EDA-Missing Value Treatment

Here is a summary of Missing Value Treatment in Exploratory Data Analysis (EDA) along with an example:

What are missing values?

Missing values, also known as missing data or NA (Not Available), occur when a value for a particular variable is not recorded or observed.

Why is it important to treat missing values?

In EDA, missing values can lead to incorrect conclusions and biased results. If left untreated, they can distort statistical analysis and machine learning models.

Types of Missing Value Treatment:

1. **Listwise Deletion**: Remove all observations with missing values.
    - Example: Consider a dataset with 100 rows (observations) and one column with missing values for 20 observations. Listwise deletion would remove these 20 observations from the analysis, leaving only 80 observations to analyze.
2. **Pairwise Deletion**: Remove only the observations that have missing values in a specific column or columns being analyzed.
    - Example: Consider a dataset with 100 rows and three columns (A, B, and C). Column A has missing values for 10 observations, but not for Columns B and C. Pairwise deletion would remove only those 10 observations from analysis when working with Column A.
3. **Mean/Median Imputation**: Replace missing values with the mean or median of the non-missing values in that column.
    - Example: Suppose we have a dataset with salaries (in thousands) for different employees, and some salaries are missing. Mean imputation would replace the missing values with the average salary of all employees (non-missing values).
4. **Regression Imputation**: Use a linear regression model to predict the value for an observation with missing data.
    - Example: Using historical sales data, we can create a regression model that predicts the sales amount based on other factors like seasonality and customer demographics.

Which method to use?

The choice of missing value treatment depends on:

1. **Type of analysis**: For exploratory analysis or machine learning models, pairwise deletion or mean/median imputation might be sufficient.
2. **Number of missing values**: If the number of missing values is small (less than 5% of total observations), listwise deletion might not significantly impact results.
3. **Nature of data**: For sensitive data like salaries or medical records, mean/median imputation could distort relationships between variables.

Remember to always document and justify your chosen method for treating missing values in EDA reports.

Curated by Brajesh Kumar