Iris Dataset Assignment

David Berberena

01-21-2024

Iris Dataset Analysis

```
# Datasets library is loaded to access built-in R datasets including Iris

library(datasets)

# Variable declaration to hold the Iris dataset

iris_data <- iris
```

Average Sepal Length by Species

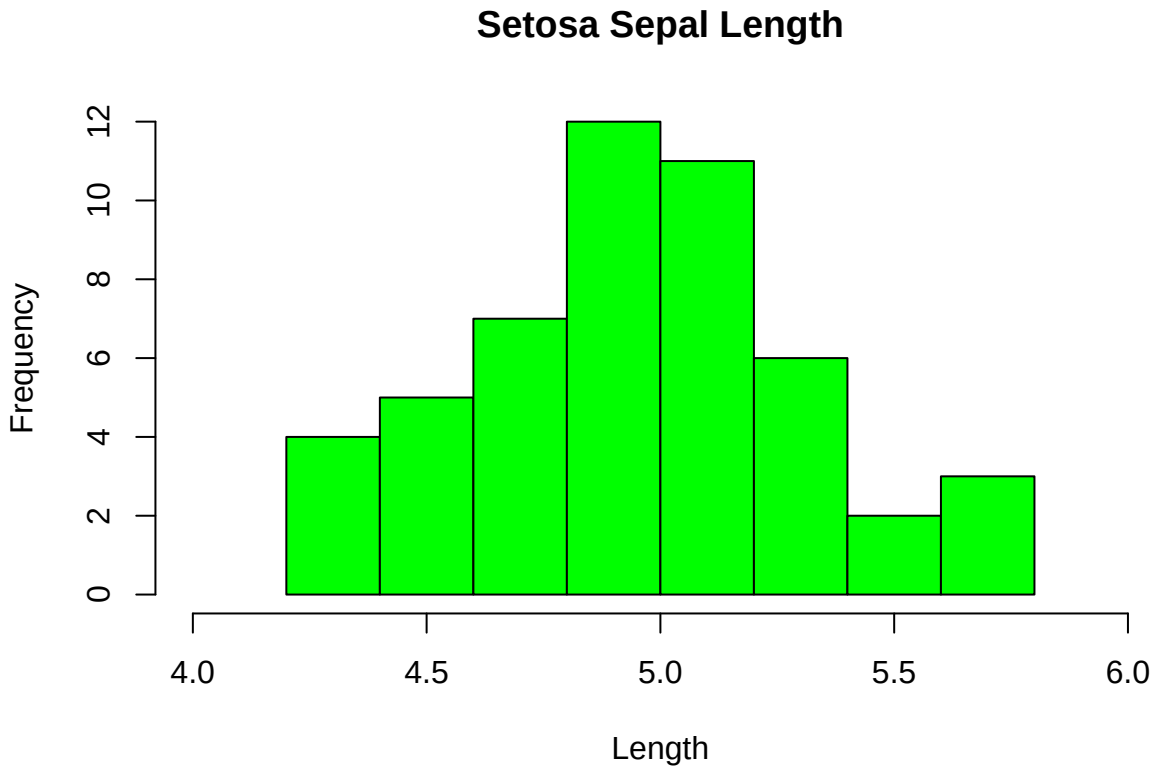
```
library(dplyr)

# group_by() and summarize() functions from dplyr are used in a pipe to gather
# the mean of the sepal length for each species

iris_species <- iris_data %>% group_by(Species) %>%
  summarize(AvgSepalLength = mean(`Sepal.Length`))
iris_species
```

```
## # A tibble: 3 x 2
##   Species    AvgSepalLength
##   <fct>               <dbl>
## 1 setosa               5.01
## 2 versicolor           5.94
## 3 virginica            6.59
```

