Chapter 2: Data Frames with R

Jixiang Wu

Associate Professor of Quantitative Genetics/Biostatistics

Agronomy, Horticulture, and Plant Science Department

South Dakota State University

Email: Jixiang.wu@sdstate.edu

Phone: (605)688-5947

Introduction

Statistical data analysis requires handling various data sets. It is important to get some general ideas to deal with dataframe in R. The following R scripts will show you how to play with data sets.

Read a data set

There are two important ways to read/load a data set. You can read a data set from your hard drive or you can load a data set from an R package.

Read a data set from hard drive

For example the following R script shows that we can read a txt file from a hard drive (the data set is available in the pc).

```
worms=read.table("c:\\ps792\\worms.txt",header=TRUE)

worms
```

Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
Oak.Mead	3.1	2	Grassland	3.9	FALSE	2

Church.Field	3.5	3	Grassland	4.2	FALSE	3
Ashurst	2.1	0	Arable	4.8	FALSE	4
The.Orchard	1.9	0	Orchard	5.7	FALSE	9
Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
North.Gravel	3.3	1	Grassland	4.1	FALSE	1
South.Gravel	3.7	2	Grassland	4.0	FALSE	2
Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
Pond.Field	4.1	0	Meadow	5.0	TRUE	6
Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
Cheapside	2.2	8	Scrub	4.7	TRUE	4
Pound.Hill	4.4	2	Arable	4.5	FALSE	5
Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

Note: If the data set is in csv file format, you need use read.csv function to load the data.

Load a data from an R package

The above loaded file is now packed in coursedata package. So it is easy to get this data set as follows.

```
require(coursedata)

## Loading required package: coursedata

data(worms)
worms
```

Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
Church.Field	3.5	3	Grassland	4.2	FALSE	3
Ashurst	2.1	0	Arable	4.8	FALSE	4
The.Orchard	1.9	0	Orchard	5.7	FALSE	9
Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
Garden.Wood	2.9	10	Scrub	5.2	FALSE	8

North.Gravel	3.3	1	Grassland	4.1	FALSE	1
South.Gravel	3.7	2	Grassland	4.0	FALSE	2
Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
Pond.Field	4.1	0	Meadow	5.0	TRUE	6
Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
Cheapside	2.2	8	Scrub	4.7	TRUE	4
Pound.Hill	4.4	2	Arable	4.5	FALSE	5
Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

The two data sets loaded from my hard drive and the R package are exactly the same.

Play with a data set

Once the file has been imported to R we can do what you want to do.

Use attach to make the variables accessible by name within the R session

Using the attach function will make all variables in the worms file become globle variables. You can use these variables anytime once you globlize these variables. However, it could make the data analysis a little messy sometimes becuase you could use the same variable names repeatedly. This can make very difficult to debug/check your R codes if you have a lengthy R file.

```
attach(worms)
```

After you run the above R code, you can use the following R codes to access all variables in the worms file. For example,

```
Field.Name
```

```
##  [1] Nashs.Field       Silwood.Bottom    Nursery.Field
##  [4] Rush.Meadow       Gunness.Thicket   Oak.Mead
##  [7] Church.Field      Ashurst           The.Orchard
## [10] Rookery.Slope     Garden.Wood       North.Gravel
## [13] South.Gravel      Observatory.Ridge Pond.Field
## [16] Water.Meadow      Cheapside         Pound.Hill
## [19] Gravel.Pit        Farm.Wood
## 20 Levels: Ashurst Cheapside Church.Field Farm.Wood ... Water.Meadow
```

```
Area
```

```
##  [1] 3.6 5.1 2.8 2.4 3.8 3.1 3.5 2.1 1.9 1.5 2.9 3.3 3.7 1.8 4.1 3.9 2.2
## [18] 4.4 2.9 0.8
```

