**Dealing with Missing Data**

Missing Data occurs when the variable or observation does not have any data value. This is one of the most common problems faced by Analysts as they can result in the compromise of the statistical inference and thereby derailing from one inference to another. The concept of missing values is important to understand in order to master the skill of successful management and understanding of the data.

Intelligent Handling of Missing data would result in building a robust model capable enough in handling complex tasks, even though there are many different ways to handle missing data, Let us now looks at some of the simple ways to take care of it.

* **Removing:**
  + These are very simple and commonly used method to handle the missing values. We delete the row if the missing value corresponds to the places in the row, or we delete the column if the column has more than 70%-75% of missing data. Again, the threshold value is not fixed and depends on how much one wishes to fix
    - Advantages:
      * Removal of missing data would help us build a robust model because removal of rows or columns that give us insufficient information is far better than having them for modelling
      * This is the simplest method of handling the missing data to build a highly reliable and accurate model
    - Dis-Advantages:
      * Removing of missing data might result in loss of information from the data
      * Removing should only be done if the percentage of missing data is less than 30% of the whole dataset
* **Mean/Median/Mode Imputation:**
  + This is a method in which in case of numerical data we compute its mean or median or its mode and use the result to replace the missing value and mode in case of the categorical data to replace the missing value
    - Advantages:
      * This is a better method as these are statistical approaches in handling the missing data rather than just removing it.
      * They also prevent data losses as they just replace the missing values and therefore works well with small datasets also
    - Dis-Advantages:
      * They may add bias and variance to the data
      * Not the best among the imputation methods for missing data
* **Predicting the missing values:**
  + In this method we replace the missing data with the help of the other available data by using suitable predictive modelling technique (depending on the nature of the data either numeric or categorical)
    - Advantages:
      * This is better method of handling the missing values as the result is going to be based upon the other available data only
      * Bias may be eliminated by using this method of imputation
    - Dis-Advantages:
      * Even though we try to predict the variables statistically, we must keep in mind that this is just a proxy and not the real data
* **Using Algorithms that support missing data:**
  + Instead of improper handling of missing data we can use algorithms like KNN or Random Forest to work towards inference with the missing Data as such with the Legitimacy of original data

**Handling Categorical Data**