Explore the Iris Data Set Using R Studio

Didem Bulut Aykurt

MIS500-1 – Foundations of Data Analytics

Colorado State University-Global Campus

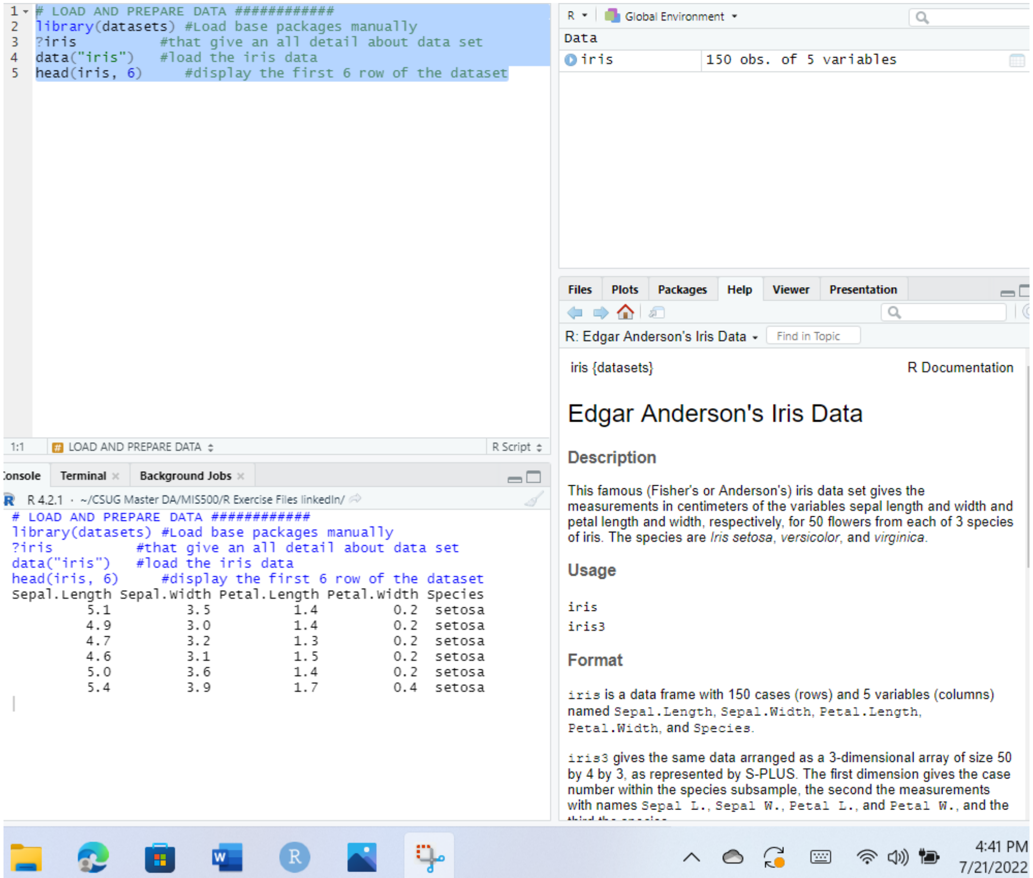
Dr. Steve Chung

July 24, 2022

Module 1: Option #1: Critical Thinking

When do I start to learn new knowledge that comes up the question of why I need that tool? Where will I use it? What kind of problem would I solve? Why do most companies work with the to R program? First, I would like to give some minor information about all my questions. I think the best reason to work R has over 12K packages and libraries available and handling big unstructured data, all free or lower than another program. Also, safe and security are the reasons all companies need them. Let's look at my work; R has an analysis tool to create graph&chart with big and heavy data. That has an excellent console to manage code, data, and results and display the graph and chart. Therefore, let's deep dive into the iris dataset with the R programing language.

Load and display the first six rows.



```
# LOAD AND PREPARE DATA ############
library(datasets) #Load base packages manually
?iris          #that give an all detail about data set
data("iris")   #load the iris data
head(iris, 6)      #display the first 6 row of the dataset
```

```
# LOAD AND PREPARE DATA ############
library(datasets) #Load base packages manually
?iris          #that give an all detail about data set
data("iris")   #load the iris data
head(iris, 6)      #display the first 6 row of the dataset
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
         5.1         3.5          1.4         0.2  setosa
         4.9         3.0          1.4         0.2  setosa
         4.7         3.2          1.3         0.2  setosa
         4.6         3.1          1.5         0.2  setosa
         5.0         3.6          1.4         0.2  setosa
         5.4         3.9          1.7         0.4  setosa
```

```
Data
iris     150 obs. of 5 variables
```

