Assignment-based Subjective Questions

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)

Categories features will have demands as following

- Seasonality: Bike demand is heavily seasonal, with peaks during warm months.
- Weather Sensitivity: Clear weather boosts rentals, while rain and snow decrease them.
- Holidays and Working Days: These indicators affect user type rather than total rentals, as both categories may compensate for each other.

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)

drop_first=True ensures :

- No perfect multicollinearity in regression models.
- Clear interpretation of coefficients relative to a baseline category.
- Decreases the number of features

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)

Temp(warm temperature) feature had highest correlation with cnt

---

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)

To validate the assumptions I used r2_score and also tried the model on test dataset which gave a promising result

R2_score on train dataset - **0.7886**

R2_score on test dataset - **0.7794**

I also used residual analysis which showed error distribution at zero

---

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)

Top three features are

- Temp
- Yr
- spring

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)

<Your answer for Question 6 goes here>
Linear Regression – Simplified Explanation

Linear regression models the **relationship between a dependent variable** (target) and **one or more independent variables** (features) by fitting a **straight line** through the data.

The general equation is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- y: Target
- x_i: Independent variables
- β_0: Intercept
- β_i: Coefficients (effect of each feature)
- ϵ: Error term

How it Works:**
1. Find the best-fit line by minimizing the cost function (Mean Squared Error - MSE):

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

2. Gradient Descent or analytical methods (like Normal Equation) are used to find the optimal coefficients.

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)

<Your answer for Question 7 goes here>

The quartet consists of four datasets, each with:

- **Identical mean** for the x and y variables.
- **Identical variance** of both x and y.
- **Identical linear regression line** (similar slope and intercept).
- **Identical correlation coefficient** (usually around 0.82).

Despite these similarities, the datasets are drastically different in structure, which becomes clear only through visualization.

For data sets

- A typical linear relationship
- A non-linear relationship
- Linear relationship with an outlier
- Vertical outlier affecting correlation

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)

<Your answer for Question 8 goes here>

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)

<Your answer for Question 9 goes here>

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)

<Your answer for Question 10 goes here>

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)

<Your answer for Question 11 goes here>
