# E0410 Fundamentals of Statistics for Scientific Data Using R

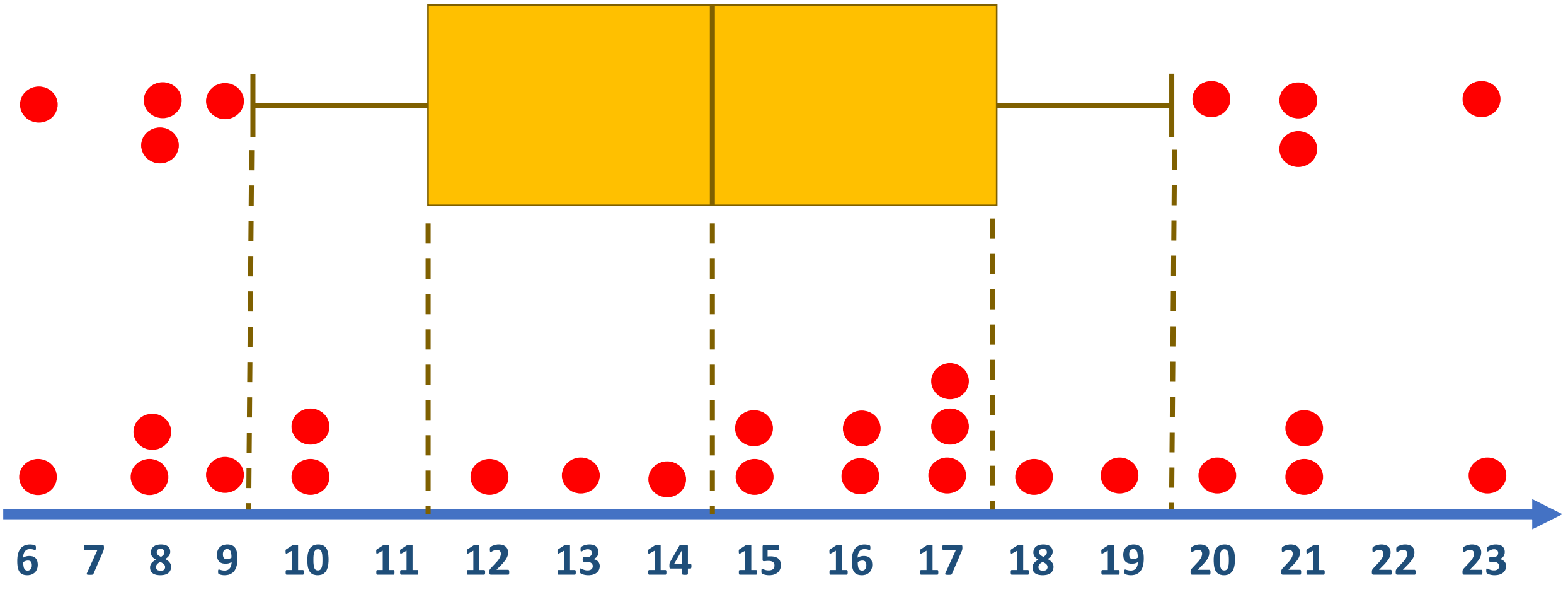
by Daria Sapunova, PhD student, RECETOX

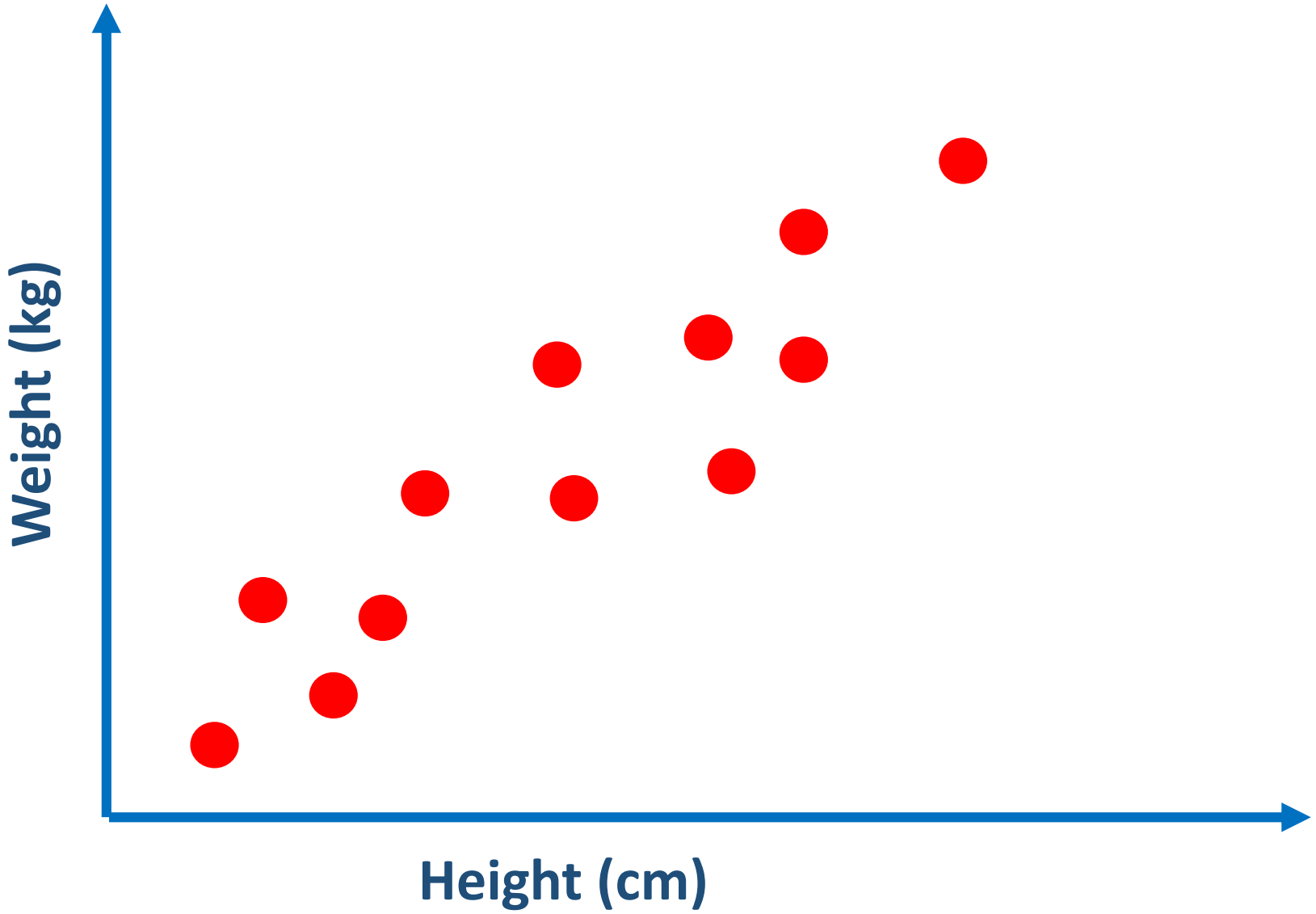
