# Chapter 1 Step into R programming–the iris flower dataset

# 1.3 Data frames contain rows and columns: the iris flower dataset

## Exercise 1.1

a.Display the first few rows of the data set. **NOT** all values in your data set.

> head(attitude)

rating complaints privileges learning raises critical advance

1 43 51 30 39 61 92 45

2 63 64 51 54 63 73 47

3 71 70 68 69 76 86 48

4 61 63 45 47 54 84 35

5 81 78 56 66 71 83 47

1. 43 55 49 44 54 49 34

b.Show the dimension of the data set.

> str(attitude)

'data.frame': 30 obs. of 7 variables:

$ rating : num 43 63 71 61 81 43 58 71 72 67 ...

$ complaints: num 51 64 70 63 78 55 67 75 82 61 ...

$ privileges: num 30 51 68 45 56 49 42 50 72 45 ...

$ learning : num 39 54 69 47 66 44 56 55 67 47 ...

$ raises : num 61 63 76 54 71 54 66 70 71 62 ...

$ critical : num 92 73 86 84 83 49 68 66 83 80 ...

$ advance : num 45 47 48 35 47 34 35 41 31 41 ...

c.Extract a subset which contains values in the 2nd through the 5th rows and the 1st through the 4th columns. If your data set contains fewer rows or columns, please choose another data set.

> attitude[2:5,1:4]

rating complaints privileges learning

2 63 64 51 54

3 71 70 68 69

4 61 63 45 47

5 81 78 56 66

## Exercise 1.2

a.Compute the mean of the sepal length in the data set **iris**.

> SL <-iris$Sepal.Length

> mean(SL)

[1] 5.843333

b.Compute the mean of speed in the built-in data set **cars**.

> View(cars)

> mean(cars$speed)

[1] 15.4

## Exercise 1.3

a.Google “R square root function” to find the R code, then compute the value of √56.7.

> sqrt(56.7)

[1] 7.52994

b.Use R to find the maximal value of the variable *mpg* in the data set **mtcars**.

> View(mtcars)

> summary(mtcars$mpg)

Min. 1st Qu. Median Mean 3rd Qu. Max.

10.40 15.43 19.20 20.09 22.80 33.90

# 1.4 Analyzing one set of numbers

## Exercise 1.4

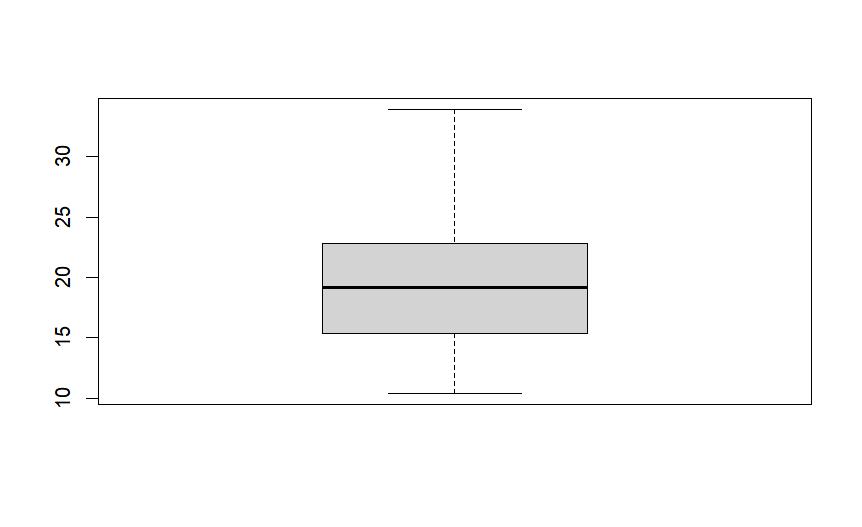
a. Check the data structure of the built-in data set **mtcars**.

R’s basic data structures include the vector, list, matrix, data frame, and factors.

> class(mtcars)

[1] "data.frame"

b. Get the boxplot of Mile Per Gallon (*mpg*) in the data set **mtcars**.



## Exercise 1.5

a.Run x = rnorm(500) to generate 500 random numbers following the Standard Normal distribution

> x = rnorm(500)

> x

[1] -0.422767866 -0.208532530 1.175800149 0.488412454 -1.014491081 -0.064697175

[7] 1.001565197 0.026546888 -0.812482678 -0.158438167 -0.064514140 1.185950066

[13] ……

b.Generate scatter plot, histogram, lag plot, and Q-Q plot. Your plots should like those in Figure [1.7](https://gexijin.github.io/learnR/step-into-r-programmingthe-iris-flower-dataset.html#fig:1-5).

|  |  |
| --- | --- |
| Scatter plot | Histogram |
|  |  |
| Lag plot | Q-Q plot |
|  |  |

## Exercise 1.6

Generate scatter plot, histogram, lag plot, and Q-Q plot for the Septal Length in the iris dataset.

|  |  |
| --- | --- |
| Scatter plot | Histogram |
|  |  |
| Lag plot | Q-Q plot |
|  |  |

## Exercise 1.7

Compute 95% confidence interval of sepal length of setosa.

> t.test(SL[1:50],mu=5)

One Sample t-test

data: SL[1:50]

t = 0.12036, df = 49, p-value = 0.9047

alternative hypothesis: true mean is not equal to 5

95 percent confidence interval:

4.905824 5.106176

sample estimates:

mean of x

5.006

*df = degrees of freedom*

## Exercise 1.8

shapiro.test: **the null hypothesis is that the data is from a normal distribution**.

*# normality test*

shapiro.test(PL)

a.Run Shapiro’s test on sepal width. Does it follow a normal distribution given the significant level α=0.05?

> SW <- iris$Sepal.Width

> shapiro.test(SW)

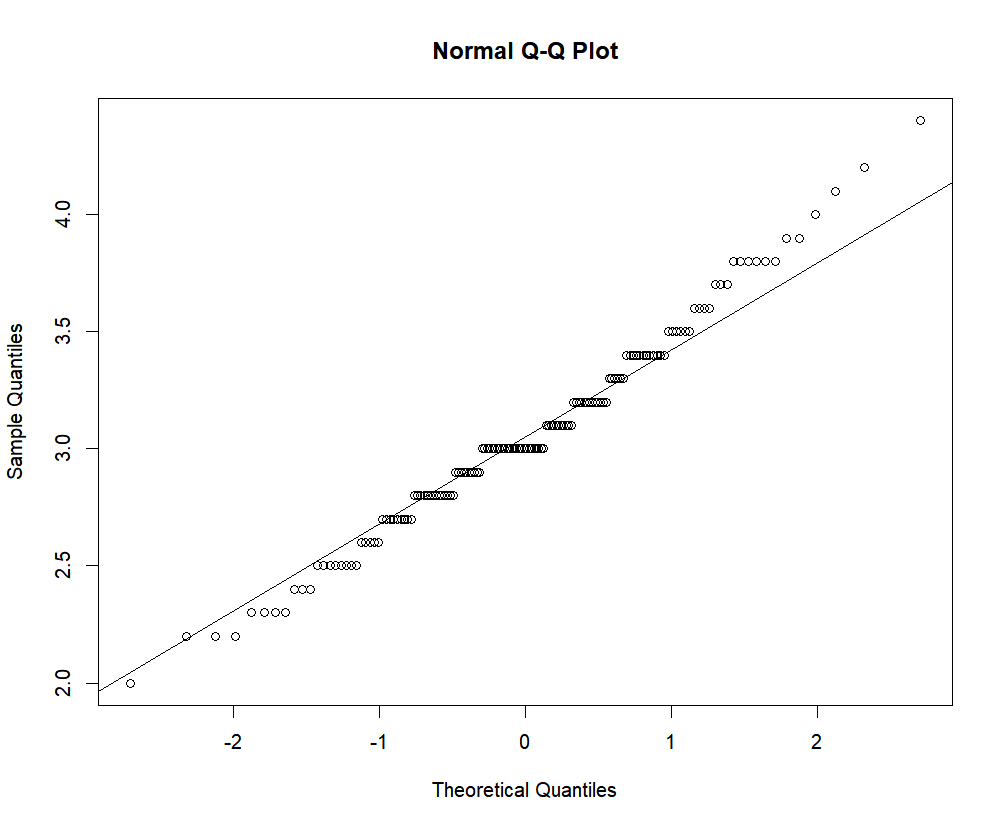
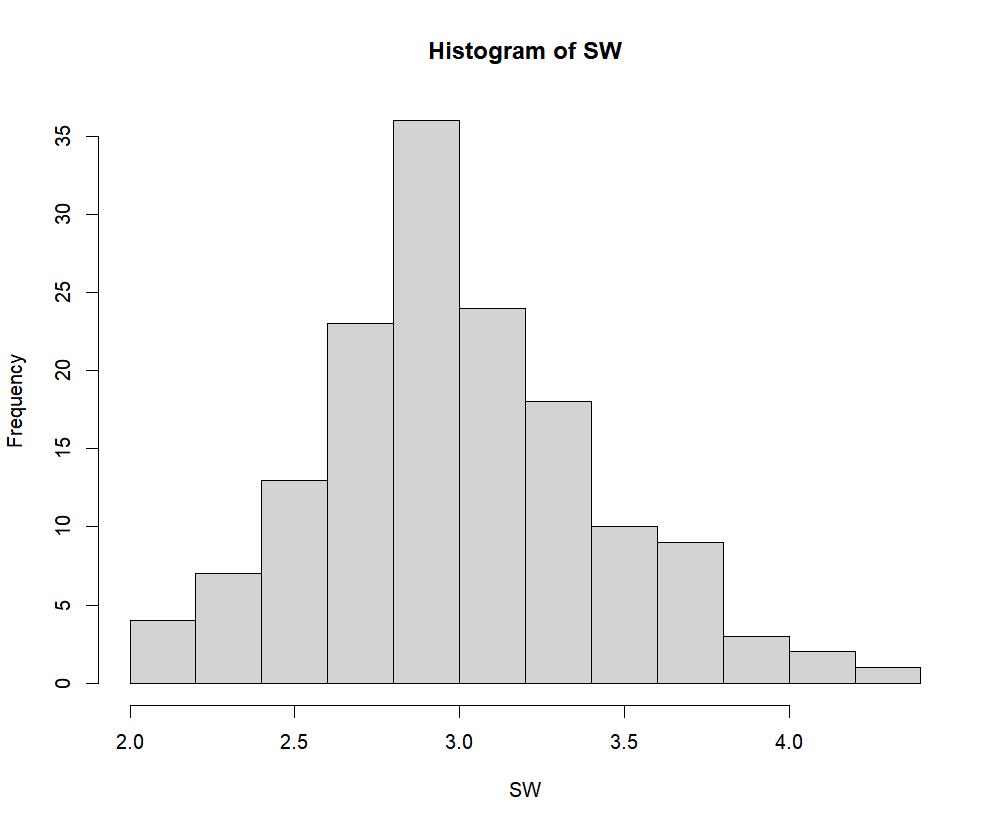
Shapiro-Wilk normality test

