**Data Manipulation**

**Data Preprocessing :** Handling missing data, normalization, standardization, feature engineering.

**Data Visualization :** Plotting Techniques, Exploratory Data Analysis(EDA)

**Data Preprocessing**

# **Handling Missing Data**

**Types of Missing Data**

1. **Missing Completely at Random (MCAR)**:
   * The missing values are independent of both the observed and unobserved data. The missing data has no systematic pattern.
   * Example: A respondent forgets to answer a question in a survey by accident.
2. **Missing at Random (MAR)**:
   * The missingness is related to the observed data but not the missing data itself. It means that the likelihood of a value being missing is related to other variables that are observed.
   * Example: A person with a low income may be less likely to report their salary in a survey, but the missingness is not related to the salary itself.
3. **Missing Not at Random (MNAR)**:
   * The missingness is related to the missing values themselves. There’s a pattern or reason that explains why the data is missing.
   * Example: People with high incomes might be more likely to omit reporting their salary.

**Methods of Handling Missing Data**

**1. Ignoring or Deleting Missing Data**

* **Listwise Deletion (Complete Case Analysis)**:
  + If a row (observation) contains missing data, the entire row is removed from the dataset.
  + Suitable when the amount of missing data is small and deleting them will not introduce bias.
  + **Disadvantage**: It can lead to loss of valuable information if many rows are missing data, and it may bias the results if the missingness is not random.
* **Pairwise Deletion**:
  + In pairwise deletion, instead of dropping rows completely, you only drop values from the analysis when a specific value is missing in a pair of variables involved in the analysis.
  + **Disadvantage**: Can lead to inconsistencies or imbalances in the dataset for different analyses.

**2. Imputation Methods**

* **Mean/Median/Mode Imputation**:
  + Replace missing values with the mean (for continuous data), median (for skewed data), or mode (for categorical data) of the respective feature/column.
  + **Disadvantage**: This can reduce variability in the data and introduce bias, especially if the missing data is not missing completely at random.
* **Forward or Backward Filling (for time series)**:
  + For time-series data, missing values can be replaced with the preceding value (forward fill) or the succeeding value (backward fill).
  + **Disadvantage**: It assumes the values don't change significantly over time, which might not always be true.
* **K-Nearest Neighbors (KNN) Imputation**:
  + Impute missing values by looking at the nearest neighbors (k-nearest data points) based on other features, and then taking the average (for continuous) or majority (for categorical) of the neighbors.
  + **Disadvantage**: Computationally expensive and works best when there are enough neighboring data points.
* **Regression Imputation**:
  + Predict the missing values based on other available features using regression models.
  + For continuous data, a linear regression model can be used, and for categorical data, a logistic regression model can be applied.
  + **Disadvantage**: Assumes a strong relationship between features, and if the relationship is weak, the imputed values may be inaccurate.
* **Multiple Imputation**:
  + Involves creating multiple versions of the dataset with different imputed values for the missing data. Each version is analyzed, and the results are combined to get a final estimate.
  + **Disadvantage**: Complex and computationally intensive but more robust as it accounts for the uncertainty of missing data.
* **Interpolation**:
  + For continuous variables, interpolation methods (like linear interpolation) estimate missing values based on other available data points.
  + **Disadvantage**: Assumes that missing values can be reasonably estimated from neighboring data, which may not always be the case.

**3. Predictive Modeling for Imputation**

* Use machine learning models (e.g., decision trees, random forests, etc.) to predict missing values based on available features. This approach can often provide more accurate imputations than simple statistical methods.
* **Disadvantage**: Requires complex models and more computational power.

**4. Using Domain-Specific Methods**

* In certain cases, domain knowledge can guide the best way to handle missing data. For example, in healthcare data, missing values for age or weight may be more appropriately imputed using domain-specific medical data.

**Handling Categorical vs. Numerical Missing Data**

1. **For Numerical Data**:
   * **Mean/Median Imputation**: Often used for numerical features when missing data is at random.
   * **Multiple Imputation**: When missing data is MAR, multiple imputation is a more robust method.
   * **KNN Imputation**: Can also work well for numerical data by considering the relationships between variables.
2. **For Categorical Data**:
   * **Mode Imputation**: Replaces missing categorical data with the most frequent category.
   * **KNN Imputation**: Works well for categorical features too, by looking for the most frequent value among the nearest neighbors.
   * **Regression Imputation**: For some categorical variables, logistic regression can be used to predict missing categories based on other features.

**Case-Specific Handling of Missing Data**

1. **When Missing Data is MCAR**:
   * Ignoring or deleting missing data is generally safe, as the missing data does not affect the analysis.
2. **When Missing Data is MAR**:
   * Imputation methods like KNN or regression-based imputation work well since the missingness is related to observed data, so imputing based on the observed data can still give a good approximation.
3. **When Missing Data is MNAR**:
   * Handling missing data becomes more complex. In such cases, multiple imputation or model-based techniques (such as Expectation-Maximization) may be used to model the missingness process and correct for biases.

**Evaluating the Impact of Missing Data**

* **Visualize Missing Data**: Tools like heatmaps or bar plots (e.g., missingno in Python) help to visualize patterns of missingness in the data, which can guide the appropriate handling method.
* **Model Testing**: It’s important to test different methods of handling missing data and evaluate their impact on the performance of the machine learning model. Cross-validation can help ensure the robustness of the method chosen.