daria.sapunova@recetox.muni.cz

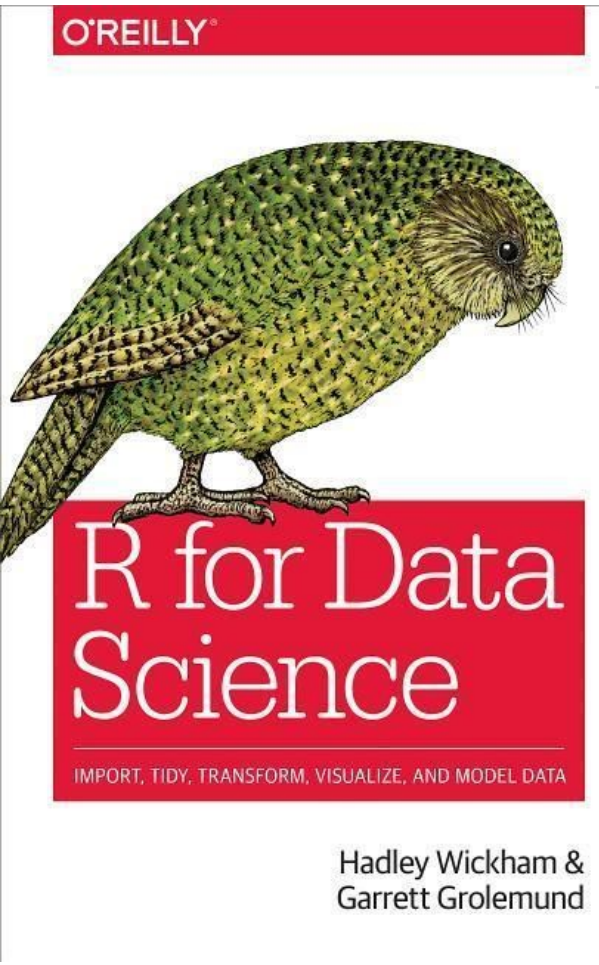
Bohunice, D29, room 123

# Theoretical part

The data is only for visual demonstration!




