Mini Project #3

Yanjiao Jia (Contribution 50%)

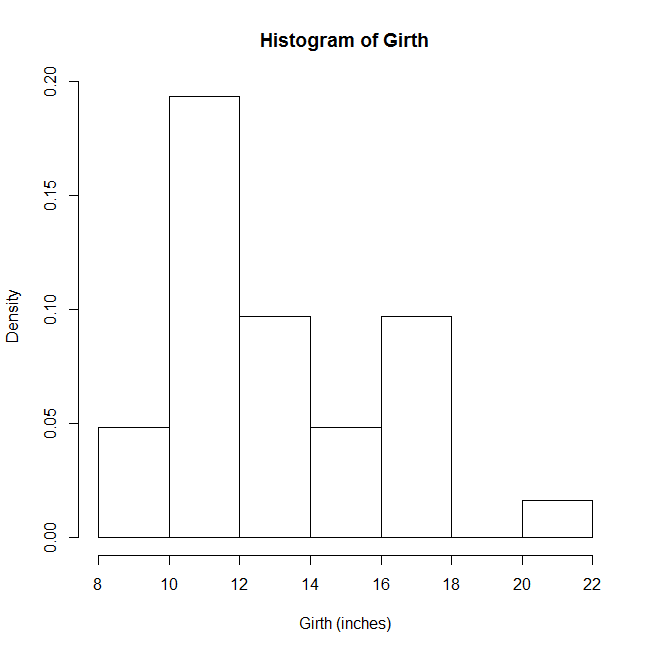
Yuzhuang Feng (Contribution 50%)

Object

In this project, we use R to make histogram, boxplot, 5-number summary, QQplot, scatterplot and correlation calculation to analysis the data of the three variables in *trees.csv* file. The variables are: girth (defined as diameter of the tree in inches measured at 4.5 feet above the ground); height (in feet); and volume of timber (in cubic feet).

Procedure and Analysis

1. We made a histogram and a boxplot, calculate the 5-number summary for each variable and decide the appropriate measure of (center, spread).

****

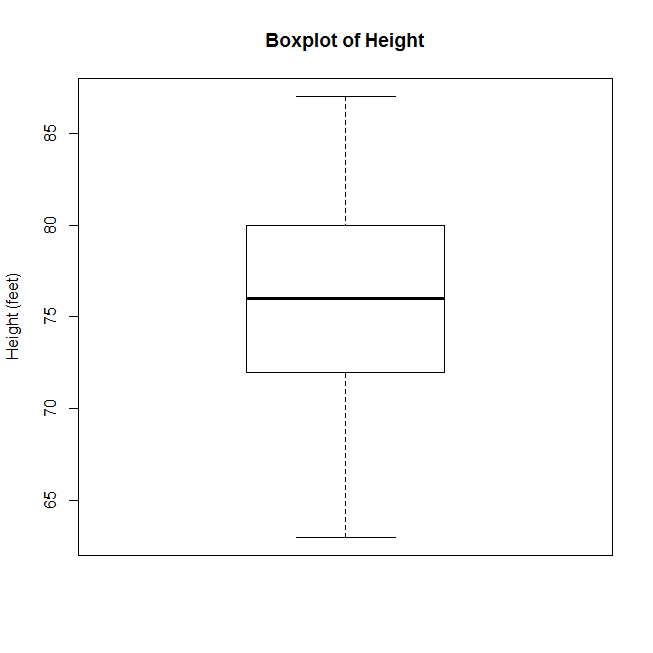
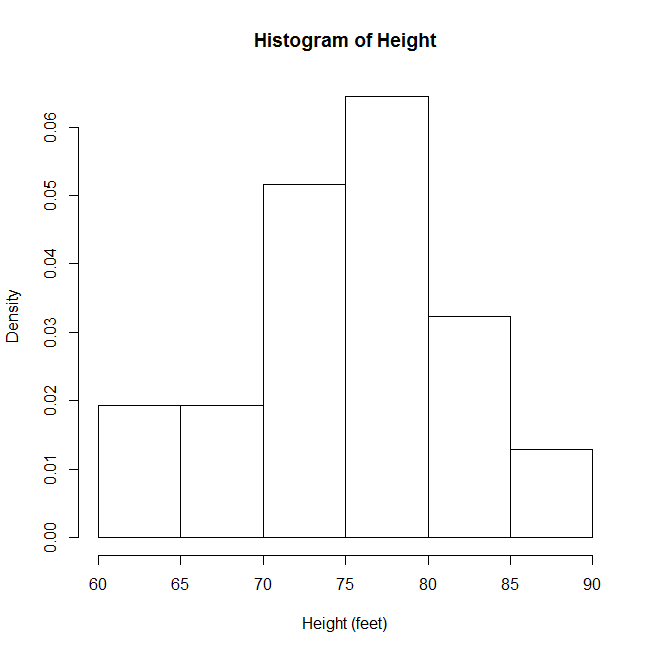
**Figure 1** Histogram and Boxplot of Girth

