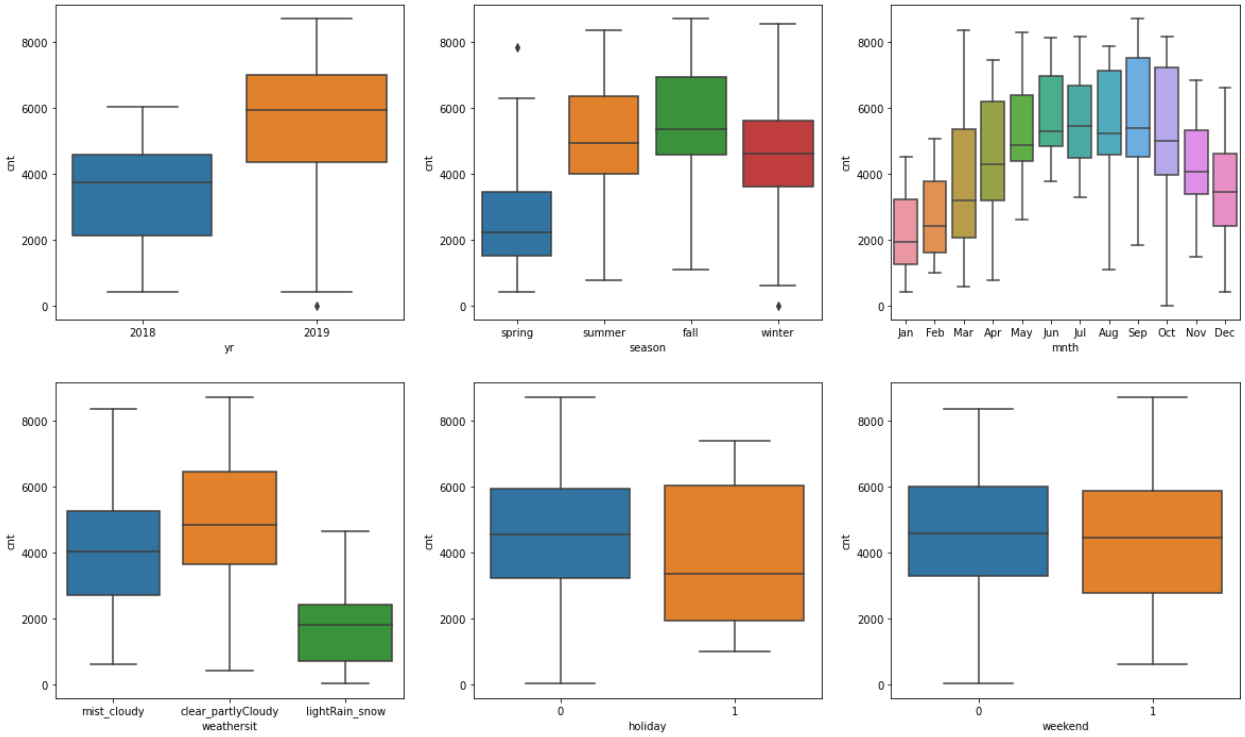
Assignment-based Subjective Questions

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

From the boxplots below, we can see that the demand for bikes is more in the summer and the fall. Seeing at a more granular level, the months April - October sees high demand. Also, the variability of dependent variable is more in the months of March, April, September and October. Looking at the weather situation, bikes are rented more on days with clear sky or partly cloudy with mist. We can also see how the demand increases from 2018 to 2019. This shows how the company has developed in 1 year.
Holiday and weekend variable don't have much significance on cnt.



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

Dummy variables are created out of categorical variables to separate out each of the features so that it becomes useful to correlate in the model. drop_first=True is used to remove redundancy in the data. As per thumb rule, if there are n levels of categories, we should use n-1 dummies for that variable. As an example, if the gender column contains male and female, we should only use a single dummy variable either male or female (but not both). Using both of them might affect the model adversely.

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

`wind_chill` has a strong negative correlation (-0.66) and **`temp`** and **`atemp`** has strong positive correlation (0.64 and 0.65 respectively).