Obtain a list of the variable names

You obtain the header names for a datafile easily.

```
names(worms)
```

```
## [1] "Field.Name"   "Area"         "Slope"        "Vegetation"
## [5] "Soil.pH"      "Damp"         "Worm.density"
```

or

```
colnames(worms)
```

```
## [1] "Field.Name"   "Area"         "Slope"        "Vegetation"
## [5] "Soil.pH"      "Damp"         "Worm.density"
```

Look at particular rows of a data set

You can use the function head to look at the first six rows of a data set.

```
head(worms)
```

Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
Oak.Mead	3.1	2	Grassland	3.9	FALSE	2

Or you can look any row(s) of the data set. For examples:

```
id=c(1:6)
worms[id,]
```

Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
Oak.Mead	3.1	2	Grassland	3.9	FALSE	2

You can use the function tail to check the bottom six rows.

```
tail(worms)
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

Data summary

```
summary(worms) ##

##         Field.Name      Area           Slope           Vegetation
##  Ashurst     : 1   Min.   :0.80   Min.   : 0.00   Arable   :3
##  Cheapside   : 1   1st Qu.:2.17   1st Qu.: 0.75   Grassland:9
##  Church.Field: 1   Median :3.00   Median : 2.00   Meadow   :3
##  Farm.Wood   : 1   Mean   :2.99   Mean   : 3.50   Orchard  :1
##  Garden.Wood : 1   3rd Qu.:3.73   3rd Qu.: 5.25   Scrub    :4
##  Gravel.Pit  : 1   Max.   :5.10   Max.   :11.00
##  (Other)     :14
##     Soil.pH        Damp          Worm.density
##  Min.   :3.50   Mode :logical   Min.   :0.00
##  1st Qu.:4.10   FALSE:14        1st Qu.:2.00
##  Median :4.60   TRUE :6         Median :4.00
##  Mean   :4.55   NA's :0         Mean   :4.35
##  3rd Qu.:5.00                   3rd Qu.:6.25
##  Max.   :5.70                   Max.   :9.00
##
```

Values of continuous variables are summarized under six headings: one parametric (the arithmetic) mean and five non-parametric (maximum, minimum, median, 25th percentile or first quartile, and 75% percentile or third quartile). TUkey's famouse five -number faction (fivenum) is slightly different, with hinges rather than first and third quartiles. Levels of categorical variables are counted. Note that the field names are not listed in full because they are unique to each row, six of them are named, then R says "plus 14 others".

You may also use the function str to summarize each field of a dataframe.

```
str(worms) ##

## 'data.frame':    20 obs. of  7 variables:
##  $ Field.Name  : Factor w/ 20 levels "Ashurst","Cheapside",..: 8 17 10 16
7 11 3 1 19 15 ...
##  $ Area        : num  3.6 5.1 2.8 2.4 3.8 3.1 3.5 2.1 1.9 1.5 ...
##  $ Slope       : int  11 2 3 5 0 2 3 0 0 4 ...
##  $ Vegetation  : Factor w/ 5 levels "Arable","Grassland",..: 2 1 2 3 5 2 2
1 4 2 ...
```

```
##  $ Soil.pH     : num  4.1 5.2 4.3 4.9 4.2 3.9 4.2 4.8 5.7 5 ...
##  $ Damp        : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Worm.density: int  4 7 2 5 6 2 3 4 9 7 ...
```

Subscripts and indices

Sometimes you are only interested in checking particular cells, rows, columns. The following R codes show that you can look at particular values.

