Analisa Dataset state.x77

Husni Mubarok Ramadhan

2023-02-17

# Analisa Dataset state.x77

## Jalankan RStudio dan di R Console atau Code Editor. Ketik dan jalankan perintah berikut.

# Memanggil objek state.x77  
state.x77

Population Income Illiteracy Life Exp Murder HS Grad Frost  
Alabama 3615 3624 2.1 69.05 15.1 41.3 20  
Alaska 365 6315 1.5 69.31 11.3 66.7 152  
Arizona 2212 4530 1.8 70.55 7.8 58.1 15  
Arkansas 2110 3378 1.9 70.66 10.1 39.9 65  
California 21198 5114 1.1 71.71 10.3 62.6 20  
Colorado 2541 4884 0.7 72.06 6.8 63.9 166  
Connecticut 3100 5348 1.1 72.48 3.1 56.0 139  
Delaware 579 4809 0.9 70.06 6.2 54.6 103  
Florida 8277 4815 1.3 70.66 10.7 52.6 11  
Georgia 4931 4091 2.0 68.54 13.9 40.6 60  
Hawaii 868 4963 1.9 73.60 6.2 61.9 0  
Idaho 813 4119 0.6 71.87 5.3 59.5 126  
Illinois 11197 5107 0.9 70.14 10.3 52.6 127  
Indiana 5313 4458 0.7 70.88 7.1 52.9 122  
Iowa 2861 4628 0.5 72.56 2.3 59.0 140  
Kansas 2280 4669 0.6 72.58 4.5 59.9 114  
Kentucky 3387 3712 1.6 70.10 10.6 38.5 95  
Louisiana 3806 3545 2.8 68.76 13.2 42.2 12  
Maine 1058 3694 0.7 70.39 2.7 54.7 161  
Maryland 4122 5299 0.9 70.22 8.5 52.3 101  
Massachusetts 5814 4755 1.1 71.83 3.3 58.5 103  
Michigan 9111 4751 0.9 70.63 11.1 52.8 125  
Minnesota 3921 4675 0.6 72.96 2.3 57.6 160  
Mississippi 2341 3098 2.4 68.09 12.5 41.0 50  
Missouri 4767 4254 0.8 70.69 9.3 48.8 108  
Montana 746 4347 0.6 70.56 5.0 59.2 155  
Nebraska 1544 4508 0.6 72.60 2.9 59.3 139  
Nevada 590 5149 0.5 69.03 11.5 65.2 188  
New Hampshire 812 4281 0.7 71.23 3.3 57.6 174  
New Jersey 7333 5237 1.1 70.93 5.2 52.5 115  
New Mexico 1144 3601 2.2 70.32 9.7 55.2 120  
New York 18076 4903 1.4 70.55 10.9 52.7 82  
North Carolina 5441 3875 1.8 69.21 11.1 38.5 80  
North Dakota 637 5087 0.8 72.78 1.4 50.3 186  
Ohio 10735 4561 0.8 70.82 7.4 53.2 124  
Oklahoma 2715 3983 1.1 71.42 6.4 51.6 82  
Oregon 2284 4660 0.6 72.13 4.2 60.0 44  
Pennsylvania 11860 4449 1.0 70.43 6.1 50.2 126  
Rhode Island 931 4558 1.3 71.90 2.4 46.4 127  
South Carolina 2816 3635 2.3 67.96 11.6 37.8 65  
South Dakota 681 4167 0.5 72.08 1.7 53.3 172  
Tennessee 4173 3821 1.7 70.11 11.0 41.8 70  
Texas 12237 4188 2.2 70.90 12.2 47.4 35  
Utah 1203 4022 0.6 72.90 4.5 67.3 137  
Vermont 472 3907 0.6 71.64 5.5 57.1 168  
Virginia 4981 4701 1.4 70.08 9.5 47.8 85  
Washington 3559 4864 0.6 71.72 4.3 63.5 32  
West Virginia 1799 3617 1.4 69.48 6.7 41.6 100  
Wisconsin 4589 4468 0.7 72.48 3.0 54.5 149  
Wyoming 376 4566 0.6 70.29 6.9 62.9 173  
 Area  
Alabama 50708  
Alaska 566432  
Arizona 113417  
Arkansas 51945  
California 156361  
Colorado 103766  
Connecticut 4862  
Delaware 1982  
Florida 54090  
Georgia 58073  
Hawaii 6425  
Idaho 82677  
Illinois 55748  
Indiana 36097  
Iowa 55941  
Kansas 81787  
Kentucky 39650  
Louisiana 44930  
Maine 30920  
Maryland 9891  
Massachusetts 7826  
Michigan 56817  
Minnesota 79289  
Mississippi 47296  
Missouri 68995  
Montana 145587  
Nebraska 76483  
Nevada 109889  
New Hampshire 9027  
New Jersey 7521  
New Mexico 121412  
New York 47831  
North Carolina 48798  
North Dakota 69273  
Ohio 40975  
Oklahoma 68782  
Oregon 96184  
Pennsylvania 44966  
Rhode Island 1049  
South Carolina 30225  
South Dakota 75955  
Tennessee 41328  
Texas 262134  
Utah 82096  
Vermont 9267  
Virginia 39780  
Washington 66570  
West Virginia 24070  
Wisconsin 54464  
Wyoming 97203

