databricks

Data Preparation and Feature Engineering

LECTURE

# Data Imputation

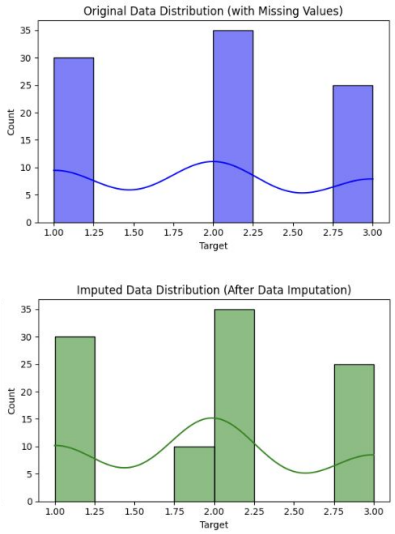


In this lecture, we discuss data imputation, exploring problems with missing data, factors that influence imputation methods, techniques for replacing missing values, and best practices like marking imputed data.

# Data Imputation

Data imputation is **the process of filling in missing values** in a dataset with estimated or predicted values.

The goal of data imputation is to enhance the quality and completeness of the dataset, ultimately improving the performance and reliability of the machine learning model.






Data imputation is the process of filling in missing values in a dataset using estimated or predicted values.

The goal is to improve the completeness of the data so that models can be trained without errors or loss of important information.

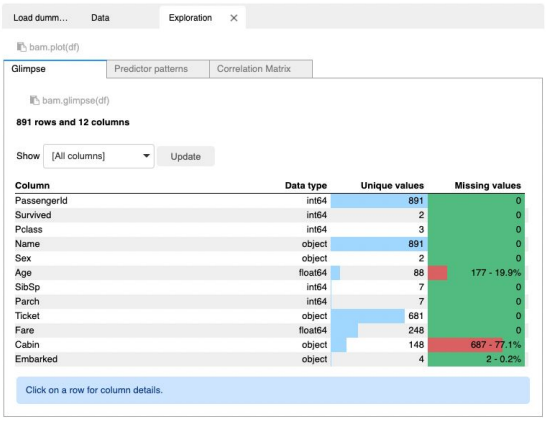
On the right, the top chart shows a dataset with missing values—note the gaps in distribution. The bottom chart shows the same data after imputation, where the missing values have been filled, resulting in a more consistent distribution.

This helps maintain the structure of the data and supports better model performance.

# Problems with Missing Data

Impacting the performance and reliability of ML models

- Reduced Model Performance
- Biased Inferences
- Imbalanced Representations
- Increased Complexity in Model Handling



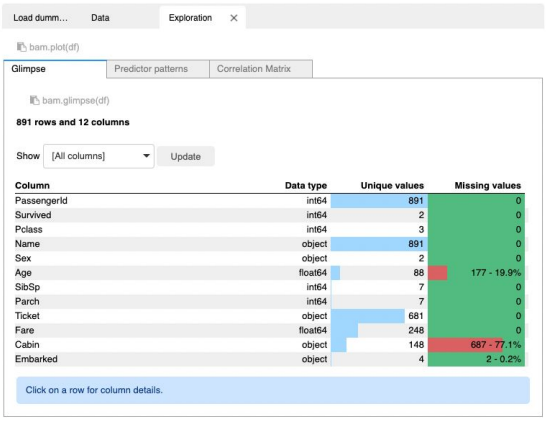



- **missing data** = **impacting the performance and reliability**
  - **Reduced Model Performance:** reduce the accuracy and precision , incomplete insights and skewed predictions
  - **Biased Inferences:** can **introduce** systematic biases, Can **result** in erroneous conclusions
  - **Imbalanced Representations**: impact the model's ability to learn and make accurate predictions
  - **Increased Complexity** in Model Handling: Necessitates advanced preprocessing technique, increases computational overhead and complexity
- Best practices for handling missing data include:
  - **Understanding** Missing Data **Patterns** = can **help in selecting** appropriate techniques
  - Selection of Handling Techniques
    - Implementing **Multiple Imputation**
    - **Evaluating** Techniques' **Impact**
    - **Document**

