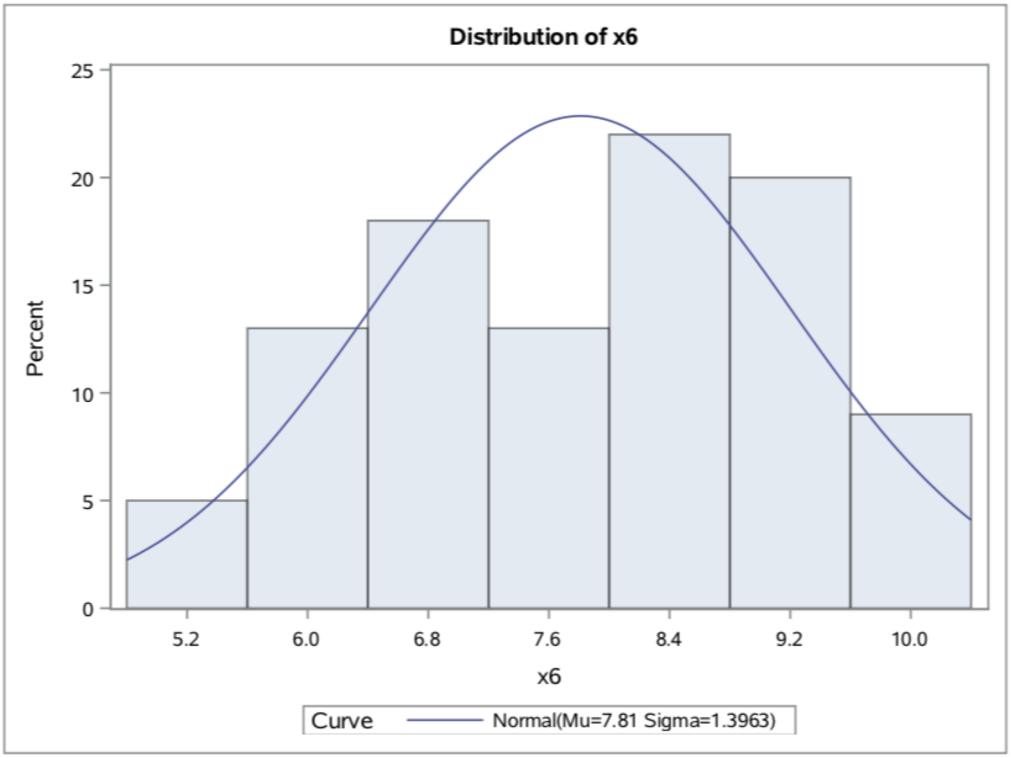
1. Missing Data

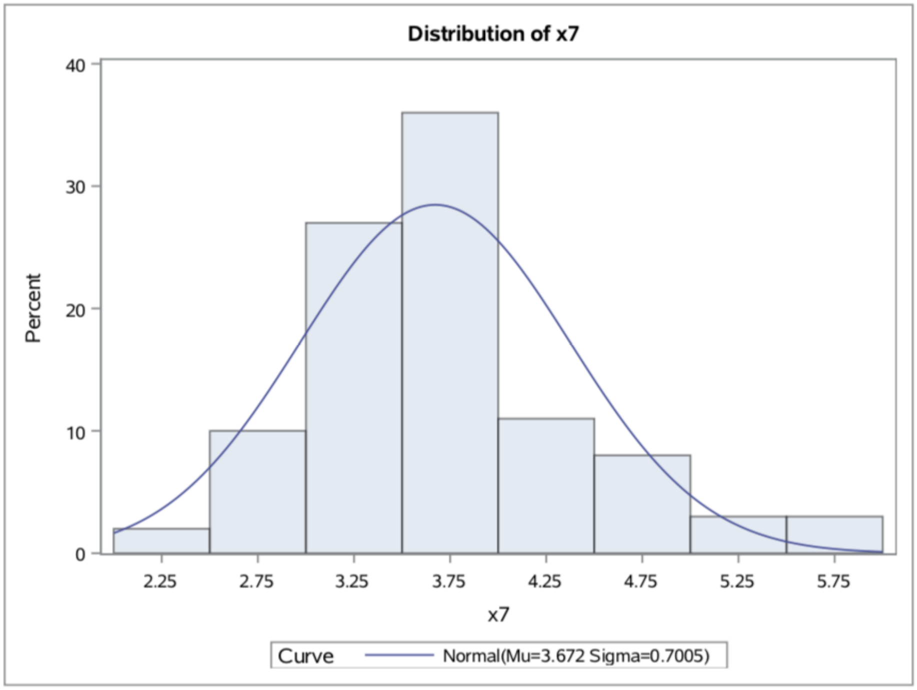
Missing data is an important topic because data sets often will not be perfect when you consider the real world. Missing data is defined as information that is not available for a particular subject or case when there is other information about that subject/case available. This is a relatively common phenomenon.

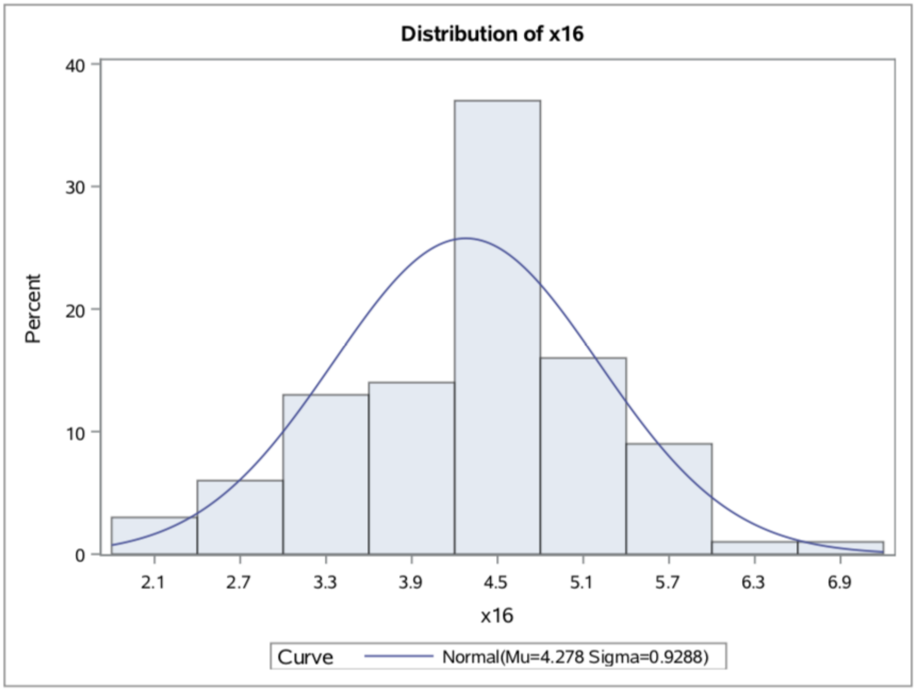
There are several causes of missing data. First, nonresponse can occur in surveys, when the subject answers some of the questions but not all of them. Many times the omitted responses come on personal questions. Next, attrition can occur when a subject begins an experiment but does not complete it. Another example is improper data collection. This is a more generic category. It can occur when the researcher loses data, mislabels data, improperly records data, improperly conducts the experiment, etc. Lastly, governments and private companies frequently publish data on many different topics, but sometimes there is data missing. This can be due to the entities choosing not to publish certain data if it does not look favorable for them, or it can simply be due to them not having the data or having missing data themselves.

