Seminar 2

Justas Mundeikis

2019-04-26

Turinys

1 Apie seminarą

1.1 Seminaro tikslai

- pasikartoti Git, Github, CLI (lieka aktualu iki pat egzamino)
- susipažinti su LSD
- Pasikartoti ir praktiškai pritaikyti paskaitų metu įgytas žinias (R 2.1 ir R 2.2)
- Toliau gilinti savo žinias rašant funkcijas

1.2 Seminaro aptarimas:

- 2019-04-25 I ir II srautai kartu

2 Nedarbo lygis Lietuvoje

Eikite į Lietuvos Statistikos departamento (LSD) tinklapį. Jame esančioje duomenų bazėje susiraskite Nedarbo lygio statistiką ("Amžius (tikslinės grupės")| Gyvenamoji vietovė | Lytis (1998-2018)..). Pasirinkite visus laikotarpius, bei pritaikykite pakeitimus. Parsisiųskite duomenis .csv formatu. Kaip naudotis duomenų baze rasite https://osp.stat.gov.lt/documents/10180/637156/RDB_naudotojo_vadovas.pdf

Importuokite duomenis

Pasitikrinkite ar importavimas įvyko teisingai bei matote:

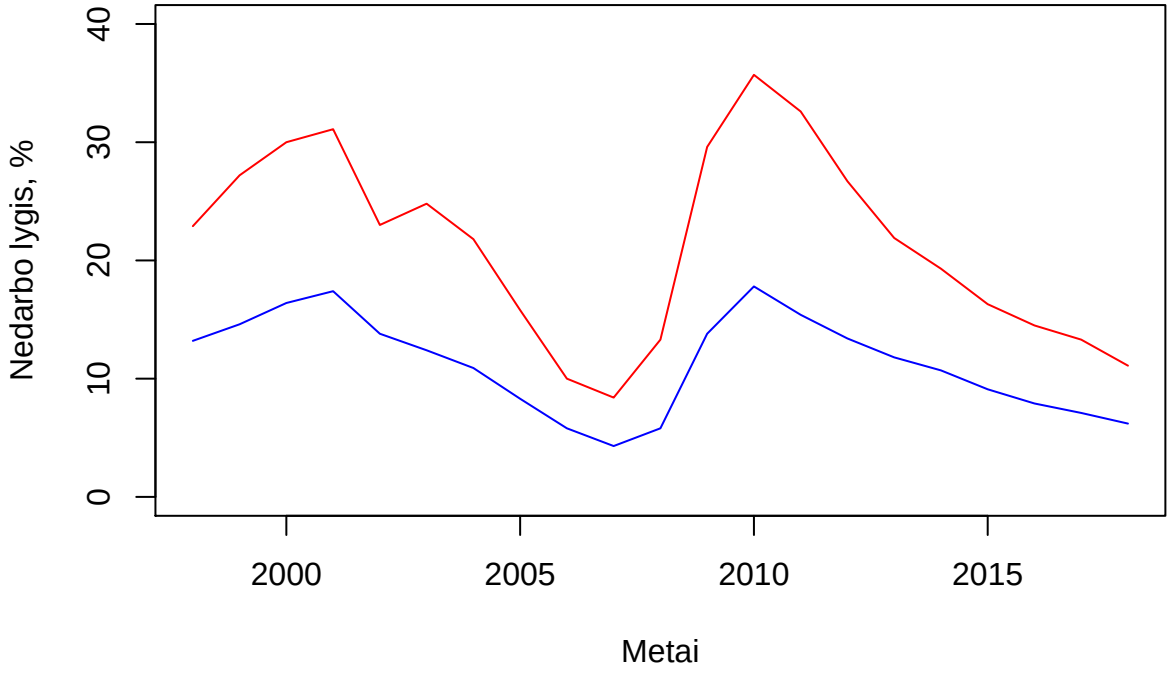
```
str(data)
## 'data.frame':    1701 obs. of  7 variables:
##  $ Laikotarpis                     : int  1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 ...
##  $ Rodiklis                        : chr  "Nedarbo lygis" "Nedarbo lygis" "Nedarbo lygis" "Nedarbo l
##  $ Amžius..tikslinės.amžiaus.grupės.: chr  "Iš viso pagal amžiaus grupes" "Iš viso pagal amžiaus grup
##  $ Gyvenamoji.vietovė              : chr  "Miestas ir kaimas" "Miestas ir kaimas" "Miestas ir kaimas
```

```
##  $ Lytis                           : chr  "Vyrai ir moterys" "Vyrai ir moterys" "Vyrai ir moterys" "
##  $ Matavimo.vienetai               : chr  "proc." "proc." "proc." "proc." ...
##  $ Reikšmė                         : num  13.2 14.6 16.4 17.4 13.8 12.4 10.9 8.3 5.8 4.2 ...
```

