Data Analysis and Visualization

CentraleDigitalLab@Nice

Deborah Dore - PhD - ddore@i3s.unice.fr
MARIANNE, I3S, CNRS, INRIA, Université Côte d'Azur

Handling Missing Values

Introduction

Definition: Missing values occur when no data is recorded for a variable in an observation, creating gaps in datasets

Significance:

- Impacts the quality and reliability of data analysis.
- Affects the performance of predictive models.
- Can introduce bias if not handled properly.

Examples:

- Missing age in a patient dataset.
- Missing survey responses.

Handling Missing Values

Types of missing values

- **MCAR (Missing Completely at Random):** Missingness has no relationship to any values, observed or missing.
- The probability of missingness is the same for all observations. In other words, the missingness is independent of both observed and unobserved data.

Example: Random equipment failure during data collection.

Handling Missing Values

Types of missing values

- **MAR (Missing at Random):** Missingness is related to observed data but not the missing data itself.
- The probability of missingness is systematic and can be predicted by other observed data, but not by the missing data itself.

Example: Income data missing but correlated with age.

Handling Missing Values

Types of missing values

- **MNAR (Missing Not at Random):** Missingness depends on the missing data itself.

- The probability of missingness is related to the missing data itself, even when controlling for other observed variables.

Example: People with high income may choose not to disclose it.

Handling Missing Values

Why Handle Missing Values?

Impact on Analyses:

- Leads to incorrect statistical conclusions
- Reduces model accuracy and robustness

Broad Approaches:

- Deletion Methods
- Imputation Methods
- Hybrid Approaches

Key Goal: Preserve as much useful information as possible.

Handling Missing Values

Deletion Methods Overview

Definition: Removing observations or variables with missing data

When to Use: Deletion methods are used when the missing data is small and random and the loss of data does not critically affect analysis

Advantages:

- Simple and quick to implement.
- Ensures that the remaining dataset is complete and consistent.
- Reduces the risk of introducing biases due to imputation assumptions.

Disadvantages:

- Can result in significant data loss, especially with high proportions of missing values.
- May introduce bias if the missing data mechanism is not MCAR (e.g., if missingness is related to other variables).
- Reduces statistical power and the generalizability of results.

Handling Missing Values

Deletion Methods Overview: row-wise deletion

Definition: Removes rows with any missing values

Advantages:

- Simplifies analysis by ensuring complete cases.
- Commonly used in statistical software.

Disadvantages:

- Significant loss of data.
- Can introduce bias if not MCAR.

Example: In a dataset with 100 rows and 10 rows have missing values, row-wise deletion retains 90 rows.

Handling Missing Values

Deletion Methods Overview: column-wise deletion

Definition: Removes columns with high proportions of missing data

Advantages:

- Retains rows and focuses on well-populated features

Disadvantages:

- Loss of potentially important variables
- Reduces feature space for modelling

Example: Removing a column with 80% missing values from a survey dataset

Handling Missing Values

Imputation Methods Overview

Definition: Replace missing values with estimated values computed on available data

When to Use: Imputation is used when the proportion of missing data is not excessive, and the missing data mechanism is either MCAR or MAR

Advantages:

- Retains the size of the dataset
- Reduces bias in comparison to deletion methods
- Reduces risk of loss of information

Disadvantages:

- Introduces assumptions into the dataset
- Risk of underestimating the true spread of data

Handling Missing Values

Imputation Methods Overview: Mean Imputation

Definition: Replace missing values with the mean of the observed data in the column.

Advantages:

- Easy to implement.
- Maintains overall mean of the dataset.

Disadvantages:

- Reduces data variability.
- Introduces bias in skewed distributions.

Example: For a column of ages [25, 30, 35, NaN], replace the missing value with the mean (30)

Handling Missing Values

Imputation Methods Overview: Median Imputation

Definition: Replace missing values with the median of the observed data in the column

Advantages:

- Robust to outliers
- Maintains central tendency for skewed data

Disadvantages:

- Reduces data variability.

Example: For incomes [30,000, 50,000, NaN, 80,000], replace the missing value with the median (50,000).

Handling Missing Values

Imputation Methods Overview: Mode Imputation

Definition: Replace missing values with the mode (most frequent value) of the observed data.

Advantages:

- Useful for categorical variables.
- Simple to apply.

Disadvantages:

- Over-represents the mode value.

Example: For colors ["red", "blue", "blue", NaN], replace the missing value with "blue."

Handling Missing Values

Imputation Methods Overview: KNN Imputation

Definition: Estimates missing values using the k-nearest neighbours based on similarity

Advantages:

- Retains data relationships and variability
- Works for both numerical and categorical data

Disadvantages:

- Computationally expensive for large datasets
- Sensitive to the choice of k (number of neighbours)

Example: Predict missing house prices based on nearby houses with similar features

Handling Missing Values

Imputation Methods Overview: KNN Imputation

Who does it works?