# Problems with Missing Data

Impacting the performance and reliability of ML models

- Reduced Model Performance
- Biased Inferences
- Imbalanced Representations
- Increased Complexity in Model Handling





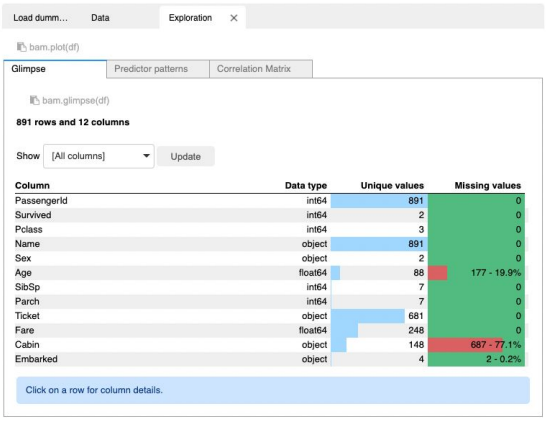

**Missing data** creates major challenges in machine learning, including:

- **Reduced Performance:** Gaps in key features hinder pattern learning and lower accuracy.
- **Biased Inferences:** Non-random missingness can skew predictions toward certain groups.
- **Imbalanced Representation:** Uneven missingness across classes distorts model perception.
- **Added Complexity:** Handling missing values (e.g., imputation, time-series gaps) complicates model development.

# Problems with Missing Data

Impacting the performance and reliability of ML models

- Reduced Model Performance
- Biased Inferences
- Imbalanced Representations
- Increased Complexity in Model Handling



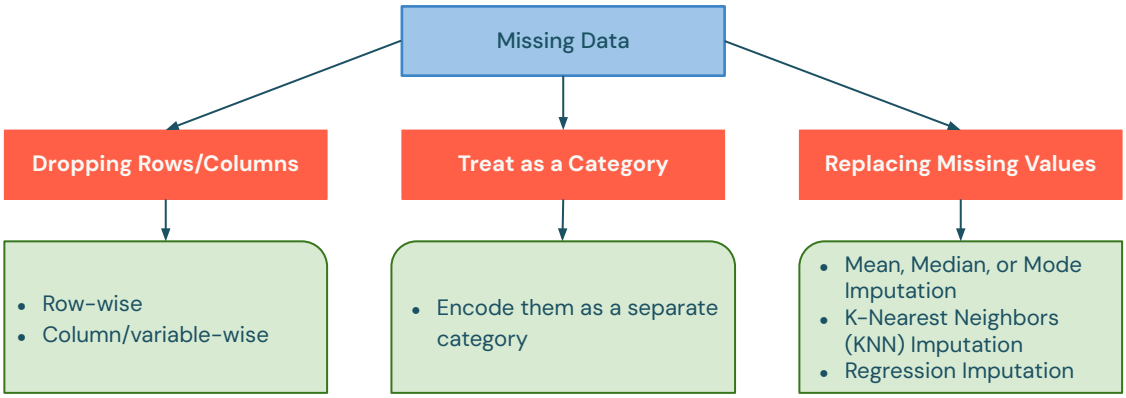



**Example:**

Consider a dataset for predicting employee performance, where variables such as education level, years of experience, and project involvement are crucial. If the "years of experience" feature has missing data, the model might struggle to accurately predict performance, as this information is highly relevant. Biased inferences may occur if the missing data is not random; for instance, if employees with higher levels of experience are more likely to have missing values. This could lead to the model favoring less-experienced candidates, impacting the overall hiring decision process. Addressing missing data is essential to ensure fair and accurate model outcomes.