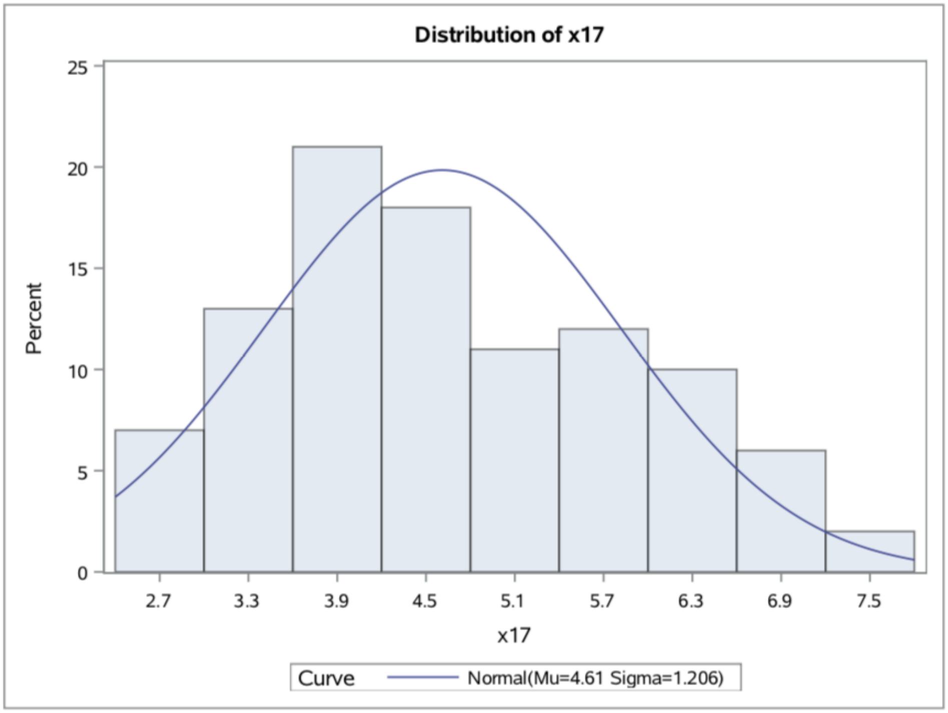
After identifying missing data, one should consider any patterns. Is the missing data completely random, random, or not random at all? The answer to this question will help determine how one should treat the missing data. Another task after identifying missing data is quantifying it. If under 10% of the data for an individual observation is missing, this can be ignored. However, if the data is missing in a manner that is not random, this must be reconsidered. If the data is still going to be used, the missing data needs to be handled, and there are several strategies for this. One can only use observations with complete data, certain cases or variables with the missing data can be deleted, or the missing values can be estimated.

2a.

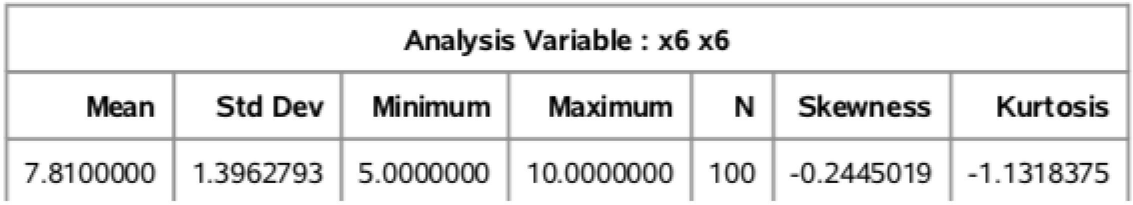



Distribution of x7
Percent
40
30
20
10
0
2.25
2.75
3.25
3.75
4.25
4.75
5.25
5.75
x7
Curve
Normal(Mu=3.672 Sigma=0.7005)

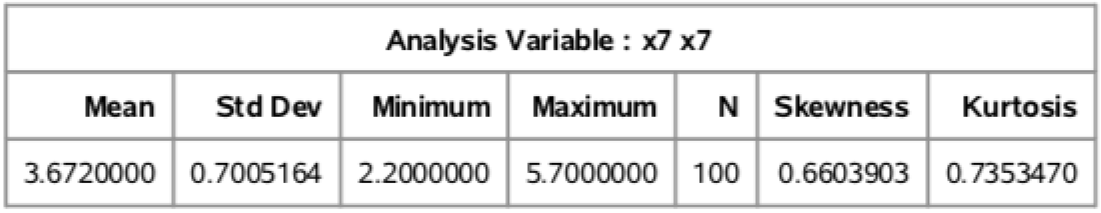

Distribution of x16
Percent
40
30
20
10
0
2.1
2.7
3.3
3.9
4.5
5.1
5.7
6.3
6.9
x16
Curve
Normal(Mu=4.278 Sigma=0.9288)



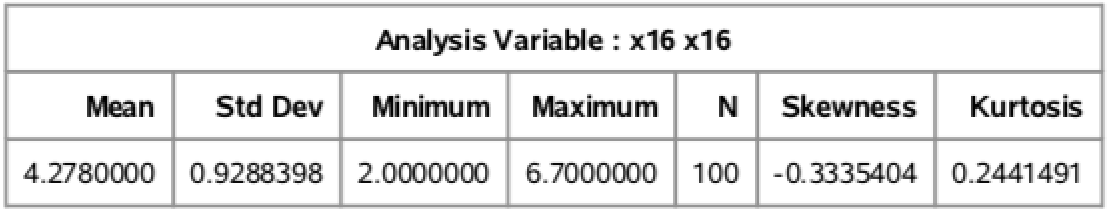
2b.



