

Seminar 2

Justas Mundeikis

2019-04-20

Turinys

1	Apie seminarą	1
1.1	Seminaro tikslai	1
1.2	Seminaro aptarimas:	1
2	Nedarbo lygis Lietuvoje	1
3	Funkcijų rašymas	2
3.1	Funkcija <code>best</code>	3
3.2	Funkcija <code>rankhospital</code>	4
3.3	Funkcija <code>rankall</code>	4
3.4	GitHub	5

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:

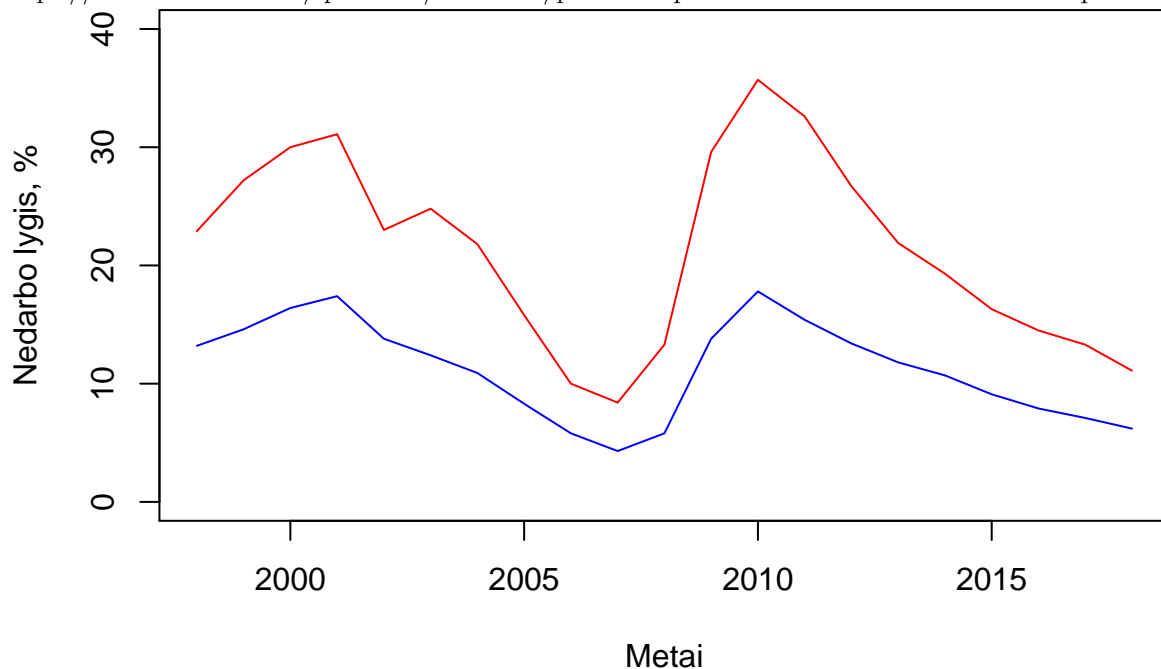
```
str(data)
## 'data.frame':    1701 obs. of  7 variables:
## $ Laikotarpis      : int  1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 ...
## $ Rodiklis         : chr  "Nedarbo lygis" "Nedarbo lygis" "Nedarbo lygis" "Nedarbo lygis"
## $ Amžius..tikslinės.amžiaus.grupės.: chr  "Iš viso pagal amžiaus grupes" "Iš viso pagal amžiaus grupes"
## $ 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 gali 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 pagalbos 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) pasirinktoje 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 identiška 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,
  ## grąžinti pranešimą apie klaidingą įvedimą,
  ## pvz., "invalid state" arba atitinkamai "invalid 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"): invalid state
best("NY", "heart attack")
## Error in best("NY", "heart attack"): invalid 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 identišškai 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,  
  ## grąžinti 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 identišškai 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)  
tail(rankall("pneumonia", "worst"), 3)  
tail(rankall("heart failure"), 10)
```

3.4 GitHub

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