5-number summary of Girth

0% 25% 50% 75% 100%

8.30 11.05 12.90 15.25 20.60

From **Figure 1**, we can see the in the histogram the bins on the left side has higher density and in the boxplot more data stay on the lower part of the plot, so we can obtained that the distribution of data of girth is right-skewed. The 5-number summary also proofs that. Based on that, the appropriate measure of (center, spread) of girth is (median, IQR).



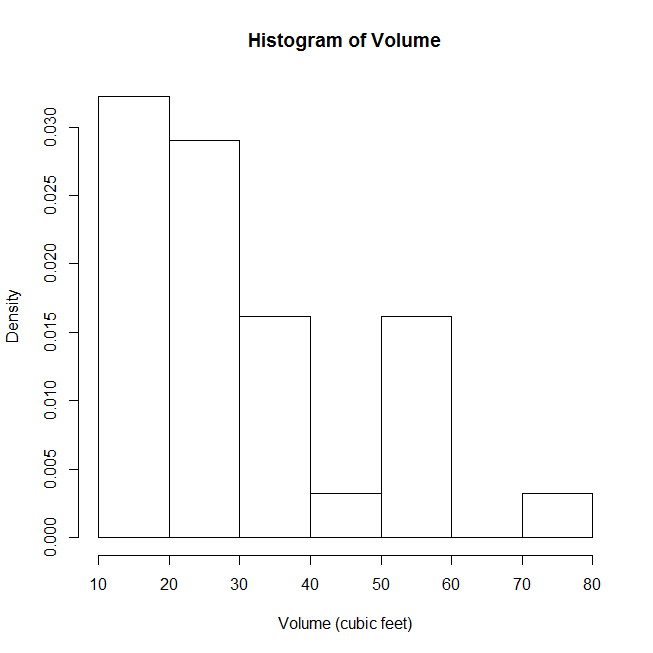
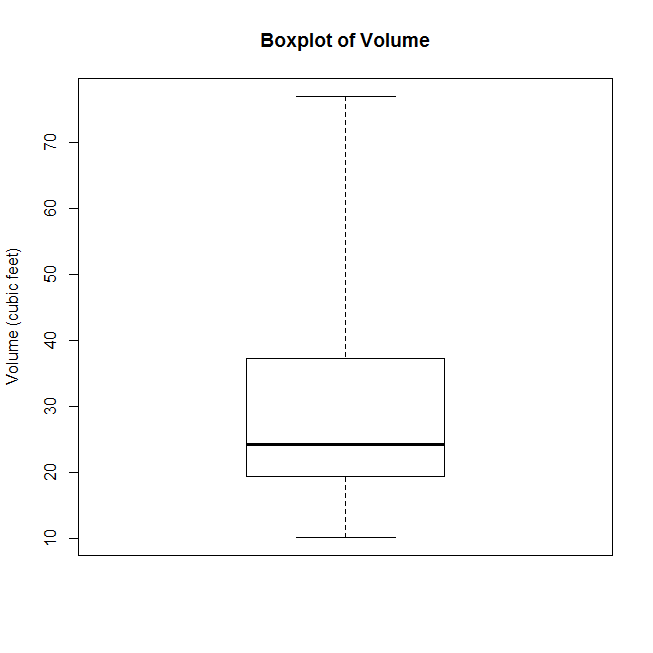
**Figure 2** Histogram and Boxplot of Height

5-number summary of Height

0% 25% 50% 75% 100%

63 72 76 80 87

From **Figure 2**, although we can see the in the histogram bins on the right side has higher density and in the boxplot more data stay on the upper part of the plot, the distribution of height is only a little bit left-skewed and we still think it is symmetric. The 5-number summary also show that we can take it as symmetric. Based on that, the appropriate measure of (center, spread) of height is (mean, SD).

