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)
    1. Usage is low in spring
    2. Usage is low in the beginning of the year
    3. Usage is low during rain
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    1. Setting drop_first creates n-1 dummy variables which in turn reduces the amount of data to be processed
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    1. With target variable cnt, temp and atemp have the highest correlation
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    1. Error terms have a normal distribution
    2. Multi-colinearity not present
    3. VIF values under 5
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    1. Bike demand is based on temperature, weather and year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

    Linear regression algorithm is a supervised learning algorithm which performs regression task. It makes prediction based on independent variables. It is used to make a prediction for continuous variables.

    Goal of linear regression is to find the best fit line with least errors. Cost function is used to estimate the values of the coefficient of the best fit line. It optimizes the coefficients and is used to fund the accuracy of the hypothesis function.

2. Explain the Anscombe's quartet in detail.
3. What is Pearson's R?
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

    Scaling is used to normalize the range of numeric/continuous variables. Normalization is preferred in algorithms like neural networks as impact of outliers is high and standardization in case of clustering analysis

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

    This happens due to the presence of multi-collinearity in the data.

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