Weight (kg)
Height (cm)




r4ds.had.co.nz https://r4ds.had.co.nz/

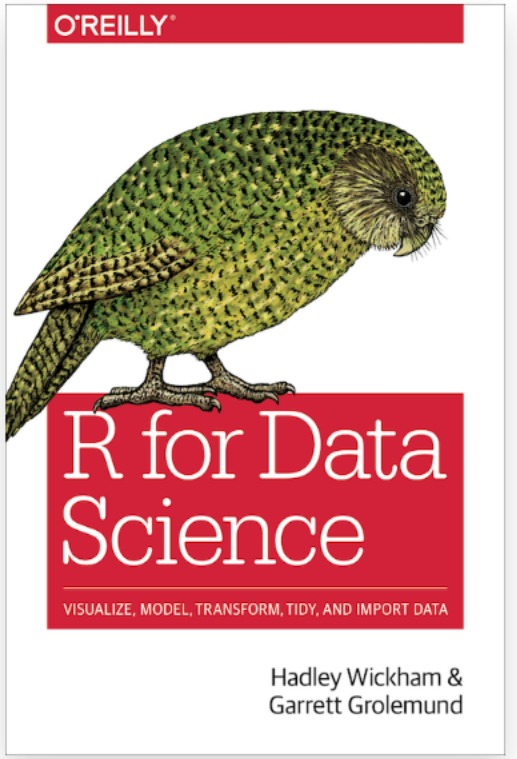
R for Data Science

Search

Table of contents

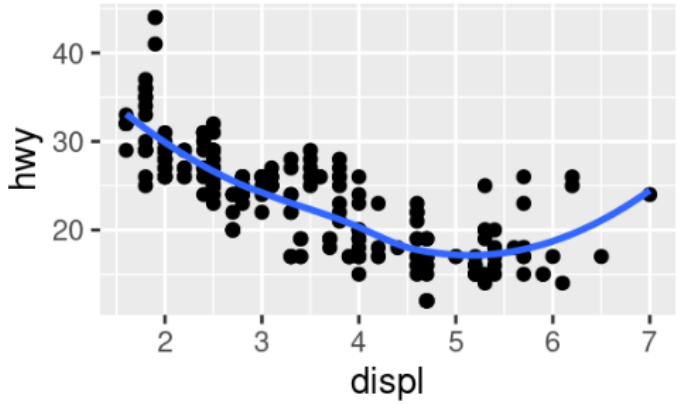


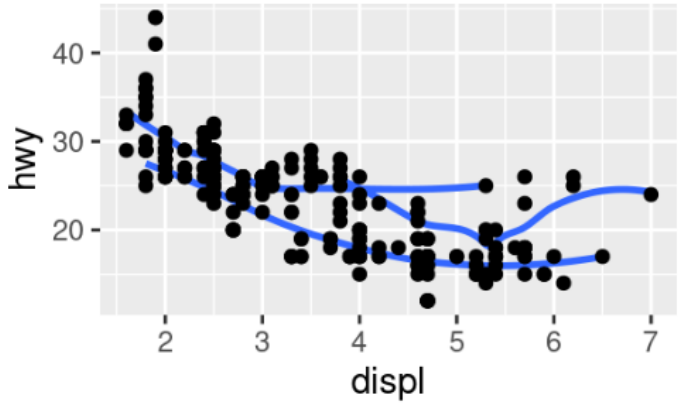
# Welcome

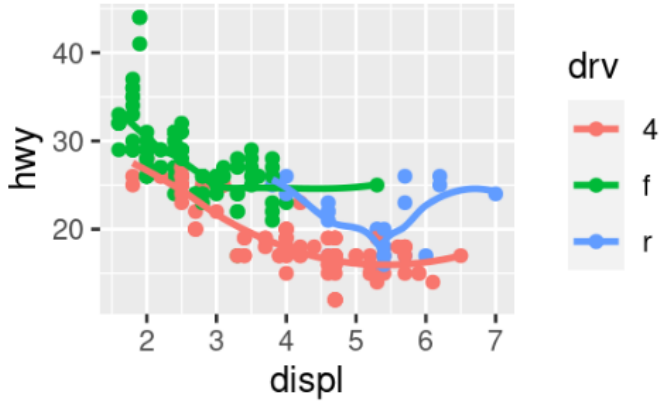


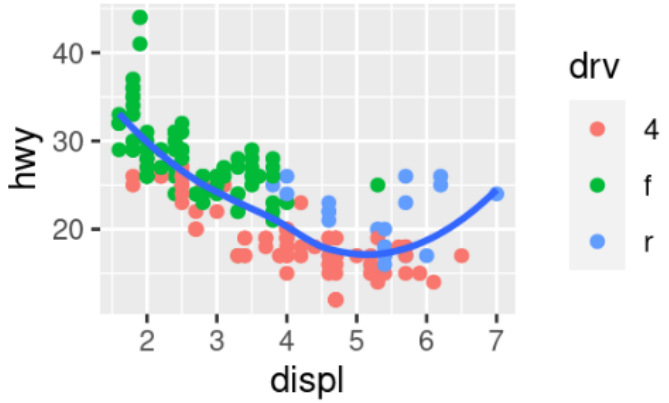

This is the website for **"R for Data Science"**. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

