Assignment-based Subjective Questions

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

Answer: Holiday: Since it's not listed among the coefficients, it wasn't included in the final model, indicating it might not have had a significant effect or was excluded during feature selection.

Workingday: Similarly, this variable isn't in the final model, suggesting a similar rationale.

Months:

sep has a positive coefficient (756.65), indicating higher bike rentals in September.

Seasons:

spring has a negative coefficient (-500.85), indicating lower bike rentals in spring compared to the reference season.

summer and winter have positive coefficients (541.48 and 809.87, respectively), suggesting higher bike rentals in these seasons compared to the reference season.

Weather Conditions:

Light_snowrain and Misty have negative coefficients (-2447.77 and -674.24, respectively), indicating significantly lower bike rentals during these weather conditions compared to clear weather.

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

Answer: Using drop_first=True is important to avoid the dummy variable trap, which occurs when there's perfect multicollinearity among the dummy variables. Dropping the first category ensures that each category's effect is measured relative to the omitted category, leading to more stable and interpretable models.

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

Answer: Among the numerical variables, `temp` has the highest positive coefficient (4169.65). This suggests that temperature is the variable most strongly correlated with bike rentals, indicating that higher temperatures lead to increased bike usage.

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

Answer: Linearity: Scatter plots of observed vs. predicted values showed a linear relationship.

Normality of Residuals: Histogram and Q-Q plots of the residuals confirmed they follow a normal distribution.

No Multicollinearity: Variance Inflation Factor (VIF) values were checked to ensure they were within acceptable limits.

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

Answer: Temp: With a coefficient of 4169.65, temperature is the most significant positive contributor to bike rentals, indicating that warmer temperatures lead to higher demand.

Year: The coefficient of 2043.03 indicates a strong positive trend in bike rentals over time, reflecting growing popularity or availability of shared bikes.

Windspeed: With a coefficient of -1301.11, wind speed is a significant negative contributor, meaning higher wind speeds reduce bike rentals.

General Subjective Questions

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

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fit line (or hyperplane in higher dimensions) that minimizes the difference between the predicted and actual values of the target variable.

Steps:

1. **Assumptions**:
   - **Linearity**: The relationship between the dependent and independent variables is linear.
   - **Independence**: Observations are independent of each other.
   - **Homoscedasticity**: Constant variance of the errors.
   - **Normality**: The residuals (errors) of the model are normally distributed.
   - **No Multicollinearity**: Independent variables are not highly correlated.
2. **Model Representation**:
   - Simple Linear Regression: $y=\beta0+\beta1x+\epsilon$
   - Multiple Linear Regression: $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$
3. **Fitting the Model**:
   - The goal is to estimate the coefficients ($\beta0,\beta1,...,\beta n$) that minimize the sum of squared residuals (SSR).
   - This is done using the Ordinary Least Squares (OLS) method: Minimize
   - $\sum(yi-yi\hat{})2$,
4. **Evaluation**:
   - **R-squared**: Proportion of variance in the dependent variable explained by the independent variables.
   - **Adjusted R-squared**: Adjusted for the number of predictors in the model.
   - **F-statistic**: Tests overall significance of the model.
   - **P-values**: Tests significance of individual predictors.
5. **Assumptions Validation**:
   - **Residual Plots**: To check for homoscedasticity and independence.
   - **Q-Q Plots**: To check for normality of residuals.
   - **VIF (Variance Inflation Factor)**: To check for multicollinearity.

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

Answer: Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) and linear regression lines, but appear very different when graphed.

- **Purpose**: Demonstrates the importance of graphing data before analyzing it, as datasets with similar statistical properties can have very different distributions and relationships.
- **Datasets**:
    1. **Dataset 1**: A typical linear relationship.
    2. **Dataset 2**: Non-linear relationship.
    3. **Dataset 3**: Linear relationship but with an outlier.
    4. **Dataset 4**: Vertical line with an outlier influencing the statistics.

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

Answer: Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two variables, providing a value between -1 and 1.

- **Value**:
    - **1**: Perfect positive linear correlation.
    - **0**: No linear correlation.
    - **-1**: Perfect negative linear correlation.
- **Interpretation**:
    - Positive values indicate a direct relationship.
    - Negative values indicate an inverse relationship.
    - Magnitude indicates the strength of the relationship.

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

Answer: Scaling is the process of transforming data to fit within a specific range or distribution, often to improve the performance of machine learning algorithms.

- **Scaling is performed because**:
    - Ensures all features contribute equally to the model.
    - Speeds up convergence of optimization algorithms.
    - Improves accuracy of distance-based algorithms.
- **Types**:
    - **Normalized Scaling (Min-Max Scaling)**: Rescales features to a [0, 1] range.
    - **Standardized Scaling (Z-score Scaling)**: Rescales features to have zero mean and unit variance.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: VIF (Variance Inflation Factor) measures the degree of multicollinearity among features in a regression model.

- **It could be Infinite when**:
    - Occurs when a predictor is a perfect linear combination of other predictors.
    - Indicates perfect multicollinearity, leading to an inability to estimate regression coefficients.

- **Implication**: A VIF value of infinity means the model is unstable and the predictors are not independent.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a specified distribution, usually normal.

- **Use:** Plots quantiles of the data against quantiles of a standard normal distribution.
- **Importance in Linear Regression**:
    - **Check Normality of Residuals**: Ensures residuals are normally distributed, a key assumption of linear regression.
    - **Interpretation**:
        - Points on or near the line indicate normal distribution.
        - Deviations from the line indicate departures from normality.
- **Significance**: Helps validate model assumptions and ensure reliability of statistical tests