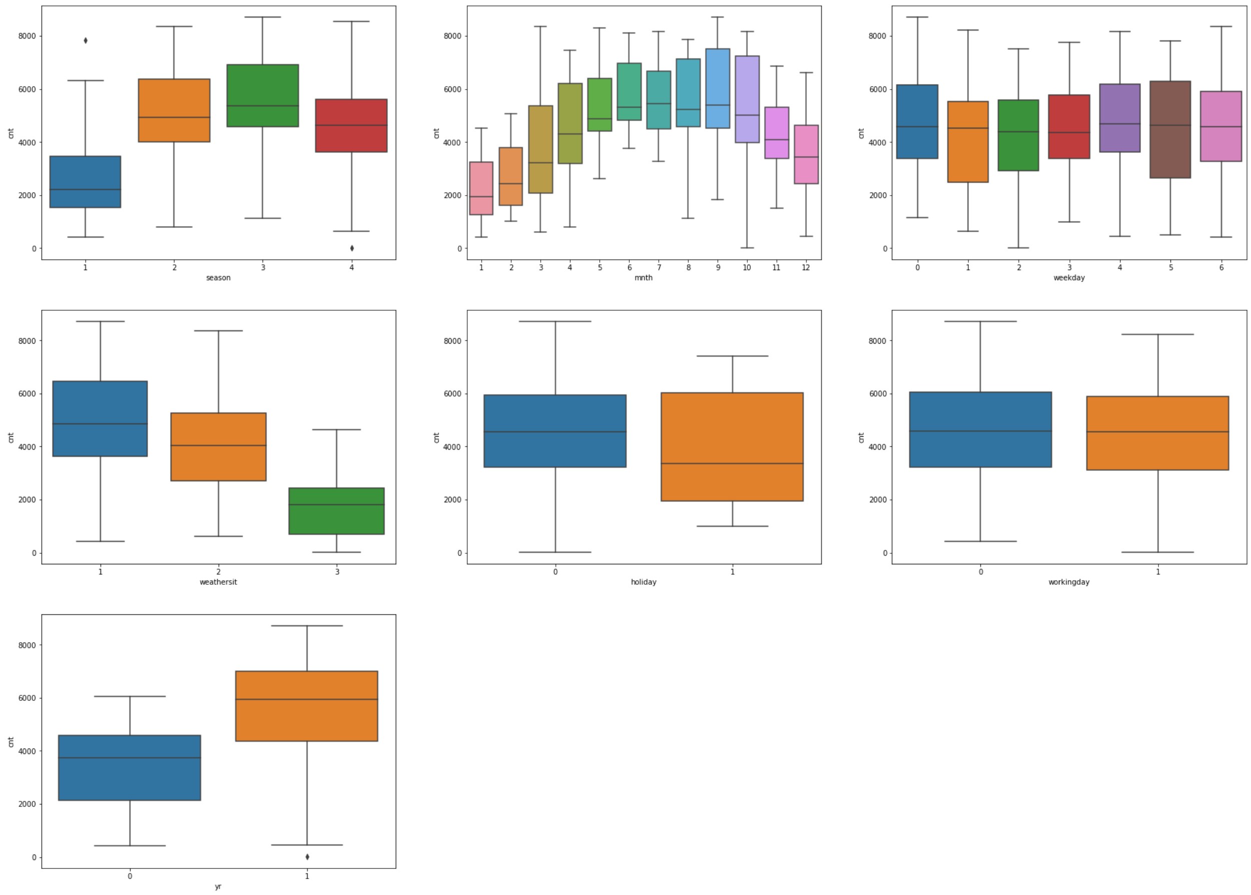
Solution for Subjective Questions

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

Ans:



From the boxplots of categorical variables we can observe that,

- In season value 3 i.e. ***fall*** has the highest cnt.
- In month there is a gradual increase till ***september***
- We can observe the decrease in cnt as weather becomes more harsh
- non holiday days tend to have more cnt and on holidays cnt decreases
- the year 2019 has higher values than previous yr so we can see a positive trend here

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

Ans:

Often when we create dummy variables we can explain the values in **(n-1)** ways , i.e. if there are n values it can be explained with 1 less than the total count.

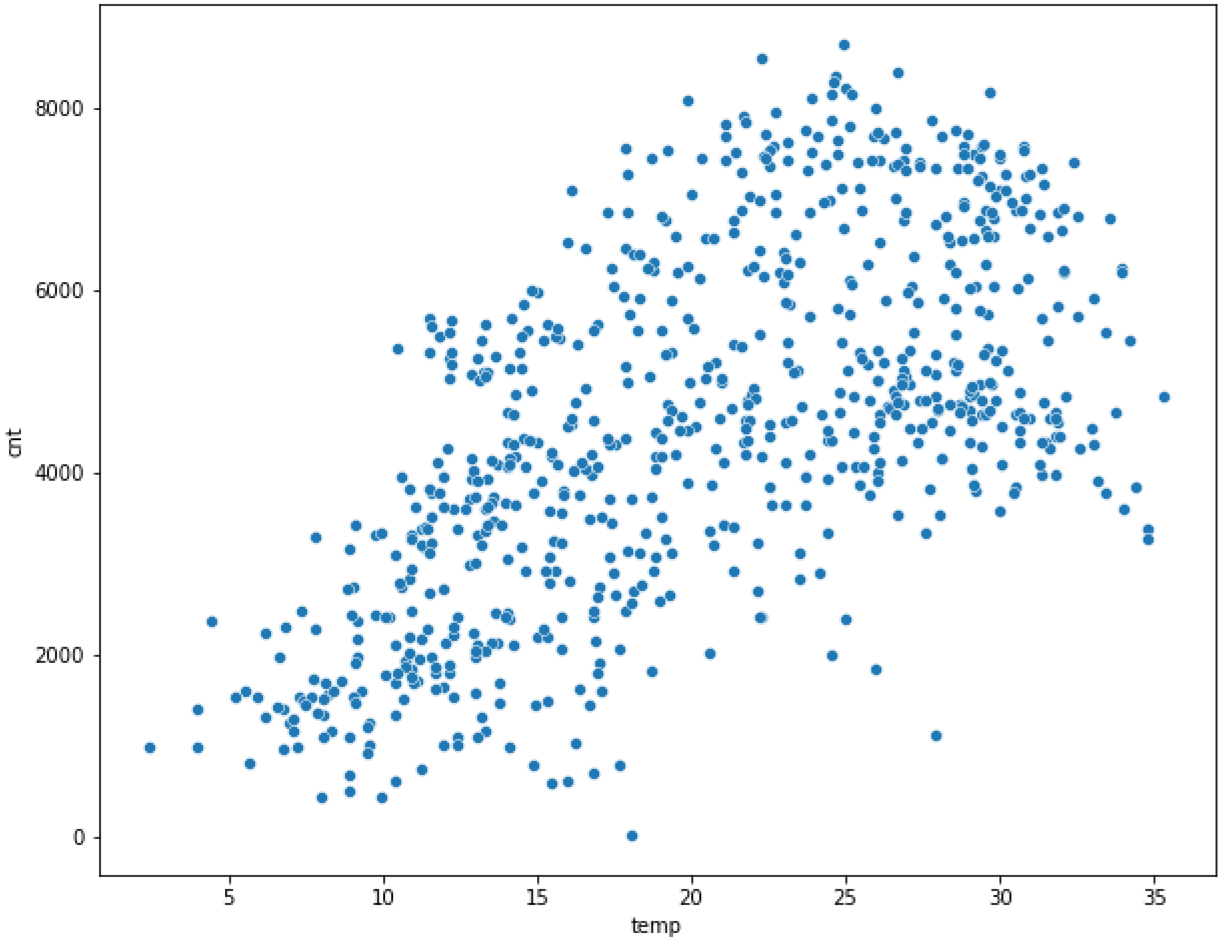
Consider an example of gender where male and female are the 2 values so when we create dummy variables it can be as follows,

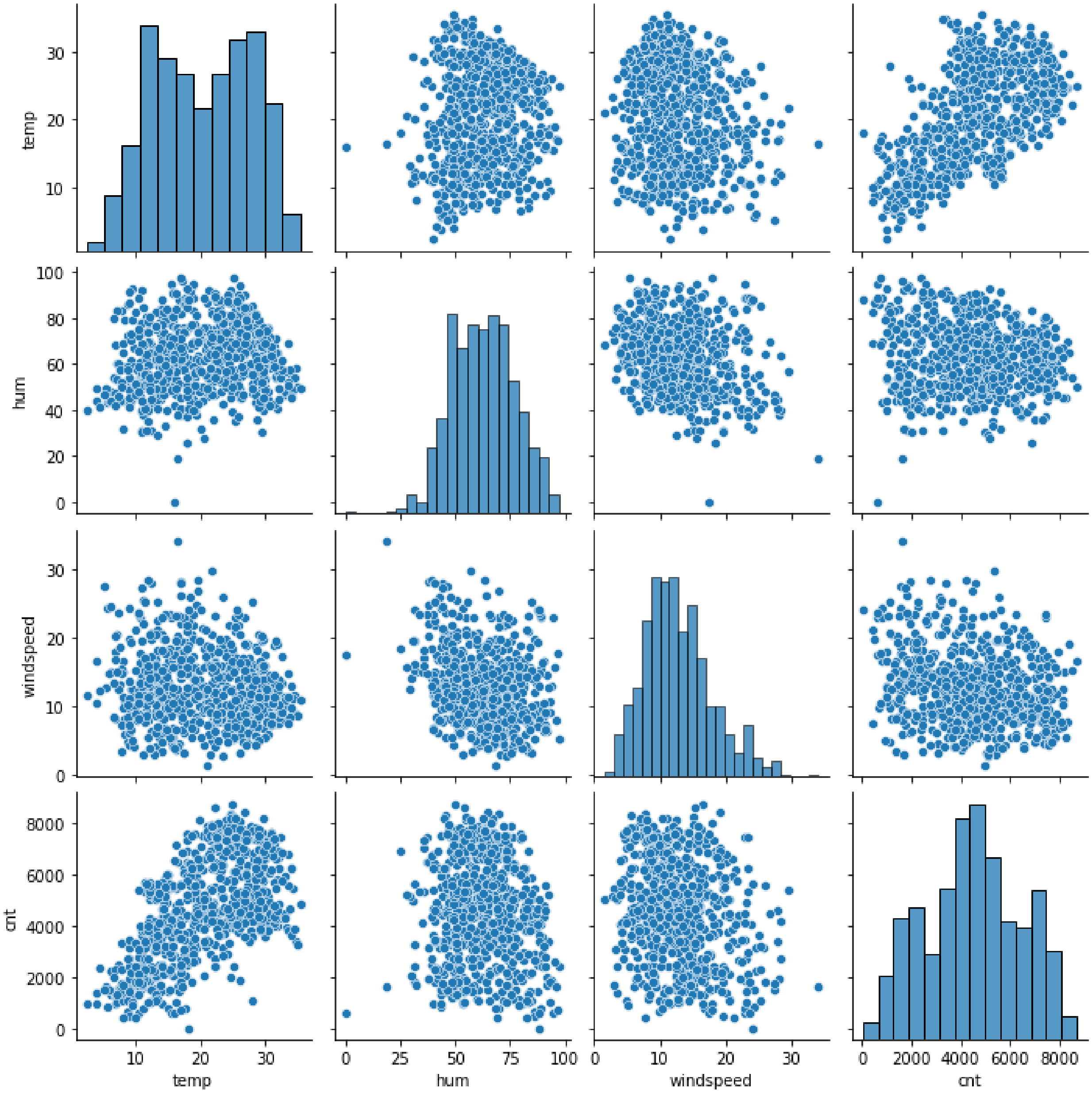
	D1	D2
Male	**0**	**1**
Female	**1**	**0**

So instead of creating two variables we create one where value 1 indicates a Male and 0 as Female.
Similarly if there are more than 2, in case of **n** variables we create **n-1** dummy variables to explain all the values.

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

Ans:





As we can observe from the above numerical variables pair-plots that temp has 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?

Ans:

We can have the assumptions on hypothesis test,

Consider the Null hypothesis as,
$H_{\circ}$ = all the coefficients = 0
H, = one of the coefficients !=0

Since we have our coefficients from the final model,

const	**0.248194**
yr	**0.232605**
temp	**0.376844**
season_spring	**-0.087401**
season_winter	**0.085330**
mnth_dec	**-0.079826**
mnth_feb	**-0.058284**
mnth_jan	**-0.080713**
mnth_nov	**-0.081932**
mnth_sept	**0.065159**
weathersit_bad	**-0.333090**
weathersit_moderate	**-0.072449**

So we reject the Null hypothesis as we have the coefficient values !=0

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

Ans:

As we can observe from the coefficients of different variable,

Mainly 3 predictors dominate the model,

- temp - which has a coefficient of 0.3768, i.e. when a unit increase in temperature will increase the count by 0.3768 units.
- yr - which has a coefficient of 0.2326, i.e. when a unit increase in the yr will increase the count by 0.2326 units.
- weathersit_bad - which has a coefficient of -0.333, i.e. when a unit increases in bad weather conditions, will decrease the count by -0.333 units.

