Obrada podataka

Dijana Stojić

Fakultet tehničkih nauka u Čačku,

6. april 2019. godine

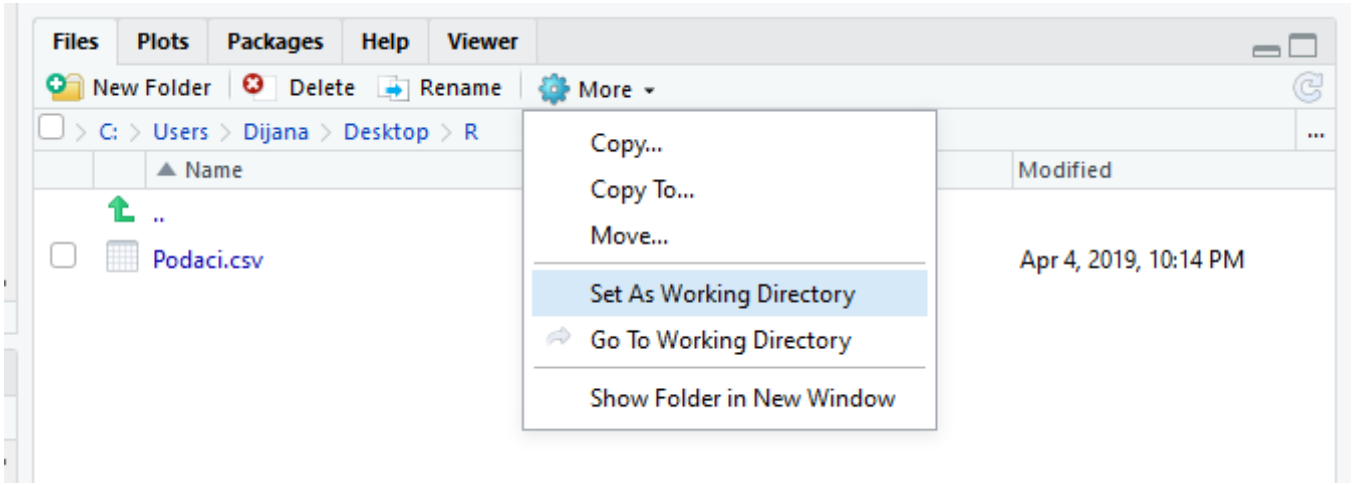
Postavljanje radnog direktorijuma

- Postavljanje radnog direktorijuma može se izvršiti na dva načina:
  - 1. Pomoću naredbe:

    ```
    setwd("C:/Users/Dijana/Desktop/R")
    ```

    - -pod navodnicima se piše putanja do radnog direktorijuma

  - 2. Tako što se u tabu Files u donjem desnom prozoru odabere Go to directory (…) i odabere direktorijum. Da bi ovaj direktorijum postao radni klikne se na More i odabere Set As Working Directory.

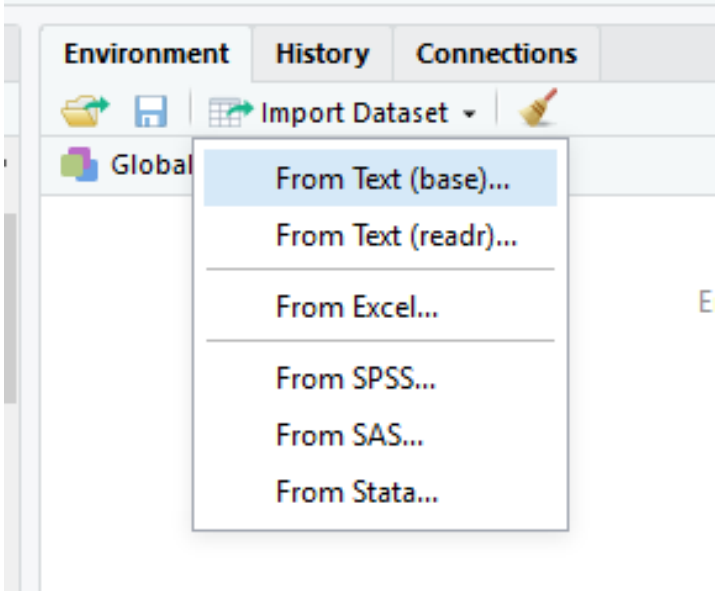


Uvoz csv datoteke

- Uvoz csv datoteke može se izvršiti na dva načina:
  - 1. Pomoću naredbe:
    - -pod navodnicima se piše ime datoteke, header je TRUE jer postoji zaglavnje u datoteci (prvi red u fajlu su nazivi kolona), podaci je ime promenljive u koju smeštamo podatke

```
podaci = read.csv("Podaci.csv", header = TRUE)
```

  - 2. Tako što se u tabu Enviroment u gornjem desnom prozoru odabere Import Dataset i odabere From Text.



