Q 1. What is Linear regression?

```
Linear Regession :

1. It is a predective analysis method, uesd to findout the linear realtionship between
dependent variable(y) and independent variable(x).
2. It is one of the easiest and most popular machine learning algorithm.
3. Linear Regerssion model makes predictions for contineous/real or numeric variables
such as salary, age, product price, etc.
4. we can find linear realtionship between x and y by following equation: -

            y = mx + c
where,

  y =  dependent variable
  x = independent variable
  m = slope
  c = Intercept

5. If there is only one input variable (x), then such linear regression is called simple
linear regression. And if there is more than one input variable, then such linear
regression is called multiple linear regression.
```

Q 2. How do you represent a simple linear regression?

```
Simple linear Regression :-

    In simple linear regression there is linear realtionship between only one independent
variable and one dependent variable. The realtionship is shown by sloped straight line or
linear line so it is called as simple linear regression.

example : experience and Salary, Income and Expenditure
```

Q 3. What is multiple linear regression?

```
Multiple linear Regression :-

    In multiple linear regression there is linear realtionship between multiple
independent variable and one dependent variable. The realtionship is shown by sloped
straight line or linear line so it is called as Multiple linear regression.
    In multiple linear regression the dependent or target variable(Y) must be the
continuous/real, but the predictor or independent variable may be of continuous or
categorical form.
```

Q 4. What are the assumptions made in the Linear regression model?

- Folowing are assumptions made for Linear Regression Model:

1. Linearity :-

-There should be linear realtionship between dependent variable(y) and independent variable(x)

2. Small or No Multicolinearity :-

-All the independent variables are independent to each other.
-There should not be linear realtionship between independent variables(x1 and x2)

3. Homosceadcity :-

-It is a situation when the error term is the same for all the values of independent variables.
-With homoscedasticity, there should be no clear pattern distribution of data in scatter plot.
-'Homo' means 'same' and 'scedasticity' means 'varienace'. In statistics of all the random variables in sequence have same infinite varience.

4. Noramality of residual(Yactual - Ypredicted) :-

-There should be minimum error margin of actual value and predicated value.

Q 5. What if these assumptions get violated?

1.Violation of linearity:-
    A linear model to data which are nonlinear, predictions are likely to be seriously in error, especially when remove beyond the range of the sample data.

2. Violation of No multicollinearity:-
    Violating multicollinearity does not impact prediction, but can impact inference. For example, p-values typically become larger for highly correlated covariates, which can cause statistically significant variables to lack significance. Violating linearity can affect prediction and inference.

3. Violation of homoscedasticity:-
    Violation of the homoscedasticity assumption results in heteroscedasticity when values of the dependent variable seem to increase or decrease as a function of the independent variables. Typically, homoscedasticity violations occur when one or more of the variables under investigation are not normally distributed.

4. Violation of normality:-
    If the population from which data to be analyzed by a normality test were sampled violates one or more of the normality test assumptions, the results of the analysis may be incorrect or misleading.

Q 6. What is the assumption of homoscedasticity?

1. The assumption of equal variances (i.e. assumption of homoscedasticity) assumes that different samples have the same variance, even if they came from different populations.

2. The assumption of homoscedasticity (meaning “same variance”) is central to linear regression models. Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables.

Q 7. What is the assumption of normality?

    Assumption of normality means that you should make sure your data roughly fits a bell curve shape before running certain statistical tests or regression.
    It’s assumed that the residuals from the model are normally distributed.

Q 8. How to prevent heteroscedasticity?

1. Transform the dependent variable:-

    One way to fix heteroscedasticity is to transform the dependent variable in some way. One common transformation is to simply take the log of the dependent variable.

2. Redefine the dependent variable:-

    Another way to fix heteroscedasticity is to redefine the dependent variable. One common way to do so is to use a rate for the dependent variable, rather than the raw value.

3.Use weighted regression:-

    Another way to fix heteroscedasticity is to use weighted regression. This type of regression assigns a weight to each data point based on the variance of its fitted value

Q 9. What does multicollinearity mean?

    Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.

    For Multicollinearity in linear regerssion is an independent variables in regression model is independent to each other. There should not be linaer realtionship between between x1 and x2.

Q 10. What are feature selection and feature scaling?

