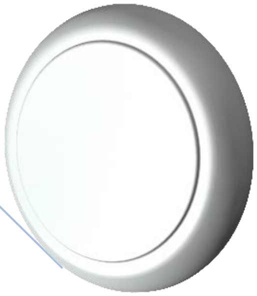


# Prepare the document for Handling missing values

How to handles the missing values by various Technologies

**Joysi Immacuate A**

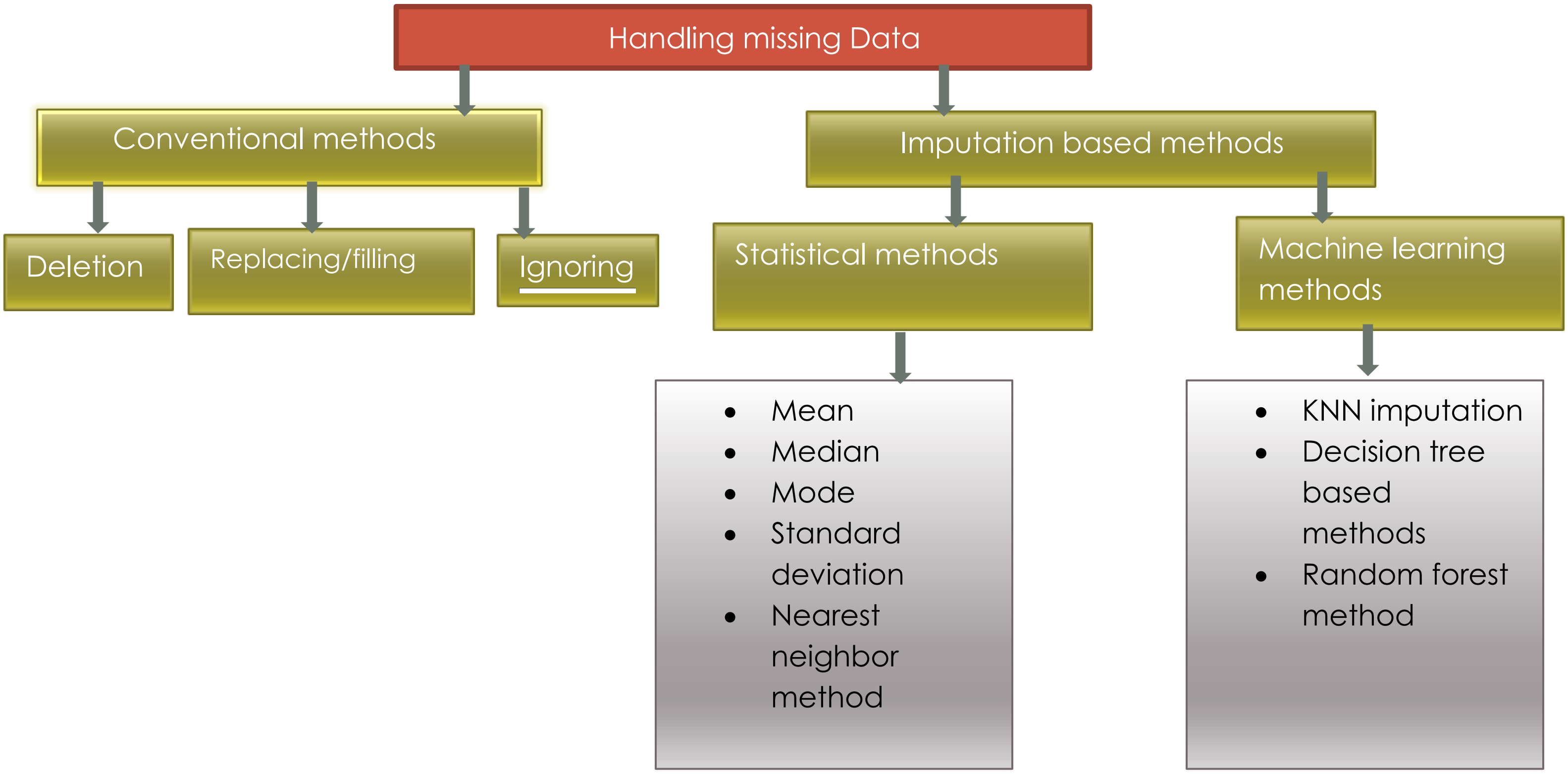
**6/29/2023**

## Introduction

Handling missing values is a crucial task in data analysis and machine learning. In any dataset, missing values can occur due to various reasons, such as data entry errors, incomplete data collection, or even deliberate omissions. These missing values can significantly impact the quality and accuracy of the analysis, as well as the reliability of any conclusions or predictions drawn from the data.

The goal of handling missing values is to minimize their impact on the analysis while maintaining the integrity of the data. This involves understanding the types and patterns of missingness, selecting suitable handling techniques, and ensuring that the chosen methods align with the specific characteristics of the dataset and the analysis objectives.

**Process of handling missing values:-**
**(I use below methods in my project work. It may enhance in future)**



**What could happen if we ignore the missing data:-**

- Biased results
- Loss of information
- Reduced sample size
- Distorted statistical measures
- False assumptions of independence
- Inaccurate predictions or classifications

## Various technologies to handles the missing values:-

- Python
- SQL
- Excel
- Power BI

### Python

**Pandas:** Pandas is a popular data manipulation library in Python that provides functions and methods to handle missing values. It offers functionalities like filling missing values, dropping rows or columns with missing values, and imputation techniques.
**NumPy:** NumPy is a fundamental library for scientific computing in Python. It provides various functions to work with missing values, such as identifying missing values, replacing them, or performing calculations while ignoring missing values.
**Scikit-learn:** Scikit-learn is a widely used machine learning library in Python. It includes modules for preprocessing data, which can be helpful in handling missing values through techniques like imputation or feature scaling.

```python