Potom se odabere datoteka, podese se parametri (Ime promenljive u koju smeštamo podatke, šta upisati umesto podataka koji ne postoje, itd.) i klikne se na Import.

- Provera tipa podataka:

```
class(podaci)
```

- Rezultat

```
[1] "data.frame"
```

- Prikaz kolona, broja kolona, broja redova i ukupan broj podataka

```
names(podaci)
```

```
[1] "date_time"
[2] "station_id"
[3] "so2"
[4] "pm10"
[5] "o3"
[6] "no2"
[7] "nox"
[8] "co"
[9] "benzene"
[10] "toluene"
[11] "no"
[12] "pm2_5"
[13] "pm1"
[14] "wind_velocity"
[15] "wind_direction"
[16] "pressure"
[17] "temp"
[18] "humidity"
```

```
dim(podaci)
```

```
[1] 58632 18
```

```
r = nrow(podaci)
r
```

```
[1] 58632
```

```
c = ncol(podaci)
c
```

```
[1] 18
```

```
total = r * c
total
```

```
[1] 1055376
```

- Broj podataka razlicitih od NA:

```
sum(!is.na(podaci))
```

```
[1] 462758
```

- Broj podataka jednakih NA:

```
sum(is.na(podaci))
```

```
[1] 592618
```

- Najveća vrednost u koloni (so2) različita od NA:

```
max(podaci$so2, na.rm = TRUE)
```

```
[1] 1357.448
```

- Najvećae vrednosti po svim kolonama:

```
apply(podaci, 2, function(x) max(x, na.rm = TRUE))
```

```
            date_time            station_id                   so2 
"2019-04-04 20:00:00"                  "59"       "1357.44818323" 
                  pm10                    o3                   no2 
       "1733.6731667"          "307.706000"          "431.3589500" 
                   nox                    co               benzene 
        "1483.654804"       "110.613322188"           "66.9792178" 
               toluene                    no                 pm2_5 
        "190.9600260"          "930.178080"          "490.7683333" 
                   pm1          wind_velocity        wind_direction 
        "309.9066667"         "10.323272843"          "9.999738e+01" 
              pressure                  temp              humidity 
         "1034.4320"         " 9.998167e+00"          "100.000000" 
```

- Sortiranje po koloni (so2) u rastućem poretku:

```
podaci[order(podaci$so2),]
```

- Uvoz biblioteke tidyverse:

```
library(tidyverse)
```

- Ekstrakcija podataka za Cacak (station_id = 31) pipe-forward operator %>% omogucava prosledjivanje medju-rezultata sledecoj funkciji

```
cacak = podaci %>% filter(station_id == 31)
```

- Ovim je kreiran novi podatak cacak (tabela)

cacak

Filter

	date_time	station_id	so2	pm10	o3	no2	nox	co	benzene	toluene	no	pm2
1	2019-02-07 00:00:00	31	NA	NA	NA	21.73373	81.10377	0.8477171	NA	NA	38.71920	N
2	2019-02-07 01:00:00	31	NA	NA	NA	22.43194	80.95712	0.8406889	NA	NA	38.19920	N
3	2019-02-07 02:00:00	31	NA	NA	NA	22.41919	82.06979	0.8219534	NA	NA	38.92928	N
4	2019-02-07 03:00:00	31	NA	NA	NA	21.88676	80.95074	0.7437304	NA	NA	38.56112	N
5	2019-02-07 04:00:00	31	NA	NA	NA	21.68591	80.33861	0.7151515	NA	NA	38.28240	N
6	2019-02-07 05:00:00	31	NA	NA	NA	21.12160	78.57874	0.7119869	NA	NA	37.49616	N
7	2019-02-07 06:00:00	31	NA	NA	NA	24.72104	85.03797	0.9217465	NA	NA	39.26832	N
8	2019-02-07 07:00:00	31	NA	NA	NA	42.66723	111.25426	1.3113473	NA	NA	44.76576	N
9	2019-02-07 08:00:00	31	NA	NA	NA	37.66500	104.37420	1.3291121	NA	NA	43.53024	N
10	2019-02-07 09:00:00	31	NA	NA	NA	40.98069	108.51882	1.2805552	NA	NA	44.01280	N

