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)

->Significant Effects: Weather conditions and months of the year play a crucial role in determining bike demand

-> Limited Impact: Variables like weekends may have a smaller or insignificant effect on the target variable

---

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 very important to avoid the dummy variable trap, which occurs when a dummy variable can be perfectly predicted by others which leads to multicollinearity in regression models.

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)

"Registered"

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)The scatter plot of actual vs. predicted values (in the second image) shows a positive correlation. This indicates that the model is capturing the linear relationship between the predictors and the target variable

2) The residuals vs. predicted values plot (left plot in the first image) was used to check for constant variance. The residuals are scattered randomly around the horizontal line (0), and no clear patterns are observed. This suggests that the variance of residuals is constant across all levels of predicted values

3) The histogram of residuals (right plot in the first image) was analyzed to ensure that residuals follow a normal distribution. The histogram shows a roughly bell-shaped curve.

4) Variance Inflation Factor (VIF) values were calculated (visible in the third image). All VIF values are below 5.

5) The regression summary (third image) shows a high F-statistic and low p-value for the model, indicating overall model significance

6) The assumptions of Linear Regression were validated using graphical methods (e.g., residual plots, histogram) and statistical measures (e.g., VIF, F-statistic)

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)
1) **Year (yr)**

- **Coefficient:** 0.2560
- **P-value:** 0.000
- **Interpretation:** The year feature has a significant positive impact on bike demand, as time progresses (e.g., moving to a new year), bike demand increases.

2) **Peak Month (peak_month)**

- **Coefficient:** 0.1838
- **P-value:** 0.000
- **Interpretation:** The "peak month" variable significantly affects bike demand indicating higher demand during specific months.

3)**Wind Speed (windspeed)**

- **Coefficient:** -0.1510
- **P-value:** 0.000
- **Interpretation:** Wind speed has a significant negative impact on bike demand, higher wind speeds reduce the number of bike users.

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)

1. **Understand and Prepare the Data**:

   ->Identify the target variable and independent variables.

   ->Clean and preprocess the data (handle missing values, remove outliers, standardize or normalize if needed).

2. **Fit the Model**:

->The algorithm finds the best-fitting line by minimizing the error between actual and predicted values. This is done using the **Ordinary Least Squares (OLS)** method.

3. **Ordinary Least Squares (OLS)**:

   ->OLS minimizes the **Sum of Squared Errors (SSE)**: SSE=∑i=1n(yi−y^i)2SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2SSE=i=1∑n(yi−y^i)2

   ->The model calculates the coefficients (β0,β1,…,βn\beta_0, \beta_1, \ldots, \beta_nβ0,β1,…,βn) that minimize the SSE

4. **Prediction**:

   ->Once the coefficients are calculated, predictions can be made using the equation: y^=β0+β1x1+β2x2+…+βnxn\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_ny^=β0+β1x1+β2x2+…+βnxn

5. **Evaluate the Model**:

   Evaluate the performance using metrics like:

   - **R-squared**: Proportion of variance in the dependent variable explained by the independent variables. R2=1−SSresidualSStotalR^2 = 1 - \frac{SS_{residual}}{SS_{total}}R2=1−SStotalSSresidual
   - **Mean Squared Error (MSE)**: Average squared difference between actual and predicted values.
   - **Root Mean Squared Error (RMSE)**: Square root of MSE.

6. **Residual Analysis**:

   ->Analyze residuals to check for violations of assumptions like normality, homoscedasticity, and independence.

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)

Anscombes quartet is a set of four datasets that have nearly identical simple statistical properties (mean, variance, correlation, and linear regression line) but appear very different when visualized graphically.

Characteristics of Anscombe's Quartet:

1. Each dataset consists of 11 (x, y) data points
2. The four datasets share the following statistical properties:
   - Mean of x: 9

- Variance of x: 11
- Mean of y: 7.5
- Variance of y: Approximately 4.12
- Correlation between x and y: ~0.816
- Regression line equation: y=3+0.5xy = 3 + 0.5xy=3+0.5x

Despite these similarities the datasets differ drastically when plotted, emphasizing the importance of graphical analysis in data exploration. Anscombes quartet is a powerful example of why data visualization should always accompany statistical analysis. It teaches us to go beyond numbers and look at the data's shape and structure to make informed decisions

---

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)

It is a is a measure of the linear relationship between two variables. It quantifies how strongly the two variables are linearly related and whether their relationship is positive or negative. **Applications used :**

1. **Statistics**:

   To quantify the strength and direction of a linear relationship between two variables.

2. **Machine Learning**:

   To identify feature relationships and remove redundant features in regression models.

3. **Finance**:

   To measure correlations between asset returns or economic indicators.

4. **Healthcare**:

   To explore relationships between variables like age and disease severity.

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 is a preprocessing technique used to adjust the range of features in a dataset so that they are comparable and fit within a specific scale. It ensures that numerical variables have a similar scale, which can significantly improve the performance of machine learning algorithms.
->Normalization is ideal when you want to confine the feature values to a specific range.
->Standardization is better for algorithms sensitive to the distribution of data (e.g., requiring zero mean and unit variance). The choice depends on the dataset, algorithm, and problem requirements.

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 becomes infinite when there is perfect multicollinearity among the independent variables.
-> Infinite VIF indicates severe multicollinearity, which can make regression coefficients unstable, unreliable, and difficult to interpret
->it can also lead to numerical instability in calculations, resulting in model performance issues

---

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 plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (commonly the normal distribution). It plots the quantiles of the dataset against the quantiles of the theoretical distribution to visually assess whether the data follows the expected distribution.
=>Q-Q plot does not quantify the degree of normality but provides a visual assessment.

=> If residuals significantly deviate from normality, alternative regression models (e.g., robust regression) or transformations might be needed.

=>While normality of residuals is important for hypothesis testing, linear reg

---