Lecture 4

–

Missing Values

Haussperling (Spatz) / House Sparrow

(*Passer domesticus*)

- It is a small bird that has a typical length of 16 cm. Females and young birds are coloured pale brown and grey, and males have brighter black, white, and brown markings.
- The house sparrow is strongly associated with human habitation, and can live in urban or rural settings.
- Because of its numbers, ubiquity, and association with human settlements, the house sparrow is culturally prominent.
- Most house sparrow vocalisations are variations on its short and frequent chirping call. Transcribed as *chirrup*, *tschilp*, or *philip*, this note is made as a contact call by flocking or resting birds; or by males to proclaim nest ownership and invite pairing.

Sources:
* Photo by Susanne Jutzeler on pixabay
* https://en.wikipedia.org/wiki/House_sparrow

Motivation – Missing Values

Missing values

- **very** common in real-world dataset
- can cause severe problems when not treated properly
- Are very often are not treated properly

Missing Values

1. Reasons for Missing Values
2. Consequences of Missing Values
3. Detecting Missing Values
4. Types of Missing Values
5. Handling Missing Values

Sources of Missing Values

- Typical reasons for / sources of missing data
  - Something crashed and data was not recorded for a certain amount of time
  - Users chose not to fill out a field
  - Data was lost or corrupted while transferring manually from a legacy database.
  - Bug in the code

- Some of these sources are just random mistakes

- Others caused by a deeper, underlying reason

Missing Values

1. Reasons for Missing Values
2. Consequences of Missing Values
3. Detecting Missing Values
4. Types of Missing Values
5. Handling Missing Values

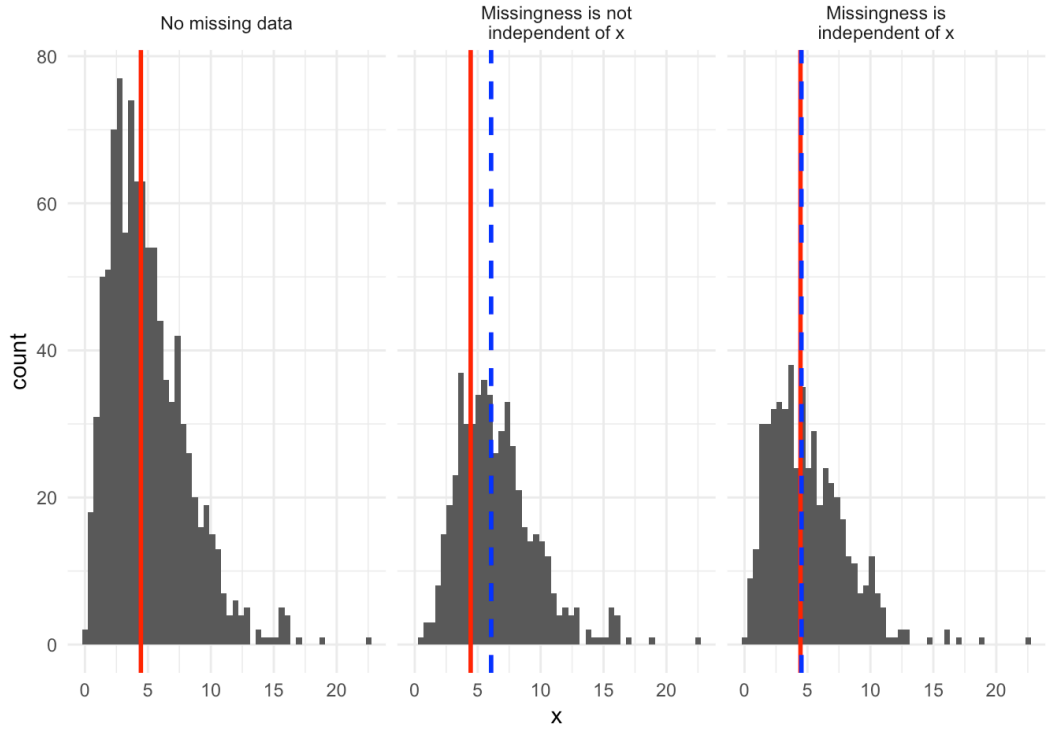
Consquences of Missing Values → Bias

- Missing data can induce bias in the estimates of descriptive patterns and causal effects!

