Az R adatelemzési nyelv alapjai I.

September 14, 2018

Gyakorlati feladat megoldása

A feladat megoldása során egy olyan adattáblával fogunk dolgozni, amely hipotetikus személyek testmagasság és testsúly adatait tartalmazza.

Gyakorlandó ismeretek:

• adattábla létrehozása

• statisztikai függvények használata

• adattáblák összefűzése

• hiányzó adatok kezelése

• adatok származtatása

• leíró statisztikák készítése adattáblákról

• részhalmazok szűrése

• adattábla exportálása / importálása

1. Hozzon létre egy adattáblát, amely 50 hipotetikus férfi testmagasságát és testsúlyát tartalmazza. Az adatok előállításához sorsoljon normális eloszlású véletlen számokat (testmagasság átlaga = 175 [cm], szórása = 10 [cm]; testsúly átlaga = 80 [kg], szórása = 10 [kg]).

Parancsok: data.frame, rnorm

```
df.ferfi <- data.frame( Magassag = rnorm( n = 50, mean = 175, sd = 10 ),
                        Suly = rnorm( n = 50, mean = 80, sd = 10 ) )

head( df.ferfi )
```

```
##   Magassag      Suly
## 1 190.1325 103.94432
## 2 167.1744  87.59986
## 3 167.0702  91.91458
## 4 186.7983  66.12818
## 5 159.6555  80.22123
## 6 171.9466  73.89548
```

2. Hozzon létre egy másik adattáblát, amely 50 hipotetikus nő testmagasságát és testsúlyát tartalmazza. Az adatok előállításához sorsoljon normális eloszlású véletlen

számokat (testmagasság átlaga = 165 [cm], szórása = 8 [cm]; testsúly átlaga = 65 [kg], szórása = 10 [kg]).

Parancsok: data.frame, rnorm

```
df.no <- data.frame( Magassag = rnorm( n = 50, mean = 165, sd = 8 ),
                     Suly = rnorm( n = 50, mean = 65, sd = 10 ) )

head( df.no )
```

```
##   Magassag     Suly
## 1 174.1159 55.52956
## 2 147.3872 46.49836
## 3 171.8997 69.40236
## 4 161.4531 68.24115
## 5 163.8828 83.78859
## 6 157.1744 61.33684
```

3. Mindkét táblához fűzzön hozzá egy új oszlopot, amely a személy nemét tartalmazza, majd fűzze össze a két táblát egy új táblává. A további feladatok megoldása során ezt az új táblát használja.

Parancsok: rbind

```
df.ferfi$Nem <- "Ferfi"
df.no$Nem <- "No"

df <- rbind( df.ferfi, df.no )

df$Nem <- as.factor( df$Nem )

head( df )
```

```
##   Magassag      Suly   Nem
## 1 190.1325 103.94432 Ferfi
## 2 167.1744  87.59986 Ferfi
## 3 167.0702  91.91458 Ferfi
## 4 186.7983  66.12818 Ferfi
## 5 159.6555  80.22123 Ferfi
## 6 171.9466  73.89548 Ferfi
```

4. Generáljon véletlenszerű hiányzást az adattáblában a testmagasság és a testsúly oszlopba (~10%-nyi hiányzó adatot generáljon). Tipp: sorsoljon véletlen indexeket a megadott oszlopokban és töltse ki “NA”-val.

Parancsok: sample

```
sample( 1:100, 100 * 0.1 )
```

```
##  [1] 24  8 40 83  9 71 63 99 92 39
```

```
df[ sample( 1:100, 10 ), "Magassag" ] <- NA
df[ sample( 1:100, 10 ), "Suly" ] <- NA

df[ !complete.cases( df ), ]
```

```
##    Magassag     Suly   Nem
## 2  167.1744       NA Ferfi
## 5        NA 80.22123 Ferfi
## 6        NA       NA Ferfi
## 12 171.5093       NA Ferfi
## 25       NA 87.42545 Ferfi
## 38 170.3953       NA Ferfi
## 40       NA 86.39088 Ferfi
## 43 171.6596       NA Ferfi
## 46       NA 92.38831 Ferfi
## 50 188.5451       NA Ferfi
## 61       NA 58.23514    No
## 62       NA 62.78106    No
## 63 159.5505       NA    No
## 64       NA 80.42573    No
## 69       NA 76.15001    No
## 72       NA 77.35119    No
## 76 160.6712       NA    No
## 87 151.4992       NA    No
## 88 162.3160       NA    No
```

