Professor Bear - Importing Data in R

Bear

September 1, 2018

The first step in data analysis is getting the data in to R. Small datasets often come in the form of Excel (.xls), a comma delimited (Comma-Separated Value/CSV or .csv) or tab delimited (Tab-Separated Value/TSV/TXT e.g. .txt) files.

## Paths and the Working Directory

First one needs to identify your *working directory*. This is the directory or folder in which R will save or look for files by default. As a reminder, you can see your working directory by typing:

getwd()

## [1] "/Users/bear/Documents/INFO\_6105/Week\_1"

You can also change your working directory using the function setwd(). Or you can change it through RStudio by clicking on “Session”.

## Functions to read in data into R

The are several functions in base R that are available for reading data.

## read.csv

read.csv reads a file in csv format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

?read.csv

Type ?read.csv to learn how to use its arguments.

read.csv(file, header = TRUE, sep = ",", quote = "\"",  
 dec = ".", fill = TRUE, comment.char = "", ...)

Using read.csv to load some data.

# Load our data using read.csv  
  
data\_url <- 'https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/psych/galton.csv'  
galton <- read.csv(url(data\_url))  
class(galton)

## [1] "data.frame"

head(galton)

## X parent child  
## 1 1 70.5 61.7  
## 2 2 68.5 61.7  
## 3 3 65.5 61.7  
## 4 4 64.5 61.7  
## 5 5 64.0 61.7  
## 6 6 67.5 62.2

summary(galton)

## X parent child   
## Min. : 1.0 Min. :64.00 Min. :61.70   
## 1st Qu.:232.8 1st Qu.:67.50 1st Qu.:66.20   
## Median :464.5 Median :68.50 Median :68.20   
## Mean :464.5 Mean :68.31 Mean :68.09   
## 3rd Qu.:696.2 3rd Qu.:69.50 3rd Qu.:70.20   
## Max. :928.0 Max. :73.00 Max. :73.70

## read.table

read.table reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

?read.table

Type ?read.table to learn how to use its arguments.

read.table(file, header = FALSE, sep = "", quote = "\"'",  
 dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),  
 row.names, col.names, as.is = !stringsAsFactors,  
 na.strings = "NA", colClasses = NA, nrows = -1,  
 skip = 0, check.names = TRUE, fill = !blank.lines.skip,  
 strip.white = FALSE, blank.lines.skip = TRUE,  
 comment.char = "#",  
 allowEscapes = FALSE, flush = FALSE,  
 stringsAsFactors = default.stringsAsFactors(),  
 fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

Using read.table to load some data.

# Load our data using read.table  
# Balloons Data Set  
data\_url <- 'https://archive.ics.uci.edu/ml/machine-learning-databases/balloons/adult+stretch.data'  
balloons <- read.table(url(data\_url))  
class(balloons)

## [1] "data.frame"

head(balloons)

## V1  
## 1 YELLOW,SMALL,STRETCH,ADULT,T  
## 2 YELLOW,SMALL,STRETCH,ADULT,T  
## 3 YELLOW,SMALL,STRETCH,CHILD,F  
## 4 YELLOW,SMALL,DIP,ADULT,F  
## 5 YELLOW,SMALL,DIP,CHILD,F  
## 6 YELLOW,LARGE,STRETCH,ADULT,T

summary(balloons)

## V1   
## PURPLE,LARGE,STRETCH,ADULT,T: 2   
## PURPLE,SMALL,STRETCH,ADULT,T: 2   
## YELLOW,LARGE,STRETCH,ADULT,T: 2   
## YELLOW,SMALL,STRETCH,ADULT,T: 2   
## PURPLE,LARGE,DIP,ADULT,F : 1   
## PURPLE,LARGE,DIP,CHILD,F : 1   
## (Other) :10

Whoops, what happened? Look at the [Balloons Data Set](https://archive.ics.uci.edu/ml/machine-learning-databases/balloons/adult+stretch.data)

balloons <- read.table(url(data\_url), sep = ",")  
class(balloons)

## [1] "data.frame"

head(balloons)

## V1 V2 V3 V4 V5  
## 1 YELLOW SMALL STRETCH ADULT TRUE  
## 2 YELLOW SMALL STRETCH ADULT TRUE  
## 3 YELLOW SMALL STRETCH CHILD FALSE  
## 4 YELLOW SMALL DIP ADULT FALSE  
## 5 YELLOW SMALL DIP CHILD FALSE  
## 6 YELLOW LARGE STRETCH ADULT TRUE

summary(balloons)

## V1 V2 V3 V4 V5   
## PURPLE:10 LARGE:10 DIP : 8 ADULT:12 Mode :logical   
## YELLOW:10 SMALL:10 STRETCH:12 CHILD: 8 FALSE:12   
## TRUE :8

## read.delim

read.delim reads a file in tab delimited table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

# set your working directory - normally where you data are  
setwd('path/to/your/data')  
data = read.delim('data.file',  
 header = TRUE,   
 sep = '\t')

Type ?read.delim to learn what the header and sep arguments do.

?read.delim

read.delim(file, header = TRUE, sep = "\t", quote = "\"",  
 dec = ".", fill = TRUE, comment.char = "", ...)

## Quiz - load some data with read.delim

Find some data on the [UC Irvine Machine Learning Repository](http://archive.ics.uci.edu/ml/) and load it with read.delim