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?

From the boxplots drawn for categorical variables (**season, yr, holiday, weekday, workingday, weathersit, mnth**):

• **Season:** Rentals are higher during **summer (2)** and **fall (3)** seasons, while lower in **spring (1)** and **winter (4)**. This shows a strong seasonal effect on bike demand.

• **Year (yr):** The year **2019 (yr = 1)** shows a significant increase in bike rentals compared to **2018 (yr = 0)**, indicating overall growth in usage.

• **Weather situation (weathersit):** Clear or partly cloudy days (**weathersit = 1**) have the highest rentals, whereas misty (**2**) and rainy/snowy days (**3 or 4**) drastically reduce demand.

• **Month (mnth):** Rentals peak during warmer months (June–September) and drop during winter months (December–February).

• **Holiday:** Demand is slightly lower on holidays compared to non-holidays.

• **Weekday/Workingday:** Rentals are consistent across weekdays, with working days showing slightly higher demand than weekends.

Inference:
Categorical variables such as **season, year, month, and weather situation** have a significant impact on bike rental demand, reflecting the influence of climate and time-related factors on user behavior.

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

When creating dummy variables, one category serves as a baseline.

For example, if a variable has three categories, they can be represented as:

- Category 1 → (1, 0)
- Category 2 → (0, 1)
- Category 3 → (0, 0)

This means that for any categorical variable with *n* categories, it can be represented using *n – 1* dummy variables.
Hence, `drop_first=True` is used during dummy creation to remove one redundant column and avoid multicollinearity.

Inference:
Using `drop_first=True` avoids the dummy variable trap by keeping only *n – 1* dummy variables, ensuring a stable regression model without multicollinearity.

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

From the pair-plot among numerical variables (`temp`, `atemp`, `hum`, `windspeed`, `casual`, `registered`, `cnt`):

• Both **registered** and **casual** users show a very strong linear relationship with total rentals (`cnt`).
• Among all independent variables, **registered** has the **highest correlation** with the dependent variable `cnt`.

But this is by definition of registered and casual. The next variable to consider is temp

Inference:
The variable **temp** is most strongly correlated with the target variable **cnt (total bike rentals)**.

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

After fitting the regression model, the following assumptions were validated:

• **Linearity:** Verified using scatter plots and pair plots — confirmed a linear trend between variables like `temp` and `cnt`.

• **Normality of residuals:** Checked using distribution plots (`sns.histplot(y_train - y_train_pred)`), ensuring residuals are centred around zero and roughly bell-shaped.

• **Homoscedasticity:** Examined scatter plots between predicted values and residuals — confirmed variance consistency.

• **Multicollinearity:** Computed **Variance Inflation Factor (VIF)** for all predictors and dropped variables with high VIF (>5).

• **Model robustness:** Compared R^2 values between training and test data to check for overfitting.

Inference:
Residual plots, VIF values, and R^2 comparisons confirmed that the linear regression assumptions were satisfactorily met for the final model.

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

From the final regression model (Method 3.3 / RFE-based model without `registered` and `casual`):
• **temp (temperature):** Positive relationship — higher temperatures increase bike usage.

• **yr (year):** Positive effect — rentals were significantly higher in 2019 compared to 2018.

• **weather_sit:** Negative relationship — light snow and rain negatively influences demand

Inference:
The top three features contributing most to bike rental demand are **temperature (`temp`)**, **year (`yr`)**, and **weather_sit**.

General Subjective Questions

1. Explain the Linear Regression algorithm in detail.

Linear Regression is a supervised learning algorithm that models the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a straight line.

Equation:
Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn + e
where

- b0 = intercept
- b1, b2, ... = coefficients
- e = error term

Steps:

1. Assume a linear relationship between predictors and target.
2. Estimate coefficients by minimizing the Sum of Squared Errors (SSE) using the Ordinary Least Squares (OLS) method.
3. Evaluate model performance using R^2, adjusted R^2, RMSE, etc.
4. Validate assumptions like linearity, normality, and multicollinearity.

2. Explain Anscombe's Quartet in detail.

Anscombe's Quartet is a set of **four datasets** (each with 11 points) created by **Francis Anscombe (1973)** to show that **datasets with identical statistical summaries can look very different graphically**.

Shared Statistics

All four datasets have:

- Mean of x = 9, mean of y = 7.5
- Variance of x = 11, variance of y = 4.125
- Correlation = 0.816
- Linear regression line = y = 3 + 0.5x

Graphical Differences

1. **Dataset I** – Typical linear trend; fits regression line well.
2. **Dataset II** – Non-linear (curved) relationship.
3. **Dataset III** – Outlier strongly affects correlation; other points almost constant.
4. **Dataset IV** – Single extreme x-value dominates the regression line (high leverage).

Key Lessons

- Numerical summaries can be misleading.
- Visualization (scatter plots) is essential.
- Outliers and high-leverage points can distort results.

Takeaway: Always plot your data before analyzing it!

3. What is Pearson's R?

Pearson's R (Pearson Correlation Coefficient)

Definition:
Pearson's R measures the **strength and direction of a linear relationship** between two continuous variables.

Range:

- R=+1R = +1R=+1 → Perfect positive linear correlation
- R=−1R = -1R=−1 → Perfect negative linear correlation
- R=0R = 0R=0 → No linear correlation

Interpretation:

- 0.0 – 0.3 → Weak correlation
- 0.3 – 0.7 → Moderate correlation
- 0.7 – 1.0 → Strong correlation

Formula:

$$R = (\Sigma (X_i - \bar{X})(Y_i - \bar{Y})) \div \sqrt{\Sigma (X_i - \bar{X})^2 \times \Sigma (Y_i - \bar{Y})^2}$$

Notes:

- Measures **linear relationships only**.
- **Outliers** can strongly affect the correlation.

Example:

- R = 0.85 → Strong positive correlation
- R = -0.6 → Moderate negative correlation

4. What is scaling? Why is scaling performed? Difference between normalization and standardization.

Scaling means transforming features so that they are on a similar scale.

Why scaling is done:

- To prevent large-magnitude features from dominating others.
- To help models converge faster.

Types:

1. **Normalization (Min-Max Scaling):**
   $X' = (X - \min(X)) / (\max(X) - \min(X)) \rightarrow$ values between 0 and 1.
   Used when data is not normally distributed.
2. **Standardization (Z-score Scaling):**
   $X' = (X - \text{mean}(X)) / \text{standard deviation}(X) \rightarrow \text{mean} = 0, \text{std} = 1$.
   Used when data follows a normal distribution.

5. Why can the value of VIF become infinite?

VIF (Variance Inflation Factor) measures multicollinearity between predictors.
Formula: $VIF = 1 / (1 - R^2)$

If a variable is perfectly correlated with others, then $R^2 = 1$, making the denominator zero.
Therefore, VIF = infinite.
This happens when one predictor can be exactly predicted from others.

6. What is a Q-Q Plot? Explain its use and importance in Linear Regression. *(3 marks)*

A Q-Q (Quantile-Quantile) plot compares the quantiles of sample data to those of a theoretical distribution (usually normal).

Use:

- If points lie along a 45° line → data is normally distributed.
- If points deviate → residuals are skewed or non-normal.

Importance:
In linear regression, the residuals should be normally distributed for valid statistical inferences.
The Q-Q plot helps visually check this assumption.