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

a. Linear relationship between dependent and independent variables.

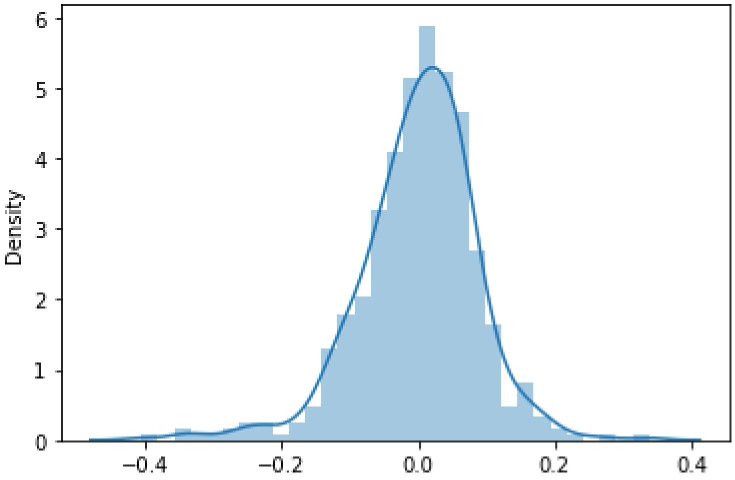
The final model which we have arrived at is:

cnt = 0.4878 + 0.2276$*yr$ − 0.0836$*holiday$ + 0.4119$*temp$ − 0.2098$*hum$ − 0.1853$*wind_speed$ − 0.2198$*heat_index$ − 0.0894 $*spring$ + 0.0514$*winter$ − 0.2677$*lightRain_snow$ − 0.0588$*mist_cloudy$ + 0.0309$*Aug$ + 0.0349$*May$ + 0.0828$*Sep$

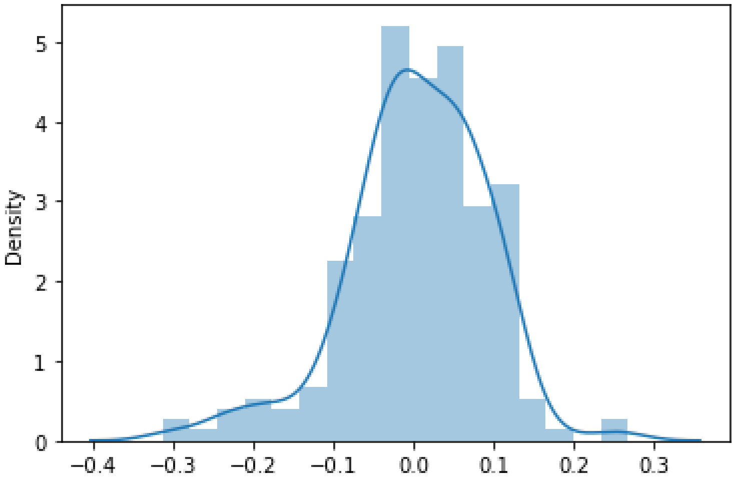
The pairplot and correlation matrix shows the relationship between these variables.

b. Residuals should have normal distribution: Upon plotting the residual data for the train and test data set, we get a normal distribution with mean=0.

Train dataset

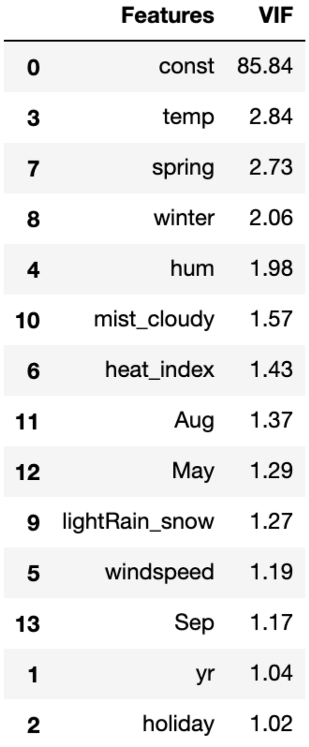


Test dataset



c. The independent variables are independent of each other, i.e., there is no multicollinearity in the data

This is evident from the VIF values we got for all the variables in the final model.



