Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Categorical variables such as season, weathersit, and holiday play a significant role in determining bike demand.

- **Season**: Different seasons exhibit varying levels of bike usage, with higher demand in summer and fall due to favorable weather conditions and lower demand in winter.
- **Weathersit**: Weather conditions significantly affect bike demand. Clear or partly cloudy days (category 1) see higher demand, while rainy or snowy days (categories 3 and 4) see reduced demand.
- **Holiday**: Bike demand tends to be lower on holidays, as commuting is reduced.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Using drop_first=True ensures that one of the categories in a categorical variable is treated as a baseline. This helps avoid the **dummy variable trap**, a scenario where multicollinearity arises because all categories are represented as separate dummy variables. By dropping the first category, we reduce redundancy and allow the model to interpret the coefficients relative to the baseline category.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The variable registered has the highest positive correlation with the target variable cnt. This is because registered users account for the

majority of bike rentals, indicating their significant influence on overall demand.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

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

1. **Linearity**: Scatter plots of residuals vs. predicted values were analyzed to confirm no discernible pattern, validating linearity.
2. **Normality**: A Q-Q plot was used to check if residuals followed a normal distribution.
3. **Homoscedasticity**: Residuals were plotted to ensure constant variance across predictions.
4. **Multicollinearity**: Variance Inflation Factor (VIF) was calculated for predictors, ensuring no excessive multicollinearity.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?

The top three features identified from the final model are:

1. **Temperature (temp)**: Higher temperatures are positively associated with increased demand.
2. **Year (yr)**: The demand increased significantly in 2019 compared to 2018, indicating growing popularity.
3. **Weather Situation (weathersit)**: Clear weather leads to higher demand, while bad weather reduces it.

General Subjective Questions

Q6. Explain the linear regression algorithm in detail.

Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (y) and one or more independent variables (X). It assumes a linear relationship:

y=β0+β1X1+β2X2+⋯+βnXn+ϵ y= \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n + \epsilony=β0+β1X1+β2X2+⋯+βnXn+ϵ

Where:

- β0\beta_0β0 is the intercept.
- βi\beta_iβi are the coefficients of predictors.
- ϵ\epsilonϵ is the error term.

The algorithm minimizes the sum of squared residuals (differences between observed and predicted values) using techniques like **Ordinary Least Squares (OLS)**. Evaluation metrics include R-squared and Mean Squared Error (MSE).

Q7. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet consists of four datasets with nearly identical statistical properties (mean, variance, correlation, etc.), yet they differ significantly when visualized. It highlights the importance of visualizing data rather than relying solely on summary statistics, as similar

numerical summaries can represent vastly different distributions or relationships.

Q8. What is Pearson's R?

Pearson's R is a measure of linear correlation between two variables, ranging from -1 to 1.

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

It helps quantify the strength and direction of a linear relationship.

Q9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling adjusts the range of features to make them comparable. It is performed to:

- Improve convergence of algorithms like Gradient Descent.
- Avoid dominance of features with larger magnitudes.

Normalized Scaling: Rescales data to [0, 1] using:

x'=x−min⁡(x)max⁡(x)−min⁡(x)x' = \frac{x - \min(x)}{\max(x) - \min(x)}x'=max(x)−min(x)x−min(x)

Standardized Scaling: Centers data around 0 with unit variance using:

x'=x−μσx' = \frac{x - \mu}{\sigma}x'=σx−μ

Normalization is used for bounded datasets, while standardization is preferred for Gaussian distributions.

Q10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity between variables (e.g., one variable is a linear combination of others). This means the determinant of the correlation matrix is zero, making VIF calculation undefined.

Q11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot compares the quantiles of residuals to a theoretical normal distribution. It helps assess whether residuals are normally distributed, a key assumption of linear regression. Points on the Q-Q plot should form a straight line if residuals are normal. Deviations indicate skewness or kurtosis in the data.