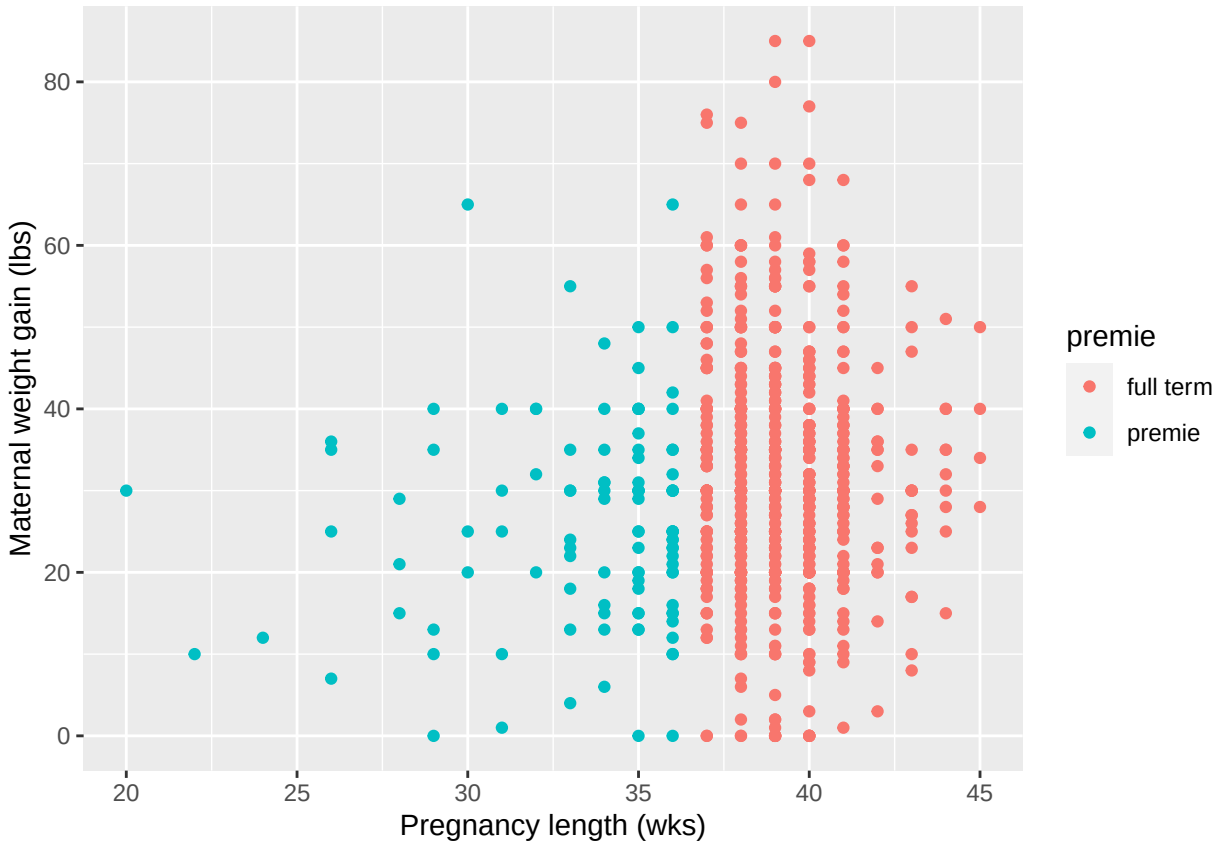
B. How many **variables** are now displayed on this plot?