data: SW

W = 0.98492, p-value = 0.1012

*Interpretation: p-value = 0.1012, which is greater than 0.05, we fail to reject the null hypothesis that the sepal width follows a normal distribution.*

b.Generate histogram and Q-Q plot for sepal width. Do the plots show a Normal approximation? *Yes.*



# 1.6 The relationship between two numerical variables

## Exercise 1.9

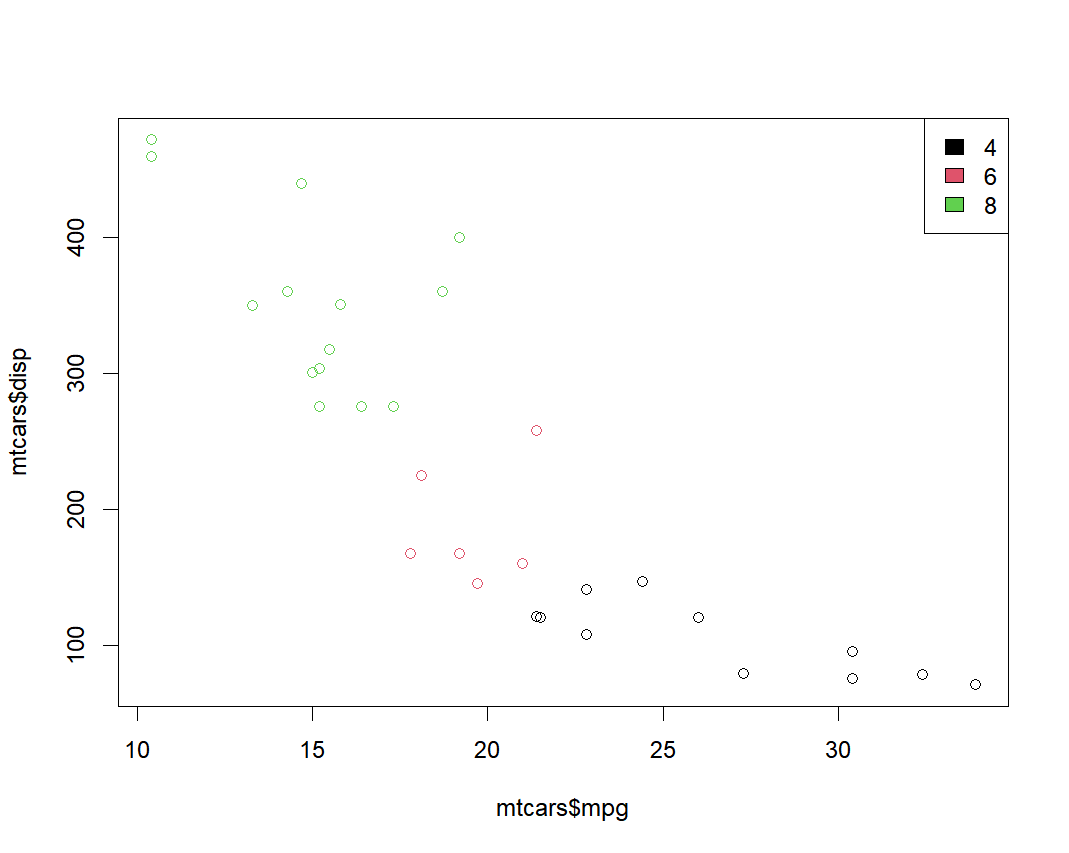
Based on mtcars data set, plot a scatter plot which shows the correlation between Miles/(US) gallon and Displacement (cu.in.). In this data set, the type of cyl is numeric. You will need to use function newcyl -> as.factor(cyl) to transfer the type to factor. Then replace all mtcars$cyl with newcyl.

1. Color the scatter plot by Number of cylinders;
2. Add legend to the top right.

> newcyl <- as.factor(mtcars$cyl)

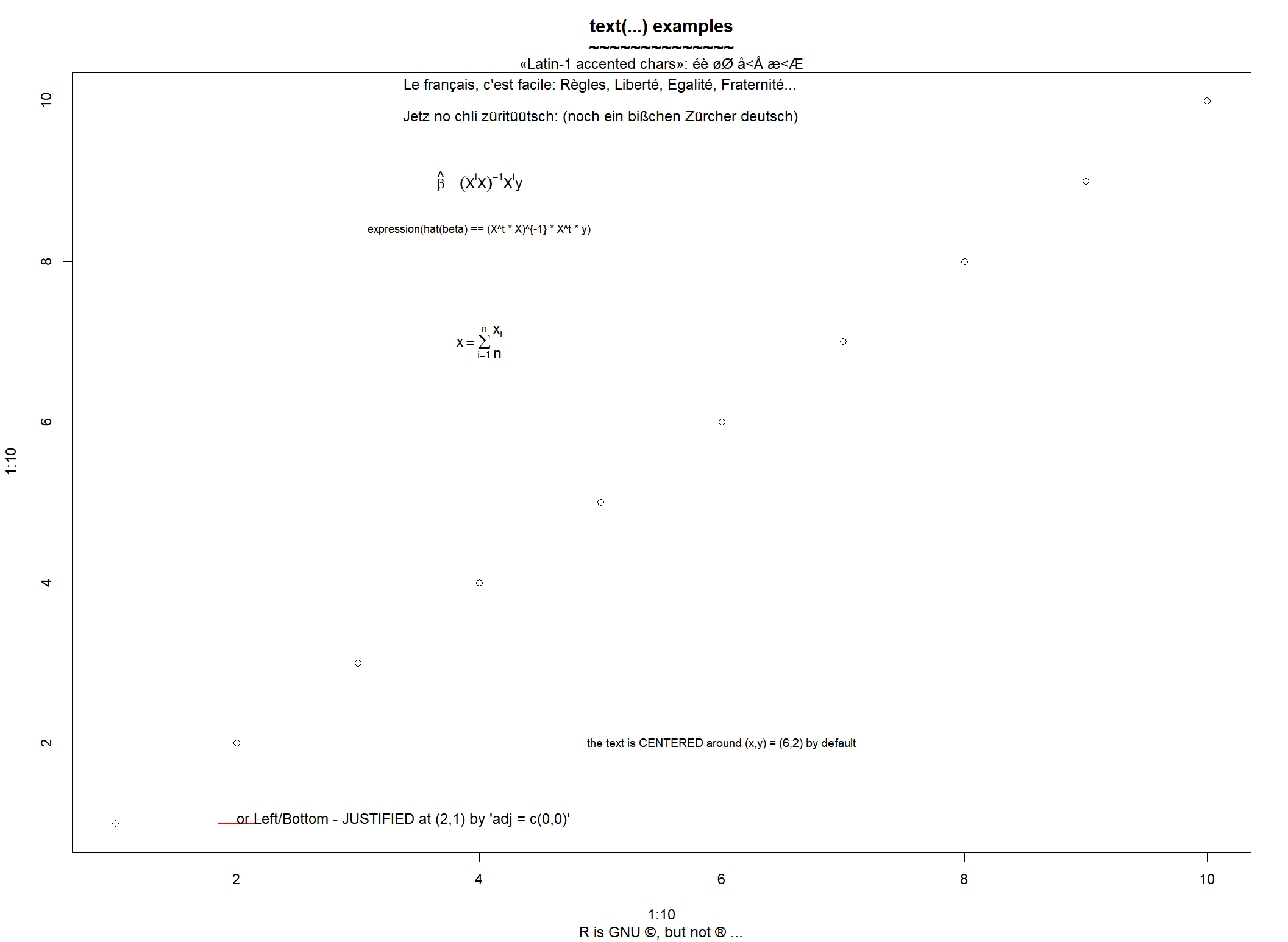
> plot(mtcars$mpg,mtcars$disp,col=newcyl)

> legend("topright",levels(newcyl),fill = 1:3)



> ?text

Example:



## Exercise 1.10

Investigate the relationship between sepal length and sepal width using scatter plots, correlation coefficients, test of correlation, and linear regression. Again interpret all your results in PLAIN and proper English.

