STA130H1S – Winter 2020

Week 1 Practice Problems - Sample Answers

N. Moon and L. Bolton [ADD YOUR NAME HERE]

Instructions

How do I hand in these problems for the January 9th deadline ?

Your complete .Rmd file that you create for these practice problems AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/138992/assignments/264008) by 11:59PM, on Thursday January 9th. Late problem sets are not accepted.

What should I bring to tutorial on January 10th?

R output (e.g., plots and written answers) for **Question 1**. You can either bring a hardcopy or bring your laptop with the output.

Tutorial Grading

Tutorial grades will be assigned according to the following marking scheme.

	Mark
Completion of required problems (due on Quercus the day before your tutorial)	1
Attendance for the entire tutorial	1
In-class exercises	4
Total	6

Practice Problems

[Question 1] The code below reads in data about books sold on Amazon (https://dasl.datadescription.com/datafile/amazon-books/). The data frame containing the book data is named `books`. Note that the height (`Height`), width (`Width`) and thickness (`Thick`) of books in this data frame are measured in inches.

```r
library(tidyverse) # Load the tidyverse package so it is available to use
books <- read.csv("amazonbooks.csv")
```

(a) Use the `glimpse()` function to view properties of the `books` dataset. How many observations does it include? How many variables are measured for each observation? How many rows and columns does the `books` data frame have?

```r
glimpse(books)
```

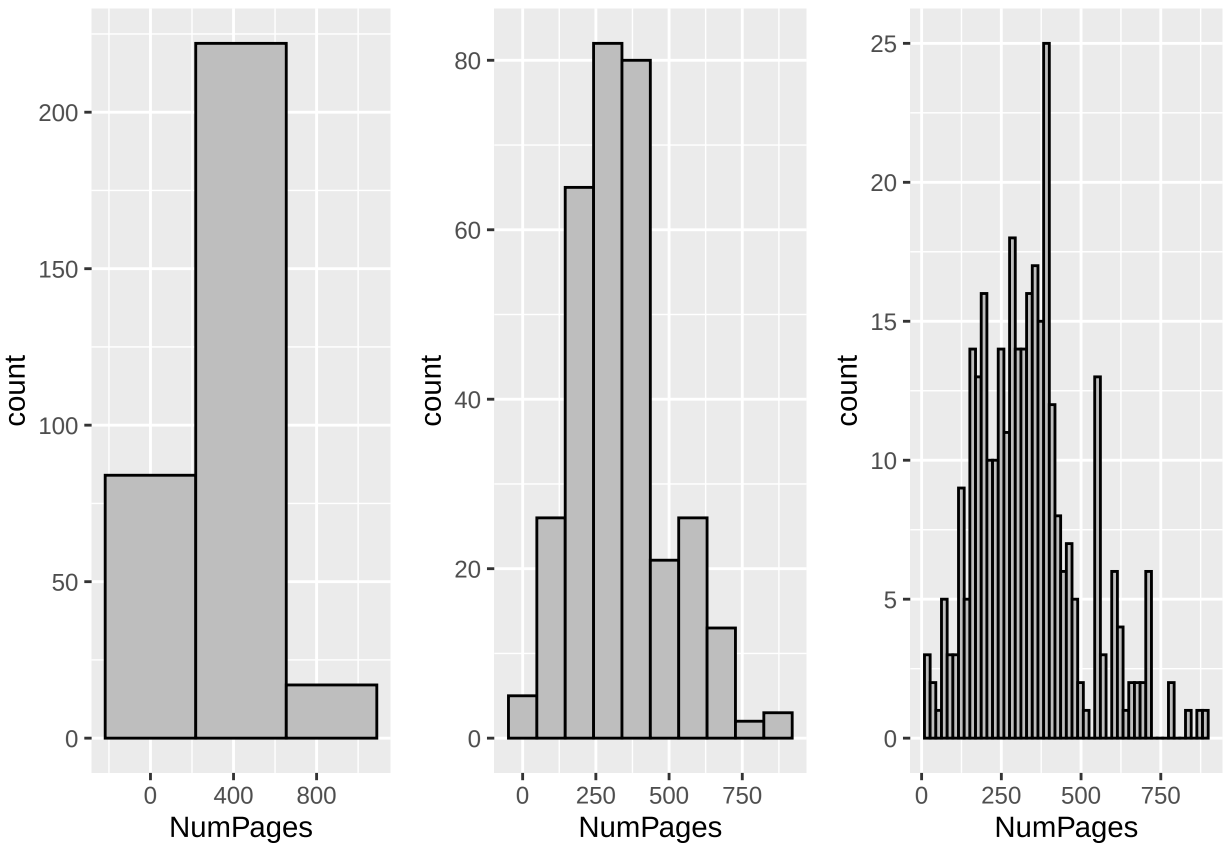
```
## Observations: 325
## Variables: 13
## $ Title         <fct> "1,001 Facts that Will Scare the S#*t Out of You: The U...
## $ Author        <fct> "Cary McNeal", "Ben Mezrich", "Smith", "Gavin Menzies",...
## $ List.Price    <dbl> 12.95, 15.00, 1.50, 15.99, 30.50, 28.95, 20.00, 15.00, ...
## $ Amazon.Price  <dbl> 5.18, 10.20, 1.50, 10.87, 16.77, 16.44, 13.46, 8.44, 18...
## $ Hard_or_Paper <fct> P, P, P, P, P, H, H, P, H, H, P, P, H, P, P, H, P, P, P...
## $ NumPages      <int> 304, 273, 96, 672, 720, 460, 336, 405, NA, 304, 624, 72...
## $ Publisher     <fct> "Adams Media", "Free Press", "Dover Publications", "Har...
## $ Pub.year      <int> 2010, 2008, 1995, 2008, 2011, 2011, 2010, 1987, 2011, 1...
## $ ISBN.10       <fct> 1605506249, 1416564195, 486285537, 61564893, 307265722,...
## $ Height        <dbl> 7.8, 8.4, 8.3, 8.8, 8.0, 8.9, 7.8, 8.2, 9.6, 9.6, 7.7, ...
## $ Width         <dbl> 5.5, 5.5, 5.2, 6.0, 5.2, 6.3, 5.3, 5.3, 6.5, 6.4, 5.1, ...
## $ Thick         <dbl> 0.8, 0.7, 0.3, 1.6, 1.4, 1.7, 1.2, 0.8, 2.1, 1.1, 1.7, ...
## $ Weight_oz     <dbl> 11.2, 7.2, 4.0, 28.8, 22.4, 32.0, 15.5, 11.2, NA, 19.2,...
```

