Assignment-based Subjective Questions and Answers.

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The categorical variables in the dataset, such as season and weather situation, have a significant impact on bike demand. For example, the data suggests that summer and fall seasons show higher demand, while adverse weather conditions like light rain and mist decrease demand. This indicates that pleasant weather positively affects bike rentals, while bad weather has a negative effect

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Setting drop_first=True during dummy variable creation avoids multicollinearity by dropping one category from each categorical variable. This prevents the dummy variable trap where the variables become linearly dependent, thus ensuring the model is not over-parameterized

---

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Based on the correlation analysis, the variable atemp (feeling temperature) shows the highest correlation (0.63) with the target variable cnt (count of bike rentals). It closely reflects the perceived temperature, making it a strong predictor

---

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of linear regression were validated as follows:

1. **Linearity** - Residuals vs. predicted values scatterplot checked for linear patterns Bike_Sharing_Assignment….
2. **Normality** - Q-Q plot ensured residuals followed a normal distribution Bike_Sharing_Assignment….
3. **Homoscedasticity** - Residuals vs. predicted values scatterplot verified constant variance Bike_Sharing_Assignment….

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 features identified as significant predictors were:

1. **yr** - Indicates yearly trends contributing positively (Coefficient: 987.12).
2. **atemp** - Perceived temperature significantly impacts demand (Coefficient: 416.56).
3. **season_Winter** - Seasonal effect indicating lower demand compared to fall Bike_Sharing_Assignment….

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a statistical method used to model the relationship between a **dependent variable (target)** and one or more **independent variables (features)**.

- **Equation**:

$Y=\beta0+\beta1X1+\beta2X2+...+\betanXn+\epsilonY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$Y=β0 +β1 X1 +β2 X2 +...+βn Xn +ϵ

where:

- YYY = Target variable
- XXX = Independent variables
- $\beta\beta$β = Coefficients showing the impact of each feature
- $\epsilon\epsilon$ϵ = Error term
- **Goal**: Minimize the sum of squared errors (SSE) to make predictions as accurate as possible.
- **Assumptions**:
    - Linearity
    - Independence
    - Normality of residuals
    - Homoscedasticity

Applications: Forecasting sales, predicting trends, and understanding feature influence.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of **four datasets** that have nearly identical descriptive statistics (mean, variance, correlation) but differ significantly when graphed.

Purpose:

- Highlights the importance of visualizing data rather than relying solely on summary statistics.
- Demonstrates how datasets can appear similar numerically but differ structurally.

Example Insights:

- Outliers or patterns may affect regression models even if numerical summaries seem valid.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

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

Formula:

$R=\sum(Xi-X^-)(Yi-Y^-)\sum(Xi-X^-)2\sum(Yi-Y^-)2R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$R=∑(Xi −X−)2∑(Yi −Y−)2 ∑(Xi −X−)(Yi −Y−)

Key Points:

- RRR ranges from **-1 to +1**:
  - **+1**: Perfect positive correlation
  - **-1**: Perfect negative correlation
  - **0**: No correlation

Application: Used in statistics and machine learning to measure dependency.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling standardizes or normalizes features to ensure they are comparable and improve model performance.

Types:

1. **Normalization** (Min-Max Scaling):
    a. Rescales data to [0, 1] or [-1, 1].
    b. Formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2. **Standardization** (Z-Score Scaling):
    a. Centers data around 0 with a standard deviation of 1.
    b. Formula:

$$X_{std} = \frac{X - \mu}{\sigma}$$

Models like Linear Regression and k-NN are sensitive to feature magnitudes. Scaling ensures all features contribute equally.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Variance Inflation Factor (VIF) measures **multicollinearity** among features.

Infinite VIF occurs when:

- Perfect linear dependence exists between features.
- Redundant variables (highly correlated predictors) are included.

Solution: Remove highly correlated features or use techniques like PCA to reduce dimensionality

---

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A **Q-Q Plot (Quantile-Quantile Plot)** compares the **distribution of data** against a **theoretical distribution** (e.g., normal distribution).

Purpose in Regression:

- Tests the **normality of residuals**.
- Ensures assumptions for linear regression are met.

How to Interpret:

- If points fall approximately on the diagonal line, residuals are normally distributed.
- Deviations indicate non-normality, which may affect predictions.
