

R and RStudio

Didzis Elferts



Interreg
Estonia-Latvia
European Regional Development Fund



EUROPEAN UNION

WaterAct


Joint actions for more efficient management
of common groundwater resources



About me

- Didzis Elferts
- University of Latvia, Faculty of Biology, professor
- Using R since 2008
- Have taught R for ~1000 students, researchers and others

 didzis.elferts@lu.lv

 <https://www.delferts.lv/en>

 @delferts

 @delferts

 0000-0002-9401-1231

Objectives for the training

- Get to know R and RStudio
- Learn to import different types of data, connect to databases
- Learn to summarise and transform data
- Learn the R package ggplot2 for the visualization
- Learn how to calculate statistics and do basic statistical tests in R

Schedule

August 23 - Introduction to R, data import

August 24 - Visualisation with ggplot2

August 26 - Tidyverse

August 30 - Dynamic documents, statistical tests

August 31 - Your ideas

Materials

All course materials (presentations, code files, data files) are available at:

https://ej.uz/r_training

Example session

Create MS Excel file with two columns - **Height** and **Weight**, fill those columns with some data (5-10 rows). Save the file with the name **Your_name.xlsx** in the directory *My documents/Documents*.

Example session

```
library(readxl)
dati <- read_excel("Didzis.xlsx")
dati
```

```
## # A tibble: 6 × 2
##   Height Weight
##   <dbl> <dbl>
## 1    175     70
## 2    180     86
## 3    165     61
## 4    163     69
## 5    172     72
## 6    169     68
```

Example session

Correlation analysis

```
cor.test(dati$Height, dati$Weight)
```

```
##  
##      Pearson's product-moment correlation  
##  
## data:  dati$Height and dati$Weight  
## t = 2.8434, df = 4, p-value = 0.04671  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
##  0.01894518 0.97937919  
## sample estimates:  
##          cor  
## 0.8179306
```


Example session

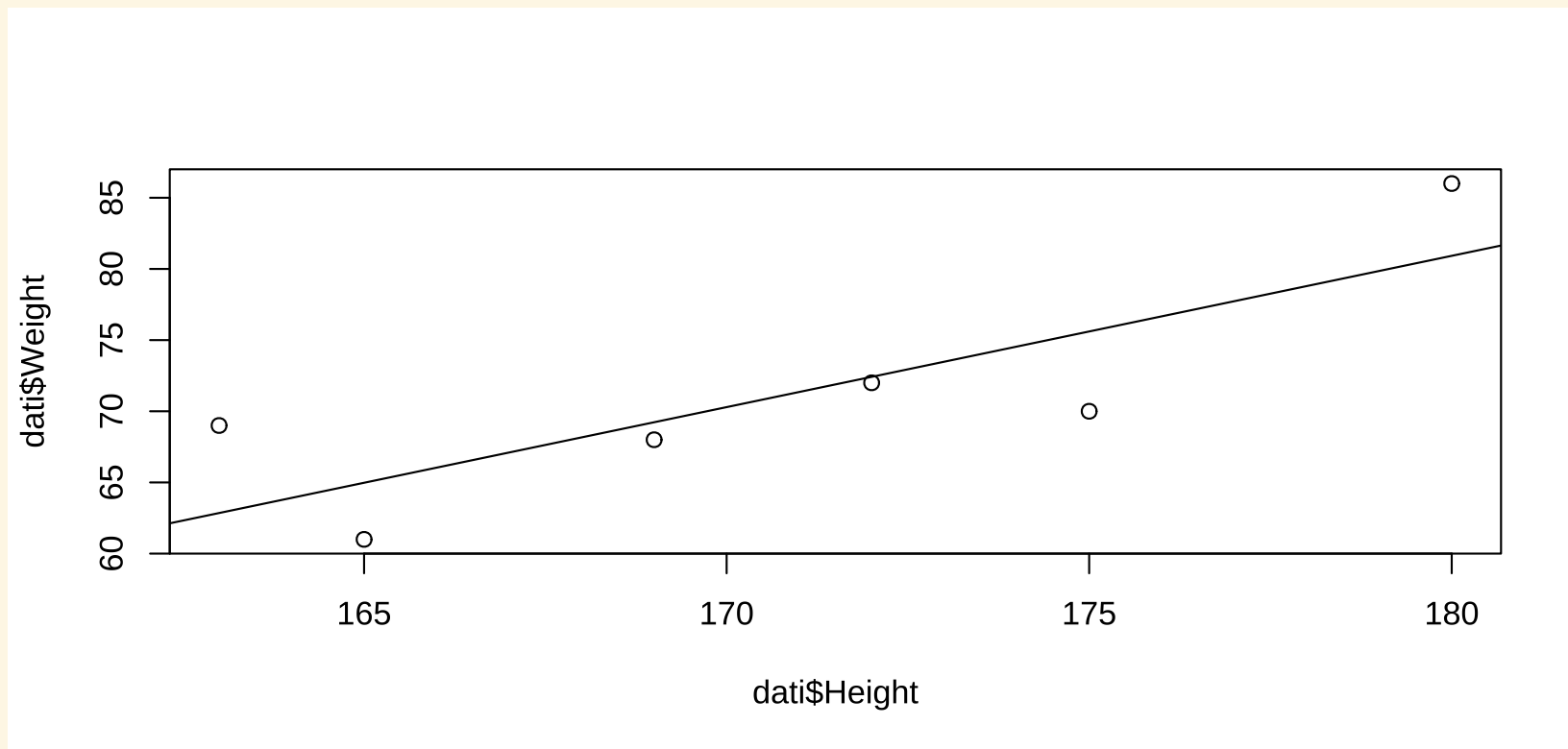
Linear regression

```
summary(lm(Weight ~ Height, data = dati))
```

```
##  
## Call:  
## lm(formula = Weight ~ Height, data = dati)  
##  
## Residuals:  
##      1      2      3      4      5      6  
## -5.6060  5.0795 -3.9768  6.1490 -0.4172 -1.2285  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -110.4040    63.8348  -1.730   0.1588  
## Height      1.0629     0.3738   2.843   0.0467 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 5.304 on 4 degrees of freedom  
## Multiple R-squared:  0.669,    Adjusted R-squared:  0.5863  
## F-statistic: 8.085 on 1 and 4 DF,  p-value: 0.04671
```