```
worms[3,5]
```

```
## [1] 4.3
```

```
worms[14:19,7]
```

```
## [1] 0 6 8 4 5 1
```

```
worms[1:5,2:3]
```

Area	Slope
3.6	11
5.1	2
2.8	3
2.4	5
3.8	0

```
worms[3,]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
3	Nursery.Field	2.8	3	Grassland	4.3	FALSE	2

```
worms[,3]
```

```
##  [1] 11  2  3  5  0  2  3  0  0  4 10  1  2  6  0  0  8  2  1 10
```

```
class(worms[3,1])
```

```
## [1] "factor"
```

```
class(worms[,3])
```

```
## [1] "integer"
```

```
worms[,c(1,5)]
```

Field.Name	Soil.pH
Nashs.Field	4.1
Silwood.Bottom	5.2

Nursery.Field	4.3
Rush.Meadow	4.9
Gunness.Thicket	4.2
Oak.Mead	3.9
Church.Field	4.2
Ashurst	4.8
The.Orchard	5.7
Rookery.Slope	5.0
Garden.Wood	5.2
North.Gravel	4.1
South.Gravel	4.0
Observatory.Ridge	3.8
Pond.Field	5.0
Water.Meadow	4.9
Cheapside	4.7
Pound.Hill	4.5
Gravel.Pit	3.5
Farm.Wood	5.1

Selecting rows from the dataframe at random

```
id=sample(1:20,8)
id
```

```
## [1] 15  6 12  3 17  4  8 10
```

```
worms[id,]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
6	Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
12	North.Gravel	3.3	1	Grassland	4.1	FALSE	1
3	Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7

Sorting dataframes

You can sort a data by a variable, for example, by Slope

```
worms[order(Slope),]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
5	Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
9	The.Orchard	1.9	0	Orchard	5.7	FALSE	9
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
12	North.Gravel	3.3	1	Grassland	4.1	FALSE	1
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
6	Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
13	South.Gravel	3.7	2	Grassland	4.0	FALSE	2
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
3	Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
7	Church.Field	3.5	3	Grassland	4.2	FALSE	3
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
14	Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
11	Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3
1	Nashs.Field	3.6	11	Grassland	4.1	FALSE	4

```
worms[rev(order(Slope)),]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
1	Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3
11	Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
14	Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7

7	Church.Field	3.5	3	Grassland	4.2	FALSE	3
3	Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
13	South.Gravel	3.7	2	Grassland	4.0	FALSE	2
6	Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
12	North.Gravel	3.3	1	Grassland	4.1	FALSE	1
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
9	The.Orchard	1.9	0	Orchard	5.7	FALSE	9
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
5	Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6

Or you can sort a data set by two variables, for example:

```
worms[order(Vegetation, Worm.density),]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
14	Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
12	North.Gravel	3.3	1	Grassland	4.1	FALSE	1
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
3	Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
6	Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
13	South.Gravel	3.7	2	Grassland	4.0	FALSE	2
7	Church.Field	3.5	3	Grassland	4.2	FALSE	3
1	Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
9	The.Orchard	1.9	0	Orchard	5.7	FALSE	9
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4

5	Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
11	Garden.Wood	2.9	10	Scrub	5.2	FALSE	8

Or you can sort a data set by three variables, for example:

```
worms[order(Vegetation, Worm.density, Soil.pH),]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
14	Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
12	North.Gravel	3.3	1	Grassland	4.1	FALSE	1
6	Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
13	South.Gravel	3.7	2	Grassland	4.0	FALSE	2
3	Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
7	Church.Field	3.5	3	Grassland	4.2	FALSE	3
1	Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
9	The.Orchard	1.9	0	Orchard	5.7	FALSE	9
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
5	Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
11	Garden.Wood	2.9	10	Scrub	5.2	FALSE	8

Sort a data set and obtain particular columns, for example:

```
worms[order(Vegetation, Worm.density),c(4,7,5,3)]
```

	Vegetation	Worm.density	Soil.pH	Slope
8	Arable	4	4.8	0
18	Arable	5	4.5	2
2	Arable	7	5.2	2
14	Grassland	0	3.8	6

12	Grassland	1	4.1	1
19	Grassland	1	3.5	1
3	Grassland	2	4.3	3
6	Grassland	2	3.9	2
13	Grassland	2	4.0	2
7	Grassland	3	4.2	3
1	Grassland	4	4.1	11
10	Grassland	7	5.0	4
4	Meadow	5	4.9	5
15	Meadow	6	5.0	0
16	Meadow	8	4.9	0
9	Orchard	9	5.7	0
20	Scrub	3	5.1	10
17	Scrub	4	4.7	8
5	Scrub	6	4.2	0
11	Scrub	8	5.2	10

