# Homework 5

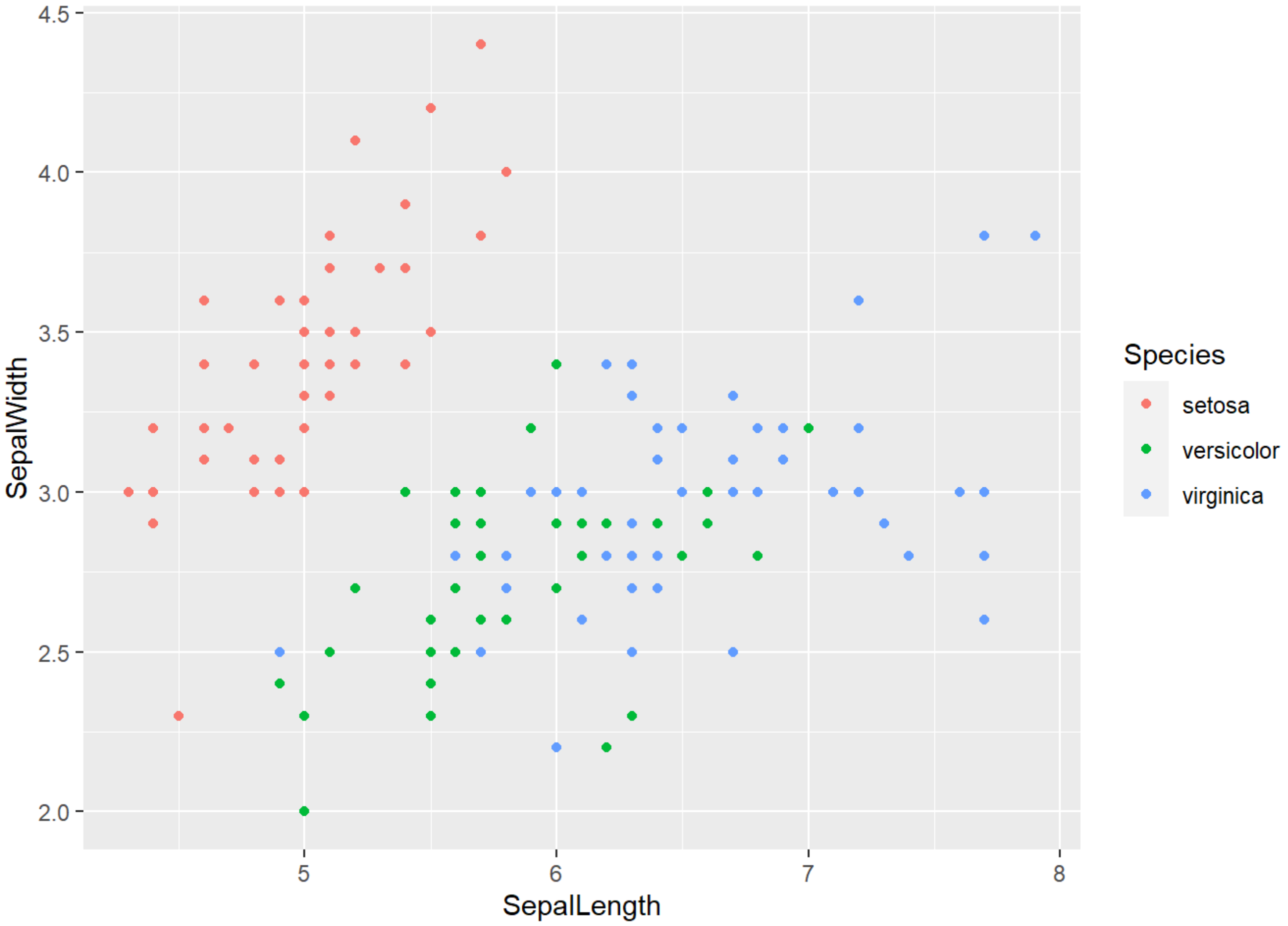
## Lisa Baumgärtner

These are my solutions to the homework 5 of the data science course in SS23.

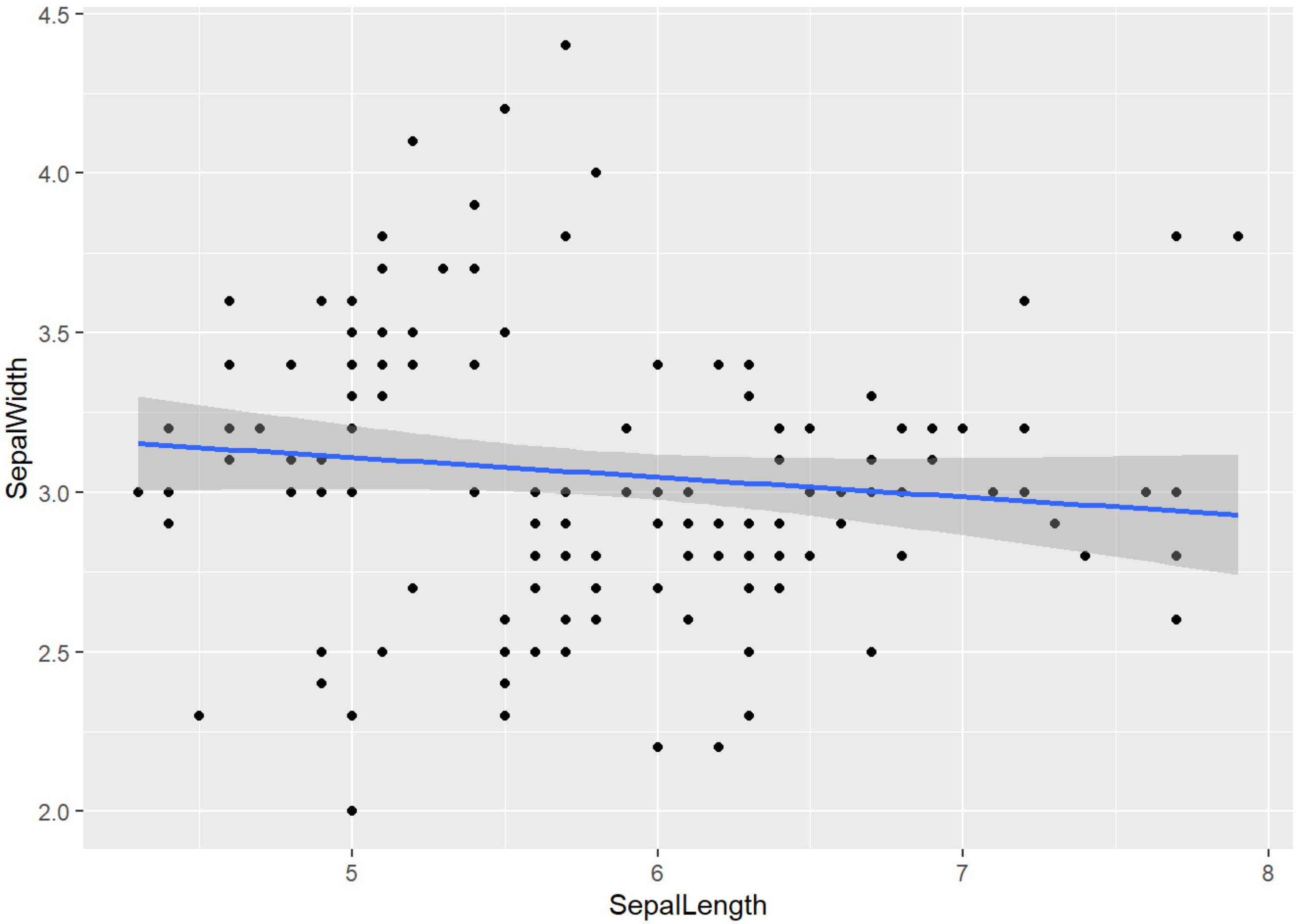
```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  
##                 
##                 
## 
```

```
## make scatter plot with x-axis: Sepal.Length and y-axis: Sepal.Width -> Species should be shown in different colors
SepalLength <- iris$Sepal.Length
SepalWidth <- iris$Sepal.Width
ggplot(iris, aes(x=SepalLength, y=SepalWidth, colour=Species)) + geom_point()
```



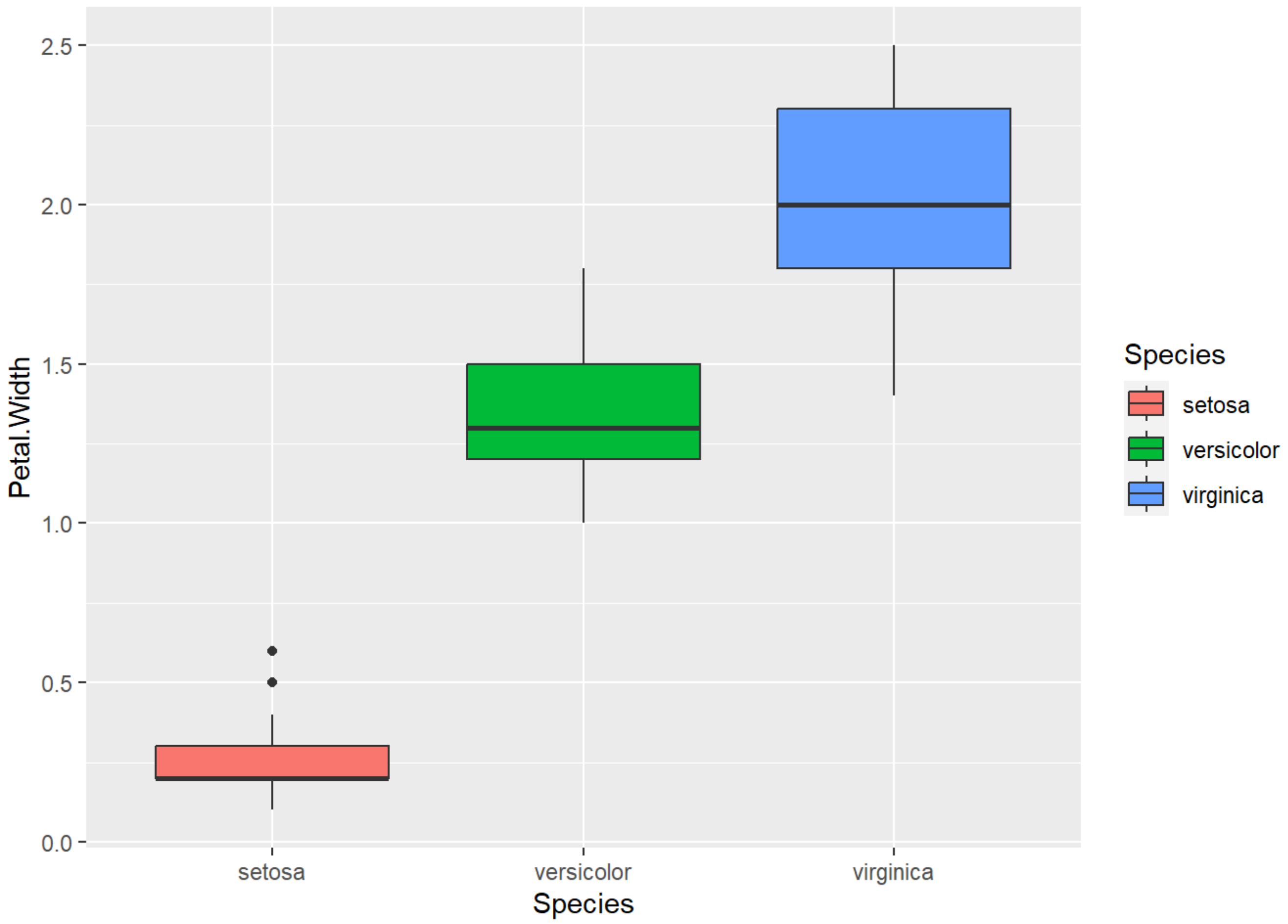
```
## add regression line for the previous plot with the whole dataset (regardless of the species)
reg_plot <- ggplot(iris, aes(x=SepalLength, y=SepalWidth)) + geom_point()
reg_plot + geom_smooth(method='lm', formula= y~x)
```



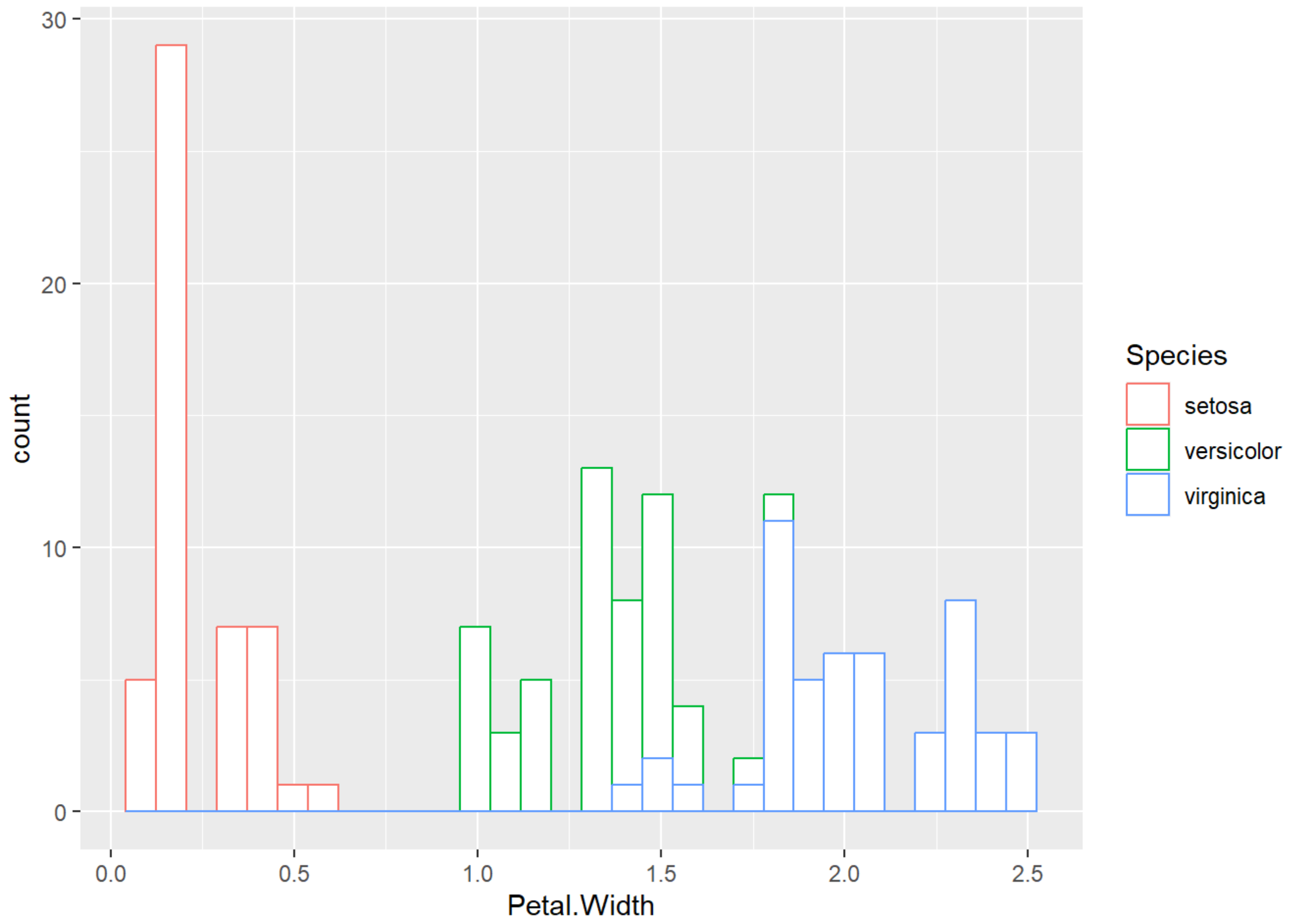
```
