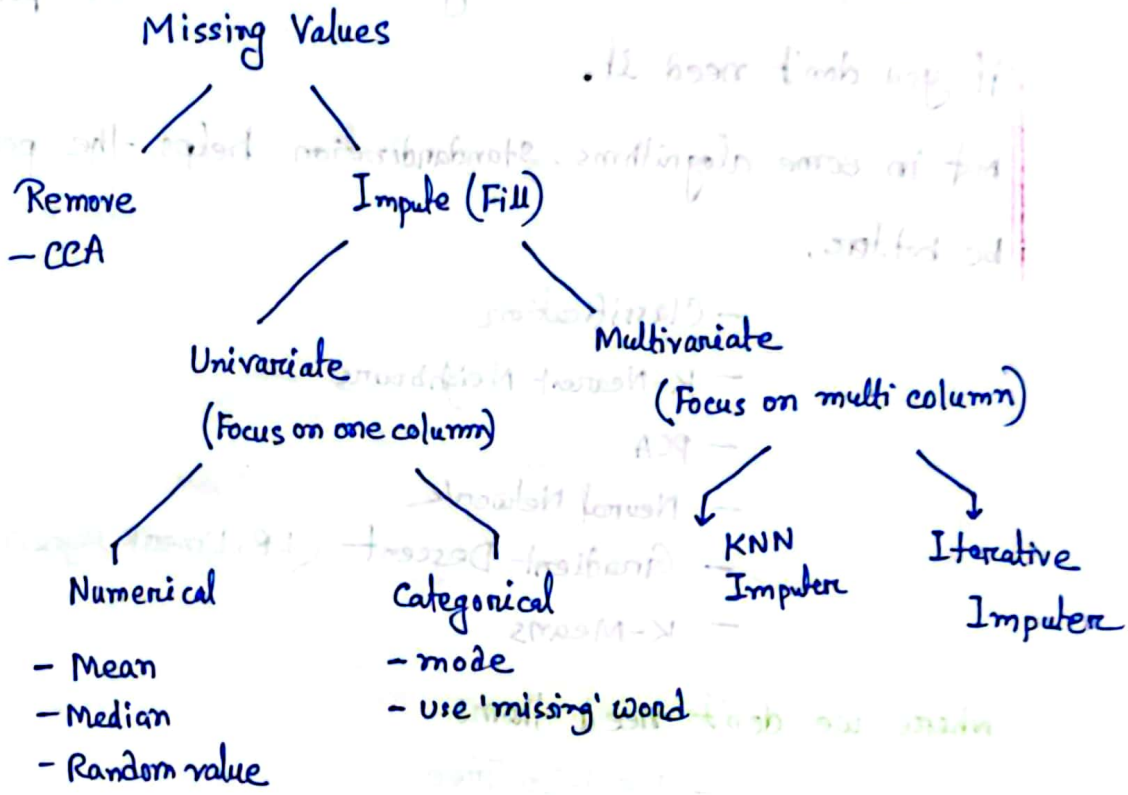
Handling Missing data:



Complete Case Analysis: (CCA)

It means analyzing only those rows for which there is information in all of the variables in the dataset.
(columns)

Here you discard the rows where values in any of the columns are missing.

- You can only do this, when the values that are missing are in random rows.
- Not like the first 50 rows, on middle 50 rows on last 50 rows.
- The values should be randomly missing.

– Because only if you delete the rows of the dataset at random, only then the distribution of the data remains same.

Advantage → 1) Easy to delete
2) Preserves the distribution

Disadvantage →
1) A large fraction of dataset can get deleted.
2) Much Info will be lost (Because the rows you are deleting many number of columns might have valuable information)
3) If you can't train your model with missing data, then the model will not know how to handle it.

When should it be used?

1) MCAR → (Column values should be missing at random rows)
2) Remove the rows or columns where maximum values are missing

How to check missing percentage in Pandas?

→ `df.isnull().mean() * 100`

3) Apply CCA on those columns where missing values at random rows are <5%

Handling Missing Numerical Data:

Techniques
- Mean, Median Imputations
- Arbitary value (~~Random val~~) (A particular value)
- End of distribution
- Random

Mean - Median Imputation: Already discussed

Disadvage → 1) Change the shape of the distribution
2) some extra outliers formed
3) Correlation chages with other columns

When to use?

1) MCAR (Data missing at random rows)
2) When missing data < 5%.

Arbitary Value Imputation:

Mostly we update categorical missing values with this technique.

- Suppose we change missing values to "missing" word.
- To use in numerical column, we can put any value that is not available in that column.

Benefit:

- Easy to apply

Disadvantage:

1) Graph (pdf) change shape
2) Variance changes
3) Correlation changes.

When to use?

- In this case, you can use this technique when data are "not randomly missing" (not randomly missing case)

End of Distribution Imputation:

- If your data is normally distributed, then use (mean + 3σ, mean - 3σ) to replace missing values
- If the data is skewed, then use (IQR Proximity)

$$Q_1 - 1.5\,IQR$$

or,

$$Q_3 + 1.5\,IQR$$

Using Arbitary Imputation or End of distribution we are presenting or we are making aware our model about the missing values.

Because using Arbitary Imputation, we are using such a value to replace missing value which is not available in the other rows. In End of distribution imputation we are replacing missing values by making them like outliers. So that ML Model can have a separate observation about them.

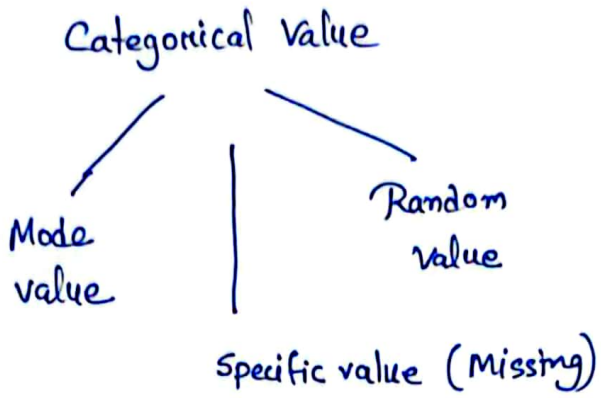
Advantage: - Easy to use

Disadvantage - Pdf, variance, correlation changes

When to use?

When your data of a particular column is not missing at random rows rather missing in rows like (Top 50 rows, middle 50, lower 50) something like that.

Handling Categorical Missing data:



Observation needed to use Mode:

- Data should be missing at random rows (MCAR)
- Mode value should be much higher than the other values.counts.

When your missing data >20% : (>10%)

- Can you specific values ("missing") to replace.
- So that your column will have now new category - "Missing".
- Your model then can observe this.