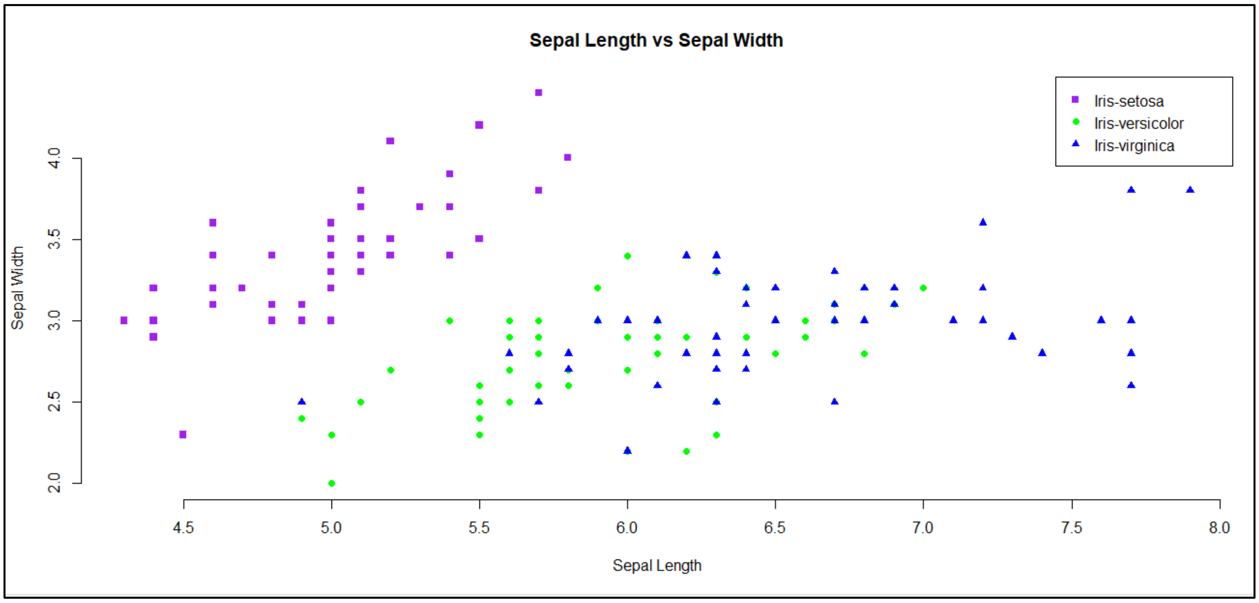
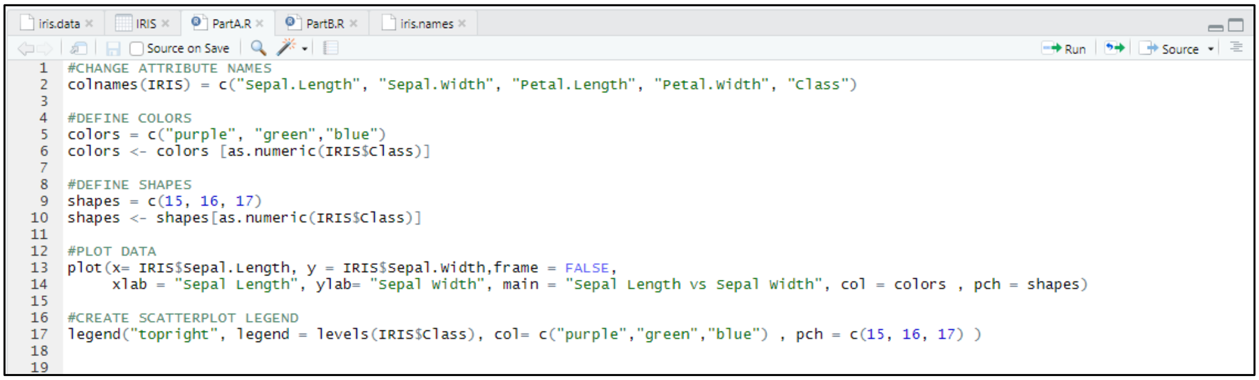
***** R was used for the analysis of the seeds dataset. The code to obtain the plots and histograms is pasted below each graph.***

