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?

Categorical variables such as season (spring, summer, winter), holiday, and weathersit (light snow/rain, mist) have statistically significant effects on the dependent variable (cnt), as indicated by their low p-values (all < 0.05).
For example:
- season_spring has a negative coefficient (-0.0746), indicating a decrease in demand compared to the reference season (autumn).
- season_summer (0.0421) and season_winter (0.0883) have positive coefficients, suggesting higher demand in these seasons relative to autumn.
- holiday_1 has a negative effect (-0.0865), meaning demand drops on holidays.
- weathersit_light_snow_rain (-0.2409) and weathersit_mist (-0.0537) both decrease demand compared to clear weather, with light snow/rain having a much stronger negative impact.

These effects are significant and indicate that seasons, holidays, and weather conditions meaningfully influence bike demand.

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

Using drop_first=True when creating dummy variables avoids the “dummy variable trap”, which is a form of perfect multicollinearity that occurs when all categories are included as separate variables.
By dropping one category (the reference), the model can uniquely estimate coefficients for the remaining categories, ensuring the regression matrix is invertible and the model is stable.

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

Based on the pair-plot as well as regression coefficients, temp (temperature) has the highest positive correlation with the target variable (cnt), as indicated by its large coefficient (0.4961) and high VIF, reflecting strong association with demand.

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

The assumptions were validated as follows:
- Linearity: Checked using residuals vs. fitted values plot to ensure no patterns, indicating a linear relationship.
- Normality of residuals: Assessed using a Q-Q plot and normality tests (e.g., Jarque-Bera), with some deviation indicated by skewness and kurtosis values.
- Homoscedasticity: Examined by plotting residuals against fitted values to check for

constant variance.
- Independence: Verified using the Durbin-Watson statistic (2.018), which is close to 2, indicating little autocorrelation.
- Multicollinearity: Checked using VIF values; some variables (hum, temp) have high VIF, suggesting multicollinearity, but most are 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?

The top 3 features, based on the magnitude and significance of their coefficients, are:
- temp (coefficient: 0.4961, highly significant)
- yr (coefficient: 0.2307, highly significant)
- weathersit_light_snow_rain (coefficient: -0.2409, highly significant negative effect)

These features have the largest absolute coefficients and are statistically significant, indicating the strongest impact on bike demand.

General subjective questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to observed data.
In simple linear regression (one independent variable), the relationship is modeled as (y = mx + b), where (m) is the slope and (b) is the intercept.
In multiple linear regression, the model generalizes to (y = beta_0 + beta_1.x 1 + beta_2.x 2 + ... + beta_k.x_k), where each (beta) represents the effect of a feature.

The algorithm works by:
- Estimating the best-fit line that minimizes the sum of squared differences (residuals) between observed and predicted values (the “least squares method”)
- Using “gradient descent” or analytical solutions to optimize the coefficients, reducing prediction error.
- Making several assumptions: linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of residuals.

Linear regression is widely used for prediction, trend analysis, and understanding relationships between variables.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is a collection of four datasets that have nearly identical simple statistical properties (mean, variance, correlation, regression line), yet appear very different when graphed.
Each dataset consists of eleven (x, y) points[rich_content:1].

The quartet demonstrates that:
- Summary statistics (like mean, variance, correlation, regression coefficients) can be misleading if used alone.
- Data visualization is essential to detect patterns, outliers, or non-linear relationships that statistics alone cannot reveal.

Anscombe's quartet is a classic example used to emphasize the importance of plotting data before interpreting statistical analyses.

3. What is Pearson's R?

Pearson's R (also called the Pearson correlation coefficient) is a statistical measure that quantifies the linear relationship between two continuous variables. Its value ranges from -1 to +1:
- +1: perfect positive linear correlation
- 0: no linear correlation
- -1: perfect negative linear correlation

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It is widely used to assess the strength and direction of a linear association.

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

Scaling refers to transforming features so they have a similar range or distribution. It is performed to:
- Ensure that features contribute equally to model training, especially for algorithms sensitive to feature magnitude.
- Improve convergence speed and model performance.

Normalized scaling (min-max scaling) rescales data to a fixed range, usually [0, 1], using:
[
x_{norm} = frac{x - x_{min}}{x_{max} - x_{min}}
]

Standardized scaling (z-score scaling) transforms data to have a mean of 0 and a standard deviation of 1:
[
x_{std} = frac{x - mu}{sigma}
]

Normalization is useful when you want features in a bounded range; standardization is preferred when data follows a Gaussian distribution or when outliers are present.

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

VIF (Variance Inflation Factor) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors.
VIF becomes infinite when there is perfect multicollinearity—that is, when one predictor is an exact linear combination of others.
In this case, the denominator in the VIF formula (1 - (R^2)) becomes zero, causing the VIF to diverge to infinity.
This indicates that the regression model cannot uniquely estimate the coefficients due to redundant information among predictors.

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

A Q-Q plot (quantile-quantile plot) is a graphical tool to compare the distribution of a dataset with a theoretical distribution (commonly the normal distribution). It plots the quantiles of the sample data against the quantiles of the reference distribution.

In linear regression, Q-Q plots are used to:
- Assess whether the residuals (errors) are normally distributed, which is an important assumption for inference (confidence intervals, hypothesis tests).
- Detect deviations from normality, such as skewness or outliers.

If the points in a Q-Q plot lie approximately on a straight line, the residuals are likely normally distributed, supporting the validity of regression inferences.