Data Visualization: Introduction to R

Parisa Niloofar

University of Southern Denmark

Fall 2022

SDU University of Southern Denmark

Overview

SDU University of Southern Denmark

R & Github

Installations

Have the followings installed:

- Git
- R and RStudio

Create a Github account.

Next Steps

1. Create an R Project
2. Create a new Repository in Github
3. Copy the related command line to the R Project's Terminal

Youtube video

SDU
University of Southern Denmark

A very Short Introduction to R

R as a Calculator

Practice1 Compute the difference between 2022 and the year you started at this university and divide this by the difference between 2022 and the year you were born. Multiply this with 100 to get the percentage of your life you have spent at this university. Use brackets if you need them.

R as a Calculator

Practice2 Repeat Practice1, but with several steps in between. You can give the variables any name you want, but the name has to start with a letter.

Functions

Practice3 Compute the sum of 4, 5, 8 and 11 by first combining them into a vector and then using the function `sum`.

Practice4 Generate 1000 samples from the N(mean=10, sd=1.4)

SDU
University of Southern Denmark

Plots

Practice5 Plot what you generated in Practice 4.

Help and Documentation

Practice6 Find help for the `sqrt` function.

Scripts

Practice7 Make a file called `firstscript.R` containing Rcode that generates 100 random numbers and plots them, and run this script several times (you can also use `source("firstscript.R")` in the console command).

Data Structures

Practice8 Put the numbers 31 to 60 in a vector named P and in a matrix with 6 rows and 5 columns named Q. Tip: use the function `seq`. Look at the different ways scalars, vectors and matrices are denoted in the workspace window.

SDU
University of Southern Denmark

Data Structures

Practice9 Make a script file which constructs three random normal vectors of length 100. Call these vectors x1, x2 and x3. Make a data frame called t with three columns (called a, b and c) containing respectively x1, x1+x2 and x1+x2+x3. Call the following functions for this data frame: `plot(t)` and `sd(t)`. Can you understand the results? Rerun this script a few times.

SDU University of Southern Denmark

Graphics

Practice10 Add the following lines to the script file of the Practice9. Try to find out, either by experimenting or by using the help, what the meaning is of `rgb`, the last argument of `rgb`, `lwd`, `pch`, `cex`.

```
plot(t$a, type="l", ylim=range(t),
lwd=3, col=rgb(1,0,0,0.3))
lines(t$b, type="s", lwd=2,
col=rgb(0.3,0.4,0.3,0.9))
points(t$c, pch=20, cex=4,
col=rgb(0,0,1,0.3))
```

SDU University of Southern Denmark

Dates

Practice11 Make a graph with on the x-axis: today, Christmas 2022 and your next birthday and on the y-axis the number of presents you expect on each of these days. Tip: make two vectors first.

SDU
University of Southern Denmark

Data Preparation

Setup

```
 pkgs <- c("ggplot2", "dplyr", "tidyr",
"mosaicData", "carData",
"VIM", "scales", "treemapify",
"gapminder", "ggmap", "choroplethr",
"choroplethrMaps", "CGPfunctions",
"ggcorrplot", "visreg",
"gcookbook", "forcats",
"survival", "survminer",
"ggalluvial", "ggridges",
"GGally", "superheat",
"waterfalls", "factoextra",
"networkD3", "ggthemes",
"hrbrthemes", "ggpol",
"ggbeeswarm")
install.packages(pkgs)
```

SDU University of Southern Denmark

Text and Excel Spreadsheets Files

Practice12 Import data from a comma delimited file: `"salaries.csv"`

Practice13 Import data from a tab delimited file: `"salaries.txt"`

Practice14 Import data from an Excel workbook: `"salaries.xlsx"`

SDU University of Southern Denmark

A Guide

Package	Function	Use
dplyr	select	select variables/columns
dplyr	filter	select observations/rows
dplyr	mutate	transform or recode variables
dplyr	summarize	summarize data
dplyr	group_by	identify subgroups for further processing
tidyr	gather	convert wide format dataset to long format
tidyr	spread	convert long format dataset to wide format

SDU
University of Southern Denmark

Selecting Variables

From the *starwars* data set:

- Practice15 Keep only the variables: *name*, *height*, and *gender*
- Practice16 keep the variables name and all variables between *mass* and *species* inclusive
- Practice17 keep all variables except *birth_year* and *gender*

Selecting Observations

From the *starwars* data set:

- Practice18 select females
- Practice19 select females that are from Alderaan
- Practice20 select individuals that are from Alderaan, Coruscant, or Endor

Creating/Recoding Variables

From the *starwars* data set:

- Practice21 convert height in centimeters to inches, and mass in kilograms to pounds
- Practice22 if height is greater than 180 then heightcat = "tall", otherwise heightcat = "short"
- Practice23 convert any eye color that is not black, blue or brown, to other
- Practice24 set heights greater than 200 or less than 75 to missing

SDU University of Southern Denmark

Summarizing Data

From the *starwars* data set:

- Practice25 calculate mean height and mass
- Practice26 calculate mean height and weight by gender

SDU University of Southern Denmark

Using Pipes

From the *starwars* data set:

Practice27 calculate the mean height for women by species using the pipe %>% operator

Reshaping Data

Practice28 Produce the following dataset and name it "Wide_data"

id	name	sex	age	income
01	Bill	Male	22	55000
02	Bob	Male	25	75000
03	Mary	Female	18	90000

and convert it to a long dataset like:

id	name	variable	value
01	Bill	sex	Male
02	Bob	sex	Male
03	Mary	sex	Female
01	Bill	age	22
02	Bob	age	25
03	Mary	age	18
01	Bill	income	55000
02	Bob	income	75000
03	Mary	income	90000

SDU University of Southern Denmark

Reshaping Data

Practice29 in Practice 28, convert the long dataset to the wide dataset.