A. Scatterplot for Sepal Variables



Code for scatterplot

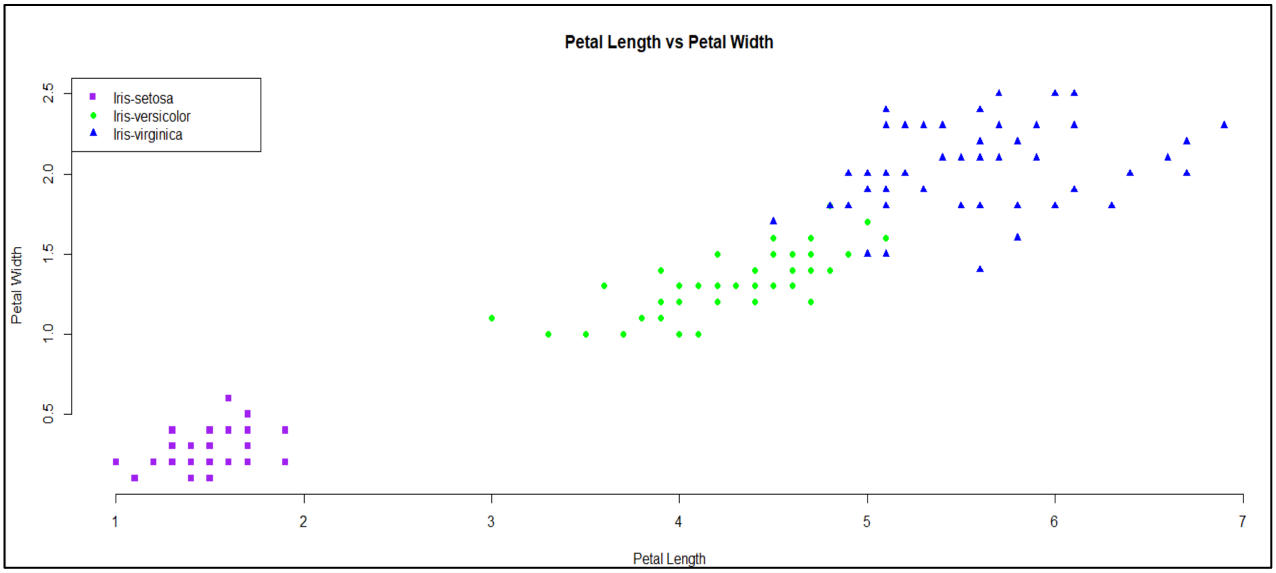


```
#CHANGE ATTRIBUTE NAMES
colnames(IRIS) = c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width", "Class")

#DEFINE COLORS
colors = c("purple", "green","blue")
colors <- colors [as.numeric(IRIS$Class)]

#DEFINE SHAPES
shapes = c(15, 16, 17)
shapes <- shapes[as.numeric(IRIS$Class)]

#PLOT DATA
plot(x= IRIS$Sepal.Length, y = IRIS$Sepal.Width,frame = FALSE,
     xlab = "Sepal Length", ylab= "Sepal Width", main = "Sepal Length vs Sepal Width", col = colors , pch = shapes)

#CREATE SCATTERPLOT LEGEND
legend("topright", legend = levels(IRIS$Class), col= c("purple","green","blue") , pch = c(15, 16, 17) )


```

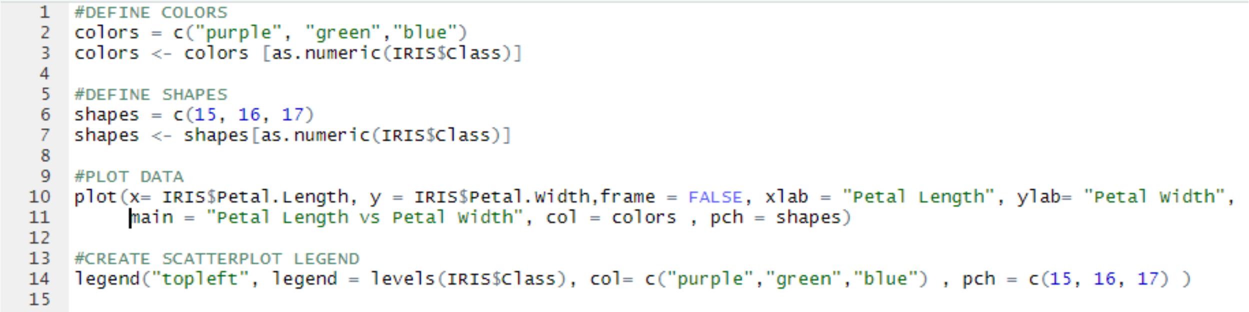
Data interpretation

The plot for the Iris-setosa class appears be more isolated and contains flowers with a wider sepal width and shorter sepal length than the other classes present in the data set. With respect to the sepal width and length variables it appears that a classification algorithm may be useful in classifying the iris-setosa class only. The other two classes appear to have some similar sepal lengths and widths which makes it harder to distinguish between the two. So, a classification algorithm to classify the classes present in the data set may not be successful when using these two variables.