**Figure 3** Histogram and Boxplot of Volume

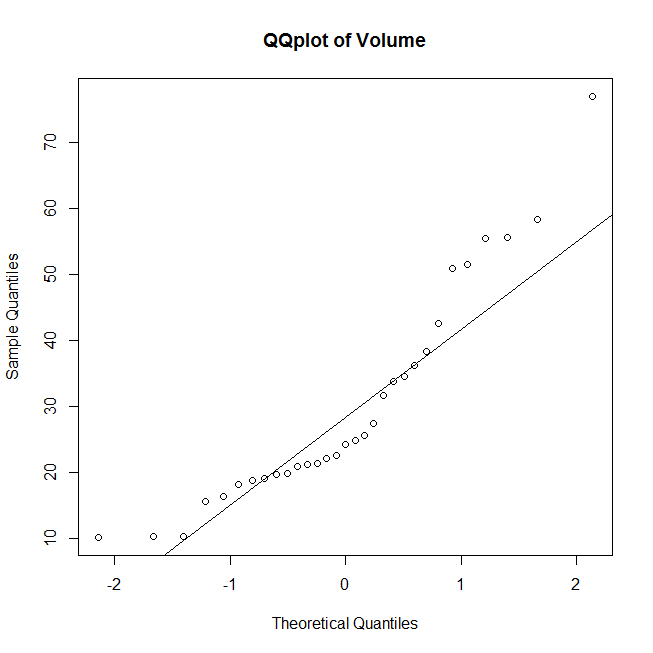
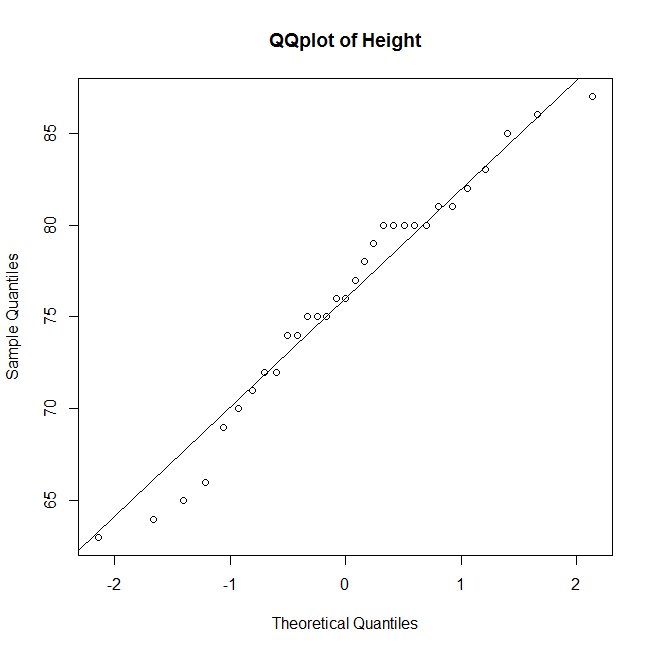
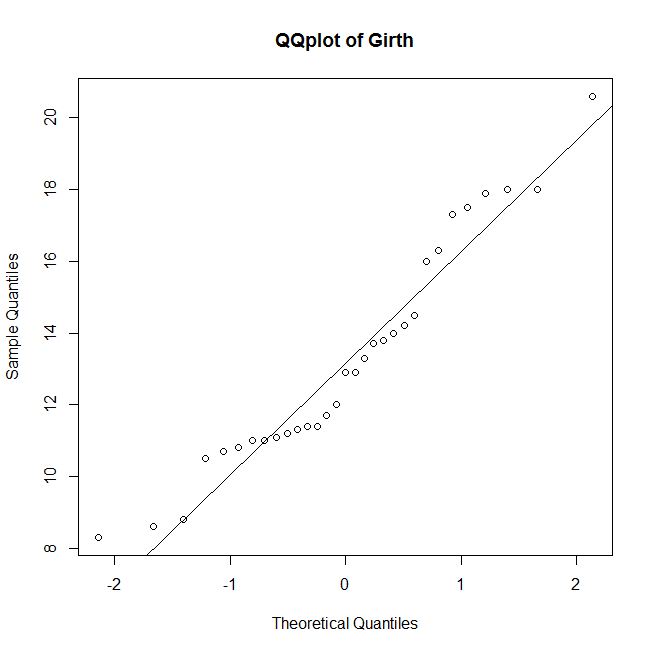
5-number summary of Volume

0% 25% 50% 75% 100%

10.2 19.4 24.2 37.3 77.0

From **Figure 3**, we can see the in the histogram the bins on the left side has higher density and in the boxplot more data stay on the lower part of the plot, the right-skewed situation of the distribution of data of volume is much more obvious than the distribution of data of girth. So we can obtained that the distribution of data of Volume is right-skewed. The 5-number summary also proofs that. Based on that, the appropriate measure of (center, spread) of Volume is (median, IQR).