iris {datasets}   R Documentation

Edgar Anderson's Iris Data

Description

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

Usage

```
iris
iris3
```

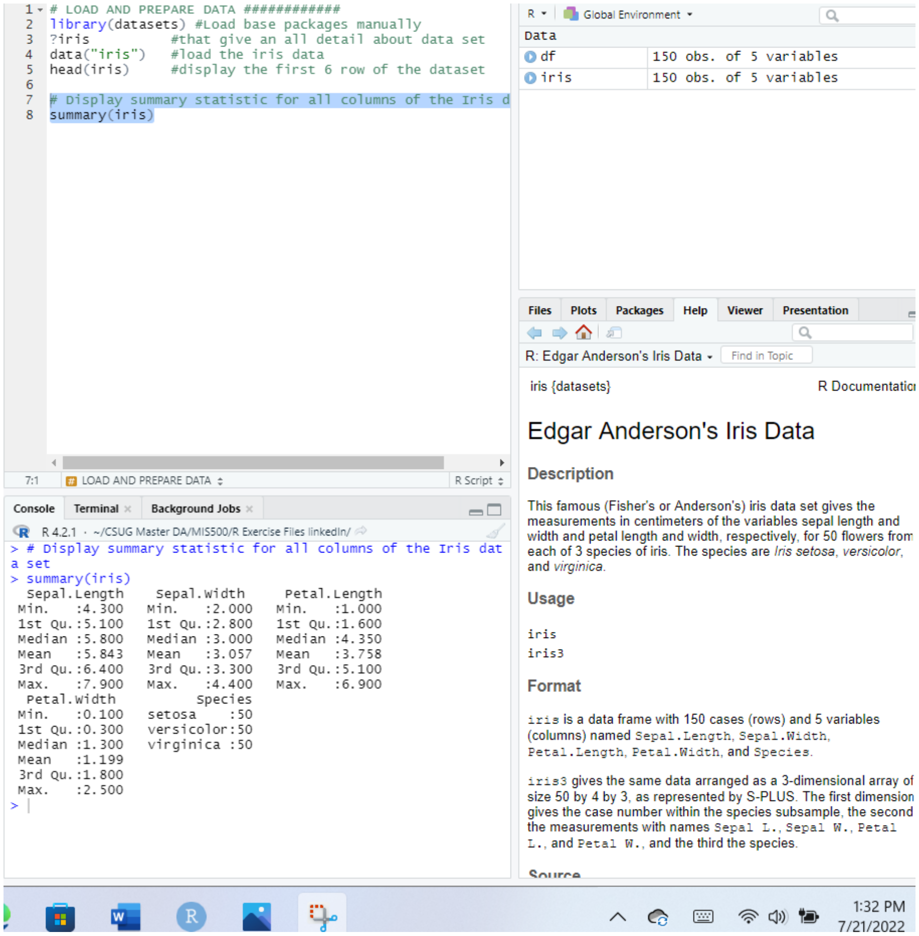
Format

`iris` is a data frame with 150 cases (rows) and 5 variables (columns) named `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`.

`iris3` gives the same data arranged as a 3-dimensional array of size 50 by 4 by 3, as represented by S-PLUS. The first dimension gives the case number within the species subsample, the second the measurements with names `Sepal L.`, `Sepal W.`, `Petal L.`, and `Petal W.`, and the

Module 1: Option #1: Critical Thinking

When displaying a statistical summary help to understand statistical information in a dataset with statistical terms; Iris data has five different variables, each of which has a close result mean and median number, which means most data point are relative to the mean, and most variable look a normal distribution thus we should check bar graph or boxplot to make sure each distribution to tell normal or not.



```
# LOAD AND PREPARE DATA ############
library(datasets) #Load base packages manually
?iris          #that give an all detail about data set
data("iris")   #load the iris data
head(iris)     #display the first 6 row of the dataset

# Display summary statistic for all columns of the Iris d
summary(iris)
```

```
> # Display summary statistic for all columns of the Iris dat
a set
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length  
 Min.   :4.300   Min.   :2.000   Min.   :1.000  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600  
 Median :5.800   Median :3.000   Median :4.350  
 Mean   :5.843   Mean   :3.057   Mean   :3.758  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100  
 Max.   :7.900   Max.   :4.400   Max.   :6.900  
  Petal.Width          Species  
 Min.   :0.100   setosa    :50  
 1st Qu.:0.300   versicolor:50  
 Median :1.300   virginica :50  
 Mean   :1.199                  
 3rd Qu.:1.800                  
 Max.   :2.500                  
> 
```