Example session

```
plot(dati$Weight ~ dati$Height)  
abline(lm(Weight ~ Height, data = dati))
```



Pros and cons of R

Pros:

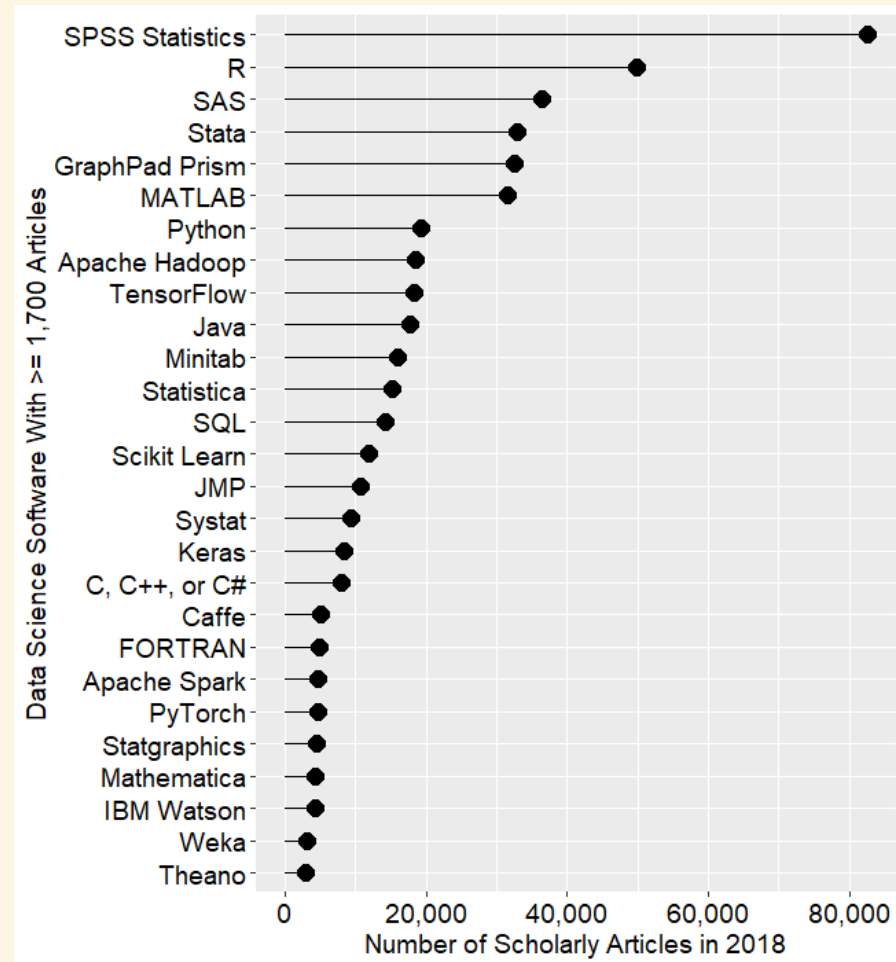
- Open-source, regularly updated, still developing program;
- Works on main platforms: Windows, MacOS, Linux
- Different statistical methods implemented, control over parameters
- Excellent graphical capabilities
- Own functions and R packages
- Development of interactive visualizations, web-applications
- Ideal to implement reproducible research

Pros and cons of R

Cons:

- Slow "learning" pace
- Partly - comandline program
- Sometimes hard to find necessary information/package/function