Apskaičiuokite vidutinį nedarbo lygį laikotarpiui 1998-2018 pagal amžiaus grupes. Jums pgali būti prasminga prieš skaičiuojant, susikurti naują R objektą, kuriame būtų tik "Miestas ir kaimas", "Vyrai ir moterys". Apskaičiuojant nepašalinkite `na.rm`. Rezultatas kurį turėtumėte gauti:

```
##                        Group.1         x
## 1                        15-24 21.395238
## 2                        15-29 15.628571
## 3                        15-64 11.414286
## 4                        15-74 11.242857
## 5                        20-64 11.180952
## 6                        25-54 10.533333
## 7                        55-64  9.204762
## 8                          65+        NA
## 9 Iš viso pagal amžiaus grupes 11.233333
```

Atgaminkite žemiau pateiktą grafiką. Naudokites internetu, jeigu reikia pgalbos formatuojant grafiką, pvz., https://stackoverflow.com/questions/14860078/plot-multiple-lines-data-series-each-with-unique-color-in-r



3 Funkcijų rašymas

Parsisiųskite zip failą pavadinimu https://github.com/justasmundeikis/duomenu_analizes_ivadas/raw/master/seminars/3_seminaras.zip. Išpakuokite turinį taip, jog visi failai esantys zip archyve būtų išpakuoti į folderį "3_seminaras". Šis folderis tarnaus Jums kaip darbinė R direktorija. Startuokite R ir pasikeiskite R darbinę direktoriją į šį folderį. Pasitikrinkite, ar R mato visus tris failus su komanda `dir()` jeigu reikia, keiskite direktoriją su `setwd("...")` komanda. Komanda `getwd()` pasako kur esate.

Šioje dalyje naudosimės *U.S. Department of Health and Human Services* surinktais duomenimis apie ligonines JAV, kuriais remiantis yra analizuojama JAV veikiančių ligoninių veikla. Šiam seminarui duomenų apimtis yra sumažinta ir direktorijoje Jūs rasite 3 failus:

- `outcome-of-care-measures.csv` kuriame yra 30 dienų mirtingumo rodikliai surinkti virš 4000 ligoninių JAV.
- `hospital-data.csv` kuriame surinkta bendrinė informacija apie kiekvieną ligoninę
- `Hospital_Revised_Flatfiles.pdf` yra duomenų `code book`, kurioje aprašomi visi failai ir juose esantys kintamieji, jų kodavimas (zip faile nėra visų duomenų, o tik 11 ir 19 failai).

3.1 Funkcija `best`

Parašykite funkciją `best`, kuri priima du argumentus: `state` - 2 ženklų ilgumo valstijos kodą, ir `outcome` ligos pavadinimą, nuskaito `outcome-of-care-measures.csv` ir pateikia character vektorių su ligoninės pavadinimu, kuriame minimos ligos mirtingumo rodiklis yra geriausias (žemiausias) pasirinkotje valstijoje. Ligoninių pavadinimai yra stulpelyje: `Hospital.Name`. Galimos ligos yra:

- `heart attack`
- `heart failure`
- `pneumonia`

Vertinant ligonines, tos ligoninės, kurios neturi norimos ligos statistikos, turėtų būti pašalintos iš reitingavimo proceso. Jeigu dvi ligoninės turi identiškai gerą statistiką pateiktai ligai, tada funkcija turi grąžinti alfabetiškai pirmą ligoninę. Parašytą funkciją (script) išsaugokite kaip failą "best.R" Rašant šią funkciją Jums gali prireikti R funkcijos `order()`. Susipažinkite su ja savarankiškai, naudokitės internete esančiais resursais.