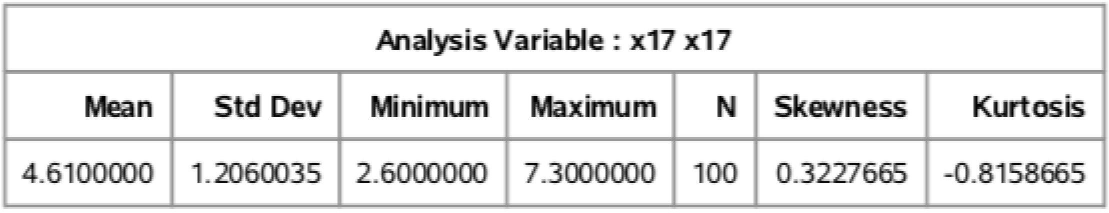
Analysis Variable : x6 x6						
Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
7.8100000	1.3962793	5.0000000	10.0000000	100	-0.2445019	-1.1318375



Analysis Variable : x7 x7						
Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
3.6720000	0.7005164	2.2000000	5.7000000	100	0.6603903	0.7353470



Analysis Variable : x16 x16						
Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
4.2780000	0.9288398	2.0000000	6.7000000	100	-0.3335404	0.2441491



Analysis Variable : x17 x17						
Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
4.6100000	1.2060035	2.6000000	7.3000000	100	0.3227665	-0.8158665

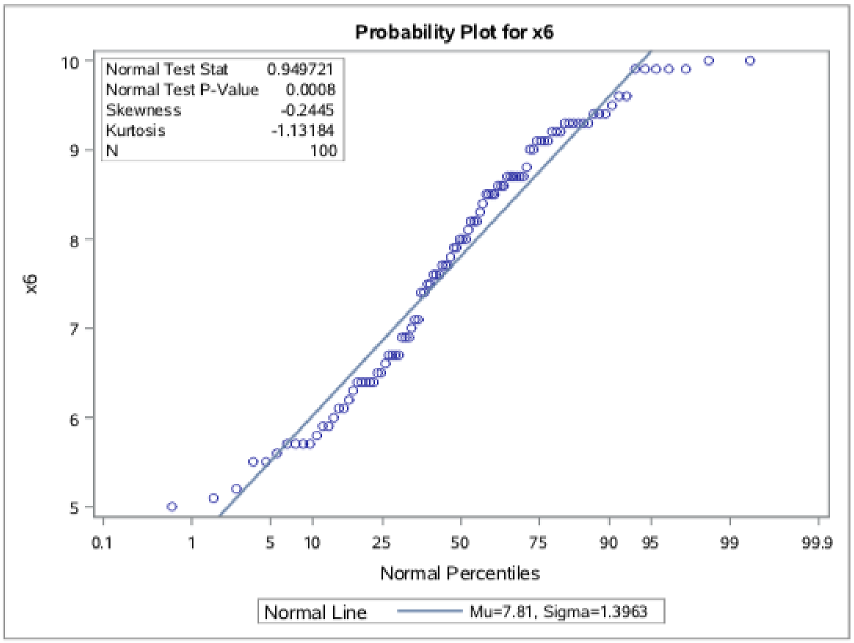
Using the formula, the z for skewness for x6 is ~1.0 and the z for kurtosis is ~2.3. The z for kurtosis may suggest a lack of normality depending on the significance level.

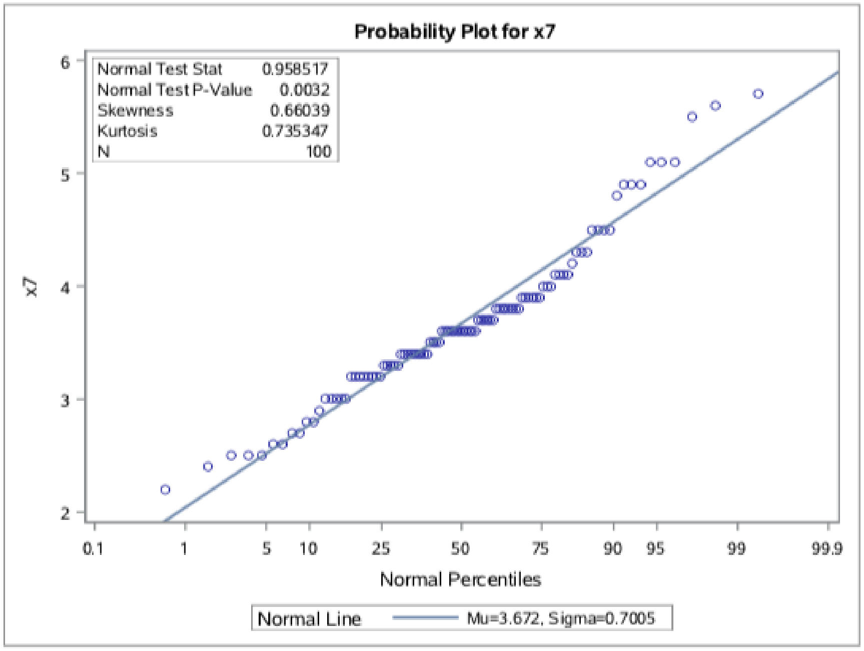
For x7, the z for skewness is ~2.7 and the z for kurtosis is ~1.5. The skewness z definitely suggests a lack of normality.

