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)

Below are my analysis on categorical variables:

- Season, year, weekday and working day has positive impact on bike rents
- Holiday and bad weather has negative impact on bike rents

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 during dummy variable creation is important to avoid multicollinearity in regression models. When we create dummy variables for a categorical feature with k unique categories, we will end up with k dummy variables. If all these dummy variables are included in the regression, the predictors become perfectly correlated because knowing the values of k - 1 dummies allows to perfectly predict the value of the last dummy. This creates what's called the dummy variable trap, leading to multicollinearity .

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 pair-plot and correlation matrix, the numerical variable with the highest correlation with the cnt variable is registered, with a correlation value of 0.95

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 predictors and the dependent variable is linear.

- Residual Analysis on train data
- No Multicollinearity: VIFs on model

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, year, and season are strong positive predictors for bike rentals.

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 supervised machine learning algorithm used to model the relationship between one or more independent variables (X) and a continuous dependent variable (Y). The objective is to find the best-fitting linear equation that minimizes the difference between predicted and actual values Steps in the Algorithm.

1. Hypothesis Function
2. Cost Function (Mean Squared Error - MSE)
3. Model Evaluation

The performance of the linear regression model is evaluated using metrics such as:

- R-squared: Proportion of variance explained by the model.
- Adjusted R-squared: Adjusted for the number of predictors.
- Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) for error measurement.

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>

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 rrr, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It describes both the strength and the direction of the linear association 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)

< Scaling refers to transforming numerical features in a dataset so that they fall within a specific range or have particular properties (e.g., zero mean and unit variance)
Why is Scaling Performed?

1. Improved Model Performance
2. Faster Convergence
3. Feature Comparison
4. Interpretability

There are two common types of scaling:

1. Normalization**:** Rescales data to a fixed range (typically [0, 1]), Compresses values between min and max
2. Standardization: Centers data with mean 0 and variance 1, Maintains relative differences but standardizes >

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)

< The value of the Variance Inflation Factor (VIF) can become infinite in certain situations, and this typically occurs when there is perfect multicollinearity among the independent variables.

In simple terms, multicollinearity refers to a situation where one or more independent variables in the regression model are highly correlated with each other. When this happens, the VIF can increase drastically, and in cases of perfect multicollinearity, the VIF will be infinite.

>

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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, commonly the normal distribution. It compares the quantiles of the dataset against the quantiles of the theoretical distribution. If the data is approximately normally

distributed, the points in the Q-Q plot will lie along a straight line (usually a 45-degree line)

The Q-Q plot is an diagnostic tool in linear regression for verifying the normality of residuals. Checking for normality helps ensure that the statistical tests used in the regression analysis are valid and that the model's predictions are reliable. >