> SP <- as.factor(iris$Species)

> plot(SW,SL,col=SP)

> model2 <- lm(SL~SW)

> summary(model2)

Call:

lm(formula = SL ~ SW)

Residuals:

Min 1Q Median 3Q Max

-1.5561 -0.6333 -0.1120 0.5579 2.2226

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.5262 0.4789 13.63 <2e-16 \*\*\*

SW -0.2234 0.1551 -1.44 0.152

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8251 on 148 degrees of freedom

Multiple R-squared: 0.01382, Adjusted R-squared: 0.007159

F-statistic: 2.074 on 1 and 148 DF, p-value: 0.1519

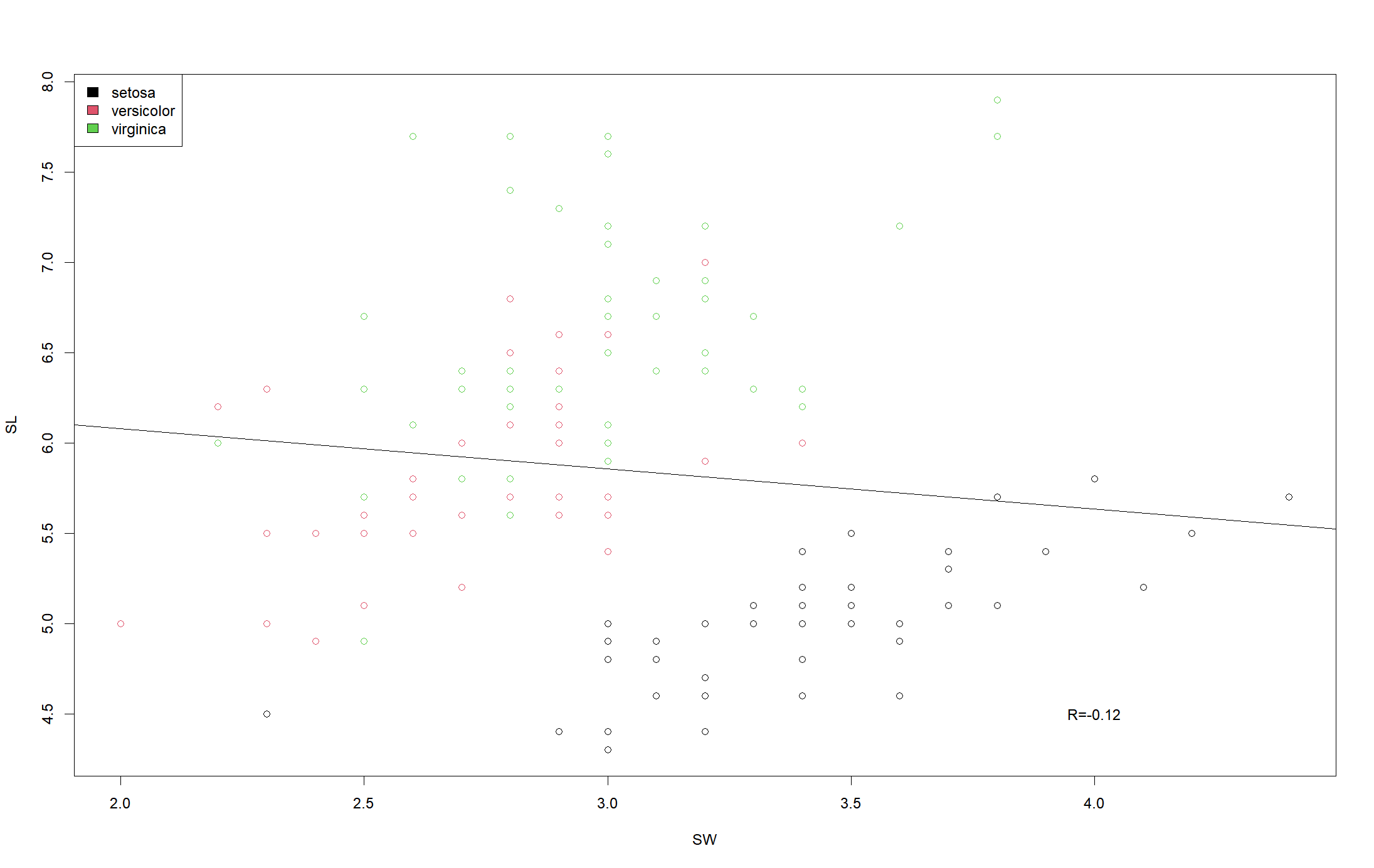
> abline(lm(SL~SW))

> cor(SW,SL)

[1] -0.1175698

> text(4.0,4.5,paste("R=-0.12"))

> plot(model2)



*Interpretation: The correlation coefficient R=-0.12, which means that there is no significant correlation in the relationship between sepal length and sepal width. As we can see from the graph, when explore the relationship between these two variables in each species, the correlation is positive. For example, the relationship between sepal length and sepal width in Setosa is strongly correlated(R=0.74).*

# 1.7 Testing the differences between two groups

## Exercise 1.11

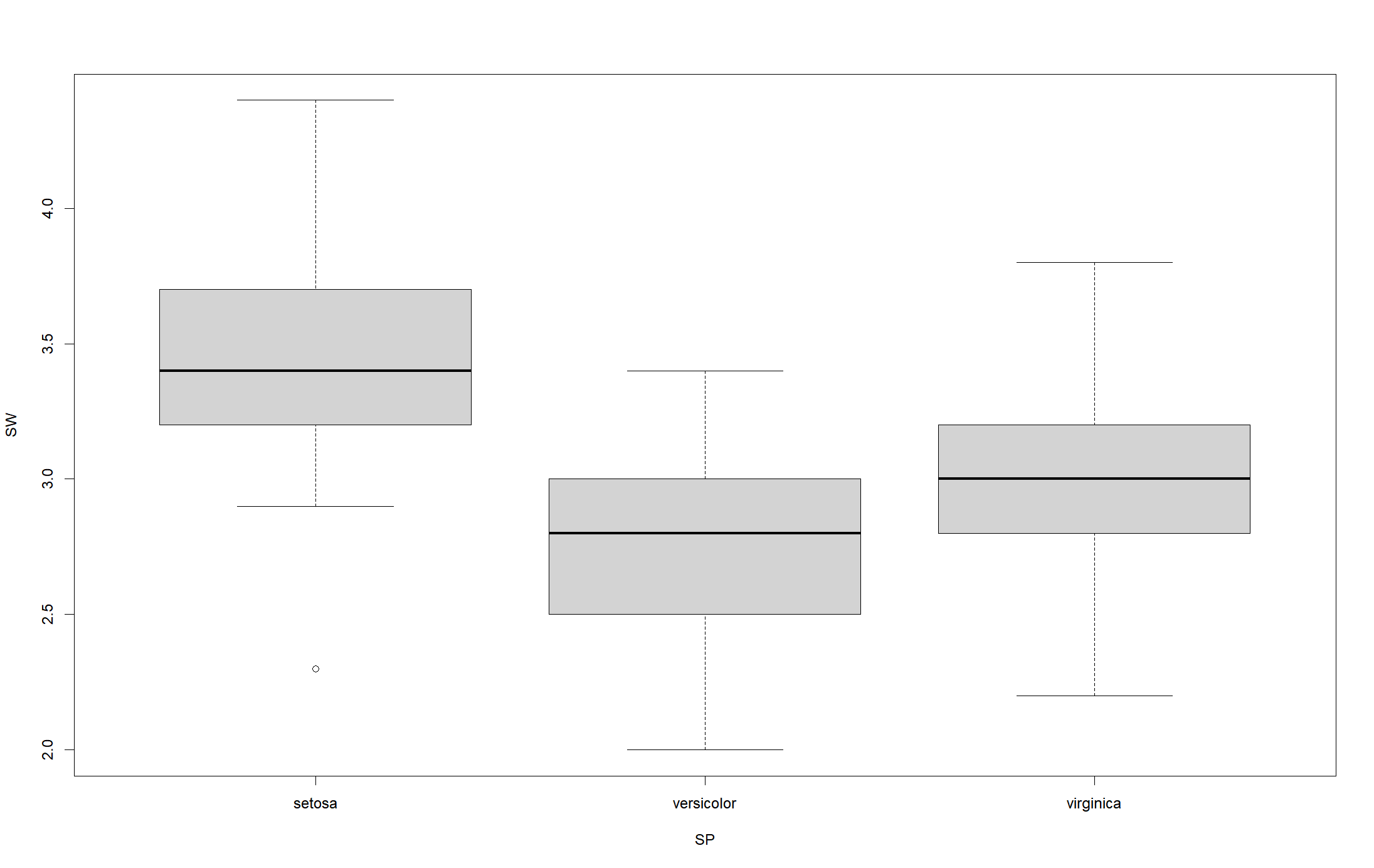
Use boxplot and t-test to investigate whether the means of sepal width between versicolor and virginica are different. Interpret your results.

