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)

Association Strength:

The strength of the relationship between a categorical independent variable and the dependent variable can be inferred using metrics like chi-square statistics or Cramér's V.
For example, if the dependent variable is binary (e.g., "Yes/No" for loan approval), categorical variables such as "Job Type" or "Marital Status" may significantly influence the outcome.
Distribution of Categories:

The frequency distribution of each category (e.g., percentage of "Single" vs. "Married") and how these correlate with the dependent variable helps understand the most impactful categories.
For instance, if the target variable is income level, certain job sectors or education levels may show disproportionately higher incomes.
Predictive Power:

In predictive models (e.g., decision trees or logistic regression), categorical variables with high importance scores or splitting criteria indicate a strong influence on 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)
Avoiding the Dummy Variable Trap
The dummy variable trap occurs when one category of a categorical variable can be perfectly predicted by the remaining categories.

By setting drop_first=True, one dummy variable (the baseline category) is dropped, ensuring that the model treats the dropped category as a reference and avoids this redundancy.

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)
Visual Assessment:

Look at the scatterplots in the pair plot to identify the strongest linear relationships between numerical variables and the target variable. A steeper, more linear trend (positive or negative) indicates a stronger correlation.

Compute Correlation Coefficients:

Use statistical methods (e.g., Pearson or Spearman correlation) to quantify the relationship numerically. The variable with the highest absolute correlation coefficient (close to ±1) is the strongest correlate with the target.

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)

Linearity
Assumption: The relationship between independent variables and the dependent variable is linear.
Validation:
Residual Plots: Plot residuals (errors) vs. predicted values. Residuals should be randomly distributed with no clear pattern.
Scatterplots: Ensure that the relationship between predictors and the target variable appears linear.

Independence of Errors
Assumption: Residuals are independent and not correlated.
Validation:
Durbin-Watson Test: A statistical test to check for autocorrelation in residuals (especially for time series data). A value close to 2 indicates no autocorrelation.

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)

Steps to Identify Key Features
Model Coefficients:

For a linear regression model, features with the largest absolute values of standardized coefficients are most influential.

Positive coefficients suggest a direct relationship with bike demand, while negative coefficients indicate an inverse relationship.

Feature Importance:

For tree-based models (e.g., random forests or gradient boosting), feature importance scores highlight the most impactful predictors.

Higher importance scores mean the feature contributes significantly to explaining variability in the target variable.

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)

Linear regression is a statistical modeling algorithm used to establish a relationship between a dependent variable (y) and one or more independent variables ($x1$,$x2$,...,xn). It is widely applied for predictive analysis and understanding variable relationships. Here's a detailed explanation:

1. Purpose

The primary goal is to predict the value of the dependent variable (y) based on the given independent variables and to quantify the strength of their relationships.

Mathematical Representation
Simple Linear Regression (one independent variable):
y=β0+β1x+ϵ
y: Dependent variable (response/output).

x: Independent variable (predictor/input).
β0: Intercept (value of y when x=0).

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)

Anscombe’s Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.), but vastly different distributions and relationships between the variables. The purpose of Anscombe’s Quartet, created by statistician Francis Anscombe in 1973, is to emphasize the importance of visualizing data before drawing conclusions from statistical analysis. Here's a detailed explanation:

1. The Four Datasets

Anscombe’s Quartet consists of four datasets, each containing 11 data points for two variables, x and y. The datasets are designed to show that statistical summaries (like means, variances, and correlations) can be identical for different datasets but may conceal important differences in the data. These datasets are often used to illustrate how data visualization can reveal underlying patterns that are not immediately obvious from 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)

Pearson's r (Pearson's correlation coefficient) is a statistical measure that assesses the strength and direction of the linear relationship between two continuous variables. It is represented by the symbol r and ranges from -1 to +1.

+1 indicates a perfect positive linear relationship (as one variable increases, the other also increases).
-1 indicates a perfect negative linear relationship (as one variable increases, the other decreases).
0 indicates no linear relationship between the 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>

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)

Scaling is a technique used in data preprocessing to adjust the range of numeric values in a dataset. It transforms features so that they have a similar scale, which can improve the performance and accuracy of machine learning models.

Why is Scaling Performed?
Scaling is performed for several reasons:

Improve Model Performance: Many machine learning algorithms, like gradient descent-based models (e.g., linear regression, logistic regression) and distance-based algorithms (e.g., k-nearest neighbors, support vector machines), are sensitive to the scale of input features. Features with larger ranges may dominate the model's behavior, leading to suboptimal results.

Convergence Speed: Algorithms that use optimization techniques (like neural networks) often converge faster if the data is scaled properly.

Consistency Across Features: Different features might have different units and ranges (e.g., age in years and salary in thousands). Scaling makes all features comparable.

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 (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. In a Q-Q plot, the quantiles of the observed data are plotted against the quantiles of the theoretical distribution.

If the data follows the theoretical distribution closely, the points on the plot will lie approximately along a straight line.

Deviations from the straight line suggest departures from the assumed distribution.

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

In the context of linear regression, a Q-Q plot is typically used to assess the normality of residuals. Residuals are the differences between the observed values and the predicted values from the regression model.