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)

Categorical variable such as season, month, weather condition and days of the week significantly influence bike rentals. Rental are generally higher in favorable weather conditions (warmer months, sunny days), highlighting the combined effect on seasonality, weather and usage pattern.

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)

When drop_first=True is used, one dummy variable is removed leaving n-1 dummy variables for a categorical feature with n categories. This eliminates multicollinearity by ensuring the dummy variable are independent of each other while preserving the necessary information.

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 variable 'atemp' has the highest correlation with the target variable 'cnt'.

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)

There are following steps to validate the assumption of Linear Regression:

1-BY plotting the predicted value vs. actual value of the dependent variable we can assess the linearity assumption of the model.

2- By checking the normality of residuals, we can assess the validity of the linearity assumption. Plotting a histogram or Q-Q plot of the residuals allows us to verify if they are approximately normally distributed, which is a key requirement for the linear regression model to provide reliable results.

3. By checking for multicollinearity, we can ensure that the linear regression model remains robust. One way to assess this is by examining the Variance Inflation Factor (VIF).

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)

Based on the final model, the top 3 features typically the ones with the highest coefficients and statistical significance (low p-values) which are yr (year),temp (temperature), workingday.

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 models the relationship between a dependent variable and independent variables using a linear equation. It aims to minimize the error between actual and predicted values using Ordinary Least Squares (OLS).
Equation: Y=β0+β1X1+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \dots + \beta_nX_n + \epsilonY=β0 +β1X1+⋯+βnXn+ϵ
Assumptions: Linearity, Independence, Homoscedasticity, Normal Residuals, No Multicollinearity.
Model Fit: Measured by R-squared, MSE/RMSE.
Interpretation: Coefficients represent the effect of each predictor on the dependent variable.

<Your answer for Question 6 goes here>

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 consists of four datasets with identical summary statistics (mean, variance, correlation) but different visual patterns.
Key Points:
Purpose: To show that descriptive statistics can be misleading without visualization.
Datasets: Linear: Clear upward slope.
Quadratic: Non-linear relationship.
Scattered: No clear pattern, high variance.
Outlier: Distorted by an extreme value.
The quartet emphasizes the importance of plotting data before drawing conclusions

<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 R is a measure of the linear relationship between two variables, ranging from -1 to +1.

- **+1**: Perfect positive correlation
- **-1**: Perfect negative correlation
- **0**: No linear correlation

It is used to assess the strength and direction of a linear relationship between two continuous

variables.

Interpretation:

- **Positive R**: As one variable increases, the other increases.
- **Negative R**: As one variable increases, the other decreases.

<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:

Scaling adjusts the range of data features to improve model performance.

Why:

- Enhances model accuracy
- Ensures fair feature contribution
- Speeds up convergence for gradient-based algorithms

Types:

1. **Normalized Scaling**:
    - Rescales to a fixed range [0, 1].
    - Used when data isn't Gaussian.
2. **Standardized Scaling**:
    - Rescales to have mean 0 and std deviation 1.
    - Used for Gaussian data or handling outliers.

Difference:

- **Normalized**: Fixed range [0, 1].
- **Standardized**: Mean 0, std 1.

<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)

The value of VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity between the independent variables. This occurs when one of the predictor variables is an exact linear combination of others, leading to a correlation of 1 or -1. In such cases, the matrix used to calculate VIF cannot be inverted, resulting in an infinite value for VIF. This indicates a problem with multicollinearity in the model, meaning one variable is redundant and should be removed to improve the model's stability.

<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 (Quantile-Quantile)** plot compares the quantiles of a dataset with the normal distribution to check if the data follows a normal distribution.

Importance in Linear Regression:

1. **Normality of Residuals**: Ensures residuals are normally distributed.
2. **Linearity Check**: Helps detect deviations from normality or outliers.
3. **Model Assumptions**: Assesses if the normality assumption holds, crucial for valid model inferences.

<Your answer for Question 11 goes here>
