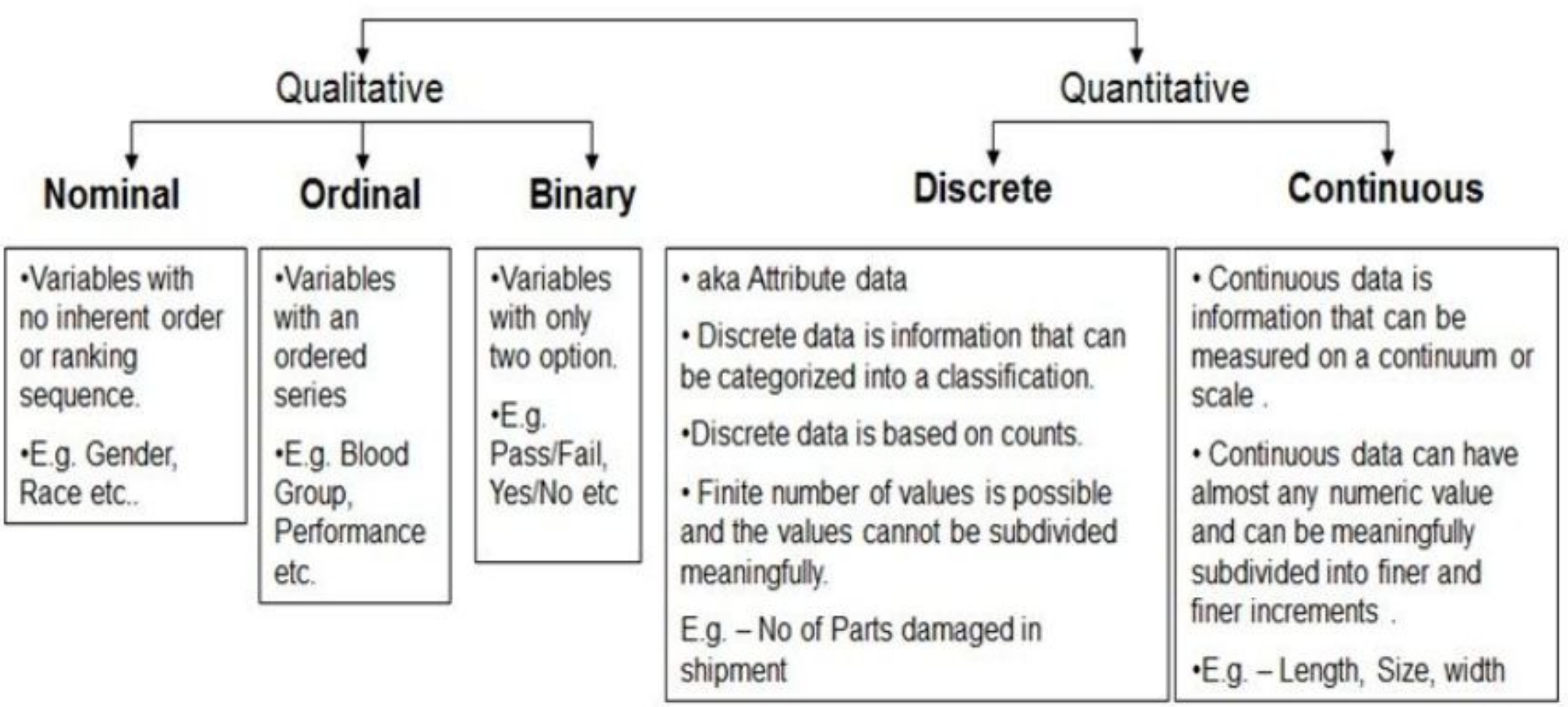
Data Visualization

Lecture 7

Data Preprocessing

Encoding and Imputation

Data Types


Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continuous
•Variables with no inherent order or ranking sequence.
•E.g. Gender, Race etc..
•Variables with an ordered series
•E.g. Blood Group, Performance etc.
•Variables with only two option.
•E.g. Pass/Fail, Yes/No etc
• aka Attribute data
• Discrete data is information that can be categorized into a classification.
•Discrete data is based on counts.
• Finite number of values is possible and the values cannot be subdivided meaningfully.
E.g. – No of Parts damaged in shipment
• Continuous data is information that can be measured on a continuum or scale .
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
•E.g. – Length, Size, width

Encoding Categorical Data

Data Encoding

- The performance of a machine learning model not only depends on the model and the hyperparameters but also on how we process and feed different types of variables to the model.

