Missing Value Imputation, Explained: A Visual Guide with Code Examples for Beginners

Samy Baladram

Source :
https://towardsdatascience.com/missing-value-imputation-explained-a-visual-guide-with-code-examples-for-beginners

What Are Missing Values?

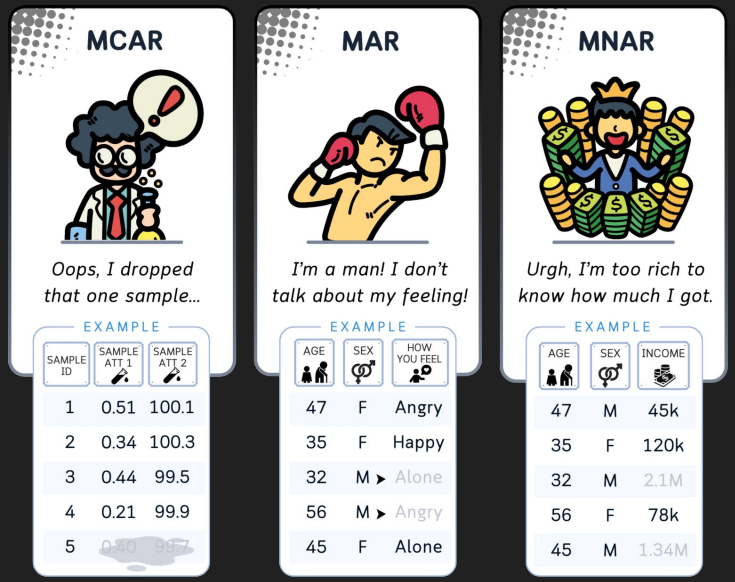
- Missing values, often represented as **NaN (Not a Number)** in pandas or **NULL** in databases
- They’re the **empty cells** in your spreadsheet, the **blanks** in your survey responses, the data points that got away.
- In the world of data, ***not all absences are created equal***, and ***understanding the nature of your missing values is crucial for deciding how to handle them***.

Why Do Missing Values Occur?

- **Data Entry Errors**: Sometimes, it's just ***human error***. Someone might forget to input a value or accidentally delete one.
- **Sensor Malfunctions**: In IoT or scientific experiments, a ***faulty sensor*** might fail to record data at certain times.
- **Survey Non-Response**: In surveys, respondents might ***skip questions*** they're uncomfortable answering or don't understand.
- **Merged Datasets**: When ***combining data*** from multiple sources, some entries might not have corresponding values in all datasets.
- **Data Corruption**: During ***data transfer*** or storage, some values might get corrupted and become unreadable.
- **Intentional Omissions**: Some data might be ***intentionally left out due to privacy*** concerns or irrelevance.
- **Sampling Issues**: The data collection method might systematically miss certain types of data.
- **Time-Sensitive Data**: In time series data, values might be missing for periods when data wasn't collected (e.g., weekends, holidays).

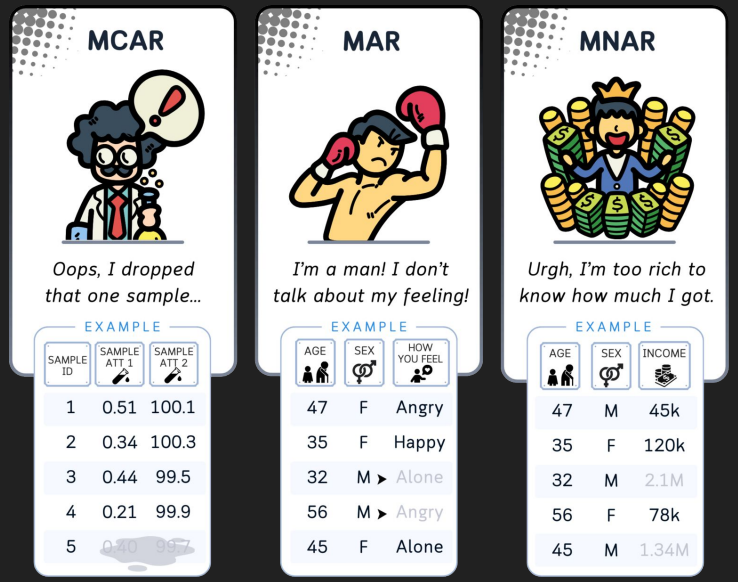
Types of Missing Data (1)

- **MCAR (Missing Completely at Random)**
  - Definition: The probability of a value being missing is independent of both observed and unobserved data.
  - Example:
    - A lab technician accidentally spills a blood sample, so that test result is missing.
    - A survey sheet is lost in the mail.
  - Implication: Analysis remains unbiased if we drop those rows (simple deletion is valid).



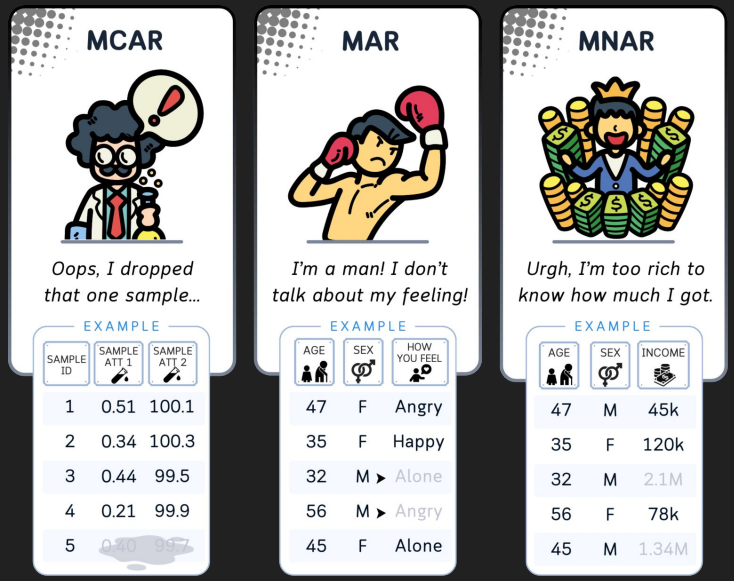
Types of Missing Data (2)

- **MAR (Missing at Random)**
  - Definition: The probability of missingness may depend on observed variables, but not on the missing value itself.
  - Example:
    - In a health survey, younger people are less likely to answer income questions → missingness depends on age (observed), not income itself.
    - Female respondents are more likely to skip a question about weight.
  - Implication: With appropriate modeling (e.g., multiple imputation using observed variables), unbiased estimates are possible.



Types of Missing Data (3)

- **MNAR (Missing Not at Random)**
  - Definition: The probability of missingness depends on the unobserved value itself.
  - Example:
    - People with very high income choose not to disclose their salary → missingness directly tied to the missing variable.
    - Patients with severe depression are less likely to answer a mental health survey item.
  - Implication: The hardest case — we can't correct with observed data alone; we need assumptions, sensitivity analysis, or explicit modeling of the missing-data mechanism.



Why Care About Missing Values?

- **Missing values can significantly impact your analysis:**
  - They can introduce bias if not handled properly.
  - Many machine learning algorithms can’t handle missing values out of the box.
  - They can lead to loss of important information if instances with missing values are simply discarded.
  - Improperly handled missing values can lead to incorrect conclusions or predictions.

The Dataset

Dataset

1	08-01	0	0	25.1	99	0		0.14
2	08-02	1	0	26.4		0		
3	08-03	2	0		96	0		0.21
4	08-04	3	0	24.1	68	0		0.68
5	08-05	4		24.7	98	0		0.2
6	08-06	5	0	26.5	98	0		0.32
7	08-07	6	0	27.6	78	0		0.72
8	08-08	0	0	28.2		0		0.61
9	08-09	1	0	27.1	70	1		
10	08-10	2	1	26.7	75			0.54
11	08-11	3	0			0		
12	08-12	4		24.3	77	1		0.67
13	08-13	5	0	23.1	77	1		0.66
14	08-14	6	0	22.4	89	1		0.38
15	08-15	0	0		80	1		0.46
16	08-16	1	0	26.5	88	0		
17	08-17	2	0	28.6	76	0		0.52
18	08-18	3	0					
19	08-19	4	0	27.0	73	1		0.62
20	08-20	5	0	26.9	73	0		0.81