Answer: 3 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))+
  geom_point() +
  labs(x = "Pregnancy length (wks)", y = "Maternal weight gain (lbs)")
```

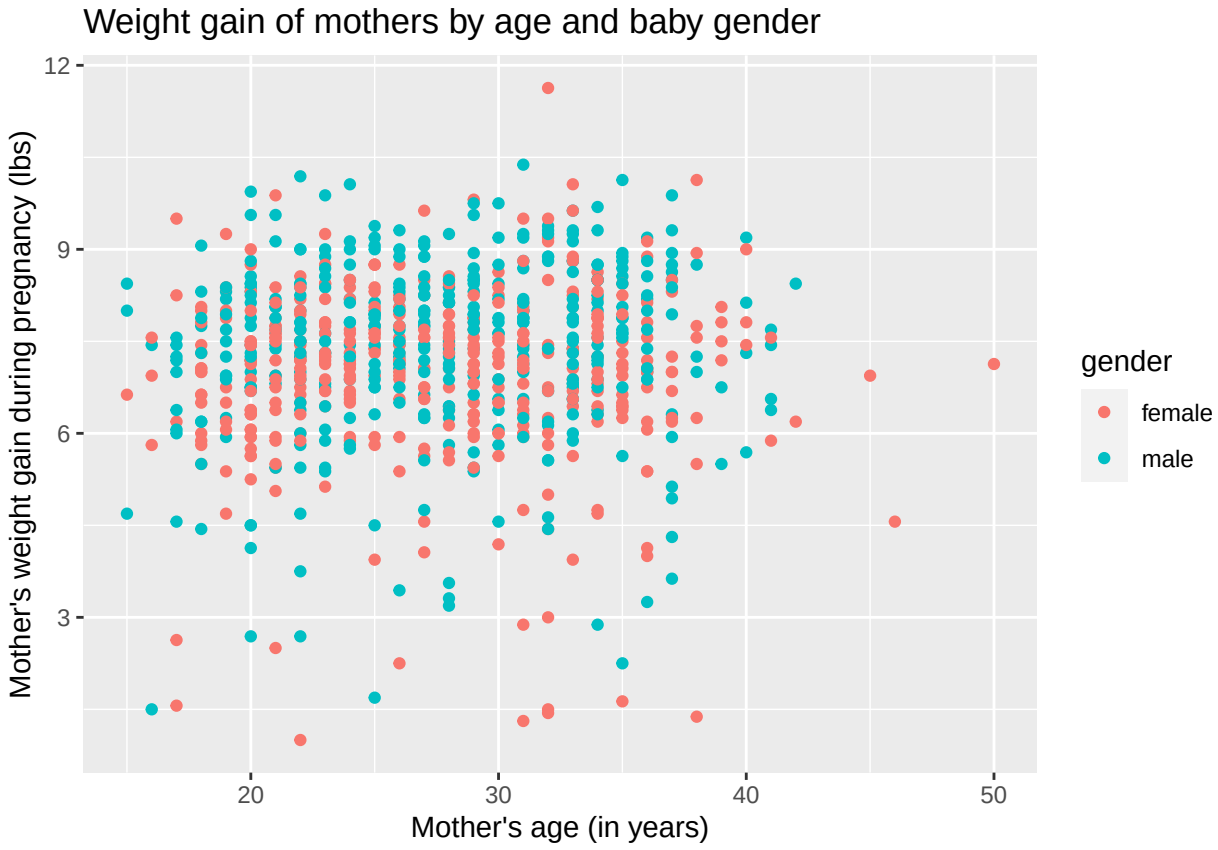


Exercise 5

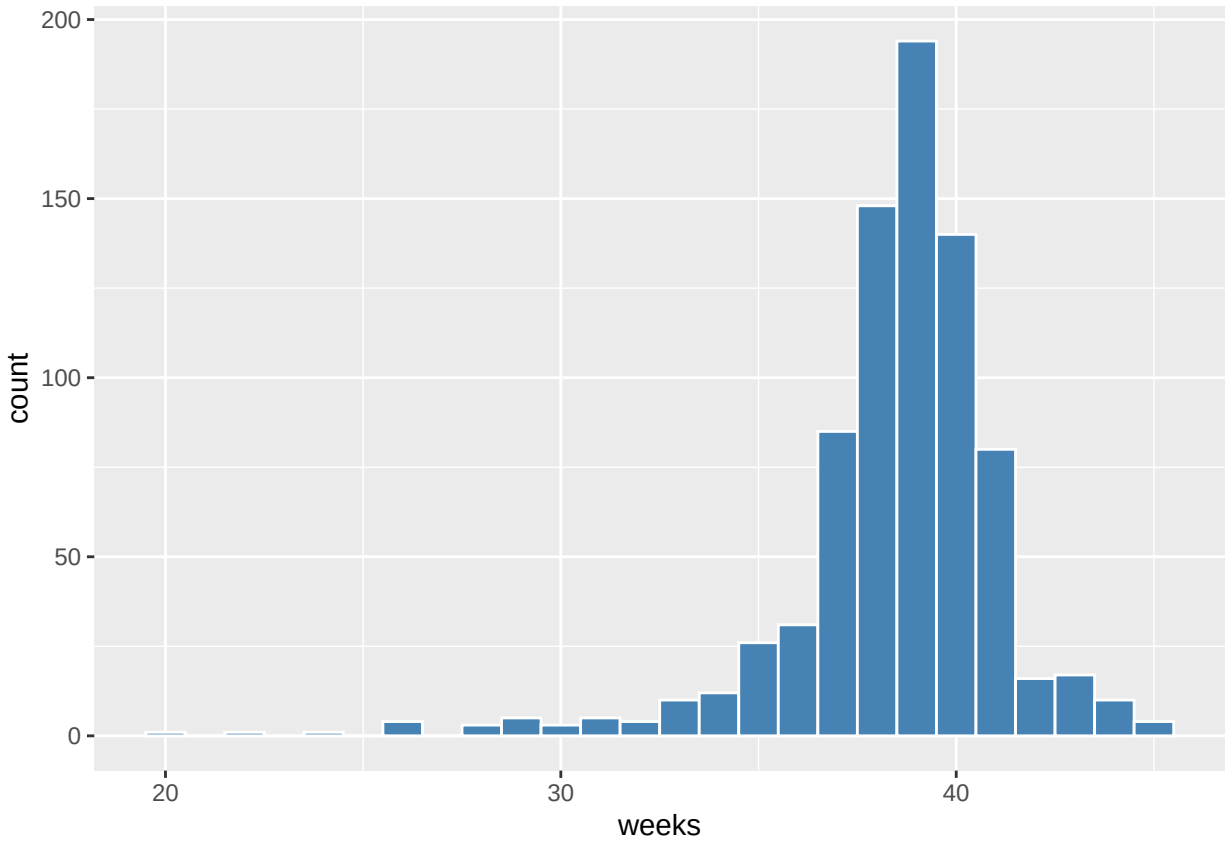
Make a new scatterplot that shows a mothers age on the x axis (variable called `mage`) and birth weight of newborns on the y axis (`weight`). Color the points on the plot based on the gender of the resulting baby (variable called `gender`). Does there appear to be any strong relationship between a mother's age and the weight of her newborn? (code and text to answer)

Answer: there does not appear to be any strong relationship between the mother's age and the weight of her newborn...the points are in a big blob