```
worms[order(Vegetation, Worm.density),c("Vegetation","Worm.density","Soil.pH"
,"Slope")]
```

	Vegetation	Worm.density	Soil.pH	Slope
8	Arable	4	4.8	0
18	Arable	5	4.5	2
2	Arable	7	5.2	2
14	Grassland	0	3.8	6
12	Grassland	1	4.1	1
19	Grassland	1	3.5	1
3	Grassland	2	4.3	3
6	Grassland	2	3.9	2
13	Grassland	2	4.0	2
7	Grassland	3	4.2	3
1	Grassland	4	4.1	11
10	Grassland	7	5.0	4
4	Meadow	5	4.9	5
15	Meadow	6	5.0	0
16	Meadow	8	4.9	0
9	Orchard	9	5.7	0

20	Scrub	3	5.1	10
17	Scrub	4	4.7	8
5	Scrub	6	4.2	0
11	Scrub	8	5.2	10

Using logical conditions to select rows from the dataframe

```
worms[Damp==TRUE,]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

```
worms[Worm.density > median(Worm.density) & Soil.pH<5.2,]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
5	Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5

Obtain the columns which are numeric.

```
worms[,sapply(worms,is.numeric)]
```

Area	Slope	Soil.pH	Worm.density
3.6	11	4.1	4
5.1	2	5.2	7
2.8	3	4.3	2
2.4	5	4.9	5
3.8	0	4.2	6
3.1	2	3.9	2

3.5	3	4.2	3
2.1	0	4.8	4
1.9	0	5.7	9
1.5	4	5.0	7
2.9	10	5.2	8
3.3	1	4.1	1
3.7	2	4.0	2
1.8	6	3.8	0
4.1	0	5.0	6
3.9	0	4.9	8
2.2	8	4.7	4
4.4	2	4.5	5
2.9	1	3.5	1
0.8	10	5.1	3

Obtain the columns which are factor.

```
worms[,sapply(worms,is.factor)]
```

Field.Name	Vegetation
Nashs.Field	Grassland
Silwood.Bottom	Arable
Nursery.Field	Grassland
Rush.Meadow	Meadow
Gunness.Thicket	Scrub
Oak.Mead	Grassland
Church.Field	Grassland
Ashurst	Arable
The.Orchard	Orchard
Rookery.Slope	Grassland
Garden.Wood	Scrub
North.Gravel	Grassland
South.Gravel	Grassland
Observatory.Ridge	Grassland
Pond.Field	Meadow
Water.Meadow	Meadow
Cheapside	Scrub
Pound.Hill	Arable

Gravel.Pit	Grassland
Farm.Wood	Scrub

Obtain data without some columns.

```
worms[-(6:15),]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
1	Nashs.Field	3.6	11	Grassland	4.1	FALSE	4
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
3	Nursery.Field	2.8	3	Grassland	4.3	FALSE	2
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
5	Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

```
worms[!(Vegetation=="Grassland"),]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
5	Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
9	The.Orchard	1.9	0	Orchard	5.7	FALSE	9
11	Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

```
worms[-which(Damp==FALSE),]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5

10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

or

```
worms[!Damp==F,]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

or even simpler

```
worms[Damp==TRUE,]
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

Omitting rows containing missing values, NA

Sometimes a data set can contain some missing values. It is important sometimes you need cleanse your data before running some data analyses.

