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)

Based on the analysis of the categorical variables such as "season", "weathersit", "mnth", "weekday", I could infer varying effects on bike demand (cnt).
Higher bike demand occurs in warmer seasons like Summer and Fall.
Weather conditions also play a important role, with clear weather, high bike rentals occurs whereas snowy or rainy weather correlate with lower bike demand.

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)

It avoids Multicollinearity by removing one category from each dummy representation of categorical variables.

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)

From pair-plot analysis, the variable "atemp" (feeling temperature) has the highest correlation with the "cnt" (target variable) with correlation coefficient of 0.631 approx.

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)

After model building, assumptions of Linear Regression were

validated through Residual analysis.
Residues vs. Predicted values was plotted to check for homoscedasticity, no clear patterns and constant variance.
A histogram of residuals was also plotted to confirm for normality.
These analysis helped to verify that residues are randomly distributed.

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 3 features are:
1. Temperature (temp) with absolute coefficient of 3063.07
2. weathersit_Light_Snow_Rain with absolute coefficient of 2347.74
3. Year (yr) with absolute coefficient of 1999.16

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 is a supervised learning algorithm used for predicting a continuous outcome. It models the relationship between one or more independent variables and a dependent variable by fitting a linear equation to observed data. The algorithm finds the best-fit line by minimizing the sum of squared residuals.
The final equation is of the form,
$Y = M1X1 + M2X2 + \ldots + MnXn$

---

where each M represents the coefficient for a feature.
Linear regression assumes a linear relationship, normality of residuals and homoscedasticity.

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 is a set of four datasets that have nearly identical statistical properties which are generally mean, variance, correlation and regression line. But differ greatly when visualized. It demonstrates the importance of visualizing data rather than relying on summary statistics. The quartet highlights the value of data visualization in identifying nuances like outliers and non-linear relationships.

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>

Also known as Pearson's correlation coefficient, is a measure of the linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship and 0 indicates no linear relationship.
It's calculated as the covariance of two variables divided by the product of their standard deviations.
Pearson's R helps understand the strength and direction of linear relationships.

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 transforms features to a common range or distribution. It is helpful for models like linear regression that can be thrown off by features with vastly different scales.

Normalized Scaling	Standardized Scaling
Squeezes values into a fixed range [0,1]	It adjusts the data so that it has a mean of zero and a std. deviation of 1
Useful when you want all values to fall within the same bounds	Useful when features have different units or ranges, allowing the model to treat them 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)

<Your answer for Question 10 goes here>

VIF becomes infinte when there is perfect multi-collinearity. That means one predictor is a perfect linear combination of others, indicating predictor provides redundant information. To resolve this, we can remove or combine highly collinear features.

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>

A Q-Q plot compares the quantiles of residuals to quantiles of a normal distribution.

In linear regression, it is used to check if residuals are normally distributed. If residuals are normally distributed, the points in a Q-Q plot will lie approximately along a straight line. Any deviations suggest non-normality, indicating the need for data transformation.