- Brišemo kolonu station_id (jer su svi podaci u njoj jednaki, pa nam nije od značaja)

```
cacak$station_id = NULL
```

- Broj podataka koji su razliciti od NA u jednoj koloni (so2):

```
sum(!is.na(cacak$so2))
```

```
[1] 0
```

- Pošto je taj broj jednak nuli, dolazimo do zaključka da su nam svi dodaci u ovoj koloni jednaki NA, i zbog toga nam ni ova kolona nije od značaja (možemo je obrisati)

```
cacak$so2 = NULL
```

- Broj podataka koji su razliciti od NA po svim kolonama:

```
apply(cacak, 2, function(x) sum(!is.na(x)))
```

```
date_time          pm10            o3           no2           nox
     1364             0             0          1356          1356
        co       benzene       toluene            no         pm2_5
     1356             0             0          1356             0
       pm1 wind_velocity wind_direction      pressure          temp
         0             0             0             0             0
  humidity
         0
```

- Brisanje svih kolona koje su cele NA:

```
for (kolona in names(cacak)){
+ if (sum(!is.na(cacak[[kolona]])) == 0){
+ cacak[[kolona]] = NULL
+ }
+ }
```

cacak

Filter

	date_time	no2	nox	co	no
1	2019-02-07 00:00:00	21.73373	81.10377	0.8477171	38.71920
2	2019-02-07 01:00:00	22.43194	80.95712	0.8406889	38.19920
3	2019-02-07 02:00:00	22.41919	82.06979	0.8219534	38.92928
4	2019-02-07 03:00:00	21.88676	80.95074	0.7437304	38.56112
5	2019-02-07 04:00:00	21.68591	80.33861	0.7151515	38.28240
6	2019-02-07 05:00:00	21.12160	78.57874	0.7119869	37.49616
7	2019-02-07 06:00:00	24.72104	85.03797	0.9217465	39.26832
8	2019-02-07 07:00:00	42.66723	111.25426	1.3113473	44.76576
9	2019-02-07 08:00:00	37.66500	104.37420	1.3291121	43.53024
10	2019-02-07 09:00:00	40.98069	108.51882	1.2805552	44.01280

- Kada pogledamo tabelu cacak detaljnije, videćemo da su neke NA ostale.
- Šta raditi sa njima?
- Možemo brisati te redove, menjati NA sa srednjom vrednošću ili pokušati predvideti vrednost

- Brojimo koliko ima NA po svakoj koloni

```
apply(cacak, 2, function(x) sum(is.na(x)))
```

```
date_time  no2  nox  co  no
        0    8    8   8   8
```

- Pošto ih ima malo, date redove možemo obrisati
- Posmatramo redove koji imaju NA

```
redovi.sa.na = apply(cacak, 1, function(x) any(is.na(x)))
```

```
redovi.sa.na
```

```
 [690] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [703] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [716] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [729] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [742] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [755] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [768] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [781] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [794]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [807] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [820] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [833] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [846] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [859] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
 [872] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [885] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [898] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [911] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [924] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [937] FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [950] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [963] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [976] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [989] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [ reached getOption("max.print") -- omitted 364 entries ]
```

- U redu gde je pronađen NA pojavljuje se TRUE vrednost.

