Correlation and Analysis on Iris Data

Correlation plot using the IRIS data. Which variables have the highest correlation coefficient?

```
corr <- round(cor(iris.dt[ , c(1:4)]),2) corr
```

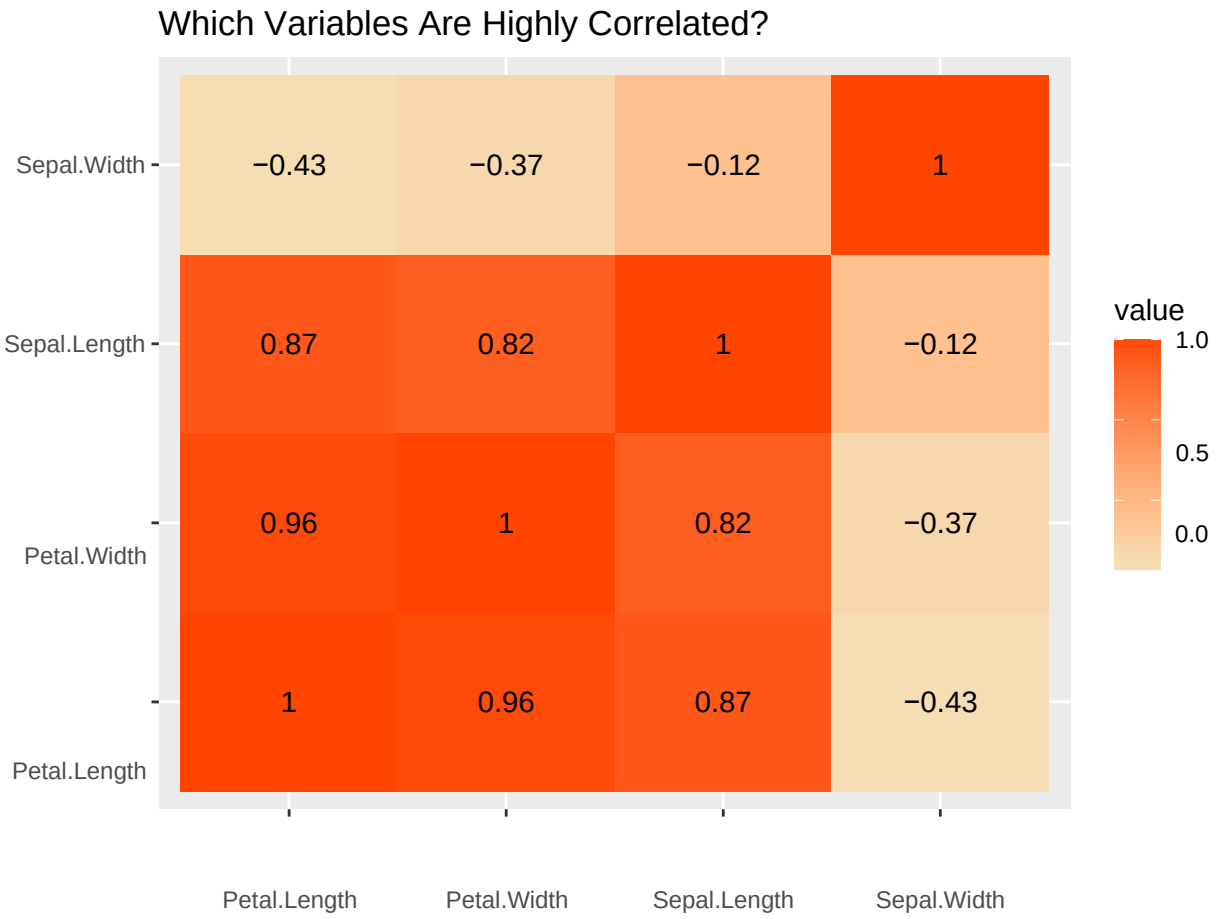
```
library(datasets)
iris.dt <- as.data.table(iris)
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length         1.00       -0.12         0.87        0.82
## Sepal.Width         -0.12        1.00        -0.43       -0.37
## Petal.Length         0.87       -0.43         1.00        0.96
## Petal.Width          0.82       -0.37         0.96        1.00
```

```
melted.corr <- melt(corr) head(melted.corr)
```

```
##          X1           X2 value
## 1 Sepal.Length Sepal.Length  1.00
## 2  Sepal.Width Sepal.Length -0.12
## 3 Petal.Length Sepal.Length  0.87
## 4  Petal.Width Sepal.Length  0.82
## 5 Sepal.Length  Sepal.Width -0.12
## 6  Sepal.Width  Sepal.Width  1.00
```

```
ggplot(melted.corr, aes(x =X1,y =X2,fill =value))                    +
  scale_fill_gradient(low="wheat",high="orangered")                  +
  geom_tile() +labs (x =NULL,y =NULL)                          +
  geom_text(data =melted.corr,          aes(x =X1,y =X2,label =value))                  +
  ggtitle("Which Variables Are Highly Correlated?")
```



Petal.Length and Petal.Width have the highest correlation with 0.96. Petal.Length and Petal.Width also have a high correlation with Sepal.Length having 0.87 and 0.82 correlation coefficients.