B. Scatterplot for Petal Variables



Code for Petal Scatterplot

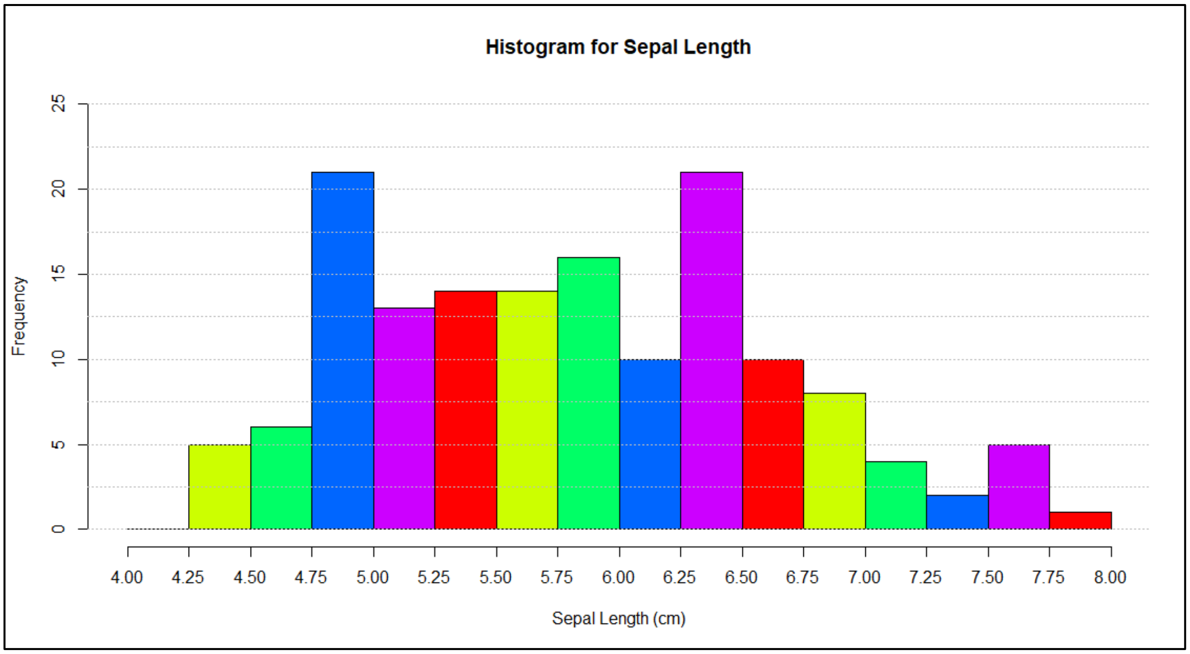


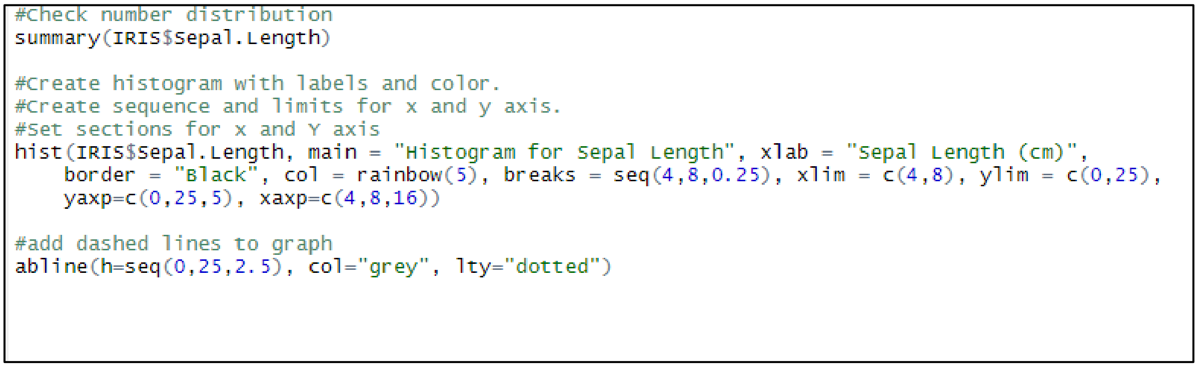
```
#DEFINE COLORS
colors = c("purple", "green","blue")
colors <- colors [as.numeric(IRIS$Class)]

#DEFINE SHAPES
shapes = c(15, 16, 17)
shapes <- shapes[as.numeric(IRIS$Class)]

#PLOT DATA
plot(x= IRIS$Petal.Length, y = IRIS$Petal.Width,frame = FALSE, xlab = "Petal Length", ylab= "Petal Width",
     main = "Petal Length vs Petal Width", col = colors , pch = shapes)

#CREATE SCATTERPLOT LEGEND
legend("topleft", legend = levels(IRIS$Class), col= c("purple","green","blue") , pch = c(15, 16, 17) )

```

Data Interpretation

The scatterplot with respect to the petal width and length appears to offer more separation of the classes. The iris-setosa class is completely distinguishable from the other two classes and iris-versicolor and iris-virginica are shown with very minimal overlap. These two variables may render a more successful result when trying to classify the data using a classification algorithm.

C. **Histograms of 4 variables**

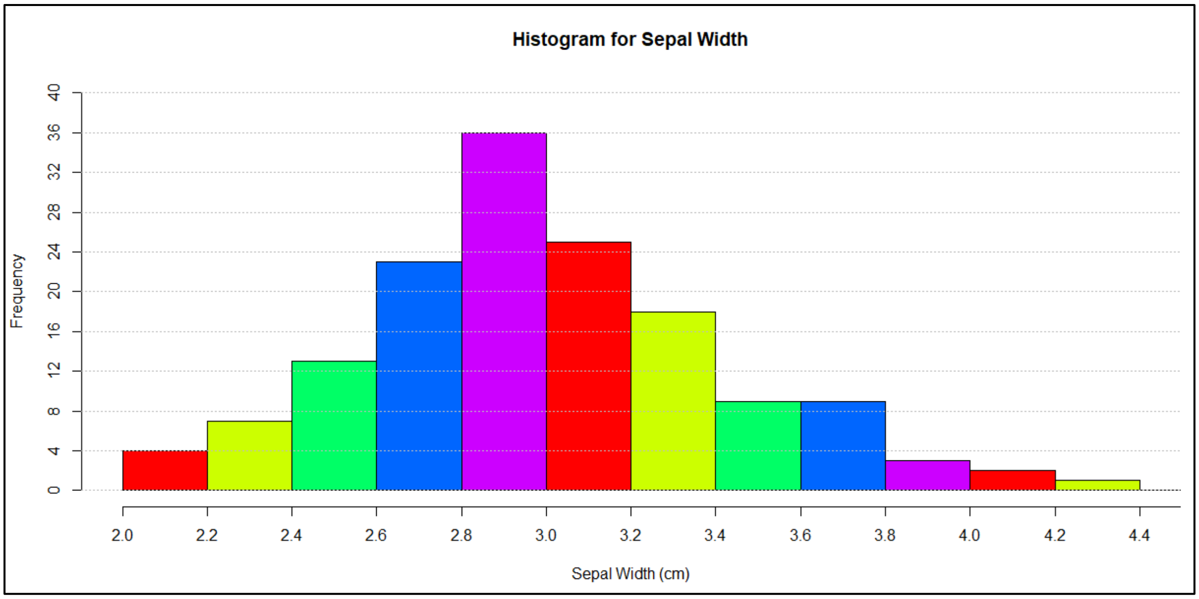


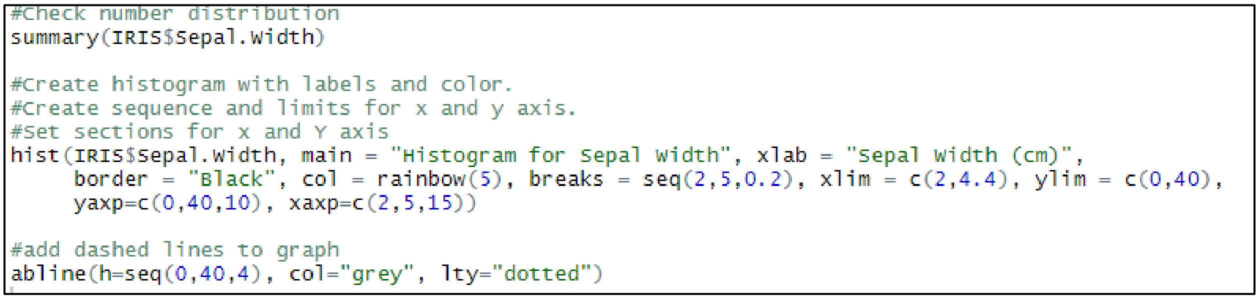


```
#Check number distribution
summary(IRIS$Sepal.Length)

#Create histogram with labels and color.
#Create sequence and limits for x and y axis.
#Set sections for x and Y axis
hist(IRIS$Sepal.Length, main = "Histogram for Sepal Length", xlab = "Sepal Length (cm)",
    border = "Black", col = rainbow(5), breaks = seq(4,8,0.25), xlim = c(4,8), ylim = c(0,25),
    yaxp=c(0,25,5), xaxp=c(4,8,16))

#add dashed lines to graph
abline(h=seq(0,25,2.5), col="grey", lty="dotted")
```

Interpretation

The distribution of the data appears to be wide-spread with a poor distribution fit. Similar frequencies appear between 4.75 to 6.50 on the histogram making the data non-symmetric. A slight skewness to the right appears.

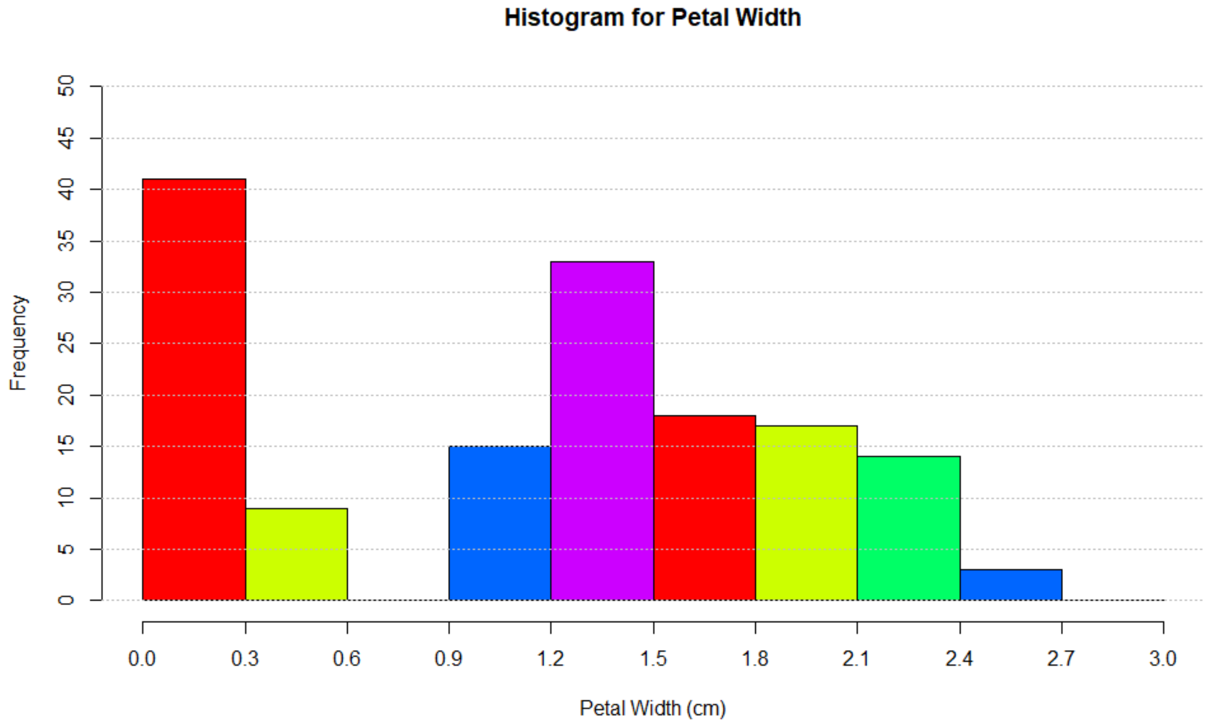


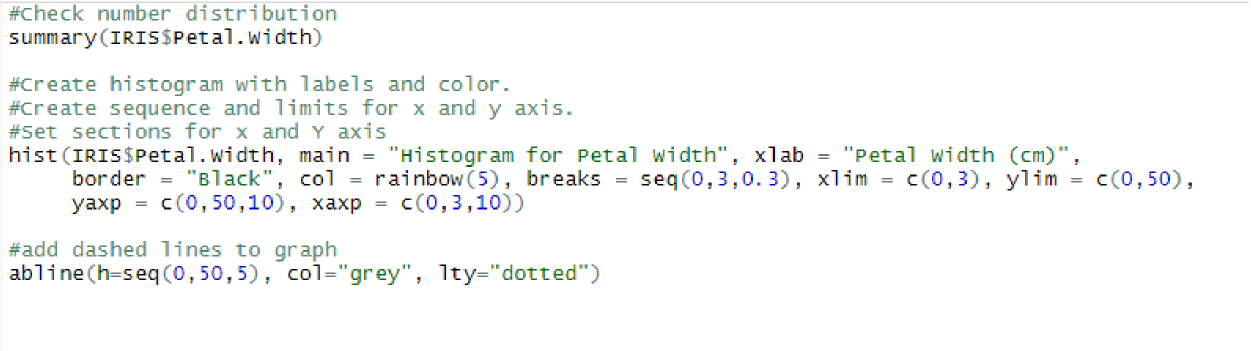


```
#Check number distribution
summary(IRIS$Sepal.Width)

#Create histogram with labels and color.
#Create sequence and limits for x and y axis.
#Set sections for x and Y axis
hist(IRIS$Sepal.Width, main = "Histogram for Sepal Width", xlab = "Sepal Width (cm)",
     border = "Black", col = rainbow(5), breaks = seq(2,5,0.2), xlim = c(2,4.4), ylim = c(0,40),
     yaxp=c(0,40,10), xaxp=c(2,5,15))

#add dashed lines to graph
abline(h=seq(0,40,4), col="grey", lty="dotted")
```

Interpretation

The data with respect to the sepal width appears to be skewed right or positively skewed. The distribution of data appears to be wide and the frequency decreases as the sepal width size increases.

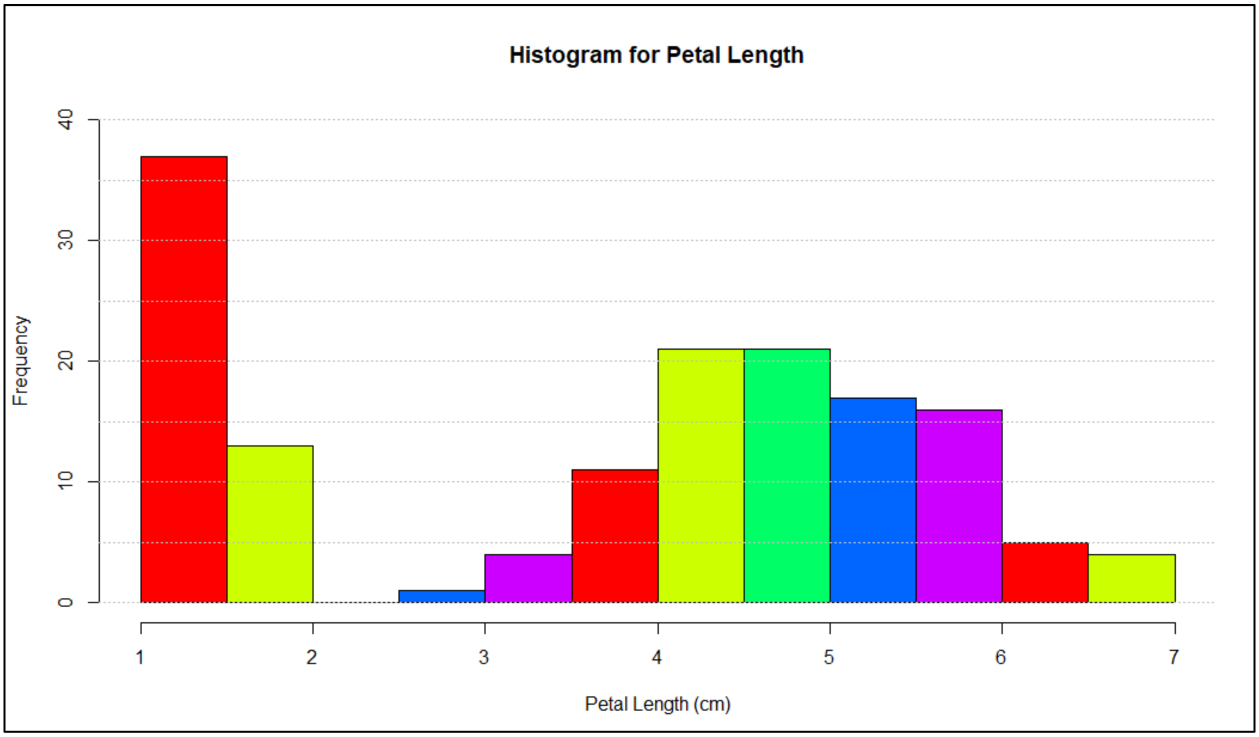


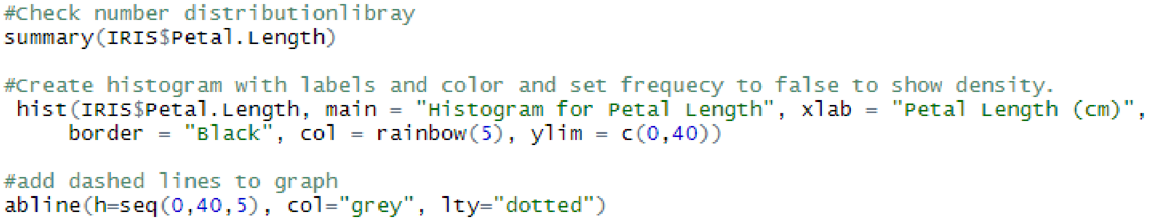


```
#Check number distribution
summary(IRIS$Petal.Width)

#Create histogram with labels and color.
#Create sequence and limits for x and y axis.
#Set sections for x and Y axis
hist(IRIS$Petal.Width, main = "Histogram for Petal Width", xlab = "Petal Width (cm)",
     border = "Black", col = rainbow(5), breaks = seq(0,3,0.3), xlim = c(0,3), ylim = c(0,50),
     yaxp = c(0,50,10), xaxp = c(0,3,10))

#add dashed lines to graph
abline(h=seq(0,50,5), col="grey", lty="dotted")
```

Interpretation

The histogram drawn with respect to the petal width appears to be skewed to the right (positively skewed). It also appears that some outliers may be present due to the bin gap in the histogram. The distribution fit is not very good.



