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)

Bike Rent is influenced by categorical variables like weathersit, Season as per the final outcomes of this assignment:

1. Bike rent majorly influenced by temperature
2. Bike rent seems to vary with seasons like Summer, Spring, Winter
3. Bike rent impacts by weather conditions like light snow, mist and gets improved by clear weather

---

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)

By drop_first=True, we can reduce the duplicate of information and thereby reduce number of variables at least by 1.

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)

The below 3 variables have high correlation:

1. atemp
2. temp
3. yr

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)
Error terms are normal distribution with mean “zero”
Error terms doesn’t have any specific patterns

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)

Bike rent heavily influenced by below top 3 features:

1. Temperature (temp)
2. Year (yr)
3. Weather (weathersit)

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>

Regression: The output variable to be predicted is a **continuous variable**, e.g. scores of a student

In this Linear Regression is a supervised learning method and it attempts to explain the relationship between a dependent and an independent variable using a straight line

---

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>

Anscombe's quartet comprises of 4 different data set plots with same statistical observations and they appear differently on scatter plots. They fools the regression model if built

---

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>

Pearsons R is a statistical measure which not only evaluates the strength but also the direction of the relationship between the continuous variables

---

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>

Scaling is performed mainly to bring all values of all variables to the common measurement range. There are 2 ways:

1. Min-Max scaler brings all variables values range from -1 till 1
2. Standard way scales based on the maximum value and so it accounts outliers as well

---

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>

VIF = 1 / (1 — R2). R-squared (R2) is a measure of how well the independent variables explain the variability in the dependent variable. If there is perfect correlation, then VIF will be infinite

---

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>

The use of Q-Q (quantile-quantile plot) is a scatter plot to compare the quantiles of two distributions and it is to compare distributions of a given set versus an ideal true gaussian distribution with same mean and deviation

---