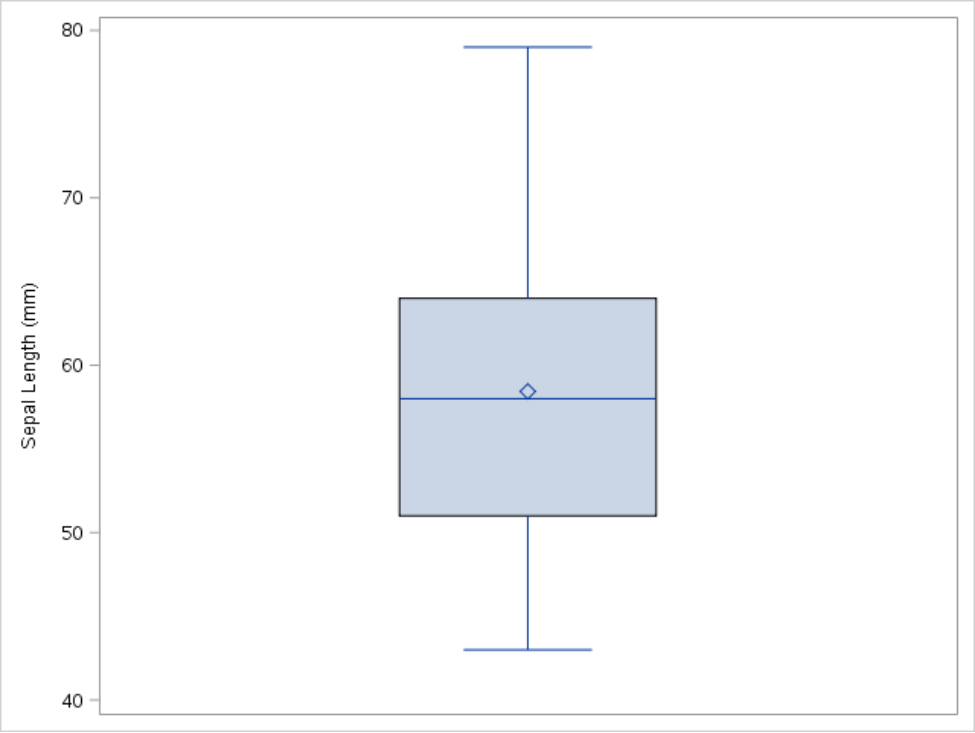
SAS HW1

Dataset snapshot for 10 observations. There are three types of species and a total of 150 observations.

Obs	Species	SepalLength	SepalWidth	PetalLength	PetalWidth
1	Setosa	50	33	14	2
2	Setosa	46	34	14	3
3	Setosa	46	36	10	2
4	Setosa	51	33	17	5
5	Setosa	55	35	13	2
6	Setosa	48	31	16	2
7	Setosa	52	34	14	2
8	Setosa	49	36	14	1
9	Setosa	44	32	13	2
10	Setosa	50	35	16	6

Section 1 Question 1: Descriptive Statistics:

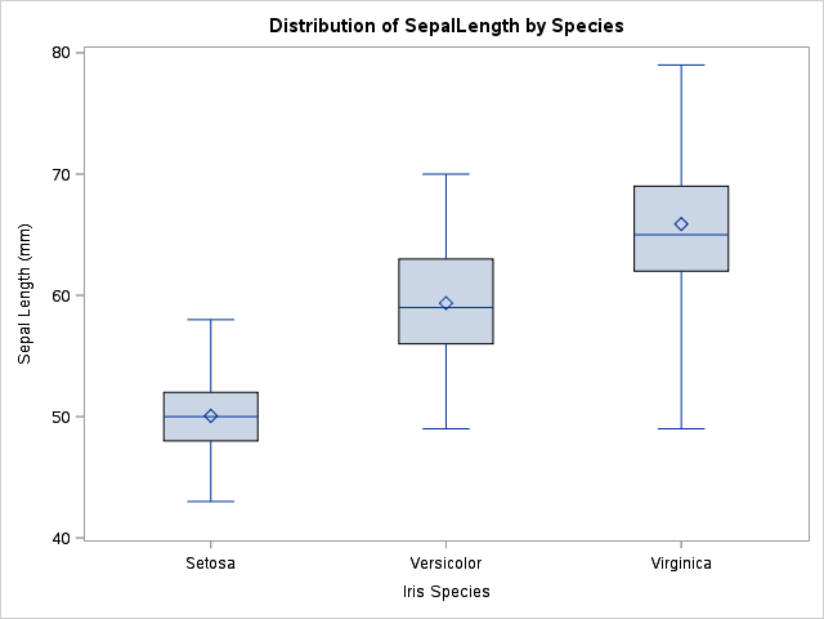


1.The box plot for the sepal length of the entire dataset is shown above. The median and the mean seem close to each other. The bottom quartile is a little larger than the top quartile, which means there is a slight skew in the data. The mean is a little larger than the median and hence this shows **skew to right** .