# How to Handle Missing Data

Data imputation methods

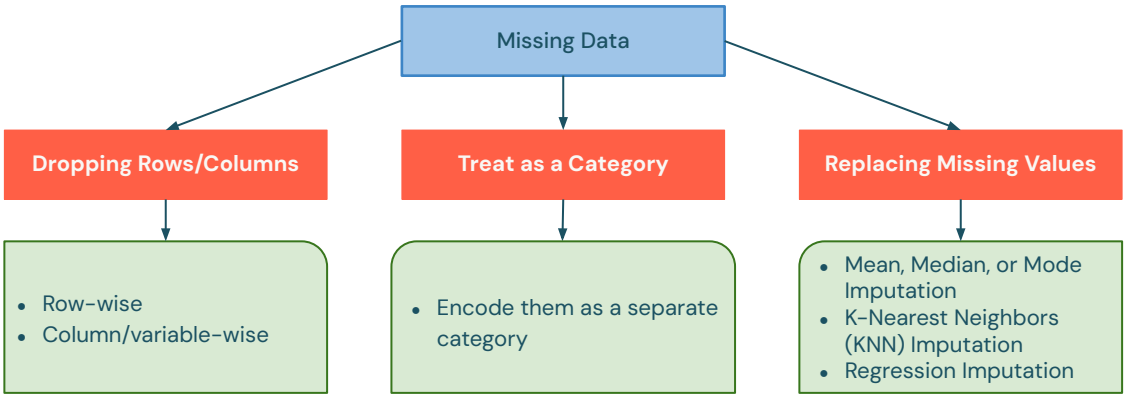





- Best practices for handling missing data include:
  - **Understanding the reasons**
  - **Consider the impact** on your analysis.
  - **Assessing the amount and pattern** of missing data
  - **Validating the impact** of handling missing data on your results
- Considering the **potential bias** introduced.
- ***several approaches to handle missing data***
  - **Dropping rows or columns**: removing rows or columns that contain missing values.
    - **functions** like dropna() in pandas or drop() in Apache Spark.
  - **Treating missing values as a separate category:** missing values may carry important information…you can treat them as a separate category
    - *patient survey responses about lifestyle choices, such as smoking or alcohol consumption, missing responses could be significant. Introducing a category like "Not Disclosed"*
    - *Might need to Create a special category such as "Information Not Available"*
  - **Replacing missing values**

# How to Handle Missing Data

Data imputation methods






**Addressing missing data** is crucial for robust machine learning models. Key approaches include:

- **Dropping Rows/Columns:** If missing data is random and minimal, removing affected rows or columns is effective. For example, in a feedback dataset, dropping missing values for a non-essential feature like favorite color has little impact.
- **Replacing Missing Values:** Imputation replaces gaps with estimates. Simple methods use mean, median, or mode, while advanced ones like KNN or regression consider relationships.
- **Factors Influencing Imputation:** Choice depends on data type (categorical/numerical), extent of missingness, and computational cost—more gaps often require complex methods.

# Replacing Missing Values

## Data imputation methods

### Mean - Mode Imputation

| Before | After |
|---|---|
| 10 | 10.0 |
| 15 | 15.0 |
| - | 18.3 |
| 20 | 20.0 |
| 25 | 25.0 |

### K-Nearest Neighbors (KNN with k=2)

| Before | After |
|---|---|
| 8 | 8.0 |
| - | 10.0 |
| 12 | 12.0 |
| 15 | 15.0 |
| - | 13.0 |

### Multiple Imputation (Regression)

| F1 | F2 | Before | After |
|---|---|---|---|
| X | X | 10 | 10.0 |
| X | X | 15 | 15.0 |
| X | $Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$ → Y | | |
| X | X | 20 | 20.0 |
| X | X | 25 | 25.0 |



- **Replacing missing values:** replace missing values with imputed values.
  - **mean, median, or mode** to fill in the missing values.
  - **KNN** = estimates missing values using the **most similar data points**
  - **Regression Imputation** = **estimate** missing values **based on relationships** inferred from other variables
- **QQ: Which method of imputing missing values (e.g., KNN, multiple imputation) have you found most effective in practice? Why?**

# Replacing Missing Values

## Data imputation methods

### Mean - Mode Imputation

| Before | After |
|---|---|
| 10 | 10.0 |
| 15 | 15.0 |
| - | 18.3 |
| 20 | 20.0 |
| 25 | 25.0 |

### K-Nearest Neighbors (KNN with k=2)

| Before | After |
|---|---|
| 8 | 8.0 |
| - | 10.0 |
| 12 | 12.0 |
| 15 | 15.0 |
| - | 13.0 |

### Multiple Imputation (Regression)

| F1 | F2 | Before | After |
|---|---|---|---|
| X | X | 10 | 10.0 |
| X | X | 15 | 15.0 |
| X | $Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$ | | Y |
| X | X | 20 | 20.0 |
| X | X | 25 | 25.0 |



three common methods for replacing missing values:

- On the left, Mean–Mode Imputation fills in missing values using the mean or mode of the column. For example, the missing value in the third row is filled with 18.3, which is the mean of the available values.
- In the middle, K-Nearest Neighbors (KNN) Imputation finds the nearest neighbors—in this case, k=2—and averages their values to fill in the missing entry. The missing value is replaced with 13.0 based on its two nearest neighbors.
- On the right, Multiple Imputation using Regression uses a regression model to estimate the missing value based on other features. Here, the missing value in column Y is predicted using a formula involving F1 and F2.

Each method shown provides a different way to handle missing data, depending on the dataset and context.

# Factors Influencing Imputation Method

## There isn't a definitive best approach

### Nature of Data

- Is data type is **continuous, ordinal or categorical**?
- Data **distribution**: For example, median imp. Might work better for non-normal distributions

### Amount of Missing Data

- How much of data is missing?
- Will imputing data improve the quality of dataset or will it add more bias?
- Multiple imp. and KNN might better for large missingness.

### Form of Missingness

- Is data is **missing at random** or is there a **missingness pattern**.
- Domain knowledge and how would imputation affect downstream analysis.



- Nature of Data =
  - **type of data =** Ex: categorical, methods like mode imputation or hot-deck *(similar)* imputation … **continuous** methods like **mean** imputation or **regression** imputation.
  - **Distribution =** ex: **skewed** or with large **outliers** = **median imputation**.
- Amount of Missing Data:
  - If minimal (less than 5%), simple methods like mean imputation or hot-deck imputation .
  - If **significant** more **advanced techniques** like **multiple imputation** or **regression** imputation.
  - ***trade-off between accuracy and computational complexity***
- Form of Missingness:
  - = pattern or structure of the missing data..
  - If missing at random then *may be* a simple replacement or leverage other observed variables to determine
  - If **not missing at random** specialized techniques like pattern mixture models or selection models.
- **QQ: Yes/No: Do you think domain knowledge should always take precedence over automated methods in determining missingness patterns**

# Factors Influencing Imputation Method

## There isn't a definitive best approach

### Nature of Data

- Is data type is **continuous, ordinal or categorical**?
- Data **distribution**: For example, median imp. Might work better for non-normal distributions

### Amount of Missing Data

- How much of data is missing?
- Will imputing data improve the quality of dataset or will it add more bias?
- Multiple imp. and KNN might better for large missingness.

### Form of Missingness

- Is data is **missing at random** or is there a **missingness pattern**.
- Domain knowledge and how would imputation affect downstream analysis.



There is no single best method for imputing missing data. The choice depends on several factors:

- Nature of Data: Consider whether the variable is continuous, ordinal, or categorical. The distribution of the data also matters. For example, median imputation might be more suitable for skewed distributions.
- Amount of Missing Data: The proportion of missing values can affect which method to choose. For larger amounts of missing data, techniques like multiple imputation or KNN may be considered.
- Form of Missingness: It's important to understand whether data is missing at random or if there's a specific pattern. Domain knowledge helps assess how imputation might influence the rest of the analysis.

# Marking Imputed Data *(Best Practice)*

## Keep track of imputed data

Important to mark imputed data, for:

- Model Evaluation
- Data Quality Assessment
- Enabling Transparency of Dataset
- Error Identification

| ID | Name | Age | Age_imputed |
|---|---|---|---|
| 1 | Alice | 25.0 | 0 |
| 2 | Bob | 30.0 | 0 |
| 3 | Charlie | 26.0 | 1 |
| 4 | David | 28.0 | 0 |
| 5 | Eva | 22.0 | 0 |



Marking which data points have been imputed is a helpful practice. It allows us to:

- Evaluate how imputed values might impact the model.
- Assess the overall quality of the dataset.
- Maintain transparency by keeping track of modified values.
- Identify and troubleshoot errors more easily.

In the example on the right, the **Age_imputed** column indicates which values in the Age column were filled in. A value of **1** shows that the data was imputed, while **0** indicates the original data.




Thank you for completing this lesson and continuing your journey to develop your skills with us.