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)

In terms of seasons, fall season is clear winner here when bike usage increases.
In terms of weather situation, under clear weather conditions bike usage is more.
During holiday's usage of rental bikes reduces.

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)

Dummy variables means categorical data which are represented in binary forms (0/1). This is done to remove the redundancy and avoids multicollinearity. For a category with k levels, we select k-1 variables. By dropping the first level, the dropped category becomes a baseline reference. Coefficients of the remaining dummy variables represent differences relative to this baseline, making the model easier to interpret.

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)

Variable - temp

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)

By checking residuals are independent of each other and its normally distributed. There should not be visible patterns in the error terms.

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)

Windspeed, temp, weekday

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>

Equation for a linear regression is

$$y=\beta 0 +\beta 1\ x+\epsilon$$

Where y = depended variable or target variable

β0 = intercept

β1 = slope (rate of change in y for a unit change in x)

x = independent variable

ϵ = error term (residual)

Step 1 - we formulate the objective function - goal is to minimise the cost function by finding the optimal values of β0, β1. Cost function is given as Mean Squared Error (MSE)

$$MSE=1/n\sum(yi -y^{\wedge} i)^{\wedge}2$$

Step 2 - Optimise the coefficient using OLS (Ordinary Least Square)

Step 3 - Make predictions - Once the coefficients are optimized, the model can predict y values for new x using

$$y^{\wedge} = \beta 0 +\beta 1\ x1 +\cdots+\beta n\ xn$$

Step 4 - Evaluate the model performance using metrics like R-squared , Adjusted R-squared, Root mean squared error

---

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 illustrates the importance of plotting data before you analyse it and build your model.Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics but there are peculiarities that fool the regression model once you plot each data set.

These 4 data sets have nearly the same statistical observations which provide the same information for each x and y point in all four data sets. However when you plot these data sets they look very different from one another.

Anscombe's quartet tells us about the importance of visualising data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers etc)

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>

The Pearson's correlation coefficient (R) is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the

relationship between two variables.It summarises the characteristics of a dataset. It describes the strength and direction of the linear relationship between b/w two quantitative 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 a step of data pre-processing which is applied to independent variables to normalise the data within a particular range.
The collected dataset may contain features varying in magnitude, units and range. If scaling is not done then algorithm will take only the magnitude in account and not units hence incorrect modelling. To solve this issue we need to do scaling to bring all variables to same level of magnitude.
Scaling only affects the coefficients and not the parameters like F-statistics,p-values, R-squared.

Normalisation - We map the minimum feature to 0 and maximum to 1. Hence, the feature values are mapped in a definite range of 0 and 1. It can be given as
Z = x - min(x)/ max(x) - min(x)
Standardised - We transform to have a mean of 0 and standard deviation of 1. Its given as
Z = x - mean(x) / std deviation

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 infinite when there is a perfect multicollinearity among the independent variables in a regression model. VIF is calculated as
1/1-R-squred.
When R-squared = 1, denominator becomes 0 causing VIF to 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>

Quantile-Quantile plots plot the quantiles of the sample distribution agains quantiles of a theoretical distribution. This helps in determining if a dataset follows any particular type of probability distribution like normal, uniform or exponential .

In linear regression, Q-Q plots are used to assess the normality of residuals, a critical assumption for hypothesis testing. Residuals in linear regression model should follow a normal distribution.A Q-Q plot of the residuals helps confirm this assumption.

It can identify deviations from normality, suggesting areas where the model might need improvement.