	Features	VIF
0	const	85.84
3	temp	2.84
7	spring	2.73
8	winter	2.06
4	hum	1.98
10	mist_cloudy	1.57
6	heat_index	1.43
11	Aug	1.37
12	May	1.29
9	lightRain_snow	1.27
5	windspeed	1.19
13	Sep	1.17
1	yr	1.04
2	holiday	1.02

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

Revisiting the same linear relationship between dependent and independent variables,

$$cnt = 0.4878 + 0.2276 \ast yr - 0.0836 \ast holiday + 0.4119 \ast temp - 0.2098 \ast hum - 0.1853 \ast wind\_speed - 0.2198 \ast heat\_index - 0.0894 \ast spring + 0.0514 \ast winter - 0.2677 \ast lightRain\_snow - 0.0588 \ast mist\_cloudy + 0.0309 \ast Aug + 0.0349 \ast May + 0.0828 \ast Sep$$

We can observe that temp has a high coefficient value of 0.4119. Additionally, weather plays an important factor in the demand for shared bikes. Demand is reduced on the days with light rain or snow or the days having comparatively higher heat_index.

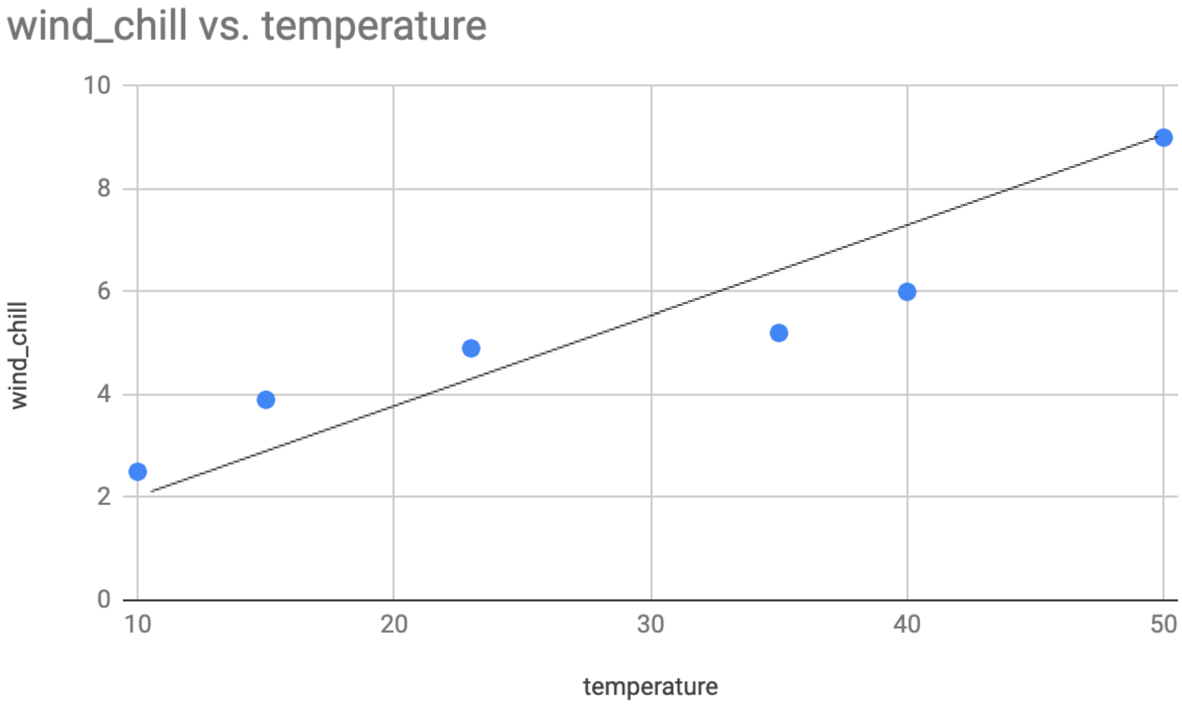
Top 3 variables affecting the demand includes: **temp**, **lightRain_snow**, **heat_index**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm which finds the linear relationship between the dependent and independent variables. In other words, linear regression helps to find the best fit line when deriving the relationship between dependent and independent variables. Dependent variable is the target variable which the business wants to derive based on the model while independent variables are the features with the help of which the business objective is fulfilled. For example, if the main objective is to predict a model which can identify the factors affecting performance of the cricket player in IPL. Performance of the player can be termed as target variable while individual factors like, venue, pitch condition, practice, etc are the independent variables.