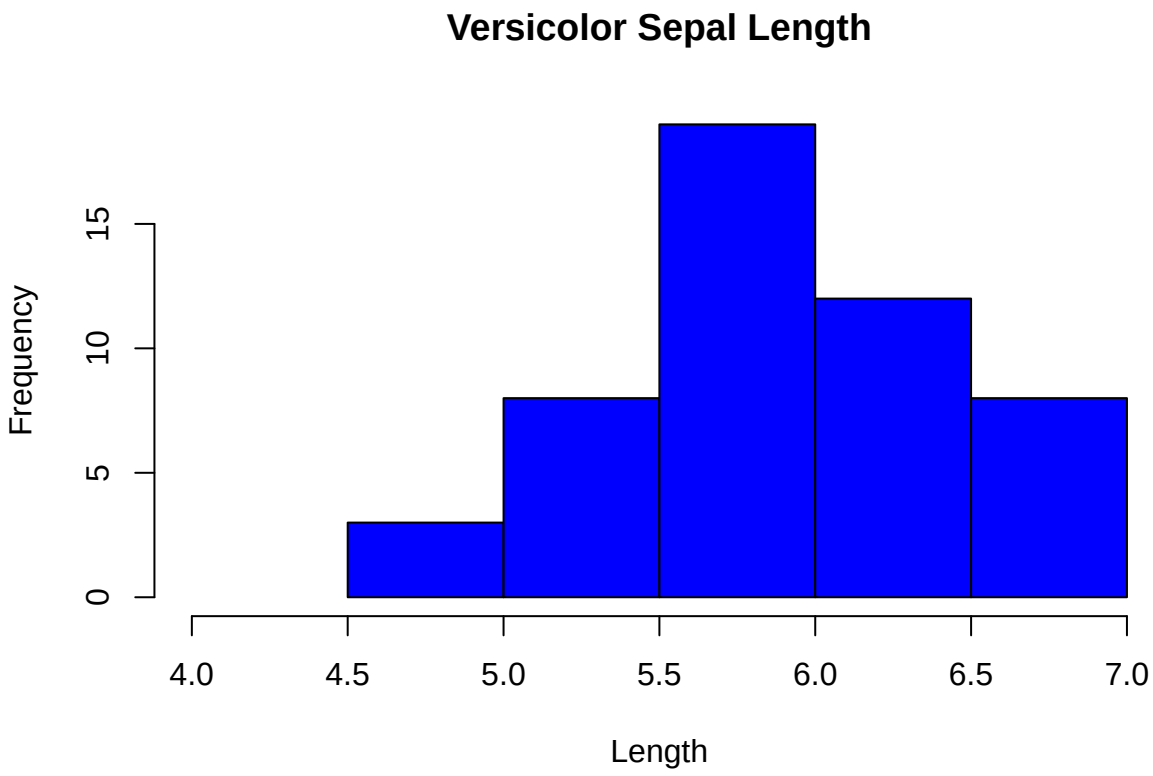
Visualizations

```
# Sepal Length Histograms

hist(iris_data$Sepal.Length[iris_data$Species == "setosa"],
     col='green',
     main='Setosa Sepal Length',
```

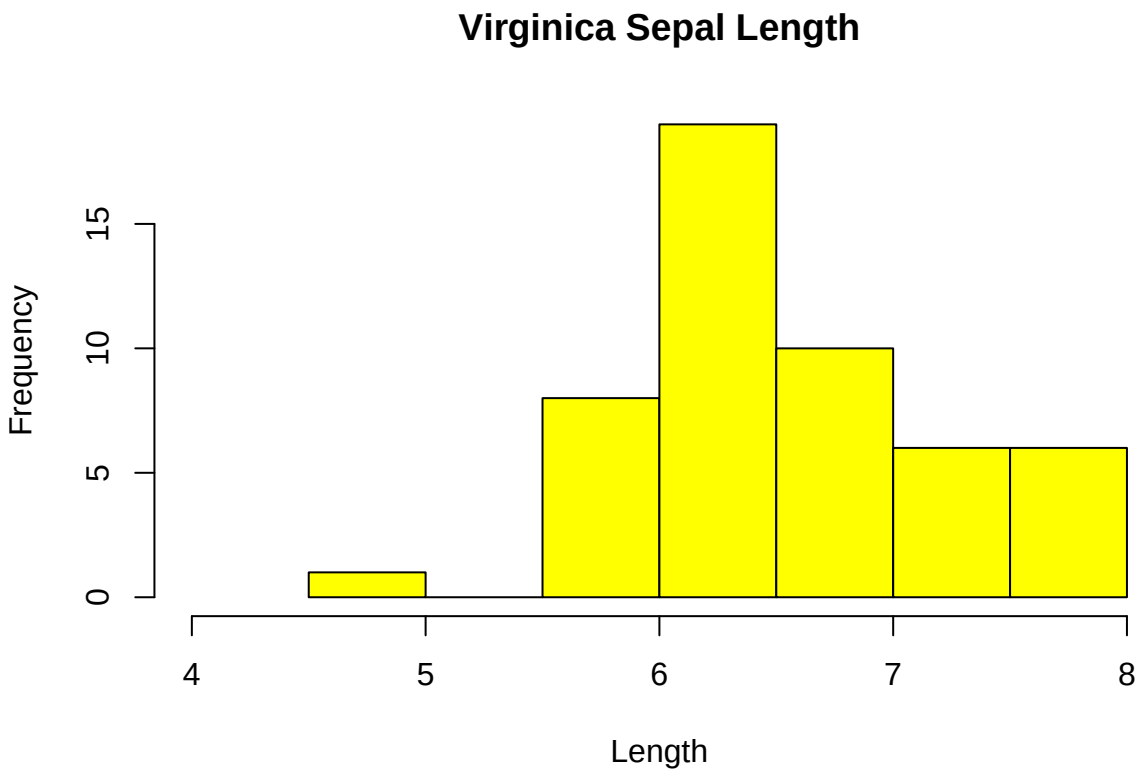
```
     xlab='Length',
     ylab='Frequency',
     xlim = c(4,6))
```




```
hist(iris_data$Sepal.Length[iris_data$Species == "versicolor"],
     col='blue',
     main='Versicolor Sepal Length',
     xlab='Length',
     ylab='Frequency',
     xlim = c(4,7))
```




```
hist(iris_data$Sepal.Length[iris_data$Species == "virginica"],
     col='yellow',
     main='Virginica Sepal Length',
     xlab='Length',
     ylab='Frequency',
     xlim = c(4,8))
```