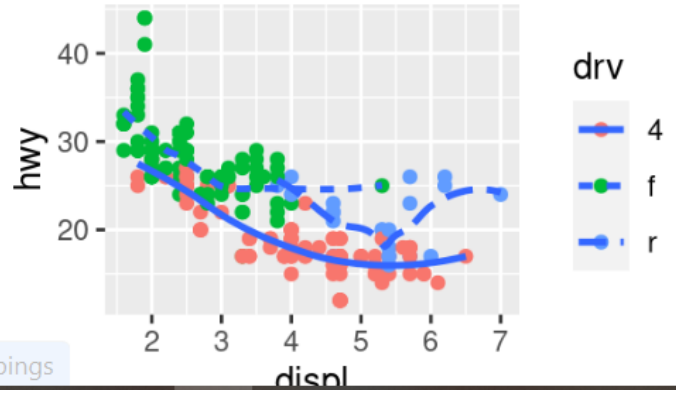
This website is (and will always be) **free to use**, and is licensed under the Creative

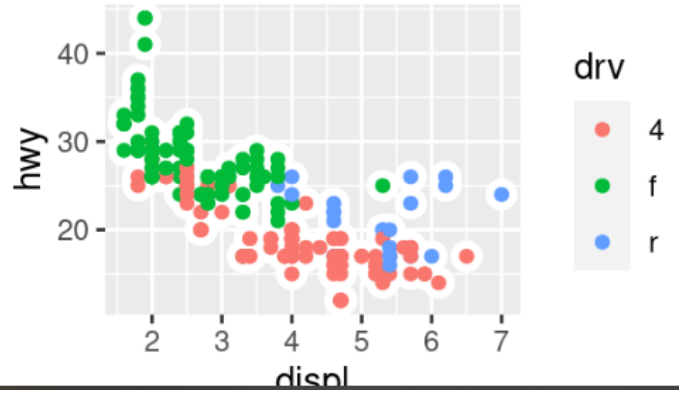

hwy
displ


hwy
displ

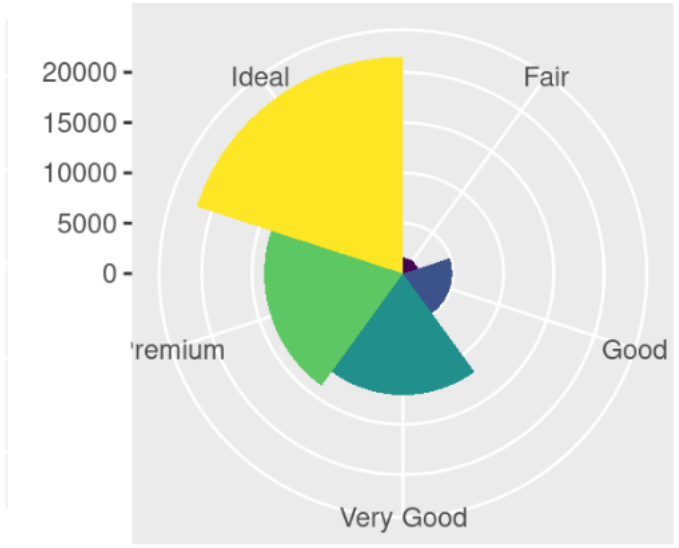

hwy
displ
drv
4
f
r

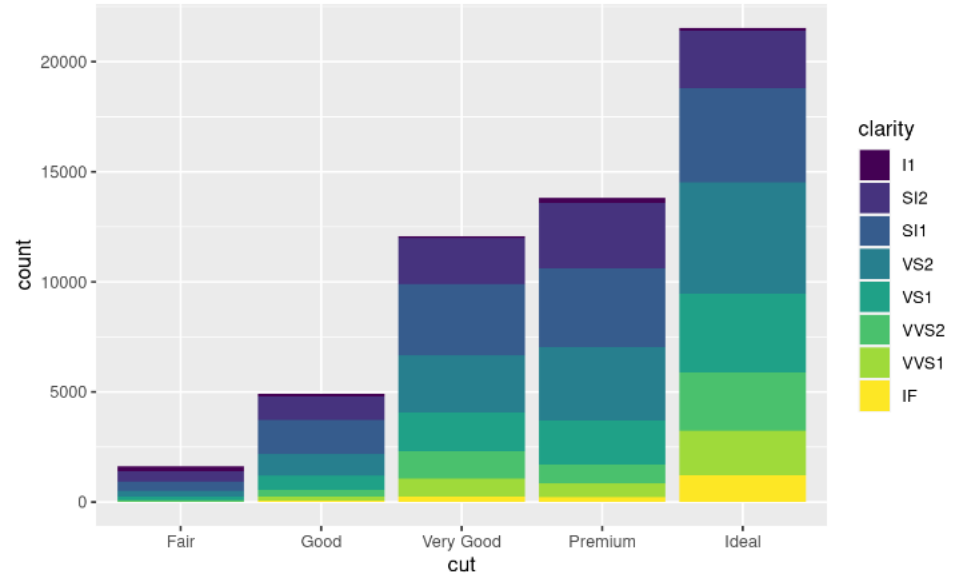

hwy
displ
drv
4
f
r

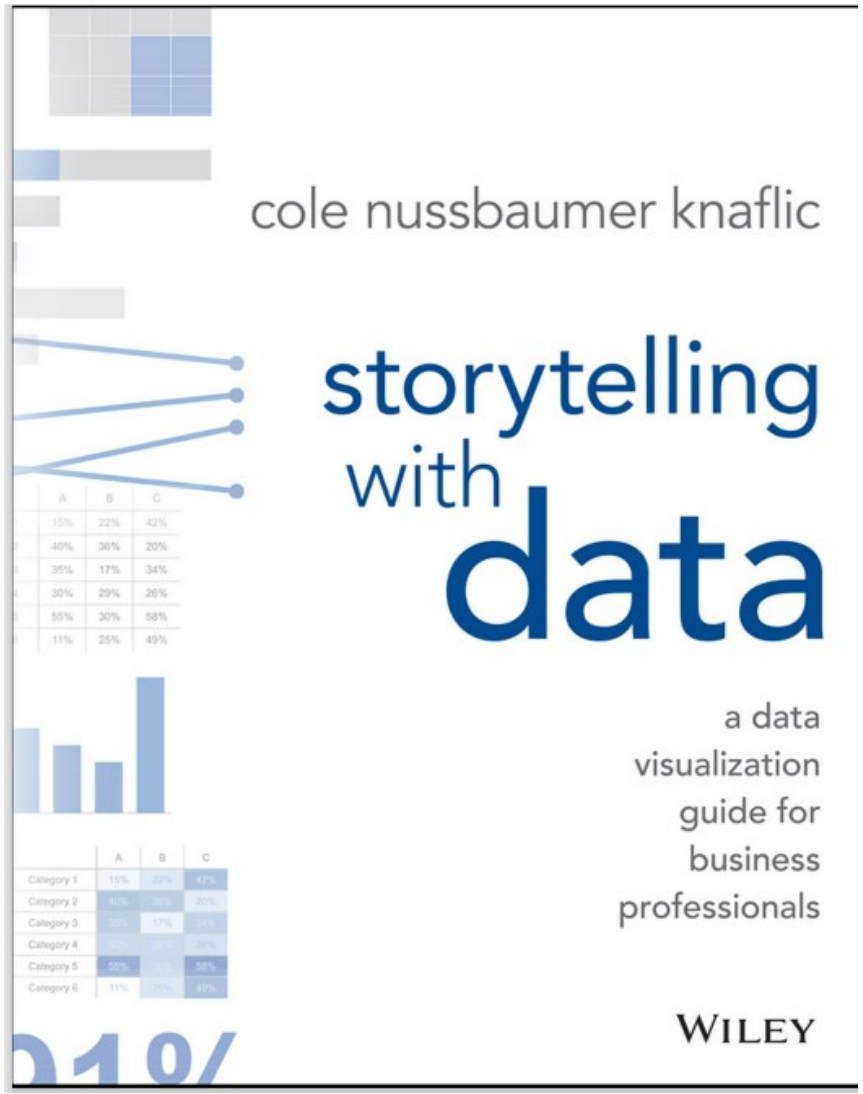

hwy
displ
drv
4
f
r

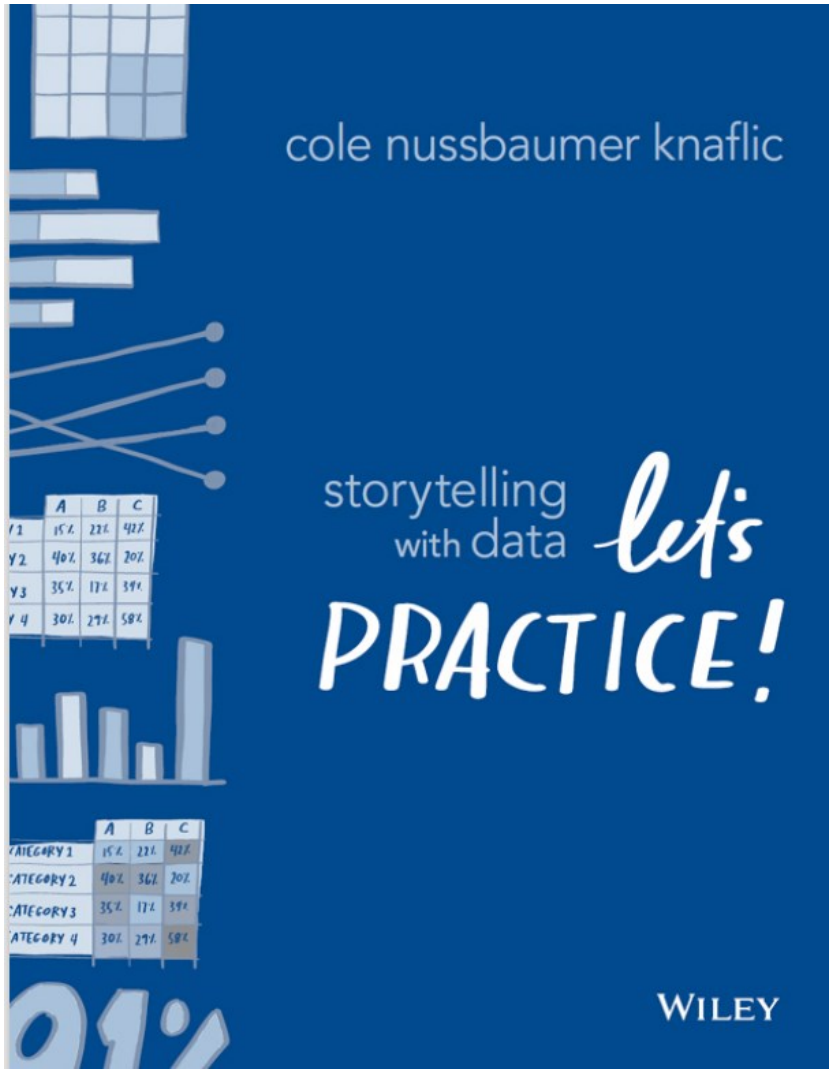

hwy
displ
drv
4
f
r

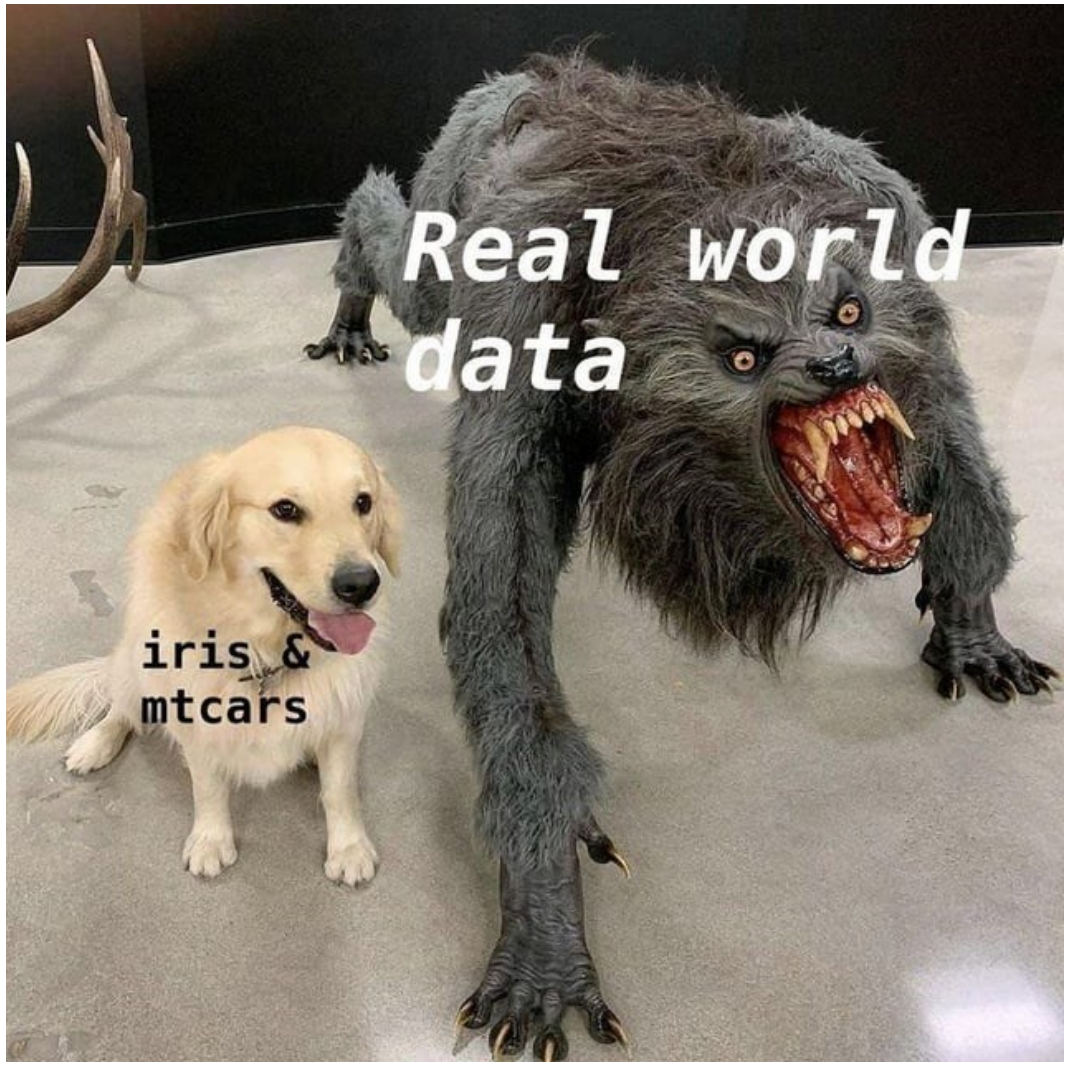
c-mappings


Ideal
Fair
Good
Very Good
remium


count
cut
Fair
Good
Very Good
Premium
Ideal
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF


cole nussbaumer knaflic
storytelling with data
a data visualization guide for business professionals
WILEY


cole nussbaumer knaflic