**Tambahkan kode seperti dibawah ini.**

state.x77 <- data.frame(state.x77)  
str(state.x77)

'data.frame': 50 obs. of 8 variables:  
 $ Population: num 3615 365 2212 2110 21198 ...  
 $ Income : num 3624 6315 4530 3378 5114 ...  
 $ Illiteracy: num 2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...  
 $ Life.Exp : num 69 69.3 70.5 70.7 71.7 ...  
 $ Murder : num 15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...  
 $ HS.Grad : num 41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...  
 $ Frost : num 20 152 15 65 20 166 139 103 11 60 ...  
 $ Area : num 50708 566432 113417 51945 156361 ...

# Kode di atas mengubah objek state.x77 menjadi data.frame dan kemudian menampilkan struktur data.frame tersebut dengan method string.

**Tambahkan kode seperti dibawah ini.**

attach(state.x77)  
Income

[1] 3624 6315 4530 3378 5114 4884 5348 4809 4815 4091 4963 4119 5107 4458 4628  
[16] 4669 3712 3545 3694 5299 4755 4751 4675 3098 4254 4347 4508 5149 4281 5237  
[31] 3601 4903 3875 5087 4561 3983 4660 4449 4558 3635 4167 3821 4188 4022 3907  
[46] 4701 4864 3617 4468 4566

quantile(Income)

0% 25% 50% 75% 100%   
3098.00 3992.75 4519.00 4813.50 6315.00

quantile(Income, c(0.5, 0.25, 0.50))

50% 25% 50%   
4519.00 3992.75 4519.00

### Penjelasan

* attach() Kode di atas melakukan attachment objek state.x77, sehingga variabel-variabel di dalamnya dapat dipanggil langsung.
* quantile() Kemudian dilakukan perhitungan quantile untuk variabel Income. Fungsi quantile() digunakan untuk menghitung persentil dari suatu vektor numerik.
* quantile(variabel, c(0.5, 0.25, 0.50)) Pada kode di atas, dilakukan perhitungan persentil ke-0.25, ke-0.5 (median), dan ke-0.75 dari vektor Income.

**Tambahkan kode seperti dibawah ini.**

# baris pertama  
cor(state.x77)[,2:5]

Income Illiteracy Life.Exp Murder  
Population 0.2082276 0.10762237 -0.06805195 0.3436428  
Income 1.0000000 -0.43707519 0.34025534 -0.2300776  
Illiteracy -0.4370752 1.00000000 -0.58847793 0.7029752  
Life.Exp 0.3402553 -0.58847793 1.00000000 -0.7808458  
Murder -0.2300776 0.70297520 -0.78084575 1.0000000  
HS.Grad 0.6199323 -0.65718861 0.58221620 -0.4879710  
Frost 0.2262822 -0.67194697 0.26206801 -0.5388834  
Area 0.3633154 0.07726113 -0.10733194 0.2283902

# baris kedua  
cor(state.x77[,2:5])

Income Illiteracy Life.Exp Murder  
Income 1.0000000 -0.4370752 0.3402553 -0.2300776  
Illiteracy -0.4370752 1.0000000 -0.5884779 0.7029752  
Life.Exp 0.3402553 -0.5884779 1.0000000 -0.7808458  
Murder -0.2300776 0.7029752 -0.7808458 1.0000000

#### Penjelasan :

Kode di atas melakukan perhitungan korelasi antara variabel-variabel di dalam state.x77.

* Pada baris pertama, seluruh variabel digunakan sebagai input, dan kemudian hanya kolom ke-2 sampai ke-5 yang ditampilkan.
* Pada baris kedua, hanya kolom ke-2 sampai ke-5 yang digunakan sebagai input. Fungsi cor() digunakan untuk menghitung korelasi antar variabel. Hasilnya adalah matriks korelasi.

