Exercises

Problem 1.1

Here is an excerpt from the baby-name data table in `"DataComputing::BabyNames"`.

name	sex	count	year
Taffy	F	19	1970
Liliana	F	162	1973
Stan	M	55	1975
Nettie	F	45	1978
Kateria	F	8	1980
... and so on for 1,792,091 rows			

Consider these five entities, that appear in the table shown above (a) through (e):

a) `Taffy` b) `year` c) `sex` d) `name` e) `count`

For each, choose one of the following:

1. It's a categorical variable.
2. It's a quantitative variable.
3. It's the value of a variable for a particular case.

Problem 1.2

What's not tidy about this table?

`president`	`in office`	`number of states`
Lincoln, Abraham	1861-1865	it depends
George Washington	1791-1799	16
Martin Van Buren	1837 to 1841	26

Table 1.7: An untidy table

Problem 1.3

Re-write Table 1.7 in a tidy form. Take care to render the information about years and about the number of states as numbers.

Problem 1.4

Here are three different organizations (**A**, **B**, and **C**) of the same data:

Data Table A

Year	Algeria	Brazil	Columbia
2000	7	12	16
2001	9	14	18

Data Table B

Country	Y2000	Y2001
Algeria	7	9
Brazil	12	14
Columbia	16	18

Data Table C

Country	Year	Value
Algeria	2000	7
Algeria	2001	9
Brazil	2000	12
Columbia	2001	18
Columbia	2000	16
Brazil	2001	14

1. What are the variables in each table?
2. What is the meaning of a case for each table? Here are some possible choices.
   - A country
   - A country in a year
   - A year

Problem 1.5

The codebook for several data tables relating to airports, airlines, and airline flights in the US is published at `https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf`.

Within that document is the codebook for the data table `airports`.

1. How many variables are there?
2. What do the cases represent?
3. For each variable, make a reasonable guess about whether the values will be numerical or quantitative.

and an object name.

```
Motors <- read.file("http://tiny.cc/mosaic/engines.csv")
```

The effect of this command is to read some data about internal combustion motors from a web site into an R object called `Motors`. Note that the URL of the data file is a quoted character string, but the function and object names are *not* quoted.

2.9 *Exercises*

Problem 2.1
The following ideas should be meaningful to you from Chapter 2:

package, function, command, argument, assignment, object, object name, data table, named argument, quoted character string, value

Construct an example R command that makes use of at least four of the ideas. Label which part of your example R command corresponds to each of those ideas.

Problem 2.2
Which of these kinds of names should be wrapped with quotation marks when used in R?

1. function name
2. file name
3. the name of an argument in a named argument
4. object name

Problem 2.3
Look at the documentation for the `CPS85` data table in the `mosaicData` package. From reading that documentation, what is the meaning of CPS?

Problem 2.4
What's wrong with this statement?

```
help(NHANES, package <- "NHANES")
```

Problem 2.5
Look at the help documentation for the `library()` function.

Without worrying about all the detail, answer these questions simply:

1. What is the other function listed under "Usage"?
2. In the "See Also" section of the documentation, what is the name of the function after `detach()`?

Problem 2.6
Some of these are legitimate object names, others are not. For the ones that are not legitimate, say what is wrong.

1. `essay14`
2. `first-essay`
3. `"MyData"`
4. `third_essay`
5. `small sample`
6. `functionList`
7. `FuNcTiOnLiSt`
8. `.MyData.`
9. `sqrt()`

Problem 2.7

Install the `nycflights13` package into R. (You can use the "Packages" tab which has an "install" button. If you are not using RStudio, given the R command `install.packages("nycflights13")`)

Once the package is installed, you can access the `flights` data table with this command:

```
data(flights, package="nycflights13")
```

The codebook is available with

```
help(flights)
```

Using the codebook and examining the data table with the `View()` command (hint: you'll need to give `flights` as an argument to `View()`), answer these questions:

1. How many variables are there?
2. How many cases are there?
3. What is the meaning of a case? ("Meaning" refers to the kind of entity, for instance, "airport" or "airline" or "date". Hint: the case in `flights` is not any of these things.)
4. For each variable, is the variable quantitative or categorical?
5. For the variables `air_time` and `distance`, what are the units?

Problem 2.8

Consider this list of some possible mistakes in an assignment operation:

1. No assignment operator
2. Unmatched quotes in character string
3. Improper syntax for function argument
4. Invalid object name
5. No mistake

For each of the following assignment statements, say what is the mistake.

a. `ralph <- sqrt 10`
b. `ralph2 <-- "Hello to you!"`
c. `3ralph <- "Hello to you!"`
d. `ralph4 <- "Hello to you!`
e. `ralph5 <- date()`

Problem 2.9

Here are a few characters: `.  ,  ;  _  -  ^  [space]  (  )`

- Which of those characters can be used in the name of an R object?
- Which of those characters can be used in a quoted character string?

Problem 2.10

These questions should be easy to answer if you use the appropriate commands to load, view, or get documentation on the datasets.

- How many variables are there in `CountryData`?
- What does the variable `tfat` measure in the `NCHS` data table? (in package `DataComputing`)
- How many cases are there in `WorldCities`?
- What's the third variable in `BabyNames`?
- What are the codes for the levels of the categorical variable `party` in the `RegisteredVoters` data table, and what does each code stand for?

3.6 Exercises

Problem 3.1
Using the object name `fireplace`, write different expressions with enough context to be able to identify the name as belonging to

1. a data frame
2. a function
3. the name of a named argument
4. a variable

Write one expression for each of the above.

Problem 3.2
Explain why the following sentence is illegitimate:

```
Result <- %>% filter(BabyNames, name=="Prince")
```

Problem 3.3
What's wrong with this statement?

```
help(NHANES, package <- "NHANES")
```

Problem 3.4
Consider these R expressions. (You don't have to know what the various functions do to solve this problem.)

```
Princes <-
  BabyNames %>%
  filter(name == "Prince") %>%
  group_by(year, sex) %>%
  summarise(yearlyTotal = sum(count))
# Now graph it!
Princes %>%
  ggplot(aes(x = year, y = yearlyTotal)) +
  geom_point(aes(color = sex)) +
  geom_vline(xintercept = 1978)
```

There are several kinds of named objects in the above expressions.

a. function name
b. data table name
c. variable name
d. name of a named argument

Using the naming convention and position rules, identify what kind of object each of the following name is used for. That is, assign one of the types (a) through (d) to each name.

1) `BabyNames` 2) `filter` 3) `name` 4) `==`
2) `group_by` 6) `year` 7) `sex` 8) `summarise`
3) `yearlyTotal` 10) `sum` 11) `count` 12) `ggplot`
4) `aes` 14) `x` 15) `y` 16) `geom_point`
5) `color` 18) `geom_vline` 19) `xintercept`

Problem 3.5
There are several small, example data tables in the `ggplot2` package. Look at the `msleep` data table by using the `View()` function with the name of the object as an argument.

- What is the meaning of the `brainwt` variable?
- How many cases are there?
- What is the real-world meaning of a case?
- What are the levels of the `vore` variable?

Problem 3.6
The data verb functions all take a data table as their first argument and return a data table as their output. The chaining syntax lets the output of one function become the input to the following function, so you don't have to repeat the name of the data frame. An alternative syntax is to assign the output of one function to a named object, then use the object as the first argument to the next function in the computation.

Each of these statements, but one, will accomplish the same calculation. Identify the statement that does not match the others.

a)
```
BabyNames %>%
  group_by( year, sex ) %>%
  summarise( totalBirths=sum(count))
```

b)
```
group_by( BabyNames, year, sex) %>%
  summarise( totalBirths=sum(count) )
```

c)
```
group_by( BabyNames, year, sex ) %>%
  summarise( totalBirths=mean(count) )
```

d)
```
Tmp <- group_by(BabyNames, year, sex)
summarise( Tmp, totalBirths=sum(count) )
```

Problem 3.7
Which characters can be used in an object name?

Problem 3.8
The `date()` function returns an indication of the current time and date.

- What arguments does `date()` take? Use `help()` to find out.
- What *kind* of object is the result from `date()`.

4.1 Exercises

Problem 4.1
Markdown provides a simple way to produce section headers and sub-headers, italic and bold text, monospaced fonts suitable for computer commands, and even web links. A reference is available in RStudio at the menu HELP/MARKDOWN QUICK REFERENCE.

For each of the following, say how it will be rendered when the Markdown is rendered to HTML. (Hint: You can figure it out by reading the documentation, or you can put the text into an Rmd document and compile it!)

```
*one*

**two**

* three

# Four

`five`

## Six

[seven](http://tiny.cc/dcf/index.html)
```

Problem 4.2
What's wrong with the markup for each of these five chunks:

(a)

```
'''{r}
9+7
'''
```

(b)

```
"""(r)
9+7
"""
```

(c)

```
```{r}
9+7
``
```

(d)

```
```{r} 9+7```
```

(e)

```
```{r}
9+7
````
```

Problem 4.3
Treat the following lines as an `.Rmd` file which will be compiled to HTML.

```
### An Introduction

Arithmetic is *easy*! For instance
```{r}
3 + 2
```
```

Using paper and pencil, sketch out what the HTML document will look like when viewed in a web browser.