storytelling with data let's PRACTICE!
WILEY


Real world data
iris &
mtcars

# How your data <u>should not</u> look before importing into R

**1.**

| BPA | BPS | BPF |
|---|---|---|
| ug/l | ug/l | ug/l |
| 1.8 | 1.035294 | 0.247059 |
| 3.722222 | 0.477778 | 1.466667 |
| 1.796026 | 32.90066 | 0.033716 |
| 0.009428 | 0.133333 | 0.028284 |

| STANDARDIZATION C | |
|---|---|
| creat | creat |
| creat_BPA | creat_BPS |
| BPA | BPS |
| ug/g | ug/g |
| 0.869281 | 0.045752 |
| 0.937853 | 0.112994 |
| 0.827815 | 0.172185 |

**5.**

| new houses fraction <250 um | | |
|---|---|---|
| Mass | PBET | BA |
| ng/g | ng/g | % |
| <LOQ | <LOQ | - |
| 500 | <LOQ | - |
| 560 | 5.1 | 0.91 |
| 170 | <LOQ | - |
| 34 | <LOQ | - |

**2.**

| Microenvironment | Dust fraction <250 µm | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mass fraction (µg/g) | | Bioaccessible mass fraction (µg/g) | | Bioaccesible fraction (%) | | Mass fraction (µg/g) | | |
| | SCCPs | MCCPs | SCCPs | MCCPs | SCCPs | MCCPs | SCCPs | MCCPs | SCCP |
| New houses | 9.8 | 42 | 0.048 | 0.2 | 0.49 | 0.48 | 5.6 | 24 | 0.: |
| Prefab flats | 17 | 38 | 0.24 | 0.53 | 1.4 | 1.4 | 8.7 | 18 | 0. |
| Schools | 25 | 63 | 0.2 | 0.52 | 0.8 | 0.83 | 18 | 42 | 0.: |
| Kindergartens | 17 | 47 | 0.17 | 0.45 | 1.0 | 0.96 | 8.7 | 22 | 0.: |
| Offices | 17 | 28 | 0.2 | 0.39 | 1.2 | 1.4 | 9.4 | 16 | 0.: |
| Public spaces | 11 | 57 | 0.22 | 0.79 | 2.0 | 1.4 | 7.5 | 35 | 0.: |
| Cars | 6.8 | 27 | 0.031 | 0.1 | 0.46 | 0.37 | 6.6 | 23 | 0.: |