From the middle 25% of data, it seems like lower quartile is a little wider than the upper. Also, considering the bottom 50%, it is taking less space /variation than top 50%. This implies the bars are higher on the left as a lot of data is accumulated in small range, and the height of bars is relatively lower on the right for a respective histogram. Thus the data is skewed to the right .

2. The plots show that the typical sepal length for three species is different.

- The highest to lowest mean value is: Virginica> Versicolor> Setosa.
  The spread of the sepal length is highest for Virginica and lowest for Setosa which also seems symmetric.
  The other two look slightly skewed as median seems to make a very slight non symmetric division.
  The ranges are also different, in the order Virginica> Versicolor> Setosa



3. For overall Sepal length:

Basic Statistical Measures			
Location		**Variability**	
Mean	58.43333	**Std Deviation**	8.28066
Median	58.00000	**Variance**	68.56935
Mode	50.00000	**Range**	36.00000
		Interquartile Range	13.00000

Basic Measures: The Mean and median are nearly equal at 58.4 and 58.0 respectively. The mode is 50. The variance is 68.57 and std deviation is 8.2. The range and IQR are 36 and 13 respectively. The data from Moments section shows skew of

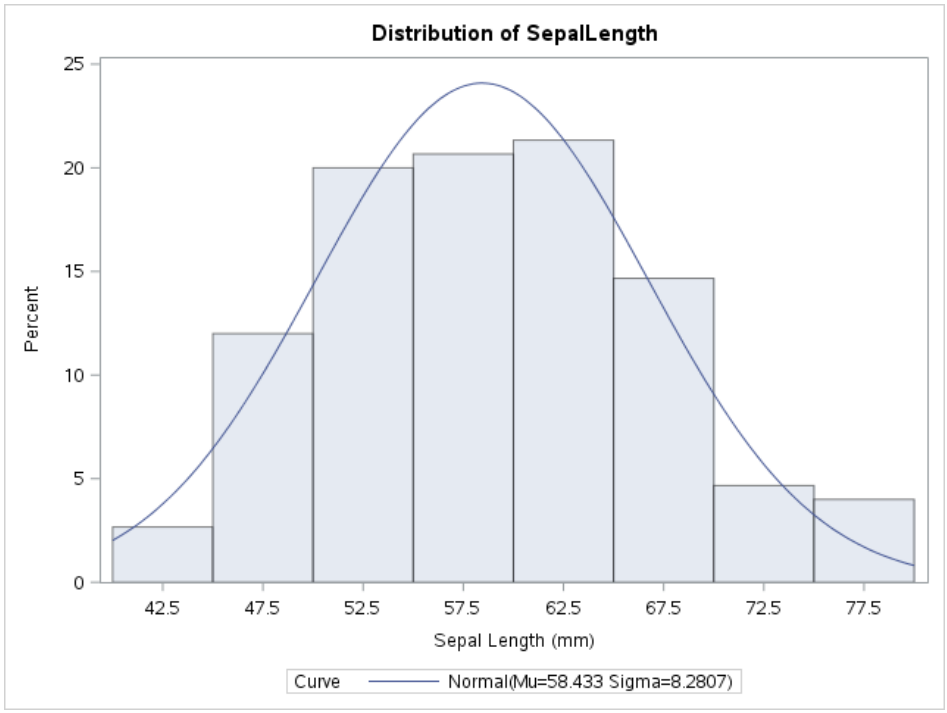
Skewness	0.31491096

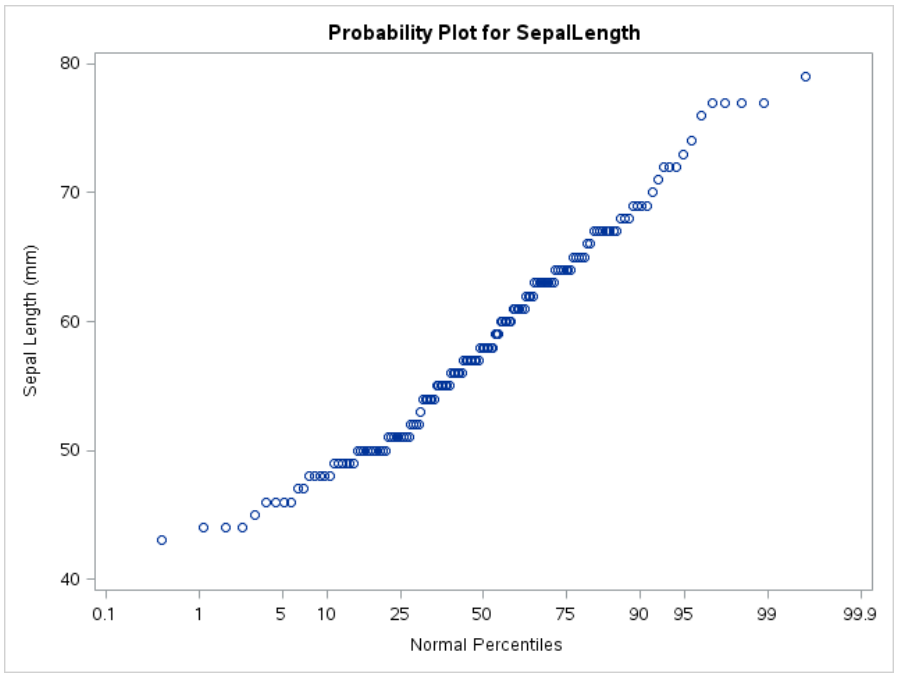
"The numerical value of the mode is the same as that of the mean and median in a normal distribution"- Wikipedia. Here mean, median mode are different which also indicates skew.

