Advanced R

Day 3

Sereina Herzog

Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz

06.03.2024

Course Content - Advanced R (Day 3)

- Data preparation

Course Content - Advanced R (Day 3)

- Data preparation
- Assignment

Data preparation

Import Data

Data cleaning with tidyverse

Data table

- each **unit** (e.g. patient, mouse, cell) equals a row
- for each unit the measured **variables** (e.g. age, blood pressure, size) equal columns

Data table

- each **unit** (e.g. patient, mouse, cell) equals a row
- for each unit the measured **variables** (e.g. age, blood pressure, size) equal columns

id	gender	age	weight	height	smoking
1	1	35	70.5	185	0
2	2	36	65.3	170	0
3	2		90.1	164	1
4	1	21	72.0	177	0
5	1	66	89.4	175	0

Repeated measurements

wide format

id	gender	syst0	syst1
1	1	120	125
2	2	118	125
3	2		110

Repeated measurements

wide format

id	gender	syst0	syst1
1	1	120	125
2	2	118	125
3	2		110

long format

id	gender	syst	time
1	1	120	0
1	1	125	61
2	2	118	0
2	2	125	60
3	2		
3	2	110	59

What is tidyverse

- tidyverse is a collection of R packages designed for data science

What is tidyverse

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure

What is tidyverse

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization

What is tidyverse

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources

What is tidyverse

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources
    - *tidyr* and *dplyr* useful for data cleaning

What is tidyverse

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources
    - *tidyr* and *dplyr* useful for data cleaning
    - ...
  - all core packages can be loaded at once: *library(tidyverse)*

What is tidyverse

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources
    - *tidyr* and *dplyr* useful for data cleaning
    - ...
  - all core packages can be loaded at once: *library(tidyverse)*
  - 'R for Data Science' (see slide with links)

Useful functions for data cleaning

- **select()** extracts columns and returns a tibble

Useful functions for data cleaning

- **select()** extracts columns and returns a tibble
- **arrange()** changes the ordering of the rows

Useful functions for data cleaning

- **select()** extracts columns and returns a tibble
- **arrange()** changes the ordering of the rows
- **filter()** picks cases based on their values

Useful functions for data cleaning

- **select()** extracts columns and returns a tibble
- **arrange()** changes the ordering of the rows
- **filter()** picks cases based on their values
- **mutate()** adds new variables that are functions of existing variables

What is %>% in Tidyverse?

%>% is used to emphasis a sequence of actions, rather than the object that the actions are being performed on

What is %>% in Tidyverse?

%>% is used to emphasis a sequence of actions, rather than the object that the actions are being performed on

```
dt_example %>%
  mutate(bmi = weight/(height^2)) %>%
  select(pat_id, sex, bmi)
```

What will we cover

- We will look at
  - importing data (example: .xlsx)
  - useful function for data cleaning
  - save R environment (.Rdata)
- We will work with .Rdata in a Rmarkdown file

Data cleaning - exercise

- Example Glucose:
  - Glucose tolerance was tested by administering 100g glucose drink
  - Glucose was tested before and 1 hour after administering
  - source: R package medicaldata

Data cleaning - exercise

- Example Glucose:
  - Glucose tolerance was tested by administering 100g glucose drink
  - Glucose was tested before and 1 hour after administering
  - source: R package medicaldata
- Download from GitHub (Course Introduction R 2023/Day2')
  - messy_glucose.xlsx
  - day2_ex2_datacleaning_v20231109.R

Data cleaning - exercise

- Example Glucose:
  - Glucose tolerance was tested by administering 100g glucose drink
  - Glucose was tested before and 1 hour after administering
  - source: R package medicaldata
- Download from GitHub (Course Introduction R 2023/Day2')
  - messy_glucose.xlsx
  - day2_ex2_datacleaning_v20231109.R
- Open R file
- Work through R file (together)