Data

df	150 obs. of 5 variables
iris	150 obs. of 5 variables

iris {datasets} R Documentatio

Edgar Anderson's Iris Data

Description

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

Usage

```
iris
iris3
```

Format

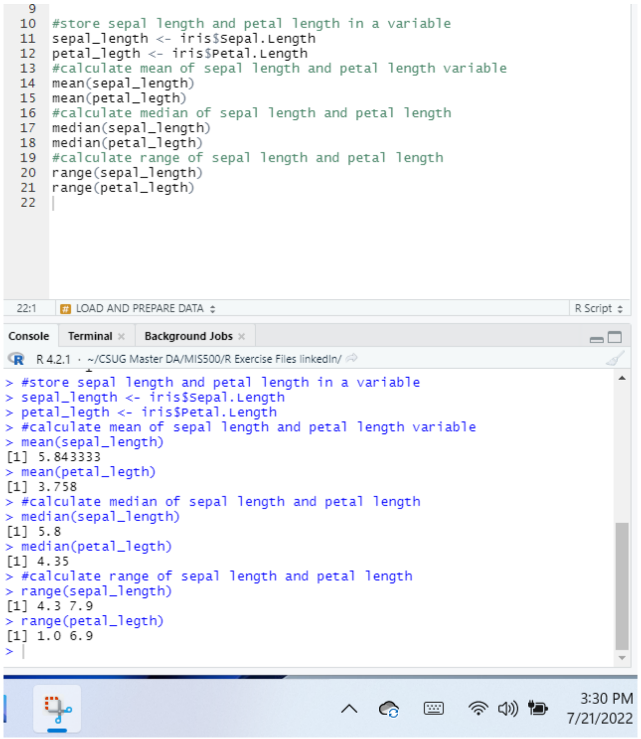
iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

iris3 gives the same data arranged as a 3-dimensional array of size 50 by 4 by 3, as represented by S-PLUS. The first dimension gives the case number within the species subsample, the second the measurements with names Sepal L., Sepal W., Petal L., and Petal W., and the third the species.

Source

Sepal.Length and Petal. Lenght is the Iris dataset of a variable as sepal length mean bigger than petal length, which implies sepal length longer than petal length. Look at the petal length range more considerable than the sepal length; thus, the petal length data points are far, and the sepal length data points close to each other.

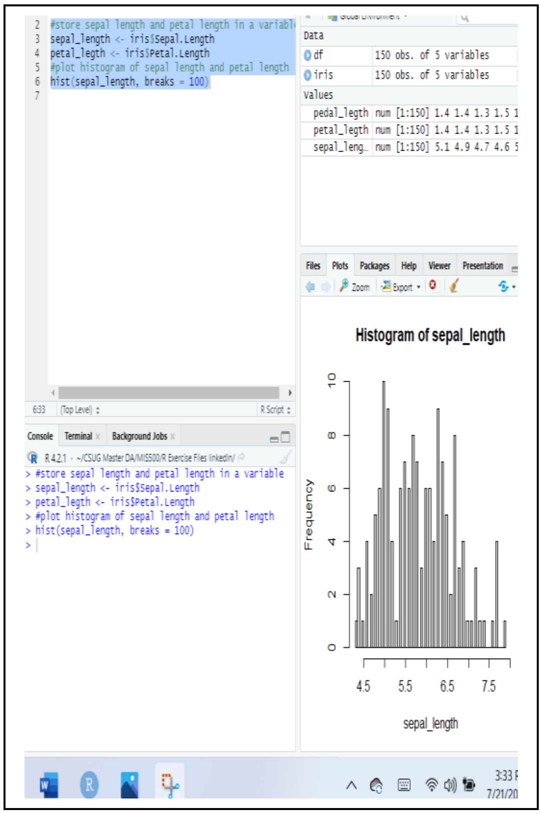
Module 1: Option #1: Critical Thinking



```
#store sepal length and petal length in a variable
sepal_length <- iris$Sepal.Length
petal_legth <- iris$Petal.Length
#calculate mean of sepal length and petal length variable
mean(sepal_length)
mean(petal_legth)
#calculate median of sepal length and petal length
median(sepal_length)
median(petal_legth)
#calculate range of sepal length and petal length
range(sepal_length)
range(petal_legth)
```

```
> #store sepal length and petal length in a variable
> sepal_length <- iris$Sepal.Length
> petal_legth <- iris$Petal.Length
> #calculate mean of sepal length and petal length variable
> mean(sepal_length)
[1] 5.843333
> mean(petal_legth)
[1] 3.758
> #calculate median of sepal length and petal length
> median(sepal_length)
[1] 5.8
> median(petal_legth)
[1] 4.35
> #calculate range of sepal length and petal length
> range(sepal_length)
[1] 4.3 7.9
> range(petal_legth)
[1] 1.0 6.9
>
```

