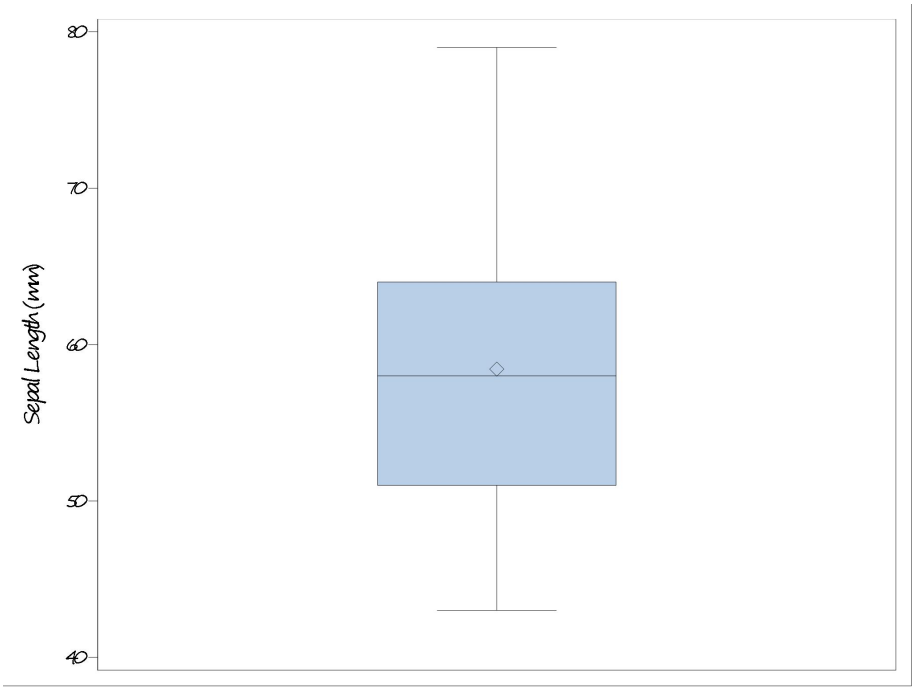
STAT 448 -Advanced Data Analysis

Homework 1

Name:Xinyan Yang NetID:xinyany2

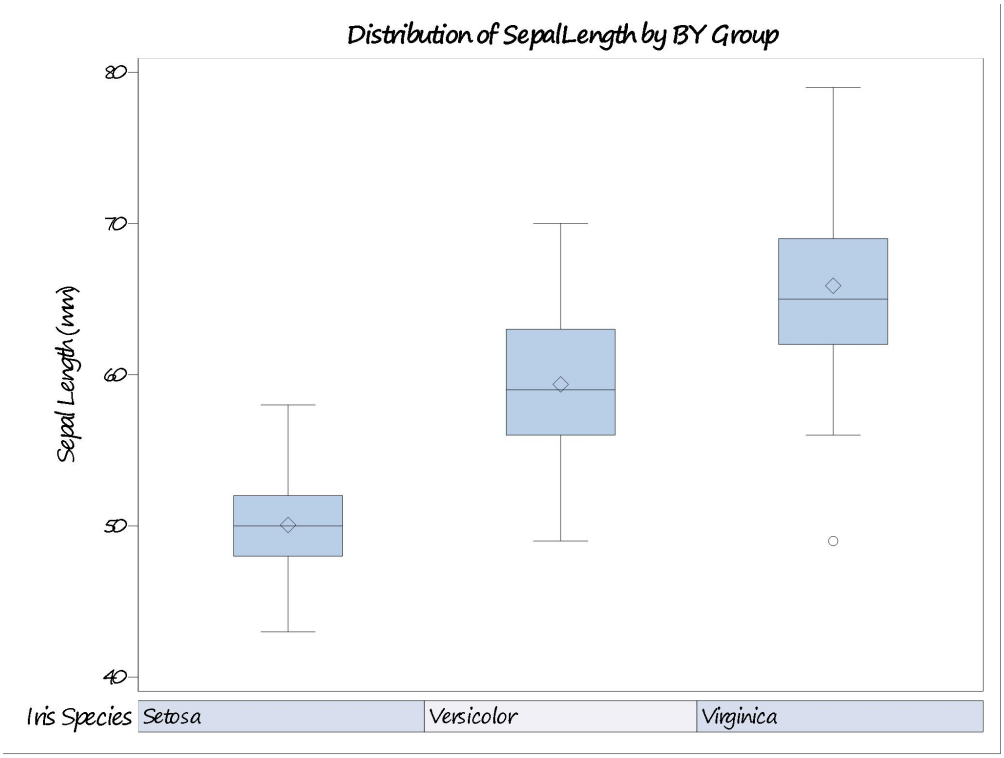
1. Descriptive Statistics Solution:

(a)



From the box plot, we can see that the median and mean of sepal length is around 57mm. Also, as the length becomes longer, the data has a larger variation. About 50% of the data has a length between 50-64mm.

