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)

We can infer the below points:

- Season, weekday, month, holiday, weathersit seems to explain some variance in cnt
- one of season and month is redundant
- workingday, weekday does not seem to have an impact on cnt

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)

For n categories n columns are typically created. However one column can be represented by all the other being 0. This would result in multi collinearity. Hence to solve this problem, one column needs to be dropped. Drop_first=True drips the first and avoids this problem by dropping the redundant 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)

temp and atemp both have very high degree of co relation with the target variable cnt. It is clear the temp and atemp are co linear and one of them will eventually be dropped from the model.

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. Residual Plot: Residuals(observed vs predicted) should be distributed around 0. Spread of residuals remains same for all fitted values. Also bell curve suggests normal distribution.
2. Pair/Box plot plot between the different predictor variables and target variable
3. VIF has been brought less that 5 to deal with multi collinearity.
4. Durbin-watson: This value close to 2 indicates that there is likely no autocorrelation in the residuals, suggesting that the model's errors are independent.
5. **F-statistic: 233.3 (p-value: 5.71e-181):** This very high F-statistic and an extremely low p-value imply that your overall model is statistically significant.
6. **R-squared: 0.824:** This indicates that 82.4% of the variance in bike demand is explained by the variables in your model. Adjusted R-squared: 0.820 This is a slight reduction from the R-squared value, accounting for the number of predictors. It still suggests a good fit, with 82% of the variance explained.
7. R-Squared for test is ~81%. Vary close to training result

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. Year: Highly significant. Positive impact on bike demand.
2. Temp: Highly significant. Warmer temperatures increase bike demand.
3. WeatherSit: Highly significant. Bad weather (snow/rain) decreases bike demand. Mist/cloudy weather reduces demand.

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>

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>

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>

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>

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>

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>