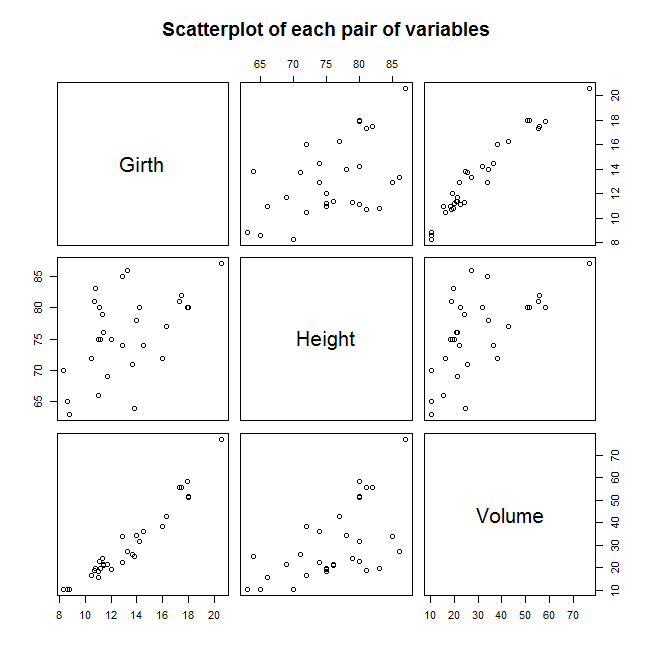
2. We made make a normal Q-Q plot, and analyze whether or not the normal distribution assumption for the data appears reasonable.



**Figure 4** QQplot of Girth, Height and Volume

From **Figure 4**, we can see the dots on all three QQplot are evenly distributed on both sides of the QQline we added on the plot, so we can obtained that the normal distribution assumption for the data of the three variables all appears reasonable.

3. We make a scatterplot for each pair of variables, analyze whether or not the correlation is an appropriate measure of strength of relationship between the two variables.



**Figure 5** Scatterplot for Each Pair of Variables

Correlation between Height and Girth

[1] 0.5192801

Correlation between Volume and Girth

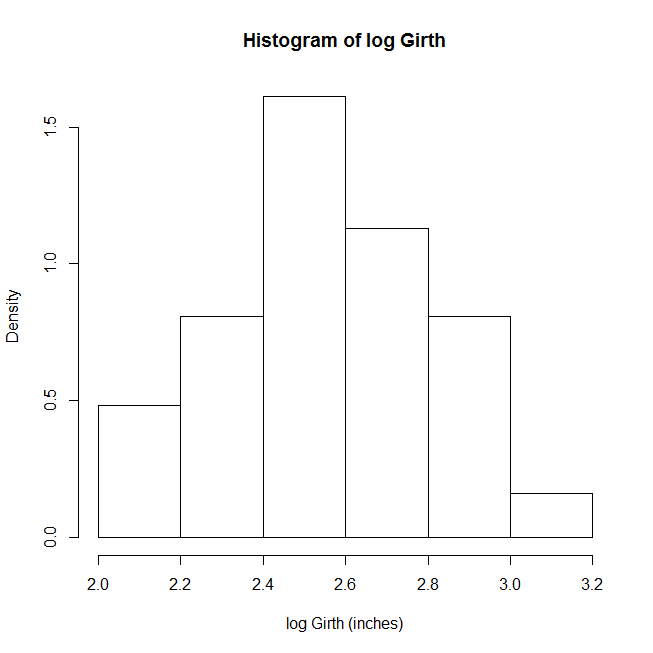
[1] 0.9671194

Correlation between Height and Volume

[1] 0.5982497

From **Figure 5**, we can see that height and girth has liner relationship which is not very strong, volume and girth has strong liner relationship, and volume and height also has moderately strong which is stronger than the liner relationship between height and girth. And the calculation of the correlations of each pair of variables is also accord with the situation we observed from **Figure 5**. So we can obtained that the correlation is an appropriate measure of strength of relationship between the two variables.

4. We made a histogram and a boxplot, calculate the 5-number summary for each variable after the log transformation and decide the appropriate measure of (center, spread).