Feature Selection :-

```
    Feature selection is the process of reducing the number of input variables when
developing a predictive model. Statistical-based feature selection methods involve
evaluating the relationship between each input variable and the target variable using
statistics and selecting those input variables that have the strongest relationship with
the target variable. These methods can be fast and effective, although the choice of
statistical measures depends on the data type of both the input and output variables.


# Feature Scaling :-

    Feature scaling is a method used to standardize the range of independent variables or
features of data. In data processing, it is also known as data normalization and is
generally performed during the data pre-processing step.

Stand = (X – Xmin)/(Xmax – Xmin)
```

Q 11. How to find the best fit line in a linear regression model?

```
Best Fit Line :-

-   It is used to find to the error between predicted value and actual value should be
minimized.
-   It has lowest Mean squared error.
-   It passes through maximum number of data points.
-   we need to calculate best m & c value for finding best fit line.
-   The least Sum of Squares of Errors is used as the cost function for Linear
Regression. For all possible lines, calculate the sum of squares of errors. The line
which has the least sum of squares of errors is the best fit line.
```

Q 12. Why do we square the error instead of using modulus?

```
Squaring always gives a positive value, so the sum will not be zero. squaring makes the
algebra much easier to work with and offers properties that the absolute method does not.
    for example, the variance is equal to the expected value of the square of the
distribution minus the square of the mean of the distribution.
```

Q 13. What are techniques adopted to find the slope and the intercept of the linear regression line which best fits the model?

```
- Gradient descent is use to find the optimum values of m and c, it differenctiate cost
function i.e MSE with respect to m and c and calculate it.
- i.e MSE with respect to m and c and calculate best values
- In that alpha is used as Learning rate and Learning step will be each step and that
will d
    Mnew = Mold - alpha*derivation(MSE) w.r.t 'm'
    Cnew = Cold - alpha*derivation(MSE) w.r.t 'c'
```

```
- This is how it will Keep on finding m and c values until iterations = 1000,  epsilon
(sqrt.(Mnew^2-Mold^2))<0.001 either of these conditions are getting satisfied
```

Q 14. What is cost Function in Linear Regression?

```
# Cost Function :-

- The difference values for coefficients of lines (m, c) gives the diiference line of
regression.
- The cost function is used to estimate the values of the coefficients for the best fit
line.
- Cost function optimize the regression coefficients.
- It measures how a linear regression model is performing.
- We can use cost function to find the accuracy of the mapping function, which maps the
input variable to output variable.

- For linear regression, MSE is a cost function. Which is average of squared error
occured between predicted values ans actual values.

    MSE = sum(Yactual - Ypred)/N

     Where,

     N = Total number of observation.
     Yactual = actual value
     Ypred = Predicted value
```

Q 15. Briefly explain gradient descent algorithm

```
Gradient Descent Algorithm :-

- It is used to find best m and c values or best fit line. Also it is used to reduce MSE.
- A regression model uses gradient descent to update the values of m and c of the line by
reducing MSE.
- Gradient Descent will try infinte m and c values and find best m and c values.
- i.e MSE with respect to m and c and calculate best values. In that alpha is used as
Learning rate and Learning step will be each step and that will d
    Mnew = Mold - alpha*derivation(MSE) w.r.t 'm'
    Cnew = Cold - alpha*derivation(MSE) w.r.t 'c'

- This is how it will Keep on finding m and c values until iterations = 10000 epsilon
(sqrt.(Mnew^2-Mold^2))<0.001 either of these conditions are getting satisfied
```

Q 16. How to evaluate regression models?

```
We generally evaluate regression model based on following metrics:
1. r2_score
2. MSE
3. RMSE
4. MAE
these values we calculate for both Train and Test dataset and see if our model is
optimal.
```

Q 17. Which evaluation technique should you prefer to use for data having a lot of outliers in it?