| Microenvironment | Dust fraction <2 mm | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Mass fraction (µg/g) | | Bioaccessible mass fraction (µg/g) | | Bioaccesible fraction (%) | | | |
| | SCCPs | MCCPs | SCCPs | MCCPs | SCCPs | MCCPs | bulk | finest |
| New houses | 5.6 | 24 | 0.19 | 0.55 | 3.4 | 2.3 | 29.6 | 51.8 |
| Old prefab houses | 8.7 | 18 | 0.3 | 0.65 | 3.4 | 3.6 | 26.7 | 55 |
| Schools | 18 | 42 | 0.36 | 0.89 | 2.0 | 2.1 | 60 | 88 |
| Kindergartens | 8.7 | 22 | 0.43 | 1.4 | 4.9 | 6.4 | 30.7 | 64 |
| Offices | 9.4 | 16 | 0.22 | 0.36 | 2.3 | 2.3 | 25.4 | 45 |
| Public spaces | 7.5 | 35 | 0.38 | 1.4 | 5.1 | 4.0 | 42.5 | 68 |
| Cars | 6.6 | 23 | 0.16 | 0.73 | 2.4 | 3.2 | 29.6 | 33.8 |

| | | | | |
|---|---|---|---|---|
| | | | 25.4 | 33.8 |
| | | | 60 | 88 |
| | SCCPs | MCCPs | | |
| SRM 1 | | | | |
| SRM 2 | | | | |
| SRM 3 | | | | |
| median | | | | |

