Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The analysis indicates that **'Clear'** weather conditions have the highest count in the dataset, suggesting a significant prevalence of clear weather. This likely correlates with increased bike rentals, as clear weather is expected to positively influence usage. Bike rentals during the fall season due to favorable weather conditions or other factors. It indicates that more bike rentals occur during this season compared to Spring, Summer and Winter.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

It avoids multicollinearity. Remains interpretable and more efficient by reducing the number of variables. Making the model simpler without losing any information.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The scatter plots between windspeed and the other two variables (temp and hum) show a clear linear pattern, indicating a strong correlation. This suggests that windspeed has a significant impact on the target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The relationship between the independent variables and the dependent variable is linear. Plotted the residuals against the predicted values. A random scatter indicates linearity. The model fits the data well. No Multicollinearity between independent variables. The histogram indicates that the linear regression model's residuals are reasonably well-behaved. The distribution is roughly symmetric, there are no major outliers, and it appears to follow a normal distribution.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Humidity and Wind Speed: Since these negatively impact demand, consider providing weatherprotective gear or improving bike design to handle such conditions better.

Seasonal and Weather-Based Promotions: Since higher temperatures are associated with higher demand, consider promoting bike rentals during warmer months. Implement discounts or special offers during colder months to balance demand. Develop strategies to mitigate the impact of bad weather on bike rentals, such as offering discounts during light rain or snow.

Yearly Trends: The demand is increasing year over year, indicating a growing market. Continue investing in expanding the service and improving infrastructure.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. The goal is to find the linear relationship that best explains the variance in the dependent variable.

Types of Linear Regression

Simple Linear Regression: Involves one independent variable and one dependent variable.

$y = mx + b + \varepsilon$

Where:

y: Dependent variable
x: Independent variable
m: Slope of the line
b: Intercept of the line
ε: Error term (residual)

Multiple Linear Regression: Involves two or more independent variables.

$y = b_0 + b_1x_1 + b_2x_2 + ... + bnXn + \varepsilon$

Polynomial Regression: A form of regression analysis where the relationship between the independent variable and the dependent variable is modeled as an n
nth degree polynomial.

Key Assumptions of Linear Regression

Linearity: The relationship between the dependent and independent variables is linear.

Independence: Observations are independent of each other.

Homoscedasticity: The residuals (errors) have constant variance at every level of the independent variables.

Normality: The residuals of the model are normally distributed.

No Multicollinearity: The independent variables are not highly correlated with each other.

Finding the Best Fit Line

The goal of linear regression is to find the values of the coefficients (β) that minimize the difference between the actual values of Y and the predicted values. This is typically done using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared residuals.

Model Evaluation:

Once the model is trained, it's crucial to evaluate its performance.

Key metrics include:

Mean Squared Error(MSE): Measures the average squared difference between predicted and actual values.

Root Mean Squared Error (RMSE): The square root of MSE, providing an error measure in the same units as the dependent variable.

Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.

R-squared (R^2): Represents the proportion of variance in the dependent variable explained by the independent variables.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Crucial Role of Visualization: Always visualize data before making conclusions to better understand its nature. Potential Misleading Nature of Summary Statistics: Relying only on numerical summaries may fail to capture the complete story behind the data. Impact of Outliers: Outliers can greatly affect the outcomes of statistical analyses, leading to potentially misleading conclusions. Variety in Relationships with Similar Statistics: Different types of relationships can yield similar summary statistics, emphasizing the need for a thorough exploration of variable relationships.

Key Points Summary: Identical Summary Statistics: The four datasets exhibit nearly the same means, variances, correlations, and linear regression lines when evaluated through numerical methods. Distinct Visual Patterns: When visualized, each dataset reveals unique patterns and relationships that differ from one another. Significance of Visualization: This demonstrates the limitation of relying solely on numerical summary statistics, as visualization uncovers hidden patterns, outliers, and the actual relationships between variables.

Overview of the Four Datasets: Linear Relationship: This dataset clearly shows a linear correlation between the x and y variables. Quadratic Relationship: Here, the y values first increase and then decrease as x increases, indicating a quadratic trend. Linear with an Outlier: While maintaining a linear relationship, this dataset includes an outlier that significantly distorts the regression line. Perfect Linear Relationship with Minimal Variation in X: This dataset illustrates a perfect linear correlation, with all x values identical except for one outlier.

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's r is a valuable tool for understanding the relationship between two continuous variables. However, it's important to remember that it only measures linear relationships and that assumptions should be met for accurate interpretation. Visualizing the data through scatter plots can also provide valuable insights into the relationship between the variables.
Range: The value of Pearson's R ranges from -1 to +1.
+1: Perfect positive correlation. As one variable increases, the other variable also increases in a perfectly linear manner.
0: No correlation. There is no linear relationship between the two variables.
-1: Perfect negative correlation. As one variable increases, the other variable decreases in a perfectly linear manner.
Interpretation:
A value of +1 indicates a perfect positive correlation, meaning that as one variable increases, the other also increases proportionally.
A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases proportionally.
A value of 0 indicates no linear relationship between the two variables.
Strength of Correlation:
The closer the value of r is to -1 or +1, the stronger the linear relationship.
Values between -0.7 and -1 or 0.7 and 1 indicate a strong correlation.
Values between -0.3 and -0.7 or 0.3 and 0.7 indicate a moderate correlation.
Values between -0.3 and 0.3 indicate a weak correlation.
Assumptions for Pearson's r:
Linearity: The relationship between the two variables should be linear.
Normality: Both variables should be normally distributed.
Homoscedasticity: The variance of one variable should be constant across all values of the other variable.
Independence: The observations should be independent of each other.
When to use Pearson's r:
When you want to measure the strength and direction of a linear relationship between two continuous variables.
When the assumptions of linearity, normality, and homoscedasticity are met.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling adjusts the range or distribution of numerical features so they can be analyzed or compared on the same scale. Ensures that features contribute equally, preventing dominance by features with larger magnitudes. Features are on a similar scale, leading to unbiased model training.

Normalization:

- When we know the exact range of the data.
- When we want to preserve the original distribution shape.
- When we have outliers that might significantly affect the scaling range.

Standardization:

- No idea about the exact range of the data.
- To reduce the influence of outliers.
- When we want to assume a normal distribution of features.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

A VIF (Variance Inflation Factor) value of infinity signifies the presence of perfect multicollinearity in a regression model, indicating that at least one independent variable can be completely predicted by a linear combination of other variables.
Why does this happen:
It can happen due to Duplicate Variables. Sometimes it may occur because of derived variables, one variable being a mapped with another variable. Mistakes leading to redundant or duplicate information in the dataset.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

Q-Q Plot: A Visual Check for Normality

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specific distribution, often a normal distribution. In the context of linear regression, it's particularly useful for evaluating the assumption of normality of the residuals.

Why is Normality Important in Linear Regression?

Valid Statistical Inference: Many statistical tests and confidence intervals rely on the assumption of normally distributed residuals.
Accurate Model Predictions: Non-normal residuals can lead to biased or inefficient predictions.
How to Interpret a Q-Q Plot:

Perfect Fit: If the data points closely follow a straight line, the data is likely normally distributed.
Deviations from Normality: Deviations from the line indicate non-normality, which can affect the model's accuracy and reliability.
By examining the Q-Q plot of the residuals, you can:

Validate Model Assumptions: Ensure that the model meets the necessary assumptions for valid statistical inference.
Identify Outliers: Detect outliers that can significantly impact the model's performance.
Assess Model Fit: Evaluate whether the model is a good fit for the data.
In essence, Q-Q plots are a valuable tool for assessing the normality assumption of linear regression models. By understanding their interpretation, you can make informed decisions about the reliability and accuracy of your model.