Analiza datelor în R

Curs 2

Pachete

- Pachete: colecţii de funcţii, documentaţie aferentă şi seturi de date
- În prezent sunt disponibile >13000 pachete R (`https://cran.r-project.org/web/packages/available_packages_by_name.html`).
- Instalare:
  `install.packages("nume_pachet")`
- Încărcare în sesiunea de lucru curentă:
  `library(nume_pachet)`
- Identificarea pachetelor încărcate pentru sesiunea de lucru curentă:
  `sessionInfo()`
- Un set de date dintr-un pachet deja încărcat se aduce în workspace cu
  `data(nume_set_date)`

Funcţii pentru vectori

1. Funcţii specifice pentru vectori logici
   - `all(x)` $\rightarrow$ T dacă toate componentele lui x sunt T; F în rest
   - `any(x)` $\rightarrow$ T dacă măcar o componentă a lui x e T; F în rest
   - `which(x)` $\rightarrow$ poziţiile componentelor T din x

2. Funcţii pentru vectori numerici
   - `min(x), max(x)`
   - `range(x)` $\rightarrow$ cea mai mică şi cea mai mare valoare din x
   - `which.min(x), which.max(x)` $\rightarrow$ poziţia primei apariţii în x pentru cea mai mică , respectiv cea mai mare valoare

- `sum(x)` → suma elementelor lui x
- `prod(x)` → produsul elementelor lui x
- `cumsum(x)` → vectorul sumelor cumulate
- `cumprod(x)` → vectorul produselor cumulate
- `diff(x)` → vectorul diferenţelor între elemente succesive

3. Funcţii generale
   - `length(x)` → numărul de elemente din x
   - `rev(x)` → vectorul elementelor în ordine inversă
   - `sort(x)`, `sort(x, decreasing=T)` - funcţii de ordonare
   - `order(x)` → poziţiile elementelor lui x ordonate crescător după valoare
   - `unique(x)` → elimină valorile care se repetă din x

Simbolul NA

`NA` (not available) - codificare pentru date lipsă

Orice operaţie care implică NA are ca rezultat NA.

- Unele funcţii au opţiunea `na.rm=TRUE`, care elimină valorile lipsă înainte de efectuarea calculelor.
- La sortare crescătoare, valorile NA sunt puse pe ultimele poziţii.
- Funcţia `is.na(x)` returnează un vector de valori logice, respectiv T acolo unde componenta corespunzătoare a lui x este NA si F în rest.

Factori

Tipuri de date statistice: $\longleftrightarrow$ Reprezentare în R:

- cantitative $\longleftrightarrow$ numeric
  - discrete
  - continue
- calitative (categoriale) $\longleftrightarrow$ factor
  - binare
  - nominale
  - ordinale

O structură factor reprezintă valorile (categoriile) variabilei calitative printr-un vector de numere întregi $\in \{1, 2, \ldots, k\}$ (k = numărul de categorii) şi un vector intern de şiruri de caractere corespunzătoare valorilor întregi.

Factori

Exemplu:

```
optiuniVot=c("C","A","B","C","C","B","A")
class(optiuniVot)
```

$\rightarrow$ *character*

```
optiuniVot=factor(optiuniVot)
class(optiuniVot)
```

$\rightarrow$ *factor*

```
levels(optiuniVot)
```

$\rightarrow A, B, C$

Variabila optiuniVot este reprezentată intern prin vectorul 3, 1, 2, 3, 3, 2, 1, cu convenţia 1=A, 2=B, 3=C. Implicit, nivelele sunt codificate în ordinea crescătoare a tipului de date din care factorul a fost construit.

