**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.

**Handling Missing Data (Numerical Data | Simple Imputer)**

Numerical Data imputation

1. Univariant imputation
2. Multivariant imputation
   1. KNN imputer
   2. Iterative imputer

**Univariant imputation**

Technique for numerical imputation:

1. **Mean | Median:** we have to use this technique when data is normally distributed. Use it when MCAR and missing values < 5% It is simple to use. If you have more than 5% missing values than its not reliable. Disadvantage of this is it changes the shape of distribution. It also changes the correlation.
2. **Arbitrary Value imputation:** In this technique you replace the missing value with any value like missing. Its mostly used in categorical data. It is easy to apply. It changes the covariance, correlation and PDF graph. *This technique is used when data is missing not at random.*
3. **End of distribution imputation:** In this technique you replace the missing value with at the end of your distribution. For normal distribution you replace with (mean + 3sigma or mean – 3sigma). And for skewed data you replace with (lower fence and higher fence values of IQR). Use it when data is not missing at random.

Technique for categorical imputation:

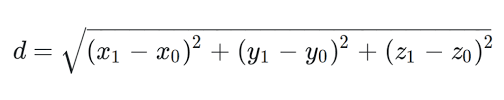
1. **Most frequent imputation**: replace the missing values with mode of data. The mode of the data should be most with respect to others. It is easy to implement. It changes the distribution of data.
2. **Missing imputation**: if values are missing more than 10% than make a new category called missing and replace the empty values with missing category.

**Random imputation**: *In this technique you fill the missing values with random numbers gathered from already available in other rows of the same feature*. It can be applied on both numerical and categorical. It is useful because it preserves the variance of the variables. It is memory heavy for deployment as we need to store the original training set to extract values from and replace the NA in coming observation. *It is well suited for linear models as it does not distort the distribution, regardless of the % of NA.*

**Missing Indicator**: In this technique you have to create a new column and indicate that either this row has missing value or not. True for missing value and false for no missing value.

**Automatic select value for imputation:** You use grid search CV technique here scikit automatically try all combinations and gives you best result for imputation.

**Multivariate Imputation:**

**KNN imputer: KNNimputer** is a scikit-learn class used to fill out or predict the missing values in a dataset. It is a more useful method which works on the basic approach of the KNN algorithm rather than the naive approach of filling all the values with mean or the median. In this approach, we specify a distance from the missing values which is also known as the K parameter. The missing value will be predicted in reference to the mean of the neighbours. It is implemented by the **KNNimputer()**method.****

Here x,y,z are the values of rows of the features for which we want to fill the value.

Nan Euclidean distance = dist(x,y) = sqrt(weight \* sq. distance from present coordinates) where, weight = Total # of coordinates / # of present coordinates

**Iterative Imputer:** Also called MICE Multivariate Imputation by chained Equations. There are some assumptions for MICE. MCAR, MAR, MNAR better result will be in MAR missing at random.

**Steps for MICE.**

Step 1. Fill all the NaN values with the mean of respective col.

Step 2. Remove all col1 missing values.

Step 3. Predict the missing values of col1 using other cols.

Step 4. Remove all col2 missing values.

Step 5. Predict the missing values of col2 using other cols.

Do the same for all other cols.