

R

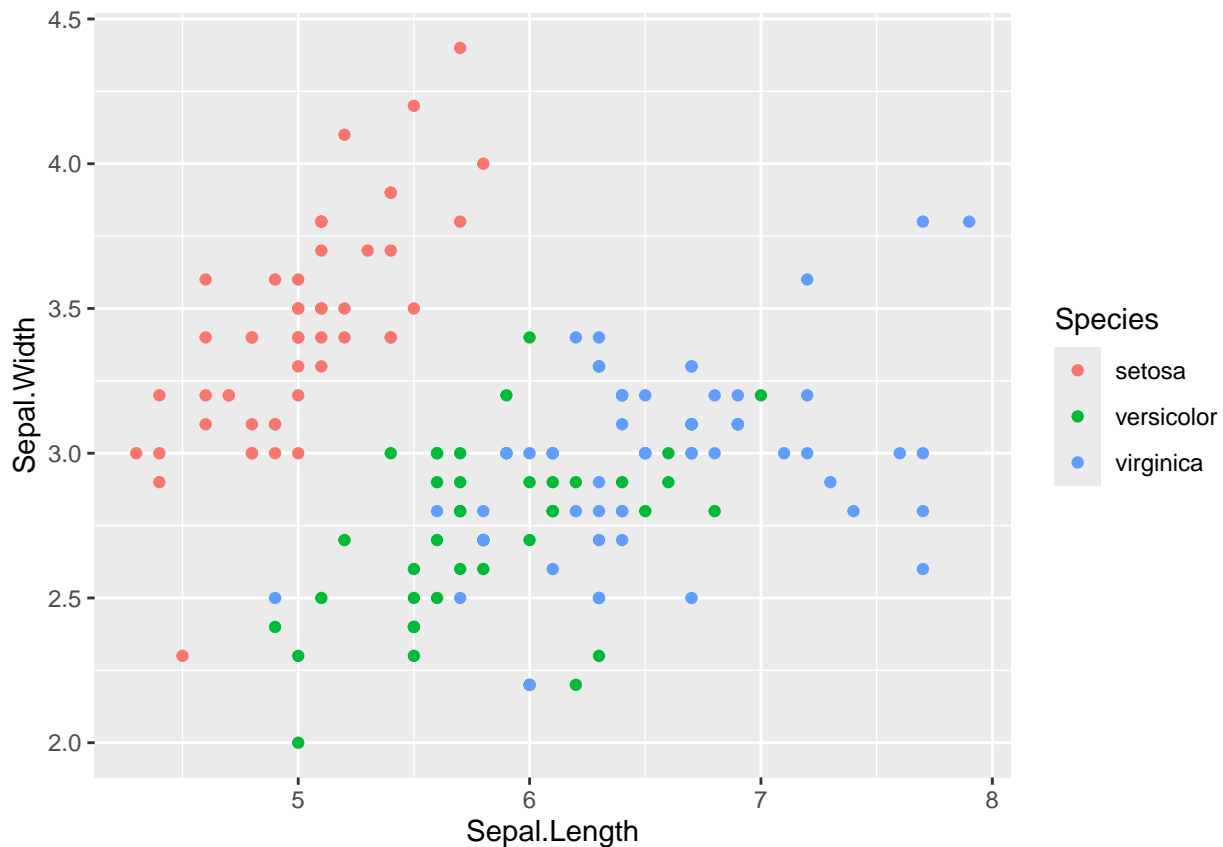
download package ggplot2

import the data("iris")

```
data("iris")
```

make a scatter plot with x-axis: Sepal.Length and y-axis: Sepal.Width, and the species should be shown in different colors