> boxplot(SW~SP)

> VersiSW <- iris$Sepal.Width[51:100]

> VirgiSW <- iris$Sepal.Width[101:150]

> t.test(VersiSW,VirgiSW)



Welch Two Sample t-test

data: VersiSW and VirgiSW

t = -3.2058, df = 97.927, p-value = 0.001819

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.33028364 -0.07771636

sample estimates:

mean of x mean of y

2.770 2.974

*Interpretation: In this two-sample t-test, our null hypothesis is that the mean of sepal width for*versicolor*is the same as*virginica*. As p=0.002 indicates under this hypothesis, it is unlikely to observe the difference of 0.240 cm through random sampling. Hence, we reject that hypothesis and conclude that the means for the two species are significantly different.*

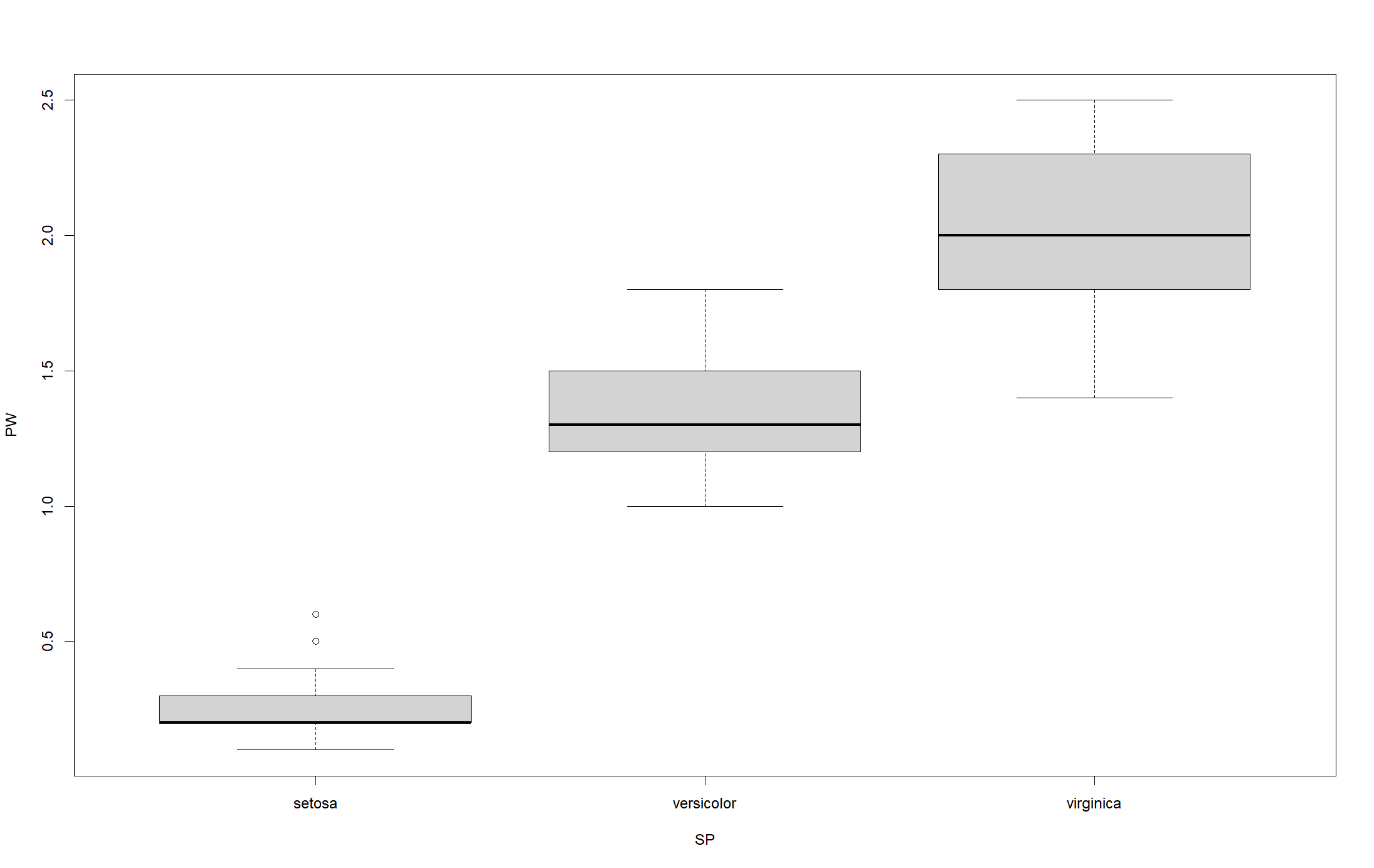
# 1.8 Testing the difference among multiple groups (ANOVA)

## Exercise 1.12

Use boxplot and ANOVA to investigate whether the means of septal width are the same for the 3 species.

> PW <- iris$Petal.Width

> boxplot(PW~SP)



> summary(aov(PW~SP))

Df Sum Sq Mean Sq F value Pr(>F)

SP 2 80.41 40.21 960 <2e-16 \*\*\*

Residuals 147 6.16 0.04

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Interpretation: The p-value is extremely smaller than 0.05, so we reject the null hypothesis. We can conclude that not all the 3 species have the same mean of petal width.*

