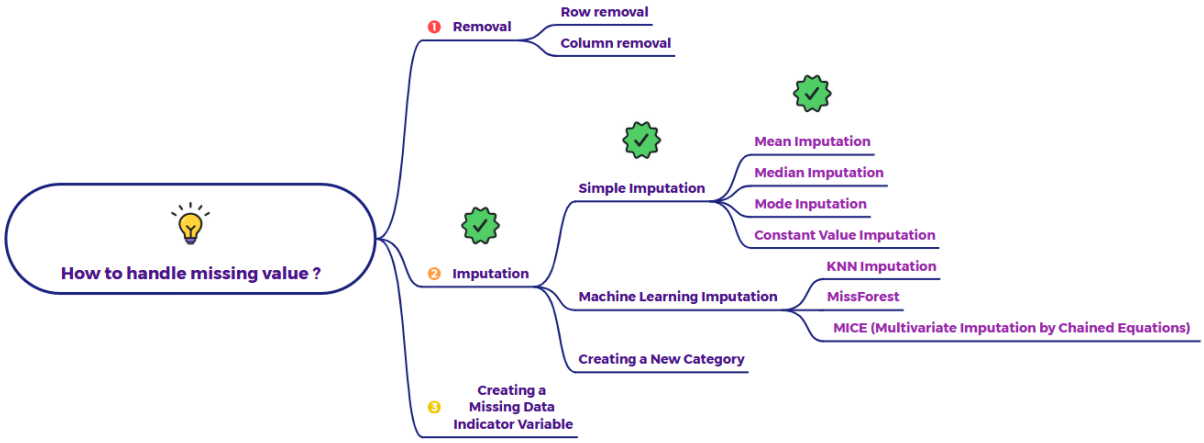
## Explain Mean Imputation with an example




### What is Mean Imputation?

Mean imputation is a simple technique for handling missing numerical data. It involves calculating the **mean (average)** of all the observed (non-missing) values for a particular numerical variable in your dataset and then using that mean value to replace all the missing values in that same variable.

### How it Works:

1. **Identify the Numerical Variable with Missing Values:** You first locate the column in your dataset that contains missing numerical entries (e.g., represented as NaN, None, or other placeholders).
2. **Calculate the Mean of the Observed Values:** For that specific column, you calculate the average of all the rows where the value is not missing.
3. **Replace Missing Values:** You then go through all the rows where the value for that column was missing and fill in the calculated mean value.

### Example:

Imagine you have a dataset of customer information, and one of the columns is "Age". Some customers didn't provide their age, resulting in missing values.

### Original Data:

| Customer ID | City | Age |
|---|---|---|
| 1 | Kolkata | 25 |
| 2 | Mumbai | 38 |
| 3 | Delhi | NaN |
| 4 | Chennai | 42 |
| 5 | Bangalore | NaN |
| 6 | Hyderabad | 30 |

### Steps for Mean Imputation on the "Age" column:

1. **Identify the variable with missing values:** The "Age" column has missing values (NaN).
2. **Calculate the mean of the observed values:** The observed (non-missing) ages are 25, 38, 42, and 30.
3. Mean Age = (25 + 38 + 42 + 30) / 4 = 135 / 4 = 33.75
4. **Replace missing values with the mean:** We now replace the NaN values in the "Age" column with 33.75.

### Data After Mean Imputation:

| Customer ID | City | Age |
|---|---|---|
| 1 | Kolkata | 25 |
| 2 | Mumbai | 38 |
| 3 | Delhi | 33.75 |
| 4 | Chennai | 42 |
| 5 | Bangalore | 33.75 |
| 6 | Hyderabad | 30 |

Now, all the missing "Age" values have been filled in with the average age of the customers who did provide their age.

### When to Consider Mean Imputation:

- **Quick and Simple:** It's a very easy method to implement.
- **Small Amount of Missing Data:** If the percentage of missing values is relatively low, the impact on the overall distribution might be minimal.

- **Rough Estimates:** When you need a quick way to fill in missing numerical values to allow algorithms that don't handle missing data to run.

**Limitations and Cautions:**

- **Reduces Variance:** Mean imputation compresses the distribution of the "Age" variable, as the imputed values are all the same. This can underestimate the true variability in the data.
- **Distorts Relationships:** It can weaken correlations between the "Age" variable and other variables in the dataset. The imputed values might not reflect the true relationship.
- **Sensitive to Outliers:** The mean itself can be heavily influenced by outliers. If there are extreme ages in the observed data, the imputed value might not be representative of the typical missing age.
- **Assumes MCAR (Missing Completely At Random):** Mean imputation implicitly assumes that the missing values are randomly distributed and not related to the actual missing age or other variables. If the missingness is MAR or MNAR, mean imputation can introduce bias.

In most real-world scenarios, especially when the amount of missing data is significant or the missingness is likely not MCAR, more sophisticated imputation techniques are generally preferred over simple mean imputation. However, it remains a basic and easily understandable method for initial data exploration or quick fixes.