All tests for normality show that we cannot assume normality. The p values are less than alpha, reject null hypothesis that the distribution is normal.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.97609	**Pr < W**	0.0102
Kolmogorov-Smirnov	**D**	0.088654	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.127398	**Pr > W-Sq**	0.0479
Anderson-Darling	**A-Sq**	0.889199	**Pr > A-Sq**	0.0231

But the plots show that the data looks nearly normal. However, we go with the statistical evidence.


Distribution of SepalLength
Percent
25
20
15
10
5
0
42.5
47.5
52.5
57.5
62.5
67.5
72.5
77.5
Sepal Length (mm)
Curve
Normal(Mu=58.433 Sigma=8.2807)


Probability Plot for SepalLength
Sepal Length (mm)
80
70
60
50
40
0.1
1
5
10
25
50
75
90
95
99
99.9
Normal Percentiles

4. For species- wise sepal length:

1. Setosa, 2. Versicolor, 3. Virginica

Basic Statistical Measures			
Location		**Variability**	
Mean	50.06000	**Std Deviation**	3.52490
Median	50.00000	**Variance**	12.42490
Mode	50.00000	**Range**	15.00000
		Interquartile Range	4.00000

Basic Statistical Measures			
Location		**Variability**	
Mean	59.36000	**Std Deviation**	5.16171
Median	59.00000	**Variance**	26.64327
Mode	55.00000	**Range**	21.00000
		Interquartile Range	7.00000

Basic Statistical Measures			
Location		**Variability**	
Mean	65.88000	**Std Deviation**	6.35880
Median	65.00000	**Variance**	40.43429
Mode	63.00000	**Range**	30.00000
		Interquartile Range	7.00000

Mean, median, mode values for 1. Setosa, 2. Versicolor, 3. Virginica:

(50.0, 50.0, 50.0) < (59.36000, 59.00000, 55.00000) < (**65.88000**, 65.00000, 63.00000).

Virginica's mean is highest.

Tests for normality: 1. Setosa, 2. Versicolor, 3. Virginica

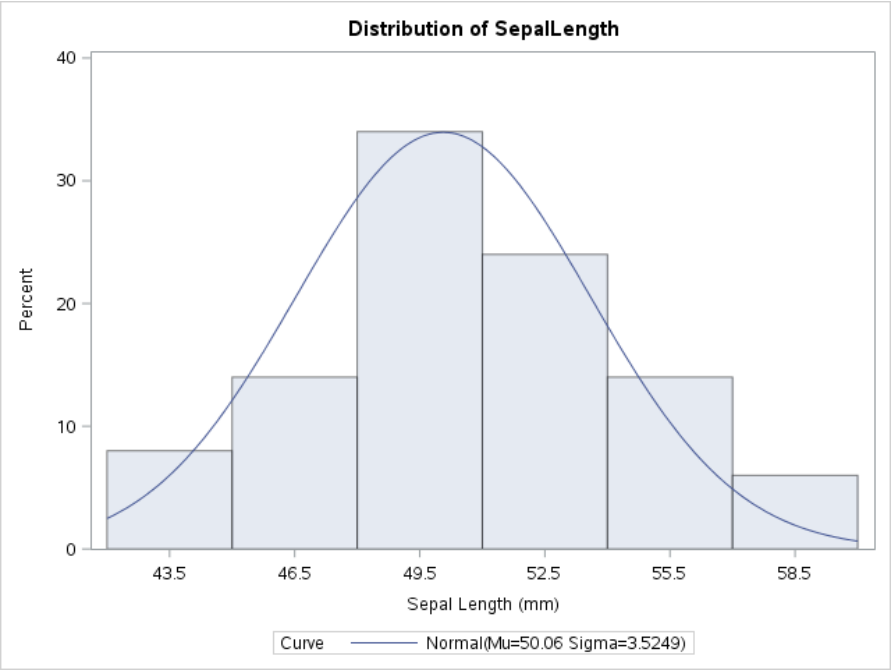
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.977699	**Pr < W**	0.4595
Kolmogorov-Smirnov	**D**	0.11486	**Pr > D**	0.0962
Cramer-von Mises	**W-Sq**	0.071753	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.407986	**Pr > A-Sq**	>0.2500

