

# Additional Material for Merging and Importing Data

*Sudhakar Kumar*

*24 March 2019*

## Built-in functions for exploring a data frame

We will use built-in dataset `iris` to explore some of the useful functions in `base` package of R language. In order to know the dimensions of `iris`, we use `dim` function. The output of `dim` function is a vector, in which the elements represent the number of rows and number of columns, respectively.

```
dim(iris)
```

```
## [1] 150 5
```

We can also use `nrow` and `ncol` to get the number of rows and number of columns, respectively.

```
nrow(iris)
```

```
## [1] 150
```

```
ncol(iris)
```

```
## [1] 5
```

Thus, `iris` has 150 rows and 5 columns, which can also be verified by using `str` function. It also returns many useful pieces of information, including the above information and the types of data for each column.

```
str(iris)
```

```
## 'data.frame': 150 obs. of 5 variables:
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
## $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

The first row in the output indicates that this dataset is a data frame with 150 observations of 5 variables. Also, `num` denotes that the variables `Sepal.Length`, `Sepal.Width`, `Petal.Length` and `Petal.Width` are numeric. `Factor` denotes that the variable `Species` is categorical with 3 levels (`setosa`, `versicolor`, `virginica`).

To know the range of values inside `iris`, we use `summary` function. In particular, this function provides a number of useful statistics including range, median and mean (Andrew Shaughnessy 2018).

```
summary(iris)
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300
## Median :5.800 Median :3.000 Median :4.350 Median :1.300
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
## Species
## setosa :50
## versicolor:50
## virginica :50
```

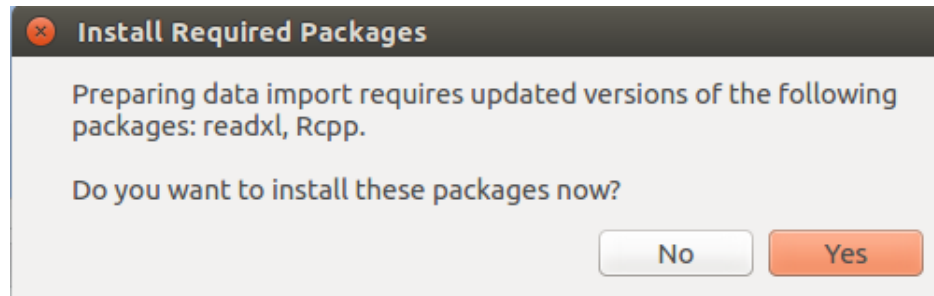


Figure 1: Installing readxl and Rcpp

```
##
##
##
```

We use `head` to obtain the first `n` observations and `tail` to obtain the last `n` observations; by default, `n = 6`. These are good commands for obtaining an intuitive idea of what the data look like without revealing the entire dataset, which could have millions of rows and thousands of columns (Cai 2013).

```
head(iris, 2)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1           3.5           1.4           0.2  setosa
## 2           4.9           3.0           1.4           0.2  setosa
```

```
tail(iris, 2)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width  Species
## 149           6.2           3.4           5.4           2.3 virginica
## 150           5.9           3.0           5.1           1.8 virginica
```

## Dependencies for reading datasets in R

In order to read XML files in R, we need to install XML package. However, the Ubuntu package `libxml2-dev` needs to be installed beforehand (Overflow 2013). On Linux operating system, open the terminal and type the following commands.

```
sudo apt-get update
```

```
sudo apt-get install libxml2-dev
```

Similarly, while importing Excel data in R, we need to install `readxl` and `Rcpp`. If these packages are not installed and you try importing Excel data, a pop-up message as shown in Figure 1 will be generated. By clicking **Yes** to this message, these packages can be installed.

## References

Andrew Shaughnessy, Elizabeth Hasenmueller, Christopher Prener. 2018. “Exploring Data in R.” <https://cran.r-project.org/web/packages/driftR/vignettes/ExploringData.html>.

Cai, Eric. 2013. “Exploratory Data Analysis: Useful R Functions for Exploring a Data Frame.” <https://chemicalstatistician.wordpress.com/2013/08/19/exploratory-data-analysis-useful-r-functions-for-exploring-a-data-frame/>.

Overflow, Stack. 2013. “Unable to install R package in Ubuntu 11.04.” <https://stackoverflow.com/questions/>

7765429/unable-to-install-r-package-in-ubuntu-11-04.