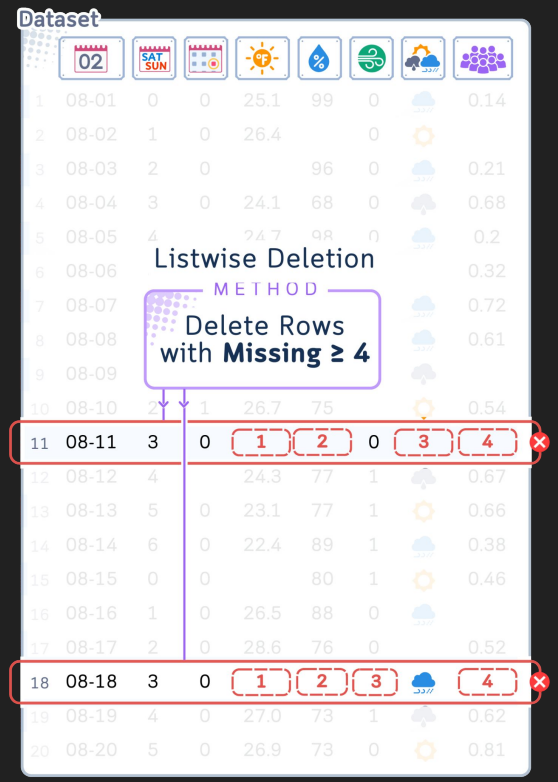
20 rows and 8 columns:

- Date: The date of the observation
- Weekday: Day of the week (0–6, where 0 is Monday)
- Holiday: Boolean indicating if it’s a holiday (0 or 1)
- Temp: Temperature in Celsius
- Humidity: Humidity percentage
- Wind: Wind condition (0 or 1, possibly indicating calm or windy)
- Outlook: Weather outlook (sunny, overcast, or rainy)
- Crowdedness: Percentage of course occupancy

This dataset is artificially made by the author (inspired by [1]) to promote learning.

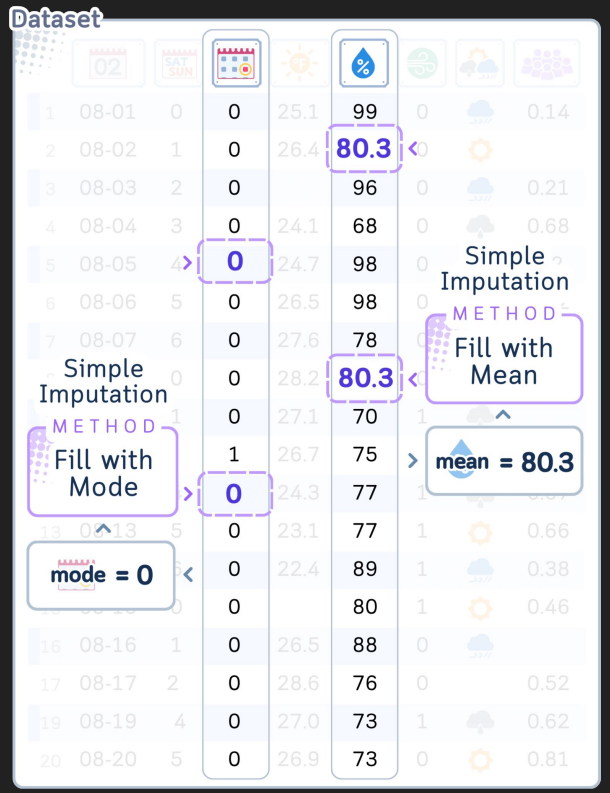
[1] T. M. Mitchell, Machine Learning (1997), McGraw-Hill Science/Engineering/Math, pp. 59