Problem 4.4
Here is a short list of names:

1. `DataComputing.org`
2. `ahab/whale.Rmd`
3. `ptth://world-bank.org`
4. `http://world-bank.org`
5. `//world-bank.org/index.html`
6. `world-bank.org/index.html`

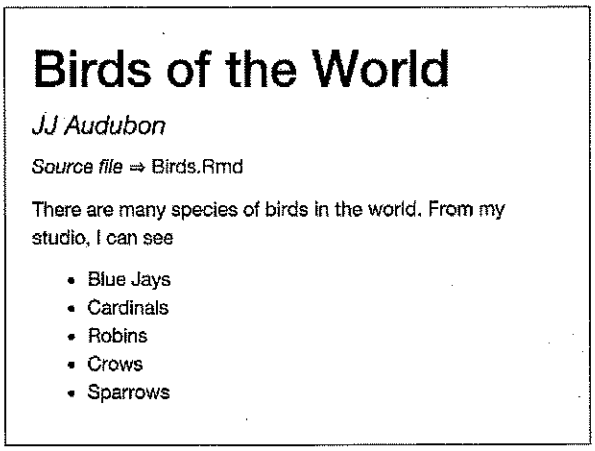
For each, say whether the name is in the allowed form for a possible URL, a possible file, neither, or both.

Problem 4.5
From the RStudio console, load the `DataComputing` package like this:

```
library(DataComputing)
```

Once this is done,

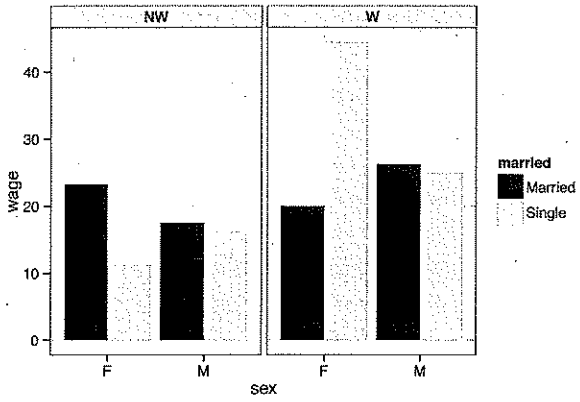
1. Open a new file using the File/New File/R Markdown ... menu item. Select "From Template" and then choose the *DataComputing simple* template.
2. Save the text that appears in the editor tab in a file named `Birds.Rmd`
3. Compile the `Birds.Rmd` file to HTML to verify that the template is working.
4. Edit the `Birds.Rmd` file to include contents that will make the compiled HTML file appear like this:



5.3 *Exercises*

Problem 5.1

Consider this bar graph of the `CPS85` data in the `mosaicData` package:

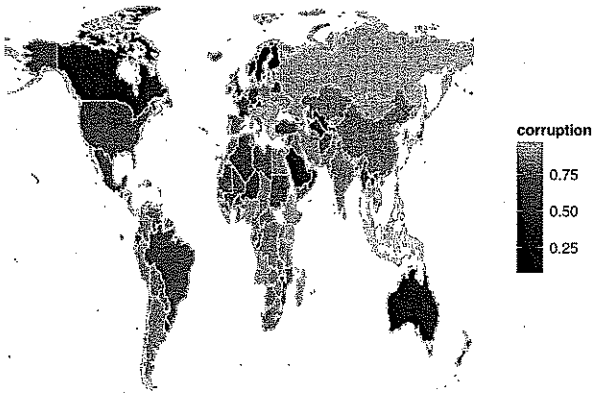


Use `barGraphHelper()` to reconstruct the graph. Start with these commands:

```
library(mosaicData)
library(DataComputing)
barGraphHelper(CPS85)
```

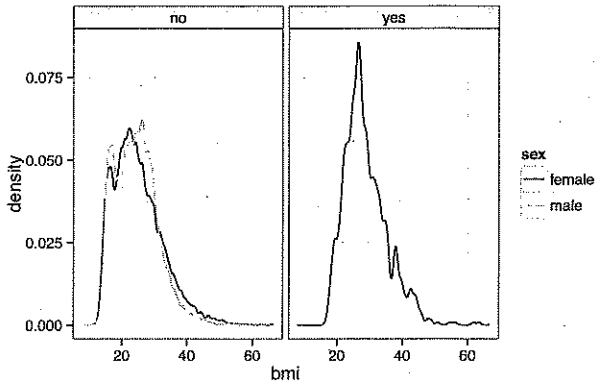
Problem 5.2

Make this map using data from `HappinessIndex` in the `DataComputing` package:



Problem 5.3

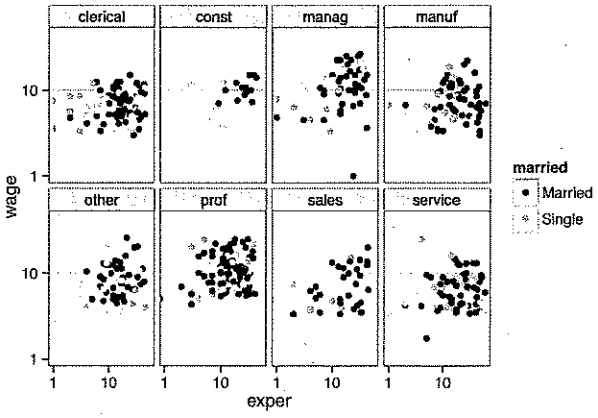
Make this graph from the `NCHS` data in the `DataComputing` package.



The "yes" and "no" in the gray bars refer to whether or not the person is pregnant.

Problem 5.4

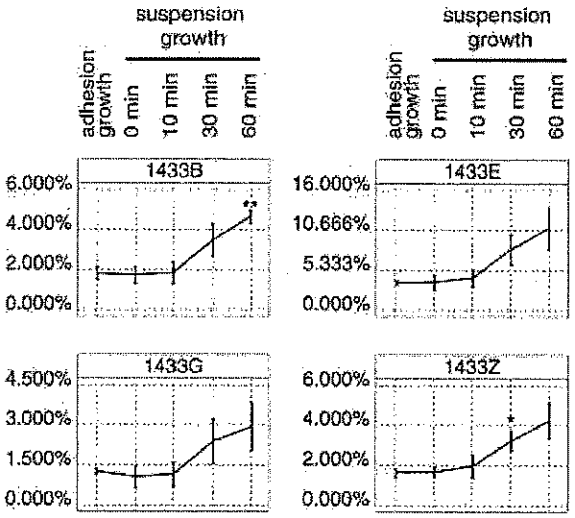
Using the `CPS85` data table (from the `mosaicData` package) make this graphic:



6.6 *Exercises*

Problem 6.1

The following chart contains four facets. Each shows the amount of a substance in different conditions:

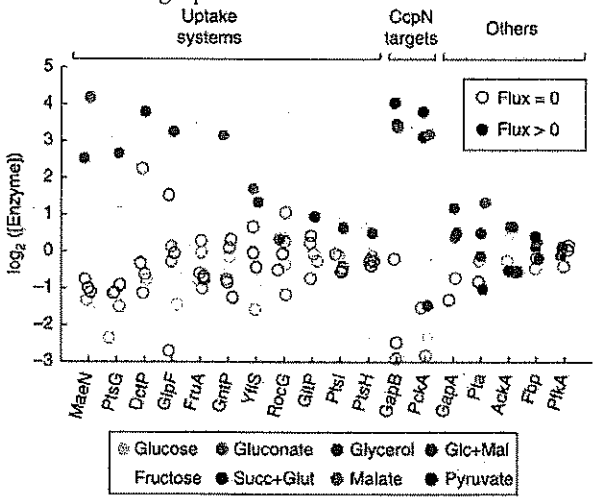


- when the cells are adhering to a surface
- when the cells are growing in suspension for different amounts of time

1. What are the labels/identifiers for the facets?
2. Are the frames the same in each facet?
3. There are three different glyphs shown in the frames. Describe each type in terms of its graphical properties.

Problem 6.2

Consider this graph



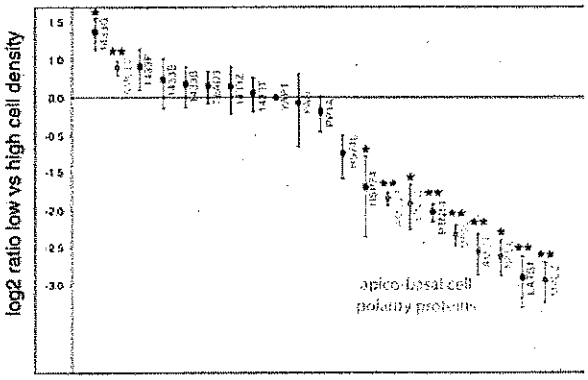
Here are some of the variables and their levels:

- Log enyzme `concentration`: numerical −3 to 5
- `target`: CcpN, Uptake, Other
- `flux`: zero or positive
- `gene`: MaeN, PtsG, DctP, ...
- `molecule`: Glocose, Fructose, Gluconate, ...