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer

# Load the data
df = pd.read_csv('data.csv')

# Identify missing values
missing_values = df.isnull()
missing_counts = df.isnull().sum()

# Dropping missing values
df_dropped = df.dropna()  # Drops rows with any missing values
df_dropped_cols = df.dropna(axis=1)  # Drops columns with any missing values

# Filling missing values
df_filled = df.fillna(value)  # Fills missing values with a specific value
df_filled_mean = df.fillna(df.mean())  # Fills missing values with the mean of each column
df_filled_median = df.fillna(df.median())  # Fills missing values with the median of each column
df_filled_mode = df.fillna(df.mode().iloc[0])  # Fills missing values with the mode of each column

# Imputation using SimpleImputer
imputer = SimpleImputer(strategy='mean')  # Can use 'median' or 'most_frequent' as well
df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)

# Handling categorical variables
df['categorical_column'].fillna('Unknown', inplace=True)

# Replacing specific values
```

```
df['column'].replace('incorrect_value', 'correct_value', inplace=True)

# Forward or Backward fill
df_ffill = df.ffill()  # Forward fills missing values using the last observed value
df_bfill = df.bfill()  # Backward fills missing values using the next observed value

# Handling missing values in time series data
df_interpolated = df['timestamp_column'].interpolate(method='time')  # Interpolates missing values in a time series column

# Working with dummy variables
df_dummies = pd.get_dummies(df, columns=['categorical_column'])  # Converts categorical columns into dummy variables

# Exporting the data
df.to_csv('cleaned_data.csv', index=False)  # Saving the cleaned data to a CSV file
```

## SQL

SQL databases, such as MySQL, PostgreSQL, or Microsoft SQL Server, offer capabilities to handle missing values in data sets.
SQL queries can be used to filter out or exclude rows with missing values, perform aggregations while ignoring missing values, or perform data transformations using functions like COALESCE or IFNULL.

### 1.Identifying missing values

```
SELECT *
FROM table_name
WHERE column_name IS NULL;
```

### 2.Dropping missing values

```
SELECT *
FROM table_name
WHERE column_name IS NULL;
```

### 3.Filling missing values

```
SELECT *
FROM table_name
WHERE column_name IS NULL;
```

**4.Replacing specificvalues**

```
UPDATE table_name
SET column_name = new_value
WHERE column_name = old_value;
```

**5.Imputing missing values**

- **Simple imputation**

```
UPDATE table_name
SET column_name = (SELECT AVG(column_name) FROM table_name WHERE column_name IS
NOT NULL)
WHERE column_name IS NULL;
```

- **Regression imputation**

```
UPDATE table_name
SET column_name = (SELECT AVG(column_name) FROM table_name WHERE column_name IS
NOT NULL)
WHERE column_name IS NULL;
```

**6. Handling Categorical Variables:**

Creating a new category for missing values in a categorical column:

```
UPDATE table_name
SET column_name = 'Unknown'
WHERE column_name IS NULL;
```

## Excel

Excel provides several built-in functionalities to handle missing values in a data set. Here's an overview of how Excel can be used to handle missing values:

**1.Identifying Missing Values:**

- Use filtering options or conditional formatting to highlight cells with missing values.
- Utilize functions like ISBLANK, ISNA, or COUNTBLANK to identify and count missing values in specific columns or ranges.

**2.Dropping Missing Values:**

- Delete rows or columns with missing values manually by selecting and deleting the respective cells or rows.
- Use filtering options to hide or exclude rows with missing values temporarily.

**3.Filling Missing Values:**

- Select the range of cells containing missing values and use the Fill feature (Home tab > Editing group > Fill) to fill the missing values with desired values, such as constants, values from adjacent cells, or a series.
- Use the Find and Replace feature (Home tab > Editing group > Find & Select > Replace) to replace specific missing values with desired values.

**4.Replacing Specific Values:**

- Use the Find and Replace feature (Home tab > Editing group > Find & Select > Replace) to replace specific values, including missing values, with desired values.

**5.Imputing Missing Values:**

- Use formulas like IF, ISBLANK, or ISNA to create conditional statements and replace missing values with desired values or calculations based on other cells' values.
- Utilize functions like AVERAGE, MEDIAN, MODE, or VLOOKUP to impute missing values with statistical measures or values from reference tables.

**6.Handling Categorical Variables:**

- Create a new category or label, such as “Unknown” or “N/A,” to represent missing values in categorical columns manually by entering the label in the respective cells.

## Power BI

In Power BI, handle missing values in a data set using various techniques. Here's an overview of how Power BI involves and performs handling missing values:

**Identifying Missing Values:**

- In the Power Query Editor, use the "Replace Values" or "Replace Errors" options to identify and replace missing values with desired values or placeholders.

**Dropping Missing Values:**

- In Power Query Editor,  filter out or remove rows with missing values using the "Remove Rows" or "Filter Rows" options.

**Filling Missing Values:**

- In Power Query Editor, use the "Fill" option to replace missing values with desired values.  choose to fill with constants, values from previous or next cells, or calculations based on other columns.

**Replacing Specific Values:**

- In Power Query Editor, use the "Replace Values" option to replace specific values, including missing values, with desired values or placeholders.

**Imputing Missing Values:**

- Use DAX (Data Analysis Expressions) functions in the Power BI data model to calculate and impute missing values based on other columns or measures. Functions like IF, ISBLANK, or AVERAGEX can be useful in these scenarios.

**Handling Categorical Variables:**

- In Power Query Editor, use the "Replace Values" or "Replace Errors" options to replace missing values in categorical columns with labels like "Unknown" or "N/A".

## Conclusion

Handling missing values is a critical step in data analysis and machine learning. By applying appropriate techniques, such as deletion, imputation, or predictive modeling, missing values can be effectively addressed. Careful consideration of the missingness pattern and the nature of the dataset is essential to select the most suitable approach. Following best practices, such as documenting decisions and conducting sensitivity analysis, ensures the integrity and reliability of the analysis results.