Handing Missing Data

What is Missing Data

Missing Data is a value that **isn't stored** for a variable in a dataset (NaN, N/A, or blank). It's different from zero — zero can be a valid recorded value, while missing data means the value was never captured.

Missing values can result from:
- Data entry errors
- Upload or system issues
- Survey questions left blank
- Data that doesn't apply to certain records

How to Handle Missing Data

Request the missing values be filled in
- Best option if possible.
- Contact data owners to gather the missing information.
- Example: Ask survey participants to complete unanswered questions.

Delete missing rows or columns
- Suitable when missing data is minimal and random.
- Avoid deleting if the missingness is intentional or widespread — it could bias results.

Create a new "NaN" category
- Useful for categorical data (e.g., responses like "Yes," "No," and "No Answer").
- Keeps missing data visible as its own group.

Fill in missing values with representative values
- Common for numerical data and forecasting.
- Methods include:
  - Forward fill: use previous value
  - Backward fill: use next value
  - Mean value: use average
  - Median value: use middle value

⚖️ Key Considerations

- Always think about the **ethical and analytical impact** of your choice.
- Ensure stakeholders know when missing data affects results.
- Discuss with peers or managers before deciding.
- Each dataset and situation requires **judgment, reasoning, and context**.

Missing data is a common issue in analytics that must be handled thoughtfully. Analysts should choose an appropriate strategy — filling, deleting, categorizing, or estimating — based on how much data is missing and how it affects the project. Communication with stakeholders, ethical awareness, and careful decision-making are essential to ensure data integrity and reliable insights.