Assignment-based Subjective Questions:

1. Categorical columns: season, mnth, weathersit ,weekday

i)Bikes rented are the most during the Fall season.

ii) Bikes rented more in the month of Sep and June.

iii) Bikes rented more in the Clear weather.

iv) Bikes rented more on Sunday and Thursday

2. 'Drop_first=True' during dummy variable creation is needed to drop the previous column values after replacing it by dummy values.

3. Temp and atemp have the highest correlation with the target variable.

4. To validate the assumptions of Linear Regression after building the model on the training set:

i) Checking whether the error terms are normally distributed or not.

ii) Checking the residuals have constant variance or not.

iii) Pair-wise scatterplots may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.

5. Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are 'yr'(year), 'Mist'(weather), 'Aug'(month August).

General Subjective Questions:

1 Linear Regression is a machine learning algorithm based on supervised learning. It is mostly used for finding out the relationship between variables. Linear regression performs the task to predict a dependent variable value based on a given independent variable.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables.

Mathematically, we can write a linear regression equation as:

$y=a+bx$ where,

b = Slope of the line.

a = y-intercept of the line.

Once we find the best a and b values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the a and b values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

2. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that numerical calculations are exact, but graphs are rough. It has been rendered as an actual musical quartet.

i)The first scatter plot appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

ii)The second graph is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

iii)In the third graph, the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

iv)Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The 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. The Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviation. If the value is greater than 0, means the two variables are positively correlated and if less than 0, that means they are negatively correlated.

4. Scaling is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range.

It's used because, sometimes, it also helps in speeding up the calculations in an algorithm.

Normalized scaling typically rescales the values into a range of [0,1]. Standardized scaling, rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

5. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model. This shows a perfect correlation between two independent variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. In statistics, a Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Since this is a visual tool for comparison, results can be very useful in the understanding underlying distribution of variables.