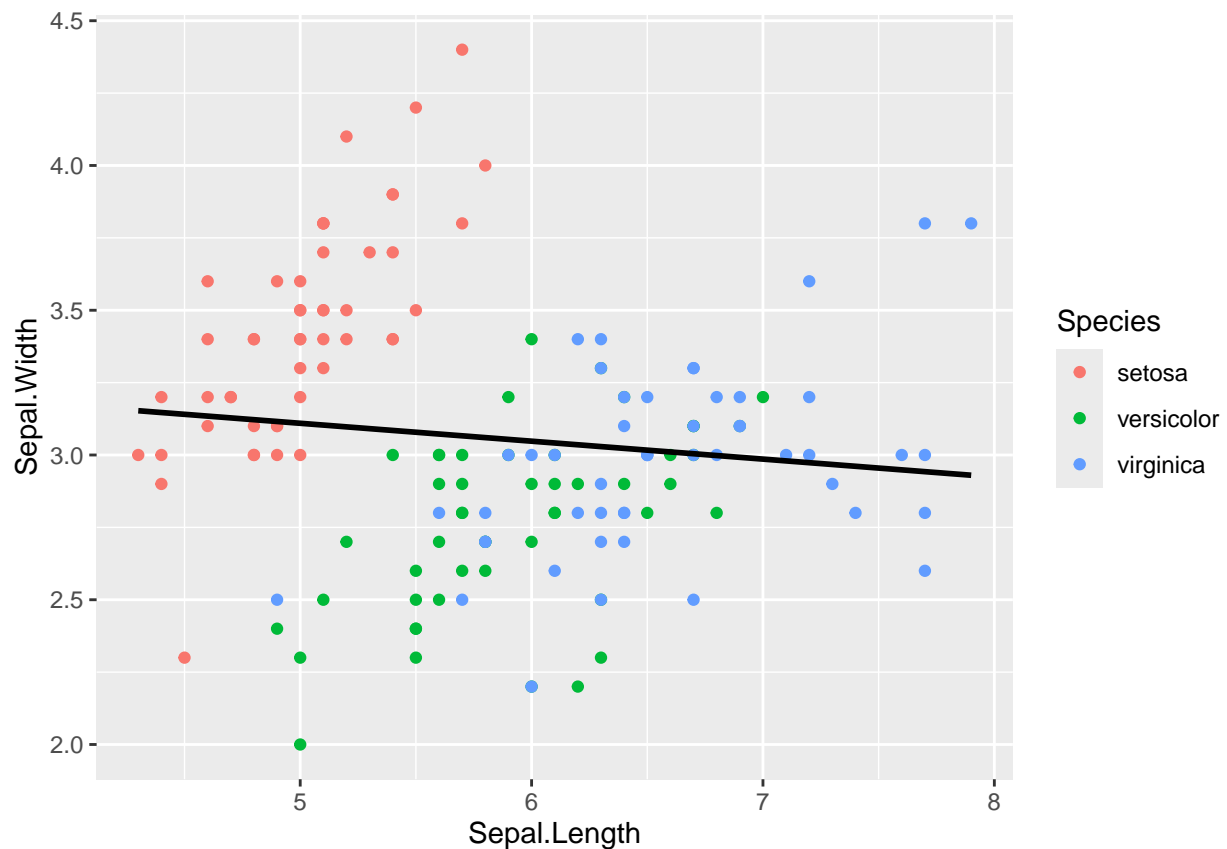
```
print( p <- ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, color=Species)) + geom_point() )
```



add regression line for the previous plot with the whole dataset (regardless of the species)

```
p + geom_smooth(method="lm", se=FALSE, color="black") # reusing above plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



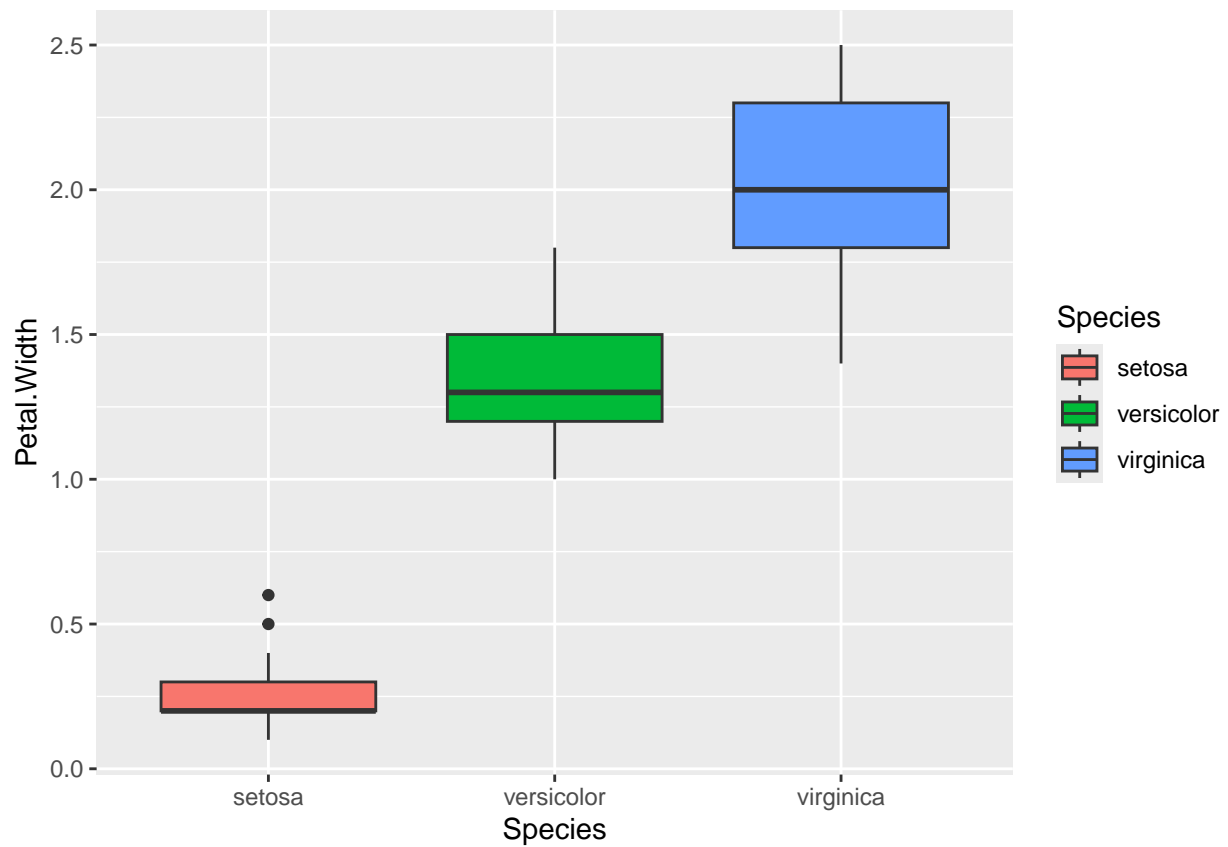
calculate the Pearson correlation for this plot

