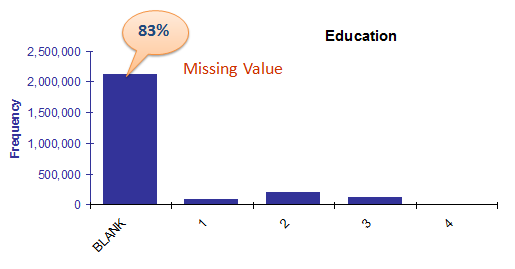
**Missing Value Analysis**

**Introduction:**

Missing values occur when no [data](https://en.wikipedia.org/wiki/Data) [value](https://en.wikipedia.org/wiki/Value_(mathematics)) is stored for the [variable](https://en.wikipedia.org/wiki/Variable_(mathematics)) in an [observation](https://en.wikipedia.org/wiki/Observation). Missing values are a common occurrence, and you need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Typically, ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values.



There are three main problems that missing data causes:

1. Missing data can introduce a substantial amount of bias
2. Make the handling and analysis of the data more arduous
3. Create reductions in efficiency

**Causes of Missing values:**

They may occur at different stages:

1. Data often are missing in research in [economics](https://en.wikipedia.org/wiki/Economics), [sociology](https://en.wikipedia.org/wiki/Sociology), and [political science](https://en.wikipedia.org/wiki/Political_science) because governments choose not to, or fail to, report critical statistics. Sometimes missing values are caused by the researcher—for example, when data collection is done improperly or mistakes are made in data entry
2. Data Extraction: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.
3. Data collection: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:
   * Missing at random: This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.
   * Missing that depends on unobserved predictors: This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included “discomfort” as an input variable for all patients.
   * Missing that depends on the missing value itself: This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

**Multiple ways to deal with missing values:**

1. Delete the variables whose missing values are > 40%
2. Ignore the records with missing values
3. Fill in missing values manually based on your domain knowledge.
4. Fill the missing values using central statistics. Mean/Median/Mode imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.
5. Use modeling techniques such as K - nearest neighbors, Bayes’ rule, decision tree, or EM algorithm.
   * KNN Imputation: In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function (Euclidean or Manhattan)

**Additional Information:**

Research Work:

* Create a small subset of data with complete observations
* Delete some values manually
* Use multiple methods to fill
* See where they are failing
* Choose the best method

In R, missing values are represented by the symbol NA (not available). Impossible values (e.g., dividing by zero) are represented by the symbol NaN (not a number)

**Interview Questions:**

1. During analysis, how do you treat missing values?
2. How do data management procedures like missing data handling make selection bias worse?