Assignment Based Subjective Question:

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**
   The categorical variables in the dataset are important to the final results, however they can end up skewing the results. This is because the categorical variables are assigned integer values which could imply a higher weight for each category when that could not be the case. This is seen in linear regression models. Hence, categorical variables must be converted into multiple fields via an encoding technique to only get the relevant data from the field. Here binary encoding or one-hot encoding can be used.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   This is an important attribute to use as it can remove the extra column created during this process. This reduces the overall dimensionality and the correlations created among the dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   The year field shows the highest correlation. This is followed by the month and the season fields.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   The relationship between the target variable and the others are linear as seen when doing pair plots of data with the target variables. The variables are independent of each other with the exceptions of month and season and to an extent the weather situation variables. Hence, the level of multicollinearity is low as well.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   The year, months_0 and the season_1 attribute are significant. These can be interpreted as the months of 8, 9, 10, 11, 12 and the seasons of fall and winter.

General Subjective Question Answers:

1. **Explain the linear regression algorithm in detail. (4 marks)**
   The algorithm involves finding the coefficients for each of the dependent variables and the overall bias required to offset the values. The process begins with data collection and standardization to normalize the data. The linear regression model is then defined.

The coefficient is calculated using gradient descent over epochs. The gradients of the loss with respect to a particular variable is calculated. The weights of the model are then updated with a learning rate. The loss function is used to determine the offset of the result with the expected answer. These steps are then repeated until the epochs are complete which is the training phase of the model.

The coefficients are then interpreted by scaling them back. This can then be used to test the model and calculate the new results.

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Anscombe's quartet is a set of four datasets that appear to be completely different however have identical descriptive results such as mean, variance, correlation and the overall linear regression model that can be generated for all.

   The datasets have different distributions as well. The takeaway is to plot the data to be analyzed before performing analysis on the dataset. The four datasets also make use of outliers to bring the descriptive values to the same points. There are other variants of the same concept including Datasaurus Dozen.

3. **What is Pearson's R? (3 marks)**
   Pearson's r value is in the range of -1 to 1 inclusive and is used to both measure the strength of correlation between two variables and the positive or the negative relationship type between the variables. The higher the value the better the correlation. It is symmetric as the relationship is based on correlation which is also a symmetric function. This is squared to sometimes only indicate the level of correlation and not the type of correlation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Scaling is the conversion of a set of values to fit a range or to change a descriptive statistic such as mean or variance. This can be used to improve the performance of a model. Normalization is used when the data does not follow a Gaussian distribution and is useful in neural networks while standardization is used when the data does follow a Gaussian distribution and it does not require the data to be bounded. This preserves outlier properties in the dataset.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   This is observed when the correlation is 1 between two variables. If R2=1 then 1/(1-R2) is infinite. The variance inflation factor can give infinite values and can be solved by removing a field to reduce multicollinearity.

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

   A Q-Q plot is a Quantile-Quantile plot and is when two quantiles are plotted against each other. This is to find if two sets of data have a particular type of probability distribution. If they are similar then the points in the plot will approximate a line. These can also help determine normality which can be an important assumption for linear regression.