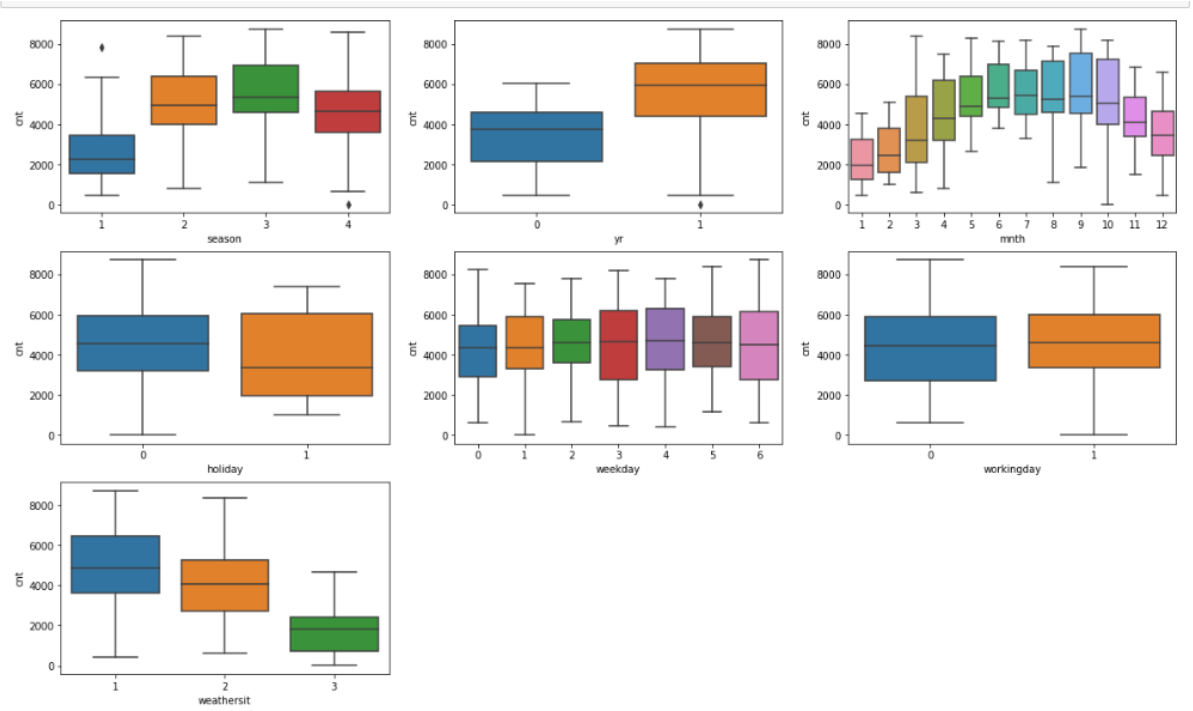
Subham Sahoo
Batch : C37

# Assignment-based Subjective Questions

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

When I plotted boxplot, we got the below plot




From the above plotting, it is observed that categorical variables like season, year, month and weathersit have significant impact on output variable i.e. output varies significantly with different levels.

But variables like holiday, weekday, workingday doesn't have much influence on output variable and output doesn't vary much with different levels.

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

To avoid dummy variable trap, we always have to drop one dummy variable while converting a categorical variable to a dummy set of variables.

If categorical variable has N levels, by default get_dummies method gives us N dummy variables. If we look at multi-collinearity, we can realise one variable is redundant i.e. when every other variable is zero, that will be one. When any other variable is 1, that is zero. Hence, we don't have to create N dummy variable, but N-1 dummy variables.