```
ggplot(data = nc, aes(x = mage, y = weight, color = gender)) +
  geom_point() +
  labs(x = "Mother's age (in years)", y = "Mother's weight gain during pregnancy (lbs)",
       title = "Weight gain of mothers by age and baby gender")
```




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



Exercise 6

Inspect the histogram of the `weeks` 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 births that fall into each bin specified on the histogram

B. What appears to be roughly the average lenght of pregnancies in weeks?

Answer: 39 weeks. . . 37, 38, or 40 also acceptable

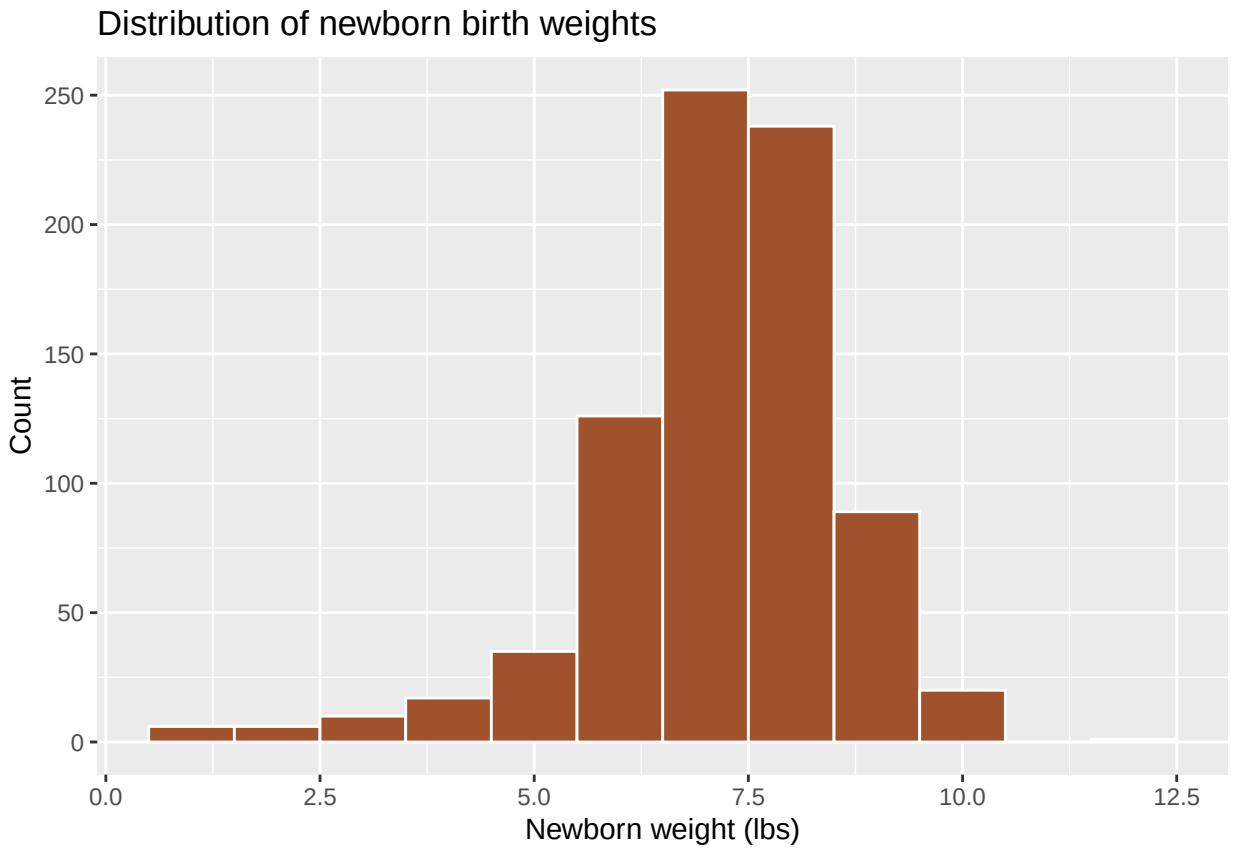
C. If we changed the binwidth to 100, how many bins would there be? Roughly how many cases would be in each bin?

Answer: there would be one bin. It would contain every case. . . so 800 births would be in it

Exercise 7

Make a histogram of the birth `weight` of newborns (which is in lbs), including a title and axis labels. (code only to 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")
```

