Missing Values in Datasets

This report covers the concept of missing values in datasets, the mechanisms behind them, and various imputation techniques. We will explore examples using the Titanic dataset.

1. Mechanisms of Missing Values

Missing values occur in datasets when some information is not stored for a variable. There are three main mechanisms:

1.1 Missing Completely at Random (MCAR)

Missing completely at random (MCAR) is a type of missing data mechanism in which the probability of a value being missing is unrelated to both the observed data and the missing data. In other words, if the data is MCAR, the missing values are randomly distributed throughout the dataset, and there is no systematic reason for why they are missing. For example, in a survey about the prevalence of a certain disease, the missing data might be MCAR if the survey participants with missing values for certain questions were selected randomly and their missing responses are not related to their disease status or any other variables measured in the survey.

1.2 Missing at Random (MAR)

Missing at Random (MAR) is a type of missing data mechanism in which the probability of a value being missing depends only on the observed data, but not on the missing data itself. In other words, if the data is MAR, the missing values are systematically related to the observed data, but not to the missing data. Here are a few examples of missing at random: Income data: Suppose you are collecting income data from a group of people, but some participants choose not to report their income. If the decision to report or not report income is related to the participant's age or gender, but not to their income level, then the data is missing at random. Medical data: Suppose you are collecting medical data on patients, including their blood pressure, but some patients do not report their blood pressure. If the patients who do not report their blood pressure are more likely to be younger or have healthier lifestyles, but the missingness is not related to their actual blood pressure values, then the data is missing at random.

1.3 Missing Not at Random (MNAR)

It is a type of missing data mechanism where the probability of missing values depends on the value of the missing data itself. In other words, if the data is MNAR, the missingness is not random and is dependent on unobserved or unmeasured factors that are associated with the missing values. For example, suppose you are collecting data on the income and job satisfaction of employees in a company. If employees who are less satisfied with their jobs are more likely to refuse to report their income, then the data is not missing at random. In this case, the missingness is dependent on job satisfaction, which is not directly observed or measured.

2. Example with Titanic Dataset

We will use the Titanic dataset to demonstrate how to handle missing values.

Code Snippet: Load Dataset

```
import seaborn as sns
```

```
df = sns.load_dataset('titanic')
df.head()
```

This code snippet loads the Titanic dataset using the seaborn library and displays the first few rows.

Code Snippet: Check for Missing Values

```
df.isnull().sum()
```

This code checks for missing values in each column of the dataset.

Code Snippet: Delete Rows with Missing Values

```
df.dropna().shape
```

This code snippet deletes rows with any missing values and shows the new shape of the dataset.

3. Imputation Techniques

We can handle missing values using various imputation techniques:

3.1 Mean Value Imputation

```
df['age_mean'] = df['age'].fillna(df['age'].mean())
```

This code fills missing values in the 'age' column with the mean age.

3.2 Median Value Imputation

```
df['age_median'] = df['age'].fillna(df['age'].median())
```

This code fills missing values in the 'age' column with the median age, which is useful in the presence of outliers.

3.3 Mode Imputation for Categorical Values

```
df['embarked'].unique()

mode_value=df[df['embarked'].notna()]['embarked'].mode()[0]

df['embarked_mode']=df['embarked'].fillna(mode_value)
```

This code fills missing values in the 'embarked' column with the mode (most frequent value).

4. Conclusion

Handling missing values is crucial for data analysis. Understanding the mechanisms behind missing data and applying appropriate imputation techniques can significantly improve the quality of the dataset.