Assignment-based Subjective Questions

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

   To analyze the effect of categorical variables on the dependent variable, we can consider the following

   Season:

   - The season variable indicates the season of the year (spring, summer, fall, winter).
   - It can be inferred that seasons may have an impact on bike sharing demand, as different seasons might affect people's willingness to ride bikes.
   - For instance, bike rentals might be higher during summer compared to winter.

   Weathersit:

   - The weathersit variable represents different weather conditions.
   - It indicates that weather can have an impact on bike sharing demand.
   - Different weather conditions, such as clear skies, misty weather, or rain, might affect people's willingness to ride bikes.

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

   Setting drop_first=True during dummy variable creation helps address multicollinearity issues, enhances interpretability by establishing a clear reference category, and improves the model's efficiency by reducing unnecessary variables.

   Multicollinearity:
   Multicollinearity occurs when there is a high correlation between two or more predictor variables in a regression model.

   Interpretability:
   When using dummy variables, the interpretation of the coefficients becomes easier when drop_first=True.

   Model efficiency:
   By dropping the first category, you reduce the number of variables in the model, which can improve computational efficiency.

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

   From the given pair plot, we could observe that temp has the highest positive correlation with target variable cnt.

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

   Validating the assumptions of linear regression is an essential step after building the model on the training set. Here are some common methods to validate the assumptions:

   a. Linearity: Plot the observed values of the dependent variable against the predicted values from the model. If the relationship between the predicted and observed values follows a roughly linear pattern, it suggests that the linearity assumption is reasonable.
   b. Normality of Residuals: Examine the distribution of the residuals (the differences between the observed and predicted values).Plot a histogram of the residuals and check if it resembles a normal distribution.
   c. Homoscedasticity: Plot the residuals against the predicted values or any independent variables. If the spread of the residuals appears roughly constant across all predicted values or independent variables, it suggests that the assumption of homoscedasticity holds.
   d. Multicollinearity: Check for high correlation between independent variables by examining the correlation matrix or variance inflation factor (VIF) values. High VIF values indicate the presence of multicollinearity, which can affect the interpretation and stability of coefficient estimates.

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

   From the final model, the top 3 predictor variables that contribute the bike sharing are:
   a. Temperature (temp), b. Year (yr), c. Humidity (hum)

General Subjective Questions

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

   Linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or more independent variables. It assumes a linear relationship between the input variables and the target variable. The algorithm works by finding the best-fitting line that minimizes the sum of squared differences between the predicted and actual target values.

   In detail, the linear regression algorithm involves the following steps:

   Data Preparation: Collect and preprocess the dataset, ensuring variables are numeric and handling missing values, outliers, and categorical variables if needed.

   Model Training: Fit the linear regression model to the training data by estimating the coefficients of the regression equation using a method like Ordinary Least Squares (OLS).

   Coefficient Estimation: Determine the coefficients (slopes) and the intercept of the linear regression equation by minimizing the sum of squared residuals between the predicted and actual target values.

   Model Evaluation: Assess the performance of the model using evaluation metrics like mean

squared error (MSE), root mean squared error (RMSE), R-squared, or adjusted R-squared to measure the goodness of fit.

Prediction: Use the trained model to make predictions on new or unseen data by plugging in the values of the independent variables into the regression equation.

Linear regression can be extended to handle multiple independent variables (multivariate linear regression) or accommodate nonlinear relationships through feature engineering or using polynomial regression. Additionally, techniques like regularization (e.g., Lasso or Ridge regression) can be applied to mitigate overfitting and improve the model's generalization ability.

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

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, including means, variances, correlations, and regression lines, but exhibit vastly different graphical representations. The quartet highlights the importance of data visualization and challenges the reliance on summary statistics alone.
The four datasets in Anscombe's quartet consist of pairs of x and y values. Here's a detailed description of each dataset:

Dataset I:
- It forms a linear relationship between x and y, following the equation y = 3x + 2.
- The data points closely align along the regression line, exhibiting a strong linear correlation.

Dataset II:
- It represents a non-linear relationship between x and y, resembling a quadratic curve.
- Despite the non-linear relationship, the summary statistics (mean, variance, correlation) closely match Dataset I.