In [ ]:

```
- For detect outliers following methods use :-
    1. Z_score method
    2. IQR method
    3. BoxPlot
    4. ScatterPlot

- We can handle outliers with the help of following methods :-

    1. By deleting Observations
    2. Imputation (mean, median, mode, zeros, any static values)
    3. Transformation (To reduce impace of outliers)
        a. Log Transformation
        b. Normalization (0 to 1 range)
        c. Standardization (No fixed range)
        d. Cuberoot Transformation
        e. Reciprocal Transformation
```

Q 18. What is residual? How is it computed?

```
# Residual :-

It is difference between actual value (Yactual) and predicted value (Ypred) it is called
as Residual.

residual = (Yactual - Ypred)

- If observed points are far away from regression line.
        - Residual will be high
        - Cost function (MSE) is high.

- If observed points are close to regression line.
        - Residual will be low
        - Cost function (MSE) is low.
```

Q 19. What are SSE, SSR, and SST? and What is the relationship between them?

```
# 1. SSE (Sum of squared error) :-

It is squared difference between actual value (Yactual) and predicted value (Ypred).

SSE = sum(Yactual - Ypred)^2


# 2. SSR (Sum of squares due to Regress) :-

It is squared difference between predicted value (Ypred) and mean value (Ymean).
```

```
SSR = sum(Ypred - Ymean)^2

# 3. SSE (Sum of Total error) :-

It is squared difference between actual value (Yactual) and mean value (Ymean). It is
total error in the model. Which is sum of SSE and SSR.

SST  = sum(Yactual - Ymean)^2 or  SST = SSE + SSR
```

Q 20. What's the intuition behind R-Squared?

```
# R-Squared :-

- R-Squared is a statical method that determines the goodness of fit.
- It measures the strength of realtionship between the dependent and independent
variables on a scale 0-100%.
- High value of R-Squared determines the less difference predicted and actual value and
hence represent a good model.
- It is also called as coefficiant of determination.

R_squared = (Explained Variation)/(Total Variation)
          = 1 - SSE / SST

# It will increase  for good predictors as well as it will also increase for bad
predictors. So we need to use Adjusted R-Squared method.
```

Q 21. What does the coefficient of determination explain?

```
# R-Squared :-

- R-Squared is a statical method that determines the goodness of fit.
- It measures the strength of realtionship between the dependent and independent
variables on a scale 0-100%.
- High value of R-Squared determines the less difference predicted and actual value and
hence represent a good model.
- It is also called as coefficiant of determination.

R_squared = (Explained Variation)/(Total Variation)
          = 1 - SSE / SST
```

Q 22 . Can R^2 be negative?

In [ ]:

```
Yes. R_squared value can be negative.
```

Q 23. What are the flaws in R-squared?

```
It will increase  for good predictors as well as it will also increase for bad
predictors. So we need to use Adjusted R-Squared method.
```

Q 24. What is adjusted R²?

```
# Adjusted R-Squared :-

- It will increase only for good predictors.
- Adjusted R-Squared value will always be less than or equal to R-Squared.

    Adjusted R-Squared <= R-Squared

Adjusted_R_Squared = 1-(1-(R_squared)^2) * (N-1)/(N-P-1)

where,

R_squared = R_squared value
N = No of samples
P = No of predictors
```

Q 25. What is the Coefficient of Correlation: Definition, Formula

```
# Coefficient of Correlation :-

- A statistical measure that defines co-relationship or association of two variables.
- Correlation coefficients are indicators of the strength of the linear relationship
between two different variables, x and y. - A linear correlation coefficient that is
greater than zero indicates a positive relationship.
- A value that is less than zero signifies a negative relationship.
- Finally, a value of zero indicates no relationship between the two variables x and y.


r = Sum(X-Xmean)(Y-Ymean)/squrt(Sum(X-Xmean)^2(Y-Ymean)^2)

r   =   correlation coefficient
X   =   values of the x-variable in a sample
Xmean=  mean of the values of the x-variable
Y   =   values of the y-variable in a sample
Ymean=  mean of the values of the y-variable
```

Q 26. What is the relationship between R-Squared and Adjusted R-Squared?

```
# R-Squared :-

- R-Squared is a statical method that determines the goodness of fit.
- It measures the strength of realtionship between the dependent and independent
variables on a scale 0-100%.
- High value of R-Squared determines the less difference predicted and actual value and
hence represent a good model.
```

```
- It is also called as coefficiant of determination.

R_squared = (Explained Variation)/(Total Variation)
          = 1 - SSE / SST

# Drawback of R_squared:-

    It will increase  for good predictors as well as it will also increase for bad
predictors. So we need to use Adjusted R-Squared method.

# Adjusted R-Squared :-

- It will increase only for good predictors.
- Adjusted R-Squared value will always be less than or equal to R-Squared.

    Adjusted R-Squared <= R-Squared

Adjusted_R_Squared = 1-(1-(R_squared)^2) * (N-1)/(N-P-1)

where,

R_squared = R_squared value
N = No of samples
P = No of predictors
```