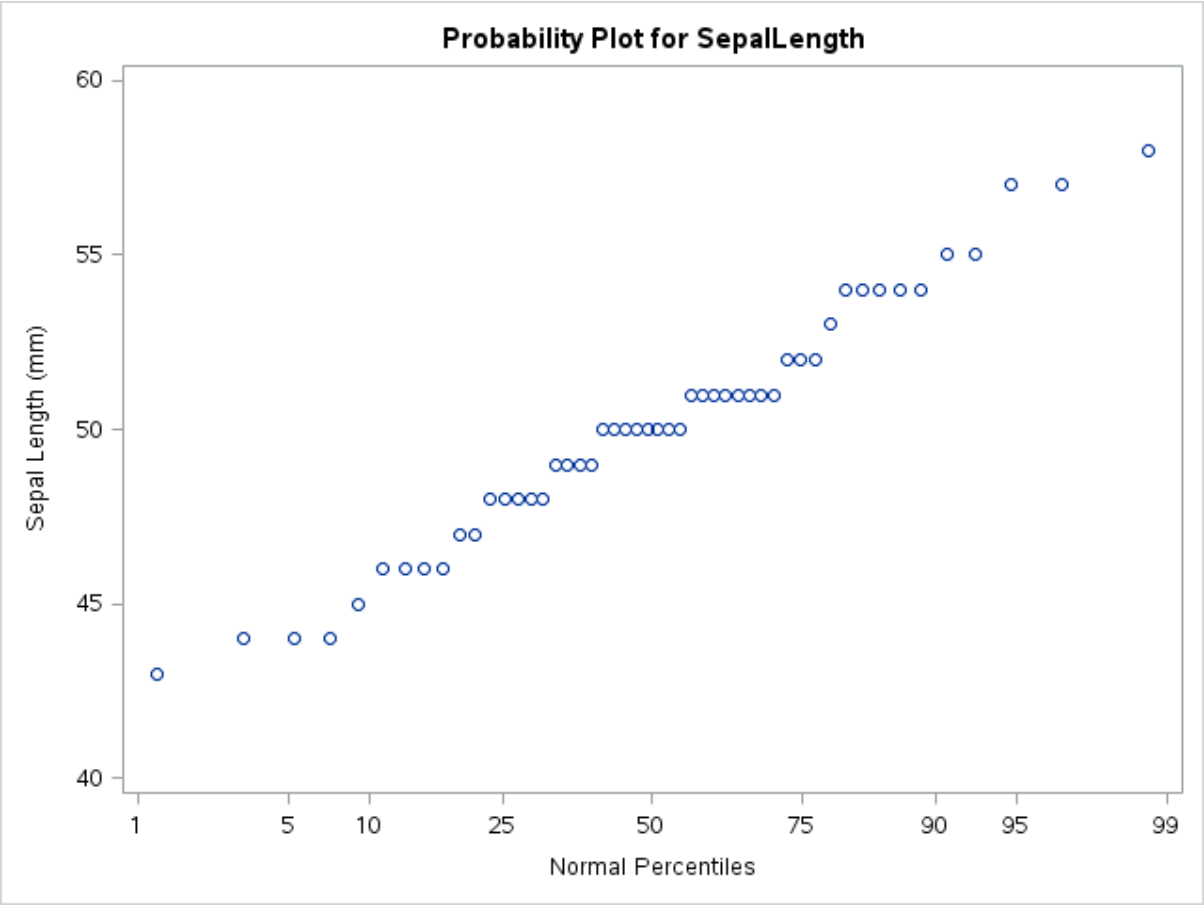
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.977836	**Pr < W**	0.4647
Kolmogorov-Smirnov	**D**	0.096241	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.057273	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.360841	**Pr > A-Sq**	>0.2500

