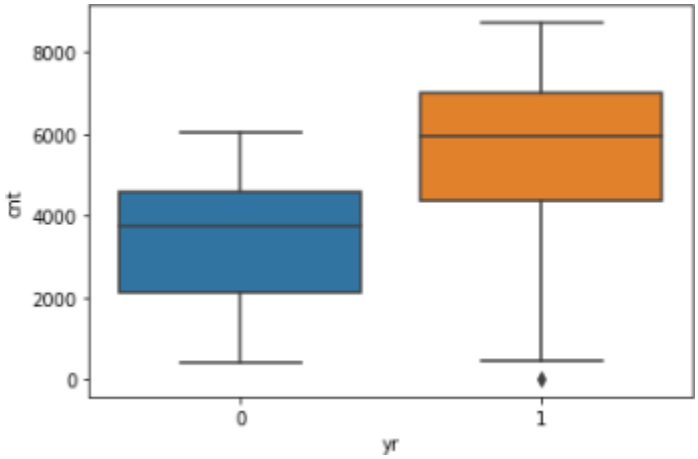
Prashant Mittal
Nov batch

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?

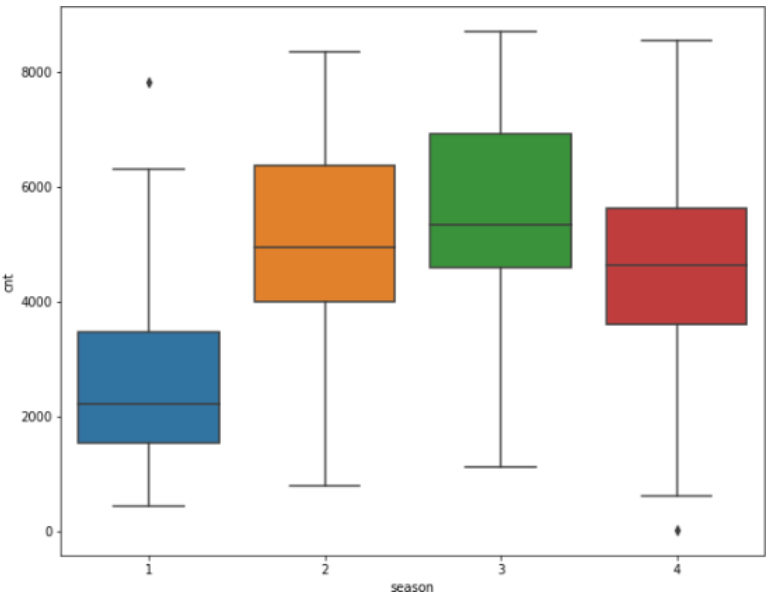
Ans –**1. Year**



0 is 2018 and 1 is 2019

2019 is having the outlier at the lower quintile range, which we can rewrite the count value at 10 percentile, so that the error reduces. But I am not removing it now. More number of counts of rental is in 2019. We can say that as the year is passing this kind of bike rental idea is spreading and more number of people was opting for this.

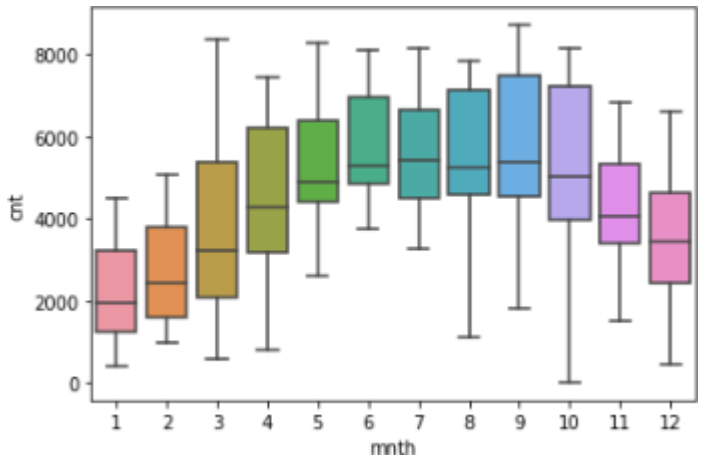
2. Season



1: spring, 2: summer, 3: fall, 4: winter.