1. List all of the **guides** in the graph. For each one, say which variable is being mapped to which graphical attribute.
2. The basic glyph is a dot. Say what are the graphical attributes of the dot (e.g. color, size, ...). For each graphical attribute found in the graph, say which variable is mapped to that attribute.
3. Which two variables set the frame?
4. The scaling of the horizontal variable (e.g. the translation of position to variable levels) is set by a combination of two variables. Which two?

Problem 6.3

Consider this graphic:



Suppose the glyph-ready data underlying the graphic were structured as follows:

protein	center	low	high	polarity	signif
1433G	1.35	1.18	1.54	plus	1
AMOL2	0.78	0.63	1.01	minus	2
1433F	0.79	0.18	1.19	plus	0
1433E	0.42	-0.15	1.01	plus	0
⋮	⋮	⋮	⋮	⋮	⋮

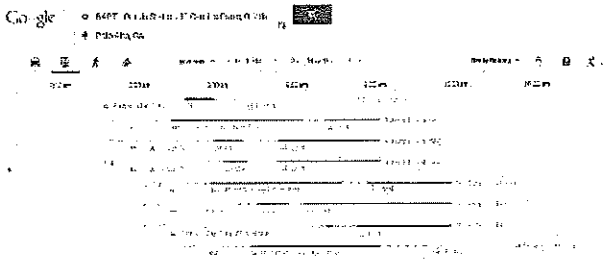
Consider these two kinds of glyph present in the graph:

and **

1. For each of the two glyphs, list the set of graphical attributes both geometrically (e.g. "dot") and in terms of the variable from the table that is mapped to that attribute (e.g., `polarity`).
2. Which variables define the frame? Give variables for both the horizontal and vertical coordinates.
3. Is color an attribute of the ** glyph?
4. What guides (if any) are displayed?

Problem 6.4

The graph, from Google Maps, shows mass transit options on a Monday morning for getting from Orinda, CA (in the East Bay), to Palo Alto, CA (in the West Bay).



(For a larger version, see `Data-Computing.org/images#C133`.)

1. Considering only that part of the graphic below the blue underlined bus and other modes of transportation, what is the frame?
2. Describe the different types of glyphs used.
3. For each different type of glyph
   - What information is encoded in the shape/style of the glyphs?
   - What information is encoded in the position of the glyph?
4. What guides are there?

Figure 6.9 presents forecasts for the US Senate elections in Nov. 2014. The numbers or words give the forecast probability of one party's candidate — Democrat or Republican — winning. The forecasts are made based on polls up through the end of August 2014. Individual results from several different polling organization are shown. The graphic is an excerpt from the full graphic at `http://www.nytimes.com/newsgraphics/2014/senate-model/comparisons.html`, which shows predictions for all 36 senate seats up for election in 2014. Source: New York Times

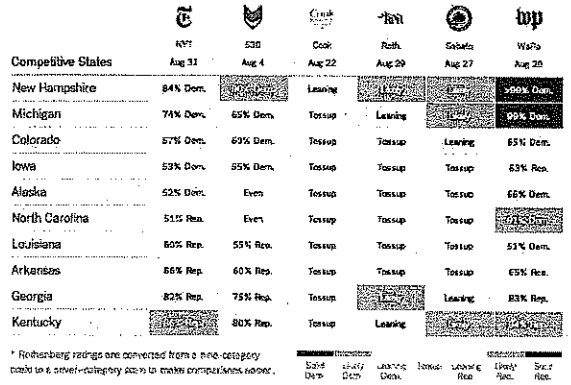


Figure 6.9: Forecasts made before the Nov. 2014 US Senate elections.

Problem 6.5

In Figure 6.9, what variables define the frame in this graphic?

a. Probability and State.

b. State and Polling Organization.

c. Democrats and Republicans.

d. Just State

e. Just Probability

Problem 6.6

In Figure 6.9, what is the glyph and its graphical attributes?

a. Glyph: names of the states. Graphical attribute: font.

b. Glyph: names of the polling organization. Graphical attribute: the organization's logo.

c. Glyph: Rectangle. Graphical attribute: color.

d. Glyph: Rectangle. Graphical attribute: color and text.

Problem 6.7

In Figure 6.9, what sets the order of the categorical variable in the scale for the vertical variable?

a) State

b) Poll

c) Roth poll probability for the Democratic candidate.

d) NYT poll probability for the Democratic candidate.

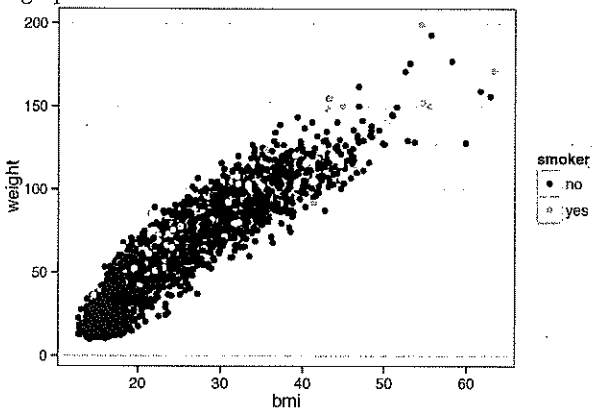
e) Date of the poll.

Problem 6.8

The `NCHS` data (in the `DataComputing` package) has 31126 rows. To speed things up, work with a small subset of `NCHS`:

```
Small <-
  NCHS %>%
  sample_n(size=5000)
```

Using the data in `Small`, make this plot with `scatterGraphHelper()` (in the `DataComputing` package). Then, write down the mapping between variables and graphical attributes.



7.4 *Exercises*

Problem 7.1
Each of these tasks can be performed using a single data verb. For each task, say which verb it is:

a. Find the average of one of the variables.

b. Add a new column that is the ratio between two variables.

c. Sort the cases in descending order of a variable.

d. Create a new data table that includes only those cases that meet a criterion.

e. From a data table with three categorical variables A, B, & C, and a quantitative variable X, produce an output that has the same cases but only the variables A and X.

f. From a data table with three categorical variables A, B, & C, and a quantitative variable X, produce an output that has a separate case for each of the combinations of the levels of A and B. (Hint: It might be easier to see the answer if the problem statement added, "and gives the maximum value of X over all the cases that have a given combination of A and B.")

Problem 7.2
These questions refer to the `diamonds` data table in the `ggplot2` package. Take a look at the codebook (using `help()`) so that you'll understand the meaning of the tasks. (Motivated by Garrett Grolemund.)

Each of the following tasks can be accomplished by a statement of the form

```
diamonds %>%
  verb1( args1 ) %>%
  verb2( args2 ) %>%
  arrange(desc( args3 )) %>%
  head( 1 )
```

For each task, give appropriate R functions or arguments to substitute in place of `verb1`, `verb2`, `args1`, `args2`, and `args3`.

1. Which color diamonds seem to be largest on average (in terms of carats)?

2. Which clarity of diamonds has the largest average "table" per carat?

Problem 7.3
For each of the operations listed here, say whether it involves a transformation function or a summary function or neither.

a. Determine the 3rd largest.

b. Determine the 3rd and 4th largest values.

c. Determine the number of cases.

d. Determine whether a year is a leap year.

e. Determine whether a date is a legal holiday.

f. Determine the range of a set, that is, the max minus the min.

g. Determine which day of the week (e.g., Sun, Mon, ...) a given date is.

h. Find the time interval in days spanned by a set of dates.

Problem 7.4
Each of these statements have an error. It might be an error in syntax or an error in the way the data tables are used, etc. Tell what are the error(s) in these expressions.

a)
```
BabyNames %>%
  group_by( "First" ) %>%
  summarise( votesReceived=n() )
```

b)
```
Tmp <- group_by(BabyNames, year, sex ) %>%
  summarise( Tmp, totalBirths=sum(count))
```

c)
```
Tmp <- group_by(BabyNames, year, sex)
summarise( BabyNames, totalBirths=sum(count) )
```

Problem 7.5

Here is a small data table based on `BabyNames`. Take this table as the input.

name	sex	count	year
Christina	M	22	1967
Rotha	F	7	1907
Wayman	M	9	1997
Song	F	11	1994
Julian	M	535	1948
... and so on for 1,792,091 rows			

For each of the following outputs, identify the operation linking the input to the output and write down the details (i.e., arguments) of the operation.

a) Output Table A

name	sex	count	year
Rotha	F	7	1907
Song	F	11	1994
Kalia	F	46	1989
Lissa	F	102	1962
Vicky	F	2945	1957
... and so on for 1,792,091 rows			

b) Output Table B

name	sex	count	year
Rotha	F	7	1907
Song	F	11	1994
Vicky	F	2945	1957
Kalia	F	46	1989
Lissa	F	102	1962
... and so on for 896,046 rows			

c) Output Table C

name	sex	count	year
Christina	M	22	1967
Julian	M	535	1948
... and so on for 416,765 rows			

d) Output Table D

total
333417770

e) Output Table E

name	count
Christina	22
Rotha	7
Wayman	9
Song	11
Julian	535
... and so on for 1,792,091 rows	

Problem 7.6

Using the `Minneapolis2013` data table, answer these questions:

1. How many cases are there?
2. Who were the top 5 candidates in the `Second` vote selections.
3. How many ballots are marked "undervote" in
   - `First` choice selections?
   - `Second` choice selections?
   - `Third` choice selections?
4. What are the top 3 combinations of `First` and `Second` vote selections? (That is, of all the possible ways a voter might have marked his or her first and second choices, which received the highest number of votes?)
5. Which `Precinct` had the highest number of ballots cast?

