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 categorical variables influence count as follows:

Year (yr): Positive impact (+0.2327) indicates increased demand over time.
Holiday: Negative impact (-0.1062) suggests reduced demand on holidays.
Weather Conditions:
Light snow/rain: Strong negative impact (-0.2877).
Misty: Moderate negative impact (-0.0802).
Seasonality:
Summer (+0.0882) and Winter (+0.1313) see higher demand than Spring.
September (Sep): Increased demand (+0.0995).
Sunday (Sun): Slightly lower demand (-0.0497).

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)
Using drop_first=True prevents multicollinearity
When we create dummy variables, it generates one variable for each category. If all categories are included, some variables will be highly correlated, leading to redundancy in the model. By dropping the first category, we avoid this issue, also prevents the dummy variable trap, where one category can be perfectly predicted by the others.

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)

From the pair-plot among the numerical variables, temp (temperature) shows the highest correlation with the target variable count.

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)

To validate the assumptions of Linear Regression after building the model on the training set, I performed the following checks:

Linearity: I plotted the residuals versus the predicted values to check for any patterns. A random scatter indicates a linear relationship between the independent and dependent variables.

Homoscedasticity: I examined the residuals plot to ensure that the variance of residuals remains constant across all levels of the predicted values. This confirms that the error terms do not exhibit heteroscedasticity.

Normality of residuals: The plot of error terms show a normal distribution, with most errors concentrated around zero. This suggests that the error terms are normally distributed.

No Multicollinearity: I calculated the Variance Inflation Factor (VIF) for each feature to ensure that no features have high multicollinearity. VIF values greater than 5 were investigated further.

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)

The top 3 features significantly contributing to shared bike demand are:

Temperature (temp): Strong positive impact (0.5475).
Year (yr): Positive effect indicating increasing demand over time (0.2327).
Holiday (holiday): Negative impact on demand (-0.1062), as demand drops on holidays.

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 used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The formula for linear regression is:

$Y=\beta 0+\beta 1X1 + \beta 2X2 + ...+ \beta nXn + \epsilon$
Where:
Y is the dependent variable (target),
X1, X2, ..., Xn are the independent variables (features),
β0 is the intercept,
β1, β2, ..., βn are the coefficients for the predictors,
ε is the error term (residual).
The algorithm estimates the coefficients β0, β1, ..., βn using the Ordinary Least Squares (OLS)

method, which minimizes the sum of squared residuals (differences between predicted and actual values). The model assumes a linear relationship between the target and the predictors, and it requires the assumptions of linearity, independence, homoscedasticity, and normality of errors.

Once the model is trained, performance is evaluated using metrics like R² (explained variance) and Mean Squared Error (MSE). The coefficients of the model can be interpreted as the impact of each predictor on the target variable, assuming all other variables remain constant.

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>

Anscombe's Quartet consists of four datasets with identical summary statistics (mean, variance, correlation, and regression results) but different distributions when plotted. It highlights the importance of visualizing data, as statistical measures alone can be misleading. The datasets show varying relationships: one is linear, another has curvature, one is influenced by an outlier, and the last shows no real relationship, demonstrating the need for careful data exploration beyond summary statistics.

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, or Pearson's correlation coefficient, measures the linear relationship between two variables. It ranges from -1 to 1:

+1 indicates a perfect positive linear relationship,
-1 indicates a perfect negative linear relationship,
0 indicates no linear relationship.

Pearson's R quantifies how well the change in one variable can be predicted by the change in another, assuming a linear relationship. A higher absolute value indicates a stronger linear correlation.

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 feature values to a similar range to improve model performance, especially for algorithms that rely on distance or optimization.

Why scale? To prevent features with larger values from dominating the model and to improve convergence speed.
Normalized Scaling (Min-Max): Rescales values to a range [0, 1]. Sensitive to outliers.
Standardized Scaling (Z-score): Rescales values to have mean 0 and standard deviation 1. More robust to outliers.

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>
Yes, that happened with me during the assignment. The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity in the dataset. This means that one or more independent variables are perfectly correlated with each other, causing redundancy.
One variable can be exactly predicted by a linear combination of others, leading to a singular matrix during calculations.
As a result, the VIF for those variables goes to infinity, indicating that they are highly dependent on each other and should be removed or combined to avoid issues in the model.
In short, infinite VIF occurs when there's exact linear dependence between predictors.

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>

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular distribution, typically the normal distribution. In a Q-Q plot:
Theoretical quantiles of a reference distribution (e.g., normal distribution) are plotted against empirical quantiles of the observed data.
If the data follows the reference distribution, the points will lie along a straight line.
Use and Importance in Linear Regression:
Normality of Errors: In linear regression, one assumption is that the error terms (residuals) are normally distributed. A Q-Q plot helps visually check this assumption.
Model Validity: If the residuals are not normally distributed, it could indicate model misspecification, heteroscedasticity, or the presence of outliers, which could affect the reliability of regression results.