Example

- describe income distribution in a country with survey data.
- Some individuals' incomes are missing.
- Suppose low-income individuals are less likely to report their income than high-income individuals → thus missingness concentrates in the lower portion of the distribution.
- → characterization of the income distribution biased!

Summary: Missingness correlated with the variable we are trying to describe → characterization of the median of the distribution is biased.



Source: https://egap.org/resource/10-things-to-know-about-missing-data/

Missing Values

1. Reasons for Missing Values
2. Consequences of Missing Values
3. Detecting Missing Values
4. Types of Missing Values
5. Handling Missing Values

How to identify missing values

- NaN (Not a Number)
- Placeholder Values
  - Invalid Values (often 0 or negative or out-of-reasonable-range like age 999)
  - Special Values (PLZ 12345, 11111, 99999 or 27498: “Handelsmärkte für Helgoland”)

➔ Domain knowledge critical

Dataset: Pima Indians Diabetes Database

- From the National Institute of Diabetes and Digestive and Kidney Diseases.
- Business Problem: diagnosing diabetes is predict whether or not a patient has diabetes
- Data Science Problem:
  - Idea: predict diabetes based on easy-to-measure tests
  - Fundamental task: Classification
  - Data:
    - Several predictor variables e.g. the number of pregnancies the patient has had, their BMI, insulin level, age, ….
    - Target variable: Outcome.
    - Constraints on the selection of the patients (taken from a larger database), e.g. all patients females at least 21 years old of Pima Indian heritage.

Number of Instances: 768

Number of Attributes: 8 plus class

Attributes:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function (scores the likelihood of diabetes based on family history)
8. Age (years)

9. Class variable (0 or 1)

Exercise

Exercise 1

Pima Indian Diabetes Dataset – Detect Missing Values

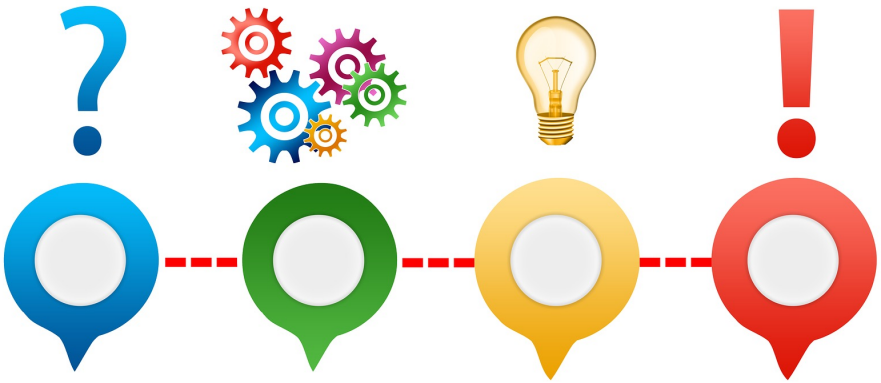


Photo by Gerd Altmann on Pixabay

Missing Values

1. Reasons for Missing Values
2. Consequences of Missing Values
3. Detecting Missing Values
4. Types of Missing Values
5. Handling Missing Values

Types of Missingness

MCAR:
Missing Completely At Random

Assumption: fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
In practice, this is assumption very rarely holds.

Example:
- in a medical study, a student drops a random blood sample

MAR:
Missing At Random

Probability that a value is missing depends on the set of **observed features**, but is neither related to the specific missing values nor to other **not observed features**.
This is the most common assumption used in practice.

Example:
- in a survey, women more often refuse to give their age then men (and the gender is recorded)

MNAR:
Missing Not At Random

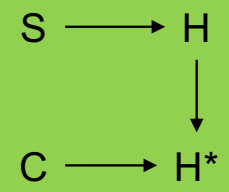
The missingness of a value either depends on the **missing value itself** or on **unobserved features**, which is problematic.

