Rtricks

Vikesh Singh Baghel

23 February 2019

Today, I am going to show some examples on how to use data.table. R’s data.table package is very fast and efficient. I will be using a simple data frame mtcars in this demo.

Let’s load the data and see how what it has.

data(mtcars)  
head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

To get the size of size, R has dim command. Let’s also get summary statistics and class

print(dim(mtcars))

## [1] 32 11

print(summary(mtcars))

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

print(class(mtcars))

## [1] "data.frame"

So, we have a data frame of 32 rows and 11 columns in this data set. All the columns are numeric.

I will store mtcars into a different object and start from there. Now I want to add another column to the data. I want to add avg\_mpg\_cg which will be mean of mpg over cyl and gear. There are several ways to do it. But my preferred approach is using data.table package.

library(data.table)  
dat <- as.data.table(mtcars)  
dat[,avg\_mpg\_cg:= mean(mpg), by = c("cyl","gear")]  
head(dat)

## mpg cyl disp hp drat wt qsec vs am gear carb avg\_mpg\_cg  
## 1: 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4 19.750  
## 2: 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4 19.750  
## 3: 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1 26.925  
## 4: 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1 19.750  
## 5: 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2 15.050  
## 6: 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1 19.750

So, It can be in just one line! Okay, now may I dont need mpg column and want remove it.

dat[,mpg:= NULL]  
head(dat)

## cyl disp hp drat wt qsec vs am gear carb avg\_mpg\_cg  
## 1: 6 160 110 3.90 2.620 16.46 0 1 4 4 19.750  
## 2: 6 160 110 3.90 2.875 17.02 0 1 4 4 19.750  
## 3: 4 108 93 3.85 2.320 18.61 1 1 4 1 26.925  
## 4: 6 258 110 3.08 3.215 19.44 1 0 3 1 19.750  
## 5: 8 360 175 3.15 3.440 17.02 0 0 3 2 15.050  
## 6: 6 225 105 2.76 3.460 20.22 1 0 3 1 19.750

Boom! Done. Okay, that’s cool. But wait mtcars had car names as rownames. It is not there in dat because data.table does not support rownames. But we can add a column to it.

dat[,car\_names:= rownames(mtcars)]  
head(dat)

## cyl disp hp drat wt qsec vs am gear carb avg\_mpg\_cg  
## 1: 6 160 110 3.90 2.620 16.46 0 1 4 4 19.750  
## 2: 6 160 110 3.90 2.875 17.02 0 1 4 4 19.750  
## 3: 4 108 93 3.85 2.320 18.61 1 1 4 1 26.925  
## 4: 6 258 110 3.08 3.215 19.44 1 0 3 1 19.750  
## 5: 8 360 175 3.15 3.440 17.02 0 0 3 2 15.050  
## 6: 6 225 105 2.76 3.460 20.22 1 0 3 1 19.750  
## car\_names  
## 1: Mazda RX4  
## 2: Mazda RX4 Wag  
## 3: Datsun 710  
## 4: Hornet 4 Drive  
## 5: Hornet Sportabout  
## 6: Valiant

Okay. So, now I want to select only the rows where car\_names has Hornet

dat[grep("Hornet",car\_names)]

## cyl disp hp drat wt qsec vs am gear carb avg\_mpg\_cg  
## 1: 6 258 110 3.08 3.215 19.44 1 0 3 1 19.75  
## 2: 8 360 175 3.15 3.440 17.02 0 0 3 2 15.05  
## car\_names  
## 1: Hornet 4 Drive  
## 2: Hornet Sportabout

Hmmm. I want Mazda too.

dat[grep(paste(c("Mazda","Hornet"),collapse="|"), car\_names)]

## cyl disp hp drat wt qsec vs am gear carb avg\_mpg\_cg  
## 1: 6 160 110 3.90 2.620 16.46 0 1 4 4 19.75  
## 2: 6 160 110 3.90 2.875 17.02 0 1 4 4 19.75  
## 3: 6 258 110 3.08 3.215 19.44 1 0 3 1 19.75  
## 4: 8 360 175 3.15 3.440 17.02 0 0 3 2 15.05  
## car\_names  
## 1: Mazda RX4  
## 2: Mazda RX4 Wag  
## 3: Hornet 4 Drive  
## 4: Hornet Sportabout

Let’s do now something more tricky. I want to apply a function (say mean) to disp, hp and drat column and store them in new columns of dat

cols\_to\_work <- c("disp", "hp", "drat")  
dat[,paste0(cols\_to\_work,"\_avg"):= lapply(.SD, mean), .SDcols =cols\_to\_work ]  
head(dat)

## cyl disp hp drat wt qsec vs am gear carb avg\_mpg\_cg  
## 1: 6 160 110 3.90 2.620 16.46 0 1 4 4 19.750  
## 2: 6 160 110 3.90 2.875 17.02 0 1 4 4 19.750  
## 3: 4 108 93 3.85 2.320 18.61 1 1 4 1 26.925  
## 4: 6 258 110 3.08 3.215 19.44 1 0 3 1 19.750  
## 5: 8 360 175 3.15 3.440 17.02 0 0 3 2 15.050  
## 6: 6 225 105 2.76 3.460 20.22 1 0 3 1 19.750  
## car\_names disp\_avg hp\_avg drat\_avg  
## 1: Mazda RX4 230.7219 146.6875 3.596563  
## 2: Mazda RX4 Wag 230.7219 146.6875 3.596563  
## 3: Datsun 710 230.7219 146.6875 3.596563  
## 4: Hornet 4 Drive 230.7219 146.6875 3.596563  
## 5: Hornet Sportabout 230.7219 146.6875 3.596563  
## 6: Valiant 230.7219 146.6875 3.596563