Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Spring, lightsnow contributes to less demand.

2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

When there are P values then P-1 dummy variables can explain the feature well.

For example if there are 3 possible values like rainy, spring, summer if value is not rainy, spring, it reflects summer.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temperature has the highest correlation with the target variable cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Multicolinearity: Made sure that VIF is less than 5.

Homoscedasticity: variance of residuals is within constant variance at every level.

Error terms are normally distributed with test & train data.

Linear dependency of target variable on few variables like temp.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Temp, year, light snow are the top 3 factors.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

2. Explain the Anscombe’s quartet in detail. (3 marks)

3. What is Pearson’s R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)