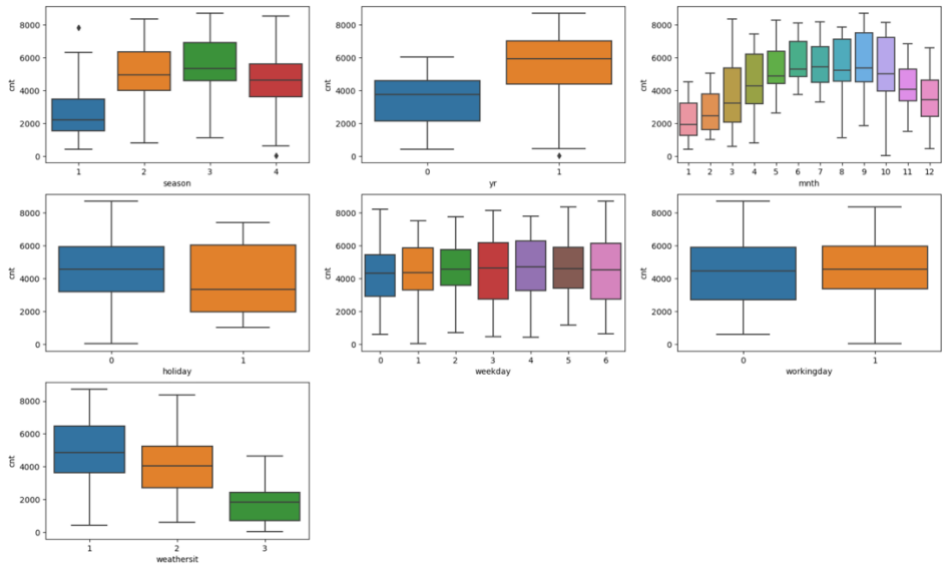
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Following is the graph of all categorical variables.



From the graphs this is what we can infer

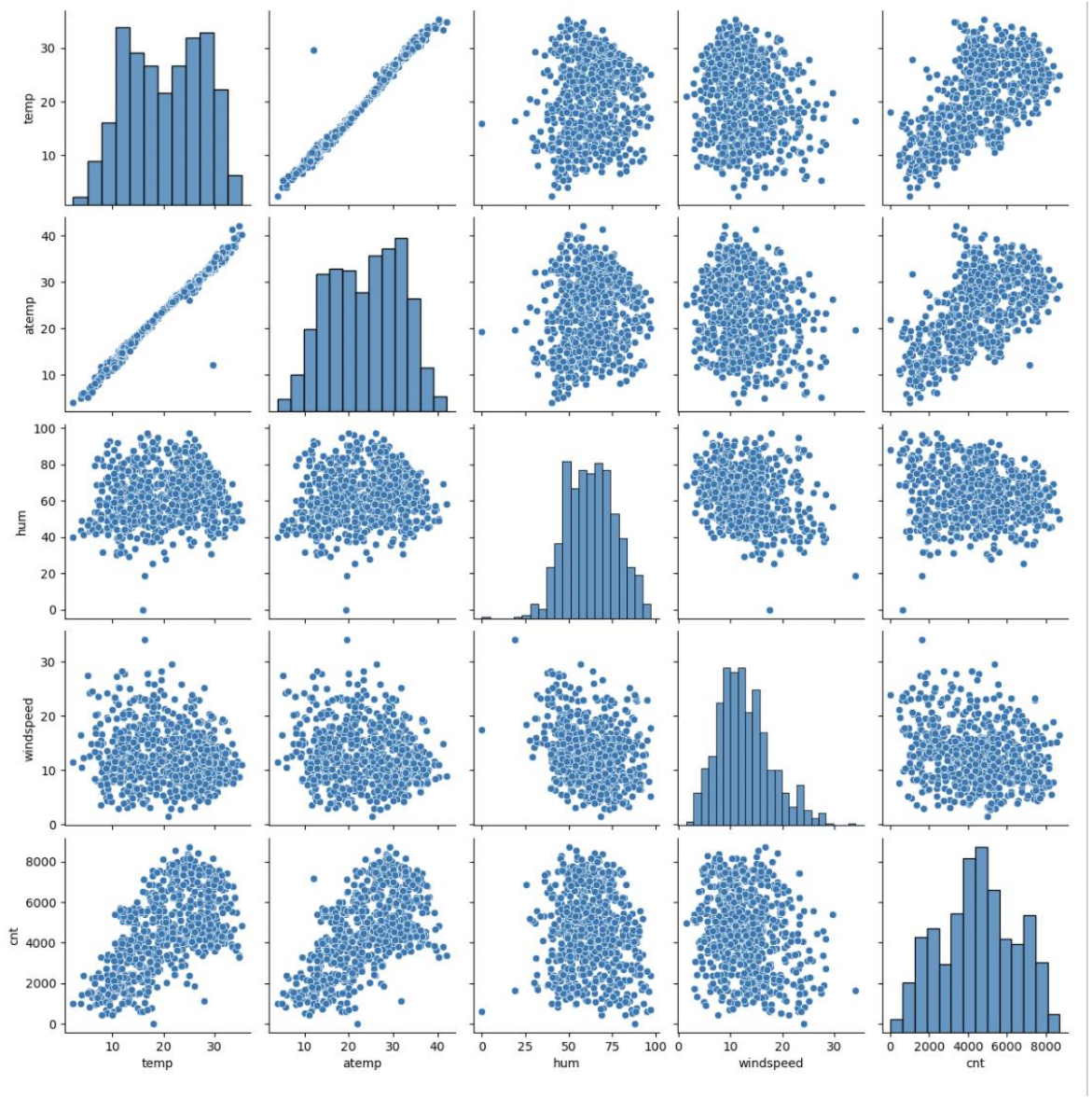
1) The seasons changes the usage count a lot.
2) There seems to be increase in demand based on weather condition as well.
3) The weekdays don't make much impact on the demand.
4) Looks like working day and holiday kind of complements each other.
5) The usage varies based on month.
6) There seems to be a corelation between month and season.