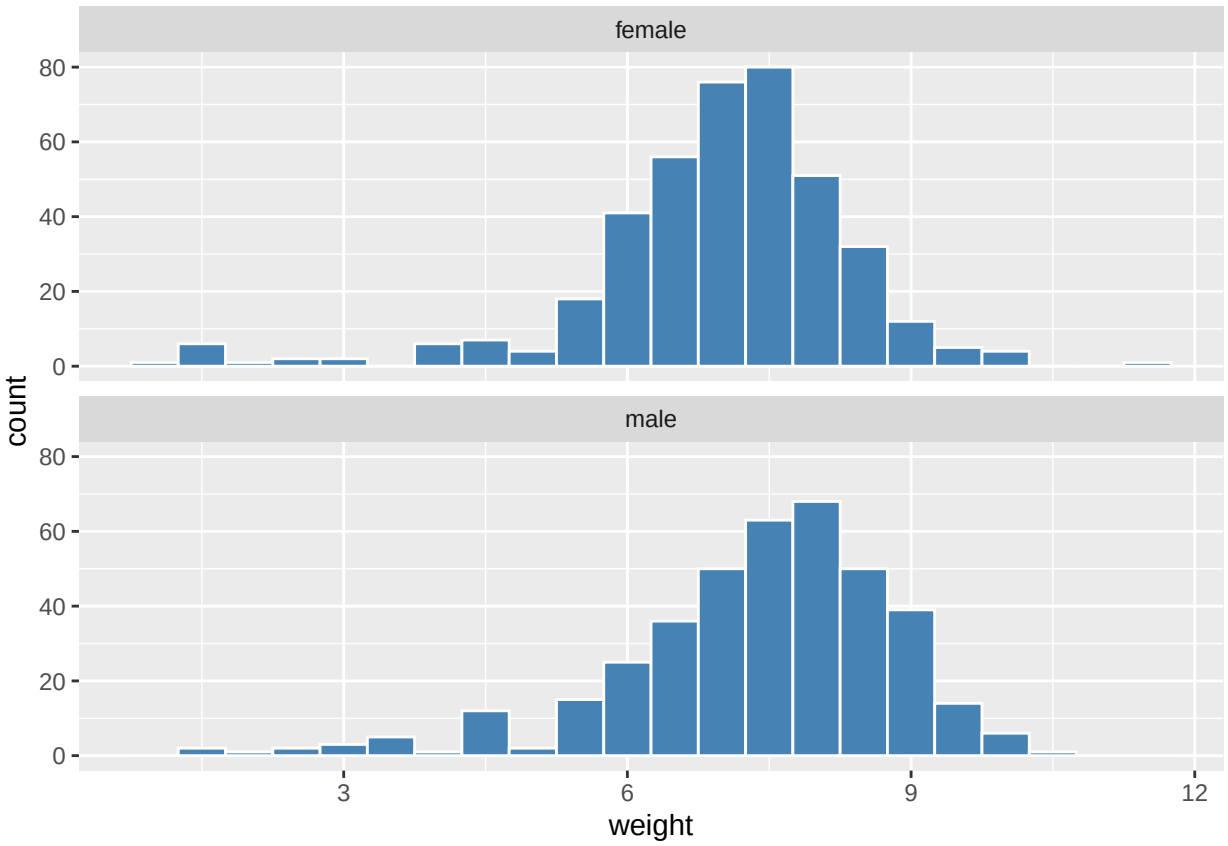


Exercise 8

Make a histogram of newborn birth `weight` split by `gender` of the child. Set the binwidth to 0.5. Which gender appears to have a slightly larger average birth weight? (code and text to answer)

Answer: Males have a slightly higher average birth weight

```
ggplot(data = nc, aes(x = weight)) +
  geom_histogram(binwidth = 0.5, color = "white", fill = "steelblue") +
  facet_wrap(~ gender, ncol = 1)
```