**3.**

| factor ng/g - ng/ml: 1,03 | | |
|---|---|---|
| PFOS (ng/l) | PFOA (ng/l) | PFHxS (ng/l) |
| 11.8 | 14.0 | < 5.8 |
| < 6.4 | 12.5 | < 5.8 |
| 33.3 | 16.7 | < 5.8 |
| 45.9 | 17.5 | < 5.8 |
| 9.4 | 6.4 | < 5.8 |
| 25.3 | 17.1 | < 5.8 |
| 16.0 | 11.7 | < 5.8 |

**4.**

| Date | Notes |
|---|---|
| 2014-2015 | Western Scheldt |
| 2014-2015 | Atlantic Ocean |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |
| 2011-2012 | Nothern Italy |

# How your data <u>should</u> look before importing into R

| Food ca | Food - description | Chemical | units | n | Mean | Stdev | Food origin | Food origin sp | Productior | Date be | Date er |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A031G | Eggs | PFOS | ng/g fw | 3 | 52.8 | 30.9 | Belgium | Gavere, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 5.3 | 3.5 | Belgium | Mechelen, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 7 | 4.7 | Belgium | Grimbergen, Belgiu | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 386.1 | 307.9 | Belgium | Zwijndrecht, Belgiu | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 2.5 | 1 | Belgium | Kessel, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 0.4 | 0.5 | Belgium | Westmalle, Belgiur | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 4.5 | 5.3 | Belgium | Nijlen, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 13 | 9 | Belgium | Hoevene, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 9.2 | 4.6 | Belgium | Arendonk, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 3.1 | 1.8 | Belgium | Olmen, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 3.3 | 1.7 | Belgium | Olmen, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 3.4 | 1.8 | Belgium | Lille, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 109.9 | 59.2 | Belgium | Zwijndrecht, Belgiu | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 7.8 | 5.9 | Belgium | Edegem, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 3 | 2.9 | Belgium | Liedekerke, Belgiur | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 0.4 | 0.7 | Belgium | Meerhout, Belgium | Home raised eg | 2011 | 2011 |
| A031G | Eggs | PFOS | ng/g fw | 3 | 4.7 | 3.1 | Belgium | Oelegem, Belgium | Home raised eg | 2011 | 2011 |

# How your data should look before importing into R

| ID | Age | Gender | Height | Weight | Education | Freq.of fish consumption (times/month) | Source of fish (majority) | Chronical disease | Last med.exam. (years ago) |
|---|---|---|---|---|---|---|---|---|---|
| Person 1 | 45 | Male | 170.4 | 76.4 | PhD | 21 | Market | Diabetes | 10 |
| Person 2 | 26 | Male | 168.3 | 65.3 | BS | 5 | Self-fishing | NA | NA |
| Person 3 | 54 | Female | 168.6 | 75.3 | MS | 23 | Grocery store | Bladder infection | 10 |
| Person 4 | 65 | Male | 156.6 | 44.2 | MS | 10 | Market | Diabetes | 5 |
| Person 5 | 21 | Female | 170.1 | 69.9 | HS | 8 | Market | No | NA |
| Person 6 | 44 | Male | 176.4 | 84.3 | MS | 43 | Self-fishing | No | 2 |
| … | | | | | | | | | |
| Person 400 | 6 | Male | 121.2 | 21.9 | NA | 15 | Market | NA | 1 |

# Data preparation – cleaning, harmonizing and structuring

- ✓ Upload your data.
- ✓ Choose variables you are going to work with.
- ✓ Inspect the variables (do descriptive statistics, check unique values, check data type, check amount of missing data).
- ✓ Harmonize your variables (esp. categorical variables).
- ✓ Structure your data.

# Preliminary analysis

- ✓Build histograms in case of numerical variables.
- ✓Build box plots in case of categorical variables.
- ✓Build scatterplots to see associations between numerical variables.
- ✓Check data distributions in case of numerical variables (normality tests).
- ✓Transform the data if necessary.

**You may proceed to the analysis**

Task!

# Data preparation – cleaning, harmonizing and structuring

- ✓ Structure your data.
- ✓ Upload your data.
- ✓ Choose variables you are going to work with.
- ✓ Inspect the variables (do descriptive statistics, check unique values, check data type, check amount of missing data).
- ✓ Harmonize your variables (esp. categorical variables).

# Preliminary analysis

- ✓ Build histograms in case of numerical variables.
- ✓ Build box plots in case of categorical variables.
- ✓ Build scatterplots to see associations between numerical variables.

- ✓ Check data distributions in case of numerical variables (normality tests).

- ✓ Transform the data if necessary.

## You may proceed to the analysis

# Practical part