Type your answer to the written question here (note that R code does not run outside of the grey R code chunks)

There are 325 observations in the `books` data frame, and each observation corresponds to one book sold on Amazon. For each book in the sample, 13 variables are measured and recorded: `Title`, `Author`, `List.Price`, `Amazon.Price`, `Hard_or_Paper`, `NumPages`, `Publisher`, `Pub.year`, `ISBN.10`, `Height`, `Width`, `Thick`, and `Weight_oz`. There are 325 rows (one for each observation) and 13 columns (one for each variable).

(b) Create 3 histograms to explore the distribution of number of pages in this sample of books. (i) one with 3 bins, (ii) one with 10 bins, and (iii) one with 50 bins. Which of these histograms is most appropriate to describe the distribution of number of pages per book? Why? Write a few sentences describing the distribution based on the histogram you chose as most appropriate.

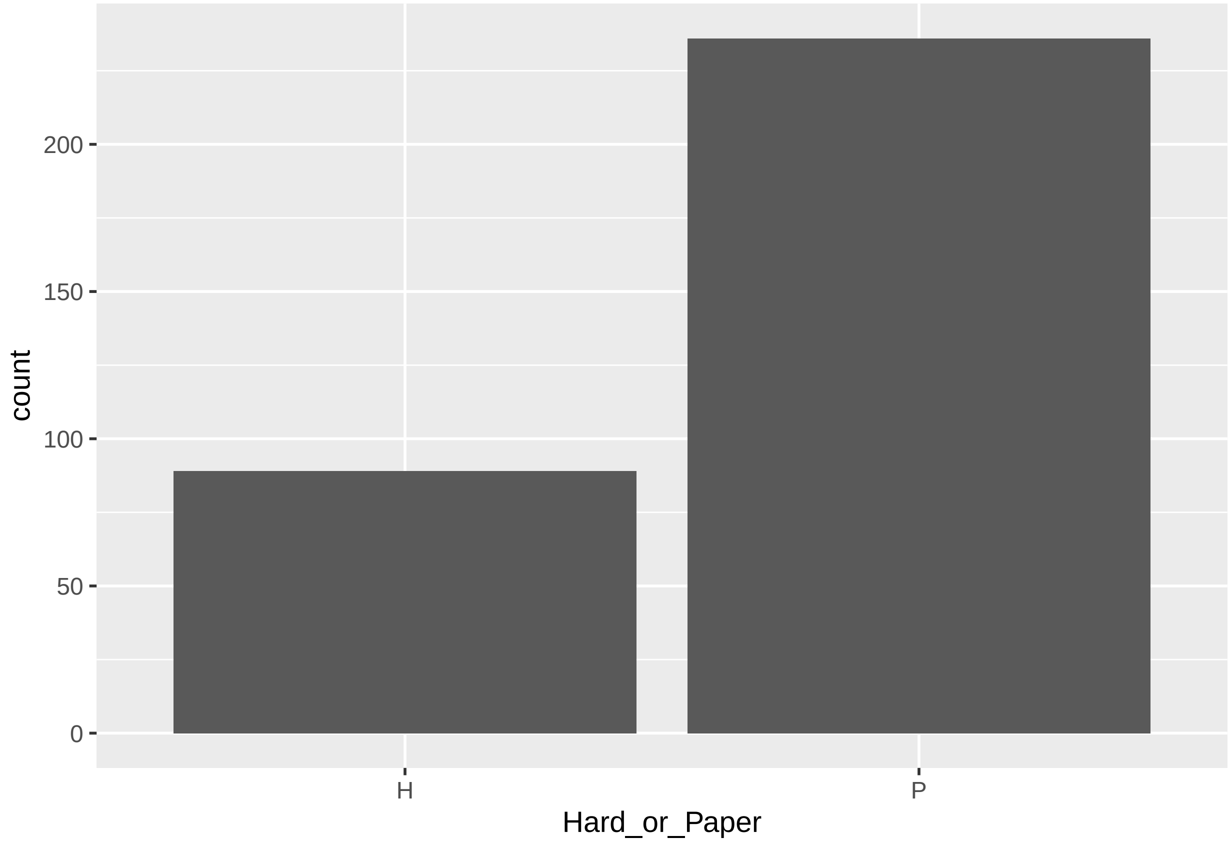
```r
hist1 <- ggplot(data=books, aes(x=NumPages)) + geom_histogram(color="black", fill="gray", bins=3)

hist2 <- ggplot(data=books, aes(x=NumPages)) + geom_histogram(color="black", fill="gray", bins=10)

hist3 <- ggplot(data=books, aes(x=NumPages)) + geom_histogram(color="black", fill="gray", bins=50)

# the gridExtra package allows for plots to be arranged in a grid layout - we'll load it here
library(gridExtra)
grid.arrange(hist1, hist2, hist3, nrow=1, ncol=3)
```



Among these three histograms, the one with 10 bins is best suited to describe the overall distribution of number of pages per book, as it best allows us to describe the center, shape, and spread. - Shape: It is hard to describe the overall shape and skewness of the distribution of book lengths from histograms 1 and 3, but from histogram 2 we see that the distribution is skewed to the right due to the right tail being longer than the left tail. - Spread: Based on histogram 1, we can only say that the number of pages ranges from 0 (even though the first bin includes some negative values in its range, we know this does not make sense) to over 1000. However, based on histograms 2 and 3, we can see that we can narrow this range down to approximately 0 to just over 825 pages. When the number of bins increases, we can get better estimates of the maximum and minimum values of a distribution. - Center: Based on histogram 2 (with 10 bins), it appears that the peak is between 250 and 400 pages (in next week's class, we'll talk more about different numerical measures of the center of a distribution). It is more difficult to estimate the center of the distribution based on histograms 1 and 3.