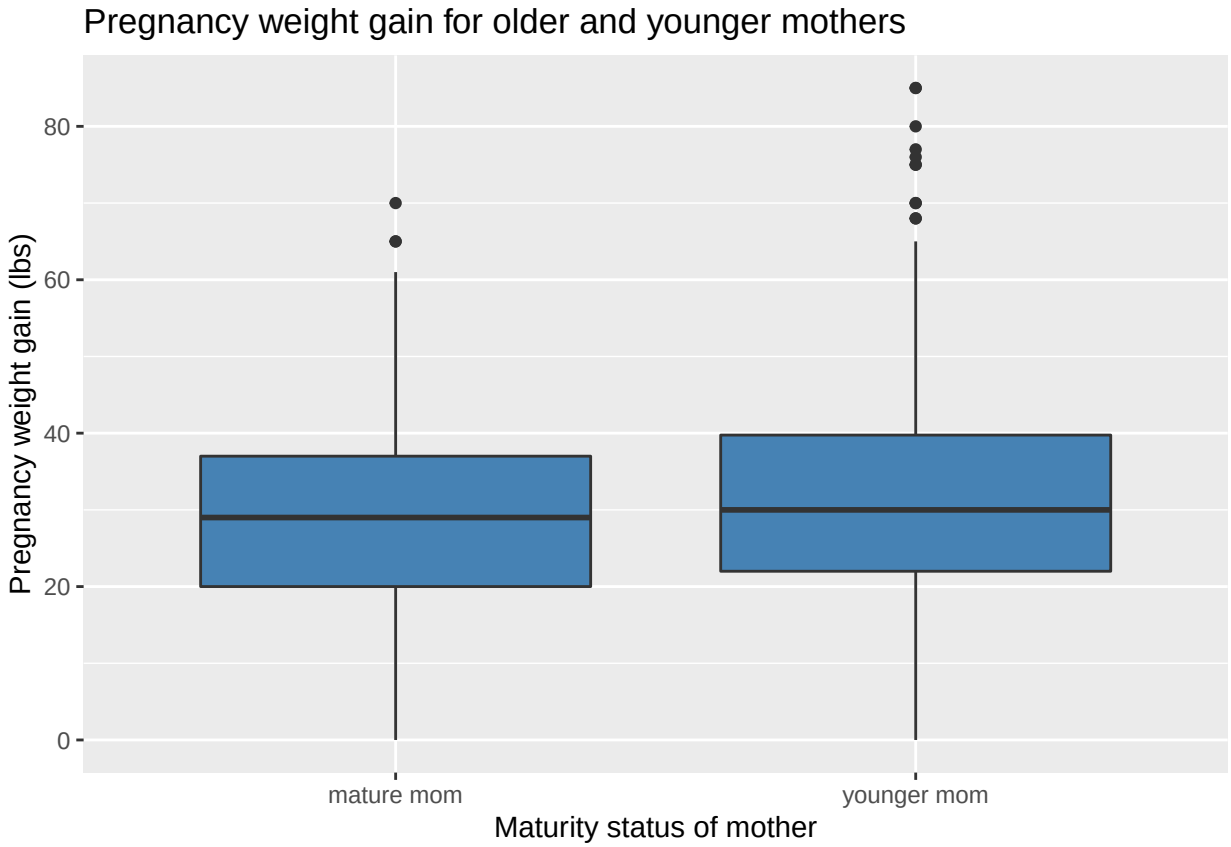


Exercise 9

Make a boxplot of the weight `gained` by moms, split by the maturity status of the mothers (`mature`). Include axis labels and a title on your plot. Is the **median** weight gain during pregnancy larger for younger or older moms? (text and code)

Answer it is slighly greater for younger moms

```
ggplot(data = nc, aes(x = mature, y = gained)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "Maturity status of mother", y = "Pregnancy weight gain (lbs)",
       title = "Pregnancy weight gain for older and younger mothers")
```