- Most machine learning models only accept numerical variables, preprocessing the categorical variables becomes a necessary step. We need to convert these categorical variables to numbers such that the model is able to understand and extract valuable information.

- A typical data scientist spends 70 – 80% of his time cleaning and preparing the data. And converting categorical data is an unavoidable activity. It not only elevates the model quality but also helps in better feature engineering.

Data Encoding

- What is categorical data?
- Categorical variables are usually represented as ‘strings’ or ‘categories’ and are finite in number. Here are a few examples:
- The city where a person lives: Delhi, Mumbai, Ahmedabad, Bangalore, etc.
- The department a person works in: Finance, Human resources, IT, Production.
- The highest degree a person has: High school, Diploma, Bachelors, Masters, PhD.
- The grades of a student: A+, A, B+, B, B- etc.

Data Encoding

- There are two kinds of categorical data:

- Ordinal Data: The categories have an inherent order
  - In Ordinal data, while encoding, one should retain the information regarding the order in which the category is provided.

- Nominal Data: The categories do not have an inherent order
  - encoding Nominal data, we have to consider the presence or absence of a feature. In such a case, no notion of order is present. For example, the city a person lives in.

Data Encoding

- Categorical Data Encoding Techniques:

- **Ordinal Data:**
  - Label Encoding
  - Ordinal Encoding

- **Nominal Data:**
  - One hot Encoding
  - Dummy Encoding
  - Frequency Encoding
  - Target Encoding

Data Encoding - Label Encoding or Ordinal Encoding

- We use this categorical data encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence.

- In Label encoding, each label is converted into an integer value.

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Data Encoding - Label Encoding or Ordinal Encoding

- Label encoding is used on alphabetically ordered data already (grades without + or -)

- Could also be used on nominal data in 2 cases:
  - X features are being used in tree-based models
  - Y feature

Data Encoding - One Hot Encoding

- We use this categorical data encoding technique when the features are nominal.
- For each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1 (0 represents the absence, and 1 represents the presence of that category)

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Data Encoding - One Hot Encoding

Advantages

- Does not assume the distribution of categories of the categorical variable.
- Keeps all the information of the categorical variable.

Disadvantages

- Not so Suitable for tree-based models.
- with statistical models such as linear regression an issue is caused (dummy variable trap or multicollinearity)

Data Encoding – Dummy Encoding

- Dummy coding scheme is similar to one-hot encoding.
- In the case of one-hot encoding, for N categories in a variable, it uses N binary variables. The dummy encoding is a small improvement over one-hot-encoding. Dummy encoding uses N-1 features to represent N labels/categories.

Column	Code
A	100
B	010
C	001

One- Hot Coding

Column	Code
A	10
B	01
C	00

Dummy Code

Data Encoding

- Drawbacks of One-Hot and Dummy Encoding
  - Expands the feature space.
  - Does not add extra information while encoding.

Data Encoding - Frequency Encoding

- It is a way to utilize the frequency of the categories as labels.

- Replace the categories by the count of the observations that show that category in the dataset.

Data Encoding - Frequency Encoding

Advantages	Limitations
• Straightforward to implement. • Does not expand the feature space. • Can work well with tree-based algorithms.	• Does not handle new categories in the test set automatically. • We can lose valuable information if there are two different categories with the same amount of observations count—this is because we replace them with the same number.

Data Encoding - Target Encoding

- replacing the category with the mean target value for that category. We start by grouping each category alone, and for each group, we calculate the mean of the target in the corresponding observations. Then we assign that mean to that category.

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

Color	Target Mean
Yellow	0.6
Blue	0.5
Red	0.66

Color	Target
0.6	0
0.6	1
0.5	1
0.6	1
0.66	1
0.6	0
0.66	1
0.66	0
0.6	1
0.5	0

Data Encoding - Target Encoding

Advantages	Limitations
• Does not expand the feature space. • Creates a monotonic relationship between categories and the target.	• May lead to overfitting. • May lead to a possible loss of value if two categories have the same mean as the target—in these cases, the same number replaces the original.

When to use each type ?

- **For most general machine learning tasks and low cardinality data (few unique values):** Start with **One-Hot Encoding**.

- **For classical statistics/econometrics linear models:** Use **Dummy Encoding** to avoid multicollinearity issues.