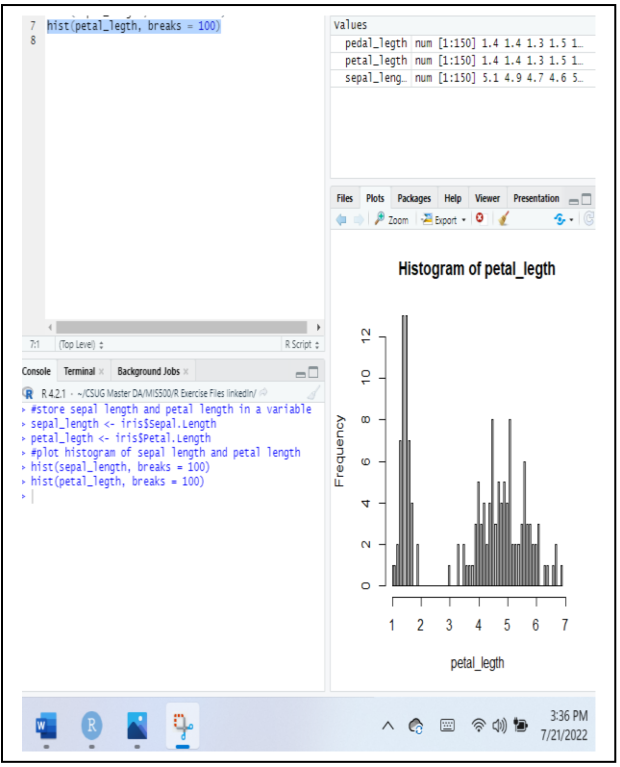
The sepal length bar graph has one pick, which means the normal distribution is symmetric from a peak of the curve and mainly observed data near the mean, but petal length variables far from the norm also have two skewed waves that make the right-skewed.



```
#store sepal length and petal length in a variable
sepal_length <- iris$Sepal.Length
petal_legth <- iris$Petal.Length
#plot histogram of sepal length and petal length
hist(sepal_length, breaks = 100)
```

```
Data
df          150 obs. of 5 variables
iris        150 obs. of 5 variables
Values
pedal_legth  num [1:150] 1.4 1.4 1.3 1.5 1
petal_legth  num [1:150] 1.4 1.4 1.3 1.5 1
sepal_leng…  num [1:150] 5.1 4.9 4.7 4.6 5
```

```
> #store sepal length and petal length in a variable
> sepal_length <- iris$Sepal.Length
> petal_legth <- iris$Petal.Length
> #plot histogram of sepal length and petal length
> hist(sepal_length, breaks = 100)
>
```



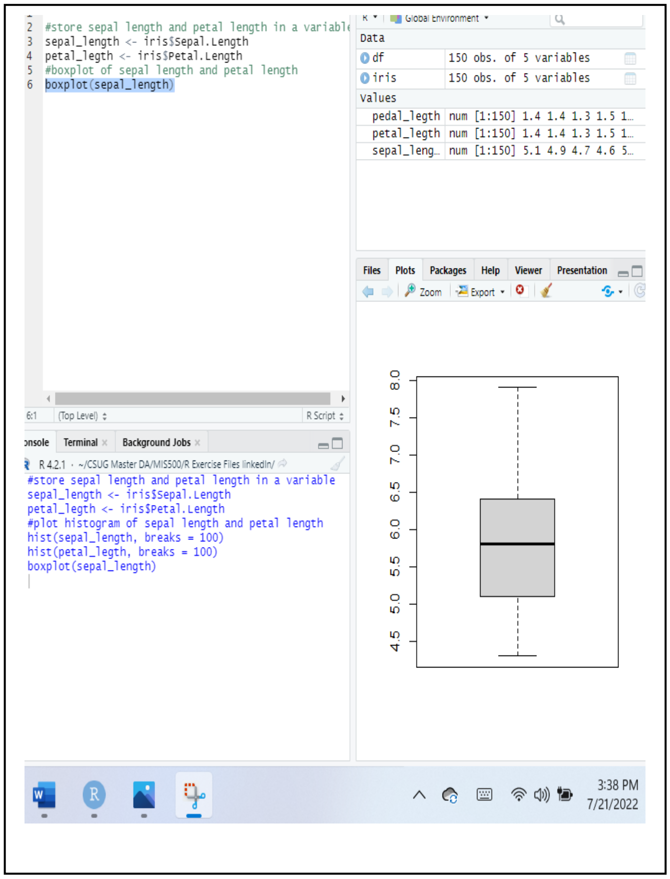
```
hist(petal_legth, breaks = 100)
```

```
Values
pedal_legth  num [1:150] 1.4 1.4 1.3 1.5 1…
petal_legth  num [1:150] 1.4 1.4 1.3 1.5 1…
sepal_leng…  num [1:150] 5.1 4.9 4.7 4.6 5…
```

```
> #store sepal length and petal length in a variable
> sepal_length <- iris$Sepal.Length
> petal_legth <- iris$Petal.Length
> #plot histogram of sepal length and petal length
> hist(sepal_length, breaks = 100)
> hist(petal_legth, breaks = 100)
>
```

