Imputation is the process of replacing missing data with substituted values. It's a critical step in data cleaning to maintain dataset integrity and enable subsequent analysis.

## 1. Imputation Methods for Numerical Data 🔢

These methods use measures of central tendency to estimate the missing value in a column. They are simple, fast, and best suited when the percentage of missing data is small (<5%) and the data is Missing Completely At Random (MCAR).

| Method | Theory | Best Use Case | Example |
| --- | --- | --- | --- |
| **Mean Imputation** | Replaces the missing value (NaN) with the **arithmetic average** of the non-missing values in that column. | When the data is **normally distributed** (symmetrical) and contains **few to no outliers**. | Replacing a missing Salary with the average salary of **$65,000**. |
| **Median Imputation** | Replaces the missing value with the **median** (the middle value of the sorted data). | When the data contains **significant outliers** or is **skewed**, as the median is less sensitive to extremes than the mean. | Replacing a missing House Price with the median price of **$300,000** to avoid distortion by a few multi-million dollar mansions. |
| **Mode Imputation** | Replaces the missing value with the **mode** (the most frequently occurring value). | Primarily used for **categorical** or **discrete numerical** data (e.g., number of children). | Replacing a missing Number of Bedrooms with the most common value, **3**. |

## 2. Imputation Methods for Time-Series Data ⌚

Time-series data, where order matters, often requires methods that preserve the temporal relationship of the data points.

| Technique | Theory | Best Use Case | Example |
| --- | --- | --- | --- |
| **Forward-Fill (FFILL)** | Replaces a missing value with the **last observed non-missing value** that came before it. Also known as **Last Observation Carried Forward (LOCF)**. | When the data is stable over short periods (e.g., stock price at a specific second is likely the same as the previous second). | If the stock price is missing at **10:01 AM**, use the price from **10:00 AM**. |
| **Backward-Fill (BFILL)** | Replaces a missing value with the **next observed non-missing value** that comes after it. Also known as **Next Observation Carried Backward (NOCB)**. | Useful when the data is known to change rapidly, and the immediate future value is a better estimate than the immediate past. Often used to fill gaps at the very start of a series. | If the temperature is missing at **1:00 PM**, use the temperature from **2:00 PM**. |
| **Interpolation** | Estimates the missing value based on a linear or non-linear relationship between the surrounding data points. | When a trend or smooth curve is expected, and abrupt jumps (like those created by FFILL/BFILL) are undesirable. | Using **linear interpolation** to estimate the missing speed reading between two time stamps. |

## 3. Advanced Imputation Methods (Machine Learning Models) 🤖

These techniques leverage the relationships between variables to predict missing values more accurately than simple statistical measures. They are computationally intensive but yield superior results, especially for data that is Missing At Random (MAR).

| Method | Theory | Principle |
| --- | --- | --- |
| **K-Nearest Neighbors (KNN) Imputation** | For a missing value in a feature, KNN finds the 'K' most similar complete observations (neighbors) based on other features and uses the mean (for numerical) or mode (for categorical) of those neighbors to impute the missing value. | **Proximity:** Values close to each other in the feature space are likely to be similar. |
| **Iterative Imputation** | Uses a regression model to estimate a missing value. It treats each feature with missing data as a dependent variable, and all other features as independent variables. The process repeats iteratively until the imputed values converge. | **Prediction:** Treats imputation as a prediction problem, leveraging feature correlations (e.g., predicting Income based on Education and Experience). |
| **MICE (Multiple Imputation by Chained Equations)** | A specific form of iterative imputation that generates multiple imputed datasets, analyzes each, and then combines the results. | **Uncertainty:** Accounts for the uncertainty of the imputation process, resulting in more robust standard errors and confidence intervals. |