

R - Getting Started CONTINUED

TRUE and FALSE (logical) data

Vectors can be assigned logical measurements, directly or as the result of a logical operation. Here's an example of direct assignment. You have seen some of this before. Here is a reminder.

```
z <- c(TRUE, TRUE, FALSE) # put 3 logical values to a vector z
```

Logical operations can identify and select those vector elements for which a condition is TRUE. The logical operations are symbolized as follows

```
== (equal to)
!= (not equal to)
< (less than)
<= (less than or equal to)
```

and so on.

The logical operators "&" and "|" refer to AND and OR. For example, put the following numbers into a vector "z",

```
z <- c(-10, -5, -1, 0, 3, 92)
```

What do the following operations yield?

```
z < 0 & abs(z) > 5

z < 0 | abs(z) > 5

z[z < 0 | abs(z) > 5]
```

Combine vectors to make a data frame

Multiple vectors representing different measurements (variables) made on the same individuals can be placed into columns of a data frame. A data frame is a spreadsheet-like object for a data set of potentially many variables. We will see more about data frames later. Here we show how to make a data frame by combining vectors of the same length. The vectors need not be of the same data type.

First, obtain your vectors. For example,

```
custno <- c(1:7)

loyalty_level <- c(100,100,90,90,40,40,50)

location<-
c("madrid","madrid","london","london","dublin","paris","paris")
```

Then combine them into a data frame named `mydata`.

```
mydata <- data.frame(customer_no = custno, loyalty_level =
loyalty_level, city = location, stringsAsFactors = FALSE)
```

The argument `stringsAsFactors = FALSE` is optional but recommended to preserve character data (otherwise character variables are converted to factors).

Getting Data Into R

Data.Frames

One of the most useful features of R is the `data.frame`. On the face of it a `data.frame` is just like an Excel spreadsheet in that it has columns and rows. In statistical terms, each column is a variable and each row is an observation.

In terms of how R organises `data.frames`, **each column is actually a vector**, each of which is the **same length**. That is very important because it lets each column hold a different type of data i.e. numeric data or character data. This also implies that within a column each element must be of the same type, just like with vectors.

There are many ways of to construct a `data.frame`, the simplest being to use the `data.frame` function. Let us create a basic `data.frame` using some of vectors `x`, `y`, `z`

```
#creating a data.frame from vectors creating a 10 X 3 data.frame
x <- 10:1
y<- -4:5
z <- c ("football", "golf", "tennis", "soccer", "hurling", "rugby",
"athletics", "fencing", "swimming", "basketball")
```

For you to do!

- Create a data frame called `sports_mydf`. The names of the columns are `sport_id`, `ranking`, and `sport`.
- Try out the following following functions `nrow()`, `ncol()`, `dim()`, `class()`, `head()`, `tail()`, `str()`
- Find out about the `rownames()` function and change the row name to one, two three etc.
- Reset the row names to the default values

Sol

```
mydf <- data.frame(x,y,z)
mydf
```

set the column names

```
mydf <- data.frame(sport_id =x, ranking =y, sport=z, stringsAsFactors = FALSE )
mydf
```

finding information about the data.frame

```
nrow(mydf)
ncol(mydf)
dim(mydf)
```

set the row names and reset the rownames

```
rownames(mydf) <- c("one", "two", "three", "four", "five", "six", "seven",
"eight", "nine", "ten")
rownames(mydf)
rownames(mydf) <- NULL
rownames(mydf)
```

Reading Data from Files

There are a number of ways of reading data into R with CSV files one of the most common

Reading CSVs

One way to read data is to use `read.table`. The result of using `read.table` is a `data.frame`. The first argument is full path of the file to be loaded. The file can be sitting on disk or even on the web. Here is a simple CSV posted on the following web site. Let us read it into R using `read.table`.