Calculate average (mean) values of the numeric variables in the data using data.table package. Which variable has the highest mean?

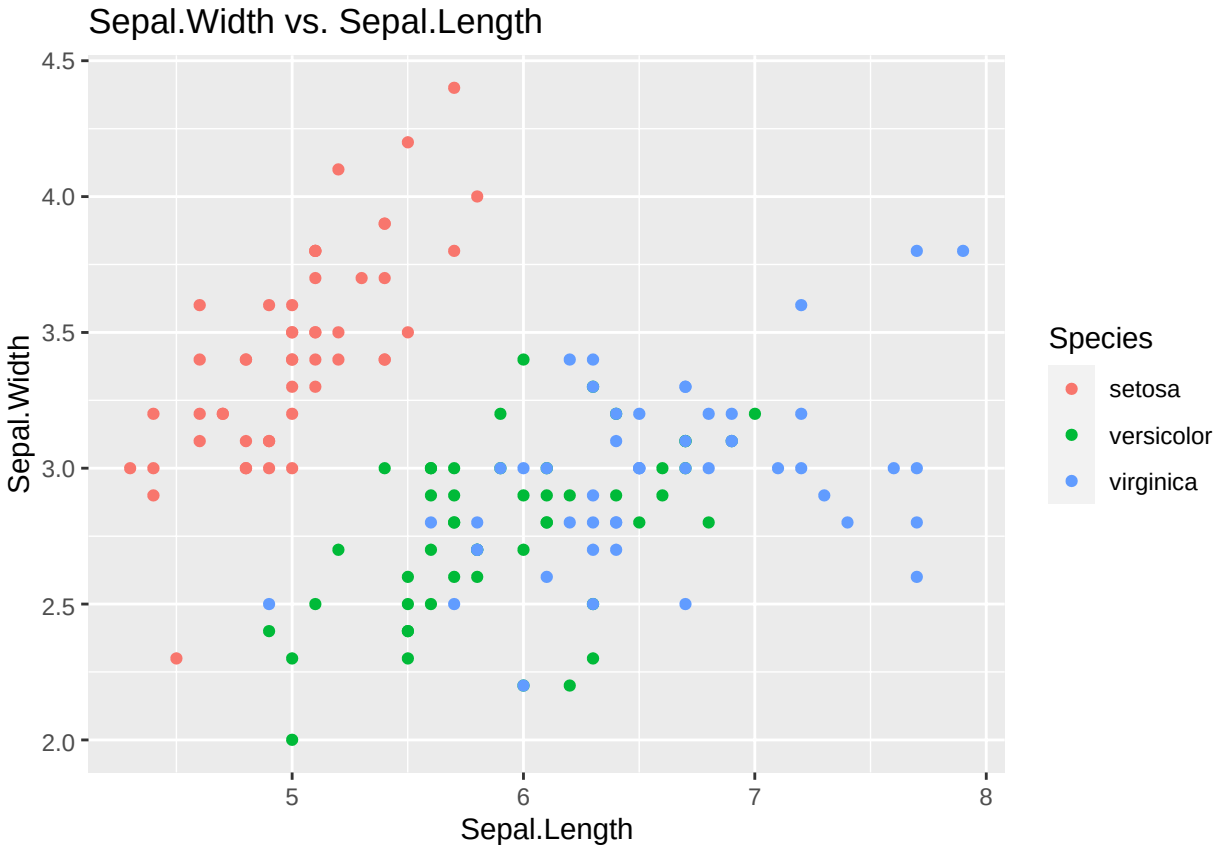
```
avg <-iris.dt[,          sapply(.SD, mean), .SDcols =1 :4] avg
```

```
## Sepal.Length    Sepal.Width Petal.Length    Petal.Width
##     5.843333       3.057333     3.758000    1.199333
```

Sepal.Length has the highest mean.