- Prebrojimo koliko ima redova sa NA

```
sum(redovi.sa.na)
[1] 16
```

- Pošto ih ima samo 16, obrisaćemo ih

```
cacak = cacak[!redovi.sa.na,]
```

- Ponovo ćemo proveriti da li je neki NA ostao u tabeli

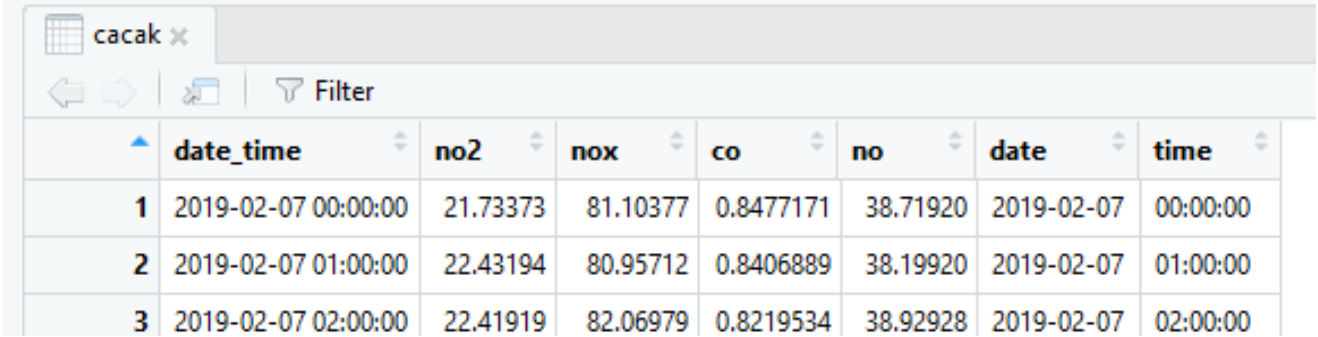
```
apply(cacak, 2, function(x) sum(is.na(x)))

date_time  no2  nox  co  no
        0    0    0   0   0
```

- Vidimo da ih više nema.
- Ostalo je još da razdvojimo kolonu date_time na datum (date) i vreme (time) jer će nam ti podaci trebati razdvojeni.

```
cacak$date = as.Date(cacak$date_time)

cacak$time = format(as.POSIXct(cacak$date_time), format = "%H:%M:%S")
```



	date_time	no2	nox	co	no	date	time
1	2019-02-07 00:00:00	21.73373	81.10377	0.8477171	38.71920	2019-02-07	00:00:00
2	2019-02-07 01:00:00	22.43194	80.95712	0.8406889	38.19920	2019-02-07	01:00:00
3	2019-02-07 02:00:00	22.41919	82.06979	0.8219534	38.92928	2019-02-07	02:00:00

- Ostalo je još da promenimo redosled kolona i dobijamo konačan izgled tabele cacak

```
cacak = cacak[,c(1,6,7,2,3,4,5)]
```

cacak

Filter

	date_time	date	time	no2	nox	co	no
1	2019-02-07 00:00:00	2019-02-07	00:00:00	21.73373	81.10377	0.8477171	38.71920
2	2019-02-07 01:00:00	2019-02-07	01:00:00	22.43194	80.95712	0.8406889	38.19920
3	2019-02-07 02:00:00	2019-02-07	02:00:00	22.41919	82.06979	0.8219534	38.92928
4	2019-02-07 03:00:00	2019-02-07	03:00:00	21.88676	80.95074	0.7437304	38.56112
5	2019-02-07 04:00:00	2019-02-07	04:00:00	21.68591	80.33861	0.7151515	38.28240
6	2019-02-07 05:00:00	2019-02-07	05:00:00	21.12160	78.57874	0.7119869	37.49616
7	2019-02-07 06:00:00	2019-02-07	06:00:00	24.72104	85.03797	0.9217465	39.26832
8	2019-02-07 07:00:00	2019-02-07	07:00:00	42.66723	111.25426	1.3113473	44.76576
9	2019-02-07 08:00:00	2019-02-07	08:00:00	37.66500	104.37420	1.3291121	43.53024
10	2019-02-07 09:00:00	2019-02-07	09:00:00	40.98069	108.51882	1.2805552	44.01280
11	2019-02-07 10:00:00	2019-02-07	10:00:00	35.65327	98.05207	0.8602204	40.79088
12	2019-02-07 11:00:00	2019-02-07	11:00:00	32.83174	94.11787	0.7733771	39.97968
13	2019-02-07 12:00:00	2019-02-07	12:00:00	31.93587	90.61726	0.7680574	38.25120

- Ovako obrađeni podaci pogodni su za vizuelizaciju.