#Read a CSV from the web and look at its contents

```
theUrl <- "http://WWW.jaredlander.com/data/Tomato%20First.csv"
```

note the arguments set i.e. the file has captions and it is comma separated

```
tomatoes <-read.table(file = theUrl, header=TRUE, sep =",")
head(tomatoes)
```

Let us use read.table and read.csv. Note read.csv calls read.table with some arguments preset. You will need to change the filepath to suit (download .csv file from moodle)

```
cars<-read.table(file = "c:/rdatasets/cars.txt", header=TRUE,
stringsAsFactors = FALSE, sep =",")
```

```
head(cars)

carscsv<-read.csv(file = "c:/rdatasets/cars.txt", stringsAsFactors = FALSE)
head(carscsv)
```

Check out cars\$brand and carscsv\$brand. What are your observations?

Frequency tables

These commands generate tables of frequencies or summary statistics. In the examples below, “x” is a single numeric variable. “A” is a categorical variable (factor or character variable) identifying different groups; “B” is a second such variable.

Frequency table and a Pie Chart

Frequency table for a categorical variable A. First let us load a file that hold weather data with a response variable of Play (yes, no) representing whether a sport was played based on the weather conditions namely outlook, temperature, humidity, and windy. There are two datasets representing the same data where one is nominal (i.e. categorical) while the other has some variables that are numeric.

For you to do!

- **Input dataset nominal weather (weathernominal.txt) into a dataframe called weathernom**
- **Show the first 10 records and ALL the columns**
- **Show the first 12 records and first 3 columns**
- **Create a table of counts for each discrete value of outlook e.g.**

```
A <- weathernom$outlook
ZZ <- table(A)
ZZ
```

- **#Let's show the data as a piechart - not the best visualisation to use! Good for getting a rough idea of proportions**

```
pie(ZZ, labels=names(ZZ), edges=200,
    col=c("yellow","red","navy"), radius= 0.9)
```

- **Try is out for some other variables**

Sol

#Input dataset nominal weather into a dataframe

```
weathernom <-read.table(file ="c:/rdatasets/weathernominal.txt",
stringsAsFactors=FALSE, sep =",", header=TRUE)
```

#Show the first 10 records and ALL the columns

```
head(weathernom)
weathernom[1:10,]
```

```

#Show the first 12 records and first 3 columns
weathernom[1:12,1:3]

#Create a table of counts for each discrete value of outlook
A <- weathernom$outlook
ZZ <- table(A)
ZZ

#Let's show the data as a piechart - not the best visualisation to use!
Good for getting a rough idea of proportions
pie(ZZ, labels=names(ZZ), edges=200, col=c("yellow","red","navy"), radius=
0.9)

```

Frequency (Contingency) table

The following command generates a frequency table for two categorical variables, A and B. The command can be extended to three or more variables.

```

B <- weathernom$play    # play is a Boolean indicating if a game was played

table(A, B)

```

To include the row and column sums in the contingency table, use the following commands.

```

mytable <- table(A, B)

addmargins(mytable, margin = c(1,2), FUN = sum, quiet = TRUE)

```

Tables of descriptive statistics

The command "tapply" creates tables of results. The function argument can be any statistical function that can be applied to a single numeric variable (e.g., mean, standard deviation, median, etc).

When creating tables for display purposes, such as in a report, you can round the results to a fixed number of decimal places. For example, to round a table of means "z" to two decimal places, use the **round** function

```
#Input dataset numeric weather into a dataframe

weathernum <-read.table(file ="c:/weathernumeric.txt",
stringsAsFactors=FALSE, sep =",", header=TRUE)

#Show the first 10 records and all the columns
weathernum[1:10,]

#Calculate the mean of 2 numeric attributes and place in a table

temp <- weathernum$temperature
hum <- weathernum$humidity
tempmean <- round(mean(temp),2)
hummean <- round(mean(hum),2)

z <- table(CC,DD, dnn = c("Temp","Humidity"))
z
```

Table display for one variable

For example, here is how to generate a one-way table of group means. weathernum\$play is the factor or character variable identifying the groups. Let us calculate the humidity means for groups **play** "yes" and "no"

```
tapply(weathernum$humidity, INDEX=weathernum$play, FUN=mean, na.rm=TRUE)
```

The na.rm option removes missing values before computing (otherwise the mean returns "NA" if there are missing values). na.rm is not a tapply option -- rather it is an option for the function mean. In general you can pass optional arguments to FUN by including them immediately afterward.

