Handling Numerical Missing Data.

If there are missing values in your data, you actually have two options. One option is to simply remove the rows with missing values, and the other way is to impute them, filling in the missing values. There are two techniques for filling in missing values: one is univariate and the other is multivariate. Now, in univariate techniques, there are different methods depending on the type of column you have, whether it's numerical or categorical.

For numerical columns, we typically have the option to fill them with the mean, median, random value, or end of distribution. In categorical columns, we can impute missing values with the most frequent value, also known as the mode, or by filling in missing values with a specific value.

In scikit-learn, there is a class called SimpleImputer which handles these tasks. I'll describe it briefly. All the things I've mentioned can be done manually or using the SimpleImputer class.

Now, let's move on to multivariate techniques. I'll cover KNN Imputer and Iterative Imputer (MICE) and the whole structure I have defined for you is also called complete case analysis.

(Assumption for CCA) When will we perform CCA?

When your data is missing completely at random, such as having 4 columns and a thousand rows where one column represents age and 50 values are missing in the age column, if you apply CCA (Complete Case Analysis), you will remove those 50 missing values. After this transformation, the shape of your data will become 950 rows while the columns remain the same. You'll only delete missing values when you're certain that the 50 missing values are randomly distributed throughout the dataset. This means that it's possible that the first 50 values are missing, or the last 50, or perhaps the first 25 and the last 25 are missing. In such a setup, you won't remove any data, which implies there's an error. Instead, you'll perform CCA only when you're sure that the data you've removed consisted of randomly missing values (MCAR - Missing Completely at Random).

CCA is somewhat of a big flop because you apply it, deploy it on a server, and when data comes from the server to your model, your model doesn't know how to handle missing values. If the incoming data from the server contains missing values, it becomes a major issue.

You will apply CCA when your column has at least 5% of missing values at Random places.

Mean/Median Imputation:

- How to apply:
    - Replace missing values with the mean (average) or median of the column.
- When to use:
    - It's useful when the missing data is assumed to be randomly distributed across the dataset.
    - It's suitable for numerical data where the distribution is not heavily skewed.
- Benefits:
    - Simple and quick to apply.
    - Maintains the overall mean/median of the dataset.
    - Doesn't distort the distribution of the data significantly.

Arbitrary Value Imputation:

- How to apply:
    - Replace missing values with a pre-defined arbitrary value (e.g., 0, -999, etc.).
- When to use:
    - It's applicable when missing values hold some meaning or when mean/median imputation is not appropriate.
    - It's useful when the missingness itself might have some significance in the data.
- Benefits:
    - Preserves the information about the missingness.
    - Simple to implement.

Arbitrary value imputation, while simple to implement, has several disadvantages. It can distort data statistics and relationships, leading to misinterpretation. Loss of information occurs as no real data is utilized. The chosen arbitrary value might not be well-documented, causing confusion. This method is not suitable for categorical variables and may introduce bias, especially if missingness is related to the variable being imputed or if there is systematic missingness. Overall, caution should be exercised when using arbitrary value imputation due to its potential to affect data integrity and analysis outcomes

End of Distribution Imputation:

- How to apply:
    - Replace missing values with a value at the far end of the distribution (e.g., the mean plus three times the standard deviation).
- When to use:
    - It's suitable when the data has outliers or the distribution is skewed.
    - It assumes that missing values are not missing completely at random, but are rather at the end of the distribution.
- Benefits:
    - It reduces the impact of outliers.
    - Works well when missingness is related to the extremes of the distribution.

When to Apply:

- Mean/Median Imputation: When missing data is random and the variable's distribution is not heavily skewed.
- Arbitrary Value Imputation: When the missingness itself might hold significance or when other methods are not suitable.
- End of Distribution Imputation: When data has outliers or when missingness is likely related to extreme values.