Giving the above data, we can set several missing data for Slope.

```
dat=worms
dat$Slope[c(1,3,5)]=NA
```

You will some missing data in the Slope column.

```
dat
```

Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
Nashs.Field	3.6	NA	Grassland	4.1	FALSE	4
Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
Nursery.Field	2.8	NA	Grassland	4.3	FALSE	2
Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
Gunness.Thicket	3.8	NA	Scrub	4.2	FALSE	6
Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
Church.Field	3.5	3	Grassland	4.2	FALSE	3
Ashurst	2.1	0	Arable	4.8	FALSE	4
The.Orchard	1.9	0	Orchard	5.7	FALSE	9
Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
North.Gravel	3.3	1	Grassland	4.1	FALSE	1
South.Gravel	3.7	2	Grassland	4.0	FALSE	2
Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
Pond.Field	4.1	0	Meadow	5.0	TRUE	6
Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
Cheapside	2.2	8	Scrub	4.7	TRUE	4
Pound.Hill	4.4	2	Arable	4.5	FALSE	5
Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

```
na.omit(dat)
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
6	Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
7	Church.Field	3.5	3	Grassland	4.2	FALSE	3
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
9	The.Orchard	1.9	0	Orchard	5.7	FALSE	9
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
11	Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
12	North.Gravel	3.3	1	Grassland	4.1	FALSE	1
13	South.Gravel	3.7	2	Grassland	4.0	FALSE	2

14	Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

or

```
new.frame=na.exclude(dat)

new.frame
```

	Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
2	Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
4	Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
6	Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
7	Church.Field	3.5	3	Grassland	4.2	FALSE	3
8	Ashurst	2.1	0	Arable	4.8	FALSE	4
9	The.Orchard	1.9	0	Orchard	5.7	FALSE	9
10	Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
11	Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
12	North.Gravel	3.3	1	Grassland	4.1	FALSE	1
13	South.Gravel	3.7	2	Grassland	4.0	FALSE	2
14	Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
15	Pond.Field	4.1	0	Meadow	5.0	TRUE	6
16	Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
17	Cheapside	2.2	8	Scrub	4.7	TRUE	4
18	Pound.Hill	4.4	2	Arable	4.5	FALSE	5
19	Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
20	Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

or

```
ok=complete.cases(dat)

newdat=dat[ok,]
```

Replacing NAs with zeros

```
dat[is.na(dat)]=0
dat
```

Field.Name	Area	Slope	Vegetation	Soil.pH	Damp	Worm.density
Nashs.Field	3.6	0	Grassland	4.1	FALSE	4
Silwood.Bottom	5.1	2	Arable	5.2	FALSE	7
Nursery.Field	2.8	0	Grassland	4.3	FALSE	2
Rush.Meadow	2.4	5	Meadow	4.9	TRUE	5
Gunness.Thicket	3.8	0	Scrub	4.2	FALSE	6
Oak.Mead	3.1	2	Grassland	3.9	FALSE	2
Church.Field	3.5	3	Grassland	4.2	FALSE	3
Ashurst	2.1	0	Arable	4.8	FALSE	4
The.Orchard	1.9	0	Orchard	5.7	FALSE	9
Rookery.Slope	1.5	4	Grassland	5.0	TRUE	7
Garden.Wood	2.9	10	Scrub	5.2	FALSE	8
North.Gravel	3.3	1	Grassland	4.1	FALSE	1
South.Gravel	3.7	2	Grassland	4.0	FALSE	2
Observatory.Ridge	1.8	6	Grassland	3.8	FALSE	0
Pond.Field	4.1	0	Meadow	5.0	TRUE	6
Water.Meadow	3.9	0	Meadow	4.9	TRUE	8
Cheapside	2.2	8	Scrub	4.7	TRUE	4
Pound.Hill	4.4	2	Arable	4.5	FALSE	5
Gravel.Pit	2.9	1	Grassland	3.5	FALSE	1
Farm.Wood	0.8	10	Scrub	5.1	TRUE	3

Conclusions

The above R scripts demonstrate the flexibility of using R in data managements. There are many more ways in which we can play with a data set before we run a real statistical data analysis.