Started on	Wednesday, June 12, 2019, 11:19 PM
State	Finished
Completed on	Thursday, June 13, 2019, 12:04 AM
Time taken	45 mins 45 secs
Points	10.00/10.00
Grade	**100.00** out of 100.00

Information

Start reading the chapter. Be sure you get what a "vector" is, on p 28:

> A vector is like a single line of data, with the position of each data item being important. As you'll see, you can reference that position, to retrieve the data.

Then try typing-in the examples on p 29, and answer the first question when you get to the bottom.

It's ok to make mistakes, by the way! You'll learn a lot, as a result, about how R likes things.

Question **1**

Correct

1.00 points out of 1.00

If you type "flu_status" into R, after doing all the work on p 29, you should see, "[1] FALSE FALSE TRUE".

Select one:

- (•) True ✔
- ( ) False

That's what I got. R simply repeats the values you typed in, toward the top of the page, since nothing got changed in between.

The correct answer is 'True'.

Information

At the top of p 30, Lantz reminds you where to get the source code and data for each chapter. You also can use the places we've provided on Moodle for this, like "Module 2 - Ch 2 data and R-code".

> You also can open the R for Ch 2, but we recommend you do that in a separate text editor. That way, if you have trouble entering a line in R, you can copy and paste it from the author's own code. However, if you download everything in the file into R, you'll be missing the activity yourself!

Read the section on Factors, and try the examples provided.

> A "factor" is really just a vector, but R stores it in an efficient way.

Then do the next question.

Question **2**

Correct

1.00 points out of 1.00

In the line,

```
symptoms <- factor(c("SEVERE", "MILD", "MODERATE"),
                   levels = c("MILD", "MODERATE", "SEVERE"),
                   ordered = TRUE)
```

what does it mean that the levels are "ordered"?

Select one:

- a. They are first, second, and third in the vector.
- b. Each one is twice as big as the previous one.
- c. Generally, they have a ranking you can test for. ✔ Yes, this is it. We don't know how MUCH bigger each is than the previous one, but we do know they are in that order.

Your answer is correct.

The correct answer is: Generally, they have a ranking you can test for.

Information

We didn't say that all the data in a vector had to be the same type, but that's the case.

If you want an ordered list of data with different types, use a "list."

Try the activities on pp 32-34, then answer the next question.

Question **3**

Correct

1.00 points out of 1.00

A list is very much like a particular line of data about some individual, as in the author's "subject1" example.

Select one:

- True ✔
- False

That's the idea. You can picture this one being a record of health info about "John Doe".

The correct answer is 'True'.

Information

You knew we had to be building up to something big, and here it is. "Data frames" are a lot like the tables you'd see in Excel, with rows and columns meaning something. They are the basic thing that most ML algorithms work on, in R.

As the author says, usually these are like rows of instances of data, with each one having the same columns. So,

- Going down the rows, you can think of data frames as a big batch of R "lists", with each list having the same columns in it. Or,
- Going across the columns, each column is an R vector, and the whole data frame is a list of those vectors.

Read about these on pp 35-37, then answer the next question.

Hint: If you have any trouble with the first line, "pt_data <- data.frame(subject_name, ...)", it might be because you didn't define some of the vectors in it, like "blood" or "symptoms". You can go back and define those, until this line works!

Hint: If you want to go back to a line you tried typing, after it didn't work, and you think you fixed the reason, you can use the up-arrow multiple times, to get back to that line.

Question **4**

Correct

1.00 points out of 1.00

After you have defined the data frame "pt_data", if you type "pt_data[, 3]", R gives you:

[1] 99.1 98.6 101.4

Select one:

- True
- False ✔

No, that would give you the flu_status for each line. The temperatures are in column 2.

The correct answer is 'False'.

Information

You should also know about matrices, because some of our ML algorithms use these.

A matrix has rows and columns, like a data set, but they all have the same kind of data in them. Usually, numbers!

Read about matrices on pp 37-39, then answer the next quiz question.

Question **5**

Correct

1.00 points out of 1.00

In R, if you were to define

```
my_matrix <- matrix(c(1,2,3,4,5,6), nrow = 2)
```

and then looked at it by typing its name

```
my_matrix
```

you would see

```
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

Select one:

- True
- False ✔

No, that's what you would see if you had defined