5. Készítsen leíró statisztikákat az adattábláról, összesítve és nemenkénti bontásban is. Figyeljen a hiányzó értékekre.

Parancsok: summary, split, lapply

```
summary( df )
```

```
##     Magassag          Suly           Nem
##  Min.   :147.4   Min.   : 46.50   Ferfi:50
##  1st Qu.:161.8   1st Qu.: 63.22   No   :50
##  Median :169.6   Median : 71.31
##  Mean   :170.0   Mean   : 71.96
##  3rd Qu.:175.9   3rd Qu.: 79.46
##  Max.   :201.9   Max.   :103.94
##  NA's   :10      NA's   :10
```

```
dfs.list <- split( x = df, f = df$Nem )

lapply( X = dfs.list, FUN = summary )
```

```
## $Ferfi
##     Magassag          Suly           Nem
##  Min.   :155.7   Min.   : 59.87   Ferfi:50
##  1st Qu.:167.4   1st Qu.: 71.60   No   : 0
##  Median :174.2   Median : 78.48
```

```
##  Mean   :175.5   Mean   : 79.15
##  3rd Qu.:183.4   3rd Qu.: 85.63
##  Max.   :201.9   Max.   :103.94
##  NA's   :5       NA's   :6
##
## $No
##     Magassag          Suly          Nem
##  Min.   :147.4   Min.   :46.50   Ferfi: 0
##  1st Qu.:158.7   1st Qu.:59.45   No   :50
##  Median :165.1   Median :66.79
##  Mean   :164.4   Mean   :65.09
##  3rd Qu.:170.6   3rd Qu.:71.33
##  Max.   :181.5   Max.   :83.79
##  NA's   :5       NA's   :4
```

6. Pótolja a hiányzó adatokat a nemeknek megfelelő átlagértékek alapján.

Parancsok: mean, is.na

```
mean( df$Magassag[ df$Nem == "Ferfi" ] )

## [1] NA

mean( df$Magassag[ df$Nem == "Ferfi" ], na.rm = T )

## [1] 175.5299

df$Magassag[ is.na( df$Magassag ) & df$Nem == "Ferfi" ] <-
  mean( df$Magassag[ df$Nem == "Ferfi" ], na.rm = T )
df[ is.na( df$Suly ) & df$Nem == "Ferfi", "Suly" ] <- mean( df$Suly[ df$Nem
== "Ferfi" ], na.rm = T )

df[ is.na( df$Magassag ) & df$Nem == "No", "Magassag" ] <- mean( df$Magassag[
df$Nem == "No" ], na.rm = T )
df[ is.na( df$Suly ) & df$Nem == "No", "Suly" ] <- mean( df$Suly[ df$Nem ==
"No" ], na.rm = T )

summary( df )

##     Magassag          Suly           Nem
##  Min.   :147.4   Min.   : 46.50   Ferfi:50
##  1st Qu.:162.6   1st Qu.: 65.04   No   :50
##  Median :169.6   Median : 71.64
##  Mean   :170.0   Mean   : 72.12
##  3rd Qu.:175.6   3rd Qu.: 79.15
##  Max.   :201.9   Max.   :103.94

lapply( X = split( x = df, f = df$Nem ), FUN = summary )

## $Ferfi
##     Magassag          Suly           Nem
##  Min.   :155.7   Min.   : 59.87   Ferfi:50
```

```
##  1st Qu.:168.4   1st Qu.: 73.13   No   : 0
##  Median :175.5   Median : 79.15
##  Mean   :175.5   Mean   : 79.15
##  3rd Qu.:181.9   3rd Qu.: 84.34
##  Max.   :201.9   Max.   :103.94
##
## $No
##     Magassag          Suly          Nem
##  Min.   :147.4   Min.   :46.50   Ferfi: 0
##  1st Qu.:159.0   1st Qu.:59.89   No   :50
##  Median :164.4   Median :65.19
##  Mean   :164.4   Mean   :65.09
##  3rd Qu.:169.6   3rd Qu.:71.13
##  Max.   :181.5   Max.   :83.79
```