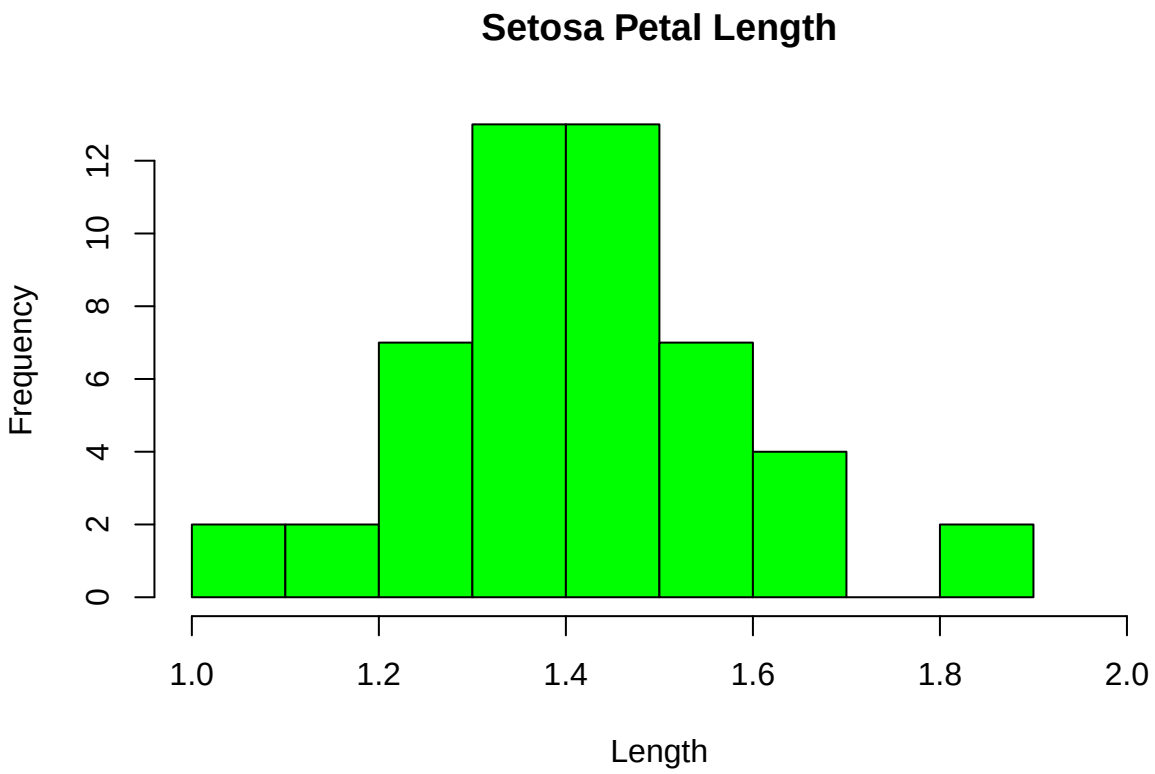


The frequency distribution of each species looks similar judging by the overall shape the bars create. The Setosa species is the most normally distributed, followed by the Versicolor which is a bit skewed and then the Virginica species being the most skewed. The Virginica species has the largest frequency gap between sepal length measurements, followed by Versicolor and Setosa, respectively.