Funkcijos prototipas:

```
best <- function(state, outcome){
        ## nuskaityti duomenis iš .csv failo

        ## patikrinti ar valstijos ir ligos pavadinimas yra teisingas, jeigu ne,
        ## gražinti pranešimą apie klaidingą įvedimą,
        ## pvz., "invalide state" arba atitinkamai "invalide outcome"

        ## grąžinti ligoninės pavadinimą pasirinktoje valstijoje su
        ## žemiausiu 30 dienų mirtingumo rodikliu pasirinktai ligai
}
```

Pastaba: norint importuoti failą "best.R" naudoti funkciją `source(best.R)` PVZ: ką turėtų grąžinti funkcija `best.R`

```
source("best.R")
best("TX", "heart attack")
## [1] "CYPRESS FAIRBANKS MEDICAL CENTER"
best("TX", "heart failure")
## [1] "FORT DUNCAN MEDICAL CENTER"
best("MD", "heart attack")
## [1] "JOHNS HOPKINS HOSPITAL, THE"
best("MD", "pneumonia")
## [1] "GREATER BALTIMORE MEDICAL CENTER"
best("BB", "heart attack")
## Error in best("BB", "heart attack"): invalide state
best("NY", "hert attack")
## Error in best("NY", "hert attack"): invalide outcome
```

3.2 Funkcija `rankhospital`

Parašykite funkciją `rankhospital`, kuri priima tris argumentus: `state` - 2 ženklų ilgumo valstijos kodą, `outcome` ligos pavadinimą bei `num`, kuris gali būti arba "best", arba "worst" arba skaitinis reitingo vietos indekso numeris. Funkcija turi grąžinti ligoninės pavadinimą pasirinktoje valstijoje su pasirinktu 30 dienų mirtingumo reitingo rodikliu pasirinktai ligai.

Galimos ligos yra:

- `heart attack`
- `heart failure`
- `pneumonia`

Vertinant ligonines, tos ligoninės, kurios neturi norimos ligos statistikos, turėtų būti pašalintos iš reitingavimo proceso. Jeigu nurodomas didesnis reitingo skaitinis numeris, funkcija turėtų grąžinti "NA". Jeigu dvi ligoninės turi identiškai gerą statistiką pateiktai ligai, tada funkcija turi grąžinti alfabetiškai pirmą ligoninę. Parašytą funkciją (script) išsaugokite kaip failą "rankhospital.R" Rašant šią funkciją Jums gali prireikti R funkcijos `order()`. Susipažinkite su ja savarankiškai, naudokitės internete esančiais resursais.

Funkcijos prototipas:

```
rankhospital <- function(state, outcome, num="best"){
        ## nuskaityti duomenis iš .csv failo

        ## patikrinti ar valstijos ir ligos pavadinimas yra teisingas, jeigu ne,
        ## gražinti pranešimą apie klaidingą įvedimą,
        ## pvz., "invalid state" arba atitinkamai "invalid outcome"

        ## grąžinti ligoninės pavadinimą pasirinktoje valstijoje su
        ## pasirinktu 30 dienų mirtingumo reitingo rodikliu pasirinktai ligai
}
```

Pastaba: norint importuoti failą "rankhospital.R" naudoti funkciją `source(rankhospital.R)` PVZ: ką turėtų grąžinti funkcija `rankhospital.R`

```
source("rankhospital.R")
rankhospital("TX", "heart failure", 4)
## [1] "DETAR HOSPITAL NAVARRO"
rankhospital("MD", "heart attack", "worst")
## [1] "HARFORD MEMORIAL HOSPITAL"
rankhospital("MN", "heart attack", 5000)
## [1] NA
```

3.3 Funkcija `rankall`

Parašykite funkciją `rankall`, kuri priima du argumentus: `outcome` - ligos pavadinimą bei `num`, kuris gali būti arba "best", arba "worst" arba skaitinis reitingo vietos indekso numeris. Funkcija turi grąžinti `dataframe`, kur pirmame stulpelyje sureitinguotas ligonines pagal 30 dienų mirtingumo rodiklius pasirinktai ligai. Dataframe turi sudaryti du stulpeliai `hospital` ir `state`.