Problem 7.7

Each of these statements has an error. It might be an error in syntax or an error in the way the data tables are used, etc. Write down a correct version of the statement.

a)
```
BabyNames %>%
  group_by(BabyNames, year, sex) %>%
  summarise(BabyNames, total = sum(count))
```

b)
```
ZipGeography <-
  group_by(State) %>%
  summarise(pop = sum(Population))
```

c)
```
Minneapolis2013 %>%
  group_by(First) ->
  summarise(voteReceived = n())
```

d)
```
summarise(votesReceived = n()) %<%
  group_by(First) <- Minneapolis2013
```

Problem 7.9

a. There's only one data verb that takes a single data table as input and produce an output that (in general) has a different meaning to the case. Which one?

b. There's only one operation that takes two data tables as input rather than just a single data table. Which one?

Problem 7.10

Using the `ZipGeography` data

Find the total land area and population in each state.

- Make a scatter plot showing the relationship between land area and population for each state.
- Make a choropleth map showing the population of each state.
- Make a choropleth map showing the population per unit area of each state.

Problem 7.11

Imagine a data table, `Patients`, with categorical variables `name`, `diagnosis`, `sex`, and quantitative variable `age`.

You have a statement in the form

```
Patients %>%
  group_by(  **some variables** ) %>%
  summarise(count=n(), meanAge = mean(age))
```

Replacing `**some variables**` with each of the following, tell what variables will appear in the output

a. `sex`

b. `diagnosis`

c. `sex, diagnosis`

d. `age, diagnosis`

e. `age`

Problem 7.12

For each of these computations, say what R function is the most appropriate:

1. Count the number of cases in a data table.
2. List the names of the variables in a data table.
3. For data tables in an R package, display the documentation ("codebook") for the data table.
4. Load a package into your R session.
5. Mark a data table as grouped by one or more categorical variables.
6. Add up, group-by-group, a quantitative variable in a data table.

8.3 Exercises

Problem 8.1

Here are several functions from the ggplot2 graphics package used in *Data Computing*.

a)	`geom_point()`	b)	`geom_histogram()`
c)	`ggplot()`	d)	`scale_y_log10()`
e)	`ylab()`	f)	`facet_wrap()`
g)	`geom_segment()`	h)	`xlim()`
i)	`facet_grid()`		

Match each of the functions to the task it performs.

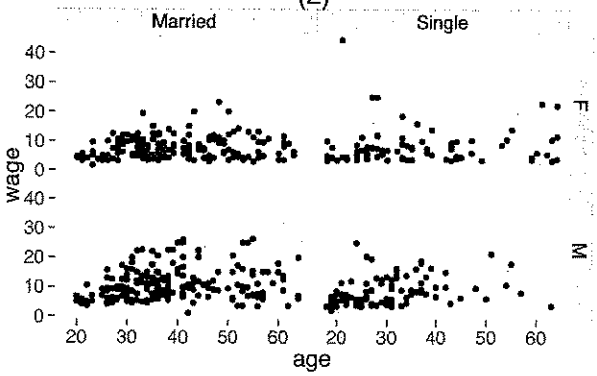
1) Construct the graphics frame
2) Add a layer of glyphs
3) Set an axis label
4) Divide the frame into facets
5) Change the scale for the frame.

Problem 8.2

Here are two more graphics based on the `mosaicData::CPS85` data table. Write `ggplot2()` statements that will construct each graphic.

(1)

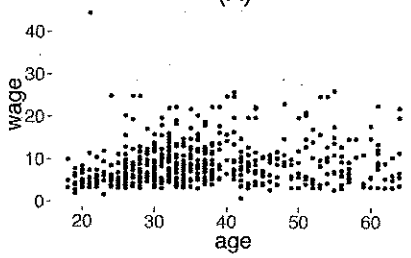
(2)

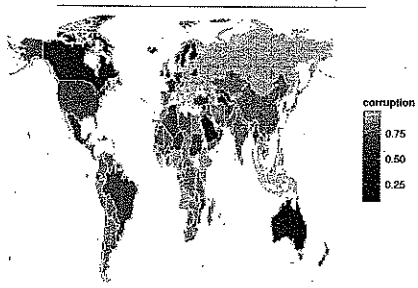


Problem 8.3

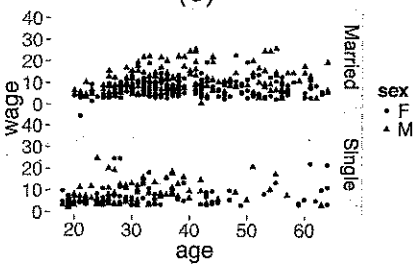
Here are four graphics based on the `mosaicData::CPS85` data table. Write `ggplot()` statements that will construct each graphic.

(A)

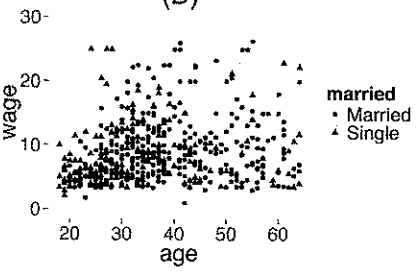




(C)



(D)



Function	Purpose
`glimpse()`	Quick summary of the table
`str()`	Quick summary
`summary()`	Quick summary
`nrow()`	How many cases in the data table
`ncol()`	How many variables in the data table
`names()`	The variable names
`View()`	Shows the table like a spreadsheet. In RStudio only.

9.7 *Exercises*

Problem 9.1

Identify each of these functions as either a **Data Verb**, a **Transformation**, a **Summary Function**, or a **Quick Presentation** or a **Comparison Expression**. (Hint: If you are unfamiliar with the function, use `help()`.)

a) `str()` b) `group_by()` c) `rank()` d) `mean()` e) `filter()` f) `summary()` g) `summarise()` h) `anti_join()` i) `glimpse()`

Problem 9.2

In the ranked-choice ballot system, once a voter has picked a first choice candidate, there is no advantage in listing that same candidate as second or third choice.

In answering each of the following, you should include three things:

1. A statement in English (using data verbs!) of your strategy for carrying out the calculation.
2. The implementation of the calculation in R.
3. The data table that is the result of the calculation. This will be printed automatically from (2). You **do not** have to format the result beautifully. It's sufficient to show the first few lines in the data table appear. (Remember `head()`).

Here are the questions.

- How many people chose the same candidate for both `First` and `Second` place?
- Of the ballots where the `First` and `Second` place choices are the same, what were the top 3 choices?
- Of the people who selected Ole Savior for `First`, what were the top three `Second` choices?

Problem 9.3

These questions refer to the `diamonds` data table in the `ggplot2` package. Take a look at the codebook (using `help()`) so that you'll understand the meaning of the tasks. (Motivated by Garrett Grolemund.)

Each of the following tasks can be accomplished by a statement of the form

```
diamonds %>%
  verb1( args1 ) %>%
  verb2( args2 ) %>%
  arrange(desc( args3 )) %>%
  head( 1 )
```

For each task, give appropriate R functions or arguments to substitute in place of `verb1`, `verb2`, `args1`, `args2`, and `args3`.

1. Which color diamonds seem to be largest on average (in terms of carats)?
2. Which clarity of diamonds has the largest average "table" per carat?

Problem 9.4

Using the `ZipGeography` data table, answer the following questions. In addition to the answer itself, show the statement that you used and the data table created by your statement that contains the answer.

- How many different counties are there?
- Which city names are used in the most states?
- Which city names with more than 5% of the state population are used in the most states?
- Does any state have more than one time zone?
- Does any city have more than one time zone?
- Does any county have more than one time zone?

10.4 Exercises

Problem 10.1

Most data verbs, when used with the chaining syntax %>%, have arguments that consist only of reduction and transformation functions, constants, and variables. For instance:

```
BabyNames %>%
  group_by(year) %>%
  summarise(total = sum(count))
```

In contrast, the join family of data verbs — `inner_join()`, `left_join()`, etc. — always have a data table as one of the arguments inside the parentheses. Explain why.

Problem 10.2

Consider these two tables containing demographic and geographic information about countries.

`Demographics`

country	pop	area
Afghanistan	31822848	652230
Akrotiri	15700	123
Albania	3020209	28748
Algeria	38813722	2381741
American Samoa	54517	199
... and so on for 256 rows		

`CountryCentroids`

name	iso_a3	long	lat
Afghanistan	AFG	66.17	33.78
Aland	ALA	19.97	60.20
Albania	ALB	20.26	41.14
Algeria	DZA	2.83	28.14
American Samoa	ASM	-170.72	-14.30
... and so on for 241 rows			

Explain why the information in the two tables cannot be successfully combined by laying the two tables side by side into a single table, that is, by simply copying the `long` and `lat` variables from one table and pasting them alongside the `country`, `pop` and `area` variables in the other table.