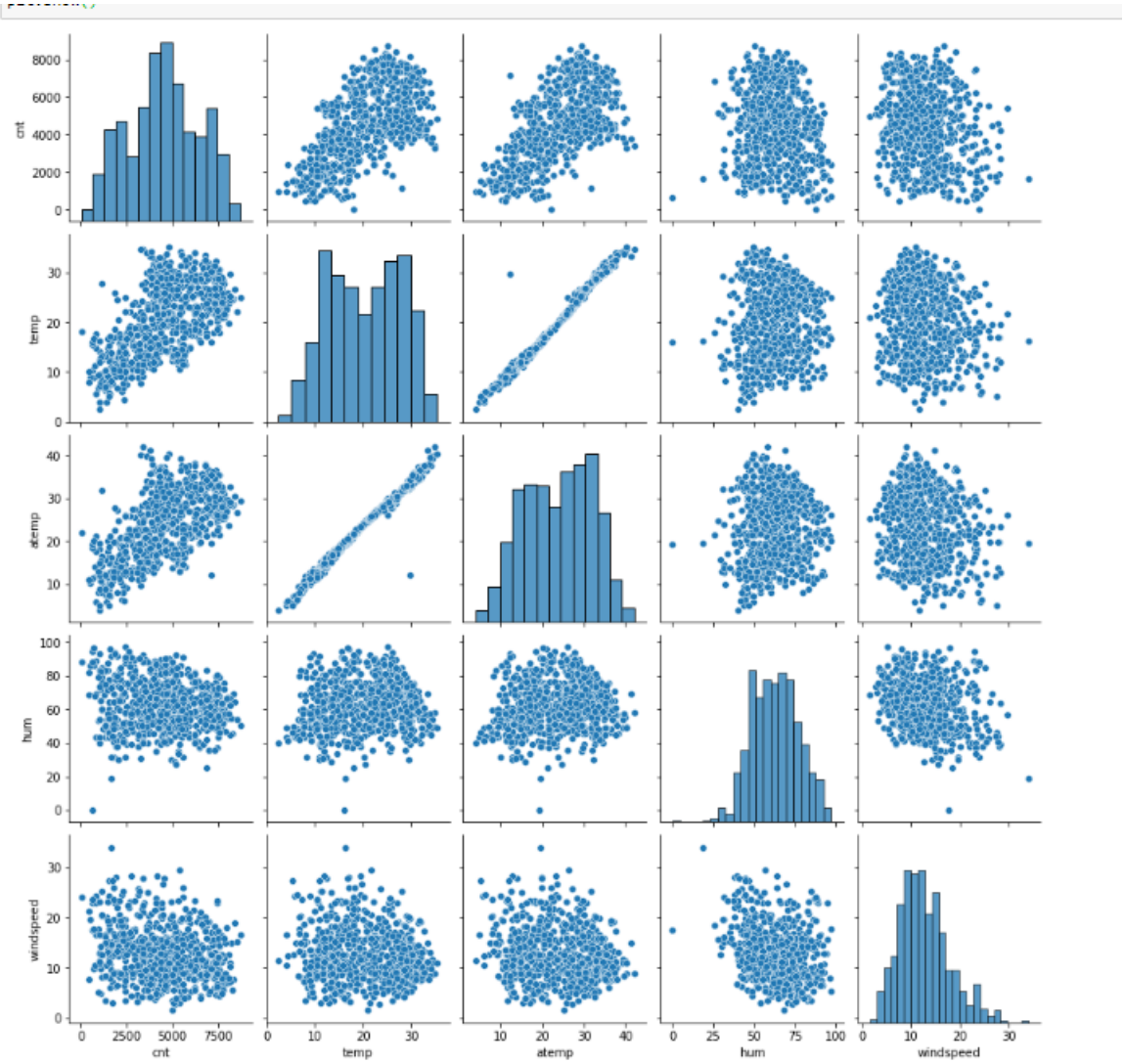
Example- gender can be male or female.

But we don't need 2 variables like is_male and is_female to store this information. Because when is_male is true, is_female automatically becomes zero. Hence we can get rid of one dummy variable let's say is_male and only use is_female as our dummy variable.

First of all, it makes input set independent of one another. And also p values won't affected by multi-collinearity.

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

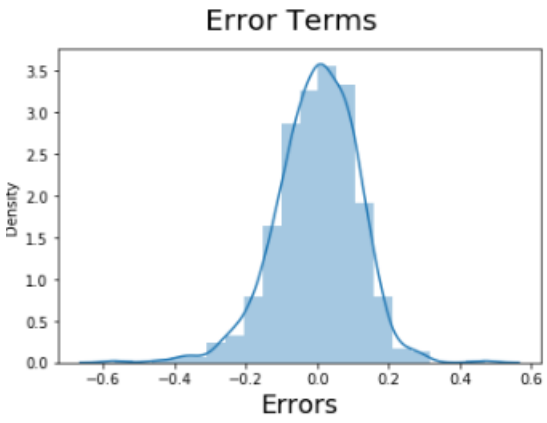
The plot is below



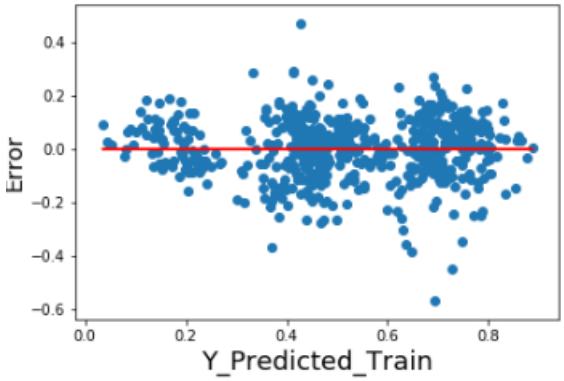

Looking at this pair plot, temp and atemp are having maximum correlation with target variable.

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

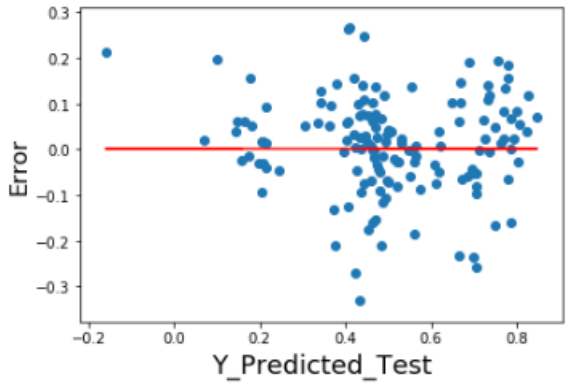
- Error term should have a zero mean
- Error term should follow a normal distribution
- Error term should have a constant variance
- There should not be any pattern of error term
- Plotting y_test vs y_pred

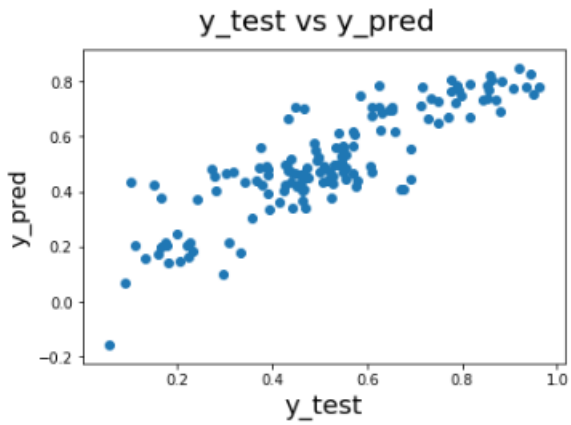



Error is centred at zero with normally distribured.




Error term is having almost constant variation and doesn't show any particular pattern




Y_pred should be linear and nearly equal to y_test.

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

Looking at my final model, year, weathersit (is_weathersit value is 3 (in my dataset I represent it as is_weather_worse)) and season (is_spring) are the three top features contributing towards explaining the demand of the shared bikes.

# General Subjective

**Questions 1**. Explain the linear regression algorithm in detail.

In Linear regression model, we try to predict the output as a linear combination of input predictor while minimizing the cost functions. Here we assume certain things to be satisfied.

- Linear dependency between output and each input variable
- Homoscedaticity. (variance is constant )
- Error term is normally distributed with mean zero
- No multi-collinearity between input variables

To train the model, we try to minimize the cost function. It can be done in 2 ways.

- By differentiation method
- By gradient descent method

To check how much a model is good, we look at R squared term which shows how much variance is explained by our model. Also, from F-statics , we verify if R squared value is actually correct or happened by chance.

We can say linear regression is of 2 types.

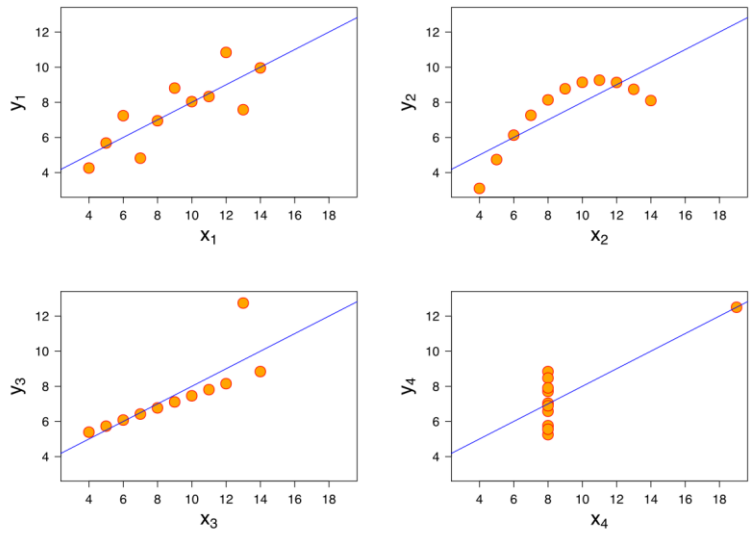
a. Simple linear regression - Model with only one independent Variable
b. Multiple linear regression - Model with multiple independent Variables

