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)

Answer: From the dataset received, we can understand that the demand varies such as:

- The demand for bike rentals has increased over time from 2018 to 2019
- The demand is higher in warm seasons such as summer and spring and less in winter season such as winter and fall.
- The rentals are lower in holidays and more during non-holidays and more demand observed during non-holidays.
- Bad weather conditions reduces demands and if it is good weather with normalized temperature then rentals are good.

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)

Answer: It is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy 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)

'TEMP' has the highest correlation among the other numerical variables.

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)

Answer: I used R-Squared which gave a value of .80 which means predictor was able to predict 80% of the variance.

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)

Answer: Year, Holiday and Temperature seems to be top 3 features contributing significantly towards explaining the demand of the shared bikes.

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>

Answer: Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

A simple way to write a linear regression equation is to use the formula

y=a+bx

or, using statistical notation,

y= β0+β1x

Here's what each part means:

y is the dependent (or response) variable you want to predict.

x is the independent (or predictor) variable.

β0 is the intercept. It represents the value of y when x x is 0.

β1 is the slope. It indicates the amount by which y changes on average when x increases by one unit.

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>

Answer: Anscombe's quartet shows four data sets with identical stats but different plots. This underscores the importance of visualizing data to detect trends and outliers.

Anscombe's quartet is a set of four carefully constructed data sets that, while sharing nearly identical summary statistics (such as means, variances, correlations, and even identical linear regression lines), display very different patterns when graphed. This example highlights that relying solely on numerical summaries can be misleading and emphasizes the importance of visualizing data to uncover underlying patterns, detect outliers, or identify nonlinearity that summary measures alone may hide.

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>

Answer: Pearson's R, also known as the Pearson product moment correlation coefficient, quantifies linear correlation between two variables on a scale from -1 to +1.

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>
Ans: Scaling in data processing refers to the process of transforming data so that it falls within a specific range or has certain statistical properties. It's an essential step in data pre-processing, particularly in machine learning and statistics.

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>
Answer: VIF becomes infinite when there is perfect multicollinearity, meaning one predictor is a perfect linear combination of others. This happens when:

- Duplicate or highly correlated features (e.g., temperature in Celsius & Fahrenheit).
- Dummy variable trap (all dummy variables included instead of dropping one).
- Mathematical dependencies (one feature is an exact sum or difference of others).

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>
Answer: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution (usually a normal distribution). It helps assess whether a given dataset follows a specific distribution.