**Tambahkan kode seperti dibawah ini.** **kode pertama**

grupIncome <- cut( Income, breaks=c(0,1000,3500,6500), include.lowest=T, dig.lab=10)  
grupIncome

[1] (3500,6500] (3500,6500] (3500,6500] (1000,3500] (3500,6500] (3500,6500]  
 [7] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500]  
[13] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500]  
[19] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (1000,3500]  
[25] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500]  
[31] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500]  
[37] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500]  
[43] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500] (3500,6500]  
[49] (3500,6500] (3500,6500]  
Levels: [0,1000] (1000,3500] (3500,6500]

#### Penjelasan :

* Fungsi cut() digunakan untuk memotong variabel numerik Income menjadi beberapa kelompok diskrit berdasarkan nilai-nilai batas yang ditentukan dalam vektor breaks.
* Dalam contoh di atas, data Income dibagi menjadi tiga kelompok, yaitu: kelompok 1 (0-1000), kelompok 2 (1000-3500), dan kelompok 3 (3500-6500).
* Argumen include.lowest=T digunakan untuk menyertakan nilai terendah (0) sebagai batas bawah kelompok pertama.
* Argumen dig.lab=10 digunakan untuk menentukan jumlah digit yang digunakan untuk mencetak label kelompok. Dalam hal ini, 10 digit digunakan untuk mencetak label kelompok.

**kode kedua**

grupIncome2 <- cut( Income, breaks=c(0,4000,4500, Inf), include.lowest=T, dig.lab=10)  
grupIncome2

[1] [0,4000] (4500,Inf] (4500,Inf] [0,4000] (4500,Inf] (4500,Inf]   
 [7] (4500,Inf] (4500,Inf] (4500,Inf] (4000,4500] (4500,Inf] (4000,4500]  
[13] (4500,Inf] (4000,4500] (4500,Inf] (4500,Inf] [0,4000] [0,4000]   
[19] [0,4000] (4500,Inf] (4500,Inf] (4500,Inf] (4500,Inf] [0,4000]   
[25] (4000,4500] (4000,4500] (4500,Inf] (4500,Inf] (4000,4500] (4500,Inf]   
[31] [0,4000] (4500,Inf] [0,4000] (4500,Inf] (4500,Inf] [0,4000]   
[37] (4500,Inf] (4000,4500] (4500,Inf] [0,4000] (4000,4500] [0,4000]   
[43] (4000,4500] (4000,4500] [0,4000] (4500,Inf] (4500,Inf] [0,4000]   
[49] (4000,4500] (4500,Inf]   
Levels: [0,4000] (4000,4500] (4500,Inf]

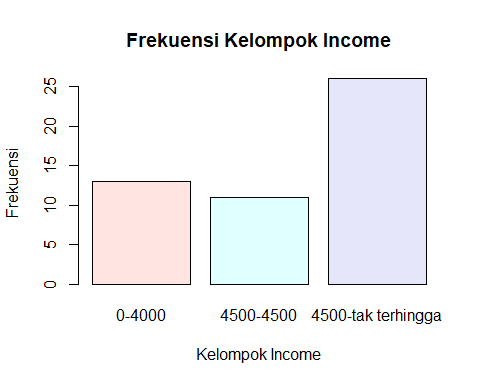
#### Penjelasan :

* Kode di atas hampir sama dengan kode sebelumnya, hanya beda dalam nilai-nilai batas kelompok yang ditentukan dalam vektor breaks.
* Dalam contoh ini, data Income dibagi menjadi tiga kelompok, yaitu: kelompok 1 (0-4000), kelompok 2 (4000-4500), dan kelompok 3 (4500 ke atas atau tidak terbatas).
* Argumen Inf digunakan untuk menunjukkan bahwa tidak ada nilai batas atas untuk kelompok terakhir.
* Argumen dig.lab=10 digunakan untuk menentukan jumlah digit yang digunakan untuk mencetak label kelompok.

Hasil dari kedua kode di atas adalah vektor dengan label-label kelompok sesuai dengan pembagian yang telah ditentukan. Vektor ini dapat digunakan untuk analisis dan visualisasi data dengan menggunakan metode yang membutuhkan variabel diskrit, seperti analisis tabel silang atau diagram batang.

# Tugas Membuat BarPlot dari data diatas

# bikin label terlebih dahulu dari data diatas  
labelGrupIncome2 <- cut( Income, breaks=c(0,4000,4500, Inf),   
 labels = c("0-4000", "4500-4500", "4500-tak terhingga"),   
 include.lowest=T, dig.lab=10)  
  
# Membuat barplot  
barplot(table(labelGrupIncome2),   
 col=c("mistyrose", "lightcyan", "lavender"),   
 main="Frekuensi Kelompok Income", xlab="Kelompok Income", ylab="Frekuensi")



#### Penjelasan :

* table(labelGrupIncome2) menghitung frekuensi data dalam setiap kelompok pada vektor labelGrupIncome2 dan menghasilkan sebuah tabel. col digunakan untuk menentukan warna pada setiap bar pada diagram batang. Di sini, kita menggunakan warna mistyrose untuk kelompok pertama, lightcyan untuk kelompok kedua, dan lavender untuk kelompok ketiga. main digunakan untuk memberikan judul pada diagram batang. xlab digunakan untuk memberikan label pada sumbu x (horizontal). ylab digunakan untuk memberikan label pada sumbu y (vertikal).