7. Számítsa ki a BMI értékét minden személyre, és ezt fűzze az adattáblához egy új oszlopba (“BMI”).

BMI = testsúly [kg] / (testmagasság [m])^2

```
df$BMI <- df$Suly / ( ( df$Magassag / 100 ) * ( df$Magassag / 100 ) )

summary( df$BMI )

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.20   22.21   25.01   25.08   27.95   36.96
```

8. A BMI értéke alapján sorolja be a személyeket az alábbi kategóriákba:

• 18.5 alatt: Alultáplált

• 18.5 és 25 között: Normális

• 25 és 30 között: Túlsúlyos

• 30 felett: Elhízott

Ezt fűzze hozzá az adattáblához egy új oszlopba (“BMI.Status”).

Parancsok: cut

```
df$BMI.Status <- cut( x = df$BMI,
                      breaks = c( 0, 18.5, 25, 30, Inf ),
                      labels = c( "Alultaplalt",
                                  "Normalis",
                                  "Tulsulyos",
                                  "Elhizott" ) )

lapply( X = split( x = df$BMI, f = df$BMI.Status ),
        FUN = summary )
```

```
## $Alultaplalt
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.20   17.23   17.37   17.51   17.58   18.32
##
## $Normalis
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.95   20.93   22.56   22.25   23.31   24.96
##
## $Tulsulyos
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.06   25.81   26.89   27.16   28.29   29.99
##
## $Elhizott
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.08   31.07   32.45   32.67   33.89   36.96
```

9. Szűrje le

• a túlsúlyos személyeket,

• az elhízott férfiakat,

• azokat a normális nőket, akiknek a testmagassága átlagon aluli

• azokat a túlsúlyos vagy elhízott nőket, akiknek a testsúlya meghaladja a normális férfiak átlagos testsúlyát

Parancsok: pl. subset, mean stb.

```
# túlsúlyos személyek
subset( df, BMI.Status == "Tulsulyos" )
```

```
##    Magassag      Suly   Nem      BMI BMI.Status
## 1  190.1325 103.94432 Ferfi 28.75333  Tulsulyos
## 2  167.1744  79.15136 Ferfi 28.32167  Tulsulyos
## 5  175.5299  80.22123 Ferfi 26.03676  Tulsulyos
## 6  175.5299  79.15136 Ferfi 25.68952  Tulsulyos
## 10 163.3361  74.47740 Ferfi 27.91643  Tulsulyos
## 12 171.5093  79.15136 Ferfi 26.90811  Tulsulyos
## 14 174.0715  81.13109 Ferfi 26.77517  Tulsulyos
## 20 178.5602  88.92385 Ferfi 27.89001  Tulsulyos
## 24 161.6493  77.81527 Ferfi 29.77947  Tulsulyos
## 25 175.5299  87.42545 Ferfi 28.37498  Tulsulyos
## 30 175.6984  79.42503 Ferfi 25.72893  Tulsulyos
## 34 171.8937  79.13792 Ferfi 26.78333  Tulsulyos
## 35 185.9269  98.15627 Ferfi 28.39446  Tulsulyos
## 37 177.7798  80.14026 Ferfi 25.35632  Tulsulyos
## 38 170.3953  79.15136 Ferfi 27.26110  Tulsulyos
## 39 183.4197  84.59552 Ferfi 25.14523  Tulsulyos
## 40 175.5299  86.39088 Ferfi 28.03920  Tulsulyos
## 43 171.6596  79.15136 Ferfi 26.86101  Tulsulyos
## 46 175.5299  92.38831 Ferfi 29.98574  Tulsulyos
```

```
## 47 173.9756  85.37346 Ferfi 28.20632  Tulsulyos
## 54 161.4531  68.24115    No 26.17902  Tulsulyos
## 57 166.6034  71.92607    No 25.91306  Tulsulyos
## 59 160.5033  66.88451    No 25.96315  Tulsulyos
## 63 159.5505  65.08511    No 25.56732  Tulsulyos
## 64 164.4123  80.42573    No 29.75270  Tulsulyos
## 67 158.6940  63.10831    No 25.05910  Tulsulyos
## 68 169.0014  73.17807    No 25.62124  Tulsulyos
## 69 164.4123  76.15001    No 28.17094  Tulsulyos
## 72 164.4123  77.35119    No 28.61531  Tulsulyos
## 73 166.2847  71.25569    No 25.77004  Tulsulyos
## 75 158.6903  69.85208    No 27.73823  Tulsulyos
## 76 160.6712  65.08511    No 25.21188  Tulsulyos
## 77 166.1219  69.97987    No 25.35824  Tulsulyos
## 78 158.7912  69.28789    No 27.47924  Tulsulyos
## 83 170.5520  77.03911    No 26.48485  Tulsulyos
## 87 151.4992  65.08511    No 28.35705  Tulsulyos
## 90 150.5156  60.87875    No 26.87217  Tulsulyos
## 95 155.7239  72.20718    No 29.77622  Tulsulyos

# elhízott férfiak
subset( df, BMI.Status == "Elhizott" & Nem == "Ferfi" )

##    Magassag      Suly   Nem      BMI BMI.Status
## 3  167.0702  91.91458 Ferfi 32.92960   Elhizott
## 7  167.7869  97.37347 Ferfi 34.58792   Elhizott
## 11 159.8953  94.49659 Ferfi 36.96108   Elhizott
## 22 164.9787  83.56930 Ferfi 30.70371   Elhizott
## 26 170.5348 100.27674 Ferfi 34.48056   Elhizott
## 27 167.3873  89.78147 Ferfi 32.04364   Elhizott
## 28 159.1946  77.42113 Ferfi 30.54940   Elhizott
## 31 157.6601  79.47437 Ferfi 31.97303   Elhizott
## 44 155.7364  79.70210 Ferfi 32.86166   Elhizott
## 49 165.3033  82.19995 Ferfi 30.08211   Elhizott

# normális nők, akiknek a testmagassága átlagon aluli
subset( df, BMI.Status == "Normalis" & Nem == "No" & Magassag <
          mean( df$Magassag[ df$Nem == "No" ] ) )

##    Magassag     Suly Nem      BMI BMI.Status
## 52 147.3872 46.49836  No 21.40515   Normalis
## 56 157.1744 61.33684  No 24.82892   Normalis
## 60 163.4124 54.10679  No 20.26196   Normalis
## 65 153.1614 54.15330  No 23.08481   Normalis
## 66 147.7659 50.60982  No 23.17856   Normalis
## 79 161.1582 61.94167  No 23.84943   Normalis
## 82 152.3506 49.24668  No 21.21724   Normalis
## 85 162.6723 56.99501  No 21.53821   Normalis
## 88 162.3160 65.08511  No 24.70354   Normalis
## 92 154.7100 59.73969  No 24.95894   Normalis
```

```
# túlsúlyos vagy elhízott nők, akiknek a testsúlya meghaladja a normális
férfiak átlagos testsúlyát
subset( df, ( BMI.Status == "Tulsulyos" | BMI.Status == "Elhizott" ) & Nem ==
"No" & Suly > mean( df$Suly[ df$Nem == "Ferfi" ] ) )
```

```
##    Magassag     Suly Nem      BMI BMI.Status
## 55 163.8828 83.78859  No 31.19740   Elhizott
## 64 164.4123 80.42573  No 29.75270  Tulsulyos
## 94 156.0953 82.09015  No 33.69078   Elhizott
```

10. Exportálja az adattáblát egy szöveges fájlba, majd töltse be onnan.

Parancsok: write.table, read.table

```
write.table( x = df,
             file = "tablazat.txt",
             quote = F,
             sep = "\t",
             row.names = F,
             col.names = T )

df.2 <- read.table( file = "tablazat.txt",
                    header = T,
                    sep = "\t",
                    quote = "" )

head( df.2 )
```

```
##   Magassag      Suly   Nem      BMI BMI.Status
## 1 190.1325 103.94432 Ferfi 28.75333  Tulsulyos
## 2 167.1744  79.15136 Ferfi 28.32167  Tulsulyos
## 3 167.0702  91.91458 Ferfi 32.92960   Elhizott
## 4 186.7983  66.12818 Ferfi 18.95137   Normalis
## 5 175.5299  80.22123 Ferfi 26.03676  Tulsulyos
## 6 175.5299  79.15136 Ferfi 25.68952  Tulsulyos
```