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

Data scaling is a method to limit the variable, so that they can be compare on a common ground.

There are 2 methods:

1. Normalization
2. Standardization /z-score

**Normalization:**

Normalization method used to rescale components of a feature vector to a complete range of vector length of 1. That usually means to change all numeric column values to a common scale without distorting differences in the range of value. All values ranges between 0-1. This pre-processing rescaling method is useful for sparse attribute features and algorithms.

**Standardization:**

Standardization method is used to get Gaussian distribution of input features and standardise to a mean of 0 and standard deviation 1. This works better with linear regression, logistic regression. This method works well with uniformly distributed data.

So if the data is uniformly distributed we choose standardization method else we use Normalization method.

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

The variance inflation factor quantifies the extent correlation between one predictor with another. To calculate VIF formula is:

VIF = 1/ (1-R2)

The formula for correlation coefficient

(-R2) = (1/VIF)-1

Then if VIF = 2

(-R2)= (½)-1 =0.5-1=-0.5

R2 = 0.5

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

Chi-square test is appreciated when there are 2 categorical variable from one population. Then this test is used to determine whether there is a significant association between those 2 variables.

So while doing this test we encounter 2 hypothesis:

* Null Hypothesis (H0)
* Alternative Hypothesis (Ha)

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

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