# Chapter 2 Visualizing the iris flower data set

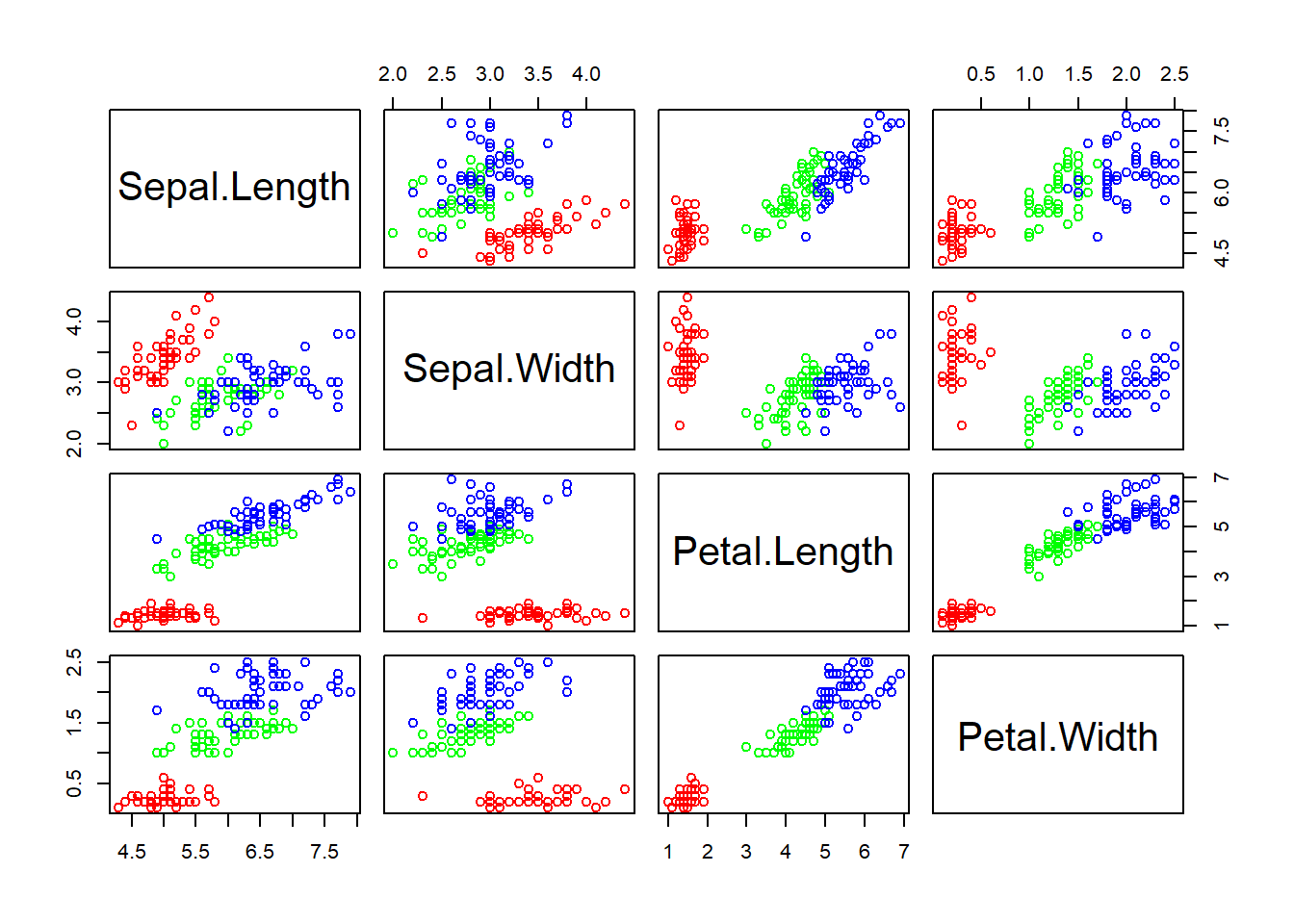
# 2.2 Scatter plot matrix

## ****Exercise 2.1****

Interpret the scatter plot matrix of Figure [2.3](https://gexijin.github.io/learnR/visualizing-the-iris-flower-data-set.html#fig:spm2-2) .

pairs(ma)

pairs(ma, col = rainbow(3)[speciesID]) *# set colors by species*



*Interpretation: This scatter plot matrix displays the relationship between any two of the variables: sepal length, sepal width, petal length, and petal width.*

## ****Exercise 2.2****

Plot a scatter plot matrix for dataset **mtcars**. Set different point symbols and colors based on the types of cyl.

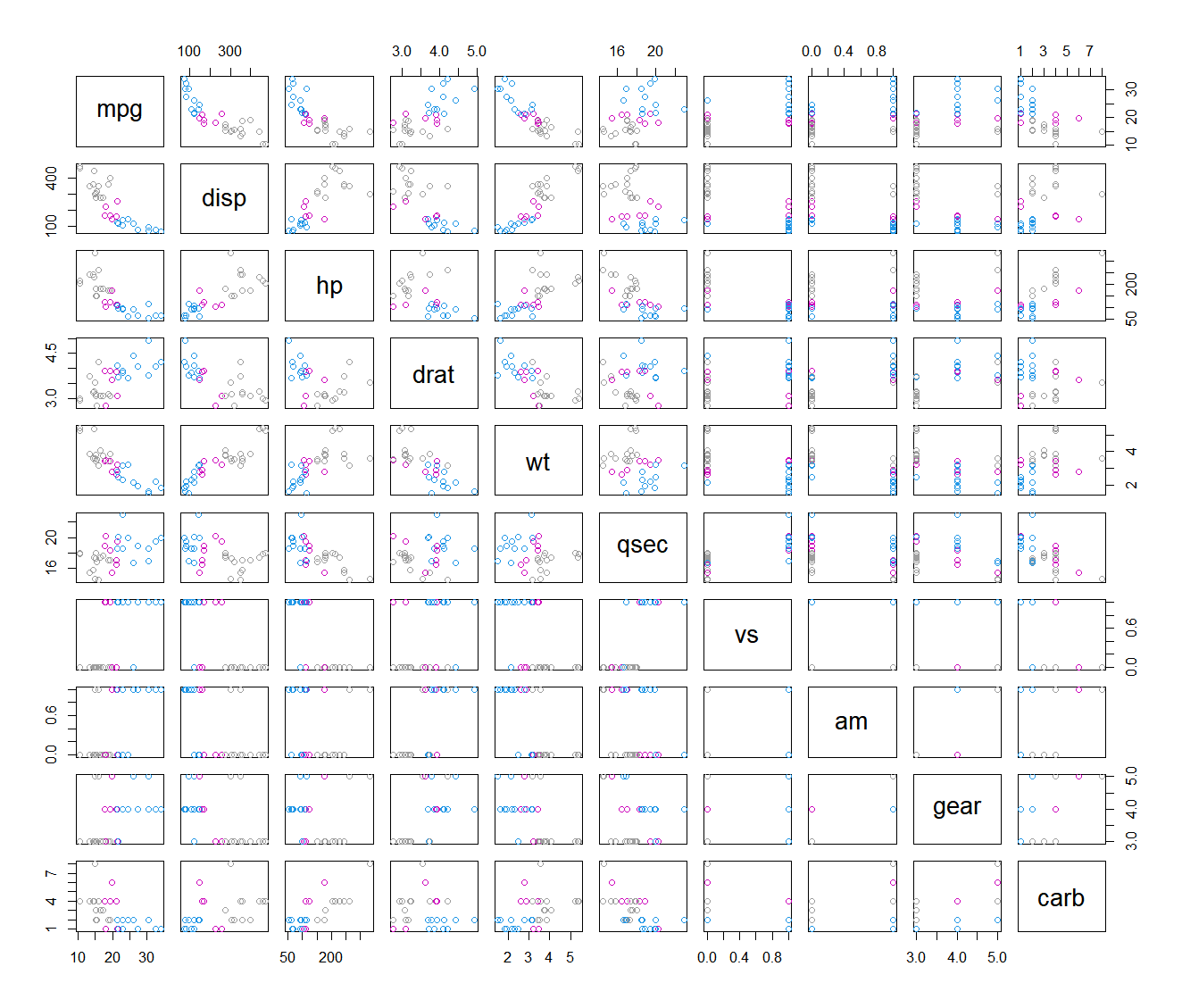
> View(mtcars)

> cylID <- as.numeric(mtcars$cyl)

> mtcars2 <- mtcars[,c(1,3:11,2)]

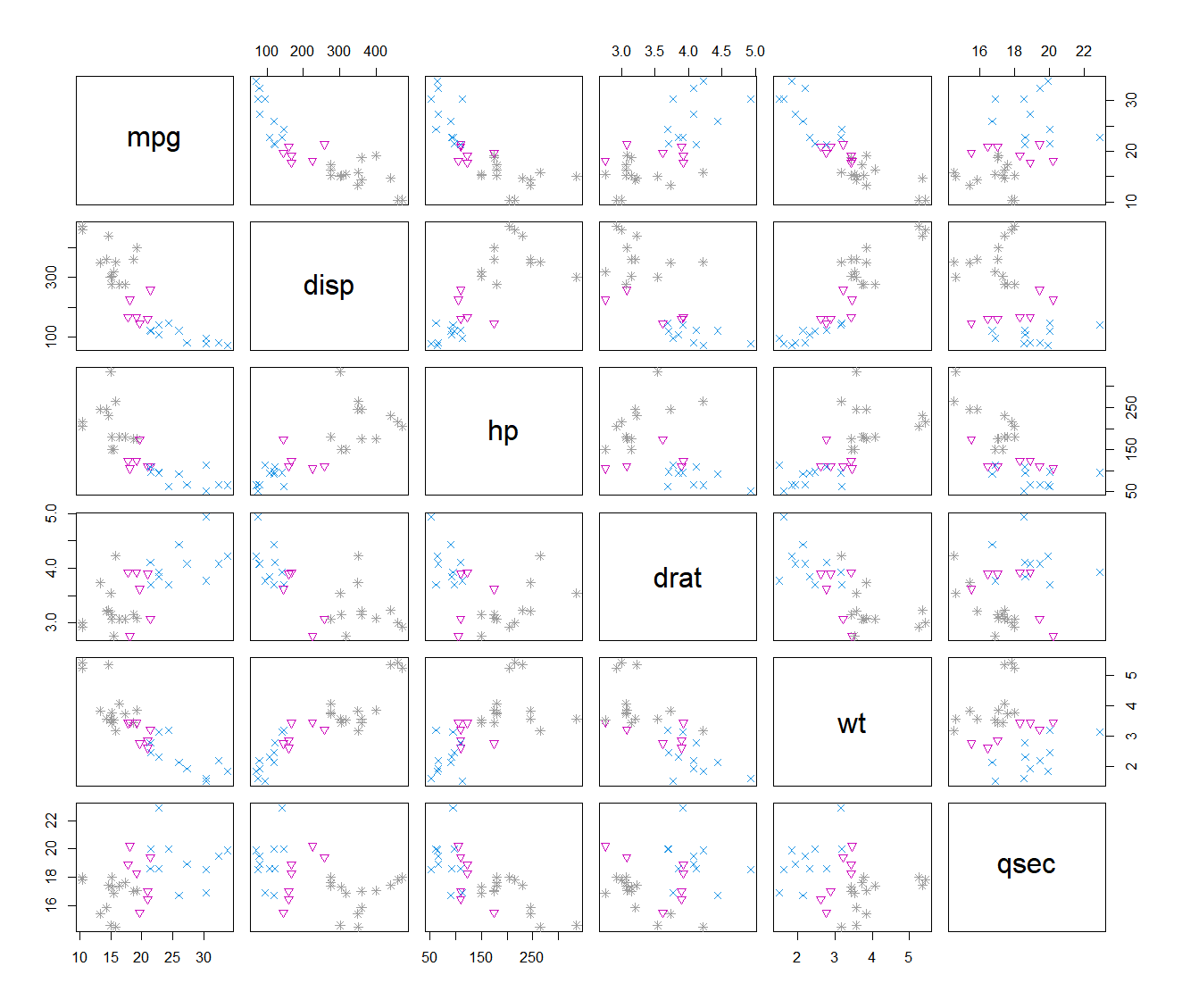
> mtx <- as.matrix(mtcars2[,1:10])

