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)

- Seasonal Impact: Demand for shared bikes is higher during the fall season, indicating a seasonal preference among users.
- Yearly Trend: The demand for shared bikes has increased significantly in year 1 (2019).
- Monthly Trend: July experiences higher median demand compared to other months, suggesting a possible peak usage period.
- Holiday Effect: The median demand decreases on holidays.
- Weekday and Working Day Effect: The median demand remains consistent across all weekdays and working days.
- Weather Influence: There is no demand during heavy rain or snowfall, while demand peaks under clear weather conditions.

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)

This is because for categorical variables with 'n' levels, only 'n-1' dummy variables are required to represent them numerically. So as a result we drop the first column.

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 variable has the highest correlation with 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 error terms are normally distributed.
- High R-squared and Adjusted R-squared values (0.805 and 0.795) calculated from testing set suggest the model captures variance well and aligns with the training set result.
- All variables have statistical significance (p-value < 0.05)
- VIF values are acceptable (almost all < 5), indicating no severe multicollinearity.

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)

- Temperature (temp) - The coefficient of temp is 0.4915.
- Year (yr) - The coefficient of yr is 0.2335.
- Light Rain or Snowfall (light_rain) - The coefficient of light_rain is -0.2852.

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.

The goal of linear regression is to find the best-fitting line that minimises the error between the predicted and actual values. The general idea is to predict the target variable by finding a linear combination of the features.

1. Simple Linear Regression (Single Feature)
In simple linear regression, we have one independent variable (feature) and one dependent variable (target). The relationship between the variables is modelled as:
$y = mx + c$

2. Multiple Linear Regression (Multiple Features)
In multiple linear regression, we have multiple independent variables (features) and one dependent variable (target). The relationship is modelled as:
$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$

Assumptions of Linear Regression:
1. Linearity: There is a linear relationship between the dependent and independent variables.
2. Independence: The residuals (errors) are independent of each other.
3. Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables.
4. Normality: The residuals should be normally distributed for inference purposes.

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 is a set of four datasets that have nearly identical simple descriptive statistics, yet very different distributions and patterns when graphed.

Visualizing the Data: When graphed, the datasets reveal their true nature:
Dataset I shows a perfect linear relationship.
Dataset II shows a quadratic curve.
Dataset III shows a linear relationship but with an influential outlier.
Dataset IV shows no relationship at all, with most of the points flat and only a few outliers.

Each dataset in the quartet has nearly identical summary statistics, such as mean, variance, correlation coefficient, and linear regression line (slope and intercept). This can mislead someone into thinking that the datasets are similar when in fact they are very different.

Anscombe's Quartet emphasizes the importance of visually inspecting data before drawing conclusions based on statistical measures. It warns against relying solely on summary statistics (like the correlation coefficient, mean, or variance) to understand the nature of the data, as they can be misleading. By plotting the data, we can better understand its underlying structure, identify potential outliers, and detect non-linear relationships that summary statistics alone may not reveal.

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, also known as the Pearson correlation coefficient, is a measure of the linear relationship or association between two continuous variables. It quantifies how strongly two variables are related and the direction of their relationship.

Interpretation:
r = 1: Perfect positive correlation (both variables increase together in perfect proportion).
r = -1: Perfect negative correlation (one variable increases while the other decreases in perfect proportion).
r = 0: No linear correlation (no linear relationship between the variables).

The closer the value of r is to 1 or -1, the stronger the linear relationship between the two variables.

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 refers to the process of adjusting the range or distribution of feature values in a dataset. It is often used in data preprocessing to ensure that numerical features have the same scale. Scaling helps improve the performance and convergence speed of machine learning algorithms, especially those that rely on distance metrics.

Normalised Scaling (MinMax Scaling):

- Values are rescaled to a fixed range, typically [0, 1].
- No assumption about the distribution of the data.
- Highly sensitive to outliers because outliers affect the min and max values.

Standardised Scaling:

- Values are rescaled to have a mean of 0 and standard deviation of 1.
- Assumes that the data is normally distributed.
- Less sensitive to outliers compared to normalisation, but outliers can still affect the mean and standard deviation.

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>

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other features. In other words, it quantifies the degree to which a feature is linearly related to other features in the model. If a feature is highly correlated with other features, the VIF will be large, indicating high multicollinearity.

A VIF becomes infinite or undefined in the following cases:

1. Perfect Multicollinearity: This occurs when one feature is perfectly correlated with another feature. In this case, the feature does not provide additional unique information and is redundant.
2. Singular Matrix in the Feature Matrix: VIF is calculated based on the inverse of the correlation matrix of the predictors. If the correlation matrix is singular (i.e., the matrix is not invertible because the predictors are perfectly linearly dependent), the VIF for the problematic feature becomes infinite.
3. Exact Linear Dependence: When one predictor variable is an exact linear combination of others in the dataset, it causes perfect multicollinearity. In such a case, the matrix used to compute the VIF becomes singular and thus results in infinite VIF for the dependent predictor.

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 plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles (or percentiles) of the sample data against the quantiles of the chosen reference distribution (often a normal distribution).

In the context of linear regression, a common use of the Q-Q plot is to check whether the residuals (errors) from the regression model follow a normal distribution. This is an important assumption in linear regression, as normality of residuals is often required for statistical inference (e.g., hypothesis tests, confidence intervals).

Importance and Use of a Q-Q Plot in Linear Regression

1. Assessing Residual Normality: A fundamental assumption in linear regression is that the errors or residuals (the difference between the observed and predicted values) should be normally distributed.

2. Detecting Skewness or Kurtosis: Deviations from the straight line in a Q-Q plot can indicate skewness (if the tails of the plot curve away from the line) or kurtosis (if the plot is too flat or too steep compared to the straight line).
3. Diagnosing Model Fit: A well-fitting model should result in residuals that are approximately normally distributed, among other assumptions like homoscedasticity (constant variance of errors). A Q-Q plot can help visually identify whether the residuals behave as expected.
4. Making Decisions: If the residuals deviate significantly from normality.