Ques: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans: From my analysis, I find out that fall season seems to have a more bookings

Ques: Why is it important to use drop_first=True during dummy variable creation?

ns: Using drop_first=True during dummy variable creation helps to avoid multicollinearity by removing one of the dummy variables from each set, ensuring linear independence among the remaining variables.

Ques: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: temp variable has the highest correlation with the target variable

Ques: How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans: I have validated the assumptions using Normality of error and multicollinearity

Ques: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans: temp, windspeed, september month

Ques: Explain the linear regression algorithm in detail.
Ans: Linear regression is a supervised learning algorithm that predicts a continuous target variable based on one or more predictor variables. It assumes a linear relationship between the predictors and the target variable. The algorithm finds the best-fitting line or hyperplane to minimize the difference between predicted and actual values. The model is trained by optimizing the coefficients using techniques like gradient descent. Evaluation metrics such as R-squared and MSE are used to assess its performance. Once trained, the model can make predictions on new data by applying the learned coefficients.

Ques: Explain the Anscombe’s quartet in detail.
Ans: Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but exhibit drastically different patterns when visualized. Each dataset consists of 11 (x, y) points, and when analyzed individually, they have nearly identical means, variances, correlations, and linear regression parameters. However, upon visual inspection, the quartet reveals that the data points are distributed differently, showcasing the importance of visualizing data rather than relying solely on summary statistics. Anscombe's quartet highlights the limitations of relying solely on numerical summaries and emphasizes the need for graphical analysis to gain a comprehensive understanding of data.

Ques: What is Pearson’s R?
Ans: Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges between -1 and +1.

The Pearson's R coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is a standardized measure, meaning it is not affected by the scale of the variables.

The value of Pearson's R indicates the strength and direction of the relationship between the variables:

If the value is close to +1, it indicates a strong positive linear relationship, meaning that as one variable increases, the other variable tends to increase proportionally.
If the value is close to -1, it indicates a strong negative linear relationship, meaning that as one variable increases, the other variable tends to decrease proportionally.
If the value is close to 0, it indicates a weak or no linear relationship, meaning that there is little to no association between the variables.

Ques: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is the process of transforming numerical variables to a common scale. It is performed to ensure variables have comparable ranges. Normalization scales values to a specific range (e.g., 0-1), while standardization transforms data to have a mean of 0 and standard deviation of 1. Normalization preserves the shape of the distribution, while standardization centers the data around zero.

Ques: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: The phenomenon of obtaining an infinite value for the Variance Inflation Factor (VIF) typically occurs when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity means that one or more independent variables can be perfectly predicted using a linear combination of other variables in the model.

When perfect multicollinearity exists, it results in the inversion of the matrix used in the calculation of VIF. In such cases, the determinant of the matrix becomes zero, leading to an infinite value for VIF.

Perfect multicollinearity is problematic because it violates the assumptions of linear regression. It means that the relationship between the variables is not identifiable, and it becomes impossible to estimate the individual effects of the correlated variables on the target variable.

To address this issue, it is necessary to identify and handle the variables causing perfect multicollinearity. This can be done by removing one or more of the highly correlated variables or finding alternative ways to represent the relationship among the variables.

Ques: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
Ans: A Q-Q plot is a graph that compares the quantiles of observed data to the quantiles of a theoretical distribution, usually the normal distribution. In linear regression, a Q-Q plot is used to assess the assumption of normality for residuals or the target variable. It helps to visually check if the data follows a normal distribution and detect departures from normality. By analyzing the Q-Q plot, we can evaluate the performance and validity of the linear regression model.