OBRADA PODATAKA

Hrvatski studiji

dr.sc. Luka Šikić

Diplomski studij sociologije

13 listopad, 2020

```
## -- Attaching packages ---------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Warning in normalizePath(path.expand(path), winslash, mu
## \Users\Lukas\anaconda3\envs\rstudio/python.exe": The sys
## specified
```

CILJEVI PREDAVANJA

- Strukturirani i ne-strukturirani podatci
- Kvantitativni i kvalitativni podatci
- Podatci u “praksi”
- Big Data

STRUKTURIRANI I NESTRUKTURIRANI PODATCI

1. **STRUKTURIRANI**

- observacije sa karakteristikama, uglavnom organizirane u tablicu (redovi i kolone)
- znanstveno prikupljeni podatci, telefonski imenik
- manji dio podataka

2. **NESTRUKTURIRANI**

- podatci bez standardne organizacijske hijerarhije
- Facebook objave, Twitter, logovi na server, genetska sekvenca nukleotida, tekstualni podatci
- vjerojatno više od 80% svih podataka
- zahtjevaju prilagodbu prije analize

KVANTITATIVNI i KVALITATIVNI PODATCI

1. **KVANTITATIVNI**

- brojevi, matematičke procedure, prosjek, vremenski trend, threshold efekti

2. **KVALITATIVNI**

- "prirodne" kategorije, jezik
- najčešća observacija, jedinstvene vrijednosti

PRIMJER

```
##                 country  beer_servings  ...  total_litres_
## 0           Afghanistan              0  ...
## 1               Albania             89  ...
## 2               Algeria             25  ...
## 3               Andorra            245  ...
## 4                Angola            217  ...
## 5  Antigua & Barbuda              102  ...
## 6             Argentina            193  ...
## 7               Armenia             21  ...
## 8             Australia            261  ...
## 9               Austria            279  ...
##
## [10 rows x 6 columns]

## count      170
```

DISKRETNI I KONTINUIRANI PODATCI

1. **DISKRETNI**

- prebrojivi
- npr. igraća kocka

1. **KONTINUIRANI**

- postoje na kontinuiranoj skali
- npr. težina ili visina

ČETIRI RAZINE PODATAKA

1. NOMINALNI
2. ORDINALNI
3. INTERVALNI
4. OMJERNI

NOMINALNA RAZINA

- podatci opisani nazivom ili kategorijom (kategorički podatci)
- npr. spol, nacionalnost, biološke vrste
- ne mogu se obavljati matematičke operacije poput zbrajanja ili djeljenja
- računanje prosjeka ili drugih statističkih momenata nema smisla

ODINALNA RAZINA

- kategorički podatci koji imaju hijerarhijsku strukturu
- iako postoji hijerarhija, nije moguće utvrditi relativne razlike među opservacijama
- matematičke operacije kao zbrajanje ili dijeljenje nisu opravdane
- usporedbe i sortiranje podataka su opravdane
- moguće je koristiti medijan (ne i prosjek)

INTERVALNA RAZINA

- npr. temperatura
- opravdane su matematičke operacije poput zbrajanja i oduzimanja
- opravdanjo korištenje mjera centralne tendencije i varijabilnosti

OMJERNA RAZINA

- opravdane matematičke operacije množenja i dieljenja
- podatci na ovoj razni ne smiju biti negativni