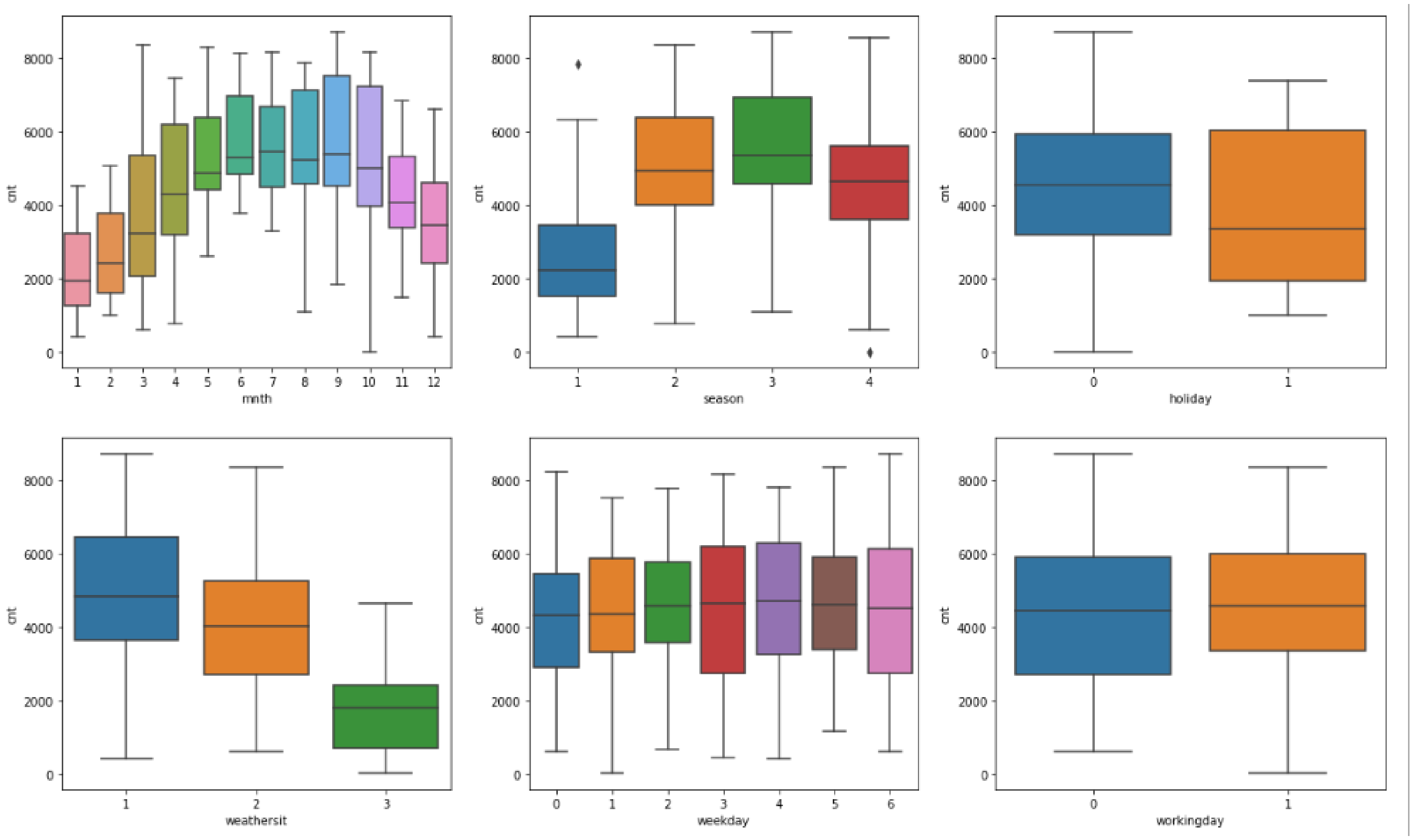
Assignment-based Subjective Questions

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



The Box plot above represents the effect of categorial variables on the dependent variable (‘cnt’).