> pairs(mtx,col=cylID)

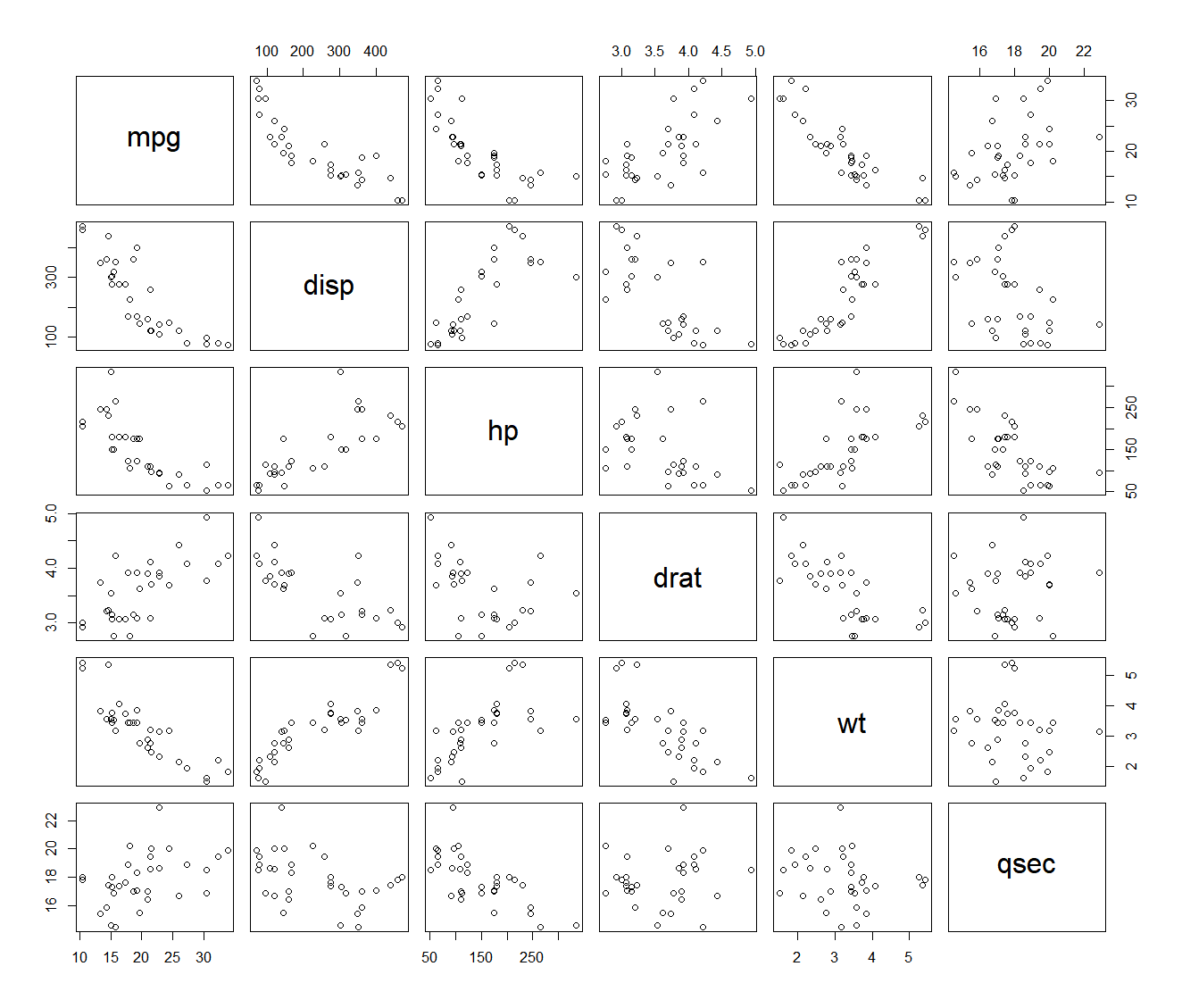


> mtx2 <- as.matrix(mtcars2[,1:6])

> pairs(mtx2,pch=cylID,col=cylID)



> pairs(mtx2)



With colors and shapes, it’s much clearer to see the differences.

# 2.3 Star and segment diagrams

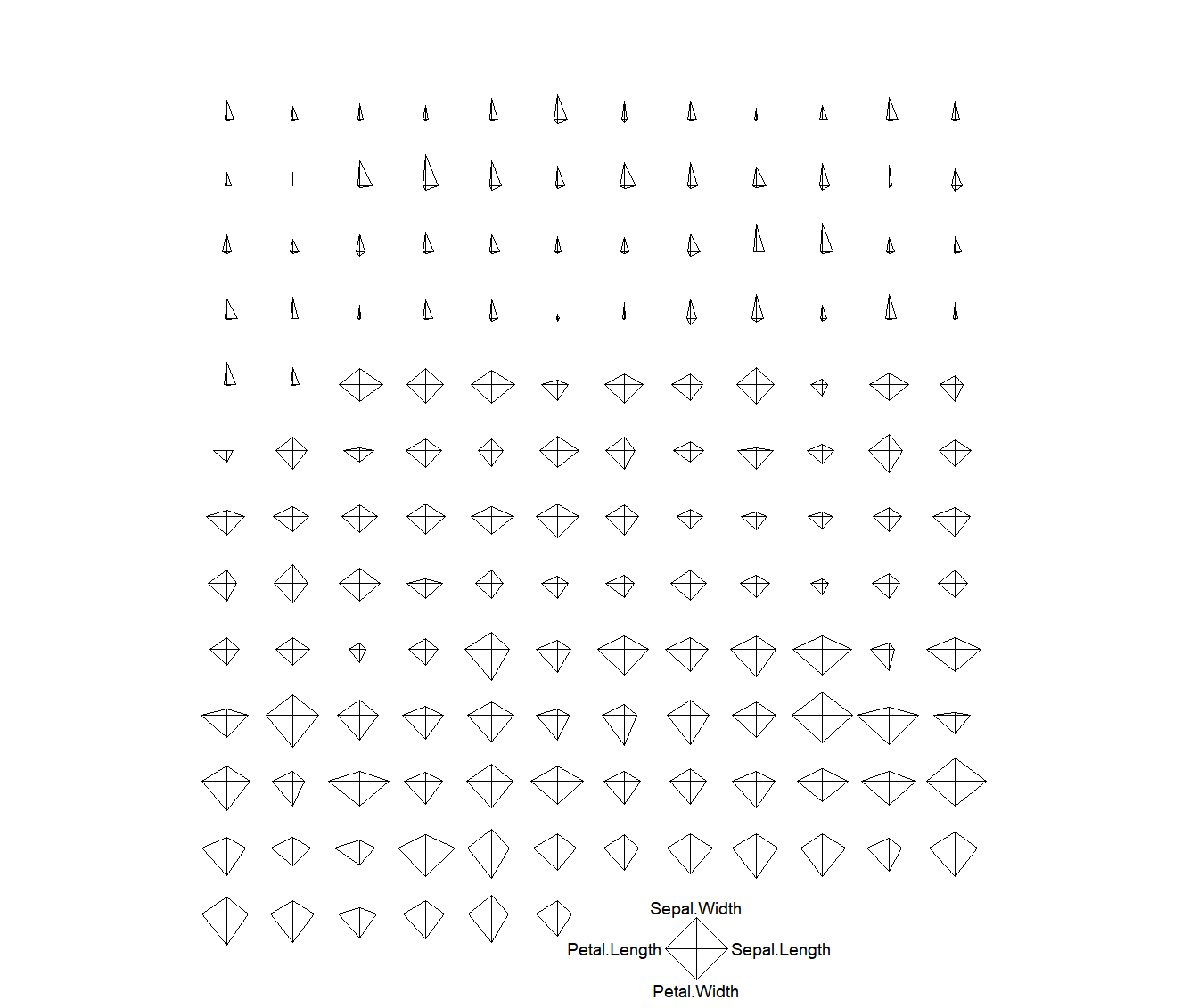
## Exercise 2.3

Based on the star plot, what is your overall impression regarding the differences among these 3 species of flowers?

