Data Cleaning

Missing Values

Data quality is most important for achieving good results in any statistical analysis. But the missing values in a dataset can degrade data quality and become a major factor in producing incorrect results, leading to incorrect conclusions. For all these reasons, it becomes a priority to deal with missing values efficiently.

There are many reasons for missing values to be present in a dataset, such as

1. **While filling in the data, some people did not respond to certain questions.**
2. **Some data got lost while reading it from equipment.**
3. **There is a skip pattern while conducting surveys, like certain types/groups of people hesitate to provide some information and many more.**

According to the nature of missing data, it can be classified into three categories:

1. **MCAR** (Missing Completely At Random)
2. **MAR** (Missing At Random)
3. **MNAR** (Missing Not At Random)

Let's take an example to understand these various missing types.

Intuition for Missing Data Classification

Suppose you are taking a survey collecting people's data for a match-making website. The mode of conducting this survey is online, and the filled data gets automatically saved after filling each section of the questionnaire. Now there can be certain cases where data can be missing; let's explore those.

1. While filling out the survey questions, some people's internet connection got interrupted, and only some parts of the survey got saved, and they could not fill in the rest of the information. This is a case of complete randomness (MCAR), meaning the fault in the internet connection is a random event while filling out the survey.
2. For the weight column, many entries were missing. This missingness can be because people who are overweight or underweight tend not to fill out weight information. This kind of missing information is called MNAR (Missing at random) because overweight or underweight people are not likely to fill out the weight information.
3. A column in the survey asks about the time of birth, and missing values are present in this column. After some analysis, it was found that those people born in cities have records of their birth time, but people born in villages still needed the exact time information. This type of data is called MAR (Missing At Random) because the reason for the missingness is dependent on other observed values.

Now, we will define these categories of Missing data formally.

MCAR (Missing Completely At Random)

When the missing data in one column is unrelated to any observed or unobserved variable, it is called MCAR. An example of MCAR can be a faulty instrument which stops recording the values randomly.

MAR (Missing At Random)

When the missing data in one column is related to any other observed variable, it is called MAR. An example of MAR is that women are most likely to hide their age. Here the gender of a person is related to the missing age column.

MNAR (Missing Not At Random)

When the missing data in a column is related to that particular feature's observed and unobserved values. An example can be that if we are observing the people who are feeling sick today, then if values for that column are

missing, this is because most of the sick people are on leave and cannot fill in the value.

Empirical Rule

After observing the missing values, the next step is to work on those missing values. Certain rules can help while removing or imputing the missing values.

<table>
<tr><td colspan="2">Less than 5% of the data is missing.</td><td>Ignore</td></tr>
<tr><td colspan="2">More than 40% of the data is missing.</td><td>Ignore and Report</td></tr>
<tr><td rowspan="2">More than 5% and less than 40% of data is missing.</td><td>MNAR / MCAR</td><td>Ignore and Report</td></tr>
<tr><td>MAR</td><td>Use Imputation Methods</td></tr>
</table>

1. If less than 5% of data is missing, then missing values can be ignored by removing those observations.
2. If more than 40% of the values are missing, then that particular column should be dropped, and the missing values should be reported to the data-gathering team.
3. If the percentage of missing values is between 5% and 40%, then check for the category of missing values.
    a. If the missing values are of type MCAR or MNAR, then ignore these missing columns and report it to the concerned team.
    b. If the missing values are of MAR type, where the missing values in one column are related to those in another, then imputation methods can be applied. Some of the imputation methods are
        i. Mean
        ii. Median
        iii. Regression
        iv. MICE