Table of Contents

# 2.3 Imputation Techniques

## Introduction

Real‑world datasets often contain missing values due to nonresponses or recording errors. Imputation techniques replace missing values with reasonable estimates to preserve the dataset’s structure and reduce bias.

## Approaches

* **Deletion (listwise/pairwise)** – Remove rows or columns with missing values. This is simple but can discard valuable information.
* **Simple imputation** – Replace missing values with a statistic such as the mean, median or most frequent value of the column【266912096589978†L125-L151】. A constant value can also be used when appropriate.
* **Multivariate imputation** – Use regression models to predict missing values from other variables. The Iterative Imputer in scikit‑learn, for example, models each variable with missing values as a function of the others【266912096589978†L125-L151】.
* **Multiple imputation** – Generates several possible values for each missing point, analyses each completed dataset separately and then combines the results, accounting for uncertainty.

## Example

Consider a dataset of patient records with missing blood pressure values. A simple approach fills in missing values with the median blood pressure. A more sophisticated approach builds a regression model using age, weight and diagnosis to predict blood pressure for missing cases.

## Summary

Imputation techniques help maintain data integrity when missingness occurs. Simple methods are easy to implement but may ignore relationships between variables, whereas multivariate and multiple imputation leverage more information to produce more accurate estimates【266912096589978†L125-L151】.

## Reflection questions

1. Why is deleting observations with missing values often undesirable?
2. Compare mean imputation with multivariate regression imputation.
3. What is the purpose of multiple imputation?

## References