- **For high cardinality features (many unique values):** Explore **Frequency Encoding** (simple and fast) or **Target Encoding** (powerful but requires careful implementation to prevent overfitting).

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.

Why Imputation is Important?

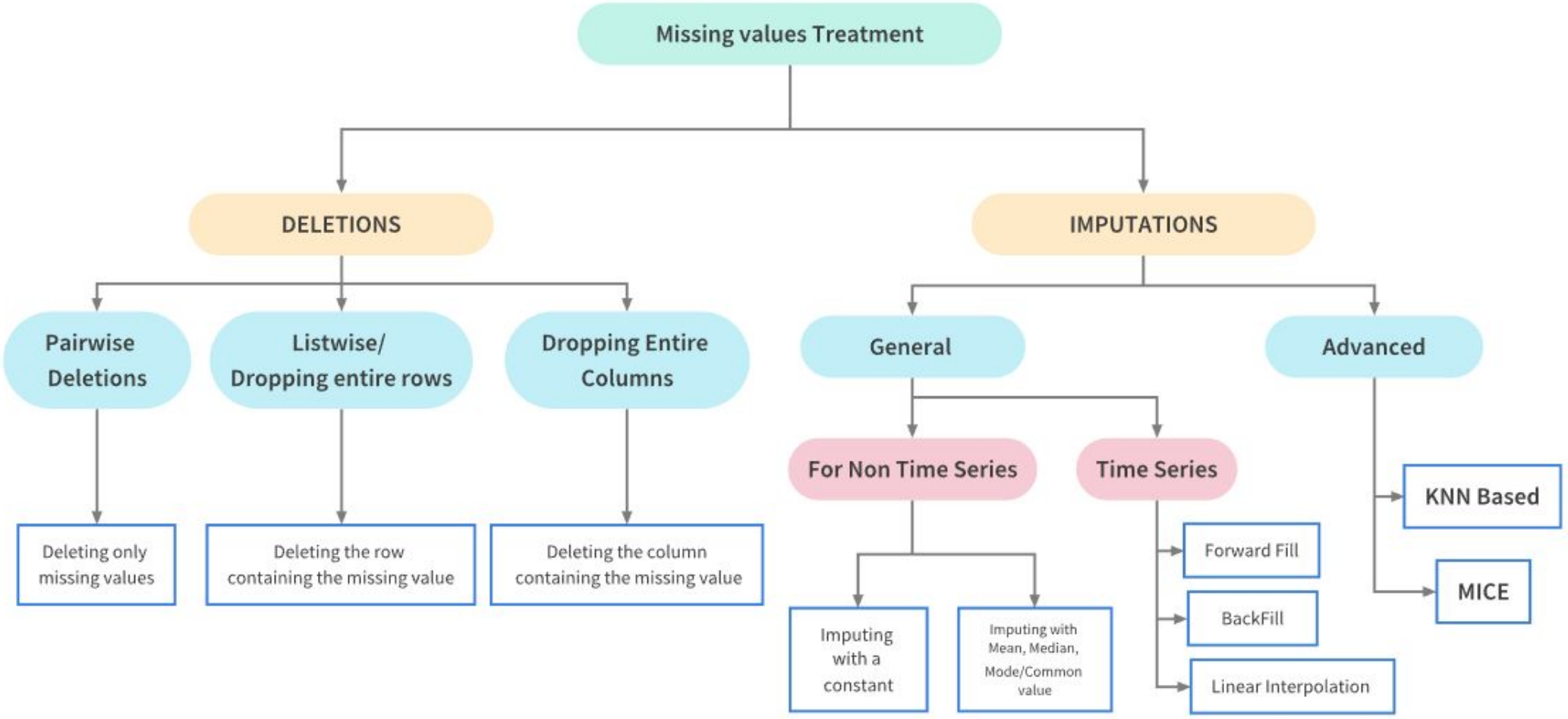
- We use imputation because Missing data can cause the below issues: –

  - **Incompatible with most of the Python libraries used in Machine Learning:-** Yes, you read it right. While using the libraries for ML(the most common is skLearn), they don't have a provision to automatically handle these missing data and can lead to errors.

