Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

From the analysis, it was observed that certain categorical variables significantly impact the dependent variable, `cnt`(count of bike rentals). For example:

- **Season**: Different seasons showed varying levels of bike rentals, with certain seasons like summer showing higher counts compared to others.
- **Month**: There is noticeable variation in bike rentals across different months, indicating seasonality in demand.
- **Weather Situation**: Weather conditions significantly impact bike rentals, with better weather conditions correlating with higher rentals.
- **Holiday**: Bike rentals were generally lower on holidays compared to regular working days.

These observations suggest that environmental and temporal factors play a significant role in influencing bike rental demand.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Using `drop_first=True` is important during dummy variable creation to avoid multicollinearity. When creating dummy variables, one category can be perfectly predicted by the others, leading to redundant information. By dropping the first category, we avoid the dummy variable trap and ensure the model does not suffer from perfect multicollinearity, which can lead to unreliable estimates.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

From the pair-plot analysis, it was observed that the variable `atemp` (apparent temperature) has the highest correlation with the target variable `cnt` (count of bike rentals).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

After building the linear regression model, the following steps were taken to validate the assumptions:

- **Linearity**: Checked by plotting the predicted values versus actual values to see if there is a linear relationship.
- **Homoscedasticity**: Verified using a residual plot to ensure residuals have constant variance.
- **Normality of Residuals**: Assessed using a Q-Q plot and histogram of the residuals to check if they follow a normal distribution.
- **Multicollinearity**: Evaluated using Variance Inflation Factor (VIF) to ensure there is no high correlation between independent variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

- **atemp (apparent temperature)**
- **yr (year, indicating a general trend over years)**
- **season_3 (indicating the season variable for the 3rd season)**

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. The goal is to find the linear equation $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$, where:

- yy is the dependent variable.
- $\beta0\beta0$ is the y-intercept.
- $\beta1,\beta2,\ldots,\beta n\beta1,\beta2,\ldots,\beta n$ are the coefficients for each independent variable $x1,x2,\ldots,xnx1,x2,\ldots,xn$.
- $\epsilon\epsilon$ is the error term.

The algorithm involves:

1. **Fitting the Model**: Using the least squares method to minimize the sum of the squares of the residuals (differences between observed and predicted values).
2. **Evaluating the Model**: Assessing the model's performance using metrics like R-squared, adjusted R-squared, and p-values of coefficients.
3. **Assumptions Check**: Validating assumptions of linearity, independence, homoscedasticity, normality of residuals, and no multicollinearity.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four datasets with nearly identical simple descriptive statistics, yet they exhibit different distributions and relationships. It highlights the importance of visualizing data:

1. **First dataset**: A typical linear relationship.
2. **Second dataset**: A parabolic relationship.
3. **Third dataset**: A linear relationship with an outlier.
4. **Fourth dataset**: A linear relationship but with a single influential point.

The quartet demonstrates that relying solely on summary statistics can be misleading, and data visualization is crucial for proper data analysis.

3. What is Pearson's R? (3 marks)

Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to 1:

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

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling transforms data to a standard range or distribution. It is performed to:

- Improve the performance of algorithms sensitive to the scale of data.
- Ensure features contribute equally to model training.

Normalized Scaling: Transforms data to a [0, 1] range. Suitable for bounded data distributions.

Standardized Scaling: Transforms data to have a mean of 0 and standard deviation of 1. Suitable for unbounded data distributions and preserving the distribution shape.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity, meaning one predictor variable is a perfect linear combination of other predictor variables. This situation leads to an undefined or infinite value for VIF, indicating redundancy in predictors and issues with model estimation.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a particular distribution, usually the normal distribution. It plots quantiles of the data against the quantiles of the theoretical distribution. In linear regression, Q-Q plots are used to:

- Check if residuals follow a normal distribution.
- Validate the assumption of normality, which is important for hypothesis testing and constructing confidence intervals.

Deviations from the diagonal line in a Q-Q plot indicate departures from normality.