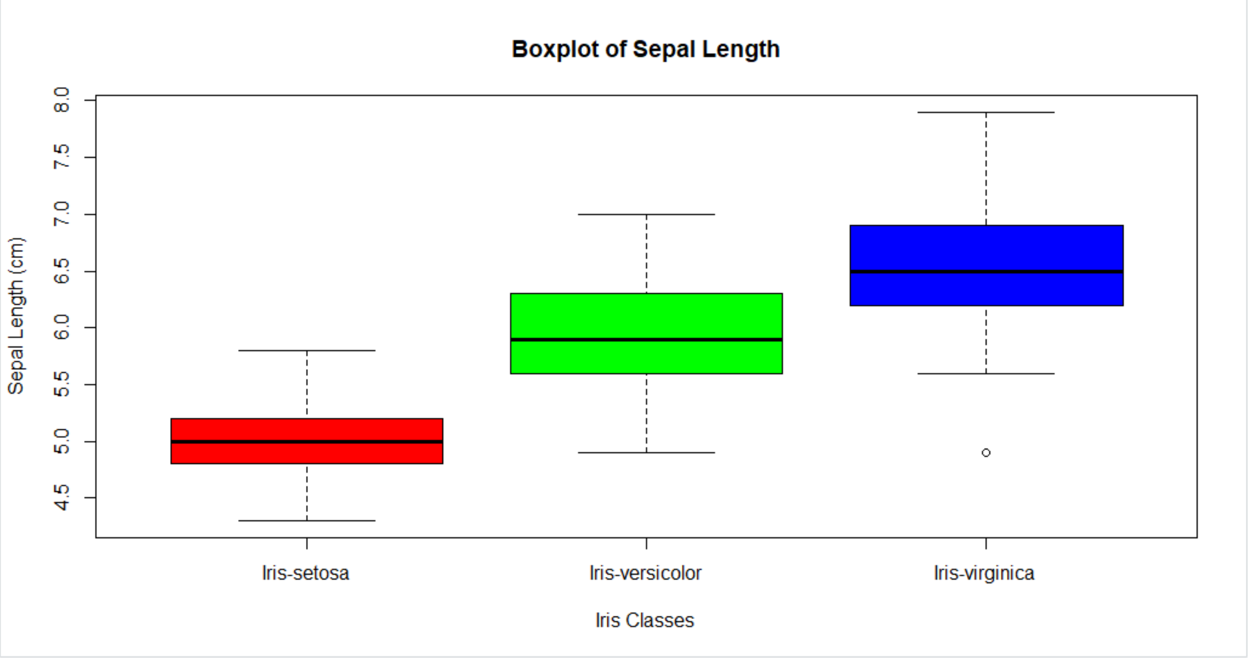


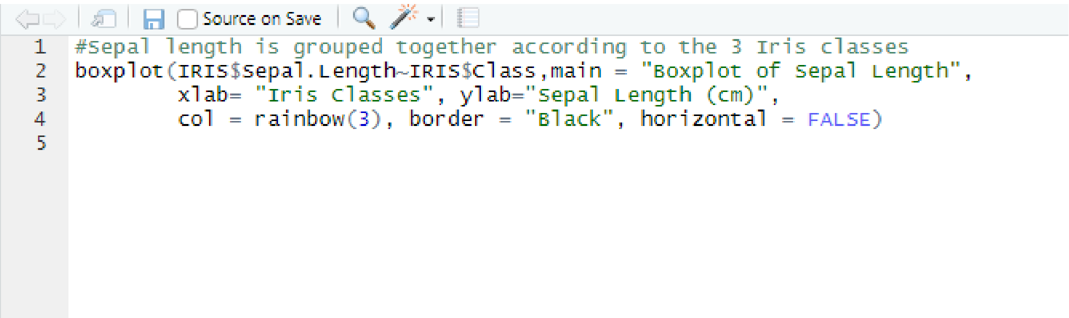
```
#Check number distributionlibray
summary(IRIS$Petal.Length)

#Create histogram with labels and color and set frequecy to false to show density.
 hist(IRIS$Petal.Length, main = "Histogram for Petal Length", xlab = "Petal Length (cm)",
     border = "Black", col = rainbow(5), ylim = c(0,40))

#add dashed lines to graph
abline(h=seq(0,40,5), col="grey", lty="dotted")
```

Interpretation

The data set for the petal length variable is more positively skewed where the mode occurs at a value that is smaller than the median. There may also be some outliers within the data as a gap is shown and the distribution appears wide.