The inference is as below,

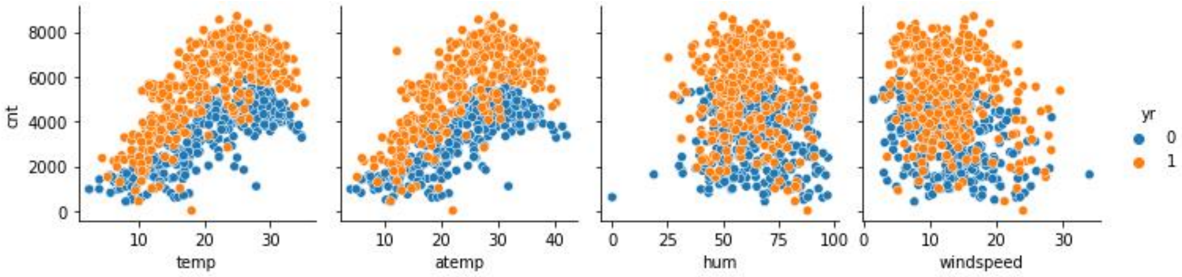
- **mnth**: Most of the bike booking are happening in the months 5,6,7,8 & 9 with a median of over 5K per month which indicates, mnth can be a good predictor for the dependent variable.
- **season**: Most of the bike bookings are happening in season2 & season3 with a median of over 5K booking (for the period of 2 years). This was followed by season4 which indicates, season can be a good predictor for the dependent variable.
- **holiday**: Most of the bike bookings are happening when it is not a holiday which indicates, holiday may not be a good predictor for the dependent variable.

- **weathersit**: Most of the bike bookings are happening during 'weathersit1 with a median of ~5000 (for the period of 2 years). This is followed by weathersit2 and indicates, weathersit can be a good predictor for the dependent variable.
- **weekday**: variable appears uniform across the week with medians between 4K to 5K bookings. This variable may not be a good predictor.
- **workingday**: Most of the bike bookings are happening in 'workingday' with a median of ~5K (for the period of 2 years) which indicates, workingday can be a good predictor for the dependent variable

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

It is important to use, as it helps in reducing the extra column created during dummy variable creation. If you don't, this may affect some models adversely and the effect is stronger when the spread/number of unique values is small. Hence, it reduces the correlations created among dummy variables. If we have categorical variable with n-levels, the resultant will be n-1 columns to represent the dummy variables.

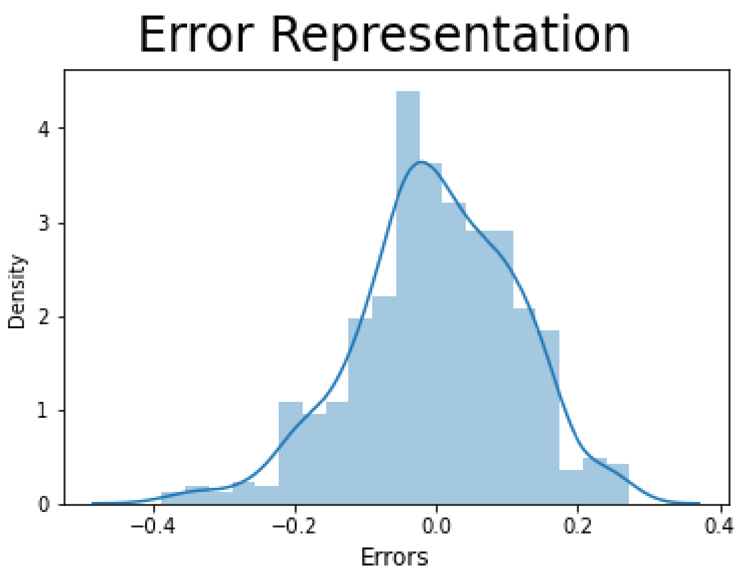
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