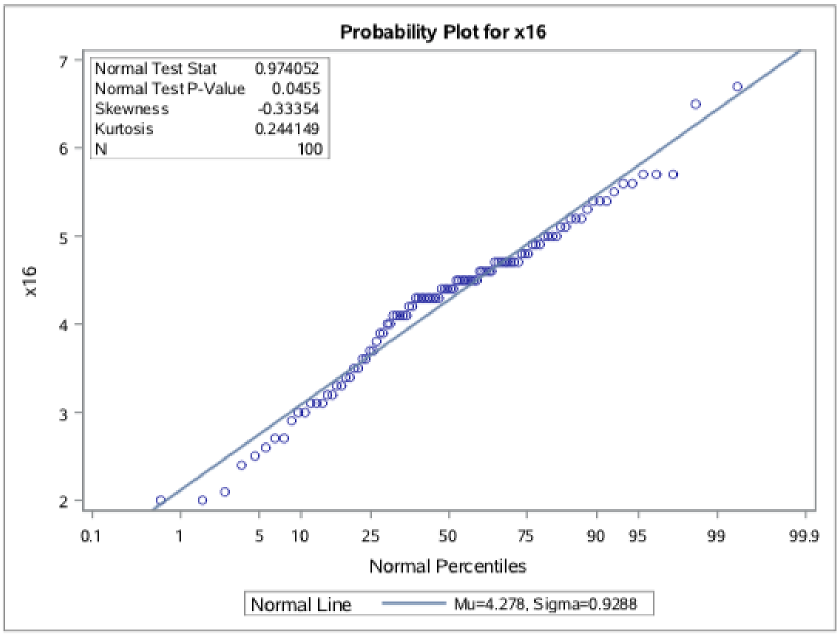
For x16, the z for skewness is ~1.4 and the z for kurtosis is ~.5. These z values do not suggest a lack of normality.

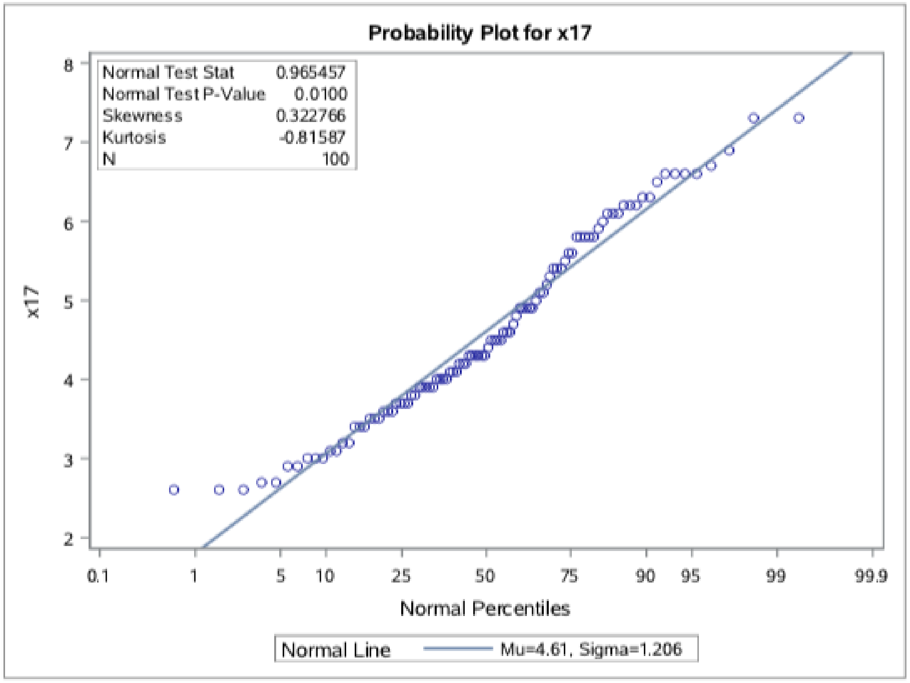
For x17, the z for skewness is ~1.3 and the z for kurtosis is ~1.7. These z values do not suggest a lack of normality.

2c.



