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?

There are total 6 categorical variables in our dataset.

Inference about effect of categorical variables on the dependent variable “cnt” are:

- **season** : category 3 i.e.,fall have highest number of bike hiring.
- **month** : month may, june, july, aug, sept have higher number of booking i.e., greater than 4000.
- **holiday** : Majority of booking are done when there is no holiday. Hence, this column can be bias towards no holiday and cannot be use for prediction.
- **weekday** : All days are having close trend.
- **workingday** : Majority of the bike booking were happening in ‘workingday’ with a median of close to 5000.
- **weathersit** : Category 1 have highest no of bike booking, category 2 have second highest no of booking while category 3 have lowest.

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

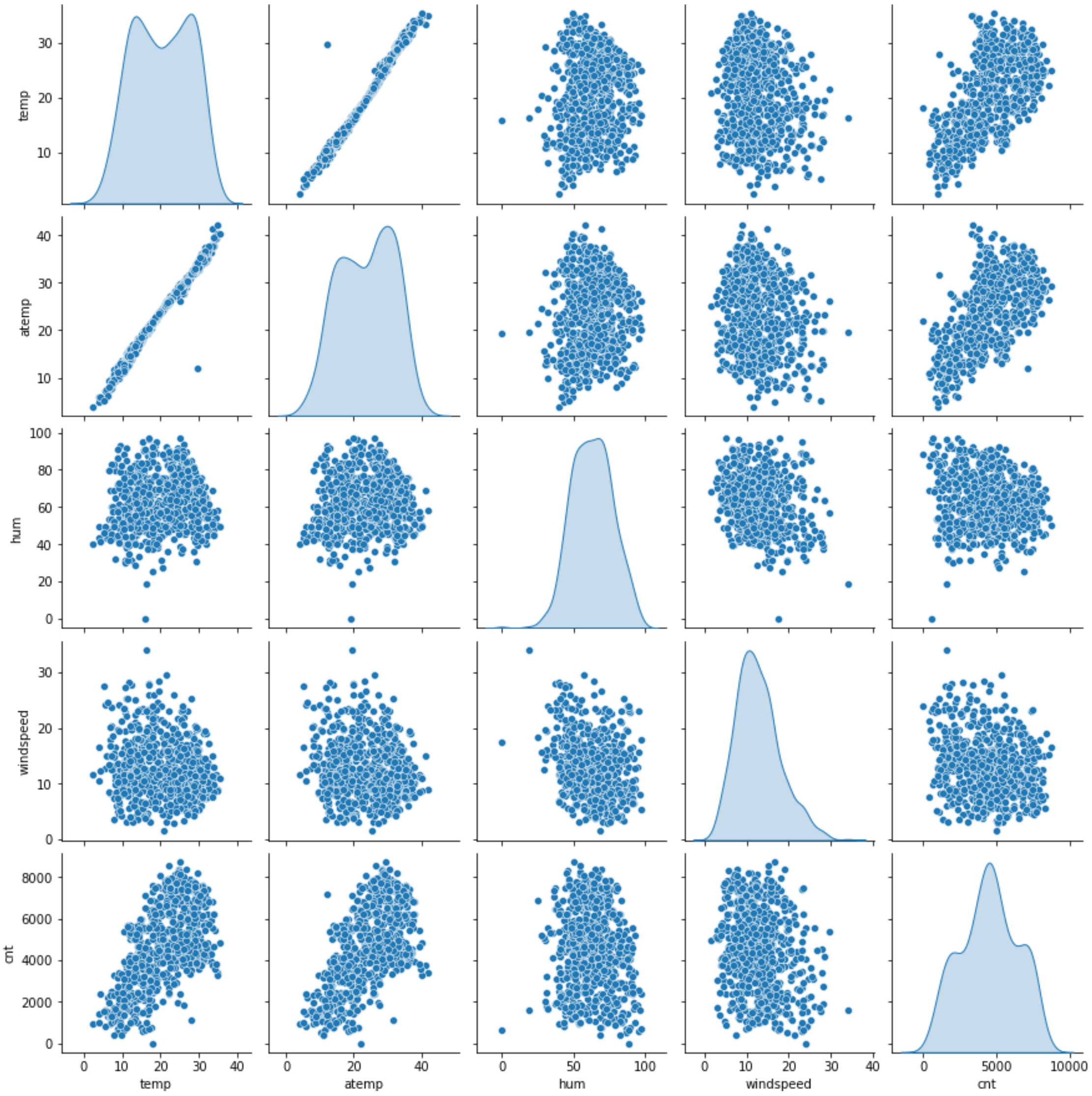
By using **drop_first=True,** it will reduce one column that will get create while creating dummy variables and due to reduction in one column the correlation created between each variable will get reduce.

