Course Outline: Comparing Strategies to Deal with Missing Data in R

Hannes Seller

July 28, 2021

Contents

Description

There are many different ways to deal with missing values. The results often vary in accuracy and time invested. In this course, the learner can compare the results of different methods on the same data to better understand their costs and benefits.

Chapter 1: Get Familiar with the Data

Lesson 1: Introduction to the Dataset (`MASS::Boston` with housing data)

- Learning objective: Load the `MASS::Boston` into a dataframe and check its dimensions and columns
- Some functions used: `read.csv`, `summary`

Lesson 2: Inconsistent Missing Value Indicators

- Learning objective: Check if the dataset contains other indicators different from NA and replace them with NA
- Some functions used: `naniar::replace_with_na_all`

Lesson 3: Analyze the Missingness

- Learning objective: Get familar with missing data specific to the data set
- Some functions used: `is.na`, `naniar::miss_var_table`

Chapter 2: Omitting and Substituting Data

Lesson 1: Delete all Incomplete Observations

- Learning objective: Detect incomplete observations and remove them
- Some functions used: `complete.cases`

Lesson 2: Substitute Continuous Variables with Mean / Median

- Learning objective: Replace missing values with mean and median values
- Some functions used: `imputeTS::na_mean`, `imputeTS::na_median`

Lesson 3: Substitute Continuous Variables with LOCF / Random

- Learning objective: Replace missing values using forward propagation or random samples.
- Some functions used: `imputeTS::na_locf`, `imputeTS::na_random`

Lesson 4: Compare Distributions of Manipulated Datasets

- Learning objective: Visually and analytically compare the results of omission and substitution methods
- Some functions used: `summary`, `ggplot2::geom_boxplot`, `ggplot2::geom_histogram`

Chapter 3: Imputing Missing Data

Lesson 1: Use MICE Functions to Impute Missing Data

- Learning objective: Use the `MICE` package to impute missing data and understand the process behind the algorithm
- Some functions used: `MICE::mice`

Lesson 2: Use kNN to Impute Missing Data

- Learning objective: Use a knn method to impute missing data and understand the process behind the algorithm
- Some functions used: `VIM::kNN`

Lesson 3: Compare Distributions of Manipulated Datasets

- Learning objective: Visually and analytically compare the results of imputation methods
- Some functions used: `summary`, `ggplot2::geom_boxplot`, `ggplot2::geom_histogram`

Chapter 4: Visualize Missing and Imputed Data

Lesson 1: Visualize the Missing Data

- Learning objective: Visualize data and highlight incomplete observations
- Some functions used: `imputeTS::ggplot_na_distribution`, `imputeTS::ggplot_na_gapsize`

Lesson 2: Visualize the Imputed Data

- Learning objective: Visualize data and highlight imputed observations
- Some functions used: `imputeTS::ggplot_na_imputations`

Lesson 3: Visual Comparison with Complete Dataset

- Learning objective: Visually and analytically compare the results with the original complete dataset
- Some functions used: `summary`, `ggplot2::geom_boxplot`, `ggplot2::geom_histogram`