**Handling Missing Data**

Options for handling missing data:

1. Removing the entire row (Not prefer able) called CCA.
2. Imputing (Filling the Values)
   1. Univariant (Using Simple imputer class)
      1. Numerical
         1. Mean or Median value
         2. Random value
         3. End of distribution value
      2. Categorical
         1. Mode value
         2. Write (Missing)
   2. Multivariant
      1. KNN imputer
      2. Iterative imputer (MICE)

**CCA (Complete case analysis):** also called list-wise deletion of cases, consists in discarding observation where values in any variables are missing. Complete case analysis means literally analyzing only those observation for which there is information in all of the variables in the database.

**When to Use CCA: MCAR, 5% < data missing**

Complete case analysis, also known as listwise deletion, is a method used in statistics and data analysis when dealing with missing data. It involves excluding cases (rows) that have missing values for any of the variables of interest. While complete case analysis has its advantages, it also comes with several disadvantages. Let's explore them:

**Advantages of Complete Case Analysis:**

**Simplicity**: Complete case analysis is straightforward and easy to implement. It involves removing incomplete cases from the dataset, making it a simple approach for dealing with missing data.

**Preserves sample size**: Since only the incomplete cases are removed, the analysis retains all the complete cases, ensuring that the sample size remains unchanged.

**Preserves original data structure**: By retaining only the complete cases, the original data structure is maintained without imputing or estimating missing values, avoiding potential distortion.

**Compatible with many statistical methods**: Complete case analysis is compatible with various statistical techniques, such as regression analysis and hypothesis testing, as it does not alter the data distribution or introduce bias due to imputation.

**Disadvantages of Complete Case Analysis**:

**Reduced sample size**: One of the most significant drawbacks of complete case analysis is that it reduces the sample size by excluding incomplete cases. This can lead to a loss of statistical power and may limit the generalizability of the findings.

**Potential bias**: Complete case analysis can introduce selection bias if the missingness of data is related to the outcome or any of the variables being analyzed. In such cases, the results may be biased and not representative of the true population.

**Loss of information**: By removing cases with missing values, valuable information from those cases is discarded, which may lead to an incomplete understanding of the underlying data patterns.

**Assumption of Missing Completely at Random (MCAR):** Complete case analysis assumes that the missing data are missing completely at random, meaning that the probability of missing data is unrelated to the observed or unobserved data. This assumption is often unrealistic and challenging to verify.

**Impact on statistical power and precision**: Removing incomplete cases can reduce the precision and efficiency of statistical estimates, as it reduces the available information for analysis.

**Inefficient for large amounts of missing data**: When there is a substantial amount of missing data, complete case analysis may result in a severe reduction of the dataset, potentially rendering the analysis impractical or unreliable.