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

From the pair-plot, we can conclude that **‘temp’** and **‘atemp’** features have highest positive correlation with target variable ‘cnt’

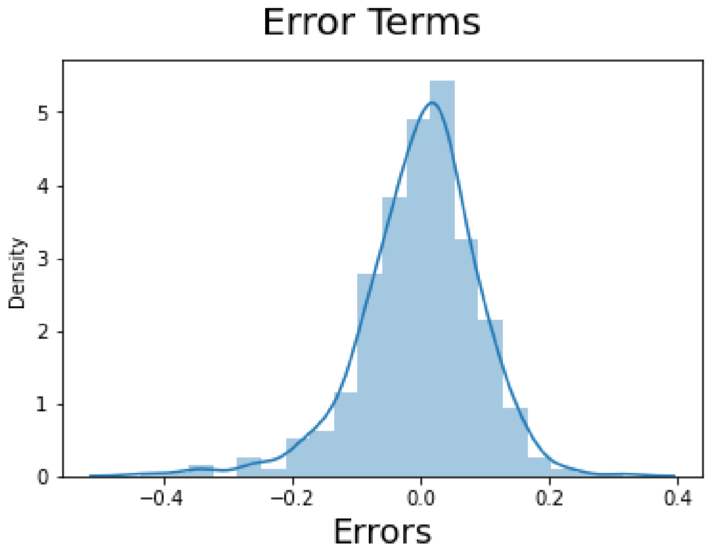
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

By plotting the pairplot, we can know the relationship between variables with target variable whether it is linear or not. Hence, the assumption that “**There is a linear relationship between X and Y**” can be validate.



In our model, temp and atemp has linear relation with target and thus, this assumption is valid.

By Residual Analysis we can know whether the error terms follows the normaldistribution curve or not. Hence, assumption that “**Error terms are normally distributed with mean zero” can be validate.**



In our model, the residual analysis plot is shown above which follows normaldistribution. Thus, this assumption is valid.

By finding the VIF value for each variable we can know whether multicollinearity is present between variables or not. We can say that If VIF<5, there is no multicollinearity present and if VIF>5, multicollinearity between variables is present. Hence, the assumption that "**No Multicollinearity between the predictor variables"** can be validate.

In our model, the VIF value of all features is below 5. Thus this assumption is valid.

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

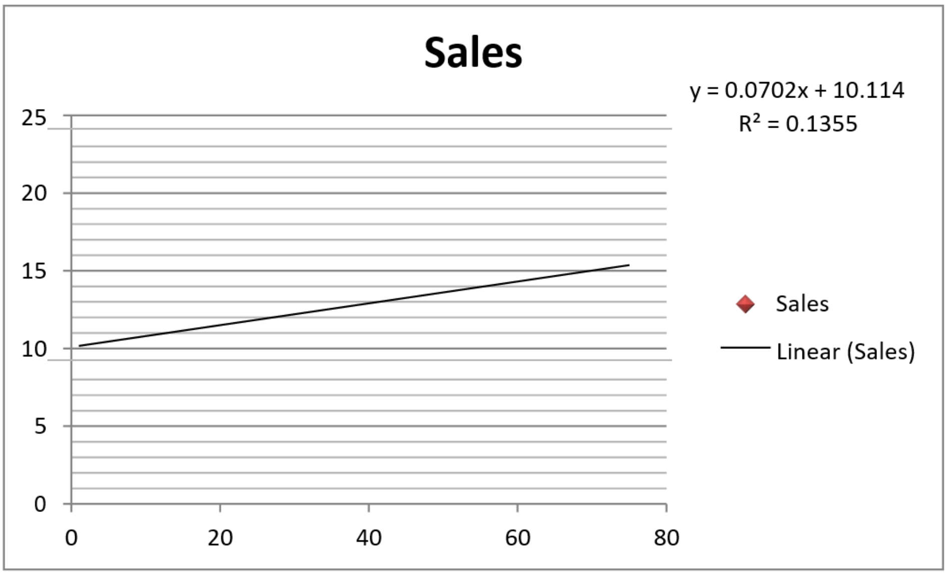
Top 3 features contributing significantly towards explaining the demand of the shared bikesare:

- **Temp** with 0.5436 as coefficient value
- **weathersit_3** with -0.2936 as coefficient value
- **yr** with 0.2333 as co coefficient value

Assignment-based General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

The line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below:

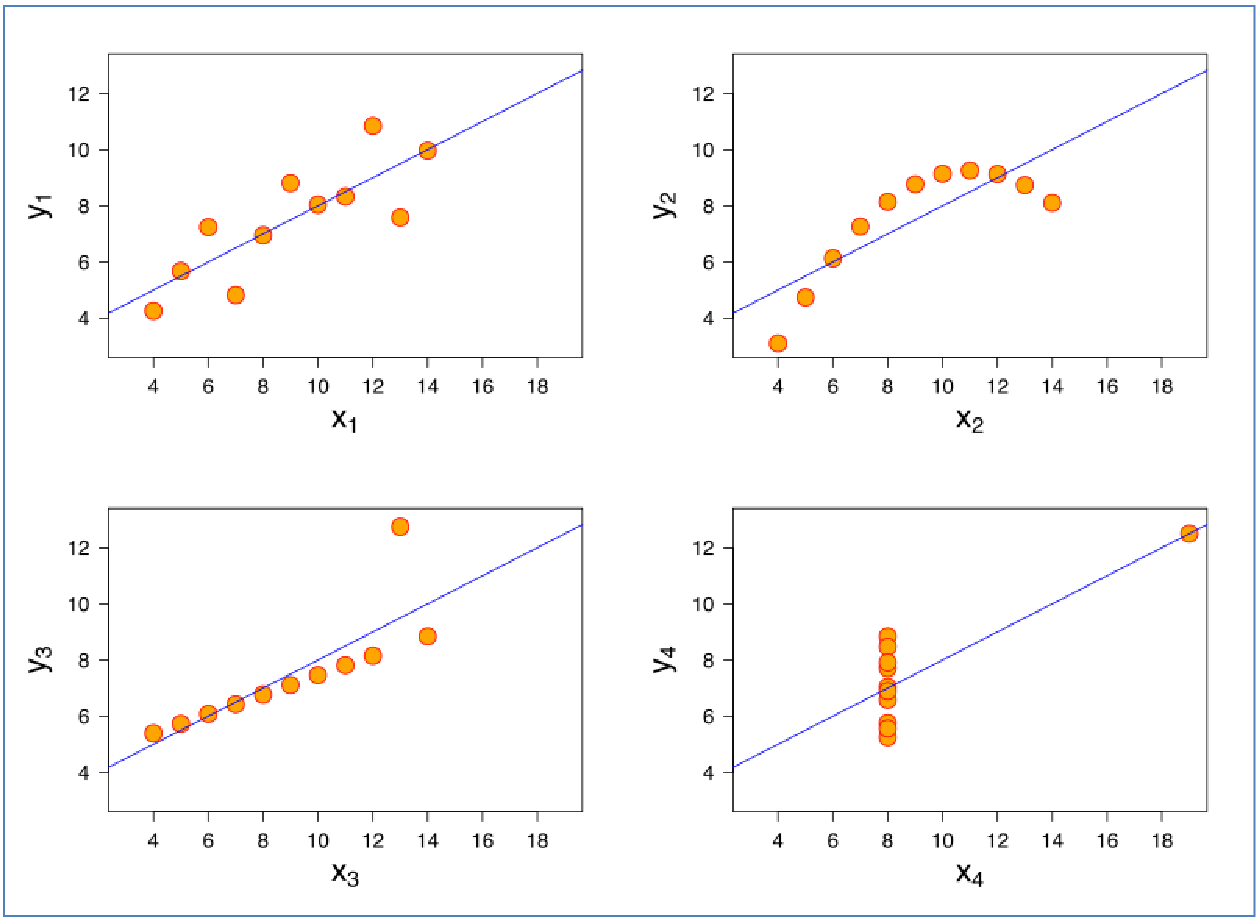
Y = m*x + c

Where, **m** = Slope of the line and **c** = Intercept.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed.