1. For each data point with missing values, the algorithm identifies the k most similar data points (nearest neighbours) based on the available features

2. The missing values are then estimated by calculating a weighted average of the corresponding values from these nearest neighbours

3. The weights are typically based on the distance between the data point with missing values and its neighbours, with closer neighbours having more influence

Handling Missing Values

Hybrid Approaches

Definition: Combining multiple methods for complex datasets.

Example: Apply median imputation for numerical data and KNN for categorical data

Handling Missing Values

Comparison of Techniques

Factors to Consider:

- Dataset size and complexity
- Proportion of missing data
- Type of variable (numerical or categorical)
- Computational resources

Handling Missing Values

Case Study 1

Scenario: Dataset with 5% missingness in numerical features (e.g., customer ages).

Handling Missing Values

Case Study 1

Scenario: Dataset with 5% missingness in numerical features (e.g., customer ages).

Technique Used: Mean imputation.

Outcome:

- Quickly resolved missingness.
- Minor bias in mean calculations

Handling Missing Values

Case Study 2

Scenario: Dataset with 20% missingness in categorical features (e.g., survey responses).

Handling Missing Values

Case Study 2

Scenario: Dataset with 20% missingness in categorical features (e.g., survey responses).

Technique Used: Mode imputation.

Outcome:

- Preserved majority category distributions.
- Slight over-representation of most frequent values.

Handling Missing Values

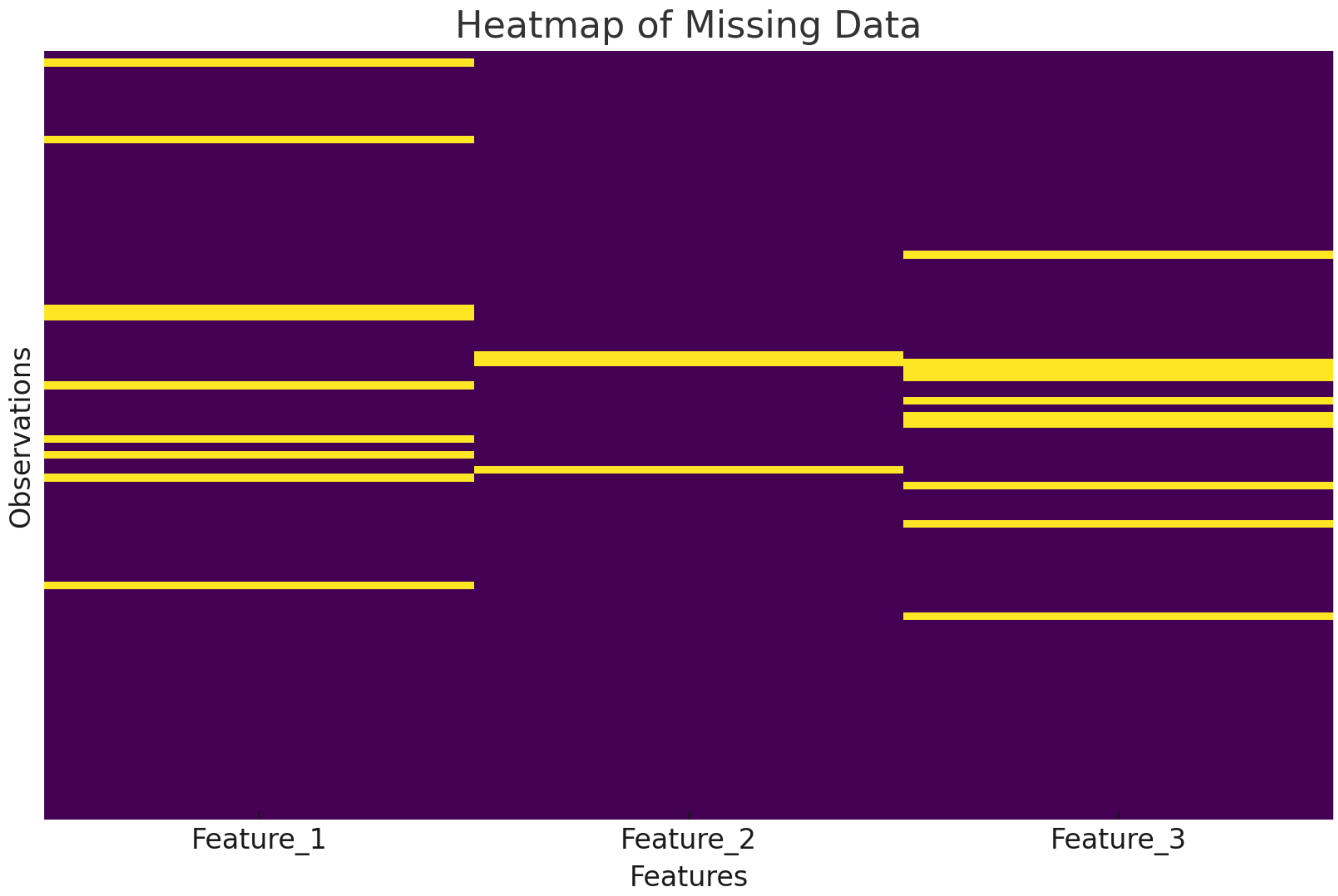
Visualising Missing Data

Techniques:

- **Heatmaps**: Highlight missing values (e.g., seaborn heatmap in Python)
- **Bar Charts**: Show proportions of missing values per column
- **Missingness Matrices**: Visualize patterns of missingness

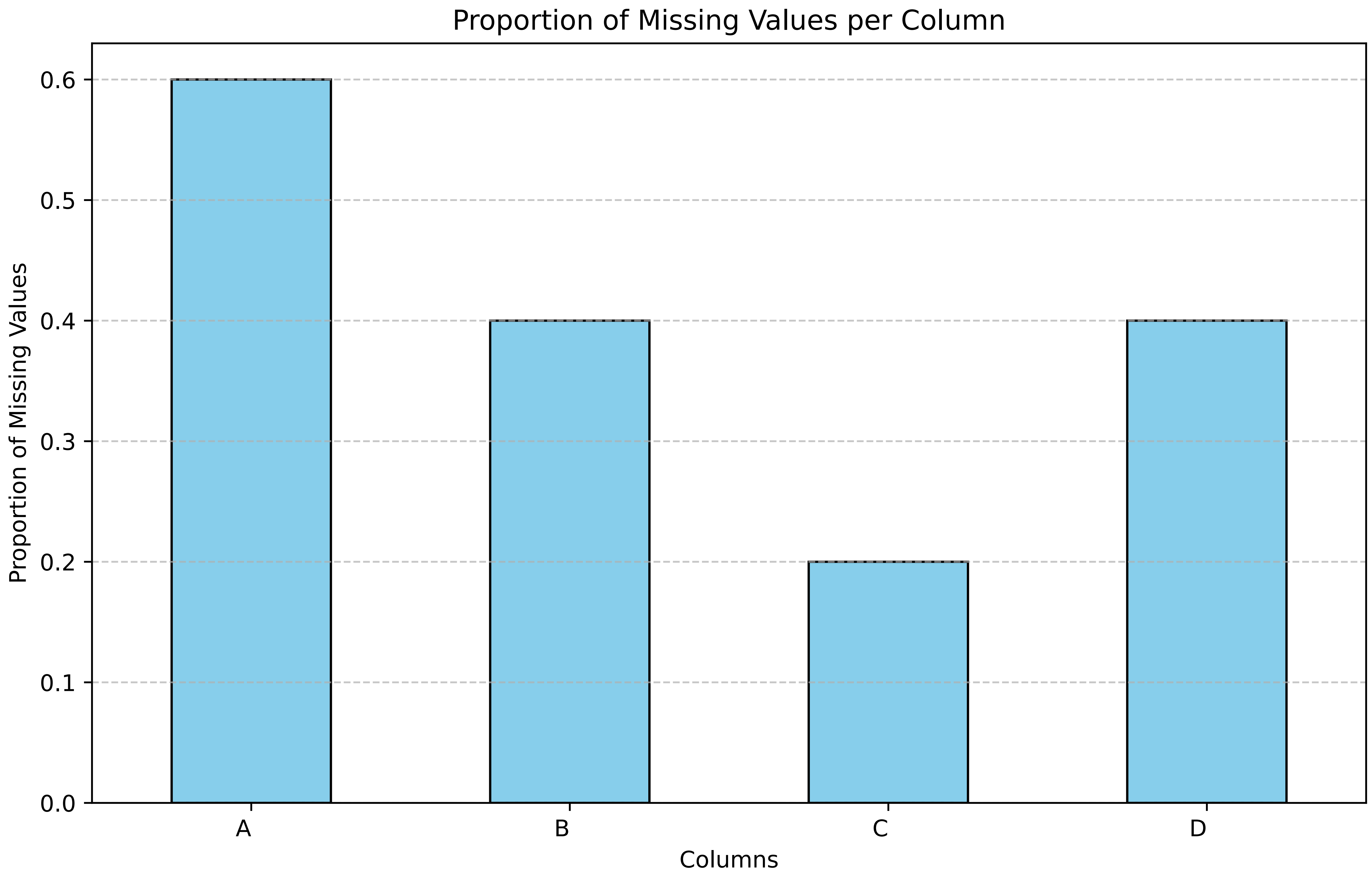
Handling Missing Values

Visualising Missing Data: Heatmap



Handling Missing Values

Visualising Missing Data: Bar plots



Demo with Notebook_Missing_Data.ipynb

Useful links

- https://scikit-learn.org/stable/modules/impute.html
- https://scikit-learn.org/stable/api/sklearn.impute.html