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 variable can also contribute more in this study. Because more variables are categorical variables contributed when comparing numerical variables here in this exercise.

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)

While creating dummy variables, its is enough to interpret or regain the original variables with only few dummy variables. We use this to remove unnecessary variable usage.

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 has high correlation with 'cnt'

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. Checked the p-Value
2. Checked the R and Adjusted R
3. Checked VIF

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)

Yr, fall and summer

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>

1. Loading Essential Libraries and Dataset
2. Visualizing the data
3. Finding the correlation
4. Splitting the dataset into Training and Testing set
5. Include Highly correlated Value
6. Build the model.
7. Check the VIF factor(to check multi-collinearity).
8. Remove the Variables one by one after checking the performance of the model
9. Repeat step 7 and 8 - till value of VIF of all included variables are under 5.
10. Test the model with High accurate LR models (Last One is chosen out of seven)
11. Do the residual analysis
12. Check the accuracy.
13. Accuracy of training data is 0.76 and test data is 0.74 in this case

---

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 is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading if not visualized.

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>

It says about the Linear relationship between the variables. The value ranges from -1 to 1
-1 says negative correlation I.e. when one variable increases the other one decreases
0 says nil correlation – no correlation
1 says strong correlation – if one goes up another one also increases

---

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 data preprocessing technique applied on the variable to handle data with different boundaries. It is performed to take a look at the dataset in a similar scale e.g. salary may range from 10k to 1000k but age ranges from 1 – 120. It will be difficult to visualize without scaling these variables. We have two methods, Min-Max and standard scaling. Min-Max scales the value between 0 and 1. Standard scaling makes the data to move towards 0.

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>
When the multi collinearity is more, VIF becomes infinite. The formula says 1 divided by (1-rsquare) . When the denominator becomes 0 (more correlation between variables says Rsquare =1 and 1 – 1 is 0), the VIF becomes 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>
Q-Q is Quantile-Quantile Plot , which helps us to compare the Observed(Practical) and Actual(Theoretical) values- for the model validation. In linear regression, we check if the errors are normally distributed which can be done with the help of Q-Q plot.