Module 1: Option #1: Critical Thinking

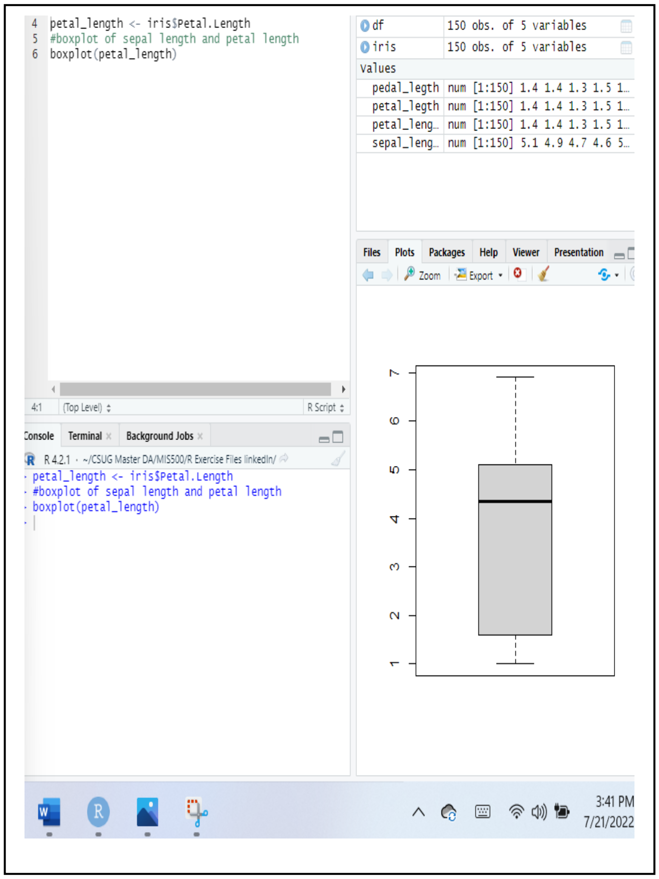
The sepal length's box plot shows the symmetric distribution. Petal length's box plot tells positively skewed distribution—sepal length's box plot shorter the fewer spread data. Petal length box plot longer the more spread data.



```
#store sepal length and petal length in a variable
sepal_length <- iris$Sepal.Length
petal_legth <- iris$Petal.Length
#boxplot of sepal length and petal length
boxplot(sepal_length)
```

```
Data
df             150 obs. of 5 variables
iris           150 obs. of 5 variables
Values
  pedal_legth  num [1:150] 1.4 1.4 1.3 1.5 1…
  petal_legth  num [1:150] 1.4 1.4 1.3 1.5 1…
  sepal_leng…  num [1:150] 5.1 4.9 4.7 4.6 5…
```

```
R 4.2.1 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/
#store sepal length and petal length in a variable
sepal_length <- iris$Sepal.Length
petal_legth <- iris$Petal.Length
#plot histogram of sepal length and petal length
hist(sepal_length, breaks = 100)
hist(petal_legth, breaks = 100)
boxplot(sepal_length)
```