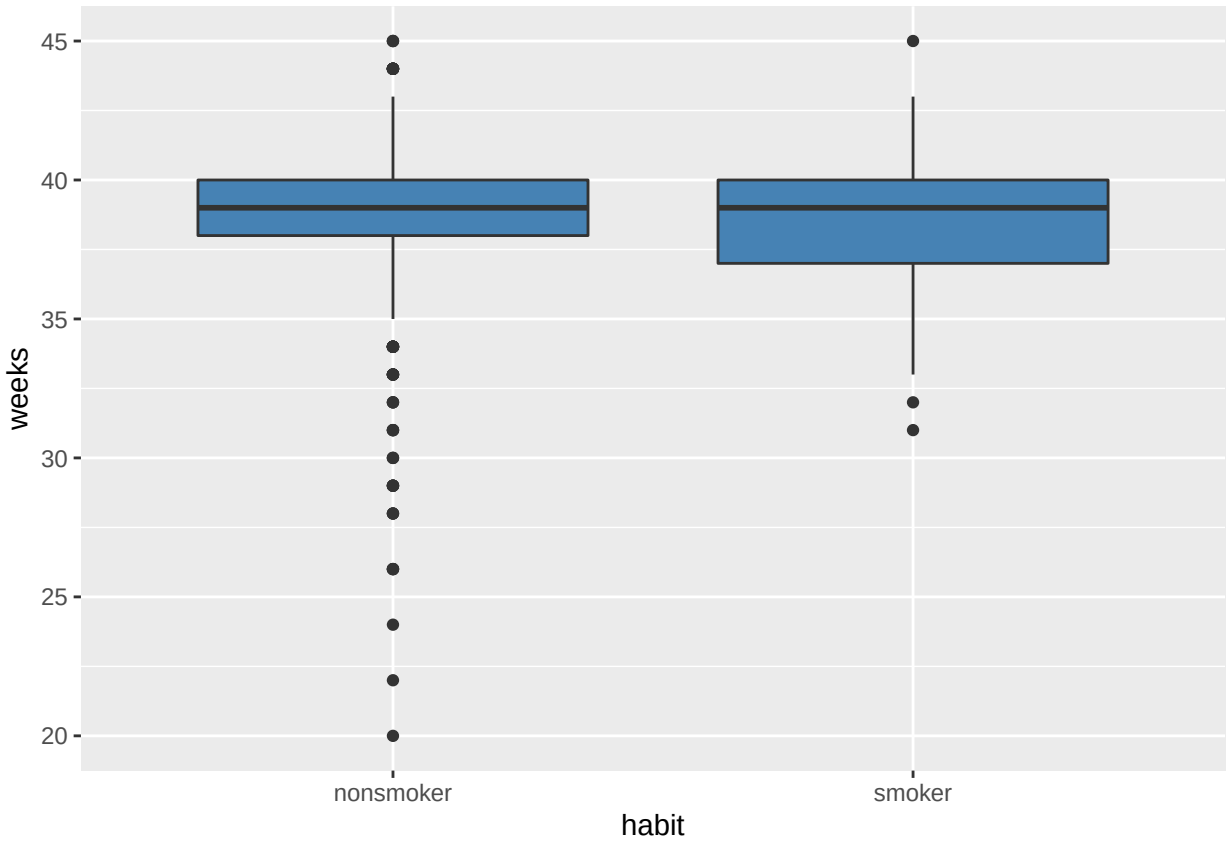


Exercise 10

Make a boxplot of pregnancy duration by smoking `habit`. Is the duration of pregnancy more **variable** for smokers or non-smokers? (i.e. which group has the greater spread for the variable `weeks`?). (code and text to answer)

Answer: based on the RANGE, pregnancy duration is more variable for nonsmokers... based on the IQR pregnancy duration is more variable for smokers