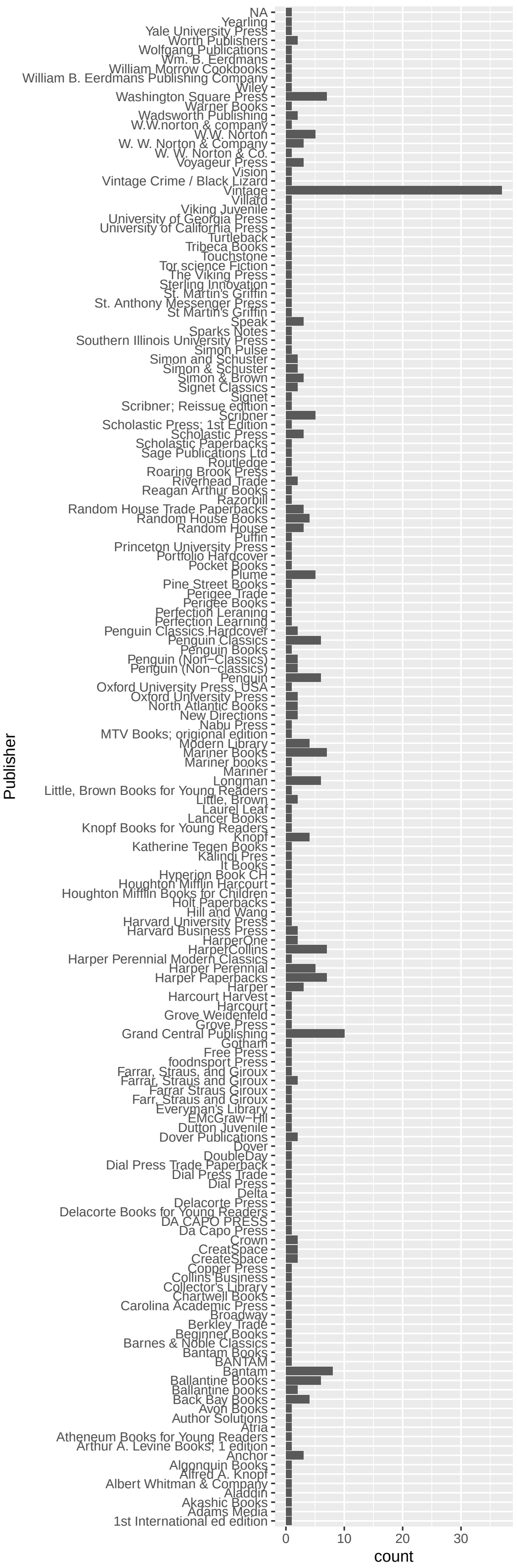
(c) Construct a plot to visualize the distribution of a categorical variable and describe the distribution in 1-2 sentences. Hint: If you choose a categorical variable with many different categories, you may find it useful to use coord_flip() to flip the bars horizontally and/or change the options in the R code chunk to make the plot large (ex: {r, fig.height=15, fig.width=5}).

```r
ggplot(data=books, aes(x=Hard_or_Paper)) + geom_bar()
```



There are about 2.5 times more books printed in paperback (P) than hardcover (H).

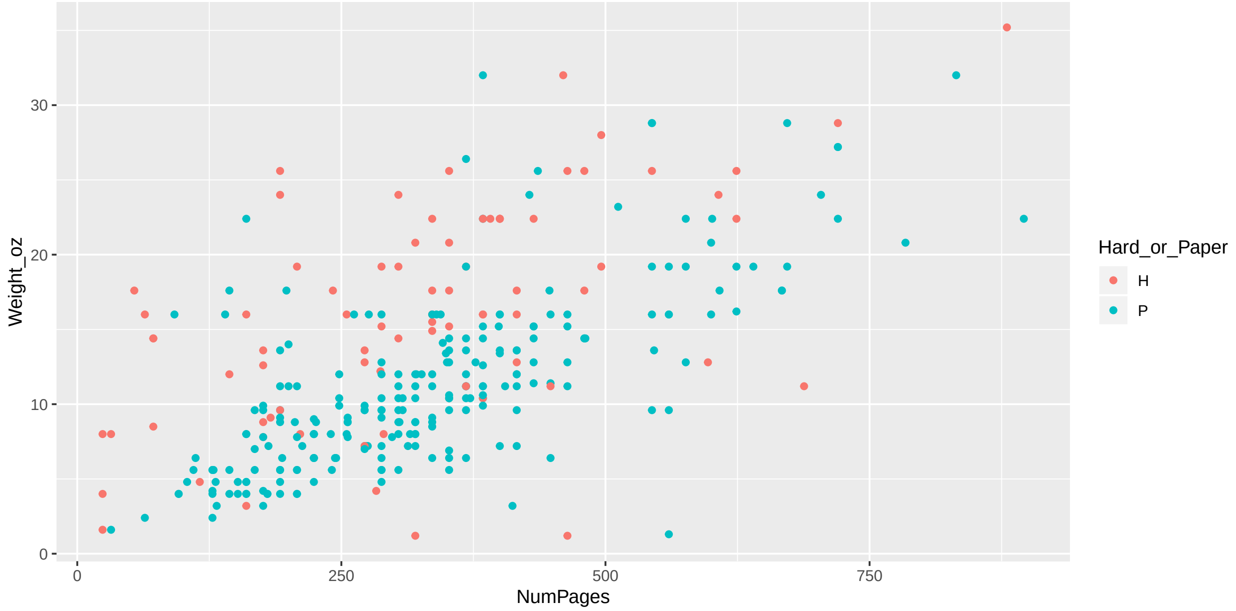
```
ggplot(data=books, aes(x=Publisher)) + geom_bar() + coord_flip()
```


