Professor Bear - Importing Data in R

Bear

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/Downloads/DAT-BOS-16/NBB"
```

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 <- 'http://www.math.uah.edu/stat/data/Galton.csv'
galton <- read.csv(url(data_url))
class(galton)

## [1] "data.frame"

head(galton)
```

```
##   Family Father Mother Gender Height Kids
## 1      1   78.5   67.0      M   73.2    4
## 2      1   78.5   67.0      F   69.2    4
## 3      1   78.5   67.0      F   69.0    4
## 4      1   78.5   67.0      F   69.0    4
## 5      2   75.5   66.5      M   73.5    4
## 6      2   75.5   66.5      M   72.5    4
```

```
summary(galton)
```

```
##      Family          Father          Mother      Gender      Height     
##  185     : 15   Min.   :62.00   Min.   :58.00   F:433   Min.   :56.00  
##  166     : 11   1st Qu.:68.00   1st Qu.:63.00   M:465   1st Qu.:64.00  
##  66      : 11   Median :69.00   Median :64.00           Median :66.50  
##  130     : 10   Mean   :69.23   Mean   :64.08           Mean   :66.76  
##  136     : 10   3rd Qu.:71.00   3rd Qu.:65.50           3rd Qu.:69.70  
##  140     : 10   Max.   :78.50   Max.   :70.50           Max.   :79.00  
##  (Other):831                                                           
##       Kids       
##  Min.   : 1.000  
##  1st Qu.: 4.000  
##  Median : 6.000  
##  Mean   : 6.136  
##  3rd Qu.: 8.000  
##  Max.   :15.000  
## 
```

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

```
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
##                                                   NA's :0
```

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 and load it with read.delim