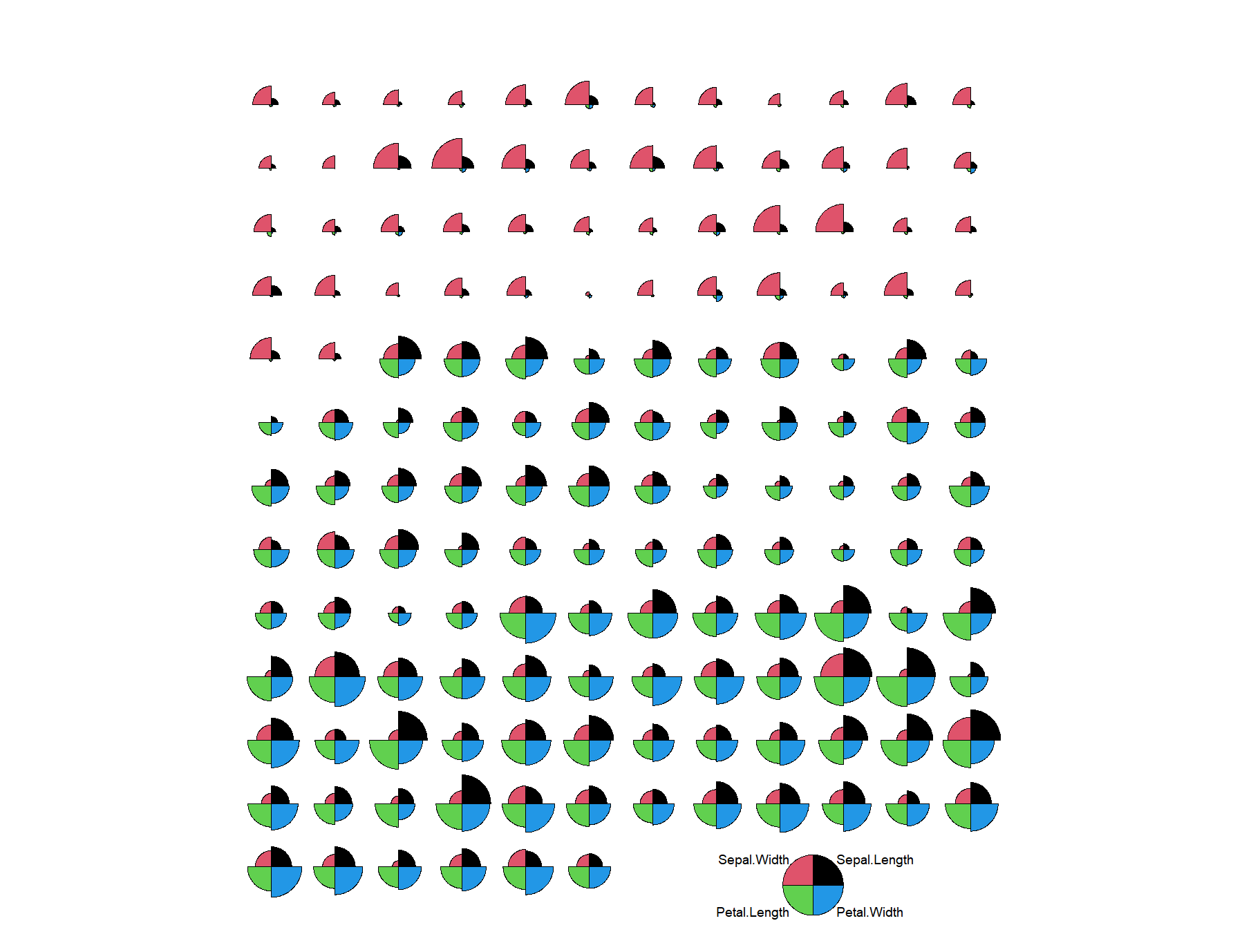
> df <- iris[, 1:4]

> stars(df)

> stars(df, key.loc = c(17, 1))



> stars(df,key.loc=c(20,1.5),draw.segments = TRUE)



*The segments diagram of iris flowers shows a clear difference between the three species. Setosa has greater sepal width than the other factors. Versicolor and Virginica are similar in the proportion of all the four variables, but Virginica has greater sepal width, sepal length, petal length, and petal width in general.*

## Exercise 2.4

Produce the segments diagram of the state data (state.x77) and offer some interpretation of South Dakota compared with other states. Hints: Convert the matrix to a data frame using *df.state.x77 <- as.data.frame(state.x77)*, then plot the segments diagram based on the dataset *df.state.x77*.

# 2.4 The ggplot2 package is intuitive and powerful

## Exercise 2.5

Create a scatter plot of sepal length vs sepal width, change colors and shapes with species, and add trend line.

> cor(iris$Sepal.Width,iris$Sepal.Length)

[1] -0.1175698

> library(ggplot2)

> ggplot(data=iris)+

+ aes(x=Sepal.Length,y=Sepal.Width)+

+ geom\_point(aes(color=Species, shape=Species))+

+ geom\_smooth(method = lm)+

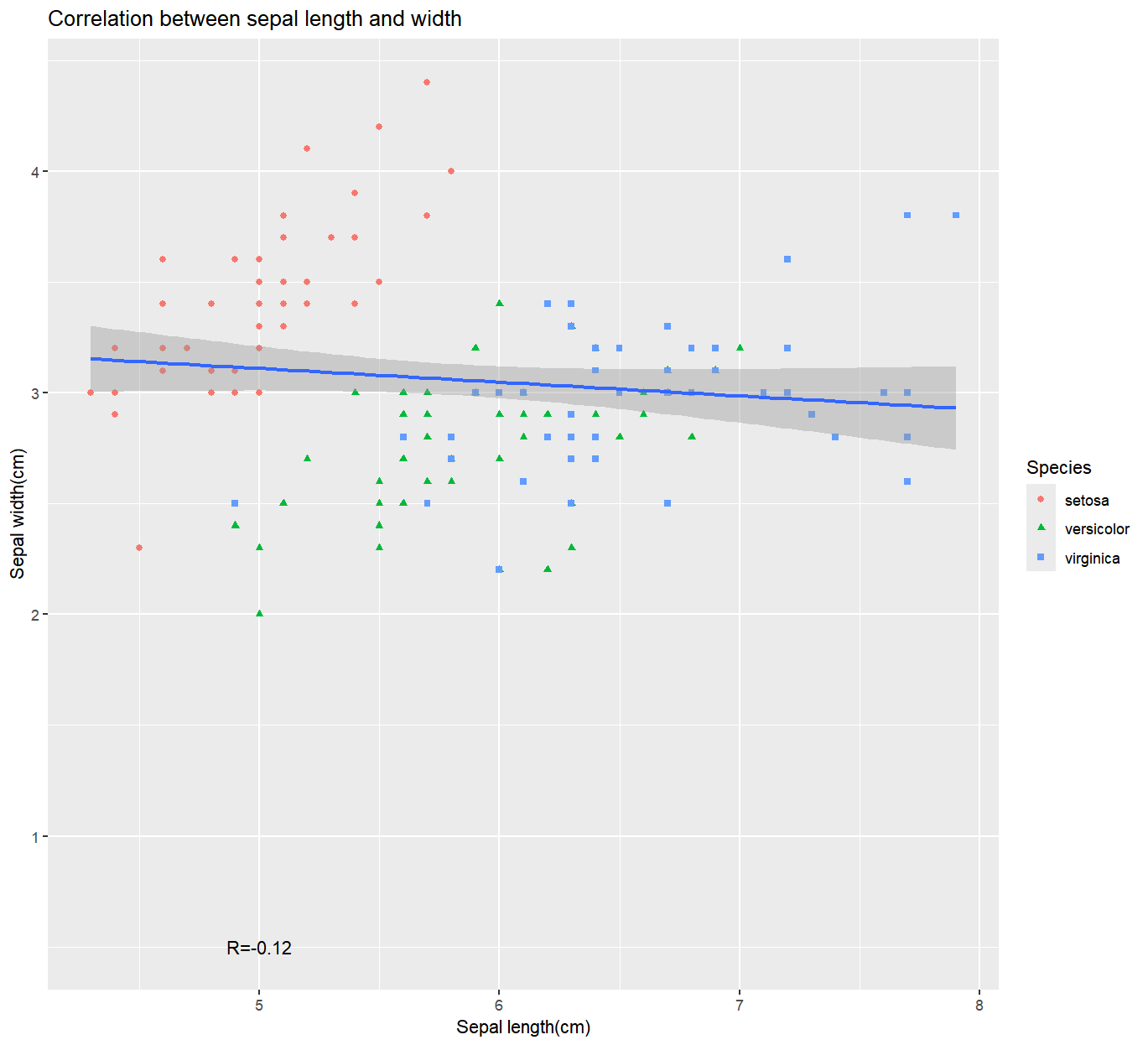
+ annotate("text",x=5,y=0.5, label="R=-0.12")+

+ xlab("Sepal length(cm)")+

+ ylab("Sepal width(cm)")+

+ ggtitle("Correlation between sepal length and width")

`geom\_smooth()` using formula = 'y ~ x'



