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?

Without being converted to dummy variables (aka numerical), I couldn't have understood their effect at all.

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

To reduce the number of derived variables to the least possible, since the dropped value will have zeros across others of the same category.

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

After dropping unnecessary variables such `casual` and `registered` for being just equivalent to `count`, and also `dtedate` and `instant` for being dummy data with no relevance to the `count`, and finally, `temp` was dropped for the favor of `atemp`, since people would decide on going out based on the real feel of the temperature.
The variable with highest correlation to count was `atemp`. However, without dropping any variable, `registered` would be of a highest correlation to `count`.

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

By drawing a heatmap to see how much did the correlation with count has changed. I found that atemp and year has slightly changed from 0.63 to 0.59 and 0.42 to 0.39 respectively. This slight change which is less than 5% suggests that the assumptions are valid.

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

Top 3 features are `atemp` with coef `2.84`, then `year` with coef `1.93`, followed by `season_summer` with coef `1.50`

General Subjective Questions :

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method that is used to establish a relationship between a dependent variable and one or more independent variables. It is a supervised learning algorithm that is used for regression analysis. The goal of linear regression is to find the best fit line that can explain the relationship between the dependent variable and the other independent variables. The line is represented by the equation `y = mx + b`, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept. The slope of the line represents the change in y for a unit change in x. The y-intercept represents the value of y when x is equal to zero.

The y-intercept is also known as ß0, while the slope is known as ß1 in hypothesis testing, which is part of how the linear regression algorithm works.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of four datasets that have the same mean, standard deviation, and regression line, but which are qualitatively different. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

The datasets are as follows:

- Dataset I: A simple linear relationship between x and y.
- Dataset II: A non-linear relationship between x and y.
- Dataset III: A simple linear relationship between x and y, but with an outlier.
- Dataset IV: A simple linear relationship between x and y, but with a different outlier.

The datasets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when the data is plotted, it is clear that the datasets are qualitatively different.

Anscombe's quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It is a number between -1 and 1. A value of -1 indicates a perfect negative correlation, a value of 0 indicates no correlation, and a value of 1 indicates a perfect positive correlation.

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

Scaling is process of changing a set of variables to fall into a specific range. It's performed to unify all values across a data set to certain range which will in turn produce realistic results in linear regression. Normalized Scaling scales the values to have a unit norm, while standardized scales the values between -1 and 1.

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). If there is perfect correlation, then VIF = infinity. 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.

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

A Q-Q plot or quantile-quantile plot is a graphical tool for determining if two data sets come from populations with a common distribution such as a Normal, Exponential, or Uniform distribution. In linear regression, a Q-Q plot is used to check if the residuals of a regression model are normally distributed. If the residuals are normally distributed, the points on the Q-Q plot will fall along a straight line. If the residuals are not normally distributed, the points on the Q-Q plot will deviate from a straight line.

The Q-Q plot is an important tool in linear regression because it helps us to check the assumptions of the regression model. If the residuals are not normally distributed, the regression model may not be appropriate for the data. In such cases, we may need to transform the data or use a different type of regression model.