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?**

Inference from Categorical Variables

From the regression model using one-hot encoded variables:

- **Season:** Spring has a **negative impact** on demand compared to the base season (likely winter or fall depending on encoding).
- **Weather Situation:** Light Snow/Rain significantly **reduces** the demand, as seen from its large negative coefficient.
- **Year (yr):** Demand was significantly **higher in 2019** (encoded as 1), indicating growth over time.

Inference: Categorical variables like season, weather condition, and year strongly influence bike demand. Spring and bad weather reduce usage, while a new year increases it.

---

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

Using drop_first=True avoids the **dummy variable trap**, a scenario of perfect multicollinearity in regression. It removes one category per categorical variable, making the model mathematically stable and interpretable.

---

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

From the correlation matrix:

- **atemp (feeling temperature)** has the highest positive correlation with demand: **0.6307**
- Slightly more than temp which is 0.6270

Answer: atemp is the most positively correlated numerical variable with bike demand.

---

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

After training the model:

- **Linearity:** Assumed through the use of a linear model.

- **Normality of Residuals:** Verified using a histogram with KDE (sns.histplot(residuals, kde=True)), which should show a roughly normal distribution.
- **Homoscedasticity:** Though not plotted, checking a residuals vs. predicted values plot would help assess constant variance.
- **Multicollinearity:** Mitigated by using drop_first=True in dummy creation and removing redundant features.

Answer: The model checks residual distribution and ensures clean data via preprocessing and encoding choices.

---

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

From the regression coefficients:

1. **yr** – Huge positive impact: ~1953 rentals increase from 2018 to 2019.
2. **weathersit_Light Snow/Rain** – Large negative impact: ~1874 rentals decrease.
3. **season_spring** – Negative impact: ~1219 rentals decrease compared to base season.

These features most strongly influence demand—either increasing or decreasing it significantly.

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**
   Linear Regression is a supervised learning algorithm used for predicting a continuous dependent variable (y) based on one or more independent variables (X).
   Key Concepts:

- Equation:
  y=β0+β1x1+β2x2+⋯+βnxn+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilony=β0+β1x1+β2x2+⋯+βnxn+ϵ
  Where:
  - β0\beta_0β0 is the intercept
  - β1,β2,...,βn\beta_1, \beta_2, ..., \beta_nβ1,β2,...,βn are coefficients
  - ϵ\epsilonϵ is the error term

  Goal:
  To find the best-fitting line (or hyperplane) by minimizing the sum of squared residuals using a method called Ordinary Least Squares (OLS).
  Steps Involved:

1. Fit the model on training data.

2. Predict the dependent variable for unseen data.
3. Evaluate using metrics like R-squared, RMSE, etc.

---

2. Explain the Anscombe's Quartet in detail.

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties (mean, variance, correlation, and regression line), yet when plotted, they look very different.

Purpose:

- To show the importance of visualizing data before drawing conclusions.
- It warns against relying solely on summary statistics and linear models without understanding the data distribution.

Each dataset in the quartet has:

- Same mean for x and y
- Same variance for x and y
- Same correlation between x and y
- Same linear regression line

Yet, their scatter plots reveal vastly different patterns—some linear, some nonlinear, and one with a clear outlier.

---

3. What is Pearson's R?

Pearson's R (correlation coefficient) measures the linear relationship between two continuous variables.

Range:

- R∈[−1,1]R \in [-1, 1]R∈[−1,1]
    - 1: Perfect positive linear correlation
    - 0: No linear correlation
    - -1: Perfect negative linear correlation

Formula:

R=Cov(X,Y)σXσYR = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}R=σXσYCov(X,Y)

It is sensitive to outliers and assumes normally distributed data with linear relationships.

---

4. What is scaling? Why is scaling performed? Normalized vs. Standardized

Scaling is the process of transforming features so they fit within a particular scale or range.

Why Scaling?

- Required for algorithms that rely on distance or gradient (e.g., KNN, SVM, Gradient Descent).
- Prevents features with large values from dominating others.

Normalization (Min-Max Scaling):

- Scales values between 0 and 1.

- Formula:
  Xnorm=X−XminXmax−XminX_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}Xnorm =Xmax−XminX−Xmin
  Standardization (Z-score Scaling):
- Scales data to have mean = 0 and std = 1.
- Formula:
  Xstd=X−μσX_{std} = \frac{X - \mu}{\sigma}Xstd=σX−μ

---

5. Why does VIF sometimes become infinite?

VIF (Variance Inflation Factor) quantifies multicollinearity—how much one predictor is explained by others.

VIF=11−R2VIF = \frac{1}{1 - R^2}VIF=1−R21

If R2=1R^2 = 1R2=1, VIF becomes infinite, which occurs when a variable is perfectly predicted by other variables (perfect multicollinearity).

Example:

- Including both temp and atemp without checking correlation—they are almost identical, causing high VIF.

---

6. What is a Q-Q plot and its importance in Linear Regression?

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

Purpose in Linear Regression:

- To check the normality of residuals, a key assumption.
- If residuals are normally distributed, points in the Q-Q plot will lie along a 45-degree line.

  Importance:
- Normal residuals validate model reliability for inference (e.g., p-values, confidence intervals).
- Deviations suggest need for transformation or alternative modelling.