Problem 10.3

Here are three tables, **A**, **B**, and **C**, with different organizations of the same data:

Data Table A

Year	Algeria	Brazil	Columbia
2000	7	12	16
2001	9	14	18

Data Table B

Country	Y2000	Y2001
Algeria	7	9
Brazil	12	14
Columbia	16	18

Data Table C

Country	Year	Value
Algeria	2000	7
Algeria	2001	9
Brazil	2000	12
Columbia	2001	18
Columbia	2000	16
Brazil	2001	14

1. Which table format do you think would make it easiest to find the change from 2000 to 2001 for each country. How would you do it?
2. Suppose you have another table, `ContinentData`, which gives the continent that each country is in. Which table format do you think would make it easiest to find the sum of the values for each continent for each of the years? How would you do it?

11.3 *Exercises*

Problem 11.1

Here are three data tables with the same information:

Version One

name	sex	year	nbabies
Harrison	F	2012	15
Harrison	M	1912	170
Harrison	M	2012	2120
Roderick	M	1912	46
Roderick	M	2012	202
... and so on for 9 rows			

Version Two

name	year	F	M
Harrison	1912	NA	170
Harrison	2012	15	2120
Roderick	1912	NA	46
Roderick	2012	NA	202
Terry	1912	17	49
... and so on for 6 rows			

Version Three

name	sex	1912	2012
Harrison	F	NA	15
Harrison	M	170	2120
Roderick	M	46	202
Terry	F	17	17
Terry	M	49	479

a. What is the meaning of a case in each of the tables?
   - Version One
   - Version Two
   - Version Three

b. Comparing Version One to Version Two, which table is narrow and which one is wide?

c. What "key" variable from the narrow table is being used?

d. There are no `NA`s in Version One, but there are in Versions Two and Three. Why?

e. Version Two has 6 cases, while Version 3 has only 5 cases. How can they contain the same information?

f. Version Three was "spread" from Version One. What variable was used to denote the spread columns?

g. Version One can be created by gathering columns from Version Two.
   - Which variables from Two were gathered into One?
   - What "key" variable, not explicitly named in Version Two, does appear in Version One?
   - Where were the values taken from Version Two to use as levels in the key variable created for Version One?

Problem 11.2

- Suppose you want to create the following table with the most popular name of either sex each year

name	sex	year	nbabies
Harrison	F	2012	15
Roderick	M	1912	46
Roderick	M	2012	202
Terry	F	1912	17

What should the chain of commands look like to make this from the data table "Version One" in the previous exercise?

- Suppose you want to calculate the ratio of male to female in each name in each year. Like this:

name	year	ratio
Harrison	1912	NA
Harrison	2012	0.01
Roderick	1912	NA
Roderick	2012	NA
Terry	1912	0.35
... and so on for 6 rows		

- Would you rather start from "Version Two" or "Version Three"?
- If you were given "Version One", would you rather work directly on that with the data verbs or, first, translate to one of the other forms?

Problem 11.3

Comparing each of the following *pairs* of tables, say which one is wide and which one is narrow. a. **A** versus **C** b. **B** versus **C** c. **A** versus **C**

Data Table A**

Year	Algeria	Brazil	Columbia
2000	7	12	16
2001	9	14	18

Data Table B

Country	Y2000	Y2001
Algeria	7	9
Brazil	12	14
Columbia	16	18

Data Table C

Country	Year	Value
Algeria	2000	7
Algeria	2001	9
Brazil	2000	12
Columbia	2001	18
Columbia	2000	16
Brazil	2001	14

Problem 11.4

Consider the data table `BP_wide` in Table 11.1. Using paper and pencil, not the computer, sketch out what will be the result of this (unfortunate) conversion from wide to narrow:

```
BP_wide %>%
  gather(key = when, value = sbp,
         subject, before, after)
```

Hint: The name `when` for the key, and `sbp` for the value don't reflect what's actually in the result.

Problem 11.5

Here are two tables containing information relevant to Table 11.3 in the text.

`Measurements`

subject	what	value	date
BHO	sbp	160	2007-06-19
GWB	sbp	115	1998-04-21
BHO	sbp	155	2005-11-08
WJC	sbp	145	2002-11-15
WJC	sbp	130	2013-09-15
... *and so on for 16 rows*			

`Treatments`

subject	treatment_date
BHO	2012-08-05
GWB	2005-11-14
WJC	1998-09-30

These are concise forms for organizing the data with several advantages:

1. The use of `treatment_date` avoids possible areas when converting the measurement date to "before" or "after".
2. Additional measurements (e.g. respiration rate, white blood cell count, ...) can easily be included.
3. Addition information about each measurement (e.g. the name of the technician performing the measurement) can be easily included.

You can access the complete tables with this statement, which will create `Measurements` and `Treatments`.

```
"http://tiny.cc/dcf/MeasTreatTables.rda" %>%
  url() %>% load()
```

The `when` variable in Table 11.3 describes whether the measurement date was before or after the treatment date.

Your task: Using the `Measurements` and `Treatment` tables, reconstruct Table 11.3. (Hint: the dates in both `Measurements` and `Treatments` are stored in a way that you can use the numerical comparison function `>` to determine whether one date is after another.)

12.1 *Exercises*

Problem 12.1
For each sex, find the 5 most popular names in `BabyNames` adding up over all the years.

Problem 12.2
Using `BabyNames`, for each year, find the fraction of all babies born in that year who were given a name in the top 100 for that year. Make a graph showing how this fraction has changed over the years.

To start, produce a data table that looks like this:

year	frac_in_top_100	total
1880	FALSE	68667
1880	TRUE	132817
1881	FALSE	66340
1881	TRUE	126360
1882	FALSE	77258
	... and so on for 268 rows	

Then you can use `spread()` and `mutate()` to find the fraction of babies with names in the top 100 each year.

```
GlyphReady <-
  PopularCounts %>%
  spread(frac_in_top_100, total) %>%
  mutate(frac_in_top_100 = `TRUE` / (`TRUE` + `FALSE`))
```

Problem 12.3
For each of the operations listed here, say whether it involves a transformation function or a summary function or neither.

a. Determine the 3rd largest.
b. Determine the 3rd and 4th largest values.
c. Determine the number of cases.
d. Determine whether a year is a leap year.
e. Determine whether a date is a legal holiday.
f. Determine the range of a set, that is, the max minus the min.
g. Determine which day of the week (e.g., Sun, Mon, ...) a given date is.
h. Find the time interval in days spanned by a set of dates.

13.4 *Exercises*

Problem 13.1

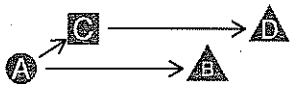
Consider this table of edges:

from	to
China	Japan
USSR	Japan
Germany	USA
France	Germany
Italy	France
Germany	UK
Japan	UK
USA	Japan
USSR	Germany

1. How many distinct vertices are there?
2. How many edges are there?

Problem 13.2

For this network ...



1. What are the vertices?
2. Which of these tables shows the edges correctly?

Table 1

A	C
B	C
A	B

Table 2

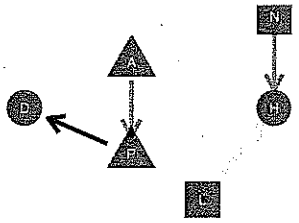
A	C
A	B
C	D

Table 3

A	C	D
A	B	

3. Explain what's wrong with each of the other tables.

Problem 13.3



Which table gives the information needed to draw the edges in the graph?

Table 1

From	To
A	P
H	D
L	H
A	N

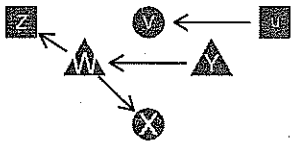
Table 2

From	To
A	P
P	D
L	H
N	H

Table 3

From	To
P	A
P	N
H	L
N	H

Problem 13.4



Here is a table of the vertices in the network above:

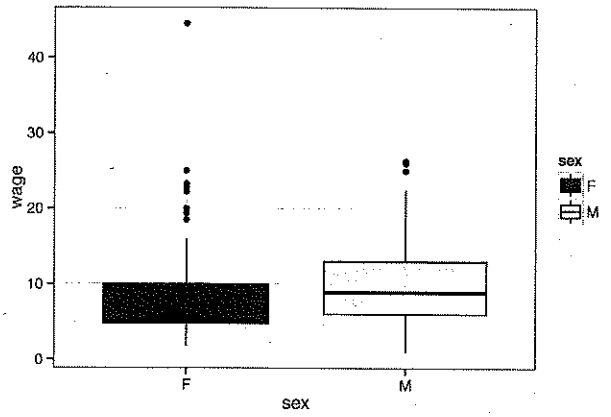
country	pop	area	exports	roads
U	3	7	1	3
V	1	4	2	2
W	2	2	4	1
X	1	1	4	3
Y	2	1	3	4
Z	3	3	2	5

1. Which variable is mapped to the size of the letter?
2. Which variable is mapped to the shape of the gray symbol?

14.6 *Exercises*

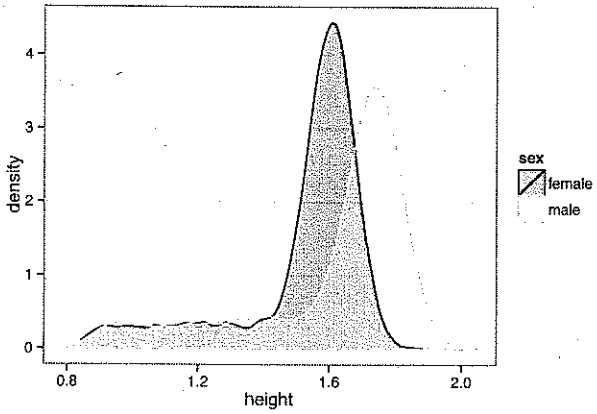
Problem 14.1

Reconstruct this graphic using the `ggplot()` system and the `mosaicData::CPS85` data table.