# 2.5 Other types of plots with ggplot2

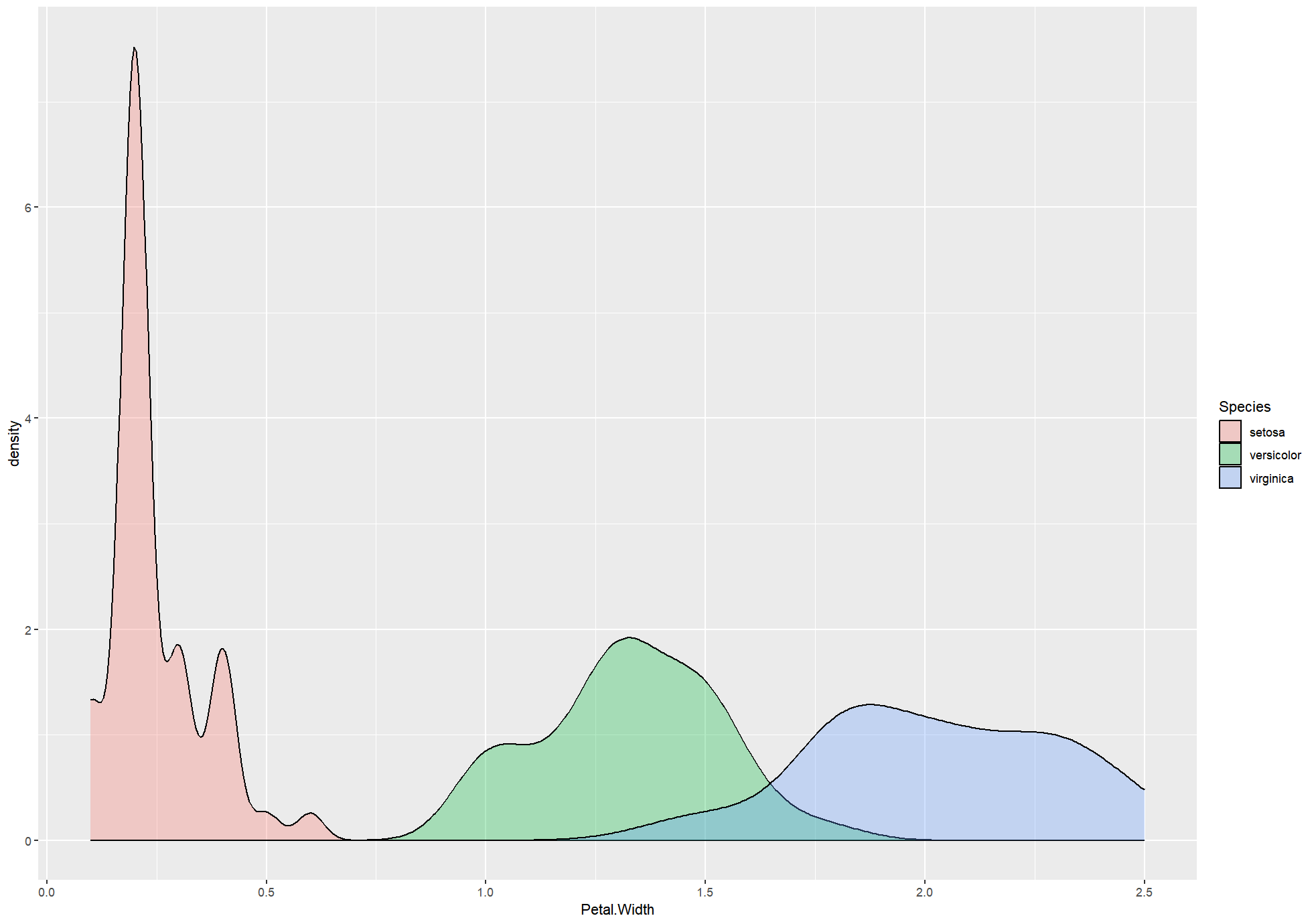
## ****Exercise 2.6****

Use boxplot, and density plots to investigate the similarity and differences of petal width of three species in the **iris** dataset.

> ggplot(iris)+

+ aes(x=Petal.Width, fill=Species)+

+ geom\_density(alpha=0.3)

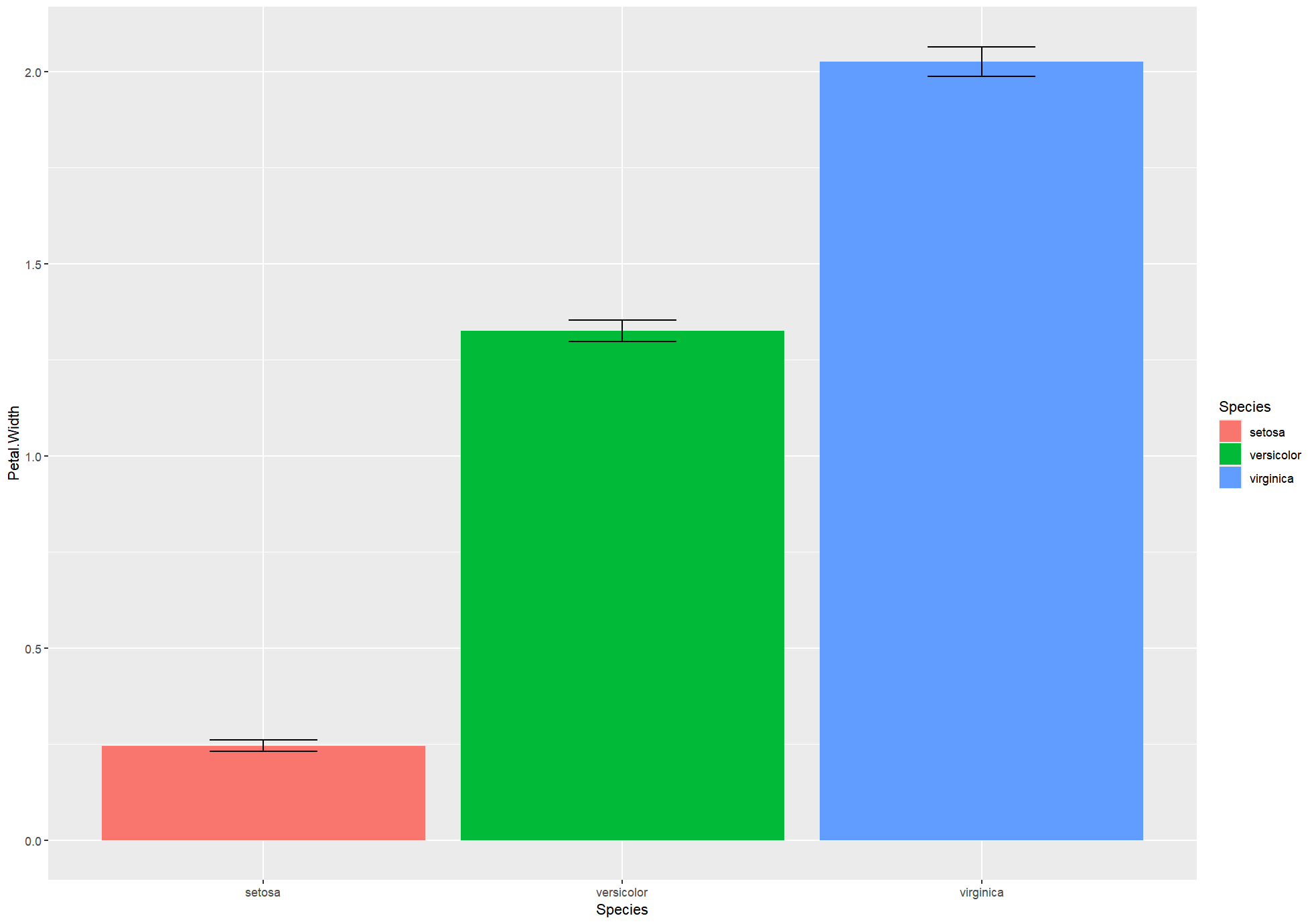


> ggplot(iris) +

+ aes(x = Species, y = Petal.Width, fill = Species) +

+ stat\_summary(geom = "bar", fun = "mean") +

+ stat\_summary(geom = "errorbar", fun.data = "mean\_se", width = .3)



# 2.7 Principal component analysis (PCA)

**Exercise 2.7**

Create PCA plot of the state.x77 data set (convert matrix to data frame). Use the state.region information to color code the states. Interpret your results. Hint: do not forget normalization using the scale option.

# 2.8 Classification by predicting the odds of binary outcomes

**Exercise 2.8**

So far, we used a variety of techniques to investigate the iris flower dataset. Recall that in the very beginning, I asked you to eyeball the data and answer two questions:

1. What distinguishes these three species?
2. If we have a flower with sepals of 6.5cm long and 3.0cm wide, petals of 6.2cm long, and 2.2cm wide, which species does it most likely belong to?

Review all the analysis we did, examine the raw data, and answer the above questions. Write a paragraph and provide evidence of your thinking. Do more analysis if needed.