

The Basics of Handling Data

- Reading and writing data from files
- Data structures
- Types of variables: numeric, character, factor
- Missing values
- Subsetting

Reading data

- `read.csv()`
- `read.table()`
- `scan()`
- List data available in R, and packages: `data()`,
`data(package="datasets")`

Data structures

- Data read in using `read.csv()` will be cast as a `data.frame()`
- A `data.frame()` is a list
- A list is an ordered collection of objects, which could consist of a numeric vector, a matrix, a character array, a function, ...
- `sla[[1]]` is the same as `sla$me`, which is the same as `sla[,1]`
- `str(sla)` describes the `data.frame`, the variable types, and the first few values of each

Your turn

- What are the three different ways that you could access the 10th row of the 3rd variable?
- What type of variable is “month”? “weight”?

Data structures

- `matrix()` contains purely numeric data
- `as.data.frame()` will convert to a data frame; reverse conversion `as.matrix()` is more complicated when there is a mix of variable types
- Many numerical computations need to be done on matrices
 - `x%%A`, `x%%t(A)`
 - `var(x)`, `prcomp(var(x))`, `apply(x,2,mean)`

Types of variables

- numeric (integer, double, single), complex, logical, character, ...
- Getting information about type: `attributes()`, `is.numeric()`, ...
- Re-casting as a different type: `as.character()`, `as.numeric()`, `as.character()`, ...

Types of variables

- factors are a way to store and organize categorical data
- `levels()`
- To re-order, or re-define levels: `factor()`

Missing values

- R assigns the label “NA” to missings
- If data has a missing value, R will return a missing from the computation, unless you explicitly instruct it to ignore missings
- Imputing missings

Your turn

- Change the levels of `slap$me` to be "splitbamboo", "rhonda" and "other"
- Impute the missing values of `slap$target.weight` using the median value

Subsetting

- Logical subsets
- Ordering