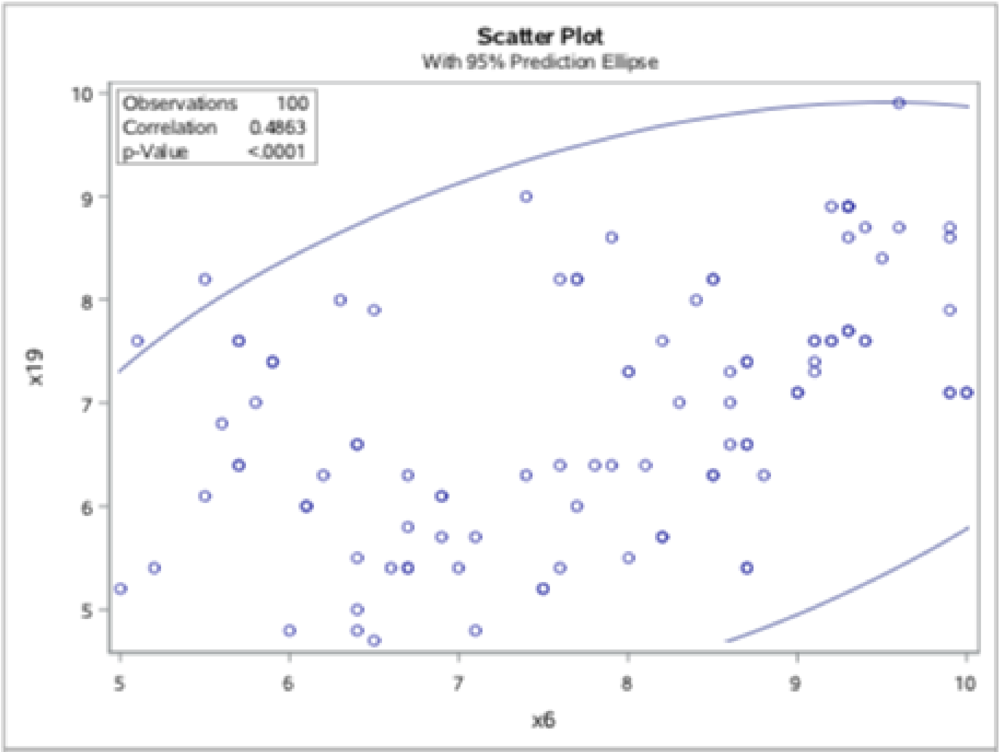


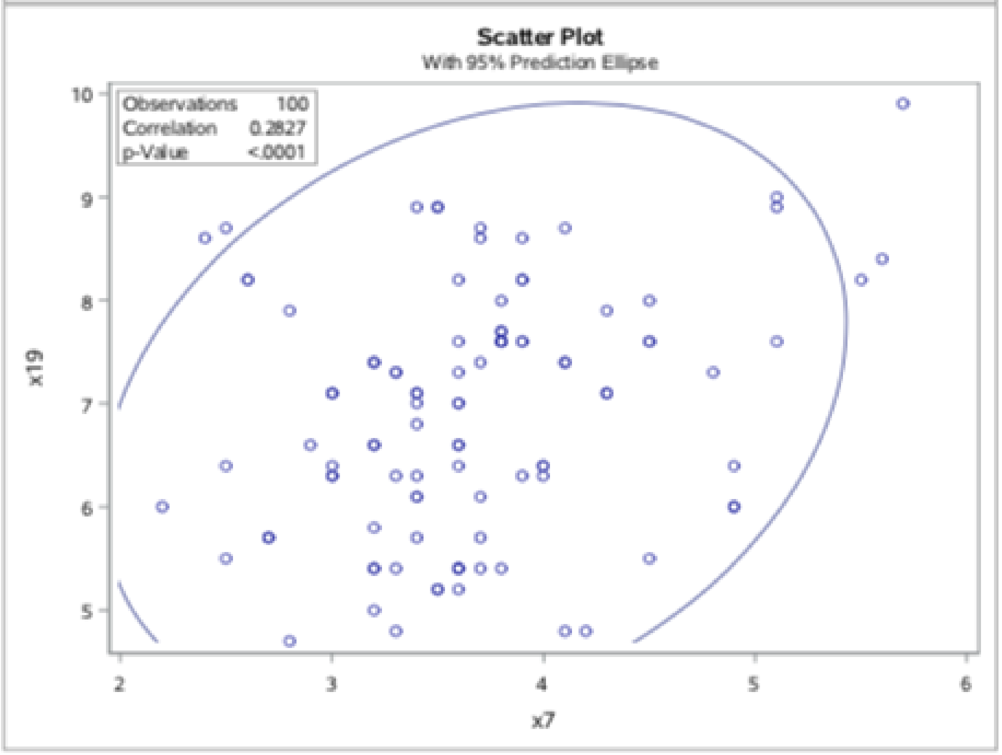




All the p-values are below 0.05. However, x16 is not below 0.01. So depending on the significance level, this is the only variable that may not be normal.

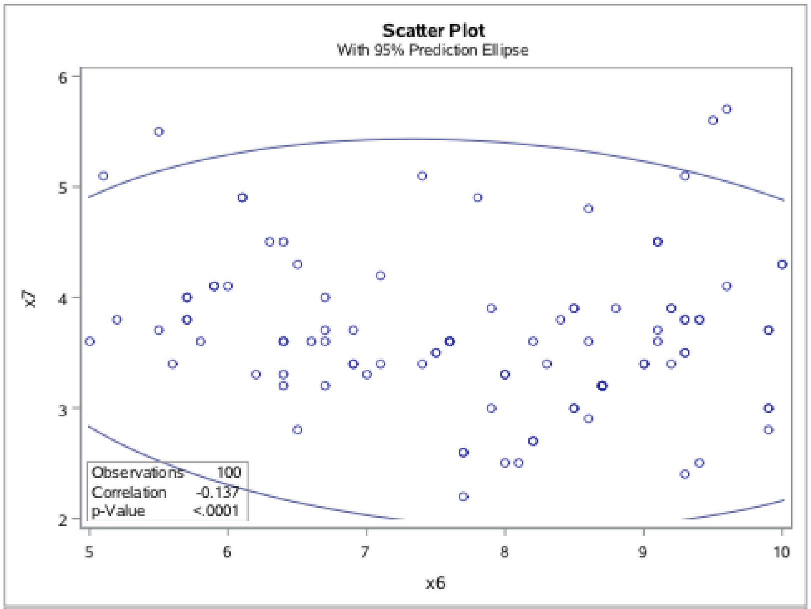
2d.