The chart below shows another example of temperature against wind chill factor.



Mathematically speaking, linear regression models are denoted with the equation similar to a linear equation.

y = $\beta 0 \;+\; \beta 1 X1 \;+\; \beta 2 X2 \;+\; \beta 3 X3 \;+$ ...... $+\; \beta n X n \;+\; \epsilon$

where y is the target or dependent variable
$\beta 0$ is the constant coefficient
X1… Xn are the independent features
ϵ is the error term

To find the relation between y and X1, we can say that if X increases by 1 unit, y increases by $\beta 1$ unit provided all the other independent variables are kept constant. Similarly, if the coefficient is negative, y will decrease by $\beta 1$ unit.

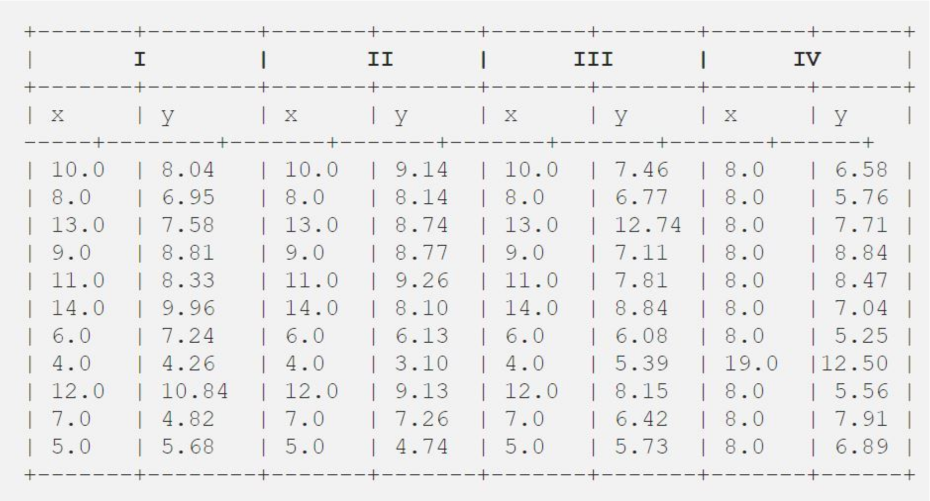
Hypothesis test:

H0 → $\beta1 = 0$
H1 → $\beta1 \neq 0$

If we reject the null hypothesis, we can say that $\beta1$ is not equal to 0 which implies that the target variable is dependent on X1. Likewise, if we fail to reject the null hypothesis, we can infer that X1 is not a significant variable in deriving y.