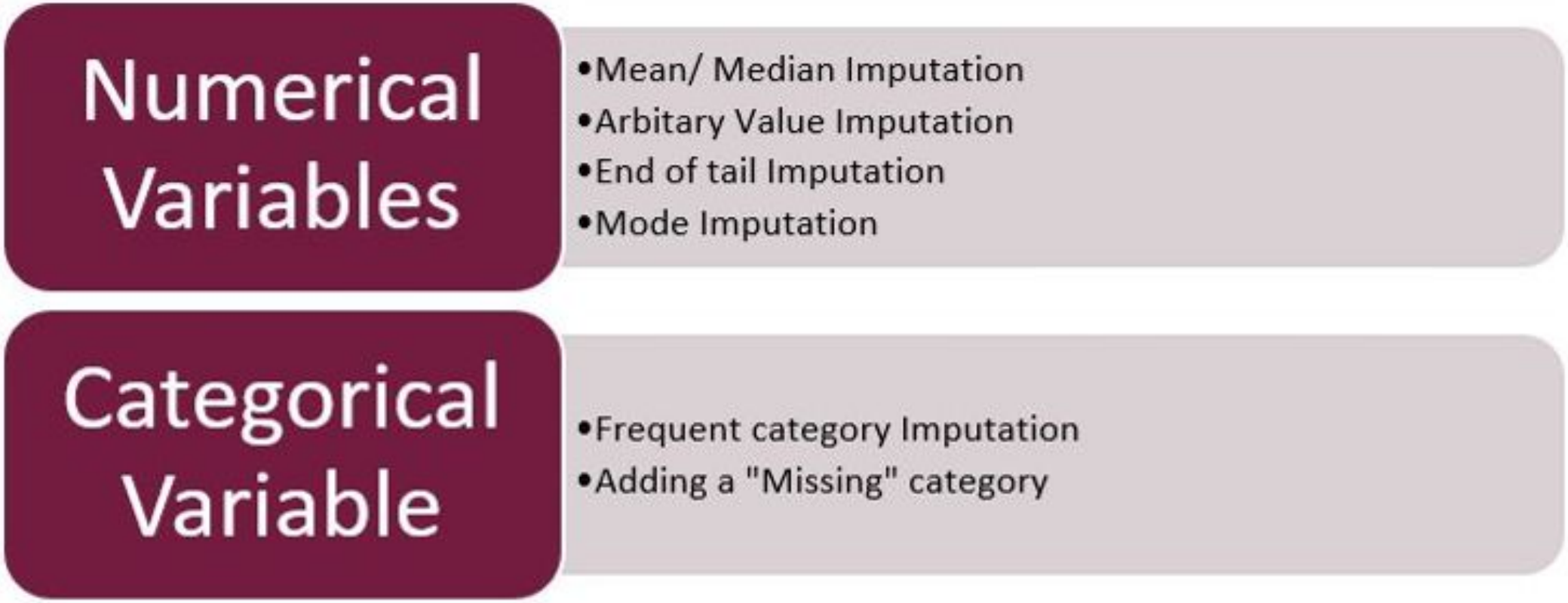
  - **Distortion in Dataset:-** A huge amount of missing data can cause distortions in the variable distribution i.e it can increase or decrease the value of a particular category in the dataset.

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

Handling Missing Values



Imputation Techniques



Types of Missing Data

- **Missing Completely At Random (MCAR)**
  - When the absence of data is completely unrelated to both the observed and unobserved data.
  - Example, missing values could occur due to technical issues like a system crash.
- **Missing At Random (MAR)**
  - When the probability of being missing is the same only within groups defined by the observed data.
  - Example, being missing is lower for younger people than for older people. the probability of missing income is related to the observed variable "age"
- **Missing Not At Random (MNAR)**
  - When the probability of missingness is related to the value of the missing data itself.
  - Example, Patients with severe symptoms are less likely to report their health status, leading to missing data that depends on their condition.

Imputation Techniques (Mean/Median)

- Mean/median imputation consists of replacing all occurrences of missing values (NA) within a variable by
  - the mean (if the variable has a Gaussian distribution)
  - or median (if the variable has a skewed distribution).