**Question 2.** Explain the Anscombe's quartet in detail.

Anscombe's quartet tells that sometimes statistical summary is not sufficient to infer the relationship. We have to visualize the distribution of data instead of relying on the data

Below given exaplme consists of 4 datasets that have same simple statistical properties like mean, SD, Correlation, , but are completely different from one another in distribution.

Every dataset consists of eleven (x,y) points. These were constructed by the statistician Francis Anscombe in 1973 to showcase that both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.




| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

All above data points have almost identical

1. X mean

2. Standard deviation of x
3. Y mean
4. Standard deviation of x
5. Correlation coefficient between X and y
6. Fitted Linear regression line

But all 4 have have totally different spread.

**Question 3.** What is Pearson's R?

It measures linear relationship strength between 2 variables. It is also called as correlation coefficient between 2 variables.

The expression can be represented as

$$r = \frac{n(\sum\ xy) - (\sum\ x)(\sum\ y)}{[n\sum\ x^2 - (\sum\ x)^2][n\sum\ y^2 - (\sum\ y)^2]}$$

X , Y are the two variables for which we want to count the correlation coefficient.

The value of this lies between -1 and 1.

If the value is positive, it means X and Y have a positive correlation which means if one increases other will increase and if one will decrease, other will decrease.

If the value is negative, , it means X and Y have a negative correlation which means if one increases other will decrease and vice versa.

If the value is 0, it means X and Y are not having any correlation. They are independent on each other.

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

- Scaling is a technique to bring a wide range of values into a particular range.

- In Linear Regression, scaling is performed for 2 purposes. First is it is better to compare the effect 2 input predictor on output variable. Second the gradient descent algorithm can converge to solution faster if inputs are in same scale.
- Without Scaling also, our model will have same goodness of prediction. Scaling just gives us a better way to look and feel the impact of each input variable on output.

There are 2 types of popular scaling technique used.

- Normalisation or Min-Max scaling
- Standardisation

In Normalisation, we map the value to between 0 and 1 using below relation

Normalized_value= $\frac{value-min}{\max - min}$

In standardisation, we make the mean of the dataset 0 and standard deviation as 1.

Standardized_value= $\frac{value-mean}{SD}$

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

VIF = $\frac{1}{1-Ri\text{^}2}$

Where Ri = good ness of fit of one input variable in terms of other input variable.

- If VIF=1, it tells zero correlation between one input variable and all other input variables in the model.
- If VIF=1 to 5, it tells some correlation between one input variable and all other input variables in the model. But it not very serious.
- If it is greater than 5 indicates severe correlation between one input variable and all other input variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

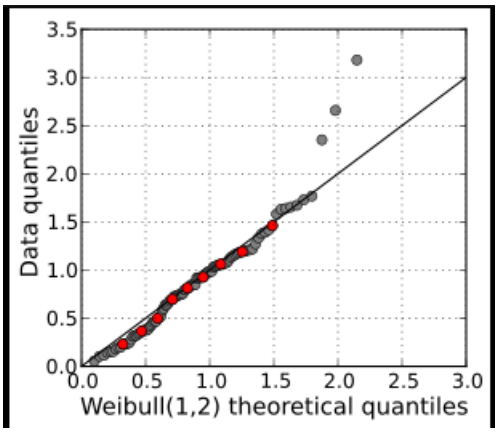
If Ri^2=1 i.e. (Ri= 1 or-1) the input variable is fully linearly dependent on all other variables, VIF will be infinite. We must drop that variable in our prediction.

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

This Q-Q plot is used to determine if 2 datasets are derived from the same population or not.

UseCase of Q-Q plot Linear Regression:
When we receive train and test dataset separately for a linear regression model, we can verify that both dataset have come from same population using this technique.

Example of QQ Plot:




While building machine learning model, we need to check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, we might have to check the distribution of the feature variables and consider transforming them into a normal shape.