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

Ans : Data Scaling : Data Scaling is one of the statistical technique that helps to maintains continuous variable data in the same range.

Normalization : Normalization allows to maintain data in the range of 0 to 1

Formula : value-minvalue/maxvalue-minvalue

Standardizatin : it explains variability of data in terms of standard deviation. It transforms the data into 0 mean and 1 unit standard deviation

Formula : Z= value-mean/standard deviation

**On what basis we choose data scaling method :**

Based on distribution of variable we have to decide scaling technique. If data is **distributed normally** then Standardization method is best for scaling.

If data is not normally distributed then Normalization technique is preferred.

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

Ans : Variable Inspection Factor(VIF) : VIF helps to determine multi colinearity between two independent continuous variable.

Formula : Vif =1/1-r^2

VIF ranges from 1 to infinity

Substitute VIF =2 in above formula

2-2r^2 = 1

R^2 = ½

R^2= 0.5

1. How do you interpret chi-square result?

Ans: chi-square test is to calculate dependency between two categorical variables.

Based on Hypothesis test we will determine dependency

Null Hypothesis : Two variables are independent

Alternate Hypothesis : Two variables are not independent

We use contingency table for determination

Chi square test calculates two values

1. Chisquare test statistic
2. Critical value

If chisquare test statistic > critical value then reject the null hypothesis

Else we accept null hypothesis

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

Ans : Outliers : The data points which are far away from regular pattern is considered as outliers.

Box plot : Boxplot gives representation of outliers in graphical way.

Outliers analysis should be perform on each and every numerical variable in a dataset.

For univariate analysis, Box plot graph gives clear representation of data. It includes 1st quarter, median, 3rd quarter,min value,max value

From the box plot graph we can easily infer the point which are far away from regular data points called outliers.

i.e lower outliers = q1-1.5(q3-q1) and upper outliers = q3 + 1.5(q3-q1)

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

Ans : There are multiple ways to impute missing values i.e mean, median,mode, KNN..etc

**Best method to impute missing value for a data :**

imputing missing value is depend on data set and data

One of the Best of imputation is :

Knowingly remove defined data(existing data point) after that Impute the missing values with various strategies and identify the strategy which is close to known value and prefer that method as best method for that data set.