(b)



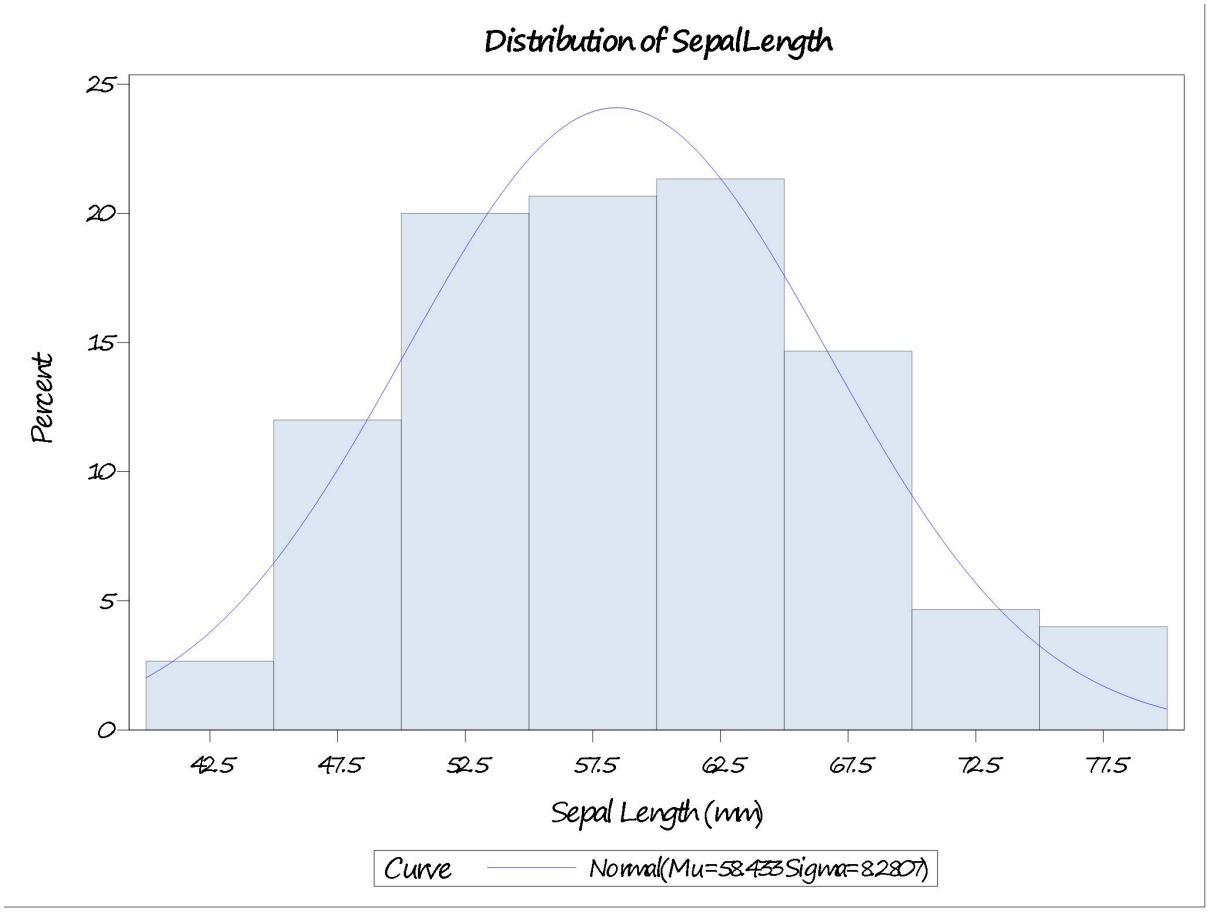
From the box plots above,we could see that the Virginica group has higher sepal length than the Versicolor group and the Versicolor group has higher sepal length than the Setosa group.
Also, the distance between the 25% quantile and the 75% quantile of the setosa is shorter than the other two species, so the distribution plot of this group will be more higher.