## calculate the Pearson correlation for this plot
cor.test(iris$Sepal.Length, iris$Sepal.Width, method = "pearson", conf.level = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Sepal.Length and iris$Sepal.Width
## t = -1.4403, df = 148, p-value = 0.1519
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.27269325  0.04351158
## sample estimates:
##        cor
## -0.1175698
```

```
## make the box plot for Petal.Width with 3 species separately in the x-axis in different colors
box_plt <- ggplot(iris, aes(x=Species,y=Petal.Width, fill=Species))
box_plt + geom_boxplot()
```



```
# make the histogram for Petal.Width with 3 species separately in x-axis in different colors
ggplot(iris, aes(x=Petal.Width, color=Species)) + geom_histogram(fill="White", bins = 30)
```



```
## run the t-test of Petal.Width between setosa and virginica, and give the conclusion if the width is a statistically significant difference between 2 species
iris = filter(iris, Species != "versicolor" ) # remove versicolor
#summary(iris)
t.test(Petal.Width~Species, data = iris)
```

```
##
##  Welch Two Sample t-test
##
## data:  Petal.Width by Species
## t = -42.786, df = 63.123, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group setosa and group virginica is not equal to 0
## 95 percent confidence interval:
##  -1.863133 -1.696867
## sample estimates:
##    mean in group setosa mean in group virginica
##                   0.246                   2.026
```

p-value < 2.2e-16 which is < 0.05 , it is significant