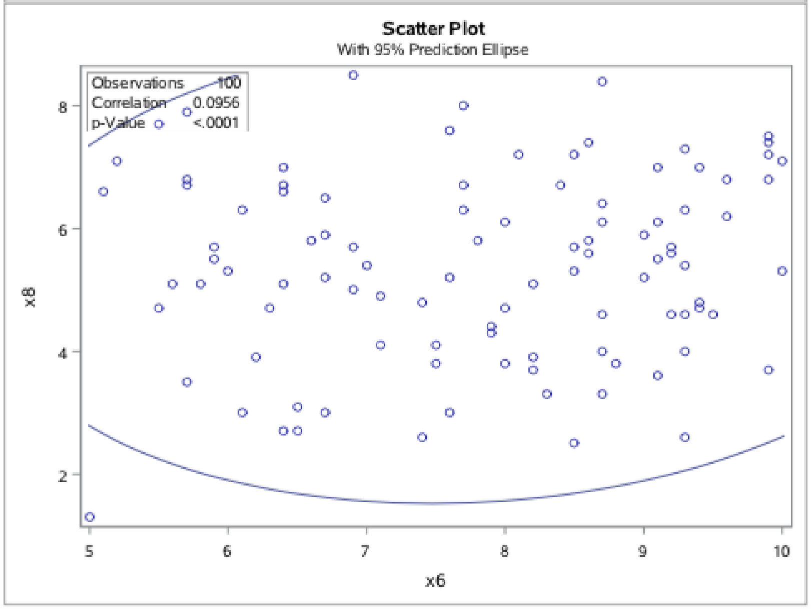


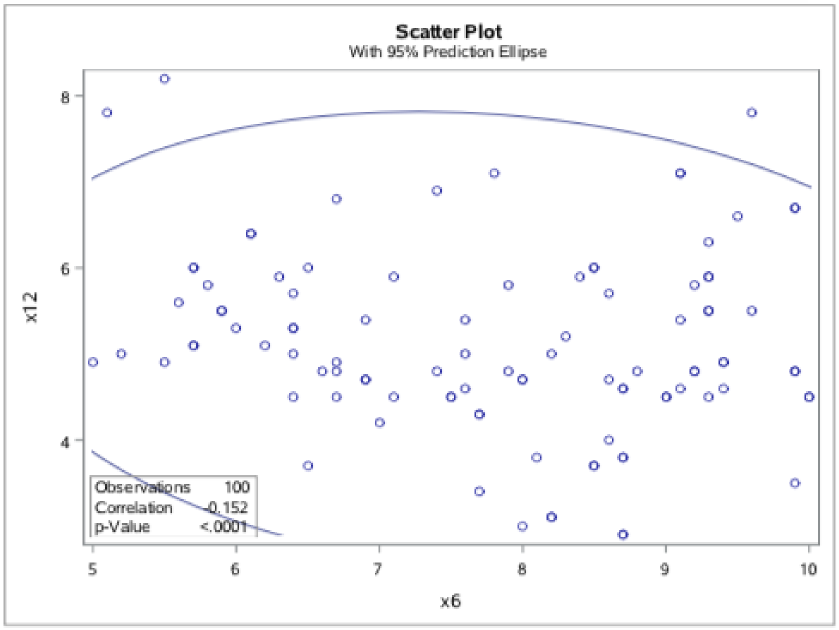


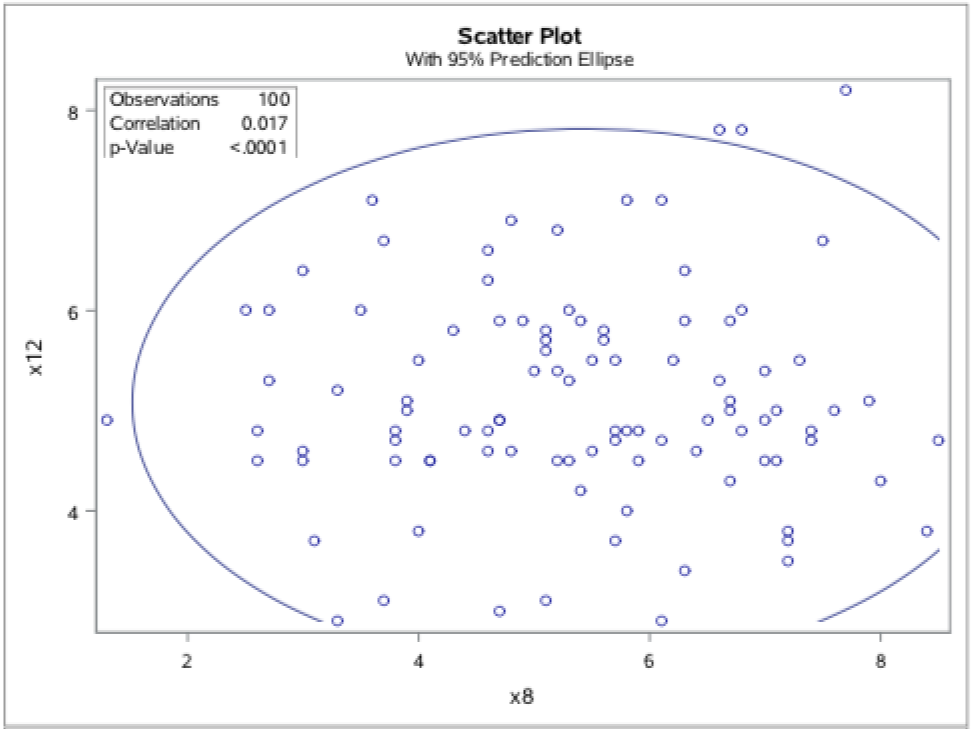
In the top plot, there are two outliers plus one that is right on the border. In the bottom plot, there are five outliers.

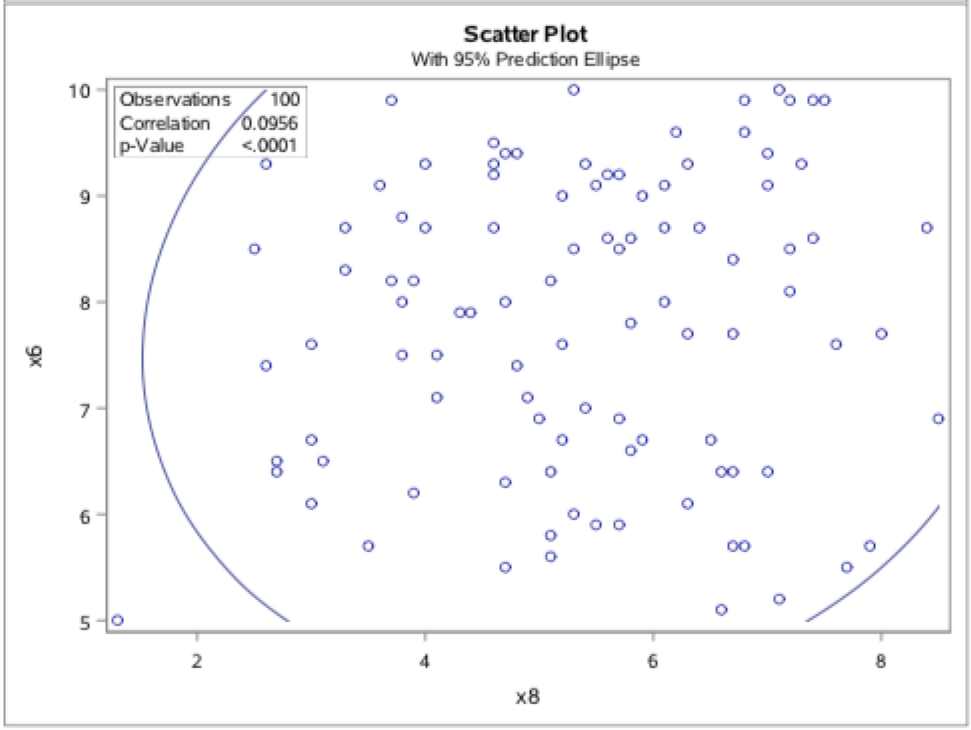
2e.