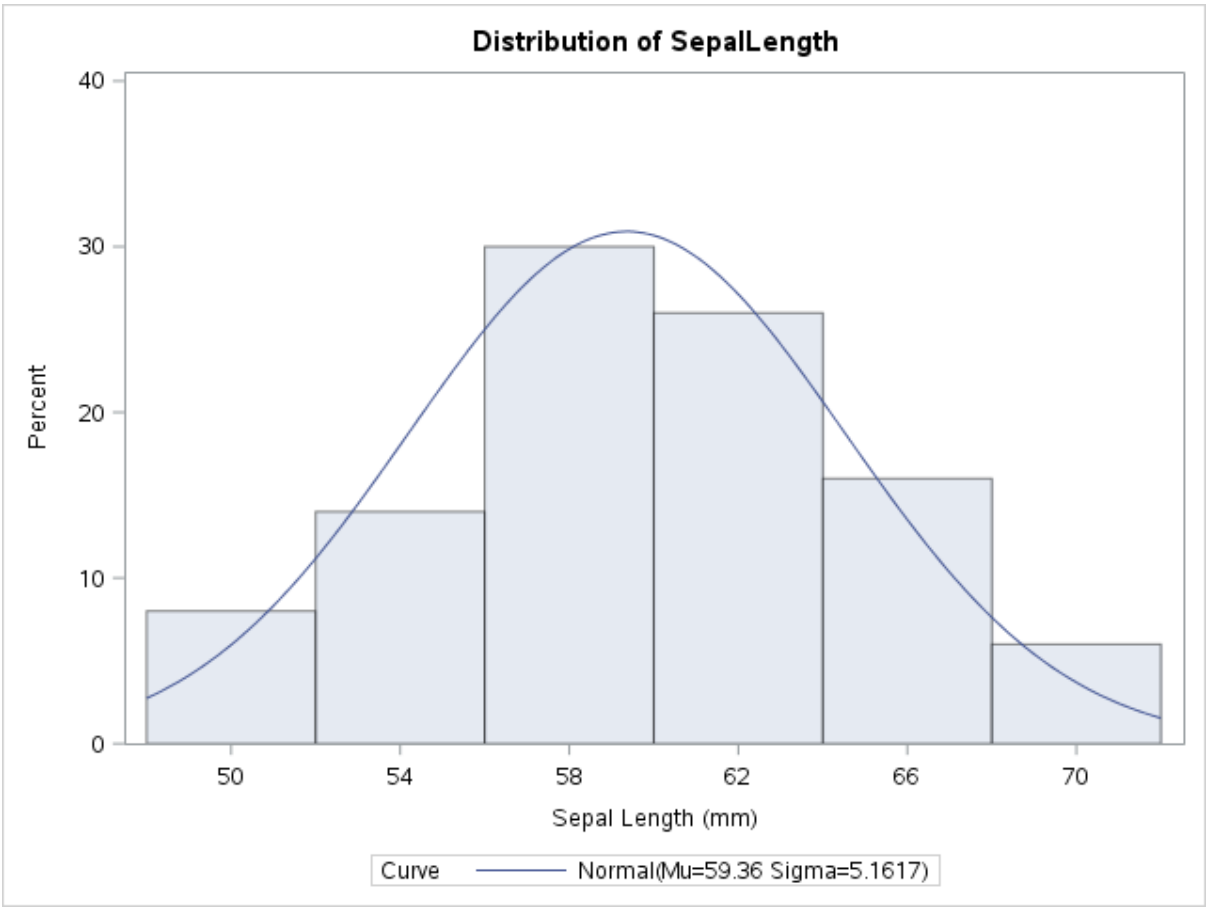
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.971179	**Pr < W**	0.2583
Kolmogorov-Smirnov	**D**	0.115034	**Pr > D**	0.0953
Cramer-von Mises	**W-Sq**	0.089467	**Pr > W-Sq**	0.1538
Anderson-Darling	**A-Sq**	0.551641	**Pr > A-Sq**	0.1506