The following shortcut works if the arguments are listed strictly in the order shown.

```
tapply(weathernum$humidity, weathernum$play, mean, na.rm=TRUE)
```

For you to do!

- What are the humidity means based on outlook?
- The command for a one-way table of group standard deviations is similar. Write an R statement to do so.

Table display for two variables

The following example produces a two-way table of group medians. `weathernum$humidity` is a numeric variable, whereas `weathernum$play` and `weathernum$outlook` are categorical variables (factors or character variables).

```
# Shows the median humidity values for 2 categorical variables outlook and play. For example for instances where the outlook was rainy and the sport was played (yes) the median humidity value was 80.0
```

```
tapply(weathernum$humidity, INDEX
= list(weathernum$play, weathernum$outlook), median)
```

```
#different display of the same data
```

```
tapply(weathernum$humidity, INDEX = list(weathernum$outlook
, weathernum$play), median)
```

How to handle Missing Data

```
#Create subset of "weather" taking the first 6 records and variables 1,2,3 & 4. Remember we have loaded the weathernum.txt into weathernum
```

```
weather.subset <- weathernum[1:6, c(1,2,3,4)]
weather.subset
```

```
#Let us simulate some missing data
```

```
weather.subset[4,1] <- weather.subset[2,1] <- weather.subset[4,4] <- NA
weather.subset[2,2] <- weather.subset[5,3] <- weather.subset[6,2] <- NA
weather.subset
```

```
#Replace a missing value with field mean e.g.
```

```
weather.subset[5,3] <- mean(na.omit(weather.subset$humidity))
weather.subset
```

```
For you to do!
```

- Replace the missing values for Outlook with field mode
- Replace the remaining missing values for windy and temperature with a value generated at random from the observed distribution. Results will vary. Checkout `sample()` for this

```
Sol
```

```
#Replace the missing values for Outlook with field mode
```

```
our.table <- table(weather.subset$outlook)
our.mode <- names(our.table)[our.table == max(our.table)]
weather.subset[2,1] <- our.mode
weather.subset
```

```
#Replace missing values with a value generated at random from the observed distribution. Results will vary
```

```
#Random sample for Nominal Data
obs.windy <- sample(na.omit(weather.subset$windy),1)
weather.subset[4,4] <- obs.windy

# Random sample for Numeric Data
obs.temp <- sample(na.omit(weather.subset$temperature),1)
weather.subset[6,2] <- obs.temp
```

HANDS-ON ANALYSIS

Using the carsTEST dataset, answer the following questions

1. How many records are there?
2. Explore whether there are missing values for any of the variables and update based on the following requirements:
 - If there are values missing for **mpg**, replace with a constant 0
 - If there are missing values for **brand** replace with mode
 - If there are missing values for **cylinders** replace with mean to the nearest integer.
 - Any other variables with missing values first find the min and max of the variable. You are also to be replace missing values with an “individual” random value from the observed distribution.
3. Provide summary statistics and standard deviation for the following variables weightlbs, cubic inches and mpg.
4. If we assume that outliers are more than 3 standard deviations from the mean (we know there are better methods to detect outliers!), how many outliers do we have for the mpg variable?
5. What mpg values are deemed outliers? What rows contain the mpg outliers?
6. Save the changes you have made to a designated directory and call it carsUPDATED.csv

Sol

1.

```
> length(cars$weightlbs)
[1] 261
```

2.

```
#mpg
which(is.na(cars$mpg))
[1] 16 29 34 259
cars[16,1] <- 0
cars[29,1] <- 0
cars[34,1] <- 0
cars[259,1] <- 0
```