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



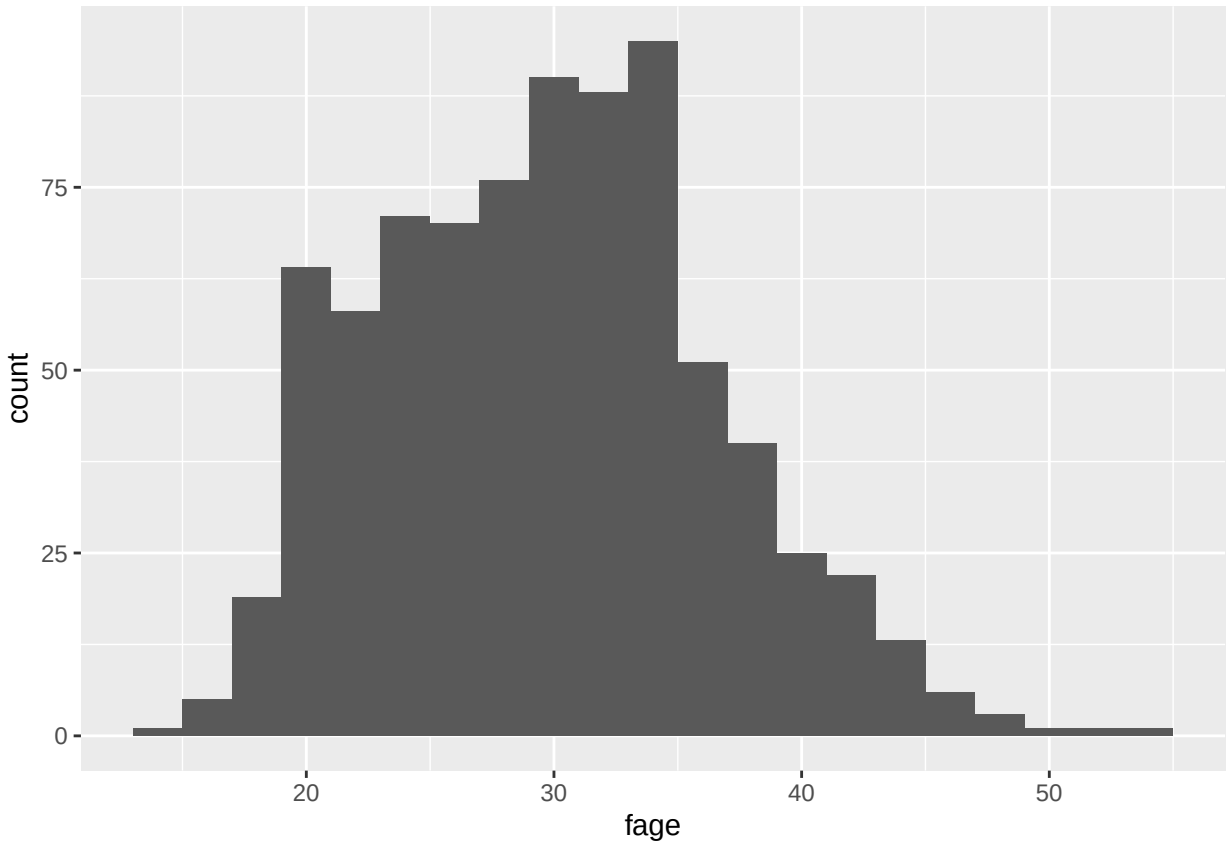
Independent Practice

For the following, you need to determine which type of plot to use, **make the plot**, and answer any questions with **text**. There is a table at the end of this document that can help you determine which plot to use, given the question/types of variables.

Exercise 11

Using a data visualization, visually assess: Is the variable for father's age (`fage`) symmetrical, or does it have a skew?

```
ggplot(data = nc, aes(x = fage)) +
  geom_histogram(binwidth = 2)
```



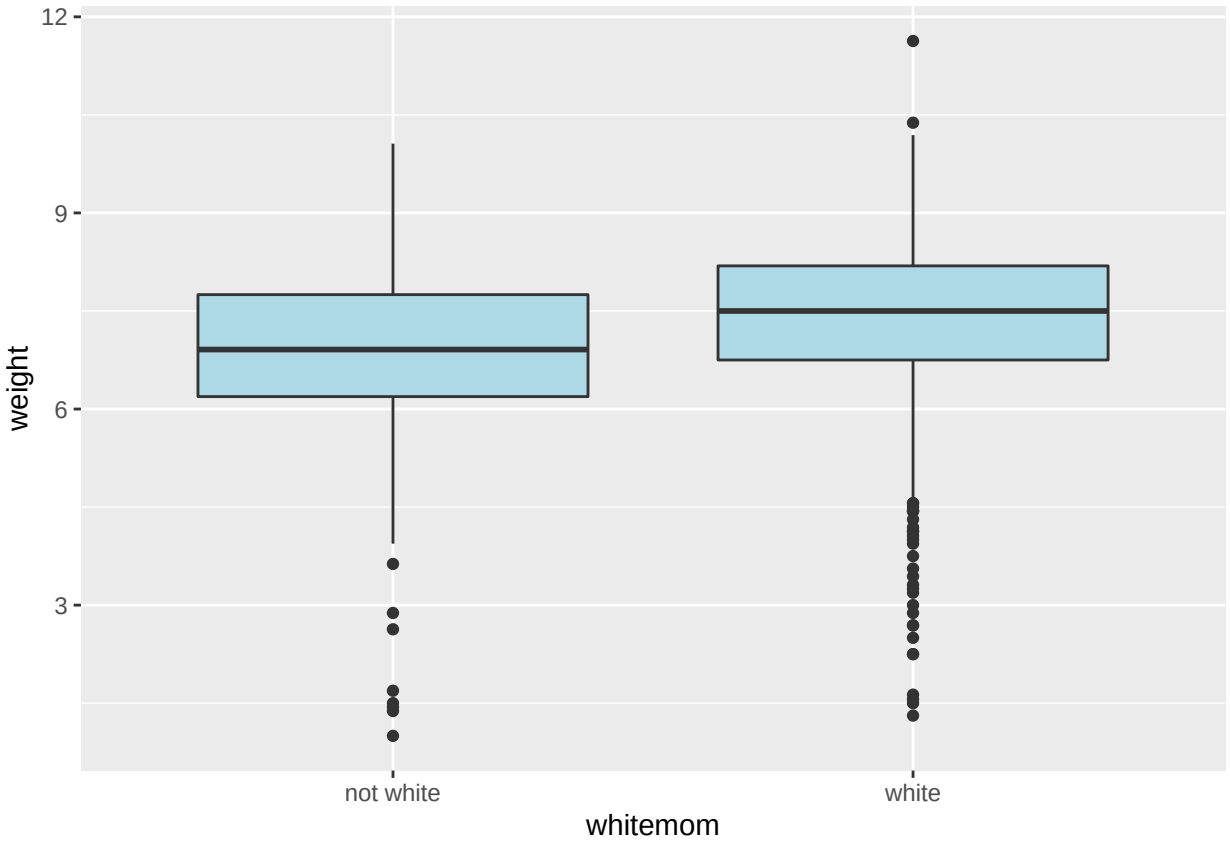
Answer: there is a right skew. Note to the grader... technically a boxplot could also be used to assess this... though we did not show the students how

