HANDLING MISSING VALUES IN DATA

Background

Missing values commonly occur in real-world datasets and can affect the quality of analysis and prediction. This project aims to apply proper handling techniques to ensure cleaner data and more accurate insights.

Steps to Handle Missing Value

01.

Copy the original dataset

the goal is to preserve the raw data

02.

Check the data types

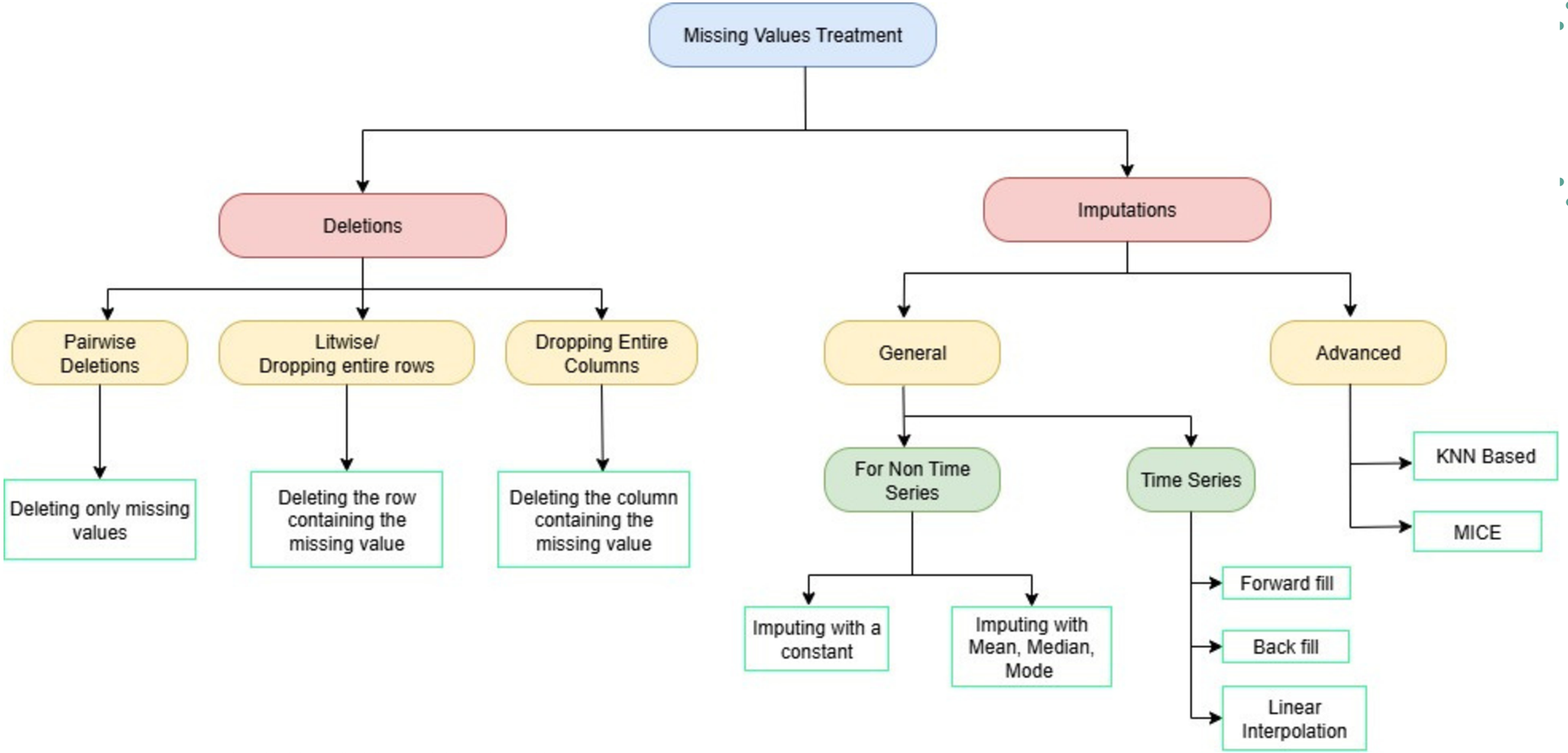
check the data in each column to ensure it is suitable for analysis

03.

Identify missing/null values

to handle data using functions like .isnull() or .info()

Missing Value Proportion	Recommended Handling
< 5%	Impute using mean/median/mode
5% - 30%	Consider imputation or advanced methods (KNN Based)
30% - 50%	Evaluate the importance of the column: drop if not essential, use model-based imputation if important
> 50%	Usually drop the column


Missing Values Treatment
Deletions
Imputations
Pairwise Deletions
Litwise/ Dropping entire rows
Dropping Entire Columns
General
Advanced
Deleting only missing values
Deleting the row containing the missing value
Deleting the column containing the missing value
For Non Time Series
Time Series
KNN Based
MICE
Imputing with a constant
Imputing with Mean, Median, Mode
Forward fill
Back fill
Linear Interpolation

Data Execution Stage

name	0.00
rating	1.00
genre	0.00
year	0.00
released	0.03
score	0.04
votes	0.04
director	0.00
writer	0.04
star	0.01
country	0.04
budget	28.31
gross	2.46
company	0.22
runtime	0.05

In this step, I handled missing values in the budget, gross, and rating columns. Appropriate imputation techniques were applied based on the proportion of missing data and the distribution characteristics of each column.

Missing Value Treatment per Column

Column Budget

The missing values in the budget column accounted for 5%–30% of the data. Since the distribution is right-skewed, the median was chosen as it is not affected by outliers. To provide more accurate imputation, I used grouped median imputation based on the genre column, considering that movie genre has a significant influence on production budget.

Column Gross

The missing values were less than 5% and the data was right-skewed, so I chose to impute the missing values using the median, as it is more robust against outliers.

Column Rating

This is categorical data and is best imputed using the mode, as it represents the most frequent value in the column.

The complete Python code for handling missing values can be viewed in my GitHub repository at the following link [GitHub link]

Thank you