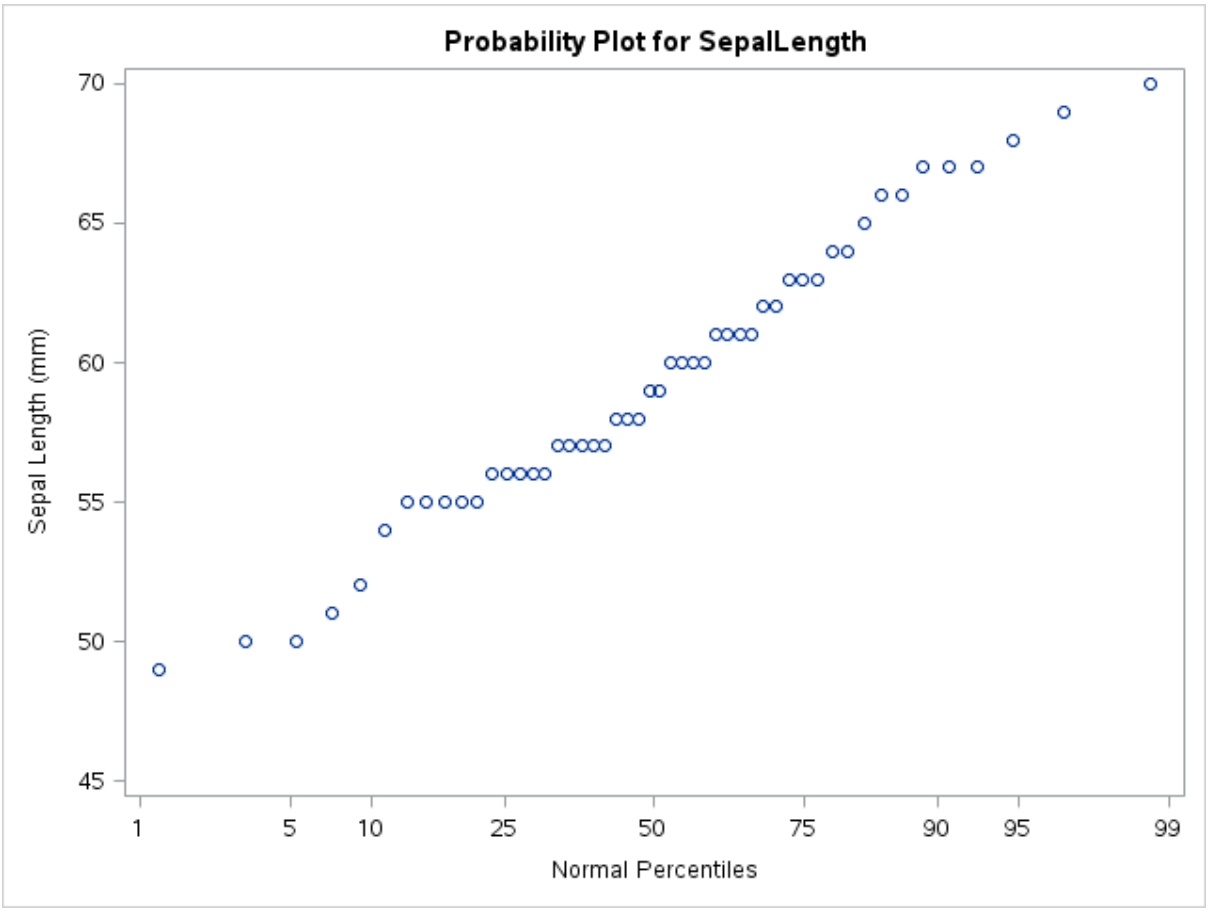
All the tests for all species show that the normality assumption is valid because p value > alpha in all cases.

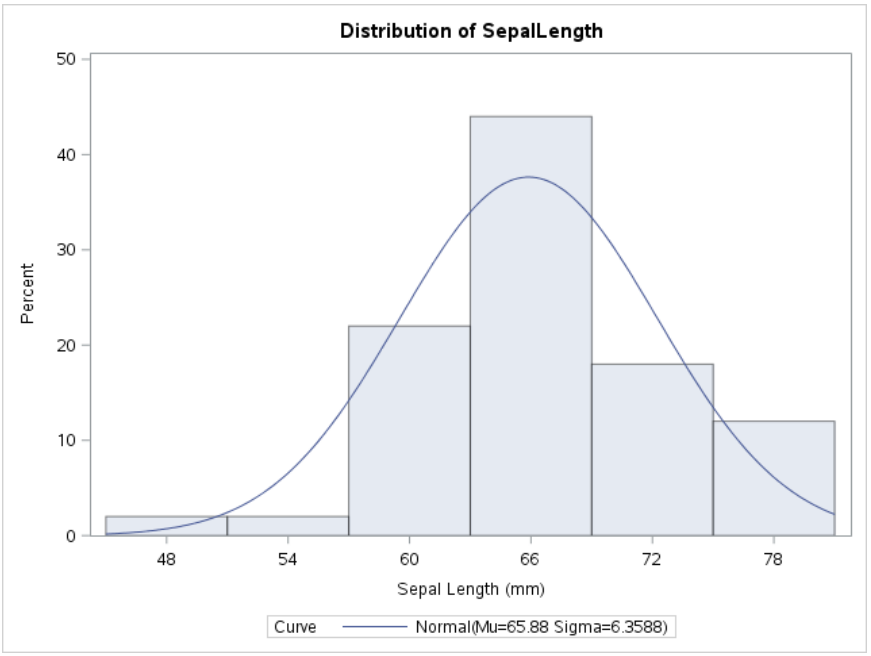
Plots in the same order: All plots show bell shaped curves supporting normality tests, hence all subsets can be assumed to be normal. The last plot for virginica seems a little skewed on the left but the tests support normality.