(c)

Moments			
N	150	**Sum Weights**	150
Mean	58.4333333	**Sum Observations**	8765
Std Deviation	8.28066128	**Variance**	68.5693512
Skewness	0.31491096	**Kurtosis**	-0.552064
Uncorrected SS	522385	**Corrected SS**	10216.8333
Coeff Variation	14.171126	**Std Error Mean**	0.67611316

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

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	79
99%	77
95%	73
90%	69
75% Q3	64
50% Median	58
25% Q1	51
10%	48
5%	46
1%	44
0% Min	43



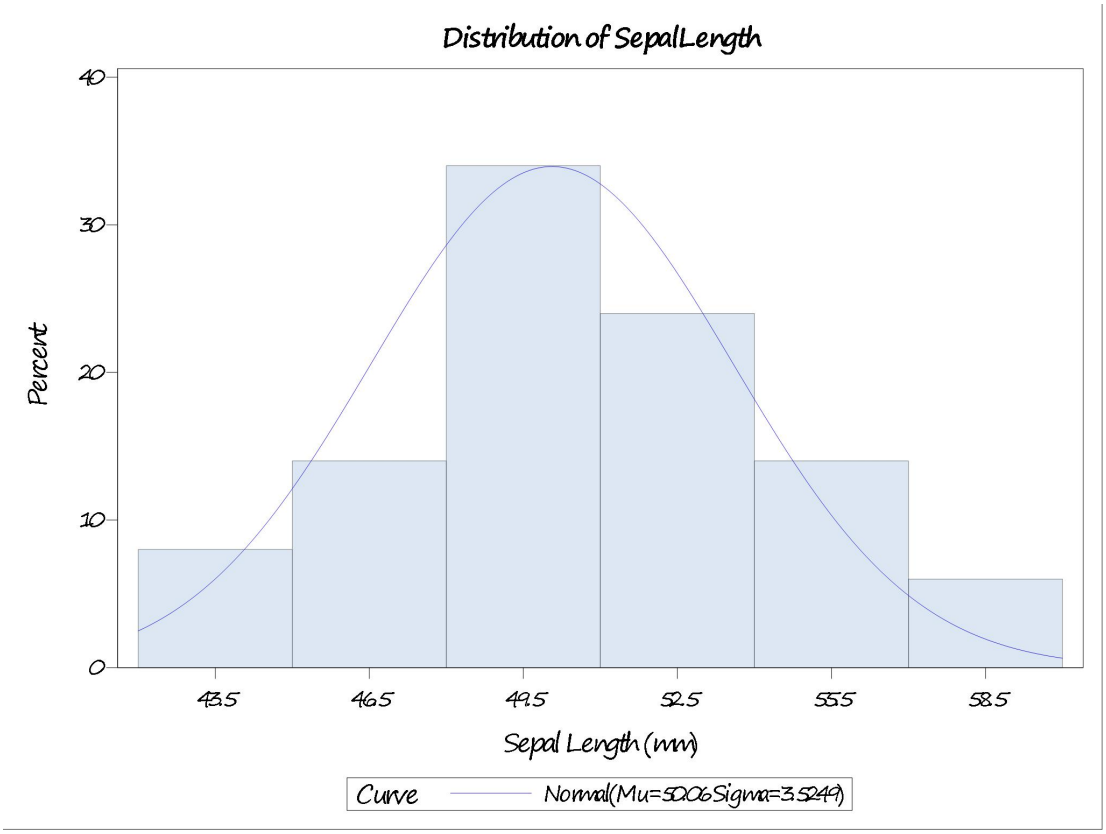
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.97609	**Pr < W**	0.0102
Kolmogorov-Smirnov	**D**	0.088654	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.127398	**Pr > W-Sq**	0.0479
Anderson-Darling	**A-Sq**	0.889199	**Pr > A-Sq**	0.0231

Frome the results above, we can conclude that the mean value of sepallength of all species is 58.43mm and the value range from 43mm to 79mm.
About the normality, qualitatively, from the histogram we can see that the data basically follows the normal distribution. And quantitatively, according to the results of Tests for Normality, we could reject the null hypothesis at the significance level of 5% and conclude that the data doesn't follow the normal distribution.

(d) **Iris Species=Setosa**

Moments			
N	50	**Sum Weights**	50
Mean	50.06	**Sum Observations**	2503
Std Deviation	3.52489687	**Variance**	12.424898
Skewness	0.12008699	**Kurtosis**	-0.2526888
Uncorrected SS	125909	**Corrected SS**	608.82
Coeff Variation	7.04134413	**Std Error Mean**	0.4984957

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

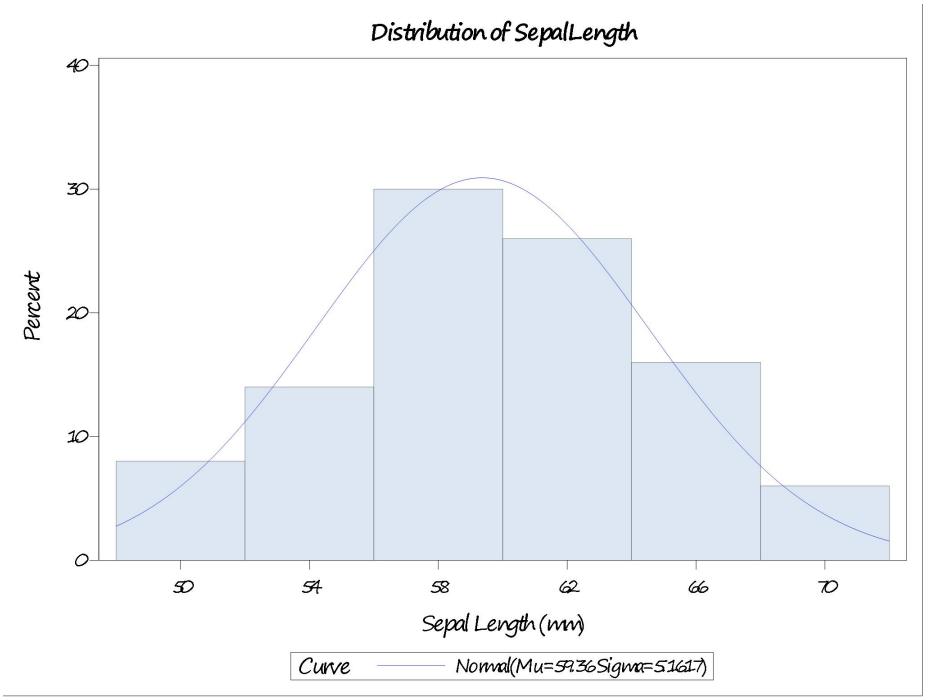


Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.977699	**Pr < W**	0.4595
Kolmogorov-Smirnov	**D**	0.11486	**Pr > D**	0.0962
Cramer-von Mises	**W-Sq**	0.071753	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.407986	**Pr > A-Sq**	>0.2500

Iris Species=Versicolor

Moments			
N	50	**Sum Weights**	50
Mean	59.36	**Sum Observations**	2968
Std Deviation	5.16171147	**Variance**	26.6432653
Skewness	0.10537762	**Kurtosis**	-0.5330095
Uncorrected SS	177486	**Corrected SS**	1305.52
Coeff Variation	8.69560558	**Std Error Mean**	0.72997624

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

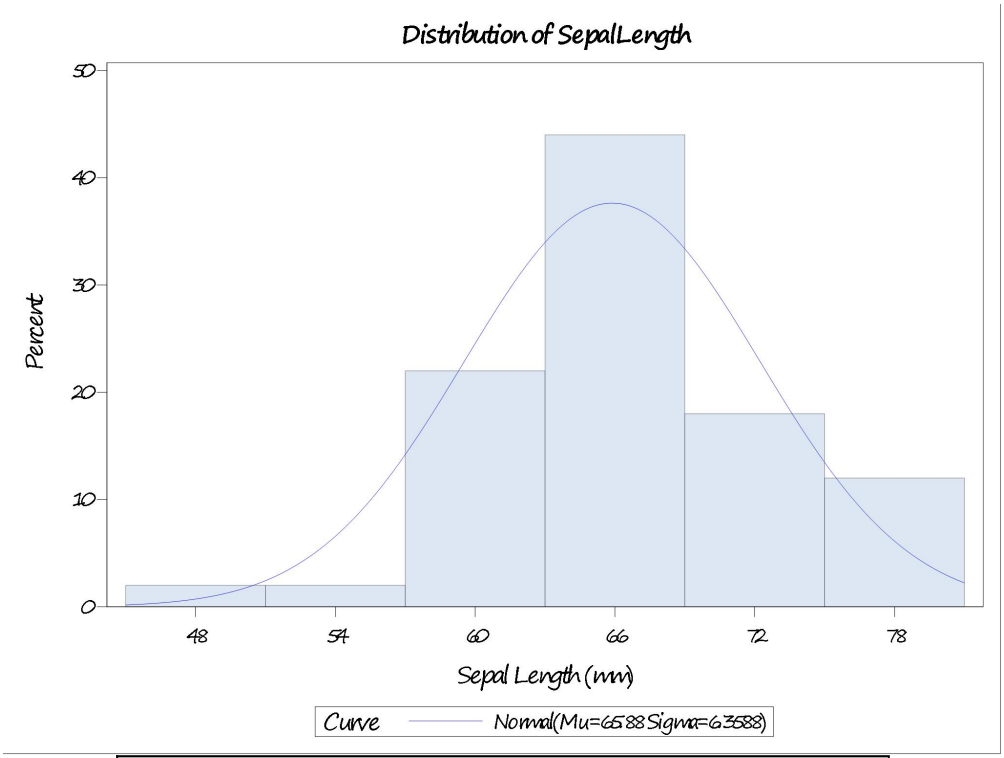


Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.977836	**Pr < W**	0.4647
Kolmogorov-Smirnov	**D**	0.096241	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.057273	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.360841	**Pr > A-Sq**	>0.2500

Iris Species=Virginica

Moments			
N	50	**Sum Weights**	50
Mean	65.88	**Sum Observations**	3294
Std Deviation	6.35879593	**Variance**	40.4342857
Skewness	0.11801512	**Kurtosis**	0.03290442
Uncorrected SS	218990	**Corrected SS**	1981.28
Coeff Variation	9.65208854	**Std Error Mean**	0.89926954

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



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

From the results above, we can find that as the group changes, not only does the group mean becomes larger, but the range, the skewness and the number of observations also increases, which means that the Virginica group tends to have a larger variation.

Both three Shapiro-Wilk tests give us a large p-value under which we couldn't reject null hypothsis and it means that both three groups follow the normal distribution. Although the Virginica group is a little right-tailed, it still shows normality.

(e)
Something unusual is that the species-wise statistics tell us that the group data shows normality

while in the last question, the overall data doesn't follow the normal distribution, which is very tricky.

2.Hypothesis Testing Solution:

(a)

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

First, the data doesn't show the normality.Second, although the skewness is greater than 0, the histogram plot actually shows that the data is symmetric. So we should choose the Signed Rank test for location . The p-value for the signed rank test is 0.0129, therefore under the significance level of 5%, we should reject the null hypothesis and conclude that the mean of sepal length is significantly different from 60.

(b)

The median value for sepal length of all species is 58 so we set h0=58 and perform the one-sided t test and get the following result.

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
65.8800	64.3723	Infty	6.3588	5.3117	7.9239

DF	t Value	Pr > t
49	8.76	<.0001

From the t-test result, since p-value is smaller than 0.001, we can conclude that under the significant level of 5%, we should reject the null hypothesis and think that the mean of virginica is significantly greater than the general population.

(c)

Since the p-value of the first table is 0.0087, we can reject the null hypothesis and think that the variances of these two groups are significantly different.

Therefore we can use the Satterthwaite test to test the mean of these two groups, since the p-value is less than 0.001, we can reject the null hypothesis and conclude that the mean of these two groups are significantly different.

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

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

3.Correlation Solution:

(a)

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

According to the results above, we notice that the correlation coefficient between the sepal length and the petal length, the correlation between the sepal length and the petal width, the one between the sepal width and the petal length, the correlation between the sepalwidth and the petal width, the correlation between the petal length and the petal width are all significant under the significant level of 5%.
In other words, except the correlation between the sepal length and the sepal width, there are relationship in all the other groups.
Among all these correlations, we can notice that the one between the petallength and the petalwidth is 0.96287, which is nearly 1. We can assume that there's some linear relationship between these two variables.

(b)

Iris Species=Setosa

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

Iris Species=Versicolor

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

Iris Species=Virginica

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

From the correlation matrix above, we can find the following results.
About the Setosa species, only the correlation between the sepal length and the sepal width is significant under the significance level of 0.1%.
About the Versicolor species, all correlation coefficients between these variables are significant.

About the Virginica species, the correlation between the sepallength and petallength and the correlation between the sepalwidth and the petalwidth are significant.

(c) We can find that the correlation matrix of Setosa species is totally different from that of the entire data set. While there's only one significant coefficient in the former, that coefficient is the only one that is not significant in the latter matrix, which is very surprising. And we can probabily infer that the setosa group has the least effect on the performance of the overall dataset. At the same time, we can find that the Versicolor group has a great effect on the overall data set.