**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?**

**Answer:**  
Categorical variables can have a significant impact on the dependent variable. For example:

* Variables like season or weathersit might influence the demand for bikes due to changes in weather patterns and user behavior during different seasons.
* Variables such as holiday or workingday are expected to have a direct impact as they influence people's commuting needs.

**Question 2: Why is it important to use drop\_first=True during dummy variable creation?**

**Answer:**  
Using drop\_first=True prevents the problem of multicollinearity by dropping one of the categories for each categorical variable. This ensures that the dummy variables are independent and avoids the "dummy variable trap," where one dummy variable is linearly dependent on others.

**Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

**Answer:**  
From the pair-plot, the variable registered has the highest positive correlation with the target variable, indicating that the number of registered users is a strong predictor of bike demand.

**Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?**

**Answer:**  
The assumptions of linear regression were validated as follows:

1. **Linearity:** Checked residual plots to ensure the linearity between predictors and the target variable.
2. **Normality:** Used a Q-Q plot to confirm that residuals follow a normal distribution.
3. **Homoscedasticity:** Verified constant variance of residuals through scatterplots of residuals vs. fitted values.
4. **Multicollinearity:** Examined VIF values to ensure no multicollinearity among predictors.
5. **Independence:** Ensured errors were independent using the Durbin-Watson test.

**Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**

**Answer:**  
The top 3 features are:

1. registered
2. casual
3. season

**General Subjective Questions**

**Question 6: Explain the linear regression algorithm in detail.**

**Answer:**  
Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more independent variables. The relationship is expressed as:  
y=β0+β1x1+β2x2+…+βnxn+ϵy = \beta\_0 + \beta\_1x\_1 + \beta\_2x\_2 + \ldots + \beta\_nx\_n + \epsilony=β0​+β1​x1​+β2​x2​+…+βn​xn​+ϵ  
Where:

* yyy is the dependent variable, x1,x2,…,xnx\_1, x\_2, \ldots, x\_nx1​,x2​,…,xn​ are the predictors, β\betaβ are coefficients, and ϵ\epsilonϵ is the error term.

The goal is to minimize the sum of squared residuals (differences between observed and predicted values) to find the best-fit line. This is achieved using techniques like Ordinary Least Squares (OLS).

**Question 7: Explain Anscombe’s quartet in detail.**

**Answer:**  
Anscombe’s quartet is a set of four datasets with nearly identical statistical properties (mean, variance, correlation, and regression lines) but significantly different distributions when plotted.  
Key takeaways include:

* Always visualize data, as relying solely on statistical summaries can be misleading.
* Different patterns (linear, nonlinear, or outliers) may affect the interpretation of statistical models.

**Question 8: What is Pearson’s R?**

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

* R=1R = 1R=1: Perfect positive linear correlation.
* R=−1R = -1R=−1: Perfect negative linear correlation.
* R=0R = 0R=0: No linear relationship.

**Question 9: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

**Answer:**  
Scaling adjusts numerical features to ensure they contribute equally to the model:

* **Why:** To prevent features with larger magnitudes from dominating the model.
* **Normalized scaling:** Rescales values to a fixed range (e.g., [0, 1]).
* **Standardized scaling:** Centers values around 0 with a standard deviation of 1.

**Question 10: You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

**Answer:**  
VIF becomes infinite when there is perfect multicollinearity, meaning one predictor is an exact linear combination of others. This causes the denominator of the VIF formula to be zero.

**Question 11: What is a Q-Q plot? Explain its use and importance in linear regression.**

**Answer:**  
A Q-Q (Quantile-Quantile) plot compares the quantiles of residuals with a theoretical normal distribution.

* **Use:** To check if residuals follow a normal distribution, which is an assumption of linear regression.
* **Importance:** Non-normality in residuals can lead to unreliable predictions and invalid significance tests.