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

Let's say we have N distinct values in our categorical variables. So, if we use get_dummies function, this does a "one hot encoding" of the values that is for N distinct values it adds N columns, and each columns showing 1 for existence of the categorical value and 0 for non-existence. Now we can always represent one categorical value by showing 0 to all the other columns. So, in this way instead of needing N columns we can use N-1 columns to represent N distinct values. Now to do this in a single step we use drop_first=True, which would by default drop the column representing the first categorical value.

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

The pair plots are looking as following –



From the graph it looks like both temp & atemp have the almost the similar high corelation with cnt (count). So, to find out the actual value we can look at the correlation numbers. Which is given below –

]:

	temp	atemp	hum	windspeed	cnt
temp	1.000000	0.991696	0.128565	-0.158186	0.627044
atemp	0.991696	1.000000	0.141512	-0.183876	0.630685

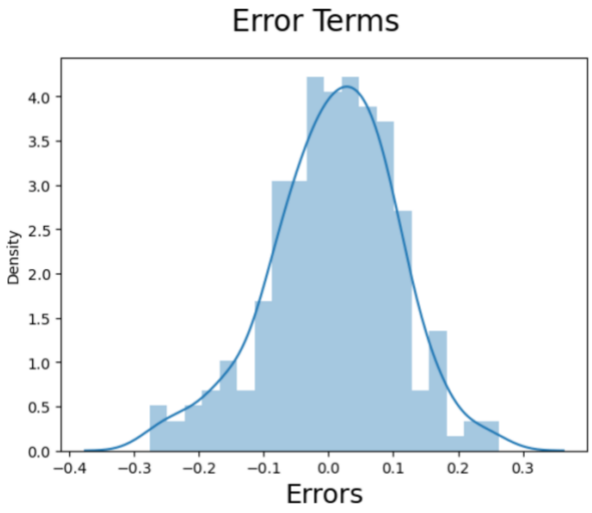
From these numbers we can see that atemp (0.63) have a slightly higher corelation with count than temp (0.627).

So, for numeric variables **atemp** have the highest correlation with the target variable.

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

To test the assumptions of linear regression model we have performed the following tests.

1) We have tested the r-squared value of the test set and we found it to be 79.96% which is close to 80%.
2) We have also tested the residual errors and found it to be normally distributed



3) We have also looked at the VIF values to make sure that there is no collinearity between the variables.

	Features	VIF
6	temp	4.06
1	workingday	3.99
0	yr	1.86
3	Sun	1.51
2	s_spring	1.48
5	s_mist	1.45
4	s_light_rain	1.06

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

Here is the coefficient of the various variables -

```
temp            0.373705
yr              0.235115
const           0.221305
Sun             0.061824
workingday      0.053364
s_mist         -0.075443
s_spring       -0.160892
s_light_rain   -0.291466
```

Now from these values we can the top 3 features are.

