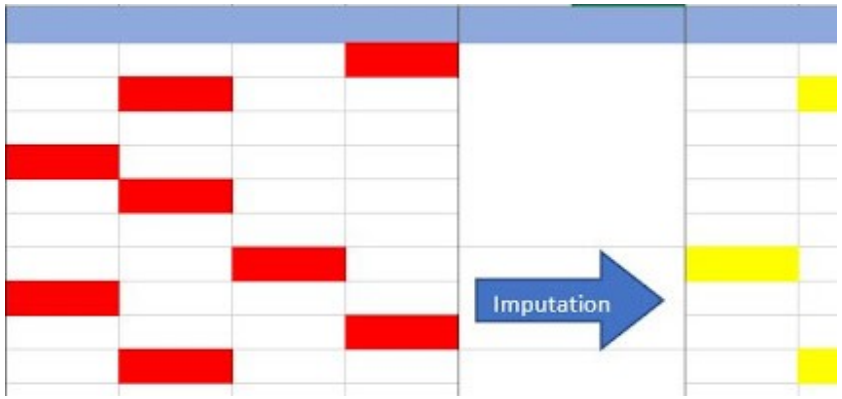
MISSING VALUES IMPUTATION

What is Imputation?

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset.

These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.



we have increased the column size, which is possible in Imputation(Adding “Missing” category imputation).

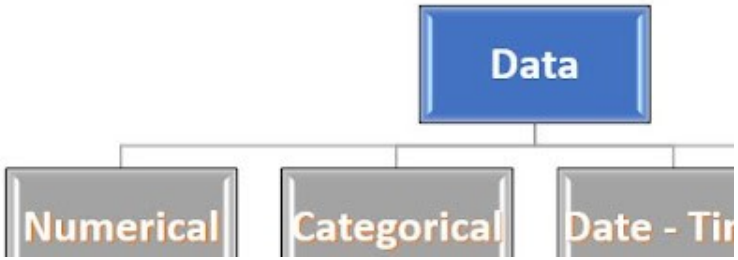
Why Imputation is Important?

We use imputation because Missing data can cause :

1.Incompatible with most of the Python libraries:- While using the libraries for Machine learning (the most common is sklearn), they don't have a provision to automatically handle these missing data and can lead to errors.

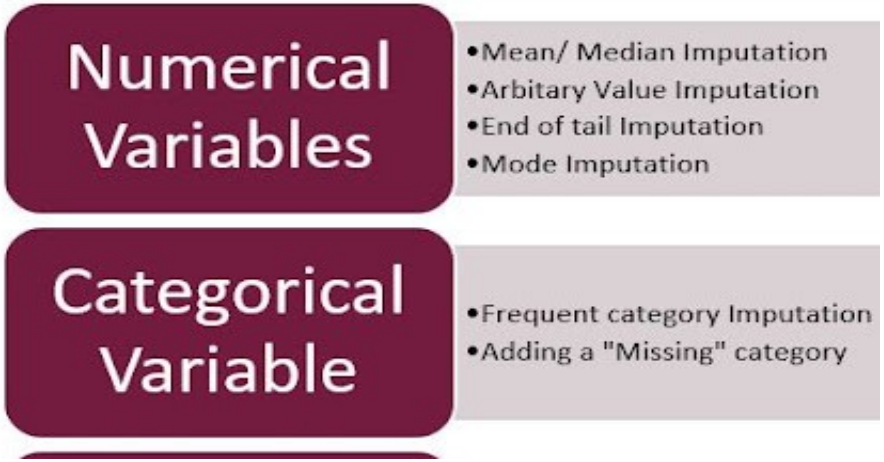
2.Distortion in Dataset:- A Huge amount of missing data can cause distortions in the variable distribution i.e it can increase or decrease the value of a particular category in the dataset.

3.Affects the Final Model:- The missing data can cause a bias in the dataset and can lead to a faulty analysis by the model.



Most data is of Four types

Imputation Techniques



taroonreddy.com

1.Complete Case Analysis(CCA):-

This is a quite straightforward method of handling the Missing Data,
which directly removes the rows that have missing data i.e we consider only those rows where, we have complete data i.e data is not missing. This method is also popularly known as **"Listwise deletion".**

- **Assumptions:-**
  - Data is Missing At Random(MAR).
  - Missing data is completely removed from the table.

- **Advantages:-**
  - Easy to implement.
  - No data manipulation is required.

- **Limitations:-**
  - Deleted data can be informative.
  - Can lead to the deletion of a large part of the data.
  - Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.
  - The production model will not know what to do with Missing data.

- **When to Use:-**
  - Data is MAR(Missing At Random).
  - Good for Mixed, Numerical, and Categorical data.
  - Missing data is not more than 5% – 6% of the dataset.
  - Data doesn't contain much information and will not bias the dataset.

2. Arbitrary Value Imputation :-

This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables.
This technique states that we group the missing values in a column and assign them to a new value
that is far away from the range ofthat column.

Mostly we use values like **99999999 or -9999999** or **“Missing” or “Not defined” for numerical & categorical variables.**

- **Assumptions:-**
  - Data is not Missing At Random.
  - The missing data is imputed with an **arbitrary value that is not part of the dataset** or **Mean/Median/Mode of data.**

- **Advantages:-**
  - Easy to implement.
  - We can use it in production.
  - It retains the importance of “missing values” if it exists.

- **Disadvantages:-**
  - Can distort original variable distribution.
  - Arbitrary values can create outliers.
  - Extra caution is required in selecting the Arbitrary value.

- **When to Use:-**
  - When data is not MAR(Missing At Random).
  - Suitable for All.

3. Frequent Category Imputation :-

This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as **Mode Imputation.**

- **Assumptions:-**
  - Data is missing at random.
  - There is a high probability that the missing data looks like the majority of the data.

- **Advantages:-**
  - Implementation is easy.
  - We can obtain a complete dataset in very little time.
  - We can use this technique in the production model.

- **Disadvantages:-**
  - The higher the percentage of missing values, the higher will be the distortion.
  - May lead to over-representation of a particular category.
  - Can distort original variable distribution.

- **When to Use:-**
  - Data is Missing at Random(MAR)
  - Missing data is not more than 5% – 6% of the dataset.