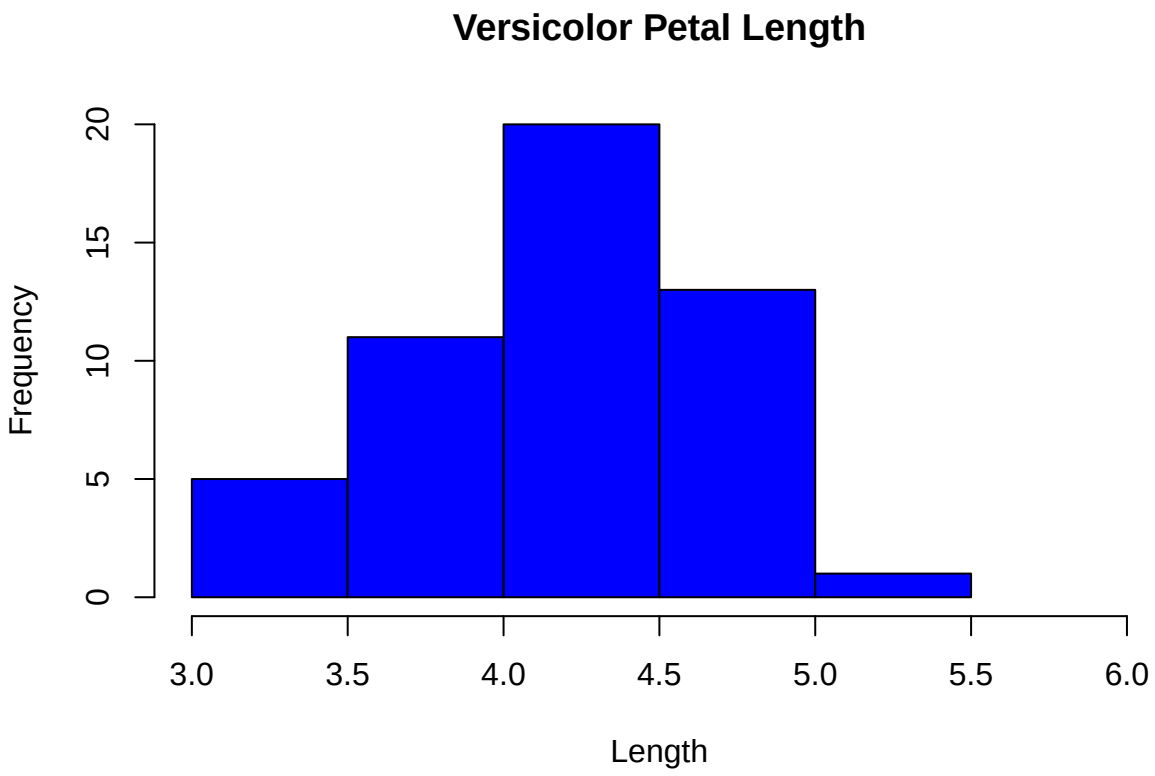
```
# Sepal Width Boxplots

boxplot(Sepal.Width~Species,
        data=iris_data,
        main='Sepal Width by Species',
        xlab='Species',
        ylab='Sepal Width',
        col='orange',
        border='black')
```

The Setosa species has the highest boxplot values (minimum, Q1, median, Q3 , and maximum) of the three species while Versicolor has the lowest values, meaning that the Setosa species generally have wider sepals than both the Versicolor and Virginica species. The Setosa species also has an outlier that falls outside the minimum value, prompting further investigation as to the existence of the data point. The median value for the Virginica boxplot looks almost equidistant from the Q1 and Q3 values, while Versicolor median lies closer to the 75th percentile and the Setosa median lies closer to the 25th percentile.