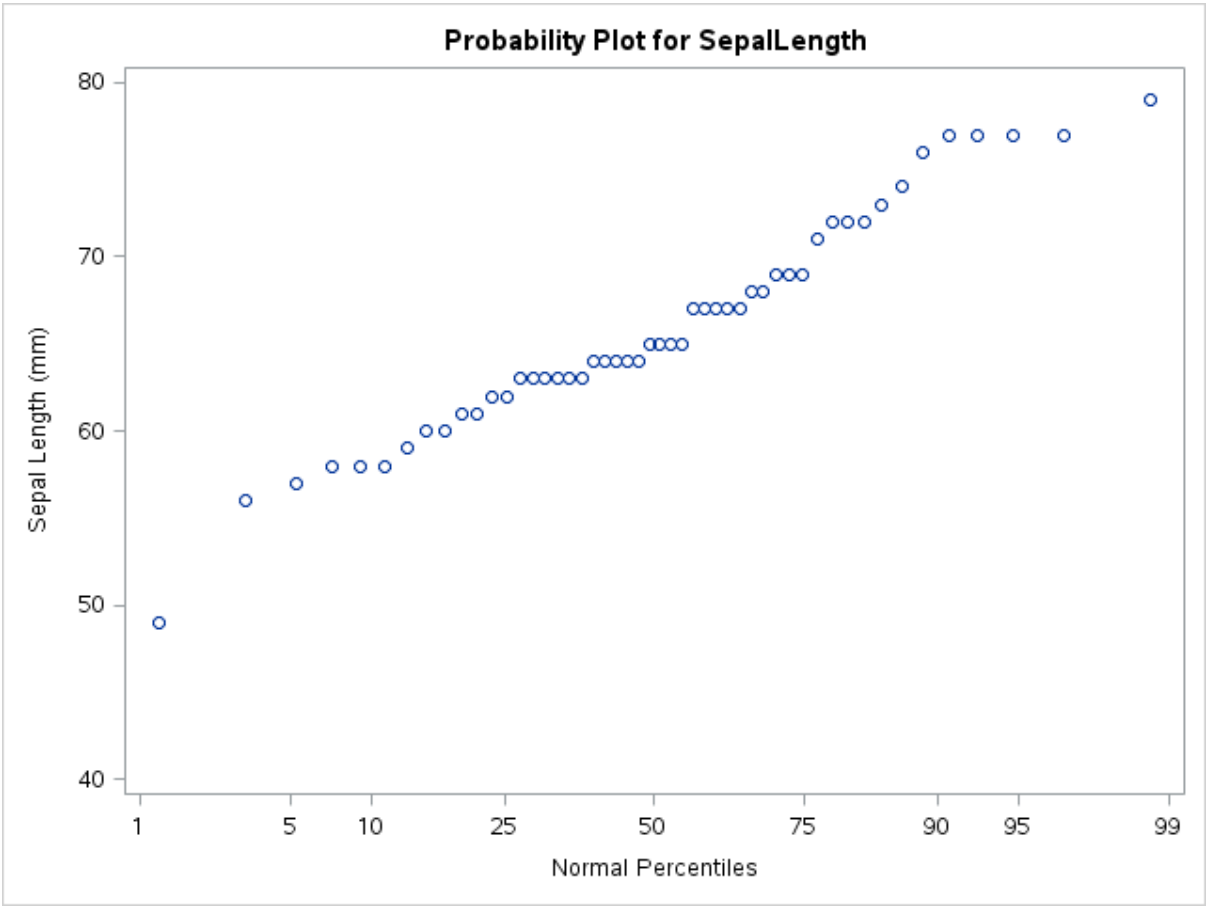





Distribution of SepalLength
Percent
40
30
20
10
0
50
54
58
62
66
70
Sepal Length (mm)
Curve
Normal(Mu=59.36 Sigma=5.1617)


Probability Plot for SepalLength
Sepal Length (mm)
70
65
60
55
50
45
1
5
10
25
50
75
90
95
99
Normal Percentiles


Distribution of SepalLength
Percent
50
40
30
20
10
0
48
54
60
66
72
78
Sepal Length (mm)
Curve
Normal(Mu=65.88 Sigma=6.3588)


Probability Plot for SepalLength
Sepal Length (mm)
80
70
60
50
40
1
5
10
25
50
75
90
95
99
Normal Percentiles

5. Comparing overall data with subsets:

The median, mean and mode of virginica and setosa for sepal length is greater than the population mean(58.4). The range of each subset is smaller than overall range of data. The variance of whole data is greater than for each of the subsets.

The plots look bell shaped for all the subsets as well as the all species dataset. The overall plot looks more dense. Each of the subsets are statistically significant for normality assumption but the all species set is not.

SECTION 2

1. **Signed rank p value is less than alpha so reject null hypothesis that mu=60**
   The previous section shows normality assumption rejected for overall data. So based on previous tests, assuming no symmetry, we should use the Signed rank test for location test.

