Answers to 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?**

   From the analysis of the categorical variables from the dataset, it can be inferred that the total count of the bike rentals (users) is dependent on Various categorical variables like season, weather situation, working day and holiday.
   There is no clear demarcation of the dependence of total bike rentals on different categorical variables but overall they affect the number of bike rentals in some or the other way.

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

   drop_first=True is quite important as it helps in reducing the extra column created during dummy variable creation. So this thing reduces the correlations created among dummy variables. So if we have categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables (as values of one of the dummy variables become insignificant e.g. From the 'season' column, four dummy variables are extracted- 1:spring, 2:summer, 3:fall, 4:winter. The values of the first dummy variable is not of any use. So it is dropped from the dataframe.

   ```
   pd.get_dummies(df, columns=[season], drop_first=True)
   ```

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

   The highest correlation of the target variable is with is with **atemp** variable.

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

   The basic assumption is that that there should be a linearity between the the independent and dependent variables.
   No multicollinearity should be there among the independent variables.
   Homoscedasticity should be there have constant variance at every level of x.
   There should be a normal distribution of error terms (the error terms' normal distribution).

   So after building the model on the training set, it can be said that most of the assumptions are valid as there exists linearity. Although multicollinearity can't be overlooked but still the model is quite acceptable.

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

   The top 3 features contributing significantly towards the demand of the shared bikes: temp/atemp, season(summer, winter)

Answers to General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear regression is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variables. It determines the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. In case of a simple linear regression, least squares method is used to discover the best-fit line for a set of paired data. After that, it estimates the values of the dependent variable from the independent variable.

Management of any organization can make better decisions by using linear regression techniques. Organizations collect ample amount of data and use that data to better manage reality instead of relying on intuitive predictions. Large amounts of raw data can be taken and it can be transformed to actionable information.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four data sets which are nearly identical in simple descriptive statistics, but there are some specialities in the dataset that makes the regression model error prone if built. They have very different distributions and appear differently when plotted on scatter plots. The Quartet tells us about the importance of visualising the data before using any ways for generating the models.

If these models are plotted on a scatter plot all datasets generates a different kind of plot that is not interpretable by any regression algorithm.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

3. What is Pearson's R?

It is a very popular correlation coefficient used in in linear regression. It can take can take Value between -1 and + 1

If the correlation Coefficient is +1 that means for every unit increase in a variable there will be a positive increase in the value of the other variable also.

If the correlation Coefficient is -1 that means for every unit increase in a variable there will be a decrease in the value of the other variable.

Zero(0) means that the variables are not related at all.

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

While performing regression, sometimes the units of the variables which are used for analysis are different. So the variables are scaled so that easily comparison can be done between the variables and the plots give a meaningful output.

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

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset