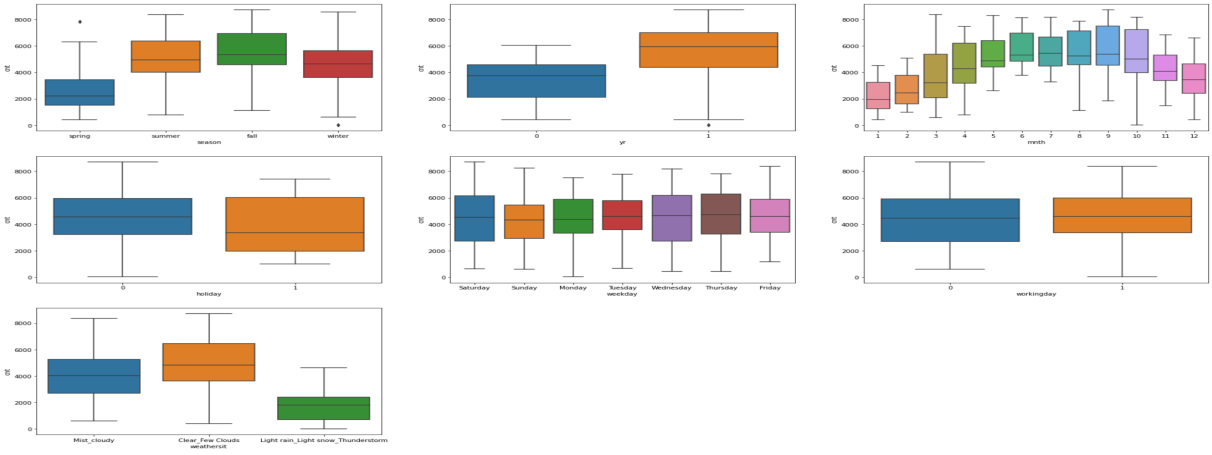
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



2019 has more bikes rented than previous year

The seasons summer and fall has more number of bikes rented

workingday column does not have much impact on target variable

April to November looks like busy time for renting

As expected clearer weather is optimal and thunderstorms not for bike rides

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

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

- `100` will correspond to `spring`
- `010` will correspond to `summer`
- `001` will correspond to `winter`
- `000` will correspond to `fall`

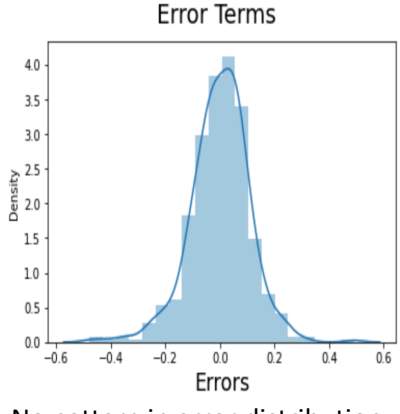
As seen from above example any categorical variable with n unique values can be explained using n-1 dummy classes . So drop_first helps in removal of extra column created during dummy variable creation.

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

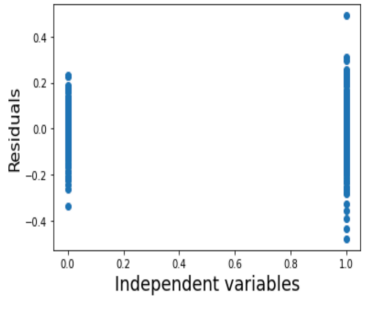
Columns “atemp” and “temp”

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

- Error distribution follows a normal distribution



- No pattern in error distribution



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

Light_rain_light_snow_thunderstorm, yr, spring

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

Scaling is conversion of data values to a similar or comparable range. scaling doesn't impact the model. If the values are in different range, It is important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

There are two common ways of rescaling:

1. Min-Max scaling x-min(x)/max-min
2. Standardisation x-mean/standard deviation given SD=1 mean=0

In Min-Max scaling the anamolies gets included min, max range where as it is not the case in standardization.

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

A large value of VIF indicates that there is a correlation between the variables. If VIF value is infinite , then there is perfect correlation between the variables.