```
my_matrix <- matrix(c(1,2,3,4,5,6), nrow = 3)
```

giving the matrix 3 rows, instead of 2.

The correct answer is 'False'.

Information

Read about saving, loading and removing R data structures, on pp 39-40.

Remember the directory where you put the two Ch 2 files?

And remember that you set your R "working directory" to point to that?

Check that it's still pointing there, by typing "getwd()".

Type "list.files()" into R, and you should see the two data files you put in there from Moodle --

[1] "MLwR_v2_02.r" "usedcars.csv"

If you were to save a file, the way Lantz shows, at the bottom of p 39, these would go into that same working directory.

If you were to load a file, the way he shows on p 40, R would try to go get them there.

Try doing the "ls()" function -- you can verify that you have a lot of the same "objects" as shown in the book.

--> Take a screenshot of your RStudio window, displaying the results of this ls() function, to save for the Module 2 Homework.

Now, read about "Importing and saving data from CSV files, on pp 41-42.

Lots of our data files will be in "csv" format, like Lantz says.

Read the next section, "Exploring and understanding data".

Try the command shown. It should work without complaining!

Then answer the next question!

Question **6**

Correct

1.00 points out of 1.00

When you now do the "ls()" function, you will see all the same objects as before, plus one called "usedcars".

Select one:

- True ✔
- False

Absolutely. It's now in there. The next section lets us explore it.

The correct answer is 'True'.

Information

Read the 3 sections on pp 43-46, and do the actions like Lantz does.

As he suggests, str(usedcars) is very powerful, showing you the structure of the data, and examples from the first few rows.

The data actually goes the other way -- "year", "model", "price", and so on column headers. And the examples of each are from the initial rows.

As Lantz says, the people creating this data were kind -- they told us meaningful names for all those columns!

When you look at simple R functions like "summary(usedcars$year)", which give you some pretty nice information, you can see why R is preferred for doing statistical studies, as well as ML.

The fact R has basic functions like "mean" and "median" -- that's just a start!

Now read the next section, about "quartiles" etc., pp 47-49.

Make sure you try the "diff(range(usedcars$price))" function. It's a starting example of the real programming you can do in R. You can call one function on top of another like this, assuming the intermediate data is like what the second function, "diff", is looking for.

Now read about doing "boxplots", on pp 49-51.

See if you can get the two box-and-whiskers figures to appear, that Lantz shows.

In R-studio, these images appear in pop-up windows. You can do "Edit - Copy" to copy that image and put it into your own file. If you click on RStudio and Save, it will save the image to a pdf, in your directory of choice.

Read about "histograms", pp 51-53.

Try to create the histograms that Lantz does.

--> Save this histogram file to a good location (the ..Rstuff/Ch2 directory should be fine), to turn in as part of your Module 2 Homework.

Then try the next quiz question.

Question **7**

Correct

1.00 points out of 1.00

The histogram Lantz shows for used car prices shows "left skew".

Select one:

- True
- False ✔

The correct answer is 'False'.

Information

Read "Understanding numeric data..." and "Measuring spread... on pp 53-56.

Uniform distribution -- How often does that happen!?

Normal distribution -- We see approximations of this more often.

Don't get thrown off by the formulas for variance and standard deviation. You don't need to memorize these!

But, you should know that these are important and commonly used ways to measure how spread out a batch of data is.

If a distribution of data is approximately normal, then, as shown in the figure on p 56, about 34% of the data is 1 standard deviation above the mean, and another 34% is 1 standard deviation below the mean.

Read "Exploring categorial variables" and do the activities shown.

--> Note, the code at the bottom of p 57 in the book is incorrect, but it's correct in Lantz's Ch2 code file. Should be:

```
color_table <- table(usedcars$color)
color_pct <- prop.table(color_table) * 100
round(color_pct, digits = 1)
```

Read "Measuring the central tendency -- the mode" on pp 58-59.

Then answer the next question.

Question **8**

Correct

1.00 points out of 1.00

Use the "mode()" function to find the mode of a batch of data.

Select one:

- True
- False ✔

No, Lantz warns us that R uses "mode()" for something completely different! He suggests looking at the table output, and spot the category or categories with the greatest number of values.

The correct answer is 'False'.

Information

Read about "scatterplots" on pp 59-61.

Try the example Lantz provides, for "usedcars".

Once again, RStudio provides a very nice image, which you can save in different ways.

Note -- If the figure looks chopped off in any direction, try pulling on the edges to make it bigger.

You can be your own ML algorithm looking at scatterplots. Like, does it seem to have a straight-line relationship? Or, is there too much noise to call it anything? We'll get more scientific about such things.

Read about two-way cross-tabulations on pp 61-64.

Notice you need to install a package to do these! But, if you are online, it's easy.

A bunch of R lines in red go by. That's ok. It should end up saying something like, "The downloaded binary packages are in " and the name of a directory. Success!

Don't forget to type the "library(gmodels)" line, too, to actually load this into R!

The steps after that, on most of p 62, are "data preparation" for creating the "CrossTable" we would like. As Lantz says, we need to reduce the number of car colors to something more general, to answer the question of interest, about people choosing "conservative" colors depending on the model of car they bought.

This data preparation is very important -- if you try to do this kind of table with your own data, for the term project, you will likely be facing some similar kinds of steps. (Probably not exactly the same ones.)

After you've done the "usedcars$conservative ..." operation shown on p 62, try the next quiz question.

Question **9**

Correct

1.00 points out of 1.00

What does the command --

```
usedcars$conservative <-
  usedcars$color %in% c("Black", "Gray", "Silver", "White")
```

do to change "usedcars" itself?

It adds more data to it!

Select one:

- True ✔
- False

Absolutely. If you do "str(usedcars)" after running this, you'll see a column called "conservative", on every row, with TRUE or FALSE.

The correct answer is 'True'.

Information

Now do the CrossTable function at the bottom of p 62. You should get the table shown on p 63, appearing in the same RStudio window where you typed the command.

--> Take a screenshot of your Crosstable, to save for the Module 2 homework.

This is a text image, but you can of course copy it and put it somewhere else, if you like.

Lantz does explain the "chi square contribution", on p 64. They have to do with the independence of the variables in each of the cells.

The row totals under them are what tell the percentage of cars, by model, that had "conservative" colors, or not.

Try the last question in the quiz, about this table!

Question **10**

Correct

1.00 points out of 1.00

Lantz looks at the very close row values, to say that there isn't much difference. But he then says you can do the "chi-squared" test to decide this more precisely.

Select one:

- True ✔
- False

Yes, and he concludes there was about a 93% probability that the variations in the cell counts, by row, were due to chance.

The correct answer is 'True'.

Question **11**

Complete

Not graded

We want to base online and remote face-to-face discussions on the topics of most value to you.

Please think carefully about all the material you read, then write a prompt for discussion you would like to hear - either:

a. Something that you aren't sure about, which you'd like to have explained in class, or

b. A topic you liked a lot, that you'd like to discuss in class.

I would love to hear more about Pearson's Chi-squared test and how this method is developed.

Thanks!