## ***On what basis we choose data scaling method (Normalization/Standardization)?***

**Answer: -**

🡪**Normalization** is the process of restructuring a relational database in accordance with a series of so-called normal forms in order to **reduce data redundancy** and **improve data integrity**. This pre-processing rescaling method is useful for sparse attribute features and algorithms using distance to learn such as KNN, Mean, Median and Mode. Normalization usually means to scale a variable to have a values **between 0 and 1**.

🡪In **Standardization**, attribute data assumes a Gaussian distribution of input features and "standardizes" to a **mean of 0 and a standard deviation of 1**. Standardization works better with linear regression, logistic regression and linear discriminate analysis.

Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks.

Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression and linear discriminant analysis.

To choose a data scaling method, we need to check the distribution of data. If the data is normally/uniformly distributed, then Standardization is the suitable method for the scaling purpose. Otherwise, if the data is not normally distributed, we should select Normalization.

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a [mean](http://www.statisticshowto.com/mean/)of zero and a [standard](http://www.statisticshowto.com/probability-and-statistics/standard-deviation/)deviation of 1.

1. ***If the VIF is 2 then what is value of correlation coefficient (r^2)?***

**Answer: - 🡪** Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. It is used to explain how much amount multicollinearity (correlation between predictors) exists in a regression analysis.

We know,

VIF = ; where = -value obtained by regressing the *kth* predictor on the remaining predictors.

Or, VIF(1-) = 1,

Or, 2(1-) =1,

Or, 2- 2=1,

Or, 2 = 1,

Or, = 0.5

By definition, so here required = 0.5 (Answer).

1. ***How do you interpret chi-square result?***

**Answer: - 🡪** Chi-square test is to determine whether there are significant relationship between two categorical variables.  We apply when we have [categorical variable](https://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable) from a single population, this test is used to determine whether sample data are consistent with a hypothesized distribution.

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](https://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

1. ***Why do we choose boxplot method than other for outlier detection and removal?***

**Answer: - 🡪**

* Box plots are used to show overall patterns of response for a group. They provide a useful way to visualize the range and other characteristics of responses for a large group.
* Boxplot is especially useful for indicating whether a distribution is skewed and whether there are potential unusual [outliers](https://www.statcan.gc.ca/edu/power-pouvoir/glossary-glossaire/5214842-eng.htm#outlier) in the data set. Box and whisker plots are also very useful when large numbers of observations are involved and when two or more data sets are being compared.
* Box and whisker plots are ideal for comparing distributions because the center, spread and overall range are immediately apparent.

So that we use Boxplot for various Statistical Data Models than any other Outlier detection and removal methods.

1. ***How do we choose best method to impute missing value for a data?***

**Answer: - 🡪** In real world data, there are some instances where a particular element is absent because of various reasons. This leads to Missing values in our Data Set.

Sometimes it is better to remove the null values directly and, for some tasks, it is better to use mining techniques to impute the values.

We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Another way is to approximate it with the deviation of neighboring values which is known as KNN imputation. This works better if the data is linear.