Publisher
NA
Yearling
Yale University Press
Worth Publishers
Wolfgang Publications
Wm. B. Eerdmans
William Morrow Cookbooks
William B. Eerdmans Publishing Company
Wiley
Washington Square Press
Warner Books
Wadsworth Publishing
W.W.norton & company
W.W. Norton
W. W. Norton & Company
W. W. Norton & Co.
Voyageur Press
Vision
Vintage Crime / Black Lizard
Vintage
Villard
Viking Juvenile
University of Georgia Press
University of California Press
Turtleback
Tribeca Books
Touchstone
Tor science Fiction
The Viking Press
Sterling Innovation
St. Martin's Griffin
St. Anthony Messenger Press
St Martin's Griffin
Speak
Sparks Notes
Southern Illinois University Press
Simon Pulse
Simon and Schuster
Simon & Schuster
Simon & Brown
Signet Classics
Signet
Scribner; Reissue edition
Scribner
Scholastic Press; 1st Edition
Scholastic Press
Scholastic Paperbacks
Sage Publications Ltd
Routledge
Roaring Brook Press
Riverhead Trade
Reagan Arthur Books
Razorbill
Random House Trade Paperbacks
Random House Books
Random House
Puffin
Princeton University Press
Portfolio Hardcover
Pocket Books
Plume
Pine Street Books
Perigee Trade
Perigee Books
Perfection Leraning
Perfection Learning
Penguin Classics Hardcover
Penguin Classics
Penguin Books
Penguin (Non−Classics)
Penguin (Non−classics)
Penguin
Oxford University Press, USA
Oxford University Press
North Atlantic Books
New Directions
Nabu Press
MTV Books; origional edition
Modern Library
Mariner Books
Mariner books
Mariner
Longman
Little, Brown Books for Young Readers
Little, Brown
Laurel Leaf
Lancer Books
Knopf Books for Young Readers
Knopf
Katherine Tegen Books
Kalindi Pres
It Books
Hyperion Book CH
Houghton Mifflin Harcourt
Houghton Mifflin Books for Children
Holt Paperbacks
Hill and Wang
Harvard University Press
Harvard Business Press
HarperOne
HarperCollins
Harper Perennial Modern Classics
Harper Perennial
Harper Paperbacks
Harper
Harcourt Harvest
Harcourt
Grove Weidenfeld
Grove Press
Grand Central Publishing
Gotham
Free Press
foodnsport Press
Farrar, Straus, and Giroux
Farrar, Straus and Giroux
Farrar Straus Giroux
Farr, Straus and Giroux
Everyman's Library
EMcGraw−Hil
Dutton Juvenile
Dover Publications
Dover
DoubleDay
Dial Press Trade Paperback
Dial Press Trade
Dial Press
Delta
Delacorte Press
Delacorte Books for Young Readers
DA CAPO PRESS
Da Capo Press
Crown
CreatSpace
CreateSpace
Copper Press
Collins Business
Collector's Library
Chartwell Books
Carolina Academic Press
Broadway
Berkley Trade
Beginner Books
Barnes & Noble Classics
Bantam Books
BANTAM
Bantam
Ballantine Books
Ballantine books
Back Bay Books
Avon Books
Author Solutions
Atria
Atheneum Books for Young Readers
Arthur A. Levine Books; 1 edition
Anchor
Algonquin Books
Alfred A. Knopf
Albert Whitman & Company
Aladdin
Akashic Books
Adams Media
1st International ed edition
0
10
20
30
count