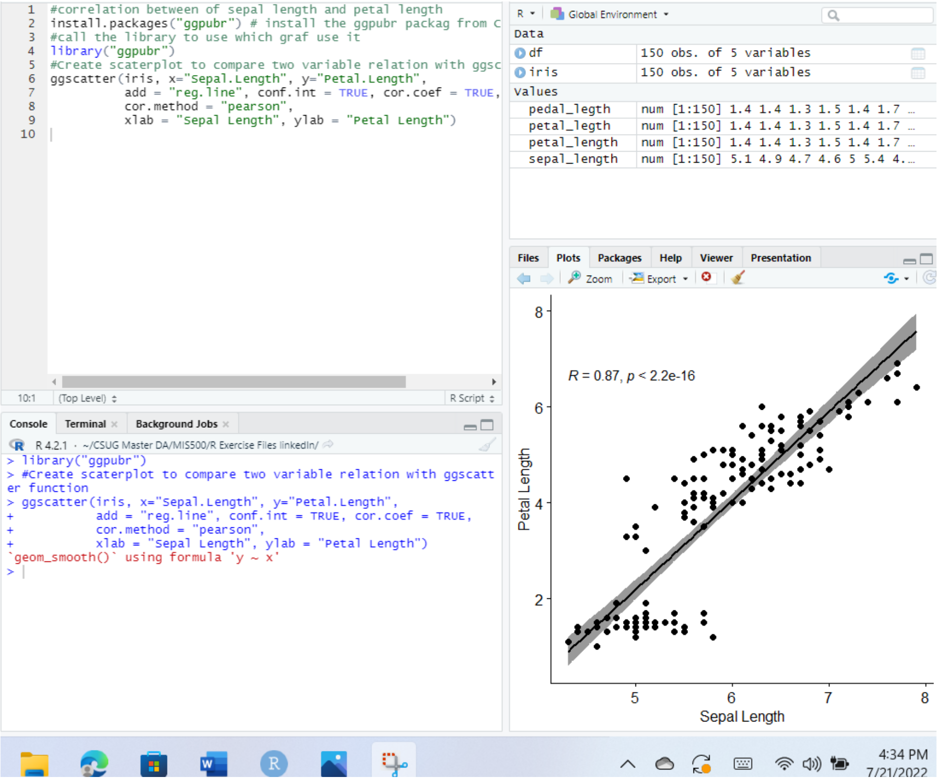
```
petal_length <- iris$Petal.Length
#boxplot of sepal length and petal length
boxplot(petal_length)
```

```
df             150 obs. of 5 variables
iris           150 obs. of 5 variables
Values
  pedal_legth  num [1:150] 1.4 1.4 1.3 1.5 1…
  petal_legth  num [1:150] 1.4 1.4 1.3 1.5 1…
  petal_leng…  num [1:150] 1.4 1.4 1.3 1.5 1…
  sepal_leng…  num [1:150] 5.1 4.9 4.7 4.6 5…
```

```
R 4.2.1 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/
> petal_length <- iris$Petal.Length
> #boxplot of sepal length and petal length
> boxplot(petal_length)
>
```

Sepal length and petal length have positive relation, and R square is 87%, which means the regression model fits the data. Also, a p-value lower than .05 says to reject the null hypothesis. The null hypothesis implies that there is no relation between the two variables.

Module 1: Option #1: Critical Thinking



```r
#correlation between of sepal length and petal length
install.packages("ggpubr") # install the ggpubr packag from C
#call the library to use which graf use it
library("ggpubr")
#Create scaterplot to compare two variable relation with ggsc
ggscatter(iris, x="Sepal.Length", y="Petal.Length",
          add = "reg.line", conf.int = TRUE, cor.coef = TRUE,
          cor.method = "pearson",
          xlab = "Sepal Length", ylab = "Petal Length")
```

```
R 4.2.1 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/
> library("ggpubr")
> #Create scaterplot to compare two variable relation with ggscatt
er function
> ggscatter(iris, x="Sepal.Length", y="Petal.Length",
+           add = "reg.line", conf.int = TRUE, cor.coef = TRUE,
+           cor.method = "pearson",
+           xlab = "Sepal Length", ylab = "Petal Length")
`geom_smooth()` using formula 'y ~ x'
> 
```

```
Data
df            150 obs. of 5 variables
iris          150 obs. of 5 variables
Values
pedal_legth   num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 …
petal_legth   num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 …
petal_length  num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 …
sepal_length  num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.…
```

Conclusion

R is a new tool to add my skills. That is like python syntax. Where I got stuck starting the find R syntax. I am still looking cut sheet for syntax or any sources. Other statical terms make me so excited to find the meaning.

Reference

[Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA]

[Summary or Descriptive statistics in R - DataScience Made Simple]