Modul 3 DS

Rafly Pradana Putra

12/2/2020

## Dasar Teori

Variasi tipe data pada R memfasilitasi keberagaman jenis variabel data. Sebagai contoh, terdapat data yang terdiri dari sekumpulan angka dan data lain yang berisi sekumpulan karakter. Pada contoh lain, ada pula data yang berbentuk tabel maupun kumpulan (list) angka sederhana. Dengan bantuan fungsi class, kita akan mendapatkan kemudahan dalam mendefinisikan tipe data yang kita miliki:

a <- 2  
class(a)

## [1] "numeric"

Agar dapat bekerja secara efisien dalam menggunakan bahasa pemrograman R, penting untuk mempelajari terlebih dahulu tipe data dari variabel-variabel yang kita miliki sehingga akan mempermudah dalam penentuan proses analisis data yang dapat dilakukan terhadap variabelvariabel tersebut.

## Data Frames

Cara paling umum yang dapat digunakan untuk menyimpan dataset dalam R adalah dalam tipe data frame. Secara konseptual, kita dapat menganggap data frame sebagai tabel yang terdiri dari baris yang memiliki nilai pengamatan dan berbagai variabel yang didefinisikan dalam bentuk kolom. Tipe data ini sangat umum digunakan untuk dataset, karena data frame dapat menggabungkan berbagai jenis tipe data dalam satu objek. Untuk memahami tipe data frame, silahkan mengakses contoh dataset pada library(dslabs) dan pilih dataset “murders” menggunakan fungsi data:

library(dslabs)  
data(murders)

Untuk memastikan bahwa dataset tersebut tipenya adalah data frame, dapat digunakan perintah berikut:

class(murders)

## [1] "data.frame"

Untuk memeriksa lebih lanjut isi dataset, dapat pula digunakan fungsi str untuk mencari tahu lebih rinci mengenai struktur suatu objek:

str(murders)

## 'data.frame': 51 obs. of 5 variables:  
## $ state : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ abb : chr "AL" "AK" "AZ" "AR" ...  
## $ region : Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...  
## $ population: num 4779736 710231 6392017 2915918 37253956 ...  
## $ total : num 135 19 232 93 1257 ...

Dengan menggunakan fungsi str, dapat diketahui bahwa dataset “murders” terdiri dari 51 baris dan lima variabel: state, abb, region, population, dan total. Selanjutnya, untuk melihat contoh enam baris pertama pada dataset, dapat digunakan fungsi head:

head(murders)

## state abb region population total  
## 1 Alabama AL South 4779736 135  
## 2 Alaska AK West 710231 19  
## 3 Arizona AZ West 6392017 232  
## 4 Arkansas AR South 2915918 93  
## 5 California CA West 37253956 1257  
## 6 Colorado CO West 5029196 65

Untuk analisis awal tiap variabel yang diwakili dalam bentuk kolom pada tipe data frame, dapat digunakan operator aksesor ($) dengan cara berikut:

murders$population

## [1] 4779736 710231 6392017 2915918 37253956 5029196 3574097 897934  
## [9] 601723 19687653 9920000 1360301 1567582 12830632 6483802 3046355  
## [17] 2853118 4339367 4533372 1328361 5773552 6547629 9883640 5303925  
## [25] 2967297 5988927 989415 1826341 2700551 1316470 8791894 2059179  
## [33] 19378102 9535483 672591 11536504 3751351 3831074 12702379 1052567  
## [41] 4625364 814180 6346105 25145561 2763885 625741 8001024 6724540  
## [49] 1852994 5686986 563626

Untuk mengetahui nama-nama dari lima variabel yang dapat dievaluasi menggunakan operator aksesor, sebelumnya, melalui fungsi str, telah kita ketahui bahwa variabel yang dimiliki dataset adalah: state, abb, region, population, dan total. Sebagai alternatif, terdapat pula fungsi name, yang dapat digunakan seperti contoh dibawah ini:

names(murders)

## [1] "state" "abb" "region" "population" "total"

print(murders)

## state abb region population total  
## 1 Alabama AL South 4779736 135  
## 2 Alaska AK West 710231 19  
## 3 Arizona AZ West 6392017 232  
## 4 Arkansas AR South 2915918 93  
## 5 California CA West 37253956 1257  
## 6 Colorado CO West 5029196 65  
## 7 Connecticut CT Northeast 3574097 97  
## 8 Delaware DE South 897934 38  
## 9 District of Columbia DC South 601723 99  
## 10 Florida FL South 19687653 669  
## 11 Georgia GA South 9920000 376  
## 12 Hawaii HI West 1360301 7  
## 13 Idaho ID West 1567582 12  
## 14 Illinois IL North Central 12830632 364  
## 15 Indiana IN North Central 6483802 142  
## 16 Iowa IA North Central 3046355 21  
## 17 Kansas KS North Central 2853118 63  
## 18 Kentucky KY South 4339367 116  
## 19 Louisiana LA South 4533372 351  
## 20 Maine ME Northeast 1328361 11  
## 21 Maryland MD South 5773552 293  
## 22 Massachusetts MA Northeast 6547629 118  
## 23 Michigan MI North Central 9883640 413  
## 24 Minnesota MN North Central 5303925 53  
## 25 Mississippi MS South 2967297 120  
## 26 Missouri MO North Central 5988927 321  
## 27 Montana MT West 989415 12  
## 28 Nebraska NE North Central 1826341 32  
## 29 Nevada NV West 2700551 84  
## 30 New Hampshire NH Northeast 1316470 5  
## 31 New Jersey NJ Northeast 8791894 246  
## 32 New Mexico NM West 2059179 67  
## 33 New York NY Northeast 19378102 517  
## 34 North Carolina NC South 9535483 286  
## 35 North Dakota ND North Central 672591 4  
## 36 Ohio OH North Central 11536504 310  
## 37 Oklahoma OK South 3751351 111  
## 38 Oregon OR West 3831074 36  
## 39 Pennsylvania PA Northeast 12702379 457  
## 40 Rhode Island RI Northeast 1052567 16  
## 41 South Carolina SC South 4625364 207  
## 42 South Dakota SD North Central 814180 8  
## 43 Tennessee TN South 6346105 219  
## 44 Texas TX South 25145561 805  
## 45 Utah UT West 2763885 22  
## 46 Vermont VT Northeast 625741 2  
## 47 Virginia VA South 8001024 250  
## 48 Washington WA West 6724540 93  
## 49 West Virginia WV South 1852994 27  
## 50 Wisconsin WI North Central 5686986 97  
## 51 Wyoming WY West 563626 5

Untuk menampilkan grafik, maka tambahkan ini :

plot(murders$population)

