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

Ans : Data scaling refers to proportionate or scale the data variables. Data scaling is important to avoid biasing. If we see the example of Age and Loan Amount where loan amount will be in lakhs or crores but age will be 1-100. If we feed the same data to a model then model may provide biasing results as Loan amount seems to be dominating. In data scaling, methods are selected based on distribution of data. Standardization assumes to be data in **normally distributed**. Most of times the data is randomly distributed. Normalised data scaling can be used for any type of data.

If we observe the formula of both the methods

Normalisation=(value-min)/(Max-min)

Standardization=(x-Mu)/sd

If we see the normalisation formula it has nothing to do with the mean or standard deviation where standardization depends on central tendency of data.

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

Ans: VIF: VIF stands for variance inflation factor. It is used to measure the correlation between the variables. If VIF is high the it indicates variables are highly correlated where VIF is low indicates low multicollinearity between variables.

If we observe the relation between VIF and correlation coefficient.

VIF=1/(1-R^2)

VIF=2

1/(1-R^2)=2 implies

1-R^2=1/2

2\*(1-R^2)=1

R^2=1/2

Correlation coefficient is (1/2) .

3. How do you interpret chi-square result?

Ans: Chi-square test is used to check the independency between two categorical variables. Chi-square testing uses contingency table. It uses two categorical variables to build contingency table. Here there are two types of hypothesis NULL and Alternate Hypothesis.

C^2=sum(((O-E)^2)/E

Here O-observed value and E is the expected value.

Chi-square statics is greater the critical value then Null hypothesis is rejected and two variables are dependent.

Chi-Square statistics is lesser the critical value then accept the null hypothesis.

Critical value is available online. This values is selected based on degree of freedom and confidence interval level.

Degree of freedom is given by (No of rows-)\*(No of cols.)

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

Ans: BoxPlot is most power full graphical tool to find the outliers. Outlier refers to unevenness in the data. Un evenness refers to some observations do not have any proportionate relationship. If we observe salary of people like 0.25$, 2500$,2600$,2100$,100000$. If we observe the data here there are two outliers because 2500,2600,2100 are varying with some valid flow. But 0.25$ or 100000$ are very different from majority of the data. So 0.25$ and 1000000$ are outliers.

Box plot is choose for oultlier detection because it does not make any assumption on data. Whereas statistical method like Grubbs test assumes the data to be normally distributed. Even the box can be easily interpreted if we see the plot. The existence of outliers are clearly indicated in box plot. Using box plot one can make decision on outlier because it is graphical. We can also get values of oulier using stats of boxplot objects. Indices can be known.

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

Ans : Lets understand first what is missing value in data. When data is received from client or taken from Data base there may be chance of error. These may be human or system error. If see the different formats of data like excel, json , csv files there may be chance of few missing observations. We cannot directly work on these type of data for model building .

Imputing here refers to treat them properly before statistical analysis. We have different methods to fill these details like filling central tendency params mean,median,mode , KNN imputing or predictive model. So choosing the best of all is important. To do this remove one observation and feed it with available imputing methods like mean, median,mode ,KNN or predicted values. So to choose the best value among these compare with actual value. Select that is closest. This can be done in iterative process for few observations to come to conclusion.