The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize completely**,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

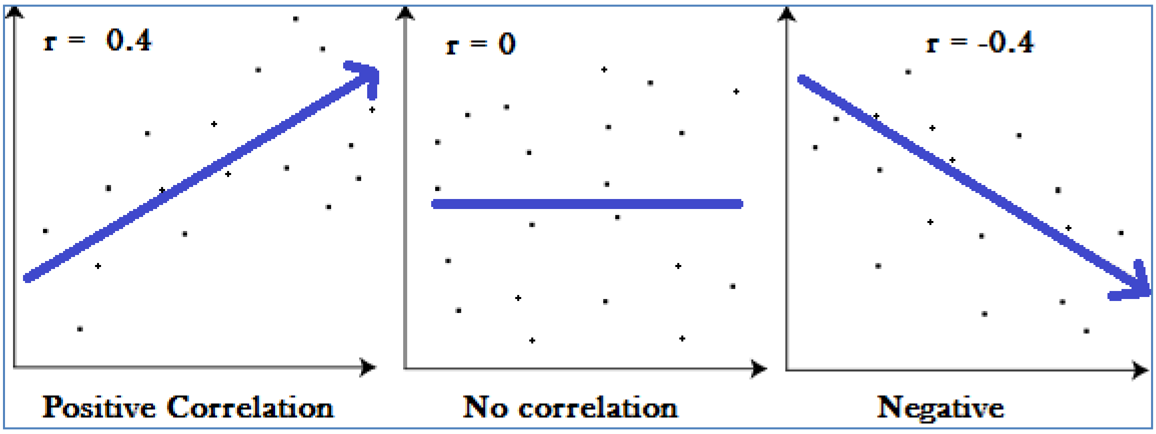


- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by anoutlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis.

3. What is Pearson’s R?

Correlation coefficients are used to measure how strong a relationship is betweentwo variables. There are several types of correlation coefficient, but the most popular is Pearson’s. Pearson’s correlation (also called Pearson’s *R*) is a correlation coefficient commonly used in linear regression. If you’re starting out in statistics, you’ll probablylearn about Pearson’s *R* first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson’s.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

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

Scaling: Feature Scaling is a technique to standardize the independent features presentin the data in a fixed range.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudesand thus, tackle this issue.

Difference between Normalization and Standardization:

- **Min-Max Normalization:** This technique re-scales a feature or observation value withdistribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that ithas distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

For example, we would fit the following models to estimate the coefficient of determination R1and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗_1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

X_2=C+ α_1 X_1+α_3 X_3+⋯

〖VIF〗 _2=1/(1-R_2^2)

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Ans: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populationswith same distributions.

Few advantages:

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and thepresence of outliers can all be detected from this plot.