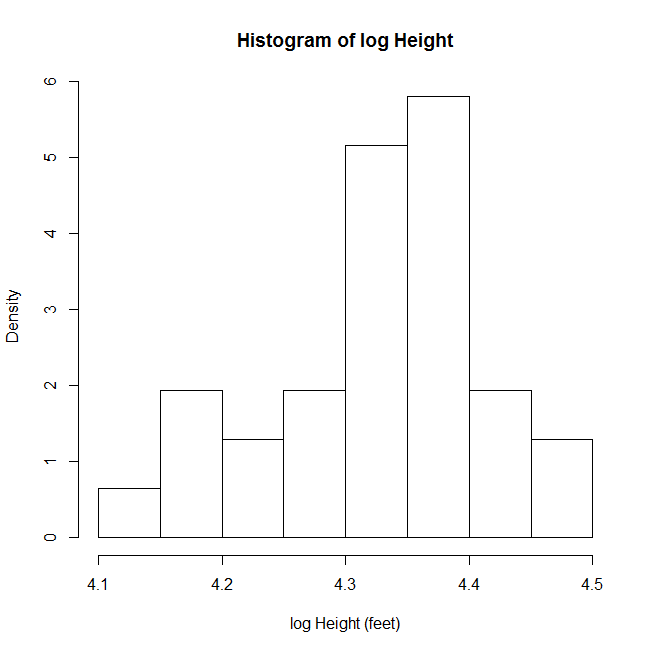
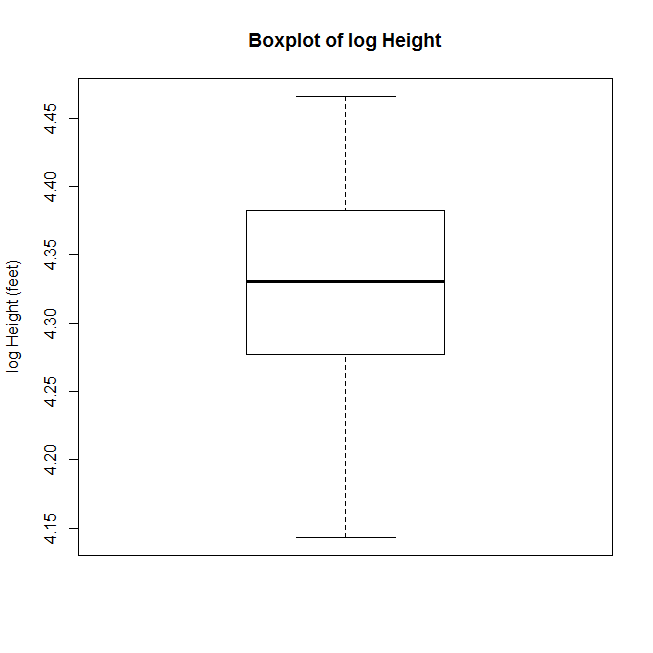
**Figure 6** Histogram and Boxplot of log Girth

5-number summary of log Girth

0% 25% 50% 75% 100%

2.116256 2.402420 2.557227 2.723369 3.025291

From **Figure 6**, we can see whichever the histogram or the boxplot looks symmetric, so we can obtained that the distribution of data of girth is symmetric. The 5-number summary also show that we can take it as symmetric. Based on that, the appropriate measure of (center, spread) of girth is (mean, SD).).

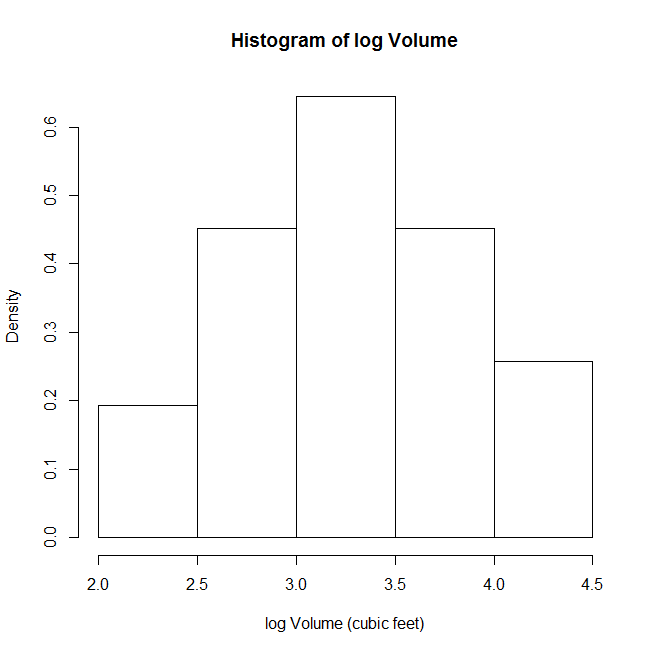
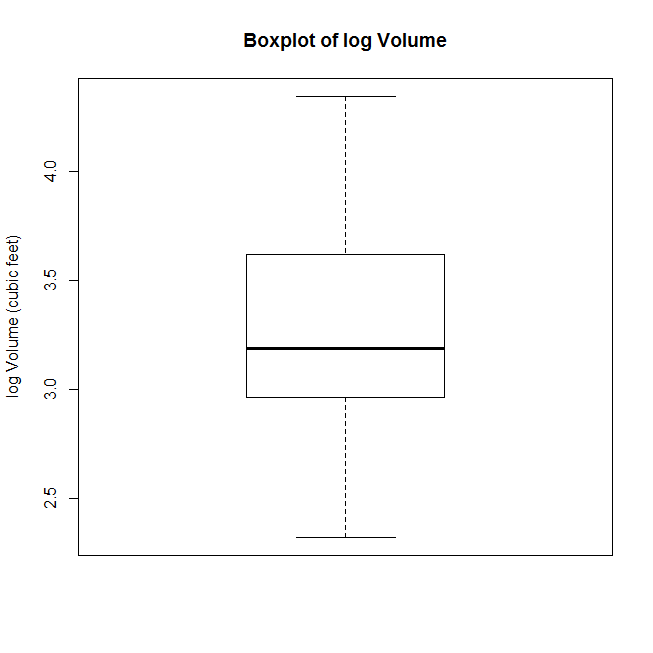
**Figure 7** Histogram and Boxplot of log Height

5-number summary of log Height

0% 25% 50% 75% 100%

4.143135 4.276666 4.330733 4.382027 4.465908

From **Figure 7**, although we can see the in the histogram bins on the right side has higher density and in the boxplot more data stay on the upper part of the plot, we obtained that the distribution of the data of height is left-skewed. The 5-number summary also show that. Based on that, the appropriate measure of (center, spread) of height is (median, IQR).

**Figure 8** Histogram and Boxplot of log Volume

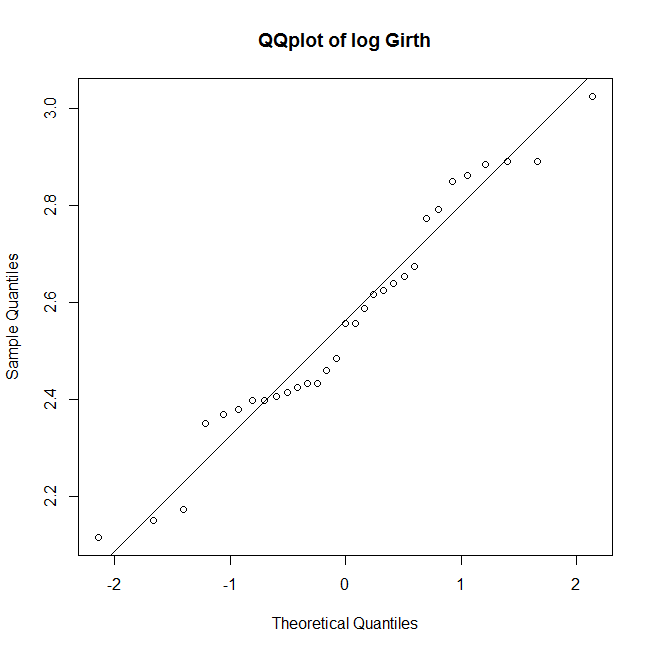
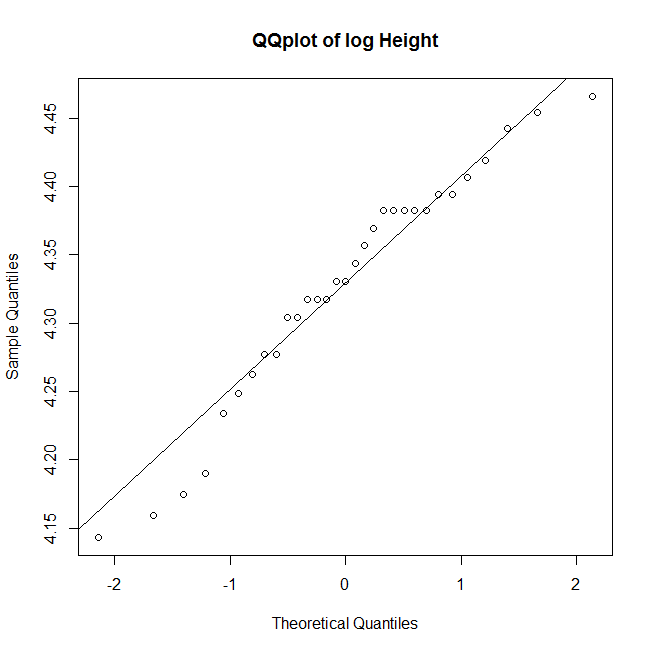
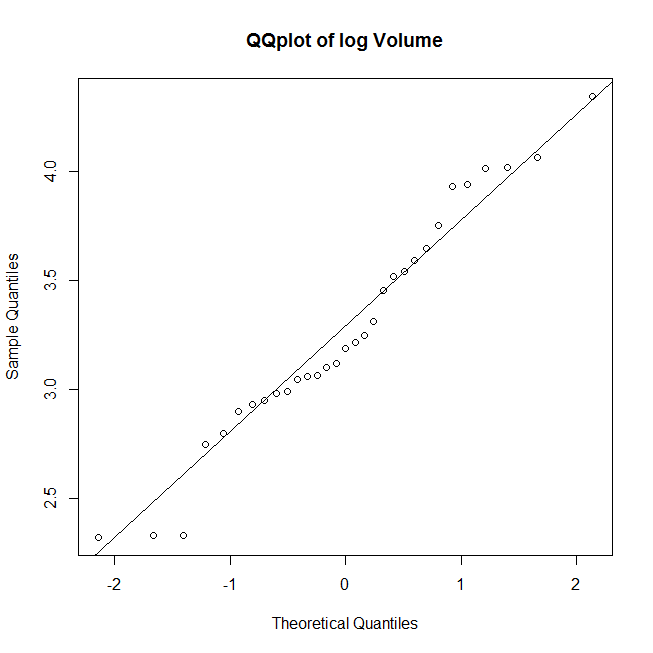
5-number summary of log Volume

0% 25% 50% 75% 100%

2.322388 2.965153 3.186353 3.618634 4.343805

From **Figure 8**, although we can see the in the histogram the distribution looks symmetric but in the boxplot it is right-skewed. And the 5-number summary also show that there is more data of the lower height value, which also shows the distribution is right-skewed. Based on that, the appropriate measure of (center, spread) of height is (median, IQR).

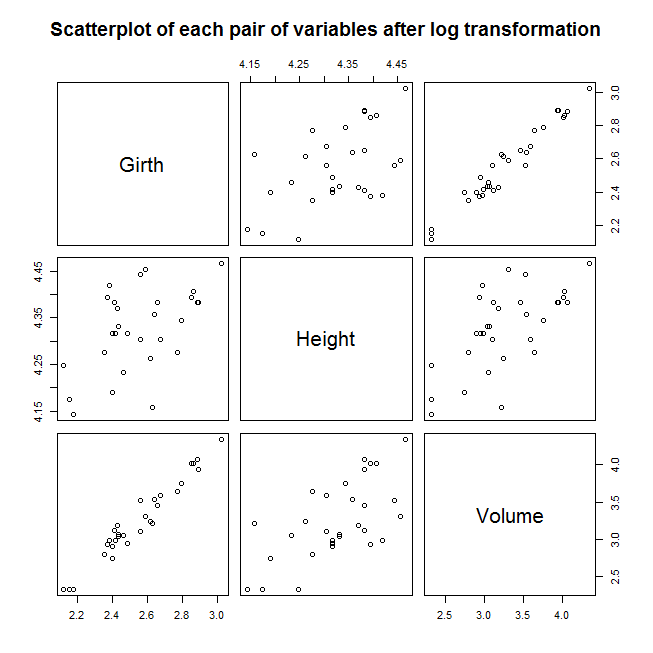
5. We made make a normal Q-Q plot for the data after the log transformation , and analyze whether or not the normal distribution assumption for the data appears reasonable.

**Figure 9** QQplot of Girth, Height and Volume

From **Figure 9**, we can see after the log transforamation, the dots on all three QQplot are still evenly distributed on both sides of the QQline we added on the plot, so we can obtained that the normal distribution assumption for the data of the three variables all appears reasonable.

6. We make a scatterplot for each pair of variables after the log transformation, analyze whether or not the correlation is an appropriate measure of strength of relationship between the two variables.



**Figure 10** Scatterplot for Each Pair of Variables

Correlation between Height and Girth after log transformation

[1] 0.5301949

Correlation between Volume and Girth after log transformation

[1] 0.9766649

Correlation between Height and Volume after log transformation

[1] 0.6486377

From **Figure 10**, we can see that after the log transformation, the situation is almost the same, height and girth has liner relationship which is not very strong, volume and girth has strong liner relationship, and volume and height also has moderately strong which is stronger than the liner relationship between height and girth. And the calculation of the correlations of each pair of variables is also accord with the situation we observed from **Figure 10**. So we can obtained that the correlation is an appropriate measure of strength of relationship between the two variables.

