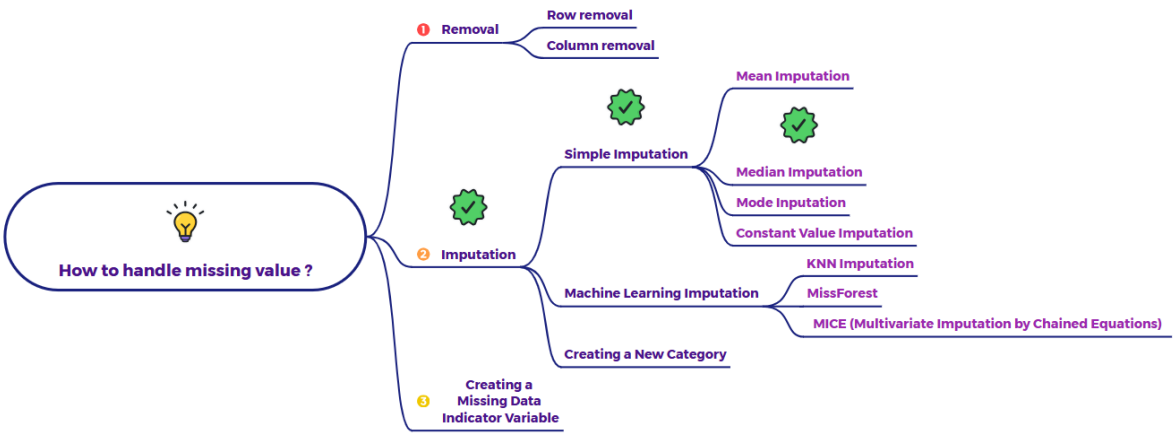
Explain Median Imputation with an example



What is Median Imputation?

Median imputation is another simple technique for handling missing numerical data. Instead of using the mean (average), it involves calculating the **median** of all the observed (non-missing) values for a particular numerical variable and then using that median value to replace all the missing values in that same variable.

How it Works:

1. **Identify the Numerical Variable with Missing Values:** Locate the column with missing numerical entries.
2. **Calculate the Median of the Observed Values:** For that column, find the middle value of all the non-missing values when they are ordered from smallest to largest. If there's an even number of observed values, the median is the average of the two middle values.
3. **Replace Missing Values:** Fill in all the missing values in that column with the calculated median.

Example:

Consider a dataset of house prices in a neighbourhood, and the "Square Footage" column has some missing entries.

Original Data:

House ID	Location	Square Footage
1	A	1500
2	B	2200
3	C	NaN
4	A	1800
5	B	NaN
6	C	1650
7	A	2500

Steps for Median Imputation on the "Square Footage" column:

1. **Identify the variable with missing values:** The "Square Footage" column has missing values (NaN).
2. **Calculate the median of the observed values:** The observed (non-missing) square footage values are 1500, 2200, 1800, 1650, and 2500.
    - First, order these values: 1500, 1650, 1800, 2200, 2500.
    - Since there are 5 observed values (an odd number), the median is the middle value, which is 1800.
3. **Replace missing values with the median:** We now replace the NaN values in the "Square Footage" column with 1800.

Data After Median Imputation:

House ID	Location	Square Footage
1	A	1500
2	B	2200
3	C	1800
4	A	1800
5	B	1800
6	C	1650
7	A	2500

Now, the missing "Square Footage" values have been filled in with the median square footage of the houses with recorded data.

When to Consider Median Imputation:

- **Robust to Outliers:** The median is less affected by extreme values (outliers) compared to the mean. This makes median imputation a better choice when the numerical variable might have skewed distributions or outliers.
- **Simple and Relatively Quick:** It's still a straightforward method to implement.
- **Skewed Distributions:** More appropriate than mean imputation for variables with skewed distributions, as the median is a better measure of central tendency in such cases.
- **Similar to Mean Imputation for Small Missingness:** If the amount of missing data is small, the impact might be less severe than in cases with substantial missingness.

Limitations and Cautions:

- **Reduces Variance (though often less than mean imputation):** Like mean imputation, it still introduces a central tendency value, potentially reducing the overall variance of the variable.
- **Can Distort Relationships:** It can still affect correlations with other variables, although potentially less so than mean imputation if the distribution is skewed.
- **Assumes MCAR (Implicitly):** While more robust to outliers in the observed data, it still implicitly assumes that the missingness is random and not related to the actual missing square footage or other variables. If the missingness is MAR or MNAR, it can introduce bias.

In situations where the numerical data is likely to have outliers or a skewed distribution, median imputation is often a preferable simple imputation method compared to mean imputation. However, for more accurate and robust handling of missing data, especially when the missingness is not likely MCAR, more advanced techniques are generally recommended.