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)

Ans: The dataset had categorical variables such as “seasons”, “weathersit”, “holiday”, “weekday”, “workingday” and “month”. When we separately plot these categorical variable with the “cnt” target we can see a pattern of dependency and influence by these categorical variables on shared bikes demand.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans: To avoid multicollinearity with the created dummies we need to use drop_first=True during dummy variable creation.

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

Ans: “atemp” & “temp” numerical variables has the highest correlation with the target variable.

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

Ans: The residuals of the linear regression model built with the training set were estimated and plotted. The residuals seem to follow a normal distribution which validates the assumptions made in linear regression model.

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

Ans: The top 3 features contributing significantly towards the demand of shared bikes were found to be “atemp”, “winter”, “mist/winter”

General Subjective Questions

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

Ans: Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task to fit a linear relationship model between dependent variable and independent variable. The linear regression model is used for predicting the dependent variable with the help of independent variable on a continuous basis.

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

Ans: Anscombe’s quartet tells us that all the data should be graphically analysed to understand the relationship, outliers, linear/non-linear behaviours and the effect of outliers in producing high r2 values. There are four plots in Anscombe’s quartet. First one tells us about the linear relationship between two variables, second one the non-linear relationship, the third one on the outliers and the fourth one on the effect of outliers in producing a false high r2 value.

3. What is Pearson’s R? (3 marks)

Ans: Pearson’s R is a measure of linear correlation between two variables. It is ratio between the covariance of two variables and the product of their standard deviations.

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

Ans: Scaling is the process of transforming all the variables into a uniform range. It is performed to better understand the model correlations and prevent errors of choosing higher value features and dropping lower value features. In normalized scaling the data is scaled between 0 & 1 using the min & max of a variable. In standardized scaling the data is scaled using the mean and standard deviation to get a 0 mean & 1 variance value.

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

Ans: When the correlation between two variables is 1, R2=1 and VIF is infinite when applied to VIF equation.

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

Ans: Q-Q plot is a graphical tool to assess if a dataset follows some theoretical distributions such as Normal, exponential or uniform distributions. It helps us to identify if training and test dataset are of same distribution or different distribution.