Data tables cleaning – checklist in R

Dana Ransby

2024-06-05

Overview

- Discuss several aspects of tabular data cleaning using R
- Presentation of a code in RStudio
- Create an untidy table
- Deal with common issues
- Goal: obtain clean data (for example for data submission to PANGAEA)

Dealing with common issues

- Date and time in separate columns
- Latitude and longitude in single cell
- Different ways of marking missing values
- Comma separated values
- Parameter with unit, that needs rescaling
- () instead of []
- Unrealisticly high number of decimal points
- Row with comments
- Row with aggregated statistics
- Feature with abbreviations (Threatened status)
- Species column with misspelled species names
- Column with NaN only -Leading/trailing/double white spaces

% http://rmarkdown.rstudio.com

Slide with Bullets

- Bullet 1
- Bullet 2
- Bullet 3

Slide with R Output

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Slide with Plot