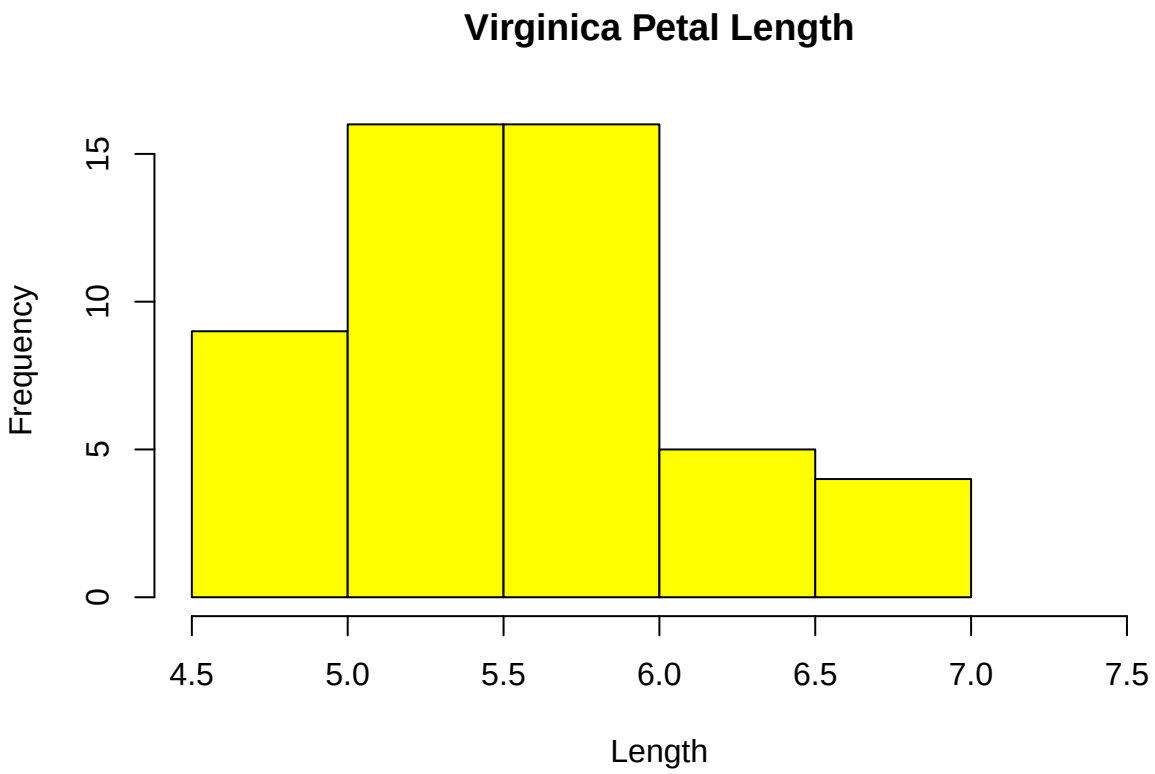
```
# Petal Length Histograms

hist(iris_data$Petal.Length[iris_data$Species == "setosa"],
     col='green',
     main='Setosa Petal Length',
     xlab='Length',
     ylab='Frequency',
     xlim = c(1,2))
```




```
hist(iris_data$Petal.Length[iris_data$Species == "versicolor"],
     col='blue',
     main='Versicolor Petal Length',
     xlab='Length',
     ylab='Frequency',
     xlim = c(3,6))
```




```
hist(iris_data$Petal.Length[iris_data$Species == "virginica"],
     col='yellow',
     main='Virginica Petal Length',
     xlab='Length',
     ylab='Frequency',
     xlim = c(4.5,7.5))
```



The frequency distributions for petal length follow the same pattern as the sepal length for each species. The Setosa species again is the most normally distributed out of the three species. The Virginica species once again has the largest frequency gap between petal length measurements.

```
# Petal Width Boxplots

boxplot(Petal.Width~Species,
        data=iris_data,
        main='Petal Width by Species',
        xlab='Species',
        ylab='Petal Width',
        col='red',
        border='black')
```

With the petal width visual aids, it is blatantly obvious that the Setosa species have the narrowest petals, with both the Versicolor and Virginica species having their minimum values be greater than the Setosa species's maximum value and the two outlier data points present that lie above the maximum Setosa boxplot value. The Virginica species has the widest petals, and the Versicolor's maximum boxplot value is approximately the same as Virginica's 25th percentile value.