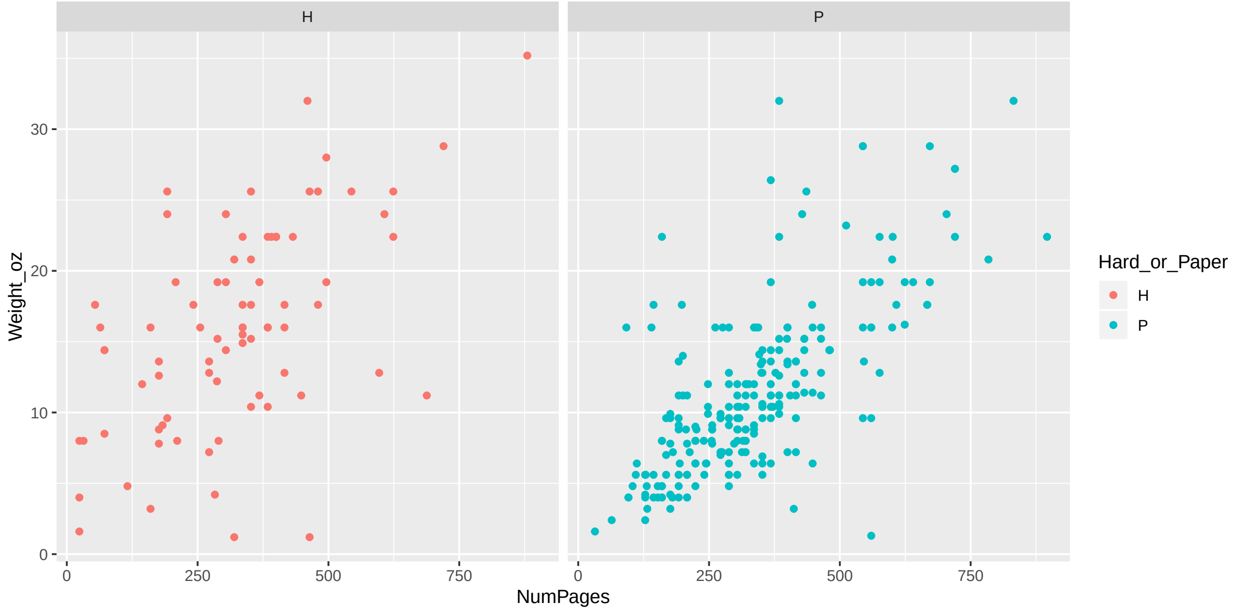
Most of the publishers only feautre a small number of books in the books data frame, but the publisher Vintage has over 35 books in this sample.

(d) Construct a plot showing the association between number of pages (NumPages), weight (Weight_oz), and the type of cover (Hard_or_Paper), and write 3-4 sentences describing the association between these three variables.

```
ggplot(data=books, aes(x=NumPages, y=Weight_oz, color=Hard_or_Paper)) +
  geom_point()
```



```
ggplot(data=books, aes(x=NumPages, y=Weight_oz, color=Hard_or_Paper)) +
  geom_point() +
  facet_wrap(~Hard_or_Paper)
```



Write your answer here

There is a positive linear association between the number of pages and the weight of a book, which is not surprising. This linear association is much stronger for paperback books than for hardcover books - again, this makes sense because there may be more variability in the weight of hardcover books, depending on the material with which the cover is made. In general, we see that hardcover books are heavier than paperback books.

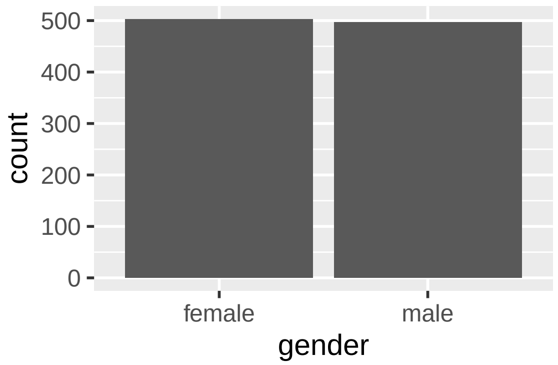
[Question 2] The `ncbirths` data set is part of the `openintro` package. It consists of observations for a sample of 1000 births in North Carolina in 2004. Type `?ncbirths` in the R console for more information about the data and to see the definition of each variable. The code below loads the required libraries for this question and provides a glimpse of the `ncbirths` data frame.

```r
glimpse(ncbirths)
```

