Track Proposal: Working with Missing Data in R

Hannes Seller

July 28, 2021

Contents

Introduction

Going through the current *DataCamp* course library, I realized that the topic of dealing with missing data can be expanded, as it is one of the most common challenges in data science. The courses use verbs from Bloom's taxonomy and become increasingly more complex.

Course description

Data sets with incomplete observations are one of the most common challenges when dealing with data. Almost every data set comes with missing values, and the reasons are manyfold. Hence, dealing with missing data is an essential skill. In this track, you will learn about different types of missingness and how to engage with them. The courses cover quick and easy strategies to substitute* missing data with expected values, and more advanced prediction methods that use regressions and Machine Learning techniques like k-nearest neighbors (knn) and decision trees. You learn how to visualize data sets with missing and imputed values. Keep in mind: Imputing data is costly, and simple methods might be preferable sometimes. If you ever work with real-life data, the insights you gain from this track should not be *missing*.

* In this outline I use the term "substitute", when a missing value is replaced by a fixed value (e.g. the column mean or the last observation carried forward), and "impute" when a value is predicted via a model for a specific observation.

Courses in the track

Course 1: Introduction to Missing Data (no coding)

Understand why data sets contain missing values and learn about common strategies to deal with them.

The learner will be able to (KNOWLEDGE / UNDERSTAND):

- Identify reasons why values might be missing (randomly or with pattern) on given examples;
- Memorize different ways to represent missing values (NA, "NA", "./.", "N/A", NULL) and contrast their usefulness;
- Memorize shortly described methods to deal with missing values (omission, substitution, imputation).
- Compare effects of different substitution methods (e.g. substitution with mean, median, and mode);
- Describe strategies used to reduce the number of missing data when collecting data; Summarize and paraphrase methods described in this course.

Course 2: Dealing With Missing Data in R (existing course)

Make it easy to visualize, explore, and impute missing data with naniar, a tidyverse friendly approach to missing data.

The learner will be able to (KNOWLEDGE / UNDERSTAND / APPLY):

- Identify reasons why values might be missing (randomly or with pattern) on given examples;
- Memorize different ways to represent missing values (NA, "NA", "./.", "N/A", NULL) and contrast their range of applications;
- Memorize shortly described methods to deal with missing values (omission, substitution, imputation).
- Use functions from the `naniar` package to summarize how many values are missing in a dataset and how to replace missing value indicators with other values;
- Use functions from the `naniar` package to visualize the missingness of values across one or multiple variables.
- Use functions from the `simputation` package to impute missing values in a dataset.

Course 3: Omitting and Substituting Missing Data in R

This course offers many hands-on exercises that showcase how to deal with missing values in categorical and continuous variables.

Prerequisites: Data Manipulation with dplyr

The learner will be able to (UNDERSTAND / APPLY):

- Use functions like `is.na` or `table` from the standard library and functions from the `naniar` package to gain an overview about a dataset's missingness;
- Use `dplyr` functions to filter and delete incomplete observations and articulate how the changes affect other variables in the dataset (e.g. distribution, mean values, standard deviation);
- Use functions like `is.na` and `ifelse` to find and replace missing values in a dataframe column with mean, median or mode values of the respective variable;
- Use functions from the `apply` family to replace missing values with mean, median or mode values of the respective variable;
- Use functions from the `imputeTS` package like `na_mean` or `na_replace` to replace missing values more easily;
- Contrast methods to substitute values in categorical (factors, logicals) and continuous (numeric) variables;
- Compare effects of different substitution methods (e.g. substitution with mean, median, and mode);

Course 4: Visualizing Missing Data in R

The naniar, VIM, and imputeTS packages provide functions to visualize missing and imputed data.

Prerequisites: Visualization Best Practices in R

The learner will be able to (UNDERSTAND / APPLY):

- Use visualization functions from the standard library and the `ggplot2` package to compare the variance between complete and incomplete observations (e.g. facets, histograms, scatterplots, or missingness illustrated by colors and shapes).
- Use functions from the `naniar` and `vim` packages to create graphs specifically designed to show missing values in different variables of a dataframe;
- Use functions from the `imputeTS` package like `plotNA.distribution` or `plotNA.imputations` to visualize statistics about missing or imputed time series data. Descriptively articulate what the respective visualizations show.

Course 5: Imputing Missing Data with Regressions in R

In this course you create regression models to impute missing data.

Prerequisites: Introduction to Regression in R, Supervised Learning in R: Regression

The learner will be able to (APPLY / ANALYZE):

- Use the lm function to create uni- and multivariate linear models to predict continuous values from a dataset;
- Use functions from the `MICE` package to perform logistic regressions to impute missing continuous data (e.g. "species" from "iris" data set).
- Use functions from the `randomForest` package to perform logistic regressions to impute categorical missing data;
- Use functions to visualize the models as scatterplots or trees; Compare the accuracy of intentionally deleted observations with predicted values;
- Discuss in which cases one of the methods shown would be useful.

Course 6: Imputing Missing Data with Supervised Learning in R

The k-nearest neighbors algorithm can be used to impute missing values based on similar observations in the data set

Prerequisites: Supervised Learning in R: Classification

The learner will be able to (APPLY / ANALYZE):

- Use the "iris" dataset to understand the concept of n-dimensional Euclidean distances;
- Use the `knn` function to build a prediction model to impute missing data;
- Compare the accuracy of intentionally deleted observations with predicted values;
- Visualize and analyze False and True positives using a ROC diagram and the AUC;
- Discuss in which cases one of the methods shown would be useful.

Course 7: Comparing Strategies to Deal with Missing Data in R

In this course you compare the costs and benefits of different methods to deal with missing data (omission, substitution, imputation).

Prerequisites: Data Manipulation with dplyr, Visualization Best Practices in R

The learner will be able to (APPLY / ANALYZE):

- Import a data set (like `MASS::Boston` Housing data) with multiple rows and columns from a CSV file (the dataset is manipulated to have continuous and categorical values missing at random; indicators are used inconsistently);
- Analyze the missingness of data (e.g. percentage of NA per column);
- Decide whether to delete, substitute or impute data;
- Apply previously learned methods (`imputeTS`, `MICE`, `knn`) to build datasets with (1) omitted observations, (2) substituted values, and (3) imputed predictions;
- Visually (`imputeTS`, `VIM`) and analytically compare the variance of previously manipulated datasets and evaluate which methods were the most cost efficient;