If two data sets —

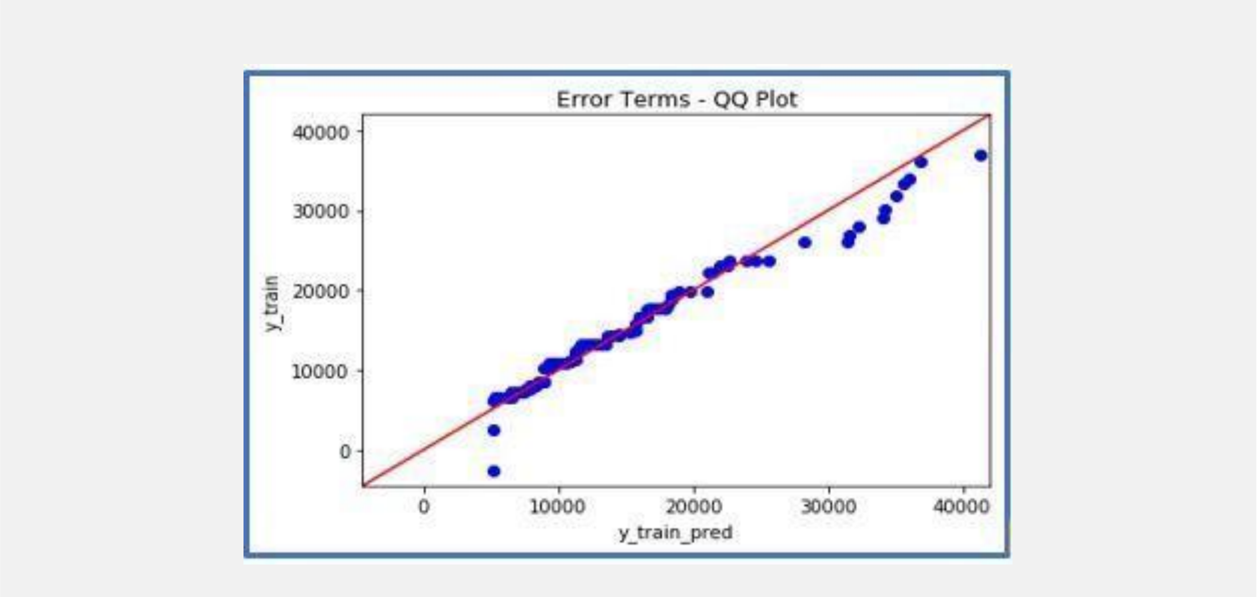
- Come from populations with a common distribution.
- Have common location and scale.
- Have similar distributional shapes.
- Have similar tail behavior.

Below are the possible regressions for two data sets:

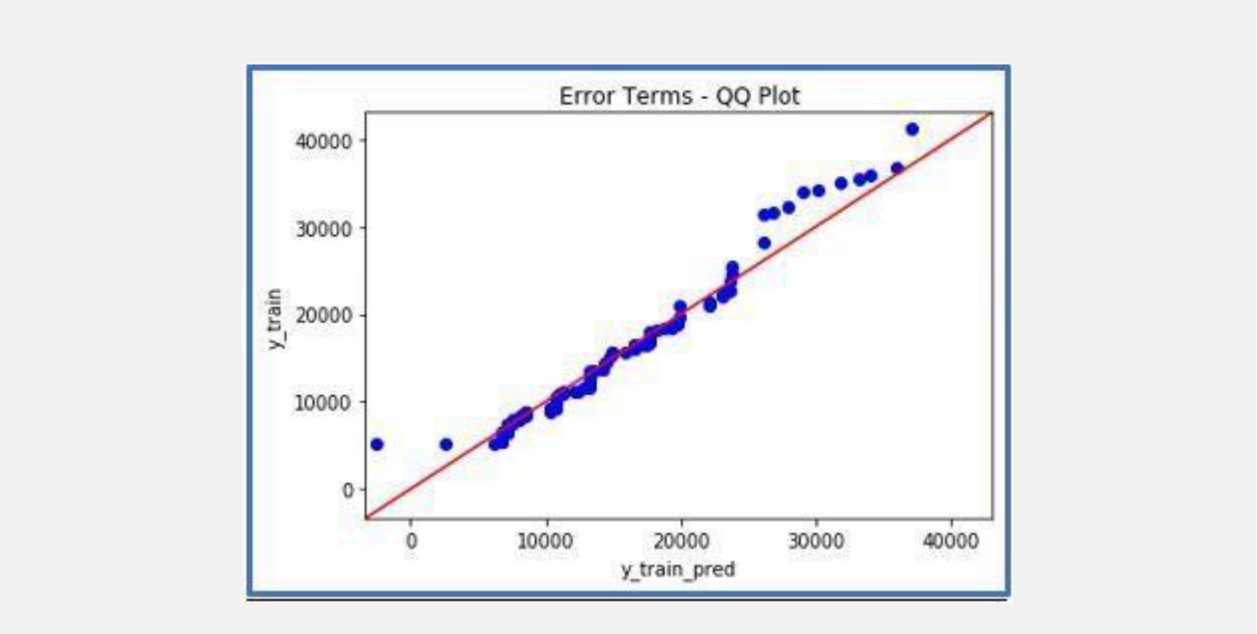
a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45degree from

x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of45 degree from x –axis.