1) Temp
2) S_light_rain
3) Year

General Subjective Questions

1. Explain the linear regression algorithm in detail

The linear regression algorithm is a statistical modelling algorithm that finds the relation between a target/dependant variable (y) with a set of independent variables (x_1, x_2, ... X_n). Often these independent variables are not dependant on each other. Also, for linear regression it is assumed that there is a linear correlation between the dependant and independent variables. The linear regression formula can be represented as follows.

$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_n x_n + e$

Where –

y – The target/dependant variable

$x_1..x_n$ – Represents the independent variables

b_0 – Represents the interceptor of a linear formula.

$b_1..b_n$ – Represents the coefficients or slopes of the linear formula

e – Represents the error in the formula. That is the variability of y which cannot be explained by independent variables.

The goal of the linear regression model is to calculate $b_0..b_n$ such that the sum of squared difference of the observed value and the predicted value by the model are minimum.

One of the well-known methods of calculating the same is called Ordinary Least Squared (OLS). There are prebuilt libraries in python which can be used for calculating the OLS value.

Once a model is built then this model can be used for future predictions.

2. Explain the Anscombe's quartet in detail?

Anscombe's quartet refers to a set of four datasets that have nearly identical simple descriptive statistics, including means, variances, correlations, and linear regression lines. However, when plotted, these datasets exhibit vastly different characteristics, challenging the assumption that summary statistics alone are sufficient to understand the underlying data.

The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphical data exploration and the potential pitfalls of relying solely on summary statistics. Each dataset consists of 11 points, which are grouped into pairs of x and y coordinates.

The four datasets within Anscombe's quartet are designed to have the following characteristics:

1. **Dataset I:** A simple linear relationship between *x* and *y*.

2. **Dataset II:** Like Dataset I but with one outlier that greatly affects the linear regression line.

3. **Dataset III**: A non-linear relationship between *x* and *y* that is not captured by simple linear regression.

4. **Dataset IV:** No apparent relationship between *x* and *y*, but the summary statistics are almost identical to those of the other datasets.

Anscombe's quartet is often used to emphasize the importance of data visualization in exploring and understanding data, as well as to highlight the limitations of relying solely on summary statistics to draw conclusions about a dataset. It serves as a reminder that visual inspection of data can reveal insights that summary statistics alone might miss.

3. What is Pearson's R?

Pearson's r, commonly referred to as the Pearson correlation coefficient, used to find the relation between two variables say X and Y. Where –

- **r = 1:** Perfect positive correlation. This means that as one variable increases, the other variable also increases proportionally.
- **r = -1:** Perfect negative correlation. This means that as one variable increases, the other variable decreases proportionally.
- **r = 0**: No linear correlation. There is no relationship between the two variables.

Pearson's correlation coefficient is widely used in statistics to assess the strength and direction of the linear relationship between two variables.

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

Scaling is the process by which we bring all the numeric independent variables in a dataset to a smaller value (between 0-1 or -1 to 1) so that none of the variable overpower other independent variables.

The difference between normalized scaling and standardized scaling are as follows –

- Normalized scaling rescales the data to a fixed range (usually 0 to 1), while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.
- Normalized scaling is useful when the range of the data is meaningful, while standardized scaling is useful for comparing variables with different scales or when assuming a normal distribution of the data.

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

If VIF is infinite this means, there is a perfect collinearity between this variable and at least another variable in the set. This means we can drop any one of the variables and the model should still be able to explain the dependant variable.

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a given dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset to the quantiles of the theoretical distribution.

Here's how a Q-Q plot is constructed:

1. First, the dataset is sorted in ascending order.
2. Then, for each data point, the corresponding quantile of the theoretical distribution is calculated.
3. Finally, these pairs of quantiles are plotted against each other.

If the dataset follows the theoretical distribution closely, the points on the Q-Q plot will approximately fall along a straight line. If the points deviate from the straight line, it indicates that the dataset's distribution differs from the theoretical distribution.

In linear regression we can use Q-Q plots for assessing the normality of a dataset. In a Q-Q plot comparing a dataset to a normal distribution, if the points closely follow the diagonal line (45-degree line), we can assume that the dataset is approximately normally distributed. Deviations from the diagonal line indicate departures from normality.