2. Tests for Location: Mu0=60				
Test	**Statistic**		**p Value**	
Student's t	**t**	-2.31717	**Pr > |t|**	0.0219
Sign	**M**	-11	**Pr >= |M|**	0.0798
Signed Rank	**S**	-1238.5	**Pr >= |S|**	0.0129

2.To check if the Virginica's mean is significantly higher than entire population we must conduct one sided t test. This is because Virginica has a normal distribution proven earlier. The mean of entire population is 58.44. Thus, the question is :

Is the mean of virginica> 58.44?
Null hypothesis is that it is = 58.44.
Alternate hypothesis: The mean is larger than overall mean.

The TTEST Procedure
Variable: SepalLength (Sepal Length (mm))

DF	t Value	Pr > t
49	8.76	<.0001

We reject null in favour of the alternative that it is significantly larger than overall mean.

3.The setosa and versicolor sepal lengths comparison.

They both have normal distributions so t test is valid.
The null hypothesis is that the means are not significantly different.
Here we see that equality of variance is rejected and so we use Satterthwite Unequal variance method. P <0.0001 so reject null hypothesis in favour of alternative that they are significantly different.

3. Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	98	-10.52	<.0001
Satterthwaite	Unequal	86.538	-10.52	<.0001

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	49	49	2.14	0.0087

Section 3

1.Pearson correlation for entire data:
The null hypothesis is that there is no correlation. For the highlighted combinations, this is rejected and we state that there is significant correlation between the respective variables.

Pearson Correlation Coefficients, N = 150 Prob > |r| under H0: Rho=0				
	SepalLength	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength Sepal Length (mm)	1.00000	-0.11757 0.1519	0.87175 <.0001	0.81794 <.0001
SepalWidth Sepal Width (mm)	-0.11757 0.1519	1.00000	-0.42844 <.0001	-0.36613 <.0001
PetalLength Petal Length (mm)	0.87175 <.0001	-0.42844 <.0001	1.00000	0.96287 <.0001
PetalWidth Petal Width (mm)	0.81794 <.0001	-0.36613 <.0001	0.96287 <.0001	1.00000

The 5 significant correlations are:

The correlation between sepal length and petal length: positive, value = 0.87175, strong. sepal length increases
with petal length.

2. The correlation between petal width and sepal length, positive , value= 0.81794, strong. petal width increases with sepal length.

3. The correlation between petal length and sepal width, negative value= -0.42844, moderate strength for petal length to decease with sepal width.

4. The correlation between petal width and sepal width, negative, value =-0.36613, moderate strength of petal width to decrease with sepal width.

5. The correlation between petal width and petal length , positive value = 0.96287, a very strong strength of petal width to increase with petal length.

2. Correlation by species: Order: Setosa, Versicolo, Virginica

Pearson Correlation Coefficients, N = 50
Prob > |r| under H0: Rho=0

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength Sepal Length (mm)	1.00000	0.74255 <.0001	0.26718 0.0607	0.27810 0.0505
SepalWidth Sepal Width (mm)	0.74255 <.0001	1.00000	0.17770 0.2170	0.23275 0.1038
PetalLength Petal Length (mm)	0.26718 0.0607	0.17770 0.2170	1.00000	0.33163 0.0186
PetalWidth Petal Width (mm)	0.27810 0.0505	0.23275 0.1038	0.33163 0.0186	1.00000

Setosa: There are two statistically significant correlations:

1. The correlation between sepal length and sepal width , positive value= 0.74255; strong.

2. The correlation between petal length and petal width , positive value= 0.33163.

For Versicolor and Virginica all pairs of four measurements are significantly correlated, positive and moderate to strong.

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0				
	SepalLength	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength Sepal Length (mm)	1.00000	0.52591 <.0001	0.75405 <.0001	0.54646 <.0001
SepalWidth Sepal Width (mm)	0.52591 <.0001	1.00000	0.56052 <.0001	0.66400 <.0001
PetalLength Petal Length (mm)	0.75405 <.0001	0.56052 <.0001	1.00000	0.78667 <.0001
PetalWidth Petal Width (mm)	0.54646 <.0001	0.66400 <.0001	0.78667 <.0001	1.00000

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0				
	SepalLength	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength Sepal Length (mm)	1.00000	0.45723 0.0008	0.86422 <.0001	0.28111 0.0480
SepalWidth Sepal Width (mm)	0.45723 0.0008	1.00000	0.40104 0.0039	0.53773 <.0001
PetalLength Petal Length (mm)	0.86422 <.0001	0.40104 0.0039	1.00000	0.32211 0.0225
PetalWidth Petal Width (mm)	0.28111 0.0480	0.53773 <.0001	0.32211 0.0225	1.00000

3.The positive relationship between sepal width and sepal length is found in all three species subsets but not in overall sample. All five significant correlations in overall sample are also found in Versicolor and Virginica but only two of them are found in Setosa. All five are significant in versicolor and virginica but not in Setosa.