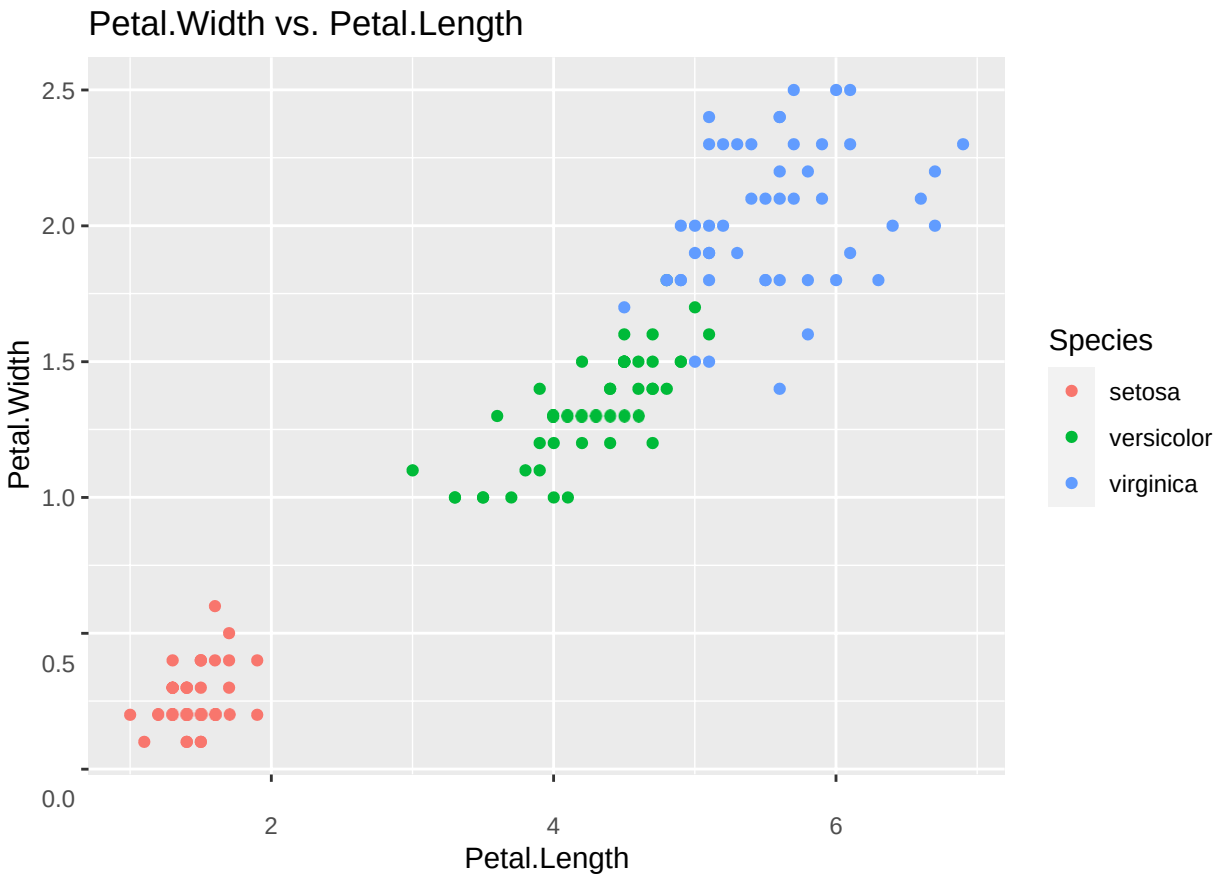
Scatterplot showing the relationship between Sepal.Length and Sepal.Width variable, using ggplot2 package. Color code the points using Species variable.

```
ggplot(data =iris.dt)          +
  geom_point(mapping = aes(x =Sepal.Length,y =Sepal.Width,color=Species))                    +
  ggtitle("Sepal.Width vs. Sepal.Length ")
```



Scatterplot showing the relationship between Petal.Length and Petal.Width variables, using ggplot2 package. Color code the points using Species variable.

```
ggplot(data =iris.dt)          +
  geom_point(mapping = aes(x =Petal.Length,y =Petal.Width,color=Species))                    +
  ggtitle("Petal.Width vs. Petal.Length ")
```



Combination of variables, creates better separation among records of different Species.

The plot comparing Petal.Length and Petal.Width create a greater separation among records. There is much less overlap betwwen species versicolor and virginica in this plot as cmpared to the plot of Sepal.Length vs Sepal.Width.

Accuracy of the model? How does it compare with the No Information Rate (NIR)?

Accuracy is (48 + 47)/100 = 0.95. The model performs much better than the 50% NIR

Sensitivity of the model, assuming the class of interest is Setosa.

Sensitivity is 48/(48 + 2) = 0.96

Specificity of the model, assuming the class of interest is Setosa.

Specificity is 47/(47 + 3) = 0.94

Precision of the model, assuming the class of interest is Setosa.

Precision is 48/(48 + 3) = 0.9412