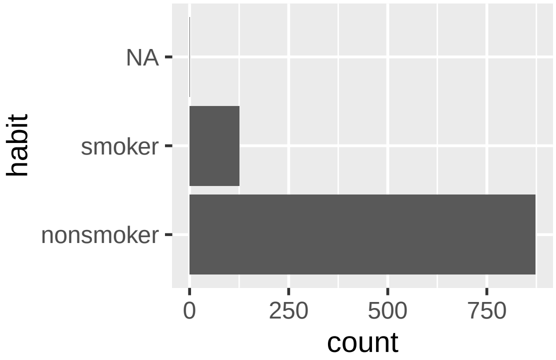
```
## Observations: 1,000
## Variables: 13
## $ fage           <int> NA, NA, 19, 21, NA, NA, 18, 17, NA, 20, 30, NA, NA, NA...
## $ mage           <int> 13, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16...
## $ mature         <fct> younger mom, younger mom, younger mom, younger mom, yo...
## $ weeks          <int> 39, 42, 37, 41, 39, 38, 37, 35, 38, 37, 45, 42, 40, 38...
## $ premie         <fct> full term, full term, full term, full term, full term,...
## $ visits         <int> 10, 15, 11, 6, 9, 19, 12, 5, 9, 13, 9, 8, 4, 12, 15, 7...
## $ marital        <fct> married, married, married, married, married, married, ...
## $ gained         <int> 38, 20, 38, 34, 27, 22, 76, 15, NA, 52, 28, 34, 12, 30...
## $ weight         <dbl> 7.63, 7.88, 6.63, 8.00, 6.38, 5.38, 8.44, 4.69, 8.81, ...
## $ lowbirthweight <fct> not low, not low, not low, not low, not low, low, not ...
## $ gender         <fct> male, male, female, male, female, male, male, male, ma...
## $ habit          <fct> nonsmoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker,...
## $ whitemom       <fct> not white, not white, white, white, not white, not whi...
```

(a) Choose two categorical variables and plot their distributions. Write one or two sentences interpreting each plot.

```r
# Construct your plots in this code chunk
ggplot(data = ncbirths) + aes(x = gender) + geom_bar()
```



```r
ggplot(data = ncbirths) + aes(x = habit) + geom_bar() + coord_flip()
```

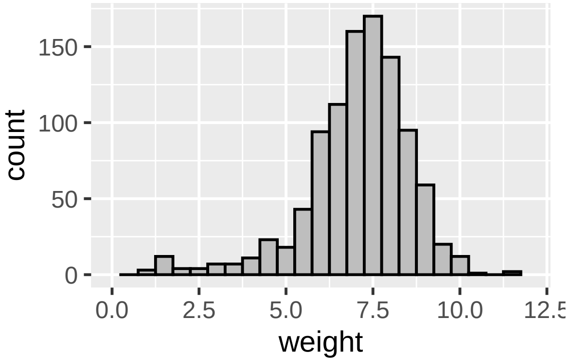


The first plot shows that there are approximately the same number of male and female babies (497 vs 503).

The second plot shows that the majority of babies were born to non-smoking mothers. There is one missing value (NA) which indicates that for one observation, it is not known if the mother was a smoker or a non-smoker.

(b) Choose a quantitative variable and plot its distribution. Interpret the plot.

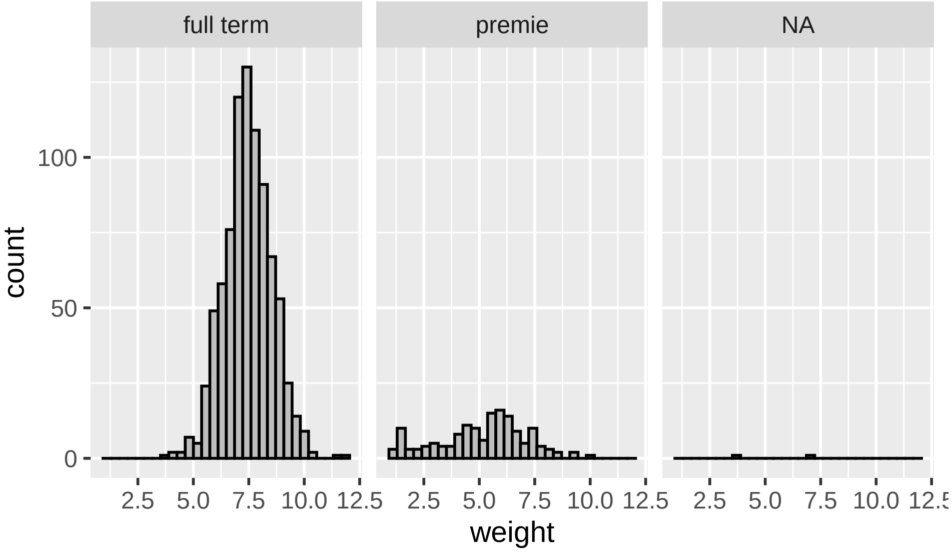
```
# Construct your plots in this code chunk
ggplot(data = ncbirths) + aes(x = weight) +
  geom_histogram(fill = "grey", colour = "black", bins=25) + xlim(0,12)
```



