**1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Categorical variables like season, weathersit, holiday, and workingday have a significant impact on the dependent variable cnt (total bike rentals).

- **Season**: The demand for bikes varies every season. For example, demand is higher in summer and fall compared to winter and spring, which is due to better weather conditions.
- **Weathersit**: Poor weather conditions like heavy rain, snow etc significantly reduce bike rentals, and clear weather increases the demand.
- **Holiday**: On holidays, the demand for bikes decrease as people may prefer to stay indoors or travel out of the city.
- **Working Day**: On working days, the demand for bikes is higher, especially registered users, as people use bikes for commuting.
- **Windspeed**: Higher the windspeed, lower will be the bike rental count, in order to ensure a safe drive. From a coding perspective, we'd consider this for visualization and prediction. Nevertheless, the correlation values for all the features are derived and they can be taken for testing and prediction purposes.

These categorical variables help in understanding how external factors influence bike rental demand, and they should be included in the model to improve prediction accuracy.

---

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

When creating dummy variables for categorical features, using drop_first=True is important to avoid the **dummy variable trap**. The dummy variable trap occurs when the dummy variables are multicollinear, meaning one variable can be predicted from the others. This multicollinearity can cause issues in linear regression models, such as unstable coefficient estimates and inflated standard errors.

By dropping the first category, we reduce redundancy and ensure that the model can interpret the categorical variables correctly without multicollinearity.

---

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

From the pair-plot, the numerical variable temp (temperature) is likely to have the highest correlation with the target variable cnt. This is because people are more likely to rent bikes when the weather is pleasant and warm.

---

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

After building the linear regression model, the following assumptions were validated:

1. **Linearity**: Checked using scatter plots of residuals vs. predicted values. The residuals should be randomly scattered around zero, indicating a linear relationship.
2. **Homoscedasticity**: Verified by plotting residuals vs. predicted values. The spread of residuals should be constant across all levels of the predicted values.
3. **Normality of Residuals**: Checked using a Q-Q plot. The residuals should follow a straight line, indicating they are normally distributed.
4. **Independence of Residuals**: Verified using the Durbin-Watson test. A value close to 2 indicates no autocorrelation.
5. **Multicollinearity**: Checked using the Variance Inflation Factor (VIF). A VIF value less than 5 indicates no significant multicollinearity.

---

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

The top 3 features contributing significantly to explaining the demand for shared bikes are:

1. **Temperature (temp)**: Higher temperatures increase bike rentals.
2. **Year (yr)**: The demand for bikes increases over the years, indicating a growing trend.
3. **Season (season)**: Depending on the season, the demand for bikes varies.

---

**General Subjective Questions**

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

Linear regression is a supervised learning algorithm used to predict a continuous target variable based on one or more independent variables. It assumes a linear relationship between the independent variables (features) and the dependent variable (target). The algorithm tries to find the best-fit line that minimizes the sum of squared errors (SSE) between the predicted and actual values.

The equation for simple linear regression is:

$y=\beta 0+\beta 1x+\epsilon y=\beta 0+\beta 1x+\epsilon$
Where:

- $yy$ is the dependent variable.
- $xx$ is the independent variable.
- $\beta 0\beta 0$ is the intercept.
- $\beta 1\beta 1$ is the slope.
- $\epsilon\epsilon$ is the error term.

For multiple linear regression, the equation extends to:

$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon$

The coefficients (β0,β1,...,βnβ0,β1,...,βn) are estimated using the **Ordinary Least Squares (OLS)** method, which minimizes the sum of squared residuals.

---

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

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties (mean, variance, correlation, and linear regression line) but are visually very different when plotted. It was created by statistician Francis Anscombe to demonstrate the importance of visualizing data before analyzing it.

The quartet highlights that:

- Summary statistics alone can be misleading.
- Data visualization is crucial for understanding the underlying patterns and relationships in the data.
- Linear regression models may not always be appropriate, even if the summary statistics suggest a good fit.

---

**3. What is Pearson's R? (3 marks)**

Pearson's R, also known as the Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship.
- **-1** indicates a perfect negative linear relationship.
- **0** indicates no linear relationship.

The formula for Pearson's R is:

R=Cov(X,Y)σXσYR=σXσYCov(X,Y)
Where:

- Cov(X,Y)Cov(X,Y) is the covariance between XX and YY.
- σXσX and σYσY are the standard deviations of XX and YY, respectively.

---

**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 data to a specific range or distribution. It is performed to ensure that all features contribute equally to the model, especially in algorithms sensitive to the magnitude of variables (e.g., linear regression, k-nearest neighbors).

- **Normalized Scaling (Min-Max Scaling)**: Rescales data to a fixed range, usually [0, 1]. The formula is:

Xnorm=X−XminXmax−XminXnorm=Xmax−XminX−Xmin

- **Standardized Scaling (Z-score Scaling)**: Rescales data to have a mean of 0 and a standard deviation of 1. The formula is:

Xstd=X−μσ*X*std=*σX*−*μ*
Where μ*μ* is the mean and σ*σ* is the standard deviation.

---

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

The Variance Inflation Factor (VIF) measures multicollinearity among independent variables. A VIF value of infinity occurs when there is **perfect multicollinearity**, meaning one independent variable is a perfect linear combination of other variables. This happens when:

- A variable is a perfect linear combination of other variables (e.g., one variable is the sum of two others).
- A dummy variable is not dropped during dummy variable creation, leading to redundancy.

---

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

A **Q-Q plot (Quantile-Quantile plot)** is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. In linear regression, it is used to check the normality of residuals.

- **Use**: The Q-Q plot plots the quantiles of the residuals against the quantiles of a normal distribution. If the residuals are normally distributed, the points will fall along a straight line.
- **Importance**: Normality of residuals is one of the key assumptions of linear regression. If the residuals are not normally distributed, it indicates that the model may not be capturing the underlying patterns correctly, and transformations or alternative models may be needed.

---