D. Determine if there are any outliers in the data with respect to sepal length



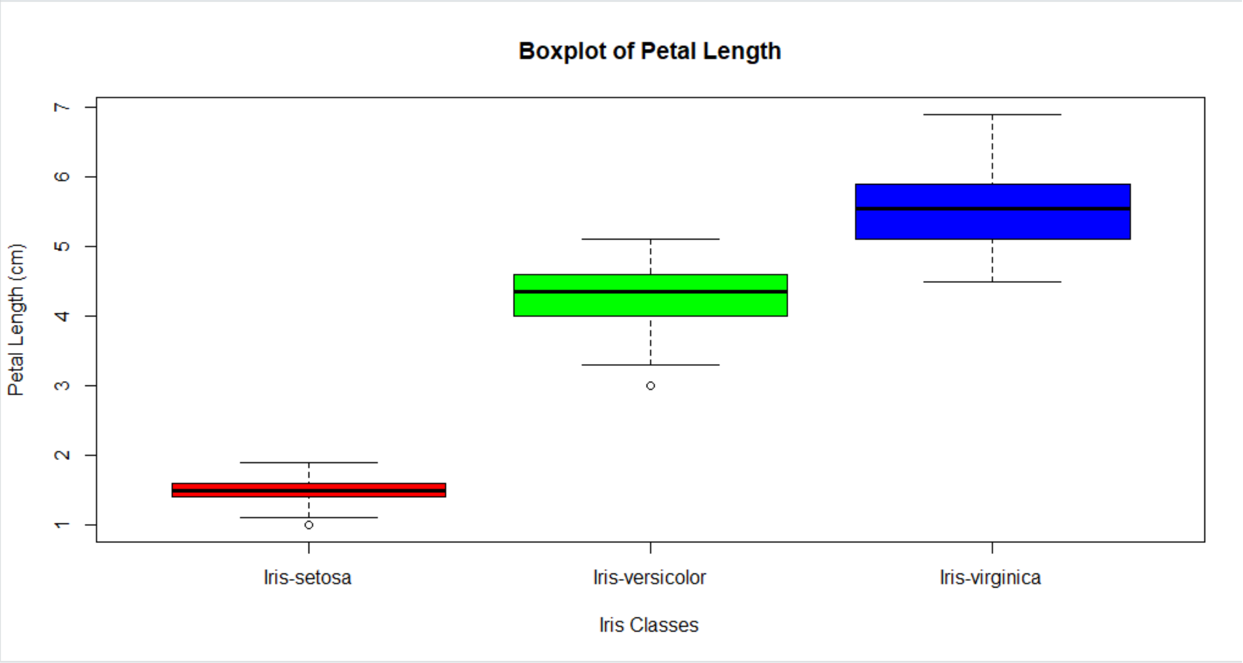


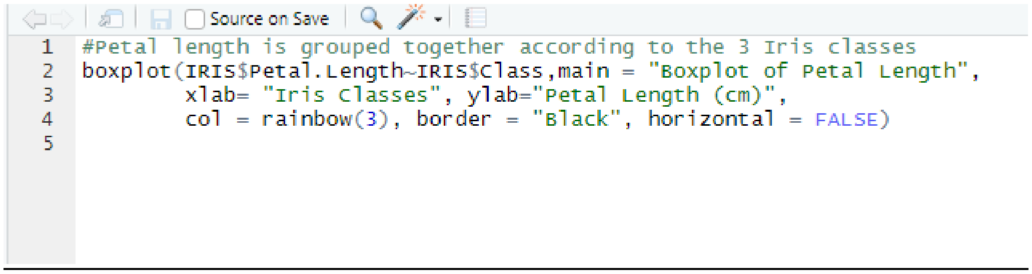
```
#Sepal length is grouped together according to the 3 Iris classes
boxplot(IRIS$Sepal.Length~IRIS$Class,main = "Boxplot of Sepal Length",
        xlab= "Iris Classes", ylab="Sepal Length (cm)",
        col = rainbow(3), border = "Black", horizontal = FALSE)

```

Interpretation

There appears to be outliers present in the Iris-virginica class with respect to the sepal length. No outliers are present for iris-setosa or iris-versicolor classes

E. Determine outliers in data with respect to petal length





```
#Petal length is grouped together according to the 3 Iris classes
boxplot(IRIS$Petal.Length~IRIS$Class,main = "Boxplot of Petal Length",
        xlab= "Iris Classes", ylab="Petal Length (cm)",
        col = rainbow(3), border = "Black", horizontal = FALSE)
```

<u>Interpretation</u>

There appears to be outliers in the Iris-setosa and Iris-versicolor classes when the data is plotted with respect to the petal length.