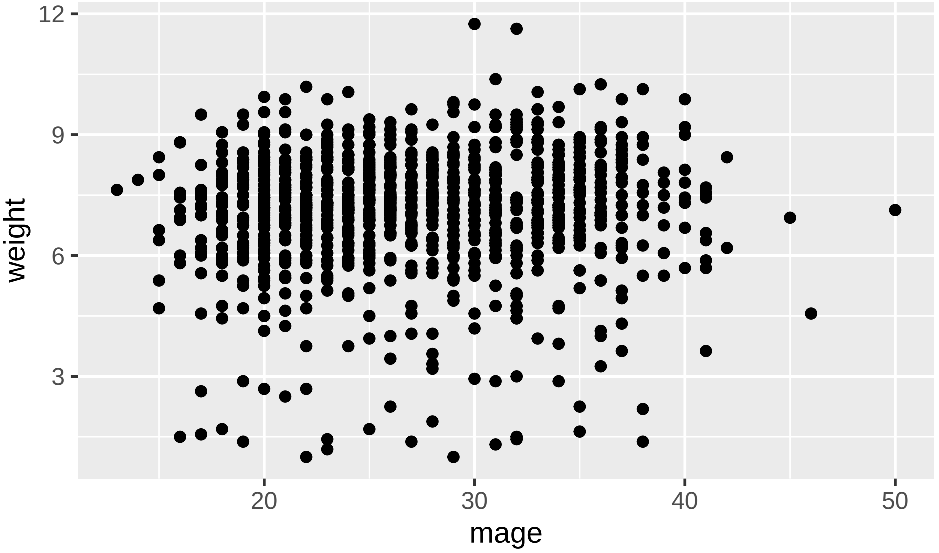
The distribution of babies' birthweight is left skeewed since the left tail is longer than the right tail. There is one prominent peak so the distribution is unimodal, with a mode around 7.5 pounds.

(c) Construct a plot that shows the relationship between two variables. What can you say about the relationship?

```
# Construct your plots in this code chunk
ggplot(data = ncbirths) +
  aes(x = weight) +
  geom_histogram(fill = "grey", colour = "black") +
  facet_wrap(~premie)
```



```
ggplot(data = ncbirths) +
  aes(x = mage, y = weight) +
  geom_point()
```



In the first plot, the distribution of birthweight is shown separately for full-term and premature (premie) babies (note that NA means ‘not available’, so for two observations we do not know if the babies were full-term or premature). We see that the distribution of weights for full-term babies is approximately symmetric, while the distribution of weights is much more left-skewed for premature babies, although both distributions are unimodal. The peak is around 7.5 pounds for full-term babies, and closer to 5 pounds for premature babies.

There is no strong association between the mother’s age and the baby’s weight. There is no obvious pattern in the dots which would suggest a relationship between these two variables.

[Question 3] In this question, you will consider another example of survivor bias. Dementia is a general term describing a decline in the mental ability of older adults which interferes with daily life. Researchers are interested in determining the average time between onset of dementia and death. To do so, they recruit individuals who have already been diagnosed with dementia for at least one year and follow them over time until their deaths.

(a) Is this sample representative of all individuals who develop dementia?

No, individuals who die soon after being diagnosed with dementia will not be our sample.

(b) Do you expect that the average survival time with dementia calculated from this sample will be close to the true value?

Since the sample only includes individuals who survived at least one year with dementia, it is reasonable to assume that the average survival time with dementia in this sample would be higher than the true average survival time of everyone diagnosed with dementia. In the plane example from class, the "missing" observations were the planes that crashed and never returned. In this example, the missing observations are the individuals who die soon after their diagnosis and therefore, are not eligible to enter the study. This is another example of survivor bias. If the patients in the sample are treated as a representative sample of all dementia patients, the study results will overestimate the average survival time. It is critical to think about the data source when drawing conclusions from data!