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?**

   Ans - From the analysis of the categorical variables in the dataset and the coefficients obtained from the best-fitted line equation, we can infer the following effects on the dependent variable (bike demand):

   a) Year (yr):

   The variable "yr" represents the year, and its coefficient is 0.2358. It indicates that each year increase leads to an increase in bike demand by approximately 0.2368 units. This suggests that bike demand has been increasing over time, possibly due to factors like the increased popularity of biking, improved infrastructure, or raising awareness about environmental benefits.

   b) Holiday:

   The variable "holiday" takes binary values (0 or 1) to represent regular days and holidays, respectively. Its coefficient is -0.1002, indicating that bike demand is slightly lower on holidays compared to regular days. This could be due to people having different travel patterns or preferring alternative modes of transportation during holidays.

   c) Seasons:

   The categorical variables "season_spring," "season_winter," and "season_summer" represent different seasons. Their respective coefficients are -0.0855, 0.0566, and 0.0429. These coefficients show the impact of each season on bike demand compared to the base season (season_fall, not explicitly shown in the equation). It suggests that bike demand is lower during spring compared to fall, slightly higher during winter, and more significantly higher during summer.

   d) Weather Situations:

   The categorical variables "weathersit_light_snow" and "weathersit_mist" represent different weather conditions. Their respective coefficients are -0.2950 and -0.0837. These coefficients indicate that both light snow and mist have a negative impact on bike demand. People are less likely to rent bikes during these weather situations, which is reasonable as adverse weather conditions can make biking less appealing and challenging.

   e) Day of the Week (Sunday):

The variable "Sunday" represents whether it's a Sunday or not. Its coefficient is -0.0512, suggesting that bike demand is slightly lower on Sundays compared to other days of the week. This may be due to different travel patterns or activities on weekends.

the analysis of categorical variables provides insights into their effects on bike demand. Different factors such as the year, holidays, seasons, weather conditions, and days of the week play significant roles in influencing the number of bike rentals. These findings can help businesses and bike-sharing services optimize their operations and tailor their services and marketing strategies based on the varying demands influenced by these categorical factors.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Ans - Using `drop_first=True` during dummy variable creation is essential to avoid multicollinearity issues in regression analysis. When creating dummy variables, one category within each categorical variable is designated as the reference category. The reference category is excluded from the set of dummy variables to prevent perfect multicollinearity, which occurs when one dummy variable is a perfect linear combination of others. By dropping the first dummy variable, we create a set of independent dummy variables that represent all but the reference category. This ensures that each category is uniquely represented, and the regression model can estimate the individual effects of each category on the dependent variable without redundancy.

3. **Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable?**
   Ans - Based on your analysis of the pair plot among the numerical variables with the target variable "cnt," it appears that the "temp" and "atemp" columns have the highest correlation with the target variable. The "temp" and "atemp" columns represent temperature-related variables, and their similarity in the pair plot indicates a strong positive correlation with the target variable "cnt," which represents bike rentals. This correlation suggests that higher values of temperature (both actual temperature "temp" and "feeling-like" temperature "atemp") are associated with increased bike demand. As the temperature rises, people are more likely to use bikes for transportation or leisure activities, leading to higher bike rentals.

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

Ans – After building the linear regression model on the training set, it is crucial to validate the assumptions of linear regression to ensure the model's reliability and accuracy. The following assumptions can be assessed:

a) Linearity: This assumption can be validated by examining the scatter plots of the dependent variable and independent variables. If the scatter plots show a linear pattern, the assumption of linearity is satisfied.

b) Homoscedasticity: Homoscedasticity refers to the constant variance of residuals across the range of predicted values. This assumption can be checked by plotting the residuals against the predicted values. If the scatter plot of residuals exhibits a random pattern without any distinct funnel shape or increasing/decreasing variance, the assumption of homoscedasticity is met.

c) Independence of residuals: The independence of residuals assumes that the residuals are not correlated with each other. This assumption can be assessed by examining the Durbin-Watson statistic, which measures the autocorrelation of residuals. A value close to 2 suggests no significant autocorrelation, indicating the independence of residuals.

d) Normality of residuals: The normality assumption can be evaluated by plotting a histogram or a Q-Q plot of the residuals. If the residuals follow a bell-shaped or approximately normal distribution, the assumption of normality is satisfied.

By validating these assumptions, we can ensure that the linear regression model is valid and reliable for making accurate predictions.

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

The top three features contributing significantly towards explaining the demand for shared bikes, we can refer to the absolute magnitudes of the coefficients in the best-fitted line equation. Larger absolute coefficients indicate stronger effects on the target variable (bike demand).

Based on the best-fitted line equation:

cnt = 0.2368 + (0.2358 * yr) + (-0.1002 * holiday) + (0.4535 * atemp) + (-0.1340 * windspeed) + (-0.0855 * season_spring) + (0.0566 * season_winter) + (-0.2950 * weathersit_light_snow) + (-0.0837 * weathersit_mist) + (-0.0512 * Sunday) + (0.0893 * 9) + (0.0429 * season_summer) + (0.0430 * 10)

The top three features contributing significantly to explaining the demand for shared bikes are:

a) Temperature (`atemp`): The coefficient of 0.4535 for `atemp` suggests that temperature (feeling-like temperature) has the most substantial positive impact on bike demand. For every unit increase in "atemp," bike demand is expected to increase by approximately 0.4535 units. Higher temperatures make biking more enjoyable and comfortable, leading to increased bike rentals.

b) Year (`yr`): The coefficient of 0.2358 for `yr` indicates that the year has a significant positive impact on bike demand. For every year increase, bike demand is expected to increase by approximately 0.2358 units. This suggests that bike rentals have been increasing over time, possibly due to factors like the increased popularity of biking and improved biking infrastructure.

c) Month (`9` or September): The coefficient of 0.0893 for the month `9` (likely representing September) suggests that September has a notable positive impact on bike demand. For every observation corresponding to September, bike demand is expected to increase by approximately 0.0893 units. This could be due to favorable weather conditions and events in September that encourage people to use bikes more frequently.

These three features (temperature, year, and the month of September) stand out as the most significant contributors to explaining the demand for shared bikes based on the coefficients in the final model.

General Subjective Questions:

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

Ans - 1. Linear regression is a supervised machine learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. The algorithm aims to find the best-fitting linear relationship between the independent variables and the dependent variable. Here is a detailed explanation of the linear regression algorithm:

a) Simple Linear Regression: In simple linear regression, a single independent variable is used to predict the dependent variable. The algorithm calculates the slope and intercept of the best-fitting line that minimizes the sum of squared differences between the predicted values and the actual values.

b) Multiple Linear Regression: In multiple linear regression, multiple independent variables are used to predict the dependent variable. The algorithm estimates the coefficients of each independent variable, representing their impact on the dependent variable while accounting for the relationships with other variables.

The algorithm employs the ordinary least squares (OLS) method to estimate the coefficients. It calculates the residuals (differences between predicted and actual values) and minimizes the sum of squared residuals to find the best-fitting line.

Assumptions of linear regression include linearity (linear relationship between variables), homoscedasticity (constant variance of residuals), independence of residuals, and normality of residuals.

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

   Ans - Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, despite having different distributions and patterns. The quartet was created by the statistician Francis Anscombe to demonstrate the importance of visualizing data and not relying solely on summary statistics. The datasets highlight the limitations of relying solely on measures such as mean, variance, and correlation.

   Each dataset in the quartet consists of two variables, x, and y, and includes the same summary statistics such as mean, variance, correlation, and regression line. However, when visualized, the datasets reveal distinct patterns like linear, non-linear, outliers, or influential points. Anscombe's quartet emphasizes the need to explore and visualize data to gain a comprehensive understanding of its characteristics and relationships.

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

   Ans – Pearson's R, also known as Pearson correlation coefficient, is a measure of the linear correlation between two continuous variables. It quantifies the strength and direction of the linear relationship between variables, ranging from -1 to 1. A value of -1 indicates a perfect negative linear correlation, 0 represents no linear correlation, and 1 signifies a perfect positive linear correlation.

   Pearson's R is calculated by dividing the covariance of the variables by the product of their standard deviations. It is commonly used to assess the strength and direction of relationships in various fields, including statistics, social sciences, and finance.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Ans - Scaling refers to the process of transforming numerical variables to a specific range or distribution. It is performed to ensure that all variables are on a similar scale, preventing certain variables from dominating the analysis due to their larger magnitude. Scaling is important because some machine learning algorithms, such as gradient descent-based algorithms, are sensitive to the scale of variables. Scaling can improve algorithm convergence and make the model more robust.

   Normalized scaling (also called min-max scaling) rescales the variables to a predefined range, typically between 0 and 1. It maintains the distribution of the variables and preserves the relationships between the data points. Standardized scaling (also called z-score scaling) transforms the variables to have a mean of 0 and a standard deviation of 1. It standardizes the variables, making them comparable and centered around the mean.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Ans - The occurrence of infinite VIF (Variance Inflation Factor) values typically indicates perfect multicollinearity among the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other. In such cases, the VIF calculation becomes unstable, resulting in infinite values.

   Perfect multicollinearity disrupts the linear regression model because it becomes impossible to estimate the individual effects of highly correlated variables accurately. It leads to numerical instability and inflated standard errors. To address this issue, multicollinearity should be identified and resolved by removing one of the correlated variables or transforming the variables to reduce correlation.

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 dataset follows a particular probability distribution, typically the normal distribution. The plot compares the quantiles of the observed data against the quantiles of the expected distribution.

   In linear regression, a Q-Q plot is useful for evaluating the assumption of normality of residuals. By plotting the quantiles of the residuals against the quantiles of the expected normal distribution, we can visually assess if the residuals deviate significantly from normality. If the residuals align closely with the diagonal line in the Q-Q plot, it suggests that the residuals follow a normal distribution, satisfying the assumption of normality in linear regression.

The Q-Q plot helps identify departures from normality, such as skewness or heavy tails, and provides insights into the distributional properties of the residuals, aiding in model evaluation and potential improvements.