7. Should the data be analyzed on log scale or original scale?

From our group’s perspective, the data should be analyzed on log scale based on the following three reasons:

1. Log transformation make the variability of the data more obvious. As we can see, in the analysis of the plot and 5-number summary of the data after the log transformation we obtained all three the distribution of all three variables are all not symmetric. While at the first time when we do the analysis of the original data. The situation of being not symmetric is not as obvious as the second time.
2. When we make the QQplot, we can see the plot of the data on log scale is more close to the QQline we added compared to the QQplot of original data.
3. For the correlation calculation of the data after the log transformation, the outcome is larger than the outcome of the original data.

In one word, the log transformation make the characters of the data more obvious, more easier for us to observe and analyze.

Code

#Get the tree data from the csv file

tree <- read.table("trees.csv", header = TRUE, sep = ",")

#Make a histogram and a boxplot of each variables

windows();hist(tree$Girth,xlab='Girth (inches)',main='Histogram of Girth',freq = F)

windows();boxplot(tree$Girth,ylab='Girth (inches)',main='Boxplot of Girth',range = 0)

windows();hist(tree$Height,xlab='Height (feet)',main='Histogram of Height',freq = F)

windows();boxplot(tree$Height,ylab='Height (feet)',main='Boxplot of Height',range = 0)

windows();hist(tree$Volume,xlab='Volume (cubic feet)',main='Histogram of Volume',freq = F)

windows();boxplot(tree$Volume,ylab='Volume (cubic feet)',main='Boxplot of Volume',range = 0)

#Calculate the 5-number summary of each variables

cat('5-number summary of Girth\n')

print(quantile(tree$Girth))

cat('5-number summary of Height\n')

print(quantile(tree$Height))

cat('5-number summary of Volume\n')

print(quantile(tree$Volume))

#Make a normal Q-Q plot of each variables

windows();qqnorm(tree$Girth,main='QQplot of Girth')

qqline(tree$Girth)

windows();qqnorm(tree$Height,main='QQplot of Height')

qqline(tree$Height)

windows();qqnorm(tree$Volume,main='QQplot of Volume')

qqline(tree$Volume)

#Make a scatterplot for each pair of variables

windows();pairs(tree,main='Scatterplot of each pair of variables')

#Calculate the correlation for each pair of variables

cat('Correlation between Height and Girth\n')

print(cor(tree$Height,tree$Girth))

cat('Correlation between Volume and Girth\n')

print(cor(tree$Volume,tree$Girth))

cat('Correlation between Height and Volume\n')

print(cor(tree$Height,tree$Volume))

#Perform a natural log transformation of the variables

#Then repeat all the calculations using the data after log transformation

tree<-log(tree)

#Make a histogram and a boxplot of each variables after log transformation

windows();hist(tree$Girth,xlab='log Girth (inches)',main='Histogram of log Girth',freq = F)

windows();boxplot(tree$Girth,ylab='log Girth (inches)',main='Boxplot of log Girth',range = 0)

windows();hist(tree$Height,xlab='log Height (feet)',main='Histogram of log Height',freq = F)

windows();boxplot(tree$Height,ylab='log Height (feet)',main='Boxplot of log Height',range = 0)

windows();hist(tree$Volume,xlab='log Volume (cubic feet)',main='Histogram of log Volume',freq = F)

windows();boxplot(tree$Volume,ylab='log Volume (cubic feet)',main='Boxplot of log Volume',range = 0)

#Calculate the 5-number summary of each variables after log transformation

cat('5-number summary of log Girth\n')

print(quantile(tree$Girth))

cat('5-number summary of log Height\n')

print(quantile(tree$Height))

cat('5-number summary of log Volume\n')

print(quantile(tree$Volume))

#Make a normal Q-Q plot of each variables after log transformation

windows();qqnorm(tree$Girth,main='QQplot of log Girth')

qqline(tree$Girth)

windows();qqnorm(tree$Height,main='QQplot of log Height')

qqline(tree$Height)

windows();qqnorm(tree$Volume,main='QQplot of log Volume')

qqline(tree$Volume)

#Make a scatterplot for each pair of variables after log transformation

windows();pairs(tree,main='Scatterplot of each pair of variables after log transformation')

#Calculate the correlation for each pair of variables after log transformation

cat('Correlation between Height and Girth after log transformation\n')

print(cor(tree$Height,tree$Girth))

cat('Correlation between Volume and Girth after log transformation\n')

print(cor(tree$Volume,tree$Girth))

cat('Correlation between Height and Volume after log transformation\n')

print(cor(tree$Height,tree$Volume))