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)

After the categorical variables were converted to dummy variables having binary values we can see from the correlation heatmap that variable 'yr' and dummy variable 'Jan' & 'Spring' have high correlation with the dependent variable.

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 is important to use drop_first=True because all the possible states of an n dimensional variable can be inferred by an n-1 dummy variable. If we don't drop one of the dummy variables it would have multicollinearity with other since it can be predicted using other variables i.e. if all the others are 0 it has to be 1.

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)
'temp' and 'atmep'

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)

We can do it by doing residual analysis i.e. we can check the difference between predicted y values and actual y values. We can see if these errors are normally distributed by plotting a distplot.

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)
Top 3 features are :

1. 'temp'
2. 'yr'
3. 'snow'

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 aims to find the best-fitting straight line through a set of data points. This line is called the regression line. The most common method to find the best-fitting line is the least squares method. This method minimizes the sum of the squared differences between the observed values and the values predicted by the model. To evaluate the performance of a linear regression model, we use metrics such as:

R-squared: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.
Root Mean Squared Error (RMSE): Measures the average magnitude of the errors.

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 nearly identical statistical properties (mean, variance, correlation, and regression line) but very different graphical representations. Created by Francis Anscombe in 1973, it demonstrates the importance of visualizing data before analysis.

Identical Statistics:

- Same mean, variance, correlation, and regression line for all datasets.

Different Graphs:

- Dataset I: Linear relationship.
- Dataset II: Non-linear relationship.
- Dataset III: Linear with an outlier.
- Dataset IV: High-leverage point affecting correlation.

Importance

- Visualization: Graphs reveal patterns and anomalies that statistics alone might miss.
- Outliers: Identifying outliers is crucial as they can skew results.
- Model Choice: Graphical analysis helps in choosing the appropriate statistical model.

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)

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. The range of the R value is -1 to 1 with -1 depicting a strong negative correlation and +1 depicting a strong positive correlation.

<Your answer for Question 8 goes here>

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)

Scaling is adjusting the range of feature values in a dataset so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.

Purpose:

1. Equal contribution of features.
2. Improved convergence of algorithms.
3. Enhanced model performance.
4. Better interpretability.

Normalization scales data to a fixed range, typically [0, 1], Standardization centers data around the mean (0) and scales it to unit variance. Normalization is useful for non-Gaussian distributions, while standardization is ideal for Gaussian distributions.

<Your answer for Question 9 goes here>

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)

An infinite VIF value occurs when an independent variable can be perfectly predicted by other variables in the model. This happens when the R-squared value in the VIF formula approaches 1, indicating perfect multicollinearity. Essentially, it means there's a perfect linear relationship between the variables, making the regression coefficients unstable.

.

<Your answer for Question 10 goes here>

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)

A Q-Q plot is a graphical tool to compare the quantiles of a dataset against the quantiles of a theoretical distribution, often the normal distribution. In linear regression, it's used to check if residuals follow a normal distribution.

<Your answer for Question 11 goes here>