Examples:
- in a survey, people with very low income leave the income field empty more often then people with higher incomes (missingness of the income depends on the missing income value itself)
- in a survey, women more often refuse to give their age then men (and the gender is not recorded) (missingness of the age depends on the gender)

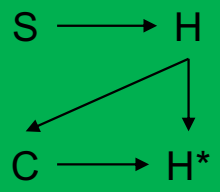
Source: Donald B. Rubin, Inference and Missing Data; Biometrika (1976), 63, 3, pp. 581-92

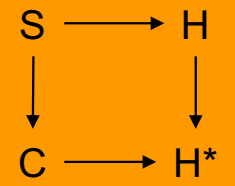
Categories of Missing Values

MCAR: Professors cat chews on random homeworks

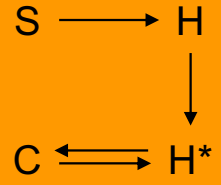


MAR: Professors cat chews on homeworks more than 2 pages long





MNAR (on not collected features): Professors cat (very sensitive) chews on smoking students homeworks



MNAR (on missing value itself): Professors cat (very smart) chews on bad homeworks

S, H, C, H*: Sets of Attributes (S of the Student, H of the complete Homework, C of the Cat, H* of the Homework with missing values), Arrows: Dependencies

Exercise

Exercise 2

Pima Indian Diabetes Dataset – MCAR, MAR or MNAR?

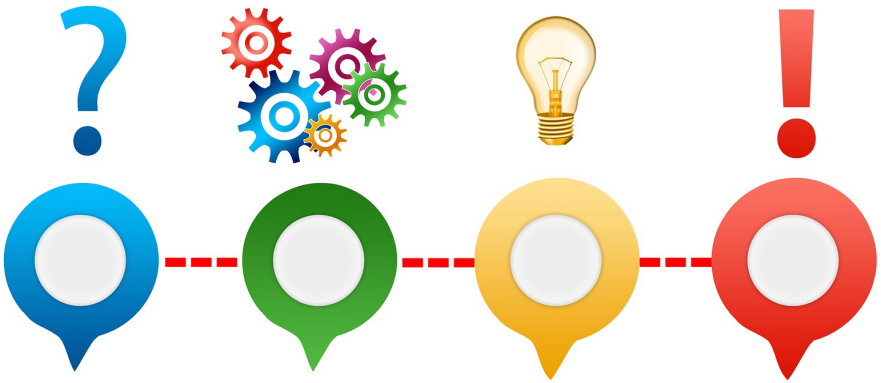


Photo by Gerd Altmann on Pixabay

Missing Values

1. Reasons for Missing Values
2. Consequences of Missing Values
3. Detecting Missing Values
4. Types of Missing Values
5. Handling Missing Values

Handling Missing Data - Overview

- The best solution to handle missing data is to have none!
  - can never be certain, if missing values are MCAR, MAR or MNAR due to *lurking variables*
- Options for Handling Missing Data
  - Avoid Generating Missing Data
  - Just leave it as it is
  - Deletion
  - Imputation
    - Simple Imputation with fixed value
    - Simple Imputation with predicted value
    - Multiple Imputation
    - (Imputation for Time Series Data)

Deletion

- Deleting Rows / Listwise Deletion / Complete Case Analysis
  - Delete all Rows (samples) with Missing Values
- Deleting Columns
  - Delete all Columns with Missing Values (→ Drop Features)
- Pairwise Deletion
  - Use all available Values
  - Mostly used for statistical analysis
    - Example:
      to compute the mean of column A, use all non-missing values in column A
      to compute the mean of column B, use all non-missing values in column B
      → different row/different number of rows may be used for each mean
      → can lead to “weird” / unexpected results