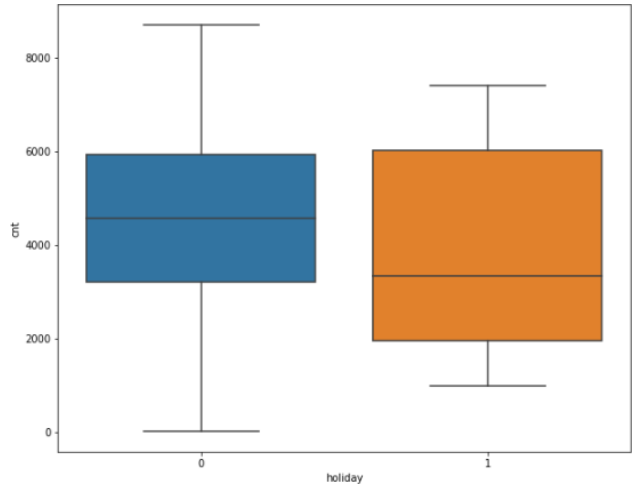
Spring and winter are having outlier on the upper and lower region respectively. Maximum number of rental was in the fall season, May be because of nice weather.

3. Month



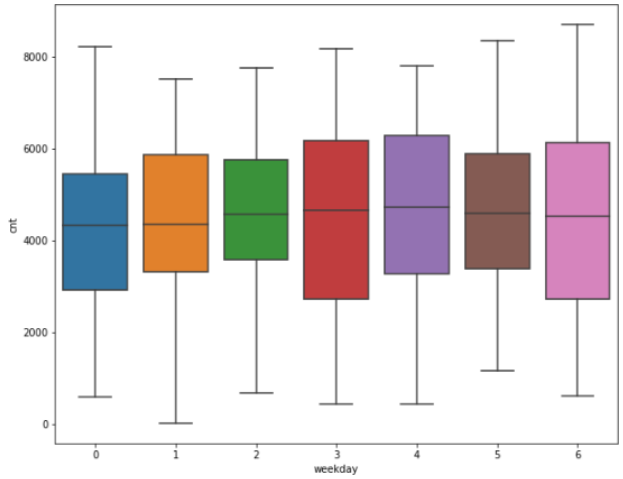
The number is according to the conventional way of numbering the months. Most bikes were rented after summer, in fall and autumn season.

4. Holiday



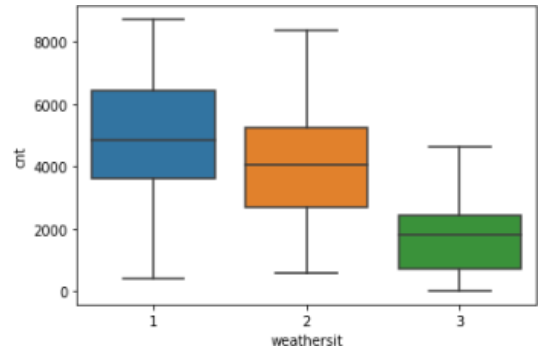
More number of bikes was rented during working day as the office and school going people might be taking it for there commuting. We can say it because we the median of a non holiday is more that the holiday

5. Weekday



Mostly it has the same count of bike rental in weekday, similar trend we have send in the above graph.

6. Weathersit



1: Clear, Few clouds, partly cloudy, partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Most bikes were rented in a clear day.

In my model the categorical variable which was used are clear, year, holiday, winter, light_snow.

1. R - Squared is coming out to be 0.803 and Adj. R-squared is coming out to be 0.80, which is pretty decent. The model is able to explain 80.3% variance in the data and Adj. R-squared value is also very near to it.

2. Probability (F-statistic) is coming out to be 1.83e-171 which is very low which means that the overall model coming out is significant.

3. Since maximum p value of the entire variable is very less (less than 0.05), this means that all the variable in this model is significant and we cannot drop any variable now.

4. VIF of all the variable is well below 5 so none of the variable is explained by other variable.

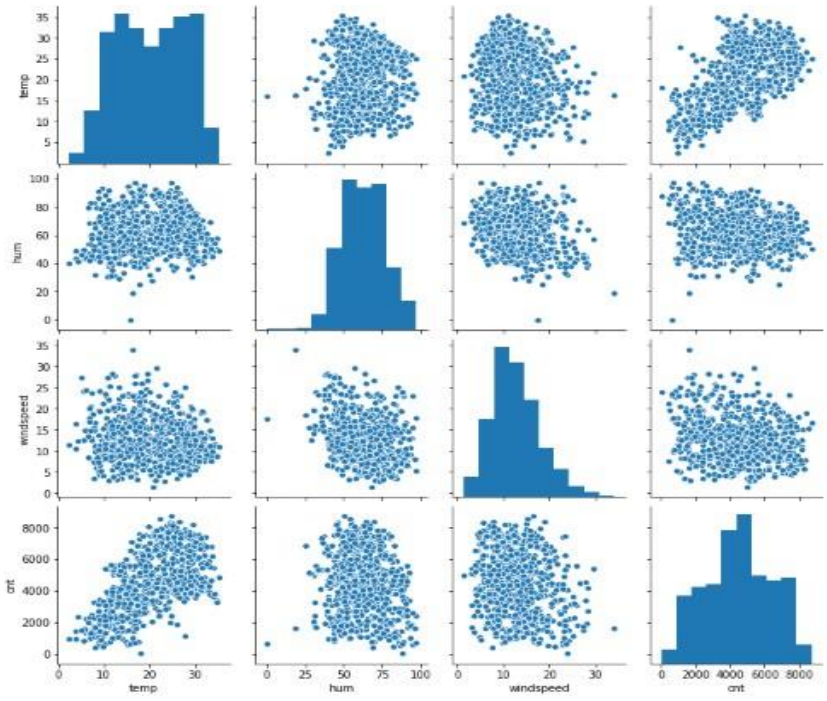
5. So this is our final model with the variable 'hum', 'clear', 'temp', 'wind speed', 'winter', 'light_snow', 'yr', 'holiday'.

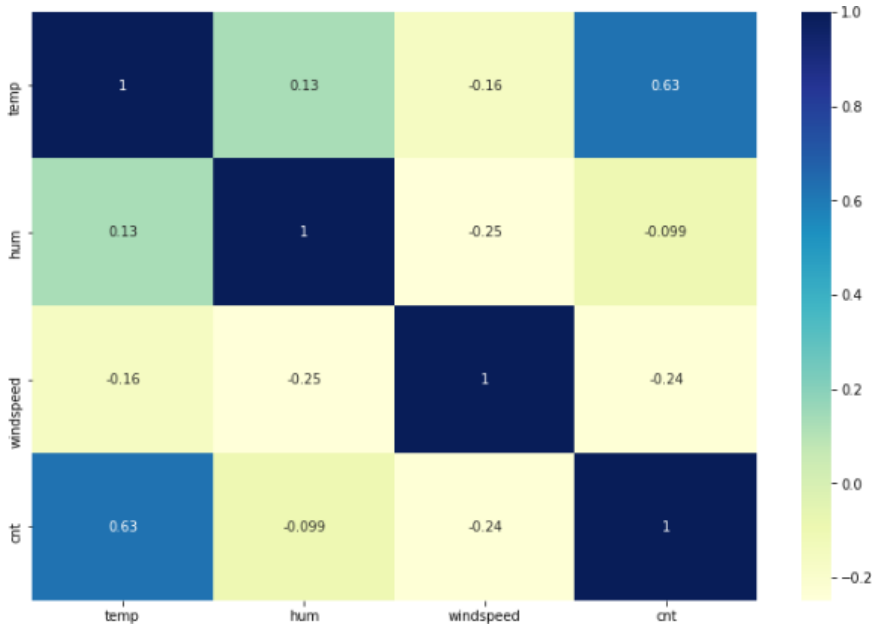
2. Why is it important to use drop_first=True during dummy variable creation?

Ans - We have dropped the first columns because we can define it with the rest of the columns as well. Suppose first column is when the dummy variable is showing value 1 and rest is showing the value is 0. We can drop it as we can find the first column if the other entire variable are showing 0 value. That's why it is not necessary to make our data complicated with more number of columns.

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

Ans –

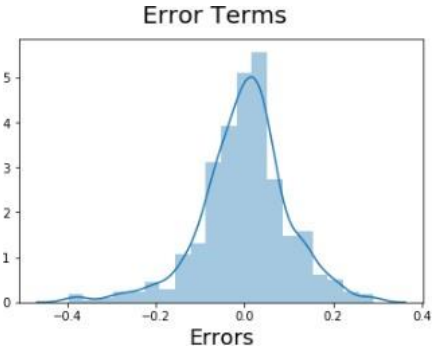




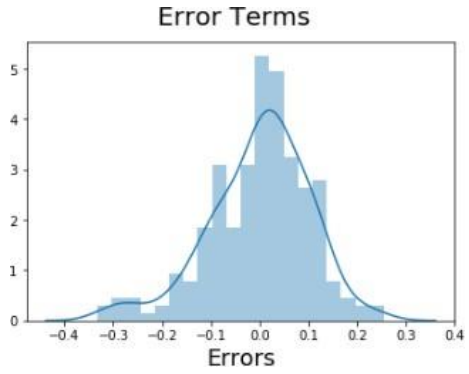
The maximum correlation is of temperature with cnt which is +0.63

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

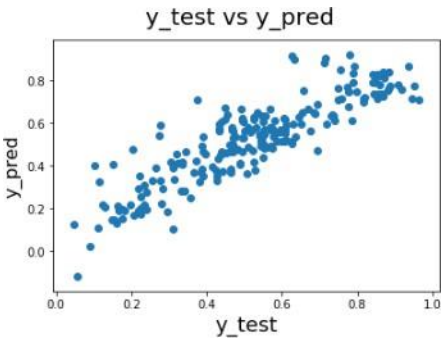
Ans – First we will calculate the error on the training data set and plot its distribution graph, if it is normal distribution and its mean is cantered at 0 with a fixed variance, then the model is good to go. We will check the same thing on the test set as well. If the above mention condition satisfied then the error is randomly distributed and having no pattern, this means that we are not missing anything in our model.



Since the error are normally distributed in the training data set with mean at zero, this mean that there is no pattern in the error, so we have not left anything in our model.



Since the error are normally distributed in the test data set with mean at zero, this mean that there is no pattern in the error, so we have not left anything in our model.



The scatter plot between y predicted and actual y value in the test data set shows a linear pattern and having a constant variance.

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

Ans – 1. Temperature with coefficient 0.6064, this means that if temperature is increased by 1 unit keeping other variable constant the dependent variable is going to increase by 0.6064.

2. Year with coefficient 0.2289, this means that if Year is increased by 1 unit keeping other variable constant the dependent variable is going to increase by 1.2289.

3. light_snow with coefficient -0.1902, this means that if light_snow is increased by 1 unit keeping other variable constant the dependent variable is going to decrease by 0.1902.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans – It is an algorithm based on learning the from the previous data and providing the interpolation result of the target variable on the unseen data.

First we have to select which is our target variable (dependent variable) and which is independent variable and then after plotting the straight line in between the points of these two variables we will get a straight line on the previous data recorded.

The Residual sum of error should be minimum that is the square of the difference between the actual y value and the y value which we are getting by our best fit line. The error terms should be randomly distributed, which means that there should not be any pattern in the error term. It means that our model is fine and we are not missing anything in our model.

