Introduction to Data Management in R

Christopher Maronga - R Programmer

NDORMS, Oxford University

Introduction

- Data is a valuable component to any industry that generates it

- Workflows in the context of data management.
  - Collecting
  - Collating (involves processing)
  - Storing
  - Archiving
  - Security, curation, access and sharing

- Focus on a single aspect - processing and manipulation using R software, to prepare a dataset for analysis
  - Changing it's dimension
  - Altering it's structure

- Assume familiarity with tidyverse - we will heavily use it.

Scope

- Manipulation
  - subsetting data
  - Type conversion
  - variable recoding & renaming
  - creating new variables
  - merging datasets (joins)

- Cleaning
  - Missing values
  - Root outliers
  - Improbable values

- Transforming
  - Reshaping
  - Group summary and manipulation
  - List columns

References

- Data wrangling sections (chapters 9 - 16) in R for Data Science book

- Rstudio cheatsheets for dplyr and tidyr

- Introduction to dplyr

- Tidy data with tidyr

- Suzan Baert's blogs on data wrangling