Manipulating Data in R

John Muschelli

January 4, 2016

Overview

You can read data into R using *read.csv*. In this module, we will show you how to:

1. Select specific elements of an object by an index or logical condition
2. Subset rows of a `data.frame`
3. Subset columns of a `data.frame`
4. Add new columns to a `data.frame`
5. Order the rows of a `data.frame`

Setup

We will show you how to do each operation in base R then show you how to use the `dplyr` package to do the same operation (if applicable).

Many resources on how to use `dplyr` exist and are straightforward * `https://cran.rstudio.com/web/packages/dplyr/vignettes/` * `https://stat545-ubc.github.io/block009_dplyr-intro.html` * `https://www.datacamp.com/courses/dplyr-data-manipulation-r-tutorial`

Select specific elements using an index

Often you only want to look at subsets of a data set at any given time. As a review, elements of an R object are selected using the brackets ([and])

Here `x` is a vector of numbers and we can select specific elements of `x` using indexing

```
x = c(1, 2, 4, 8, 10)
x[2]
```

```
## [1] 2
```

```
x[5]
```

```
## [1] 10
```

```
x[c(2,5)]
```

```
## [1]  2 10
```

Subsetting Data

You can put a - before integers inside brackets to remove these indices from the data.

```
x[-2] # all but the second
```

```
## [1]  1  4  8 10
```

Note that you have to be careful with this syntax when dropping more than 1 element:

```
x[-c(1,2,3)] # drop first 3
```

```
## [1]  8 10
```

```
# x[-1:3] # shorthand. R sees as -1 to 3
x[-(1:3)] # needs parentheses
```

```
## [1]  8 10
```

Selecting on multiple queries

What about selecting rows based on the values of two variables? We can 'chain' together logical statements using the following:

- `&` : AND
- `|` : OR