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

Ans-

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

Ans-It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.
•For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it.
•It is also used to reduce the collinearity between dummy variables .

3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans . •atemp and temp both have the same correlation with target variable of 0.63 which is the highest among all numerical variables.

4.How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans . •Linearity of the relationship between response and predictor variables.
•Normality of the error distribution (Normal distribution of error terms).
•Constant variance of the errors or Homoscedasticity.
•Less Multi-collinearity between features (Low VIF).

5.Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans . •temp
•light_rain_snow
•Sept

1) **Explain the linear regression algorithm in detail.**

Ans- Linear Regression is a machine-learning algorithm that is based on the supervised learning category. It finds the best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables.

Linear regression is of the 2 types:

i. **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

The formula for the Simple Linear Regression:

$Y=\beta0+\beta1X1+\epsilon$

ii. Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values of independent variables in X. It fits a 'hyperplane' instead of a straight line.

The formula for the Multiple Linear Regression:

$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$

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

Ans Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x ,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. The first scatter plot (top left) appears to be a simple linear regression, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and Pearson's Correlation Coefficient is not relevant. In the third graph (bottom left), the distribution is linear, but with a different regression line, which is offset by the one outlier, which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows another example of when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

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

Ans- Pearson's R was developed by **Karl Pearson** and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Pearson's correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations.

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

Ans Scaling is the process of normalizing the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range. The two most discussed scaling methods are Normalization and Standardization. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula for normalizing scaling = x = x-min(x) / max(x)-min(x)
Formula for Standardized Scaling =x = x-mean(x) / sd(x)

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans- The value of VIF is calculated by the below formula:
VIFi =1/1-R1^2
Where 'i' refers to the ith variable.
If the R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value becomes infinite. It denotes perfect correlation in variables.

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

1. The sample sizes do not need to be equal.
2. The q-q plot can provide more insight into the nature of the difference than analytical methods.