Exercise 12

Using a data visualization, visually assess: (in this sample) is the median birth `weight` of babies greater for white or non-white mothers (variable called `whitemom`)?

Answer: median weight of babies is greater for white mothers

```
ggplot(data = nc, aes(x = whitemom, y = weight)) +
  geom_boxplot(fill = "lightblue")
```

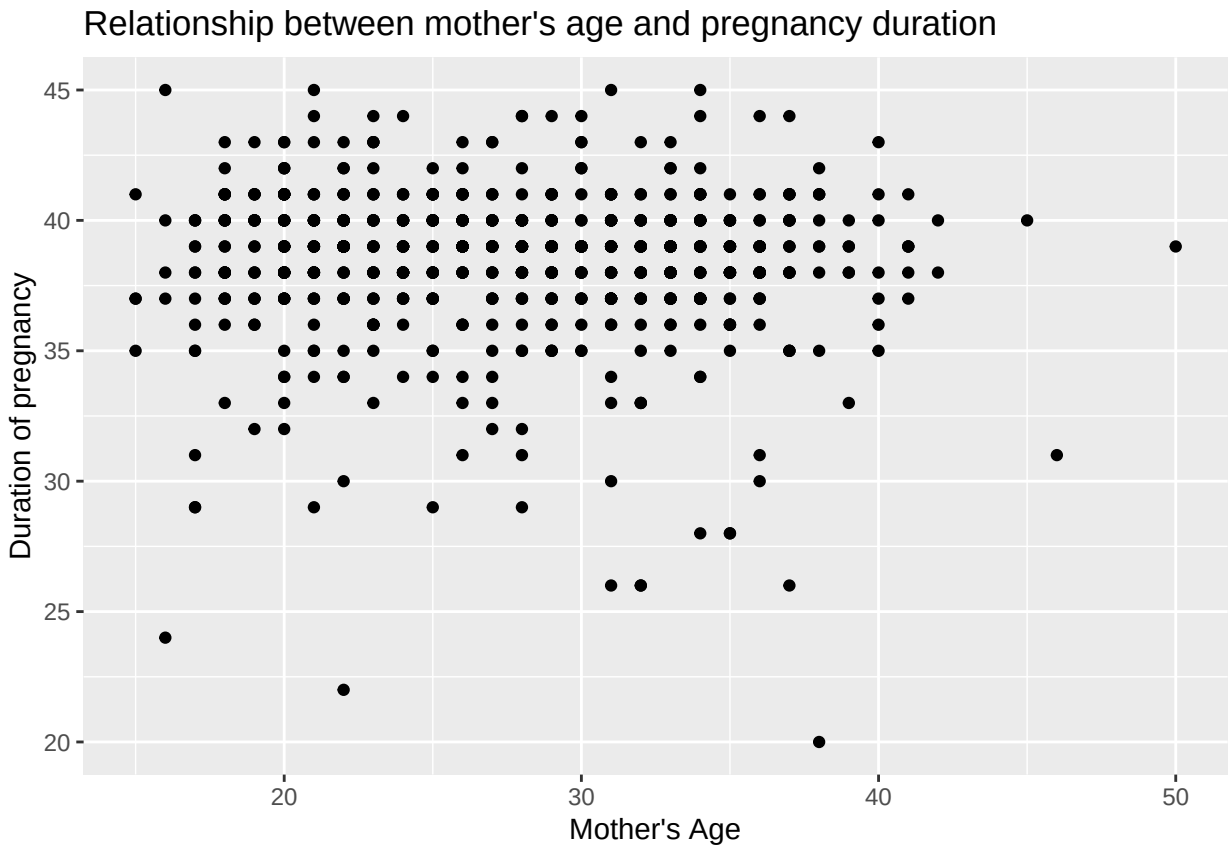


Exercise 13

Using a data visualization, visually assess: (in this sample) as a mother's age increases, does the duration of pregnancy appear to decrease?

Answer: no...there does not appear to be much of a relationship

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


Relationship between mother's age and pregnancy duration
Duration of pregnancy
45
40
35
30
25
20
20
30
40
50
Mother's Age