Method 1: Listwise Deletion



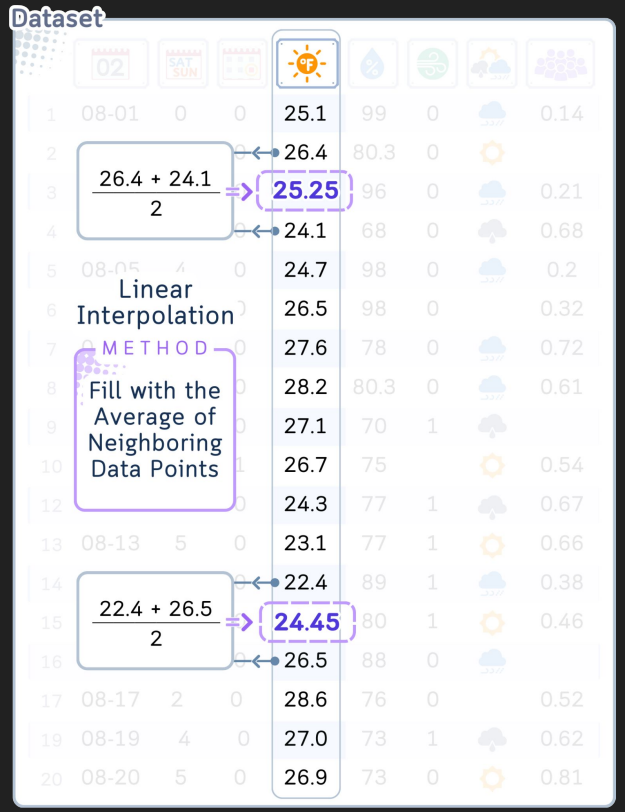
- 👍 **Common Use**:
  - Listwise deletion is often used when the number of missing values is small and the data is missing completely at random (MCAR).
  - It's also useful when you need a complete dataset for certain analyses that can't handle missing values.
- **In Our Case**:
  - We're using listwise deletion for rows that have at least 4 missing values.
  - These rows might not provide enough reliable information, and removing them can help us focus on the more complete data points.
  - However, we're being cautious and only removing rows with significant missing data to preserve as much information as possible.

Method 2: Simple Imputation – Mean and Mode



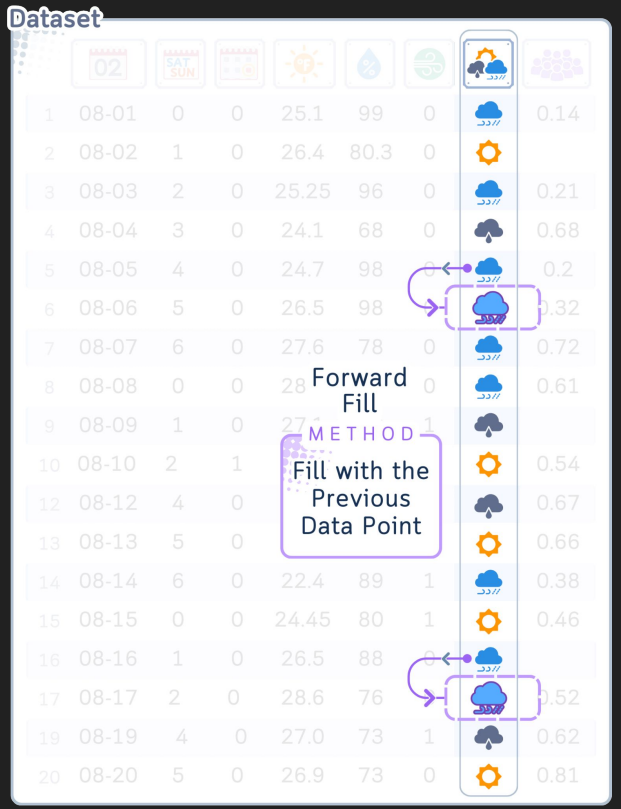
- 👍 **Common Use**:
  - **Mean imputation** is often used for continuous variables when the data is missing at random and the distribution is roughly symmetric. **Mode imputation** is typically used for categorical variables.
- **In Our Case**:
  - We're using mean imputation for Humidity and mode imputation for Holiday. For Humidity, assuming the missing values are random, the mean provides a reasonable estimate of the typical humidity. For Holiday, since it's a binary variable (holiday or not), the mode gives us the most common state, which is a sensible guess for missing values.