Simple Imputation

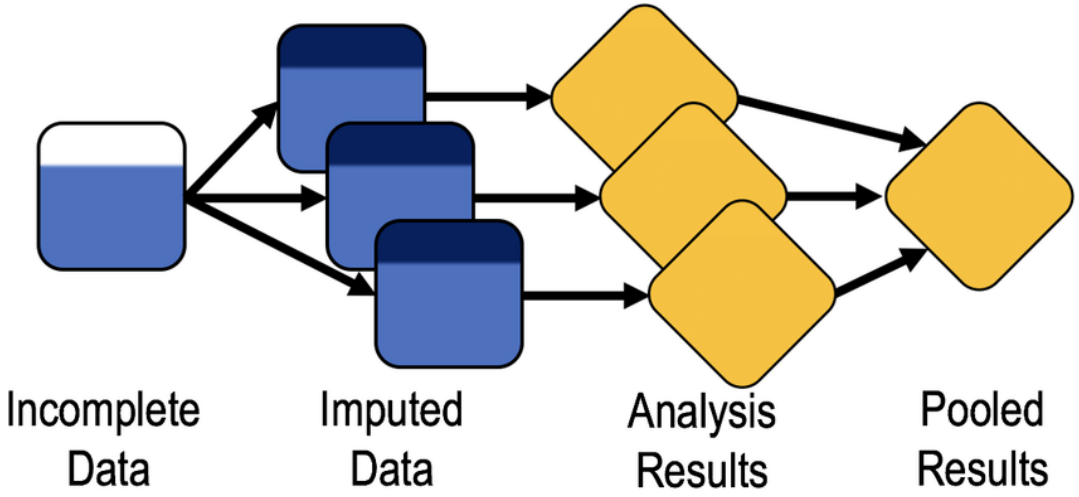
- Replace missing values with ”some” other value
- Most common
  - use mean / mean in subsets of rows
  - use median / median in subsets of rows
  - use mode / mode in subsets of rows
- Additionally, most common for time-series data
  - Previous / Next Value
  - Linear / (Moving) Average Interpolation
- Common Mistake
  - When using train/test (or k-fold-cross validation), the imputed value must be computed on the training-set only (this makes train/test and esp. k-fold-cross-validation much more complex to handle)

Imputation: Missing Value Prediction

- Predict Missing Value from the other features
  - Numerical data: often linear regression
  - Categorical data: often k-NN or Decision Trees
  - Many other machine learning algorithms can be used as well, even deep learning
  - May give very good results
  - But may be very hard to explain (“black box”)

- Common Mistake
  - Predictor must be trained on the training-set only
  - Target Variable must not be used to impute features (or you may/will have “target leakage”)
    - This should be obvious – we will have to impute missing values for the real data when the model is used later on, and the target is not available for this data…