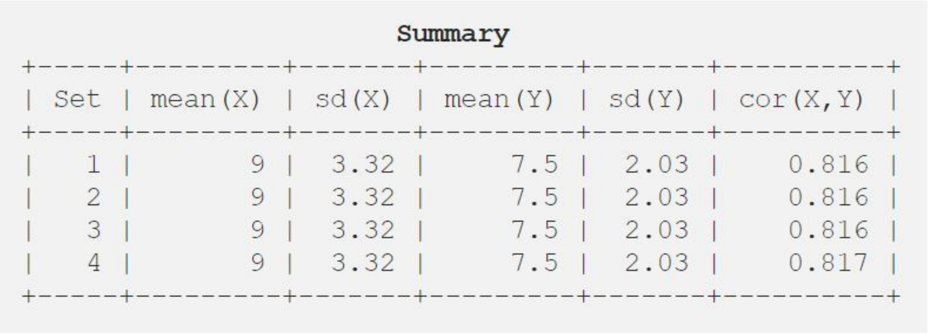
2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is based on 4 sets of data each with 11 points which follows a similar linear plot all with different values but with the same mean and standard deviation. This was constructed by the The famous statistician, Francis John Anscombe who illustrated the importance of visualizing data before performing the regression model and the impact of outliers in statistics.



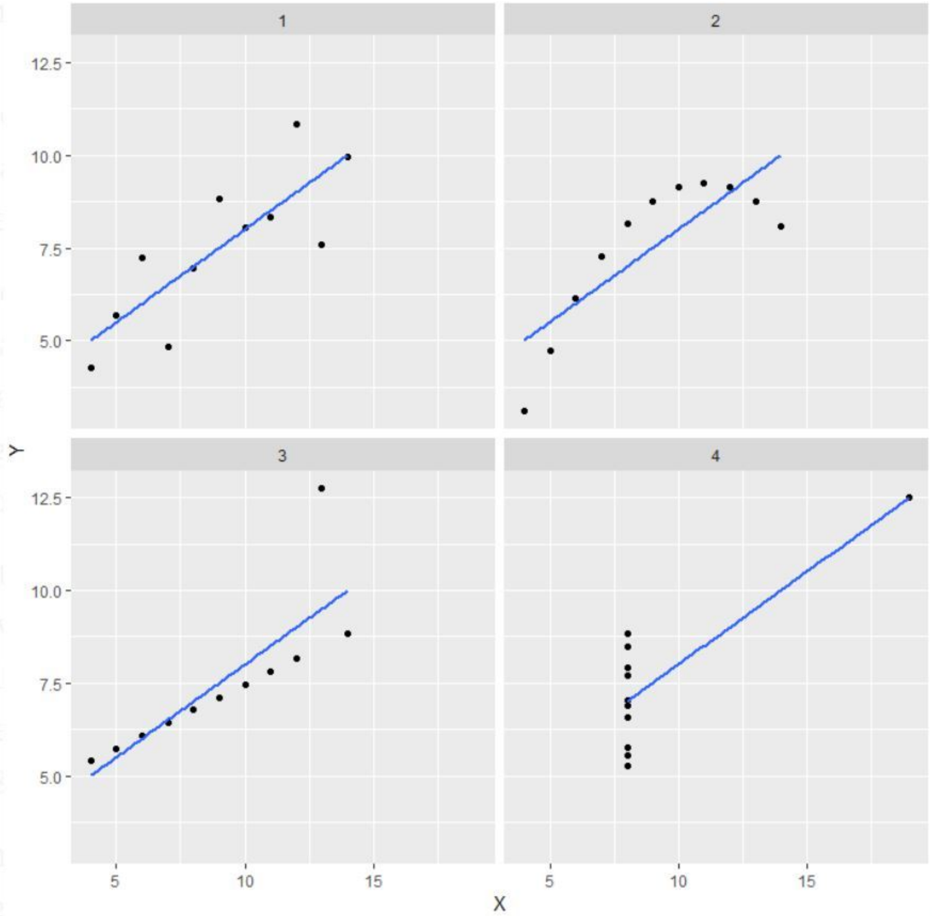
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Chart below shows the mean and standard deviation of 4 sets of data



Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817



All the above 4 plots are the same due to the fact that the presence of outliers have distorted the graphs. The 1st plot seems to be doing good with the data. The 2nd plot shows that linear regression is not capable of handling other kinds of data except the model with linear relationship. 3rd and 4th plots show that the presence of outliers have changed the plot otherwise the line would have passed through all the data points.

