**Missing Value**

Missing values are a common problem in real-world datasets. For example, if data is captured by sensors from a particular source, the sensor might stop working for a while, leading to missing data or some users unwilling to share information for a question etc.

Missing values are also referred to as NA (Not Available), None or NaN (Not a Number).

*Note: Not all missing values come in nice and clean np.nan or None format. For example, the dataset we work on may include “?” and “- -“ values in some cells. We can convert them to np.nan representation when reading the dataset into a pandas dataframe. We just need to pass these values to na\_values parameter.*

**Identifying the Type of Missingness**

The first step to implementing an effective imputation strategy is identifying why the values are missing. Even though each case is unique, missingness can be grouped into three broad categories:

1. **Missing Completely at Random (MCAR):** this is a genuine case of data missing randomly. Examples are sudden mistakes in data entry, temporary sensor failures, or generally missing data that is not associated with any outside factor. Generally, the amount of missingness is low.
2. **Missing at Random (MAR):** this is a broader case of MCAR. Even though missing data may seem random at first glance, it will have some systematic relationship with the other observed feature. Missing at Random is also called Missing Conditionally at Random because the missingness is conditional on another variable.
3. **Missing Not at Random (MNAR):** missing values may exist in large amounts, and the reason for the missingness is associated with external reasons.

**Handle Missing Values**

These are some tried and tested ways:

1. **Drop samples with missing values:** this is instrumental when both the number of samples is high, and the count of missing values in one row/sample is high. This is not a recommended solution for other cases, since it leads to heavy data loss.
2. **Replace missing value with constant value, mean, median or most frequent:** you can use statistical functions like mean, median or most frequent as a replacement for missing values. Even though they’re also assumptions, these values make more sense and are closer approximations when compared to one single constant value.
3. **Interpolate the missing values:** interpolation helps to generate values inside a range based on a given step size. For instance, if there are 9 missing values in a column between cells with values 0 and 10, interpolation will populate the missing cells with numbers from 1 to 9. Understandably, the dataset needs to be sorted according to a more reliable variable (like the serial number) before interpolation.
4. **Build a model with other features to predict the missing values:** by far the most intuitive of all techniques we’ve mentioned. Here, an algorithm studies all the variables except the actual target variable (since that would lead to data leakage). The target variable for this algorithm becomes the feature with missing values. The model, if well trained, can predict the missing points and provide the closest approximations.