Q 27. What is the difference between overfitting and underfitting?

```
# Overfitting:-

- Sometimes machine learning performs well on training data but does not perform well
with the test data.
- It means the model is not able to predict the output when seals with unseen data by
introducing noise in the output and hence the model is called overfitted.
- It has high training accuarcy and low testing accuarcy.
- It has low bias and high varience.

# How to handle avoid Overfitting:-

1. Remove Features (15 data to 10 data)
2. Use parameter tunning. Use prunung for DT.
3. Remove Outliers
4. Increase dataset
5. Regularisation.


# Underfitting :-

- Sometimes machine learning not performs well on training data and testing data.
- So this model is called Underfitted.

- It has low training accuarcy and low testing accuarcy.
- It has high bias and low varience.

# How to handle avoid Underfiiting:-

1. Add Features (10 data to 13 data)
2. Handling Missing values(maen,median,etc)
```

```
3. Remove Outliers
4. Increase dataset
5. Use correlated features.
```

Q 28. How to identify if the model is overfitting or underfitting?

```
# Overfitting:-

- Sometimes machine learning performs well on training data but does not perform well
with the test data.
- It means the model is not able to predict the output when seals with unseen data by
introducing noise in the output and hence the model is called overfitted.
- It has high training accuarcy and low testing accuarcy.
- It has low bias and high varience.


# Underfitting :-

- It has low training accuarcy and low testing accuarcy.
- It has high bias and low varience.
```

Q 29. How to interpret a Q-Q plot in a Linear regression model?

In [ ]:

```
Q-Q Plot :- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each oth
    - Q-Q Plot is used in linear regression for determining data is normally distributed or
    - If the all the data points are on the regression line then the data is norally distri
    - Also it will give information about which distribution is fit for line(Uniform distri
```

Q 30. What are the advantages and disadvantages of Linear Regression?

```
# Advantages of Linear Regression :-

    1. Linear Regression is performs exceptionally well on linearly seaprable data.
    2. It is easy to implement.
    3. Overfitting can be reduced by regularization L1 and L2.

# Disadvantages of Linear Regression :-

    1. Linearity - There should be linear realtionship between dependent variable(y) and
independent variable(x).
    2. Independence - All the independent variables are independent to each other. There
should not be linear realtionship between independent variables(x1 and x2).
    3. It is very sensitive to outliers.
    4. It is sensitive to missing values.
```

31. What is the use of regularisation? Explain L1 and L2 regularisations.

```
Regularisation :-

-    Regularisation used for prevent model from overfitting by adding extra information
to it.
-    Sometimes machine learning performs well on training data but does not perform well
with the test data.
-    It means the model is not able to predict the output when seals with unseen data by
introducing noise in the output and hence the model is called overfitted.
-    So Regularisation used to reduce overfitting.
-    This Technique can be used in such a way that it will allow to maintain all
variables or features in the model by reducing the magnitude of the variables. Hence it
maintains accuracy as well as generalization of the model.

# Regularisation is divided into two types:-

    1. Regid Regularisation (L2)
    2. Lasso Regularisation (L1)


1. Regid Regularisation (L2) :-

It is mostly used to reduce the overfitting in the model, and it includes all the
features present in the model.
It is calculated by

 c.f. = (Ycatual - Ypredicted)^2 + lambda * (slope)^2

 where,
 lambda = Hyperparameter tunning (0 to infinity range)
 if we take lambda=0 this model will act as linear regression. hence try to take lambda =
0.01


# 2. Lasso Regularisation (L1) :-

It is used to reduce the overfitting in the model, by shrinking as well as feature
selection.
It is calculated by

 c.f. = (Ycatual - Ypredicted)^2 + lambda * |slope|

In this methode slope is "+Ve"
 where,
 lambda = Hyperparameter tunning (0 to infinity range)
 if we take lambda=0 this model will act as linear regression. hence try to take lambda =
0.01
```