**Missing Values with Mean, Median & Mode**

There are three main types of missing data:  
Missing completely at random (MCAR)  
Missing at random (MAR)  
Not missing at random (NMAR)

Techniques:

* Do nothing
* Imputation Using (Mean/Median) Values
* Imputation Using (Most Frequent) or (Zero/Constant) Values
* Imputation Using k-NN
* Imputation Using Multivariate Imputation by Chained Equation (MICE)
* Imputation Using Deep Learning (Datawig)

Mean is the average of all values in a set

Median is the middle number in a set of numbers sorted by size

Mode is the most common numerical value for two or more sets.

The goal is to find out which is a **better measure of the central tendency of data** and use that value for replacing missing values appropriately. The mean and median can only be calculated on numerical variables, therefore, these methods are suitable for continuous and discrete numerical variables only.

Plots such as **box plots**and **distribution plots**come very handy in deciding which techniques to use.

**If the data is skewed.** There are **several or large numbers of data points** that act as outliers. Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model and hence gets ruled out.

 For symmetric data distribution, one can use the **mean value** for imputing missing values.

In median **imputation**in which the missing values are replaced with the **median value** of the entire feature column. When the data is skewed, it is good to consider using the median value for replacing the missing values. Note that imputing missing data with median value can only be done with **numerical data**.

If the data is skewed or if the data consists of outliers, one may want to use median.

**In mode imputation**in which the missing values are replaced with the **mode value** or **most frequent**value of the entire feature column. When the data is skewed, it is good to consider using mode values for replacing the missing values. For data points such as the salary field, you may consider using mode for replacing the values. Note that imputing missing data with **mode**values can be done with numerical and categorical data.

 If the data is skewed, one may want to use mode.

Practical Implementation using package:

from sklearn.impute import SimpleImputer

#

# Missing values is represented using NaN and hence specified. If it

# is empty field, missing values will be specified as ''

#

imputer = SimpleImputer(missing\_values=np.NaN, strategy='mean')

strategy = mean, mode, median, constant and most\_frequent

Resources:

<https://vitalflux.com/missing-data-imputation-machine-learning/>

<https://vitalflux.com/imputing-missing-data-sklearn-simpleimputer/>

Boxplot: <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>