Assignment Based Subjective Questions:

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: Dependent variable(cnt) is heavily dependent on season, weathersit and mnth categorical variables. Season and mnth exhibit multicollinearity. So, season and weathersit matters/influences the most in predicting the demand of bikes

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: To represent N categories in a categorical variable, we need only N-1 dummy variables. The get_dummies function returns N dummy variables and if we don't use drop_first=True which means one variable is correlated with other N-1 variables. So, it is needed to remove the correlation among the N dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: temp

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: 1. Plotted the distribution of error terms (Yactual - Ypredicted) on the test and train set and found to be normal with mean zero. **2.** Error terms are independent of each other. **3.** As the provided problem is MLR, the fitted linear regression curve is a hyper plane which we cannot visualize so cannot say that error terms are having constant variance.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: temp, yr and Light Rain (weathersit: 3)

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is used to predict the dependent variable which is continuous, based on the history of independent variables which exhibit linear relationship and correlation with the dependent variable. To find the optimal linear regression model, we use gradient descent algorithm in an iterative fashion until the Cost function (Mean Square of residual error) takes a minimal value.

The direction to find optimal model coefficients is provided by the gradient descent algorithm. Constant (Intercept) and model coefficients (Betas) are updated every iteration to re-compute the linear regression model. This process continues till the cost function reaches a minimal value (theoretically zero) at which gradient descent stops and the final computed coefficients are the model coefficients for the linear regression model.

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, but have very different distributions when graphed. This quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Q3. What is Pearson's R? (3 marks)

Ans: It measures a linear dependence between two variables also known as parametric correlation. In linear regression, Pearson's correlation helps us to understand the relationships between the feature values (independent values) and the target value (dependent value or the value to be predicted)

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: If the predictor variables are on different/contrast scales, then the model coefficients that we obtain from linear regression modelling will be high/very high for predictors represented in low scale vs low/very low for predictors represented in high scale. Sometimes, the model might ignore the model coefficients which are very small though it is significant just because scaling is not done on that predictor.

Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: Value of VIF being infinite means that this variable can be expressed by a linear combination of other variables

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: The Q-Q plot is a graphical method for determining whether two samples of data came from the same population or not. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

In linear regression, Q-Q plot helps to understand whether the training and test data set are received separately. To confirm this, A 45-degree line is plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. If the two data sets have come from populations with different distribution, then the data points will be far from the reference line.