Dataset III:
- It consists of a few distinct clusters of data points with the same x-values but different y-values.
- The data points form a step-like pattern and show no correlation between x and y, yet summary statistics suggest a linear relationship.

Dataset IV:
- It contains an outlier data point that significantly influences the linear regression line.
- Removing the outlier drastically changes the regression line, revealing a different relationship between x and y.

Anscombe's quartet demonstrates the limitations of relying solely on numerical summaries and the importance of visualizing data to gain a comprehensive understanding of its patterns and relationships. It emphasizes the need for exploratory data analysis and graphical representation to complement statistical analysis.

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

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 measures the degree of association between the variables on a scale from -1 to +1.
The Pearson's R coefficient is calculated using the following formula:

$$r = (\Sigma((x_i - \bar{x}) * (y_i - \bar{y}))) / \sqrt{((\Sigma(x_i - \bar{x})^2) * (\Sigma(y_i - \bar{y})^2)}$$

where:
r: Correlation coefficient
xi: Values of the x-variable in a sample
$\bar{x}$: Mean of the values of the x-variable
yi: Values of the y-variable in a sample
$\bar{y}$: Mean of the values of the y-variable

The resulting Pearson's R value indicates the strength and direction of the linear relationship:
R close to +1 indicates a strong positive linear relationship.
R close to -1 indicates a strong negative linear relationship.
R close to 0 indicates a weak or no linear relationship.
Pearson's R is widely used in statistics and data analysis to assess the correlation between variables, helping to understand the degree to which changes in one variable are associated with changes in the other.

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

Scaling, in the context of data preprocessing, refers to the process of transforming variables to a specific range or distribution. It is performed to ensure that all variables are on a similar scale, which can have several benefits for various machine learning algorithms.

There are two common scaling techniques: normalized scaling and standardized scaling.
Normalized Scaling (Min-Max Scaling):
- Normalization scales the variables to a specific range, typically between 0 and 1.
- The formula for normalized scaling is: X_normalized = (X - X_min) / (X_max - X_min), where X is the original value, X_min is the minimum value, and X_max is the maximum value.
- Normalization preserves the relative relationships between the data points and maintains the distribution shape.

Standardized Scaling (Z-score Scaling):
- Standardization transforms variables to have a mean of 0 and a standard deviation of 1.
- The formula for standardization is: X_standardized = (X - X_mean) / X_std, where X is the original value, X_mean is the mean value, and X_std is the standard deviation.
- Standardization centers the data around zero and adjusts the spread of the data.

Standardization is useful when the distribution of the variable is not necessarily normal and when there may be outliers in the data.

The choice between normalized scaling and standardized scaling depends on the specific requirements of the problem and the characteristics of the data. Normalization is typically more suitable when the data distribution is bounded, while standardization is more robust to outliers and works well with various distributions.

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

The occurrence of infinite VIF values happens when perfect multicollinearity exists in the data. Perfect multicollinearity refers to a situation where one or more independent variables can be perfectly predicted by a linear combination of other variables. This results in an exact relationship between variables, making it impossible to calculate the VIF. Perfect multicollinearity can arise from including redundant variables or data coding errors. It is important to detect and address multicollinearity as it affects the stability and interpretability of regression models.

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess the distributional similarity between a dataset and a theoretical distribution, typically the normal distribution. It compares the quantiles of the dataset against the quantiles expected from the theoretical distribution.
In linear regression, a Q-Q plot is important for checking the assumption of normality of the residuals. The residuals are the differences between the observed values and the predicted values from the linear regression model. The Q-Q plot of the residuals allows us to visually inspect whether the residuals follow a normal distribution.

The use and importance of a Q-Q plot in linear regression are as follows:

Assessing Normality: By plotting the quantiles of the residuals against the expected quantiles from a normal distribution, a Q-Q plot can reveal any deviations from normality.

Identifying Outliers: A Q-Q plot can also help identify outliers in the residuals. Outliers appear as points that deviate significantly from the expected line.

Model Assessment: The normality assumption is important for valid inference and hypothesis testing in linear regression.