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

ANS. The most common method for including categorical data in regressions is to create dummy variables for each possible category .when using this method:
one reference category must be excluded to avoid perfect multicollinearity.

The impact of each level on the dependent variables is in relationship to the reference level.

Example of categorical variable: dummy variables

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

ANS. Is important to use ,as it helps in reducing the extra column created during dummy variable Creation. Hence it reduces the correlations created among dummy variables

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

ANS. Temp

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

ANS. The simple way to determine if this assumption is met or not is by creating a scatter plot x vs Y .If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

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

ANS. thu , sat and holiday

1. Explain the linear regression algorithm in detail?

ANS. Linear regression is a machine learning algorithm based on supervised learning . It performs A regression task. Regression models a target prediction value based on independent Variables. It mostly used for finding out the relationship between variables and the Foresating.

2. Explain the Anscombe’s quartet in detail.

ANS. Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics,but there are some peculiarities in the dataset that fools the regression model if built.They have very different distributions and appear differently when plotted on scatter plots.

2. What is Pearson's R?

ANS. In statistics , the pearson correlation coefficient (PCC, pronounced /)-also known as Pearson product- moment correlation coefficient(PPMCC), the bivariate correlation Coefficient, or colloquially simply as the correlation simply as the correlation coefficient- Is a measure of linear correlation between two sets of data.

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

ANS. Scaling refers to putting the features values into the same range. Scalling is extremely Important for the algorithms considering the distances between observations like k-nearest neighbors. A normalized scalling will always that ranges between 0 and 1. A standardizied scalling of mean 0 and standard deviation 1, but no maximum and minimum values.

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

ANS. The correlation is perfect then VIF=infinity. That shows a perfect correlation between two Independent variables. In the case of perfect correlation, we get R2=1, which lead to 1/(1-R2) Infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

ANS. Q-Q plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantities against each other. A point on the plot corresponds to one of the quantities of the second distribution plotted against the same quantitile of first distribution. It is a graphical tool to help us access if a set of data plausibly came from some theoretical distribution such as a Normal, exponential.