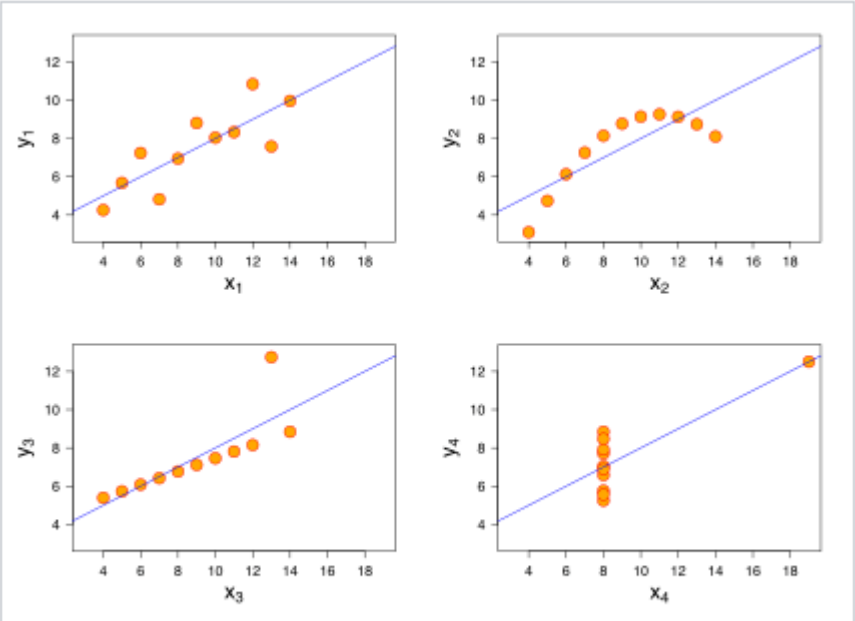
And in the scatter plot of the error term the variance should be constant to be called as a good model.

There are few things that we should check in our model to say that our model is good.

1. R2 – It will tell how well our model is capturing the variance, it should be as high as possible, max value is 1. This will always increase or remain same when you add extra variable to your model.
2. Adjusted R2 – This will penalised on adding more number of variable to our model. It depends on the number of variable you are taking to model and the size of the data. This can decrease on increasing the extra variable to our model.
3. P(F-Statistic) – This will tell that our whole model is capturing the data correctly or not. This value should be very very less, this means that we are rejecting the null hypotheses of the coefficient of variable is equal to zero. This ultimately means that with less value(less that 0.05) of P(F-Statistic) our model is significant.
4. P value – This will tell that our variable selected is significant or not. If P value is less that 0.05, this means that we are rejecting the null hypotheses of coefficient of that variable equals to be zero and saying that the variable is significant to our model and we cannot drop it.
5. VIF – Variance inflation factor value tell us that is there any variable which can be explained by one or combination of other variable, this is called multi co linearity. if the value of VIF is greater than 10 then it the variable id highly collinear with other variable and we should drop it from our model, and vice-versa if VIF is less than 5.

2. Explain the Anscombe's quartet in detail.

Ans –



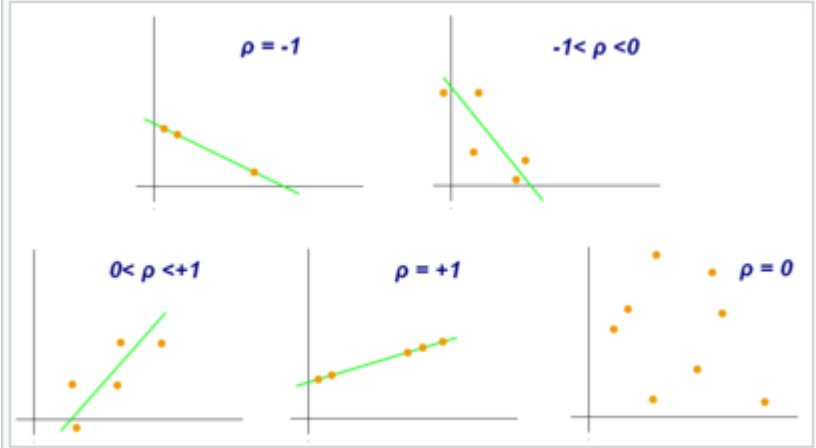
All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

The Anscombe's quartet comprises of 4 graphs which shows the same straight fit line through the data points but all the data points are completely different