Se poate opta pentru o altă codificare a nivelelor decât cea implicită:

```
optiuniVot=factor(optiuniVot,levels=c("B",
          "A", "C"))
levels(optiuniVot)
```

$$\rightarrow B, A, C$$

```
as.numeric(optiuniVot)
```

$$\rightarrow 3, 2, 1, 3, 3, 1, 2$$

Factori ordonaţi

Pentru reprezentarea variabilelor calitative ordinale, acolo unde relaţia de ordine între categorii este importantă, se utilizează la crearea variabilei factor opţiunea `ordered=T` şi se specifică nivelele în ordinea crescătoare dorită.

Exemplu:

```
x=c("mic","mediu","mic","mare","mediu","mic")
x=factor(x,ordered=T,levels=c("mic","mediu","mare"))
```

Două variabile de tip factor ordonat care au aceleaşi nivele şi aceeaşi structură de ordine pot fi comparate (element cu element).

Dataframe-uri

Cea mai uzuală modalitate de reprezentare a seturilor de date statistice este cea de tabel, unde

- liniile = obiecte (indivizi)
- coloanele = atribute (variabile)

$\longrightarrow$ În R: **dataframe**

Coloanele=vectori de aceeaşi lungime. Ele pot avea tipuri diferite.

Un dataframe poate fi creat:

- dintr-un set de vectori preexistenţi

```
nume=c("Popescu", "Ionescu", "Anton")
prenume=c("Mircea", "Alina", "Andrei")
varsta=c(24,31,45)
bd=data.frame(nume,prenume,varsta)
```

- dintr-un set de date extern

Importarea / exportarea seturilor de date

Import

- date în format .csv:
  `numeDF=read.csv("numefisier.csv")` (! opţiuni)
- date în format .txt:
  `numeDF=read.table("numefisier.txt")`
- alte formate proprietare $\rightarrow$ pachetul *foreign*

Export

`write.csv(numeDF, "numefisier.csv")`

Pentru detalii, a se vedea R Data Import/Export manual
`https://cran.r-project.org/doc/manuals/R-data.pdf`

Funcţii pentru dataframe-uri

- `attach(numeDF)` $\rightarrow$ ataşarea unui dataframe la sesiunea de lucru curentă
- `detach(numeDF)` $\rightarrow$ eliminarea dataframe-ului din sesiunea curentă
- `str(numeDF)` $\rightarrow$ informaţii despre conţinut
- `names(numeDF)` $\rightarrow$ afişarea numelor coloanelor (variabilelor)
- `dim(numeDF)` $\rightarrow$ numărul de linii şi de coloane
- `edit(numeDF)` $\rightarrow$ deschidere în format spreadsheet editabil pentru modificarea conţinutului

Exemplu: Setul de date *Cars93* din pachetul *MASS*

Selectarea datelor dintr-un dataframe

- O variabilă `v` dintr-un dataframe `d` se accesează cu `d$v`. Dacă dataframe-ul a fost deja ataşat la sesiunea de lucru curentă, variabilele sale pot fi accesate direct utilizând numele lor.
- selectarea liniei i: `d[i, ]`
- selectarea coloanei j: `d[ ,j]`
- selectarea elementului de pe poziţia (i,j): `d[i,j]`
- selectarea mai multor linii (coloane) specificate:

```
d[istart:ifinal, ]
d[c(i1,i2,...,ik), ]
d[-c(i1,i2,...,ik), ]
d[ ,jstart:jfinal]
d[ ,c(j1,j2,...,jk)]
d[ ,-c(j1,j2,...,jk)]
```

- selectare condiţionată a obiectelor:

```
d1=subset(d,conditie)
```

Exerciţii

Se consideră setul de date *Cars93* din pachetul MASS.

i) Să se construiască un dataframe care conţine doar variabilele Manufacturer, Make, Price, Passengers şi Origin.

ii) Să se selecteze şi afişeze doar datele pentru maşinile de provenienţă americană. Să se determine câte astfel de maşini sunt în baza de date.

iii) Să se afişeze modelele de maşini al căror producător este Ford.

iv) Să se selecteze şi afişeze datele pentru maşinile care pot transporta cel puţin 5 pasageri, sortate crescător după preţ.