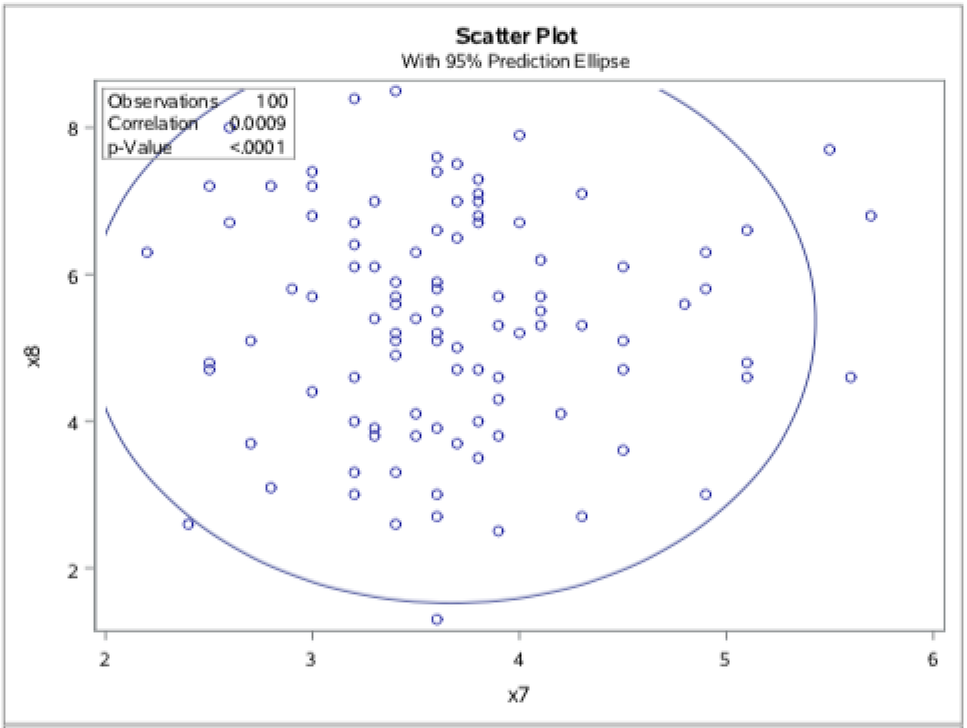

Scatter Plot
With 95% Prediction Ellipse
Observations 100
Correlation -0.137
p-Value <.0001
x7
2
3
4
5
6
x6
5
6
7
8
9
10

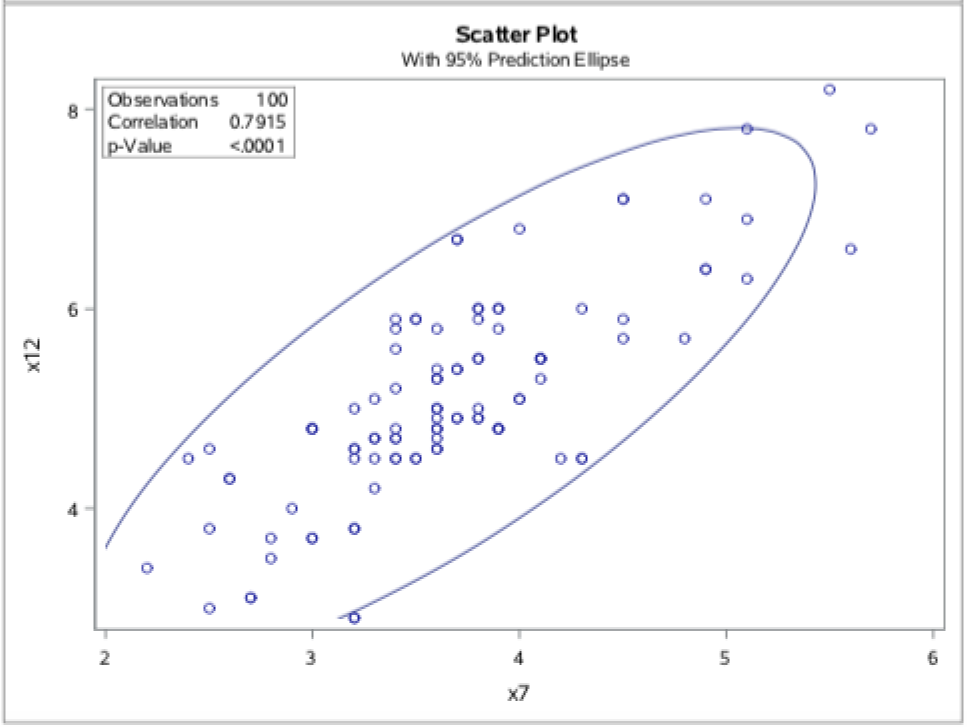

Scatter Plot
With 95% Prediction Ellipse
Observations 100
Correlation 0.0956
p-Value <.0001
x8
2
4
6
8
x6
5
6
7
8
9
10


Scatter Plot
With 95% Prediction Ellipse
Observations 100
Correlation -0.152
p-Value <.0001
x12
4
6
8
x6
5
6
7
8
9
10


Scatter Plot
With 95% Prediction Ellipse
Observations 100
Correlation 0.017
p-Value <.0001
x12
8
6
4
2
4
6
8
x8

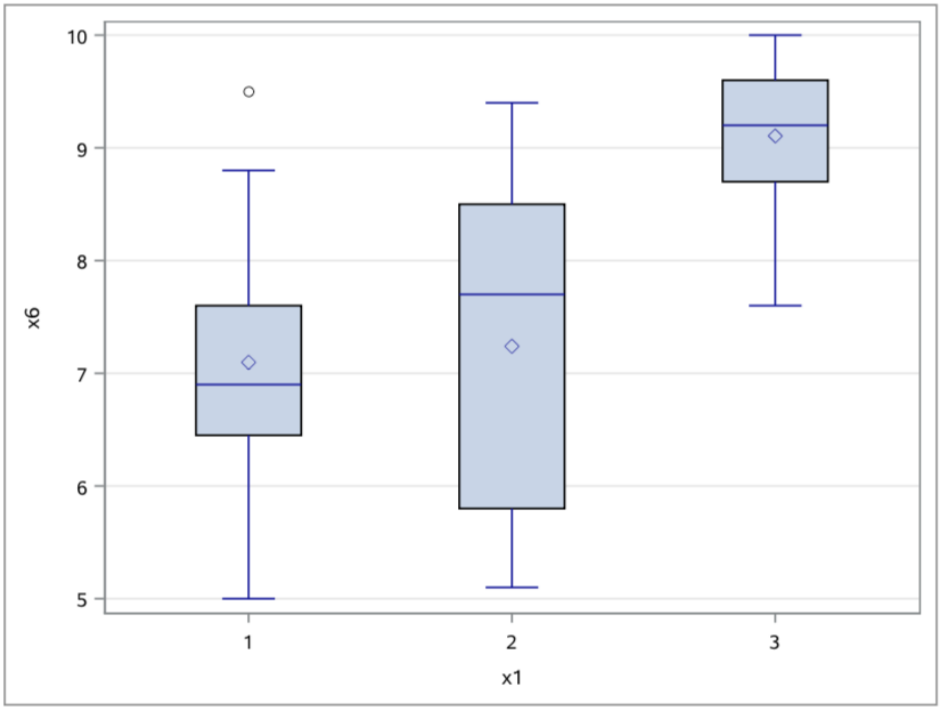

Scatter Plot
With 95% Prediction Ellipse
Observations 100
Correlation 0.0956
p-Value <.0001
x6
10
9
8
7
6
5
2
4
6
8
x8

