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

Ans: Following observations are made for the dependent variable 'cnt', based on the categorical variables-

1) Season 'Falls' has the most number of shared bike users and season 'spring' has the least.
2) The count of shared bike users are maximum from June to October months on an average.
3) Users don't prefer shared bikes during bad weather, but on a clear day the number of shared bike users increase.
   That is on a cloudy day the chances of using shared bike is very less compared to a sunny day.

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

Ans: During dummy variable creation, all the categories of a feature are identified separately by unique values. Therefore, if one dummy variable is removed then that can be easily identified by the non-existent value which would represent the dummy variable.

By removing one dummy variable, less memory is used and the result remains the same.

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

Ans: Temperature feature 'atemp' has the highest correlation with the target variable 'cnt', with the correlation of 0.99.

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

Ans: I have validated the assumption of linear regression based on the R2 score for test data set, which was coming out to be around 71% accuracy.

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

Ans: The top three features that explains the demand of shared bikes are-

a) 'temp' – temperature of the day
b) 'workingday'
c) 'year'

1. Explain Linear regression algorithm in detail.

Ans: Linear regression is a supervised machine learning algorithm used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that the dependent variable can be expressed as a linear combination of the independent variables, with some random error.

The algorithm aims to find the best-fitting line that represents the relationship between the independent variables (also called features or predictors) and the dependent variable (also called the target variable). This line is represented by the equation:

$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$

The goal of linear regression is to estimate the values of β_o, β_1, β_2, ..., β_p that minimize the sum of squared residuals between the observed dependent variable values and the predicted values from the linear equation.

To accomplish this, the algorithm typically uses the method of least squares, which minimizes the sum of the squared differences between the actual dependent variable values and the predicted values. The squared differences are summed up across all the data points, and the coefficients are adjusted iteratively to minimize this sum.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a collection of four datasets that were created by the statistician Francis Anscombe in 1973. The quartet consists of four sets of x and y values, and despite having very different patterns and characteristics, they have nearly identical summary statistics. Anscombe's quartet is often used to emphasize the importance of visualizing data and not relying solely on summary statistics.

The key takeaway from Anscombe's quartet is that summary statistics such as mean, variance, and correlation can be misleading when used in isolation. Although all four datasets have the same summary statistics, they have distinct patterns and relationships when visualized. It highlights the importance of data visualization and exploratory analysis to gain a deeper understanding of the underlying data structure.

3. What is Pearson's R?

Ans: Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a statistical measure that quantifies the linear relationship between two continuous variables. It measures the strength and direction of the linear association between the variables, ranging from -1 to 1.

The formula for Pearson's R is as follows:

$R = (\Sigma[(X_i - \bar{Y})(Y_i - \bar{Y})]) / \sqrt{(\Sigma(X_i - \bar{Y})^2 * \Sigma(Y_i - \bar{Y})^2)}$

Pearson's correlation coefficient is widely used in various fields, including statistics, social sciences, finance, and data analysis. It helps to assess the degree of association between two variables, enabling researchers and analysts to understand and quantify the relationship between them. However, it should be noted that correlation does not imply causation, and other factors may be involved in the relationship between variables.

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

Ans: Scaling, in the context of data preprocessing in machine learning, refers to the process of transforming the values of variables to a specific range or distribution. It is performed to ensure that all variables are on a similar scale, as it can greatly impact the performance of certain machine learning algorithms.

Scaling is necessary because many machine learning algorithms are sensitive to the relative magnitudes of the variables. When variables are on different scales, algorithms that rely on distance calculations or gradient-based optimization may be biased towards features with larger scales. Scaling helps to mitigate this issue and ensures that all variables contribute equally to the analysis.

There are two common types of scaling techniques: normalized scaling and standardized scaling.

1. Normalized Scaling (or Min-Max Scaling):

Normalized scaling, also known as min-max scaling, transforms the values of variables to a specific range, typically between 0 and 1. The formula for normalized scaling is as follows:

X_scaled = (X - X_min) / (X_max - X_min)

Normalized scaling preserves the original distribution of the data but ensures that the variable values fall within the specified range.

2. Standardized Scaling (or Z-score Scaling):

Standardized scaling, also known as z-score scaling or standardization, transforms the values of variables to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is as follows:

X_scaled = (X - μ) / σ

Standardized scaling transforms the data to have zero mean and equalizes the spread of the data across all variables. It makes the variables comparable and suitable for algorithms that assume a standard normal distribution.

The main difference between normalized scaling and standardized scaling lies in the resulting distribution. Normalized scaling preserves the original distribution and range of the data, while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1, irrespective of the original distribution.

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

Ans: VIF assesses the degree to which the variance of the estimated regression coefficient is inflated due to multicollinearity.

In some cases, the VIF value can be calculated as infinite. This happens when there is perfect multicollinearity present in the regression model. Perfect multicollinearity occurs when one or more independent variables in the model can be perfectly predicted by a linear combination of other independent variables. In such a scenario, the VIF for the variable(s) involved becomes infinite.

Perfect multicollinearity leads to problems in regression analysis because it violates the assumptions of the regression model. When variables are perfectly correlated, it becomes impossible to estimate unique regression coefficients for each variable. Consequently, the estimation process fails, resulting in an infinite VIF value.

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

Ans: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a given dataset follows a specific theoretical distribution. It compares the quantiles of the observed data against the quantiles of the expected distribution, usually the normal distribution. Q-Q plots are commonly used in statistics and data analysis to evaluate the assumption of normality and to identify departures from normality.

The Q-Q plot is created by plotting the sorted values of the observed data against the expected quantiles from the theoretical distribution. If the data perfectly follows the expected distribution, the points on the plot will fall along a straight line. Deviations from a straight line suggest departures from the expected distribution.