1. **Explain the linear regression algorithm in detail.**
Linear regression is a supervised learning algorithm which provides a linear relationship between predictor variables and predicted variable. The predictor variable is also known as independent variable that remains unchanged due to the change in other variables. However, the predicted variable changes with fluctuations in the independent variables. A sloped straight line represents the linear regression model also known as Best fit line.

The mathematical equation used to define it is

$y(x) = B0 + B1 * x$

Linear regression are of two kinds:

Simple Linear regression

One predictor variable and one predicted variable

Eg: $y(x) = B0 + B1 * x$

Multiple Linear regression

Two or more predictor variables and one predicted variable

Eg: $y(x) = B0 + B1 * x1 + B2 * x2$ etc.

where,

- y = Predicted variable. Variable y represents the continuous value that the model tries to predict.
- x = Predictor variable. x is the feature, while it is termed the independent variable. Variable x represents the input information provided to the model at any given time.
- B0 = y-axis intercept
- B1 = the regression coefficient. B1 is the equivalent of the slope of the best-fit straight line of the linear regression model.

Assumptions of linear regression are as follows:

1. Linearity: Linear relationship exists between X and Y variables.
2. Normality: Residuals are normally distributed.
3. Independence: Error terms are independent of each other.
4. Homoscedasticity: The variance of residual is the same for any value of X.

Procedure to perform Simple Linear regression:

1) Read the data and Visualise it to understand the Linear relationship between predictors and to be predicted variable by plotting graphs.
2) Get the best predictor bye understanding the correlation between the variables.
3) Split the given data in training data and test data
4) Train the model
5) Assess the model by checking following values in Summary
   — P-value: helps in determining whether the coefficient is significant or not.(<0.05)
   — F statistic: higher the value of F statistic, the more significant a model turns out to be.
   — R squared: tells how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.
   6) Predict the Y variables and get the residuals .
7) Check whether the residuals are normally distributed.
8) Use the same model to predict the y variable of test set.
9) Check the r2_score of test set prediction. It should be close to training set.

Procedure to perform Multiple Linear regression. It is similar to simple linear but with few more assumptions:

1. Overfitting: Training the models with more and more variables might make the model so fitting that it might memorise every value and might not generalise at all.
2. Multicollinearity: Correlation amongst the predictor variables might affect the model as well. If many independent variables are interrelated, it might cause redundancy.

There are two ways of dealing with Multicollinearity:

— Correlation: Find the correlation between the field using a pair plot or heatmap.

— VIF value: Explains the relationship of one independent variable with all the other independent variables. Value more than 5 is considered as an issue and variable is dropped.

As multiple linear regression can be built with different combinations of the variables, to asses the model, we use two new parameters along with R-squared:

— Adjusted R2 (1 - ((1- Rsquared) (N-1) / N-p-1))

— AIC

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet shows us why visualising data is so important. The quartet consists of four datasets with identical summary statistics. Each dataset has a series of x values and dependent y values. However, if you plot the datasets, they look surprisingly different compared to each other.

For example, let's say we are studying the correlation between the age of a car and its kilometre per litre (kmpl). We could calculate the correlation coefficient or other summary statistics, but these don't tell us anything about the actual pattern of the relationship. By visualising the data in a scatterplot, we can see if the relationship is linear or nonlinear, if there are any outliers, or if there are other factors at play.

Visualising data can also help us identify inconsistencies in our data. For eg, if we plot a scatterplot with a perfect linear line, this might indicate that the data is too good and needs to be re-checked.

Visualising data can help us make better decisions and avoid making incorrect assumptions based on summary statistics alone.

3. **What is Pearson's R?**

The Pearson correlation coefficient is a measure of the strength of the linear relationship between two variables. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use r to represent Pearson's correlation unless otherwise noted.

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Figure 1 shows a scatter plot for which r = 1.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling is a way of normalising the independent variables, generally performed during data processing step. The data is converted in such way that they are centered around 0 or in the range (0,1) depending on the scaling technique

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very useless coefficients that might be difficult to interpret. So we need to scale the features.

There are two ways of scaling data:

— Standardising

Standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

X = x - mean(x)/Standard deviation of x

— MinMax scaler

It is the simplest method and consists of rescaling the range of variables to scale the range in [0, 1]. The general formula for normalization is given as:

X = x-min(x) / max(x)-min(x)

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

   If VIF value of a variable is infinite then it indicates that the variable may be expressed exactly by a linear combination of other variables. To solve this, we have drop the variable and get VIF again to make sure none of the other variable is expressing similar correlation.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**

   Q stands for Quantile in Q-Q plot. Quantile is also known as percentile. Eg: Median is a percentile. Q-Q plot is a scatter plot of quantiles of two datasets against each other. The expectation is that if the two datasets come from populations with a common distribution, then the points fall approximately along the 45 degree line.

   It answers following questions:

   1) If two datasets come from populations with a common distribution.
   2) If they have similar shape
   3) if they have similar tail behaviour.

   We can use Q-Q plots to test the normal distribution in the residuals along with the distplot in Linear regression. It will help us understand what kind of distribution is followed by residuals.

———

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   Count is more during summers and fall season.
   Count is more during the month of June to August as well

2. **Why is it important to use drop_first=True during dummy variable creation?**

   It is used to drop one of dummy columns and should be dropped to avoid multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   After dropping ‘registered’ variable as it is linked to ‘cnt’, ‘temp’/‘atemp’ has more correlation with ‘cnt’.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Validated that linear relationship exists between the variables.
   Residuals are normally distributed.
   Variance of error terms is the same for any value of X

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes**

   Temperature
   Year
   Month of Sep