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)

Categorical variables influenced the dependent variable count as some of them appeared in the final model that we selected.
From categorical variables, we can infer some below things,
Seasons: The fall season has the highest contributor to the count variable.
Month: June and September months have contributed the highest number of rentals.
Weathersit: clear weather has attracted more bike rentals.
Weekday: most of the days except Sunday achieved constant rentals.
Year: 2019 contributed more rentals compared to the previous year.

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 removes any one of the redundant column to avoid any possible correlation that may occur.

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 and atemp) and (temp and count) are the highest correlation.

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)

For the above question, we will use the error terms and plot the graph which will need to form the normal distribution if it correctly follows the Linear regression. In our final selected lr_5 model, we have achieved the normal distribution in error teams.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of 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)

1. yr (year): Coefficient = 0.2480. This suggests that the demand increases significantly with each year.
2. spring: Coefficient = -0.2581. The negative value suggests that demand decreases in the spring season.
3. snowrain: Coefficient = -0.2146. This suggests that demand decreases significantly during snowy or rainy weather conditions.

These three features have the largest absolute coefficients, indicating they contribute the most to explaining the demand in the model.

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 an machine learning algorithm used for prediction of dependent variables using available independent vars. It uses the principle of linear vector i.e., y=mx+c.

Multi linear regression uses multiple variables and based on formula i.e., yi=B0+b1x1+….+bnxn

Regression is based on 4 assumptions:
1) linearity of variables
2) observations are independent of each other
3) Homoscedasicity i.e., variance of resudial errors are independent.
4) Normality of residuals i.e., normal distribution of errors.

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 quartet includes 4 scatter plot quarters in which where first plot is simple linear relationship. Second scatter plot is not normally distributed but some relationship is there. In third plot, distribution is linear but in presence of outliers. In fourth graph, graph is highly correlated which is horizontal in relationship in presence of outlier.

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>

pearson's r is a numerical representation of strength of linear relationship between the variables. It values ranges from -1 to 1. It depicts the linear relationship of two sets of data.

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 method used to normalize or standardize the data values in independent variables. If feature scaling is not done, then OLS or regression algorithm need to deal with greater values which may affect the model accuracy and can cause inconsistencies. It can be done using two methods i.e., Normalisation: uses the gaussian distribution principle and Standardization uses the MinMaxScaar method weigh from 0 to 1.

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 indicates how much collinearity has increased among variables. Below are possible reasons for VIF to achieve infinite results.

1. Perfect multicollinearity
2. Dummy variables
3. Redundant variables.

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>

Q-Q plot is a tool for comparing the shapes of different distributions. A scatterplot generated by

plotting two sets of quantiles against each other. Due to both sets of quantiles came from same distribution, the points should be form a line.

The q-q plot is used to answer following questions:

Two data sets have similar tail similar tail behaviour.
Data sets have similar distribution shapes.