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)

Fall season attracts more bookings, showing a seasonal impact. Higher bookings in May to October, decreasing towards year-end. Clear weather and weekends (Thu to Sun) lead to more bookings. Bookings are fewer on holidays, and 2019 shows a significant increase.

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

Reduces extra columns in dummy variable creation. Mitigates correlations among dummy variables. Improves model efficiency by avoiding multicollinearity.

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

'temp' variable exhibits the highest correlation with the target variable.

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

- Normality of error terms checked for normally distributed residuals.
- Multicollinearity verified for insignificant correlation among variables.
- Linear relationship validated through visual inspection.
- Homoscedasticity ensured by observing no visible pattern in residual values.
- Independence of residuals confirmed, indicating no auto-correlation

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

'temp' is the most significant feature. 'winter' and 'sep' also contribute significantly to explaining bike demand.

General Subjective Questions---

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

Linear regression finds the best-fit line, minimizing the gap between predicted and actual values. It's like drawing a line through points, showing how one thing influences another. It assumes a linear relationship and employs coefficients to represent the impact of independent variables on the dependent variable.

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

Anscombe’s Quartet has four datasets with similar stats but looks different. It reminds us to trust our eyes and not just numbers when studying data.

3. What is Pearson’s R? (3 marks)

   Pearson’s R measures how things move together. It goes from -1 to 1, telling if things go up and down together . 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling? (3 marks)

Scaling is the process of transforming variables to a standard range. It is performed to bring all variables to a similar scale, preventing one variable from dominating others in models sensitive to magnitude differences. Normalized scaling brings values between 0 and 1, maintaining the distribution shape, while standardized scaling (z-score normalization) transforms variables to have a mean of 0 and standard deviation of 1, preserving relationships but making interpretation easier.

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

(3 marks)

Infinite VIF happens when predictors are like twins, and you can predict one from the other. It’s a sign to tidy up and choose one.

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

A Q-Q plot checks if residuals (the leftovers in math) act normally. If dots make a neat line, things are good. If not, it helps us see what needs fixing in our predictions.

(3 marks)