Solutions of General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression algorithm is one of the types of Machine learning algorithms in the supervised category for continuous variables.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is a group of four data sets which are identical or similar when we just look at the statistic summary but when we actually plot these values they will be very different from the statistics and this can fool the model built on the same dataset.

It is always better to plot the values beforehand to visualize how actually the data is distributed.

3. What is Pearson's R?

Ans:

It is a type of correlation coefficient. Pearson's R measures the strength of the linear relationship between two variables whose value is always **between -1 and 1.** It is mainly used to compare two numerical variables. The full name is the Pearson Product Moment Correlation (PPMC)

This attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points are with the best fit line).

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

Where, x^ and y^ are mean of the series

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

Ans:

Scaling is a process of converting the data points to a uniform unit. When we consider different variables in a dataset which are always in different scale values, the coefficients obtained from these are also very different. All the variables should be at a comparable scale then the coefficients will also be comparable with each other i.e interpretebility. So as to avoid these issues we use the different scaling methods,

- Normalized scaling
- Standardized scaling

Normalized scaling :

In this the values are converted to a scale where every value is in between 0 and 1.

It can be given by the formula for value Xi,

Normalization = (Xi - Xmin) / (Xmax - Xmin)

Standardized scaling :

In this the values are converted to a scale where every value is in between -1 and 1.

Here the data is centered around 0 (mean) and the standard deviation is 1.

It can be given as,

Standardization = (Xi- Mu) / Sigma

Where, **Mu**= mean of the data

Sigma= standard deviation

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

Ans:

The formula to calculate the VIF,

VIF = 1 / (1-R^2)

So when the value of **R** is one VIF will be infinite. This means there is a high collinearity between the variables as one of the variables can be expressed by another or combination of other variables - a perfect collinearity. Usually we drop the variables which have a VIF >=5.

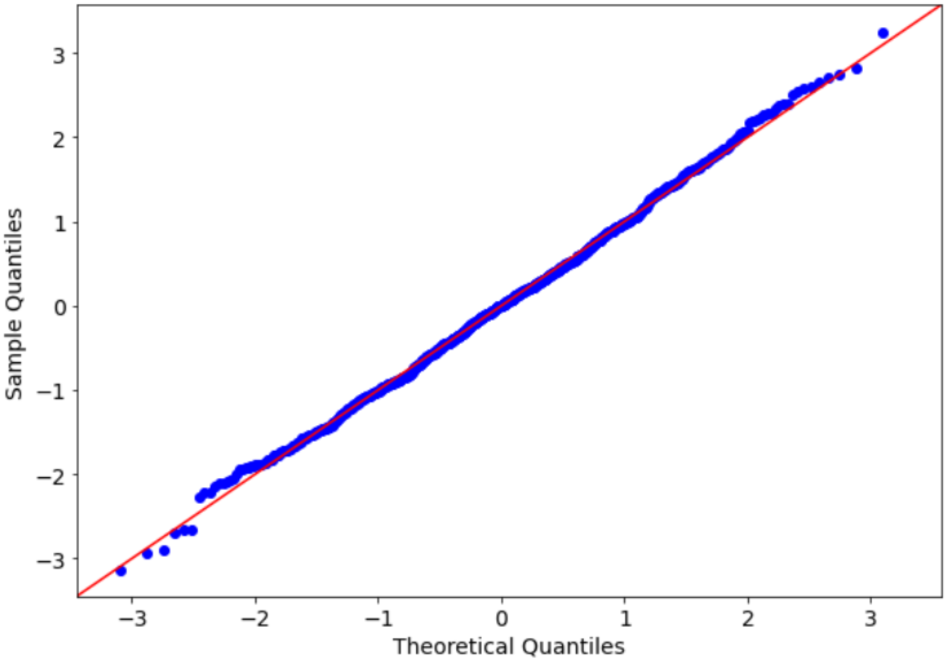
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans:

Q-Q plots are also known as Quantile-Quantile plots. They plot the quantiles of a Test or sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

In Linear regression we use Q-Q plot for the following,

- To check, If the test and theoretical data follows the same distribution
- To verify whether the residual errors give me the normal distribution as the assumption of Linear regression, we can check with the Q-Q plot.
- Skewness of the distribution.


Sample Quantiles
Theoretical Quantiles