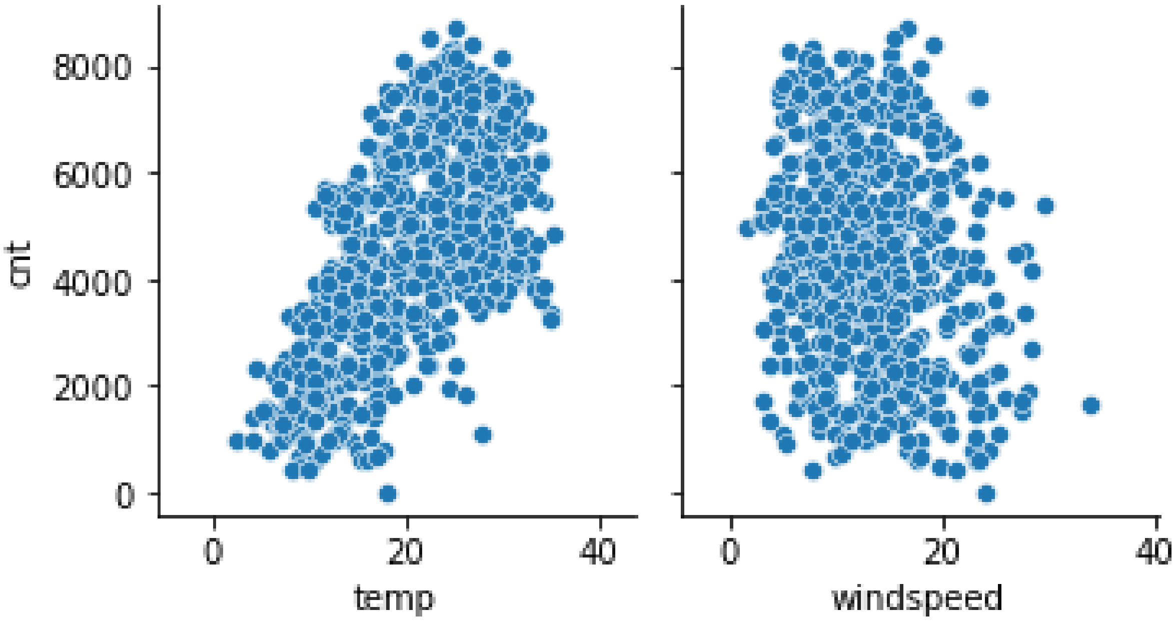
Temperature (temp) has the highest correlation of 0.63 with the target variable ('cnt').

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

- By conducting residual analysis on errors to ensure that they are normally distributed with a mean of 0
- The histogram below confirms to the assumption



- Confirm presence of linear relationship between top predictor (temperature) and target ('cnt')



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

- Temperature (temp): A coefficient value of "0.53" indicating that a unit increase in temperature results in the bike demand to grow by "0.53" units
- Year (yr): A coefficient value of "0.23" indicating that a unit increase in year results in the bike demand to grow by "0.23" units
- Weather Situation 3 (weathersit_3): A coefficient value of "-0.22" indicating that a unit increase in weather_sit3 results in bike demand to decrease by "0.22" units

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used for predictive analysis and shows the relationship between the continuous variables. It demonstrates the linear relationship between the independent variable (X-axis) and the target variable (Y-axis). If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression**.** Linear regression is represented using,