Method 3: Linear Interpolation



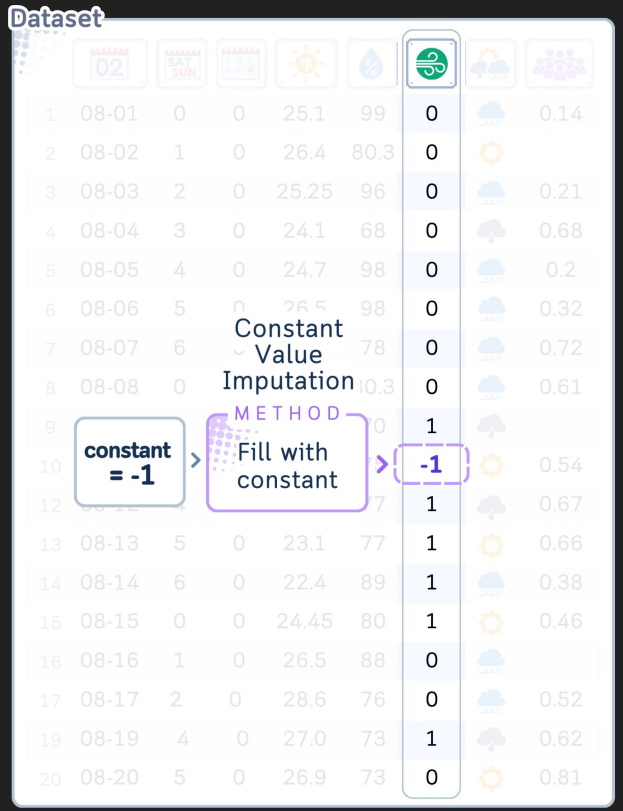
- 👍 **Common Use**:
  - **Linear interpolation** is often used for time series data, where missing values can be estimated based on the values before and after them. It's also useful for any data where there's expected to be a roughly linear relationship between adjacent points.
- **In Our Case**:
  - We're using linear interpolation for Temperature. Since temperature tends to change gradually over time and our data is ordered by date, linear interpolation can provide reasonable estimates for the missing temperature values based on the temperatures recorded on nearby days.

Method 4: Forward/Backward Fill



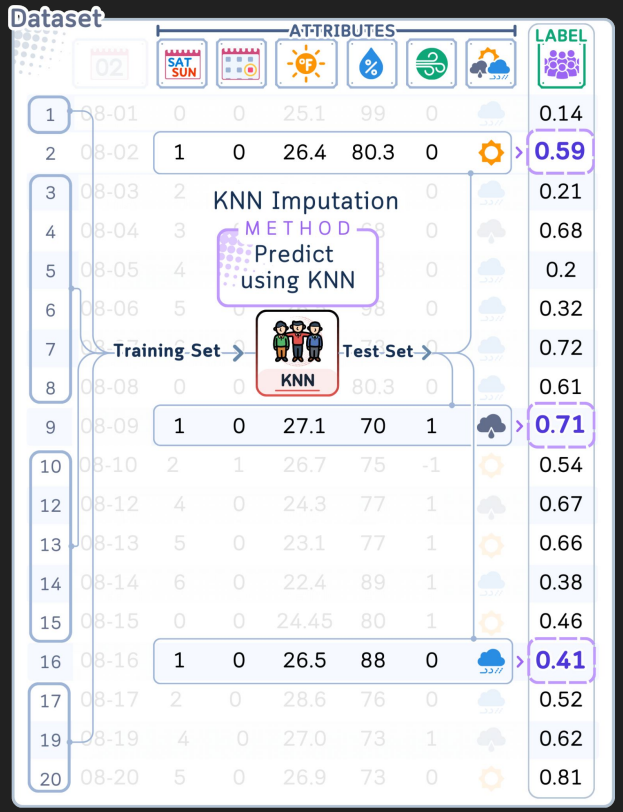
- 👍 **Common Use**:
  - **Forward/backward fill** is often used for time series data, especially when the value is likely to remain constant until changed (like in financial data) or when the most recent known value is the best guess for the current state.
- **In Our Case**:
  - We're using a combination of forward and backward fill for Outlook. Weather conditions often persist for several days, so it's reasonable to assume that a missing Outlook value might be similar to the Outlook of the previous or following day.

