Day-9
Handling Missing data

1. Removing Missing Data

Removing using CCA(Complete Case Analysis):

- Discard row where values in any columns are missing.

Assumption:

- MCAR (Missing Completely at Random) i.e
  Distribution of data before and after should be similar
- Missing data should be less than 5%.

How to check MCAR??

- Distribution of data before and after should be similar
- Ratio of data before and after should be similar

2. Imputation

Univariante

(i) Mean/median:

Fill missing values with the mean, median, or mode of the column.

(ii) Arbitrary value imputation:

Arbitrary Value Imputation involves filling missing data with a specific constant value that is chosen based on the context or specific needs of the analysis.

Eg. -1,0 ,100

(iii) End of distribution imputation:

- For normal distribution

  (mean + 3*σ) and (mean - 3*σ)

- Skewed distribution

  For left skewed → (Q1 - 1.5 * IQR)

  For right skewed → (Q1 + 1.5 * IQR)

(iv) KNN Imputer

KNN Imputer estimates missing values based on the average or most frequent values from the nearest neighbours.

(v) MICE

MICE creates multiple imputed datasets by modelling and imputing missing values iteratively, then averages the results for robust imputation.