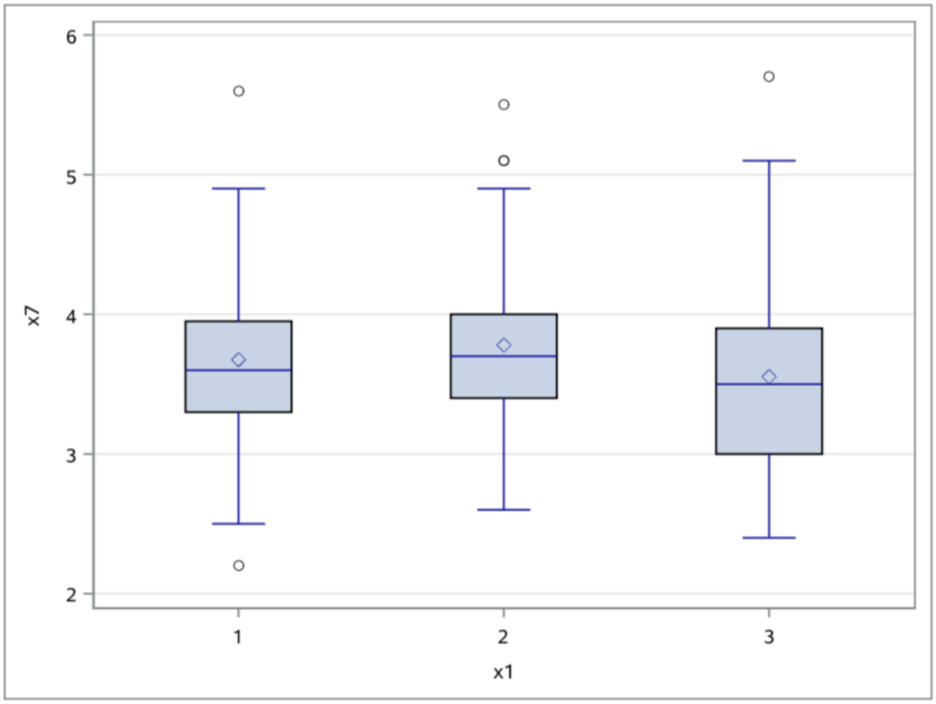




Based on the plots and the correlation numbers, x7 and x12 are the only ones that are correlated.

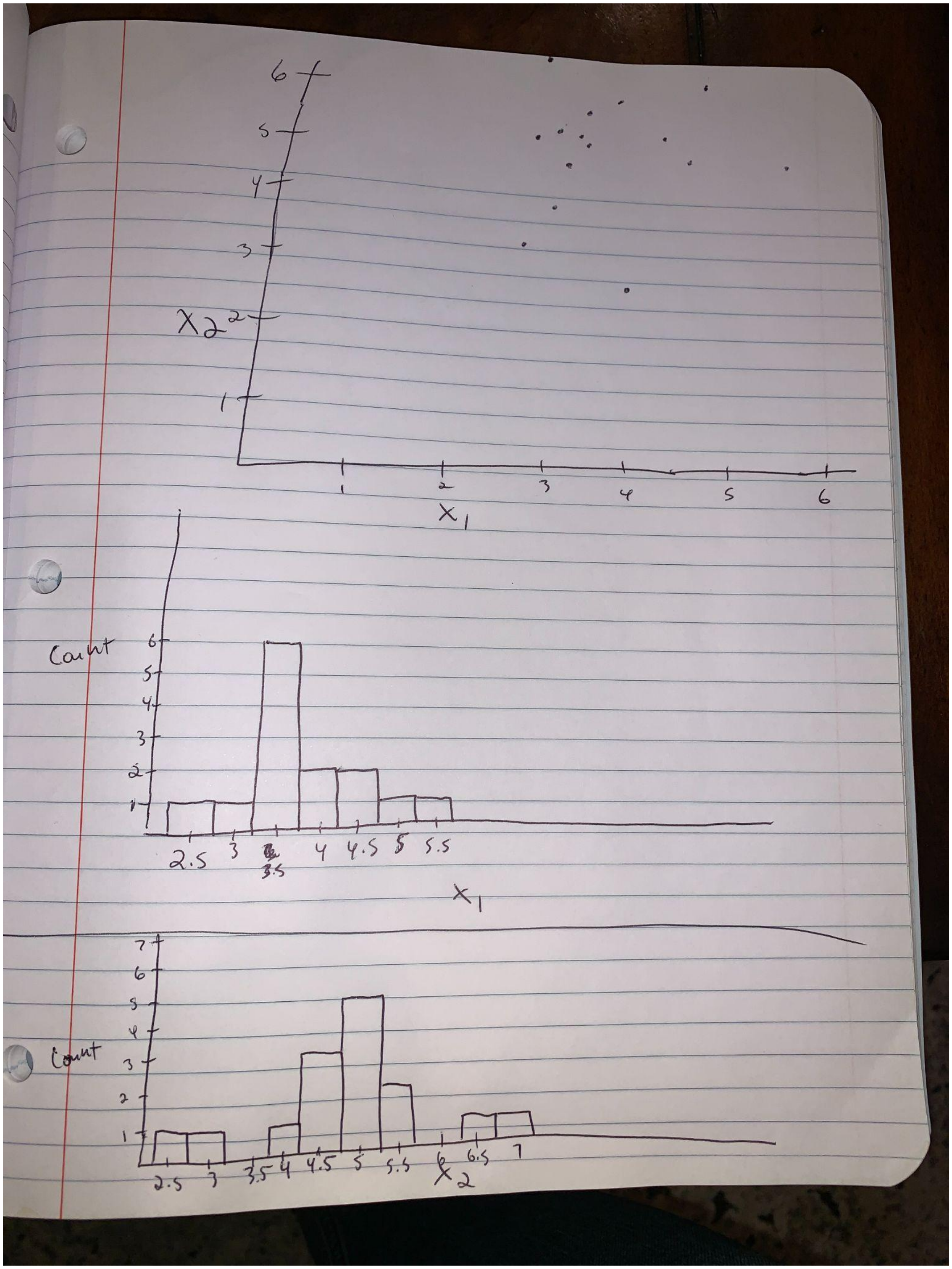
2f.





Yes, box plots can detect outliers. The ‘whiskers’ or error bars show the range of the data. If there are outliers, they will appear outside of the whiskers on the plot. In the top plot, you can see one outlier, and in the bottom there are five.

3.



X2 looks much more normal than X1 as there is much more natural distribution of numbers that fit the traditional bell curve shape.