3. What is Pearson's R?

Pearson's R or Pearson's correlation coefficient is a statistic that measures linear correlation between the two variables. It's value ranges from -1 to +1. It is not capable of capturing the non-linear relationship between two variables. In general, it is not reliable to use Pearson's R for non-linear relations. In such cases, we use Spearman's R. For example, for the relationship, $\mathbf{Y = X^3}$, the correlation coefficient given by Pearson's R comes out to be ~0.66 as this does not show any linear relationship. The same with Spearman's R comes out to be 1.

Pearson's R is given by the below formula:

$$\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

Where x_i and y_i are the data point and $\overline{x}$ and $\overline{y}$ are the respective means.

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

Scaling is a method which is used to normalize the range of feature variables for deriving better correlations. Without scaling, we can get improper correlation coefficients which can at times become very difficult to infer something about the model. This is done as a part of data preprocessing step after the train and test data is split up.

There are two types of scaling that is performed

1. Normalized scaling or Min-Max scaling (Normalization)
2. Standardized scaling (Standardization)

In normalized scaling, the values are normalized within the range of 0 to 1 whereas in standardized scaling, the values are scaled in such a way that the mean of the new data points is 0 and variance is 1.

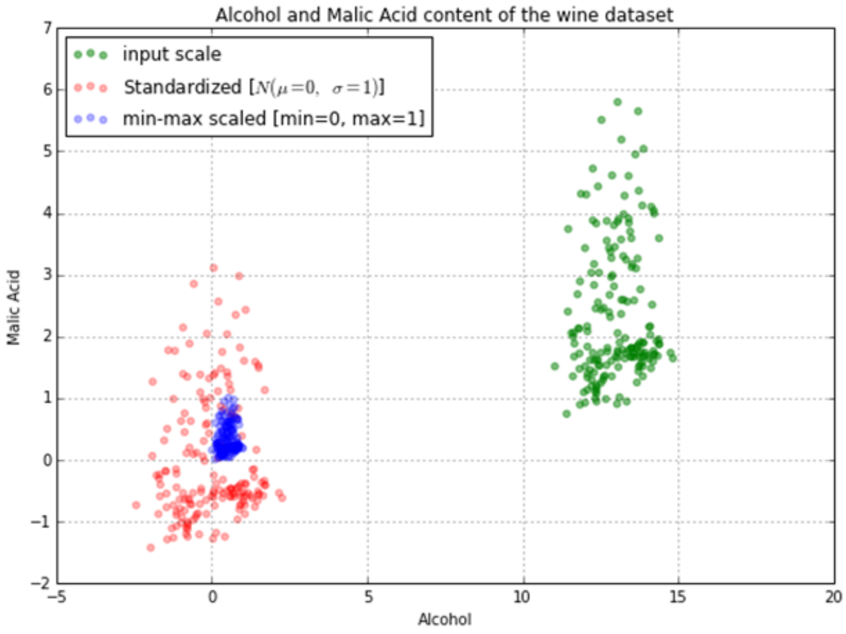
Normalized value:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardized value:

$$x' = \frac{x - \bar{x}}{\sigma}$$

The diagram below taken from explains the difference between both types of feature scaling:



Blue points are in the range 0 to 1 while red points are standardized such that mean = 0.

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

High value of VIF indicates high correlation between the variables. If VIF is infinite, it means there is a multicollinearity that needs to be dealt with.

$VIF = \frac{1}{1-R^2}$

$VIF = \infty \rightarrow R^2 = 1$

This is a case of perfect correlation. We need to drop the variables by inspecting the p-value first to remove such dependency. Else the assumption we made in linear regression won't hold good.

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

Q-Q plots are known as Quantile-Quantile plots which help to plot the quantiles of the sample distribution against that of the theoretical distribution. It helps us to assess whether the datasets are from populations of the common distribution.

Following are the interpretations of the data sets:

a. All points of quantiles lies on or close to the line y = x or at an angle of 45 degrees with X-axis
b. X quantiles are lower than Y quantiles
c. Y quantiles are lower than X quantiles
d. All points of quantiles lies away from the line y = x