OR

```

cars[c(16,29,34,259),1] <- 0

OR
x <- which(is.na(cars$mpg))
cars$mpg[x] <-0

#cylinders
which(is.na(cars$cylinders))
[1] 8 32
cylinders.mean <-round(mean(na.omit(cars$cylinders)))
cars[8,2] <- cylinders.mean
cars[32,2] <- cylinders.mean

#cubicinches OKAY
which(is.na(cars$cubicinches))

#hp OKAY
which(is.na(cars$hp))

#weightlbs Random
which(is.na(cars$weightlbs))

[1] 16 48 261
max(na.omit(cars$weightlbs))
min(na.omit(cars$weightlbs))

obs1.weightlbs <- sample(na.omit(cars$weightlbs),1)
obs2.weightlbs <- sample(na.omit(cars$weightlbs),1)
obs3.weightlbs <- sample(na.omit(cars$weightlbs),1)

cars[16,5] <- obs1.weightlbs
cars[48,5] <- obs2.weightlbs
cars[261,5] <- obs3.weightlbs

#time-to-60 OKAY
which(is.na("cars$time-to-60"))

#year OKAY
which(is.na(cars$year))

#brand
which(is.na(cars$brand))
[1] 17 34

brand.table <- table(cars$brand)
brand.table
Europe.  Japan.    US.
      48      50     161

brand.mode <- names(brand.table)[brand.table == max(brand.table)]
brand.mode

cars[17,8] <- brand.mode

cars[34,8] <- brand.mode

3.
summary(cars$weightlbs)
summary(cars$mpg)

```

```
summary(cars$cubicinches)
sd(cars$weightlbs)
sd(cars$mpg)
sd(cars$cubicinches)
```

4.

```
#calculate standard deviations
sd1 <-sd(cars$mpg)
sd3 <-sd1*3
sd3

#calculate mean
mpg.mean <- mean(cars$mpg)
#lower and outer bounds for outliers
mpg.outlier.outer <- mpg.mean+sd3
mpg.outlier.lower <- mpg.mean-sd3

# any outliers TRUE or FALSE
mpg.outliers <- cars$mpg > mpg.outlier.outer | cars$mpg < mpg.outlier
.lower

#Show counts. We have one outlier!
mpg.outliers.table <- table(mpg.outliers)
mpg.outliers.table
```

5.

```
#what values are outliers and what row
which(cars$mpg>mpg.outlier.outer)
which(cars$mpg<mpg.outlier.lower)
```

```
value 53 in row 84
cars[84,1]
```

6.

```
#Write CSV in R
write.csv(cars, file = "K:/aSEANSPEN/EDT2014/labs/R/Rdatasets/carsUPDATED.csv", row.names=FALSE)
```

Exercises

1. **Describe the possible negative effects of proceeding directly to mine data that has not been preprocessed.**

Neglecting to preprocess the data adequately before data modeling begins will likely produce data models that are unreliable and whose results should be considered dubious at best. Performing data cleaning and data transformation during the data preparation phase is absolutely necessary for successful data mining efforts.

For example, suppose you are analyzing a data set that includes a person's Age and Date_of_Birth attributes, and you want to calculate the average Age. Now, if 5% of the records contain a value of 0 for Age, the mean value would be very misleading and inaccurate. One solution to this problem would be to derive Age for the zero-based records based on information contained in the Date_of_Birth variable. Now, the mean value for Age is more representative of those persons in the data set.

2. **What is an outlier? Why do we need to treat outliers carefully?**

Consider a set of numerical observations and the center of this observation set. An outlier is an observation that lies much farther away from the center than the majority of the other observations in the set.

We must treat outliers carefully because they can cause us to misrepresent the true center of an observation set incorrectly if they lie significantly farther away from the other observations in the set.