Multiple Imputation

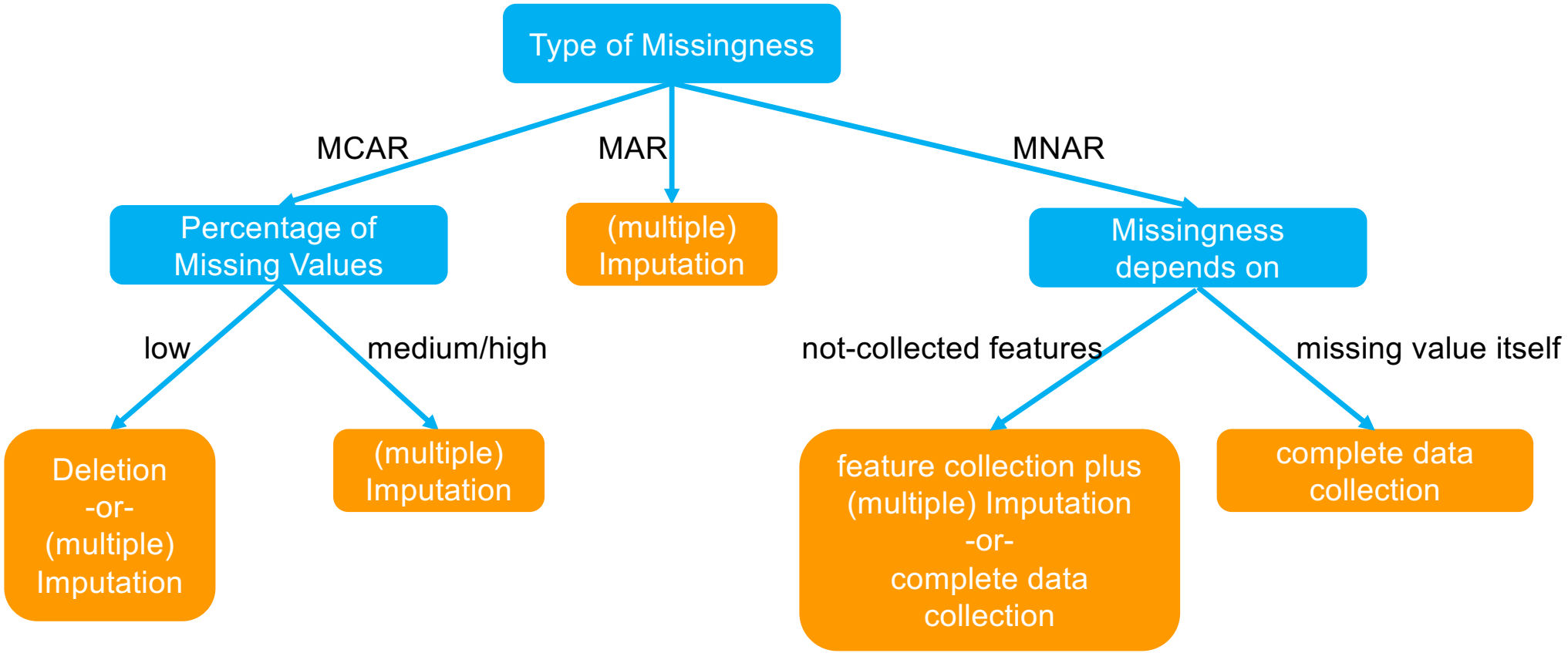


Multiple Imputation process

- missing data (shown in white) are imputed (shown in dark blue)
  ➔ M complete data sets (shown for M=3)
- Each complete imputed data set is analysed (e.g. classification, linear regression, …)
- Results are combined (Pooled)

Source: https://www.researchgate.net/figure/The-multiple-imputation-MI-process-In-the-first-step-missing-data-shown-in-white-are_fig4_334213038

Missing Value Decision Tree



Key Takeaways

- Missing values
  - very common in real-world dataset
  - can cause severe problems when not treated properly
- MCAR, MAR, MNAR: hard to decide, domain knowledge is key
- Deleting missing values
  - Only safe if MCAR
  - MAR, MNAR: will introduce bias in the model
  - Will reduce the power of the model
- Imputation, esp. multiple Imputation
  - best approach if MAR or MCAR
- If MNAR has to be assumed, there is no good solution
  - If missingness depend on not-collected feature: collect these feature (so that MAR holds)
  - If missingness depends on the missing values themselves: collect data without missing values

Image by Gordon Johnson on pixabay

Exercise

Exercise 3

Pima Indian Diabetes Dataset – Deletion and Imputation

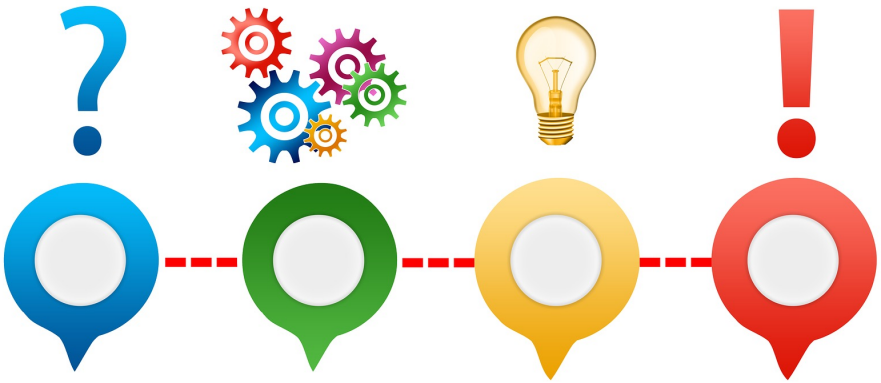


Photo by Gerd Altmann on Pixabay