Problem 14.3

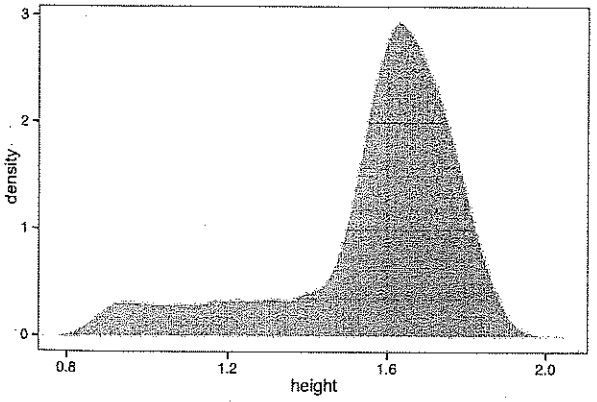
Based on the following graph, what's the most likely height for women? For men?



This and several following problems concern the `NCHS` data in the `DataComputing` package.

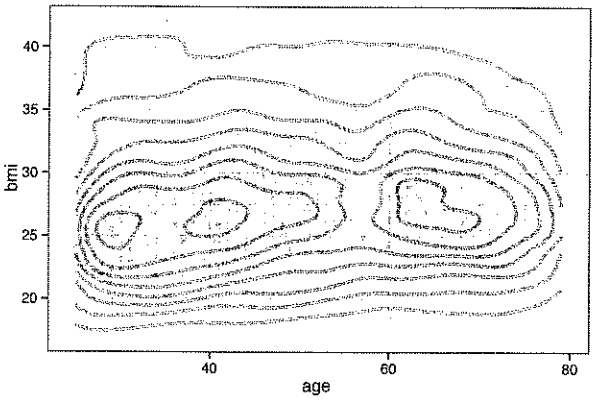
Problem 14.2

Judging from the graph, what's the most likely height among the `NCHS` people? (FYI: The heights are in meters.)



Problem 14.4

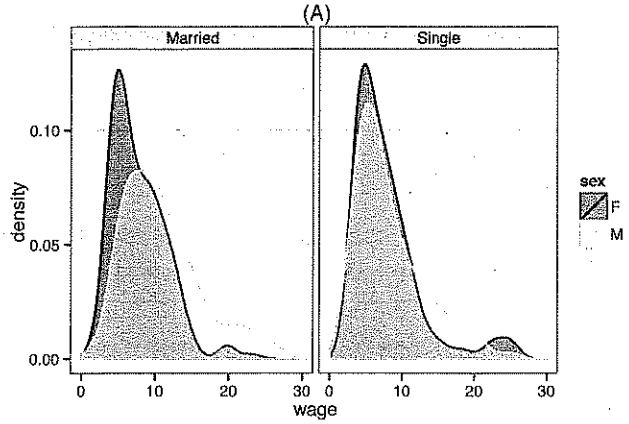
Below is a plot of the density for `bmi` and age simultaneously. The curved lines play the role of contours on a contour map of geography. Based on this graph, what's the most likely BMI for a 40-year old? For a 70-year old?



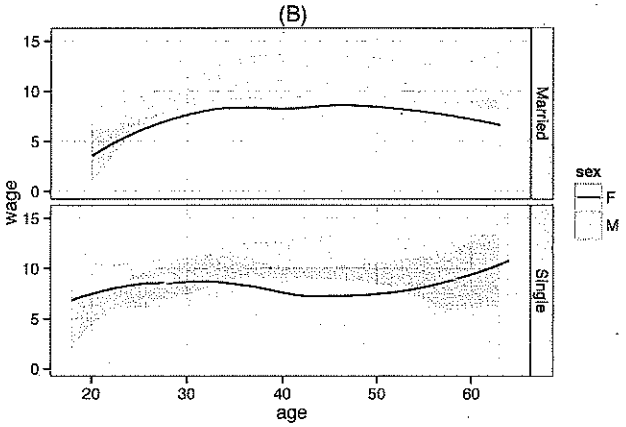
Problem 14.5

Here are several graphics based on the `mosaicData::CPS85` data table. Write `ggplot2()` statements that will construct each graphic.

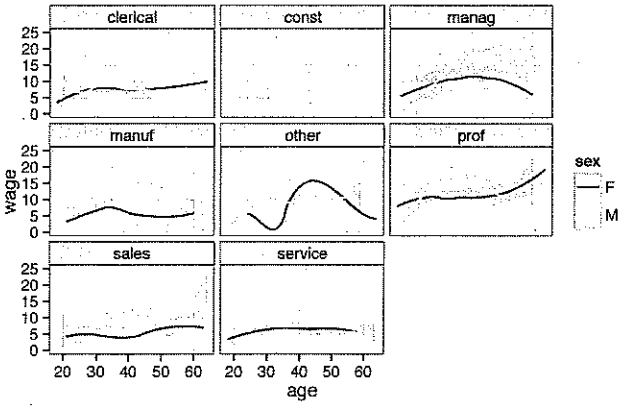
(A)



(B)

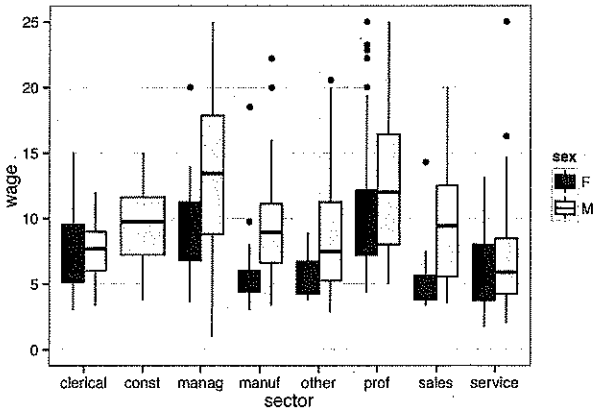


C



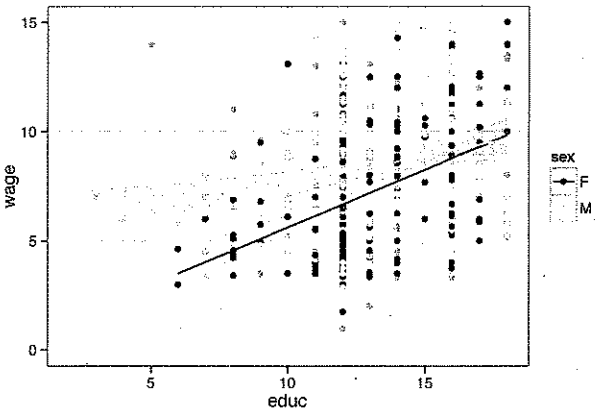
Problem 14.6

Reconstruct this graphic using the `ggplot()` system and the `mosaicData::CPS85` data table.



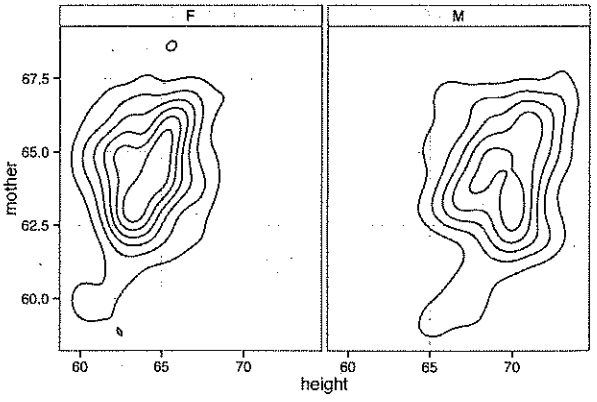
Problem 14.7

Reconstruct this graphic using the `ggplot()` system and the `mosaicData::CPS85` data table.



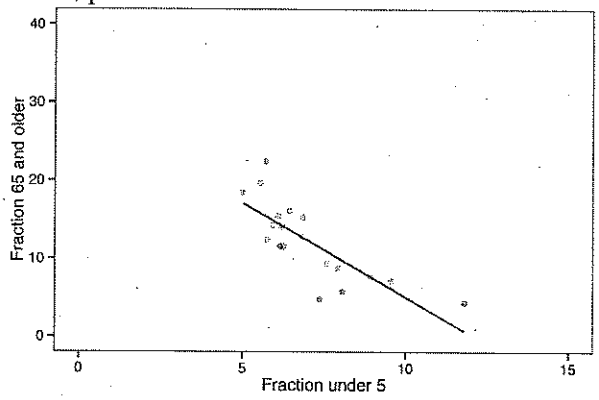
Problem 14.8

Reconstruct this graphic using the `ggplot()` system and the `mosaicData::Galton` data table.



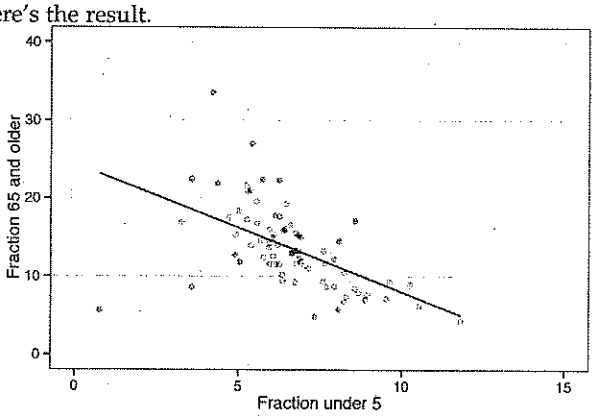
Problem 14.9

You're doing a study of whether grandparents go where the grandkids are. As a first attempt, you look at 20 ZIP codes. You find the relationship between the fraction of people in a ZIP code under 5 years old and the fraction 65 and older, plotted below.



- Do the data indicate that ZIP codes with high elderly populations tend to have high child populations?
- Looking at the confidence bands, is your data possibly consistent with there being no relationship (that is, a level line) between elderly population and child population?

You decide to get more data: study 80 ZIP codes. Here's the result.



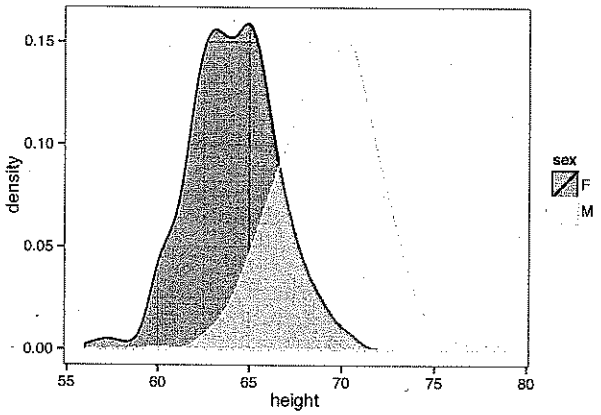
a. Is a flat line consistent with the data?

b. Compare the height of the confidence band with 20 ZIP codes to the height of the band with 80 ZIP codes: 4 times as much data in the larger sample than the smaller. Roughly, what's the ratio of confidence band heights?

c. Statistical theory indicates that the width of a confidence band based on n points goes as $1/\sqrt{n}$. Does this seem about right?

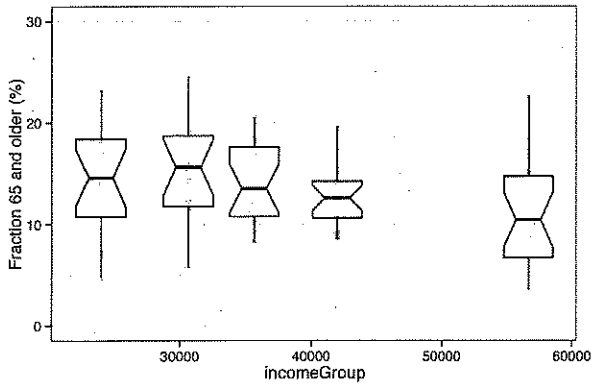
Problem 14.10

Reconstruct this graphic using the `ggplot()` system and the `mosaicData::Galton` data table.



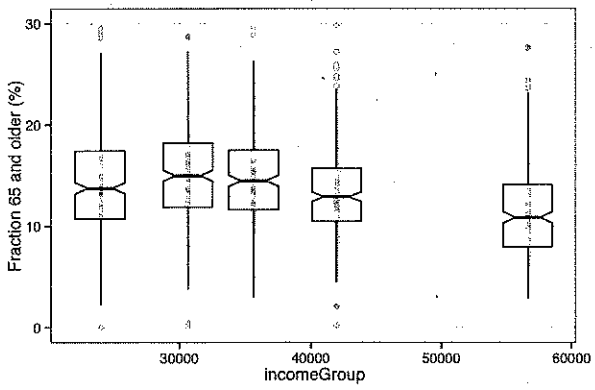
Problem 14.11

Studying a sample of 100 ZIP codes, the following graphic divides ZIP codes into 5 income groups, looking at the fraction of the population that is 65 and older.



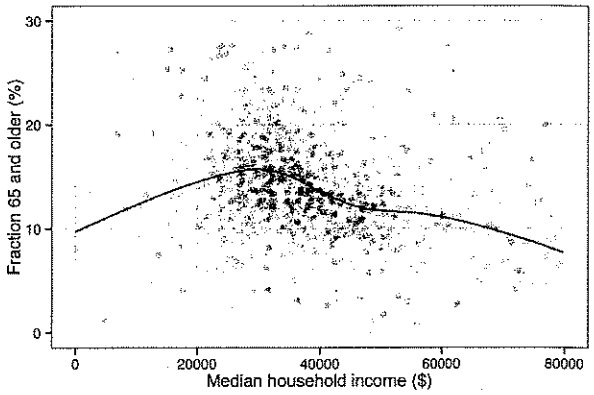
(a) What relationship do you see between income and the fraction of the population greater than 65? (The notch shows the confidence interval on the median population.)

The same analysis, but for a sample 16 times as big as before.



(b) Do the notch sizes (confidence intervals) follow the $1/\sqrt{n}$ rule?

Extra information: Often, quantitative variables such as income are divided into groups. Sometimes this reflects a lack of awareness by the data analyst about other possibilities. For instance, smoothers can provide an indication of general trends in the data without dividing into groups. For instance, this graph has `income` rather than `incomeGroup` on the x-axis.



Problem 14.12

One way to measure the "spread" of values is with the "standard deviation." The higher the standard deviation, the more spread out the values are. Use `sd()`, a summary function, to compute the standard deviation.

In the `MedicareCharges` data table, the `drg` variable describes billing codes for different types of medical conditions, called "Direct Recovery Groups."

- Taking all the hospitals together, which DRG has the highest standard deviation in Medicare charges.
- Taking all the DRGs together, which hospital has the highest standard deviation in Medicare charges.
- Medicare typically pays the hospital less than was charged by the hospital. Calculate the proportion of the charge that was paid (hint: a simple ratio), then find,
  - across all hospitals, which DRG has the
    - lowest proportion
    - highest proportion
  - across all DRGs, which hospital has the
    - lowest proportion
    - highest proportion

Problem 14.13

Figure 8.3 shows height versus age for smokers and non-smokers in the `NCHS` data.

Make a similar graphic that shows confidence bands of a smoother. Interpret the graph to conclude at what ages the smoker and non-smoker groups differ systematically in height.

15.4 *Exercises*

Problem 15.1

Here are some character strings containing times or dates written in different formats.

For each date, choose an appropriate function from the `lubridate` package to translate the character string into a date-time object. Then, using subtraction, find the number of days between that date and your birthday.

- `"April 30, 1777"` Johann Carl Friedrich Gauss
- `"06-23-1912"` Alan Turing's birthday
- `"3 March 1847"` Alexander Graham Bell's birthday
- `"11:00 am on Nov. 11th, 1918 at 11:00 am"` Armistice ending World War I on the Western Front.
- `"July 20, 1969"` First manned moon landing

Example:

```
lubridate::ymd("1941-09-01") -
  lubridate::mdy("July 28th, 1914")
```

```
Time difference of 9897 days
```

Problem 15.2

Here are some strings containing numerical amounts. For which ones does `as.numeric()` work correctly? How about `tidyr::extract_numeric()`?

a. `"42,659.30"`

b. `"17%"`

c. `"Nineteen"`

d. `"£100"`

e. `"9.8 m/seconds-square"`

f. `"9.8 m/s^2"`

g. `"6.62606957 × 10^-34 m2 kg / s"`

h. `"6.62606957e-34"`

i. `"42.659,30"` (A European style)

Problem 15.3

Grab Table 4 (or another similar table) from the Wikipedia page on world records in the mile (or some similar event). Make a plot of the record time versus the date in which it occurred. For extra credit, mark each point with the name of the athlete written above the point. (Hint: Use `geom_text()`)

Some tips To convert time entries such as "4:20.5" into seconds, use the `lubridate` package's `as.duration(ms("4:20.5"))`.

You can get rid of the footnote markers such as `[5]` in the dates with a statement like this:

```
T4 <-
  T4 %>%
  mutate(Date = gsub("\\[.\\]$", "", Date))
```

The `gsub()` transformation function replaces the characters identified in the first argument with those in the second argument. The string `"\\[.\\]$"` is an example of a "regular expression" which identifies a pattern of characters, in this case a single character in square brackets just before the end of the string. (Chapter 16 describes regular expressions in more detail.)

Problem 16.1
Using the `BabyNames` data table, find the 10 most popular

1. Boys' names ending in a vowel.
2. Names ending with "joe" or "jo" (like BettyJoe)

Problem 16.2
Here is a character string with a regular expression:
`"([2-9][0-9]{2})[- .]([0-9]{3})[- .]([0-9]{4})"`
To explain the first bit ...

- `[2-9]` means "one digit from 2 to 9.
- `[0-9]` refers to one digit from 0 to 9.
- `[0-9]{2}` refers to two consecutive digits, 0 to 9.
- `[2-9][0-9]{2}` means one digit 2 to 9 followed by two digits 0 to 9
- `[- .]` means "any of the characters dash, space, period, just once.
- The parentheses refer to the matching contents to be extracted. The whole expression has the structure `(stuff)[- .](more stuff)[- .](still more stuff)`. The three sets of parentheses mean to extract those three pieces from strings thatmatch.

Explain what familiar kinds of strings the entire general expression would match. (Hint: Call me maybe.) What components of those strings are being extracted?

Problem 16.3
Consider this regex:

```
(A[LKSZRAP]|C[AOT]|D[EC]|F[LM]|G[AU]|HI|I[ADLN]|
K[SY]|LA|M[ADEHINOPST]|N[CDEHJMVY]|O[HKR]|P[ARW]|
RI|S[CD]|T[NX]|UT|V[AIT]|W[AIVY])
```

(Ignore the line breaks)

1. How long will the strings be that match the pattern?
2. How many different strings will match?
3. People living in the United States may be able to figure out what the pattern is meant to express. Give it a try.

Problem 16.4
A list of names from the Bible can be accessed like this:

```
BibleNames <-
  read.file("http://tiny.cc/dcf/BibleNames.csv")
```

Using the names in `BibleNames`,

1. Which names have any of these words in them: "bar", "dam", "lory"?
2. Which names *end* with those words?

You need only show a few.

Problem 16.5
The city of Boston publishes various public-safety data online. A data table listing almost 300,000 crime reports from Feb. 6, 2012 up through the present is available via `https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx`

A small, convenient extract of the Boston crime data is available to you:

```
CrimeSample <-
  read.file("http://tiny.cc/dcf/Boston-Crimes-50.csv")
```

The `Location` variable contains information about latitude and longitude. Each of these is a number, but they are represented in `Location` as a formatted character string.

Write a regular expression that will extract the latitude and longitude as numbers into separate variables. To do the extraction you can use `tidyr::extract()`, e.g.

```
my_regex <- # Your regular expression goes here
CrimeSample %>%
  tidyr::extract("Location", into=c("lat", "long"),
                 regex = my_regex,
                 convert = TRUE)
```

Some hints:

- You'll need the extraction parentheses, written as plain parens: `( )`. If you want to refer to the parentheses characters, not as extraction markers but as plain text, you need to "escape" them with backslashes, e.g. `"\\(some pattern in parens\\)"` The two backslashes are needed so that R realizes that you are escaping the character that follows.
- The regex symbol for a digit is [0-9].
- A regex that will extract a single floating-point number surrounded by parentheses is: `"\\([+-]*[0-9]*[0-9\\.][0-9]*\\)"`.

17.1 Exercises

Problem 17.1

A risk factor is a characteristic associated with an increased likelihood of developing a disease or injury. In many situations, identifying risk factors provide an important means to screen those at high risk. In this exercise, you'll use the NCHS data to look for risk factors of diabetes.

In a large dataset such as NCHS with many variables, there are often some variables with much missing data. For instance, with NCHS, there are `nrow(NCHS)` cases, but many fewer that have no missing values. The `na.omit()` data verb filters out only those cases that have no missing values:

```
NCHS %>%
  na.omit() %>%
  nrow()
```

```
[1] 12013
```

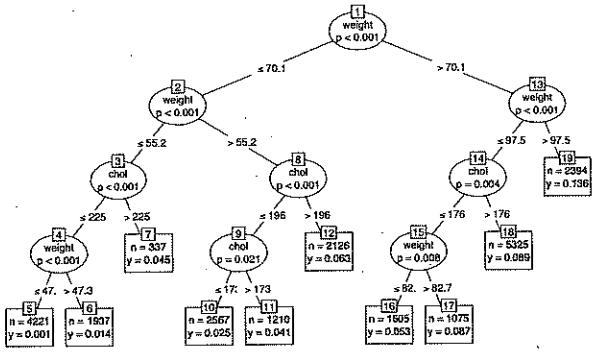
If only a few variables are of interest, it can be helpful to select out those variables when filtering out missing data.

```
CompleteCases <-
  NCHS %>%
  select(diabetic, weight, age,
         bmi, chol, smoker) %>%
  na.omit()
CompleteCases %>% nrow()
```

```
[1] 22797
```

Here is one model that considers weight and cholesterol levels as risk factors for diabetes:

```
mod1 <- party::ctree(
  diabetic ~ weight + chol,
  data=CompleteCases)
plot(mod1, type="simple")
```



In interpreting this tree, note that the quantity y indicates the probability of people in the group having diabetes, while n gives the size of the group.

A useful risk factor splits the population into groups with very high and with very low probabilities of diabetes. For instance, node 19 marks a group with a 13.6% risk of diabetes, substantially larger than other groups. Using other variables in the model, such as `smoker` or `age` might do a better job of dividing the cases into groups with high and low risk.

One way to measure the effectiveness of the model is with a quantity called the "log likelihood." The log likelihood compares, for all the cases individually, the actual outcome against the probability predicted by the model. A high log likelihood indicates a more successfully predictive model. Without going into the theory behind log likelihood, you can still calculate the value.

```
CompleteCases %>%
  mutate(probability = as.numeric(predict(mod1)),
         likelihood =
           ifelse(diabetic,
                  probability,
                  1-probability)) %>%
  summarise(log_likelihood = sum(log(likelihood)))
```

log_likelihood
-4450.64

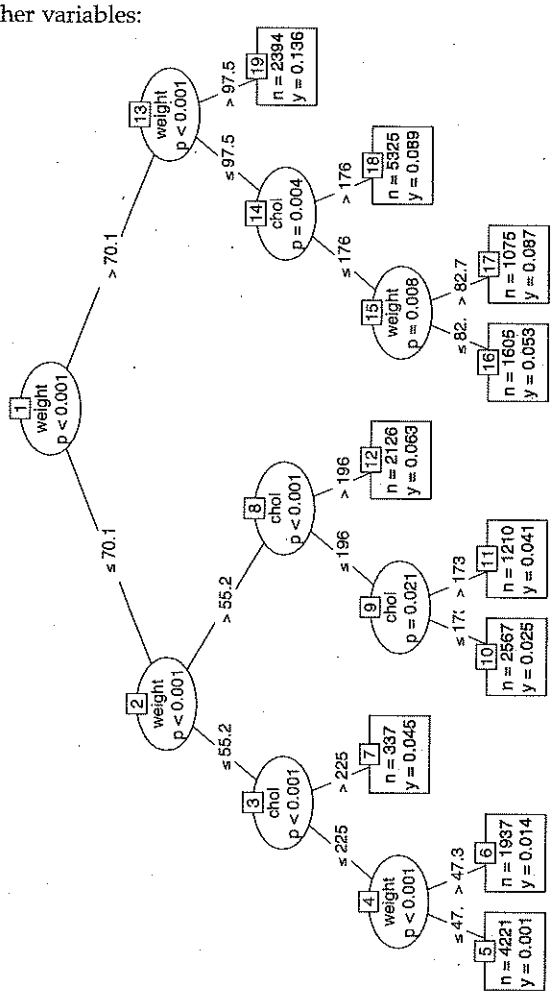
The number is meaningful only in comparison to the log likelihood for other models. Explore other models for diabetes using different combinations of potential risk factors. Find one with a higher log likelihood than `mod1`. Remember that the log likelihood number will always be negative, so -3000 is much bigger than -4000.

Problem 17.2

Consider these data on houses for sale:

```
Houses <-
  read.file("http://tiny.cc/dcf/houses-for-sale.csv")
```

Here is a model of house price as a function of several other variables:



In the model tree, the terminal node, shown as a rectangle, gives information about n the number of houses included in that group and y the mean price of the houses in that group. Based on the model tree, answer these questions:

1. What variables are in the model?
2. For houses with a living area less than 1080 square feet, does the number of bathrooms make a difference in price?
3. For houses with a living area between 1080 and 1483 square feet, how much is the typical difference in price between houses with 1 1/2 baths and houses with 2 bathrooms?
4. Is having a fireplace associated with a higher house price? What other variables does this depend on?