Imputation Techniques (Mean/Median)

Assumptions	Advantages	Limitations
Mean/median imputation has the assumption that the data are missing completely at random (MCAR).	• Easy to implement • Fast way of obtaining complete datasets	• Distortion of original variance • Distortion of covariance with remaining variables within the dataset

Imputation Techniques (Arbitrary/Constant Value)

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

Imputation Techniques (Arbitrary/Constant Value)

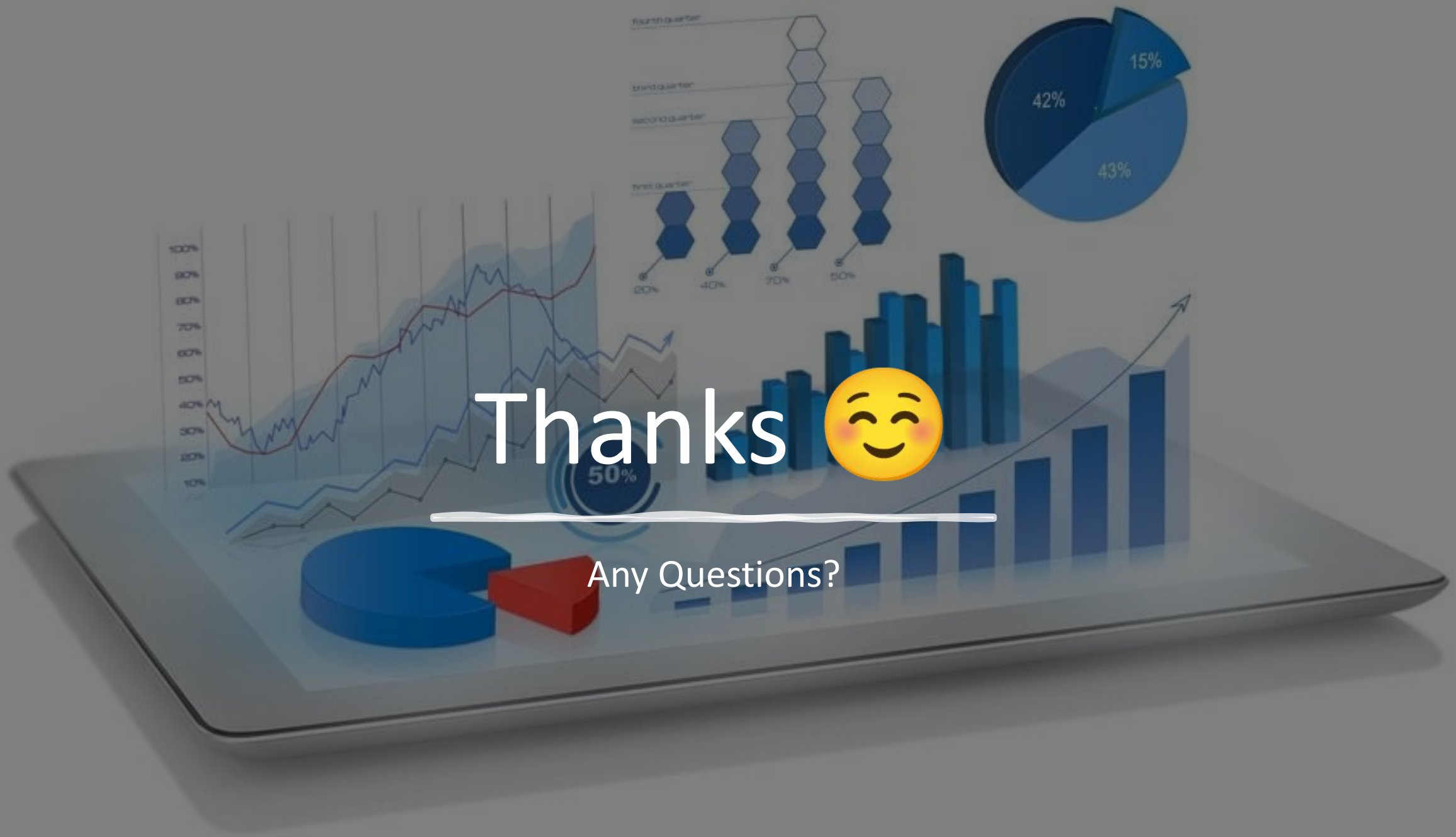
Assumptions	Advantages	Limitations
• Data is not Missing At Random(MNAR). • The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data.	• Easy to implement. • We can use it in production. • It retains the importance of “missing values” if it exists.	• Can distort original variable distribution. • Arbitrary values can create outliers. • Extra caution required in selecting the Arbitrary value.

Imputation Techniques (Mode/Frequent Category)

- 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 Techniques (Mode/Frequent Category)

Assumptions	Advantages	Limitations
• Data is missing at random (MAR). • There is a high probability that the missing data looks like the majority of the data.	• Implementation is easy. • We can obtain a complete dataset in very little time. • We can use this technique in the production model.	• 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.



Thanks 😊

Any Questions?