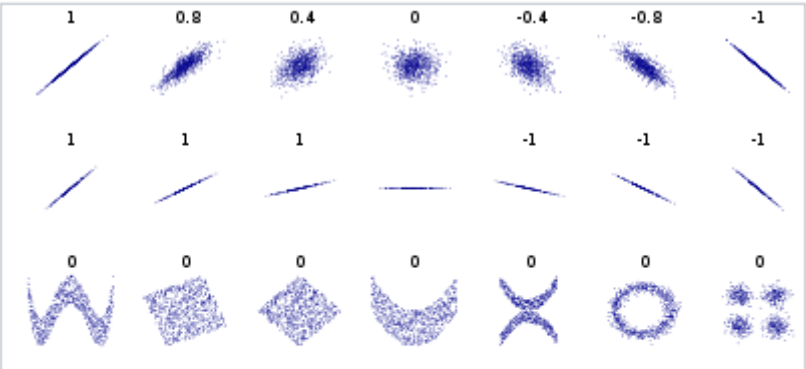
- The 1st plot (top left) is showing simple linear relationship between two variables with randomly distributed points showing a straight fit line.
- 2nd graph (top right) is not normally distributed, the data points are not normally distributed and straight line fit is not the best solution here. So we have to look for another technique to fit in these points.
- 3rd plot (bottom left), the distribution is linear, but because of one outlier the error term will be high. If this outlier would not be there then this would be a great fit straight line the data points.
- 4th (bottom right) shows that maximum number of points is gathered at one point and because of one outlier we are getting a relationship. These should be avoided and other regression method should be looked upon.

3. What is Pearson's R?

Ans – Pearson correlation coefficient show the linear correlation between two variables, it can be plotted by using pair plot in python. If the value of correlation is+1, this means that both the variable highly related to each other, if one increase other also increases by the same magnitude. If the value is zero this means that there is no relation between the two variables. And if the value is -1, this means that both the variable highly inversely related to each other, if one increases other decreases by the same magnitude. These can be shown by the graph below.



Examples of scatter diagrams with different values of correlation coefficient (ρ)



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

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

Ans - Scaling means to convert all the variables of the data set in the same range so that we can interpret the coefficient of the variable very well. There can be categorical variable in our data set which will have the value in between 0 and 1 and we can also have the data which is in the order million, so we need to scale down or scale up(not recommended for the categorical variable)to match the same range of all the data set.

Normalising scaling is also known as minmax scaling, in this all the variable of the data set is converted between 0 and 1. This can e found out by the formula.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization technique means when the values are cantered around mean with the deviation equals to unit standard deviation. In the mean of the value will be zero and unit sigma. This can be calculated by the formula.

$$X' = \frac{X - \mu}{\sigma}$$

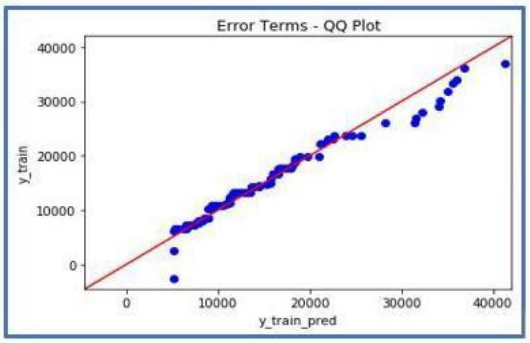
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans – VIF shows the multi co linearity between the variable. Ideally it should be less than 5, but sometimes we observe that the VIF is showing infinite value which means that the variable can be exactly explained by the combination of other variable, means this variable is 100% correlated with the combination of other variable, so in this case we need to drop that variable.

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

Ans – Quantile-Quantile plot help us to determine whether the set of data's are from the same distribution of population (normal distribution, exponential distribution, uniform distribution, etc). It helps us to find the location, scale, behaviour of the distribution, tail behaviour etc. There could be 3 cases

1. Similar Distribution – If all the points lie on or near the straight line at 45 degree, then both the data set have same distribution.
2. Y values < x values – It will happen when the y quintile values are less than the x quintile values.



3. Y values > x values – It will happen when the y quintile values are more than the x quintile values.