Cleaning Data Lecture Notes – The data.table package

Data.table

1. Inherets from data.frame
   1. All functions that accept data.frame work on data.table
2. Written in C
   1. Ergo, it’s really fast
3. Much faster at subsetting, grouping, and updating and updating variables

Creating data tables is just like creating data frames

Code:

Library(data.table)

Previously:

DF <- data.frame(x=rnorm(9), y=rep(c(“a”, “b”, “c”), each=3), z=rnorm(9))

Now:

DT <- data.table(x=rnorm(9), y=rep(c(“a”, “b”, “c”), each=3), z=rnorm(9))

See all data tables in memory

tables() gives back all

Subsetting

DT[x,] gives you the rows, same

DT[DT$y==”a”], same

Subsetting rows – different

DT[c(2,3)] gives back the whole 2nd and 3rd rows

Subsetting Columns = really different

DT[,c(2,3)] used to give you back the 2nd and 3rd columns, now it will just give back the vector = <2,3>

How to subset by Columns in data.table

1. The subsetting function is modified for data.table
2. Arguments pass after commas are called an expression

Calculating Values for variables with expressions

1. DT[,list(fxn1(x), fxn2(z))]
   1. Gives back a list with the result of function 1 and function 2 as vector functions as the element.
      1. Example: sum(x) will sum up the entries in x
   2. Quotes are not necessary
2. Passing a list on the second argument always runs a function or expression against the lsit

Adding new Columns to the talbe

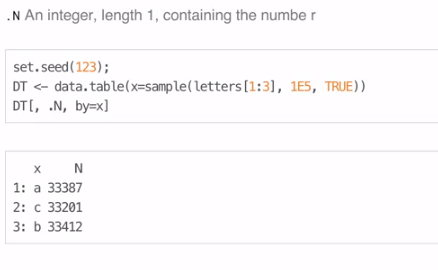
1. DT[,w:=z^2] gives back DT with a 4th column w that contains the square of each entry in z
2. := adds a column with a variable name
3. Does not create a new copy of the data table, just rewrites it. Good for freeing up RAM, bad for losing your first function.
   1. Use copy(DT) if you want copies

Multiple Operations

1. DT[, m:={tmp <- x+z}, log2(tmp+5)}] will create the column m and have the colum M be log2((x+z)+5)
   1. Log2 is the binary log
2. DT[,a:=x>0] will add the column a to the data table showing T/F depending on if x is positive or negative
3. DT[,b:=mean(x+w), by=a] will calculate the mean when a is true, and not when a is false

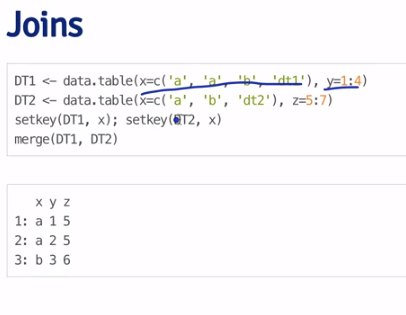
New Variables

1. .N is an integer of length 1 containin the number of times a group appears
   1. Ex DT <- data.table(x=sample(letters[1:3], 100000, TRUE))
      1. Makes a table with a length of 100000 that is randomly filled with a’s, b’s, or c’s
   2. DT[, .N, by=x] will give us a data table with the number of times a, b, and c show up in the data table



Data Tables contain Keys

1. DT <- data.table(x=rep(c(“a”,”b”,”c”), each=100), ynorm(300))
2. setkey(DT, x)
3. DT[“a”]
4. setkey() tells it the most pertinent column in the table
5. facilitates joins as well (requires the same key in both tables



list of differences between data.table and data.frame from <http://stackoverflow.com/questions/13618488/what-you-can-do-with-data-frame-that-you-cant-in-data-table>

Package instructions

<https://cran.r-project.org/web/packages/data.table/data.table.pdf>

Home page

<https://github.com/Rdatatable/data.table/wiki>