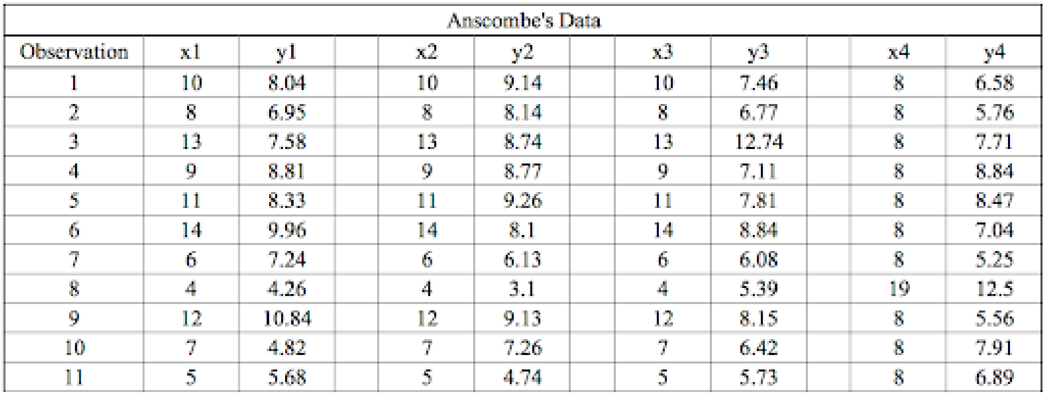
y = b0+b1x
y= Dependent Variable.
x= Independent Variable.
b0= intercept of the line.
b1 = Linear regression coefficient.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship. If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship. The goal of the linear regression algorithm is to get the best values for b0 and b1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

The cost function helps to figure out the best possible values for b0 and b1, which provides the best fit line for the data points. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values. Using the MSE function, we will change the values of b0 and b1 such that the MSE value settles at the minima. Model parameters can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

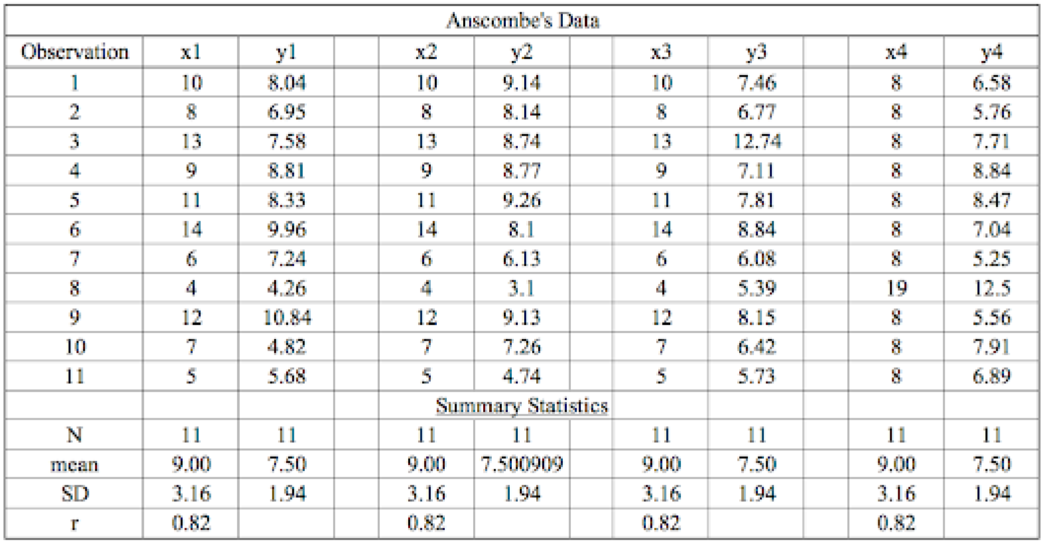
2. Explain the Anscombe's quartet in detail.

Anscombe's quartet tells us about the importance of visualizing data before attempting to build models. This suggests the data points must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, linear separability of the data). Moreover, the linear regression can only be considered a fit for the data with linear relationships. We can define these four plots as follows:



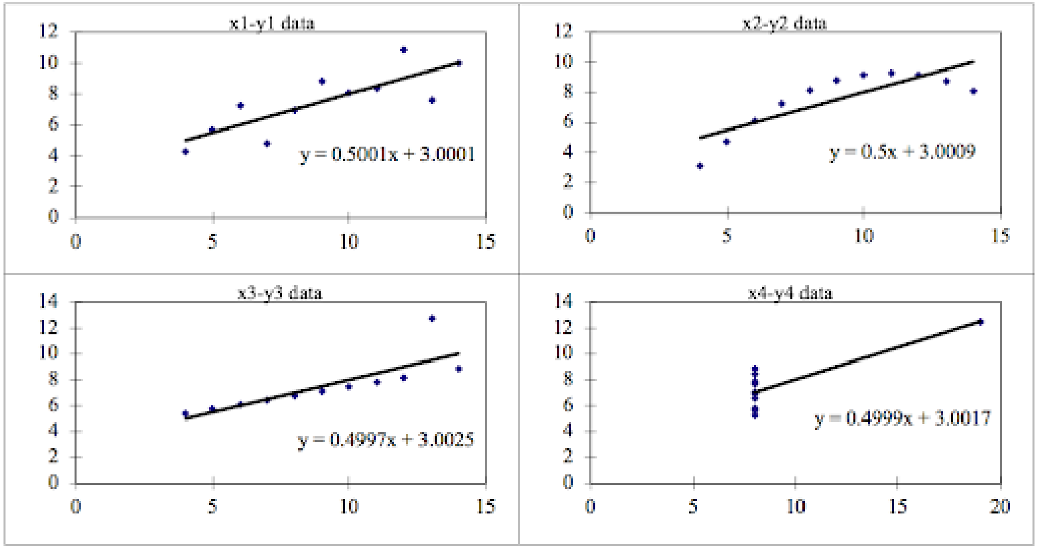
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical info for these four data sets is similar.



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by linear regression algorithm.



Dataset 1: fits the linear regression model well

Dataset 2: cannot fit the linear regression model because the data is non-linear

Dataset 3: shows outliers, which cannot be handled by the linear regression model

Dataset 4: shows outliers, which cannot be handled by the linear regression model

3. What is Pearson's R?

Pearson's correlation coefficient is used to measure the relationship between two continuous variables.

It has a numerical value that lies between -1.0 and +1.0. It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations**.** Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

The formula is as follows,

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Scaling is part of data pre-processing and is applied to independent variables to normalize the data within a particular range. It helps in speeding up the calculations in an algorithm. Occasionally, there could be independent variables highly varying in magnitudes, units and range. If scaling is not conducted, then the model would only take magnitude in account and not units hence incorrectly predicting the target values. To solve this issue, we have scale and bring all the variables to the same level of magnitude. Scaling primarily affects the coefficients and none of the statistical significance or p-value parameters.

Normalization/Min-Max Scaling: It brings all the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling: Replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

sklearn.preprocessing.scale helps to implement standardization in python.

The disadvantage of normalization over standardization is that it loses key information in the original values, especially about outliers. So, while scaling variables with outliers it is recommended to use standardization over normalization. Also, there is no hard and fast rule to determine when to normalize or standardize. We can always start by fitting the model to raw, normalized, and standardized data and compare the performance for best results.

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

VIF stands for Variance Inflation Factor and VIF = Infinity indicates a perfect correlation between two independent variables. In a perfect correlation, R2 = 1 which then leads to,

VIF = 1/(1-R2) = 1/ (1-1) = 1/0 = Infinity

In order to solve this problem, we will have to drop one of the independent variables which is causing the perfect multicollinearity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

We plot the theoretical quantiles on the x-axis and the ordered values for which we want to find whether it is normally distributed or not on the y-axis. We then focus on the ends of the straight line and if the points are in a curve, we conclude that the values are not normally distributed.

A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

Q-Q plots are also used to find the Skewness of a distribution. When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution, we want to know on the y-axis then we see a very peculiar shape of a normally distributed Q-Q plot for skewness. If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution is left-skewed but when we see the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then it is right-skewed.

Similarly, the measure of Tailedness or Kurtosis of the distribution can be determined by looking at its Q-Q plot. The distribution with a fat tail will have both the ends of the Q-Q plot to deviate from the straight line and its center follows a straight line, whereas a thin-tailed distribution will form a Q-Q plot with a very less or negligible deviation at the ends thus making it a perfect fit for the normal distribution.