Method 5: Constant Value Imputation



- 👍 **Common Use**:
  - **Constant value imputation** is often used when there's a logical default value for missing data, or when you want to explicitly flag that a value was missing (by using a value outside the normal range of the data).
- **In Our Case**:
  - We're using constant value imputation for the Wind column, replacing missing values with -1. This approach explicitly flags imputed values (since -1 is outside the normal 0–1 range for Wind) and it preserves the information that these values were originally missing.

Method 6: KNN Imputation



- 👍 **Common Use**: ***KNN imputation*** is versatile and can be used for both continuous and categorical variables. It's particularly useful when there are expected to be complex relationships between variables that simpler methods might miss.
- **In Our Case**: We're using KNN imputation for Crowdedness. Crowdedness likely depends on a combination of factors (like temperature, holiday status, etc.), and KNN can capture these complex relationships to provide more accurate estimates of missing crowdedness values.

Conclusion: The Power of Choice (and Knowledge)

Dataset (Imputed)

	02	SAT SUN			%			
1	08-01	0	0	25.1	99	0		0.14
2	08-02	1	0	26.4	80.3	0		0.59
3	08-03	2	0	25.25	96	0		0.21
4	08-04	3	0	24.1	68	0		0.68
5	08-05	4	0	24.7	98	0		0.2
6	08-06	5	0	26.5	98	0		0.32
7	08-07	6	0	27.6	78	0		0.72
8	08-08	0	0	28.2	80.3	0		0.61
9	08-09	1	0	27.1	70	1		0.71
10	08-10	2	1	26.7	75	-1		0.54
12	08-12	4	0	24.3	77	1		0.67
13	08-13	5	0	23.1	77	1		0.66
14	08-14	6	0	22.4	89	1		0.38
15	08-15	0	0	24.45	80	1		0.46
16	08-16	1	0	26.5	88	0		0.41
17	08-17	2	0	28.6	76	0		0.52
19	08-19	4	0	27.0	73	1		0.62
20	08-20	5	0	26.9	73	0		0.81

- **Listwise Deletion**: Helped us focus on more complete data points by removing rows with extensive missing values.
- **Simple Imputation**: Filled in Humidity with average values and Holiday with the most common occurrence.
- **Linear Interpolation**: Estimated missing Temperature values based on the trend of surrounding days.
- **Forward/Backward Fill**: Guessed missing Outlook values from adjacent days, reflecting the persistence of weather patterns.
- **Constant Value Imputation**: Flagged missing Wind data with -1, preserving the fact that these values were originally unknown.
- **KNN Imputation**: Estimated Crowdedness based on similar days, capturing complex relationships between variables.

⚠️ Warning: The Purpose and Limitations of Missing Value Imputation

- **Not a Magic Solution**: Imputation is not a cure-all for missing data. It's a tool to make your data usable, **not to create perfect data**.
- **Potential for Bias**: Imputed values are educated guesses. They can introduce bias if not done carefully, especially if the data is Not Missing At Random (NMAR).
- **Loss of Uncertainty**: Most simple imputation methods don't account for the uncertainty in the missing values, which can lead to overconfident models.
- **Data Distortion**: Aggressive imputation can distort relationships in your data. Always check if imputation has significantly altered your data's distribution or correlations.
- **Document Your Process**: Always clearly document your imputation methods. This transparency is crucial for reproducibility and for others to understand potential biases in your results.

Reference

- https://towardsdatascience.com/missing-value-imputation-explained-a-visual-guide-with-code-examples-for-beginners
- https://pandas.pydata.org/docs/user_guide/missing_data.html#
- https://scikit-learn.org/stable/modules/impute.html