```
cor.test(iris$Sepal.Length, iris$Sepal.Width, method="pearson")
```

```
##
##  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 boxplot for Petal.Width with 3 species separately in the x-axis in different colors

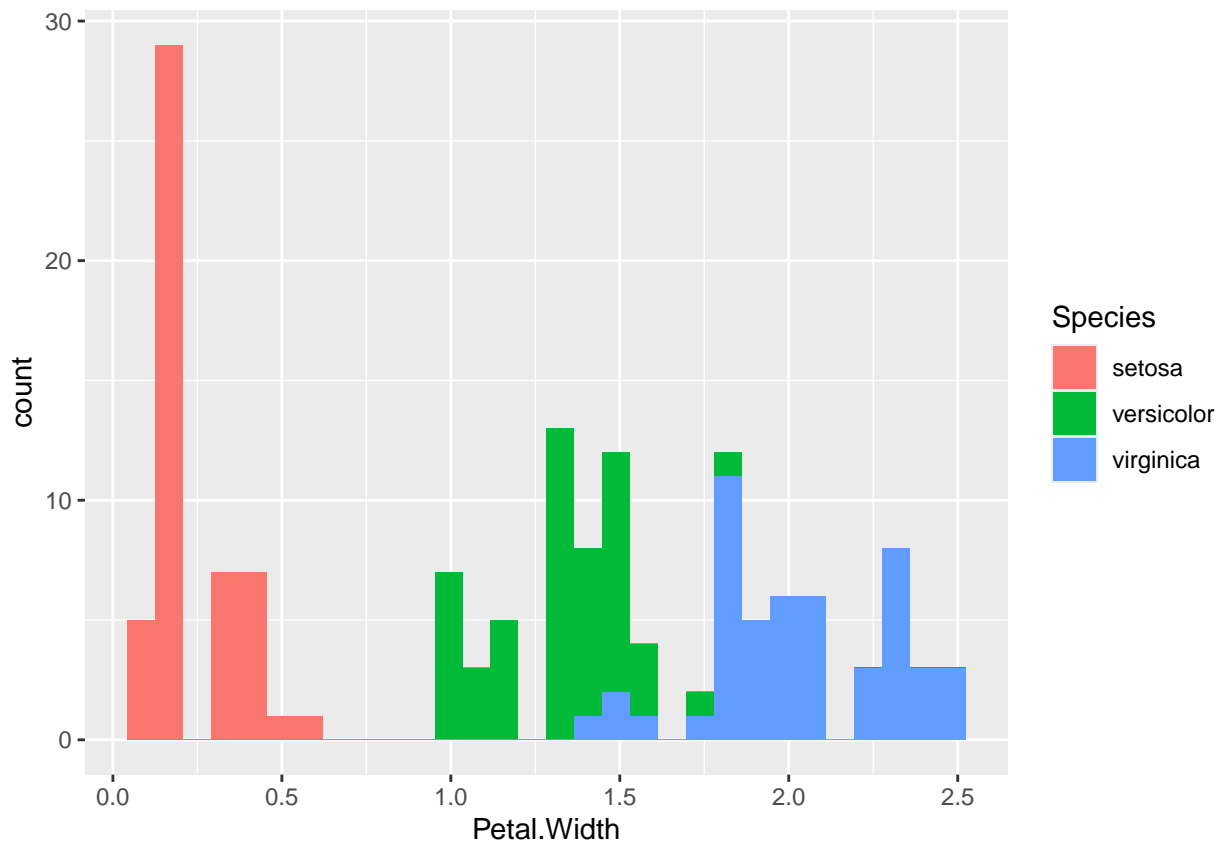
```
ggplot(iris, aes(x=Species, y=Petal.Width, fill=Species)) + geom_boxplot()
```



make the histogram for Petal.Width with 3 species separately in x-axis in different colors

```
ggplot(iris, aes(x=Petal.Width, fill=Species)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
t_test <- t.test(
  subset(iris, Species == "setosa")$Petal.Width,
  subset(iris, Species == "virginica")$Petal.Width,
)
cat("Conclusion: The petal width between setosa and virginica is statistically", if(t_test$p.value > 0.05) "not" else "is", "significant.\n")
## Conclusion: The petal width between setosa and virginica is statistically significant.
```