3. **Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis.**

It is not recommended to omit records or fields from an analysis simply because they have missing values. The rationale for this recommendation is that omitting these fields and records may cause us to lose valuable insight into the underlying relationships that we may have gleaned from the partial information that we do have.

4. **Make up a classification scheme which is inherently flawed, and would lead to misclassification,. For example, classes of items bought in a grocery store. What is wrong with this classification?**

Breakfast	Count
Cold Cereals	72
Sugar Smacks	1
Cheerios	2
Hot Cereals	28
Cream of Wheat	3

Using the table above, the “Breakfast” categorical attribute contains 5 apparent classes. However, upon further inspection the classes are discovered to be inconsistent. For example, both “Sugar Smacks” and “Cheerios” are cold cereals, and “Cream of Wheat” is a hot cereal. Below, the cereals are now classified according to one of two classes, “Cold Cereals” or “Hot Cereals.”

Breakfast	Count
Cold Cereals	75
Hot Cereals	31

Use the following stock price data (in dollars)

10	7	20	12	75	15	9	18	4	12	8	14
----	---	----	----	----	----	---	----	---	----	---	----

5. Calculate the mean, median, and mode stock price.

The **mean** is calculated as the sum of the data points divided by the number of points as follows:

$$\text{Mean Stock Price} = (10+7+20+12+75+15+9+18+4+12+8+14) / 12 = 204 / 12 = \$17.$$

The **median** is calculated by placing the prices in order and (a) selecting the middle value if the number of points is odd, or (b) taking the average of the two middle values if the number of points is even. Since we have twelve points, median is calculated as follows:

$$\text{Median Stock Price} = \text{mean of center values } \{4, 7, 8, 9, 10, \mathbf{12}, \mathbf{12}, 14, 15, 18, 20, 75\} = 24/2 = \$12.$$

The **mode** is calculated as the value that occurs the most often in the set and is calculated as follows:

$$\text{Mode Stock Price} = \text{highest frequency of } \{4, 7, 8, 9, 10, \mathbf{12}, \mathbf{12}, 14, 15, 18, 20, 75\} = \$12.$$

6. Compute the standard deviation of the stock price. Interpret what this number means.

The **standard deviation** represents the expected distance of a point chosen at random from a data set to the center of that set and is calculated by taking the square root of the **variance**. The variance is the average of the sum of squared distances of each point from the data-set mean. Given that the mean is \$17 (see Exercise #13) for this set, the variance for the set of stock prices is calculated as follows:

Stock Price Variance (Var) =

$$\begin{aligned} & (4-17)^2 + (7-17)^2 + (8-17)^2 + (9-17)^2 + (10-17)^2 + (12-17)^2 + (12-17)^2 + (14-17)^2 + (15-17)^2 + (18-17)^2 + (20-17)^2 + (75-17)^2 \\ & = \\ & (-13)^2 + (-10)^2 + (-9)^2 + (-8)^2 + (-7)^2 + (-5)^2 + (-5)^2 + (-3)^2 + (-2)^2 + (1)^2 + (3)^2 + (58)^2 = \\ & 169 + 100 + 81 + 64 + 49 + 25 + 25 + 9 + 4 + 1 + 9 + 3364 = 3900 / 12 = \mathbf{325 \$^2}. \end{aligned}$$

Taking the square root of the Variance, the Standard Deviation (SD) is calculated as follows:

Stock Price Standard Deviation (SD) of Stock Price = $\sqrt{325} = \mathbf{\pm\$18.03}$.

Since the mean is \$17 and the standard deviation is plus/minus \$18.03, the expected price of a stock drawn at random from the set of twelve stocks is expected to lie mathematically between $(\$17 - \$18.03) = \mathbf{-\$1.03}$ (i.e. \$0.01 since we assume that a stock price can never be less than one penny USD) and $(\$17 + \$18.03) = \mathbf{\$35.03}$.

As we can see, each stock with the exception of the one priced at \$75 is priced within this range.