R popularity



Source: <http://r4stats.com/2019/04/01/scholarly-datasci-popularity-2019/>

R popularity

The TIOBE Programming Community index

Programming Language	2021	2016	2011	2006	2001	1996	1991	1986
C	1	2	2	2	1	1	1	1
Java	2	1	1	1	3	30	-	-
Python	3	5	7	8	27	18	-	-
C++	4	3	3	3	2	2	2	8
C#	5	4	5	7	11	-	-	-
Visual Basic	6	13	-	-	-	-	-	-
JavaScript	7	8	10	10	9	33	-	-
PHP	8	6	4	4	19	-	-	-
R	9	17	33	-	-	-	-	-
SQL	10	-	-	-	-	-	-	-
Lisp	34	27	13	14	16	7	4	2
Ada	36	26	19	16	22	8	8	3
(Visual) Basic	-	-	6	6	4	3	3	4

Avots: <https://www.tiobe.com/tiobe-index/>

Usefull links

- R program homepage <http://www.r-project.org/>
- RStudio homepage <http://www.rstudio.com>
- YouTube channel with tutorials <https://www.youtube.com/playlist?list=PLcgz5kNZFCkzSyBG3H-rUaPHoBXgijHfC>
- Q&A page Stack overflow <http://stackoverflow.com/>
- Search in R packages <http://www.rdocumentation.org/>

R packages

- Base R - only small part of statistical analysis, base graphics
- Additional capabilities through R packages (libraries) that must be installed (if not already done, only once) and then added to the session (to be done in each session)
- Developed by users, hosted on CRAN (official), github or internally
- CRAN packages: 18031 (16.08.2021.)
- R packages are installed with function `install.packages()` and added to the session with function `library()`.

```
install.packages("cplm")  
library(cplm)
```


R reference

With the function `citation()` you can get reference for the version of R you are using.

`citation()`

```
##
## To cite R in publications use:
##
## R Core Team (2021). R: A language and environment for statistical
## computing. R Foundation for Statistical Computing, Vienna, Austria.
## URL https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
## @Manual{,
##   title = {R: A Language and Environment for Statistical Computing},
##   author = {{R Core Team}},
##   organization = {R Foundation for Statistical Computing},
##   address = {Vienna, Austria},
##   year = {2021},
##   url = {https://www.R-project.org/},
## }
##
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

R reference

Adding of the package name to the function `citation()` gives reference for that package.

```
citation("readxl")
```

```
##  
## To cite package 'readxl' in publications use:  
##  
## Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R  
## package version 1.3.1. https://CRAN.R-project.org/package=readxl  
##  
## A BibTeX entry for LaTeX users is  
##  
## @Manual{,  
##   title = {readxl: Read Excel Files},  
##   author = {Hadley Wickham and Jennifer Bryan},  
##   year = {2019},  
##   note = {R package version 1.3.1},  
##   url = {https://CRAN.R-project.org/package=readxl},  
## }
```

Thinks to remember

- To use more than one processor core, additional packages and functions needed. Can be extended to use computer clusters, cloud computing, GPU computing
- R stores all data used for calculations in RAM

Data sources

- Most data formats supported (mainly additional R packages needed) - txt, cvs, xlsx, sav, json, NetCDF, ...
- Direct download from the webpages
- Connection to databases (also to password protected), data filtering - can use SQL commands or R language
- Data from loggers - if there is no function and data have the same pattern, we can make it!

RStudio

- RStudio is a company developing free and open tools for R, as well as, enterprise-ready professional products
- Software: RStudio IDE, RStudio Server, Shiny Server
- Cloud: RStudio Cloud, shinyapps.io
- R packages: tidyverse, ggplot2, dplyr, tidyr, purrr, stringr, shiny, rmarkdown, flexdashboard, sparklyr, tidymodels, reticulate, plumber,

Source: <http://www.rstudio.com>

Work with R

R commands

- To add a comment to the command line, you must type “#” before you want to type it
- Spacing in commands is usually ignored, the exception is when writing “< -”
- If the command is too long, you can simply split it with Enter key
- The missing values in the R are indicated by NA (may be different in the data but must be specified at the time of import)

Data types

- Numeric, integer, double - 1, 5, 2.6, -123.45
- Character - AA, green, plant
- Logical - TRUE and FALSE
- Factor

Data structures

- Vector
- Matrix
- List
- Data frame
- other

Making R objects

- If you need to save the results of an action for future actions, you must create an object (name < - action, or action - > object)
- Object name cannot start with a number, cannot have spaces (or must be placed in apostrophes)
- Be careful about using special symbols (letters) because the encodings on different computers/systems may not match

Obtaining help

```
help.start()
```

```
help(plot)
```

```
args(cor)
```

```
## function (x, y = NULL, use = "everything", method = c("pearson",  
##           "kendall", "spearman"))  
## NULL
```

```
example(plot)
```

Homepage:

<http://www.rdocumentation.org/>

Questions?