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?

The only categorical variable present on this dataset is the 'dteday' which is a unique variable that matches the number of observations (730) on the dataset which I consider an insignificant variable.

The variables 'weathersit' on the dataset is already encoded to numbers. confirmed from the data dictionary.

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

mainly reduce or eliminate the redundancy and correlation that will be caused by creating the dummy variables, and also optimize the computation by removing 1 more column.

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

'registered' variable is highly correlated with 'cnt' the target variable (95%).

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

Using the adjusted R_squared to evaluate the accuracy of the model, and also using the p-value to eliminate the insignificant variables and their effect on the adjusted R_squared value.

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

The most 3 features that explain well the regression model are :

Feature: registered, Score: 0.78683
Feature: casual, Score: 0.28184
Feature: weekday, Score: 0.00371

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

A supervised learning algorithm, that performs a regression task to predict a continuous output based on relations between the target (dependent) variable and the independent variables.

2. Explain the Anscombe's quartet in detail.

Is a technique used for a group of datasets that is identical on some statistical measures but at the same time are different on the distribution forms.

3. What is Pearson's R?

A Pearson's R or Pearson correlation coefficient it's a unit of measuring the correlation coefficient between 2 datasets that return a value between -1 and +1 to determine if the correlation is a negative or positive one.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling the Data is to transform the dataset within a specific range, and the difference between normalized and standardized scaling is too sample, Normalized scaling is transforming the data in a range between 0 and 1 or -1 and 1, in another hand Standardized scaling is taking the mean 0 as a reference and use the standard deviation measures to transform the data.

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

The VIF is infinite when the correlation is perfectly matching.

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

The quantile-quantile plot is mainly a graphical technique that is used to review and determine the distribution of datasets against each other's quantiles, if the datasets are from the same distribution then we will get points forming nearly a straight line, and it's a good technique to found outliers too.