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)

As per the model prediction, below are the effect on Predicted Variable (Total Bikes Count):

- LightSnowRain -0.256693
- Spring -0.150424
- Nov -0.080866
- Cloudy -0.060485
- Jul -0.059844
- Dec -0.040381
- workingday 0.051266
- Sat 0.056511
- Sep 0.059561
- Winter 0.090467
- year 0.239096

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)

If there are n levels, we only need n-1 variables to identify and/or interpret the data, hence we can drop the unnecessary extra column.

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)

temperature

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)

1. Errors must be distributed normally

    - Calculate Errors using y_test – y_test_pred and plot a distribution plot. The errors must be centered towards the mean Zero.

2. Errors should not have any patterns as such

    - When we Plot errors against train dataset, there should not be any patterns, the errors must be completely random.

3. Errors must be independent of each other

    - Errors when plotted against Predicted data sets should not have similar values, they must be independent.

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 Demands increase by 0.41 times the Temperature (Positive)
- Bike Demands decrease by 0.25 times during Lightsnow (Negative)
- Bike Demands decrease by 0.15 times during Spring (Negative)

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)

- Linear Regression is a method where in we can use the historical data to understand patterns and the dependency of each factor/variable with the target variable and predict values for the required target variable
- There must be a linear correlation with the target variable for any of the factors which are considered for building the model
- We consider p-value, R-sq and prob(F-statistics) to evaluate the model
- p-value must be lower than or equal to 0.05 for all the attributes considered for building the model (5% for errors)
- R Squared value must be as high as possible, which depicts the percentage of data variance that can be explained by the model.
- prob(F-Statistics) must be very low

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)

I Have not used Anscombe's quartet in model building.

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)

I Have not used Pearson's R in model building.

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)

- Scaling is performed to ensure that all the variables used in the model have the values in same

---

range.
- Standardized Scaling uses mean and standard deviation for scaling the values as below
  x = x – mean(x) / standard_deviation(x)
- Normalized Scaling also know as MinMax Scaling uses min(x) and max(x) to scale the values.
  X = x – min(x) / max(x) - min(x)

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)

- VIF basically indicates multicolinearity. Higher the value, higher the correlation. As per the formula, VIF = 1 / 1 – R2.
- If it is infinite, the only explaination is R2 score is exactly 1. Which means that there is overfit on the data without any errors whatsoever.

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)

I have not used Q-Q plot in Linear Regression