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)
    a. mnth – has integer values of months. Converted to String names of months. There is not much correlation with the dependent variable.
    b. weekday – weekday also converted from integer to Day names. Later converted to dummy variables.
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    a. It will remove one column from the created dummy variables. This will simplify the dataset. Also, we don't loose any thing since the we can interprest the value of missing variable from the other variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    a. temp and atemp has the high correlation with the cnt variable.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. Residual validation. Mean should be 0. The graph should be like distribution plot.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    a. temp, hue and yr

General Subjective Questions

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

    a. Linear regression is machine learning algorithm. It derives a linear equation between explanatory variables and the dependent variable.

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

    a. Ascombe Quartet is a set of 4 datasets, which have similar statistics like mean, variance etc.

3. What is Pearson's R? (3 marks)

    a. Pearson's R indicates the direction and strength of correlation between two variables.

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

    a. Scaling is done to fit sequential values between o and 1. This helps us to compare the relative affect of variable on the target variable.

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

    a. If the colleration between variables is perfect then the VIF value will be infinite.

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

Q-Q plot is Qualtile-Quantile plot.