Galimos ligos yra:

- `heart attack`
- `heart failure`
- `pneumonia`

Vertinant ligonines, tos ligoninės, kurios neturi norimos ligos statistikos, turėtų būti pašalintos iš reitingavimo proceso. Jeigu dvi ligoninės turi identiškai gerą statistiką pateiktai ligai, tada funkcija turi grąžinti

alfabetiškai pirmą ligoninę. Parašytą funkciją (script) išsaugokite kaip failą "rankall.R" Rašant šią funkciją Jums gali prireikti R funkcijos `order()` bei `rank()`. Susipažinkite su jomis savarankiškai, naudokitės internete esančiais resursais.

Funkcijos prototipas:

```
rankall <- function( outcome, num="best"){
        ## nuskaityti duomenis iš .csv failo

        ## patikrinti ar ligos pavadinimas yra teisingas, jeigu ne,
        ## grąžinti pranešimą apie klaidingą įvedimą, "invalid outcome"

        ## kiekvienai valstijai surasti ligoninės pavadinimą
        ## pagal pasirinkta reitingo numerį

        ## grąžinti ligoninės pavadinimą ir valstijos trumpinį dataframe objekte
}
```

Pastaba: norint importuoti failą "best.R" naudoti funkciją `source(rankall.R)` PVZ: ka turėtų grąžinti funkcija `rankall.R`

```
source("rankall.R")
head(rankall("heart attack", 20), 10)
##                                                Hospital.Name State
## 59                                D W MCMILLAN MEMORIAL HOSPITAL    AL
## 211                             ARKANSAS METHODIST MEDICAL CENTER    AR
## 154                           JOHN C LINCOLN DEER VALLEY HOSPITAL    AZ
## 564                                     SHERMAN OAKS HOSPITAL    CA
## 651                                  SKY RIDGE MEDICAL CENTER    CO
## 696                                   MIDSTATE MEDICAL CENTER    CT
## 808                                SOUTH FLORIDA BAPTIST HOSPITAL    FL
## 910                                 UPSON REGIONAL MEDICAL CENTER    GA
## 1412                                      COVENANT MEDICAL CENTER    IA
## 1107 JESSE BROWN VA MEDICAL CENTER - VA CHICAGO HEALTHCARE SYSTEM    IL
tail(rankall("pneumonia", "worst"), 3)
##                                Hospital.name State
## 52 MAYO CLINIC HEALTH SYSTEM - NORTHLAND, INC    WI
## 53                     PLATEAU MEDICAL CENTER    WV
## 54           NORTH BIG HORN HOSPITAL DISTRICT    WY
tail(rankall("heart failure"), 10)
##                                                          Hospital.Name
## 3797                         WELLMONT HAWKINS COUNTY MEMORIAL HOSPITAL
## 3935                                   FORT DUNCAN MEDICAL CENTER
## 4237 VA SALT LAKE CITY HEALTHCARE - GEORGE E. WAHLEN VA MEDICAL CENTER
## 4341                                     SENTARA POTOMAC HOSPITAL
## 4278                           GOV JUAN F LUIS HOSPITAL & MEDICAL CTR
## 4275                                       SPRINGFIELD HOSPITAL
## 4399                                    HARBORVIEW MEDICAL CENTER
## 4561                                 AURORA ST LUKES MEDICAL CENTER
## 4473                                      FAIRMONT GENERAL HOSPITAL
## 4644                                    CHEYENNE VA MEDICAL CENTER
##      State
## 3797    TN
## 3935    TX
## 4237    UT
## 4341    VA
```

```
## 4278    VI
## 4275    VT
## 4399    WA
## 4561    WI
## 4473    WV
## 4644    WY
```

3.4 GitHub

Sukurkite naują repozitoriją GitHube pavadinimu “3_seminaras” ir pushinkite direktorijos turinį į GitHubą naudodamiesi Git.