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)

The effect of categorical variables such as month (mnth), year (yr), season, and weather conditions significantly impacts the total bike rental count. From the analysis, several key insights were observed:

1. **Yearly Comparison**: There was a notable growth in bike rentals in 2019 compared to 2018, indicating an upward trend in demand.
2. **Seasonal Influence**: The summer and fall seasons saw the highest rental counts. This suggests that bike rentals are more popular during these periods.
3. **Weather Conditions**: Clear weather conditions significantly boost bike rentals. When the weather is clear, the number of rentals increases substantially.
4. **Holidays**: Holidays tend to have a slightly negative impact on bike rentals, possibly due to people opting for other activities.
5. **Monthly Trends**: The months from July to September experience high rental counts, emphasizing the importance of these months for the bike rental business.

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)

If all k dummy variables are included in the model, it leads to perfect multicollinearity. This means one of the dummy variables can be perfectly predicted from the others, causing redundancy and instability in the regression coefficients.

By setting (drop_first=True), you drop the first category from the dummy variables. This means you only include k-1 dummy variables, effectively avoiding multicollinearity. The dropped category serves as the reference category, and the coefficients of the remaining dummy variables represent the difference from this reference category.

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 atemp are features that are showing a strong positive correlation of 0.6 with target(cnt) with a linear trend.

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)

1. Normality of Residuals:

I calculated the residuals (differences between actual and predicted values). Checked distribution using histogram.

2. Homogeneity of Variance (Homoscedasticity):

I plotted the residuals against the predicted values in a scatter plot.

3. Independence of Residuals

I examined the Durbin-Watson statistic to check for autocorrelation in the residuals. A value close to 2 indicates that the residuals are independent of each other.

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)
The features Year, temp and weathersit are the top 3 features contributing significantly towards outcome.

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>

Explain the linear regression algorithm in detail.

Linear regression is a used to analyze the relationship between a dependent variable and one or more independent variables. It helps in predicting the value of the dependent variable based on the values of the independent variables.

Linear regression assumes that the relationship between the dependent and independent variables can be represented by a straight line.

Types of Linear Regression**:**

1. Simple Linear Regression: Involves one independent variable.

$$Y=\beta0+\beta1X+\epsilon$$

2. Multiple Linear Regression: Involves two or more independent variables.

$$Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$$

Where:

- Y is the dependent variable.
- X1,X2,…,Xn are the independent variables.
- β0 is the intercept.
- β1,β2,…,βn are the coefficients (slopes).
- ϵ\epsilon is the error term (residuals).

Assumptions of Linear Regression:

1. Linearity: The relationship between the dependent and independent variables is linear.
2. Independence: The residuals (errors) are independent.
3. Homoscedasticity: The residuals have constant variance.
4. Normality: The residuals are normally distributed.

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 serves as a powerful reminder of the importance of data visualization in statistical analysis.
Simply relying on summary statistics can miss critical patterns and anomalies in the data.
Visualization reveals the true nature of data.
Plotting data can uncover outliers, patterns, and other features that summary statistics may obscure.

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 correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the relationship.
The coefficient ranges from -1 to 1.

Pearson's R is calculated using the covariance of the variables divided by the product of their standard deviations:

1 : Indicates a perfect positive linear relationship.
0: Indicates no linear relationship.
-1: Indicates a perfect negative linear relationship.

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 is the process of transforming the features (variables) of a dataset to a specific range, often to ensure that the features have similar scales.

Scaling is performed because:
1. **Improves Convergence in Gradient Descent**: Algorithms like linear regression, use gradient descent for optimization. Scaling helps in faster and more stable convergence.
2. **Ensures Fair Contribution**: Features with larger scales can dominate the learning process. Scaling ensures that all features contribute equally.

Normalization:
Normalization (or Min-Max Scaling) transforms the data to a fixed range, typically [0, 1] or [-1, 1].

$$xnorm = X - min(X) / (max(X) - min(X))$$

Useful when the distribution of the data is not Gaussian (normal) and you need a bounded range.

Standardized:
Standardization (or Z-score normalization) transforms the data to have a mean of 0 and a standard deviation of 1.

$$Std = X - \mu / \sigma$$

(μ-mean & σ – standard deviation)
Useful when the distribution of the data is Gaussian.

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) is a measure used to detect multicollinearity in regression models. Multicollinearity occurs when two or more independent variables are highly correlated, leading to redundancy and instability in the regression coefficients.

$$VIF(Xi)=1/1-Ri2$$

Where Ri2R^2_i is the coefficient of determination of the regression of XiX_i on all other independent variables.

If Ri2=1R^2_i = 1 (indicating perfect multicollinearity), the denominator becomes zero, and the VIF becomes infinite. So, it leads to perfect multicollinearity.

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, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution .

Horizontal Axis (Theoretical Quantiles): Represents the expected quantiles if the data follows the theoretical distribution.

Vertical Axis (Sample Quantiles): Represents the quantiles of the observed data.

It is widely used for Checking Normality of Residuals: