Project 4

Code

JONATHAN FENG

Get Your Data

Hide

```
data(iris)
iris
```

Sepal.Length <dbl>	Sepal.Width <dbl>	Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
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
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

1-10 of 150 rows Previous **1** 2 3 4 5 6 ... 15 Next

Hide

```
names(iris) <- c("Sepal.Length", "Sepal.Width", "Petal.Length",
"Petal.Width", "Species")
```

Initial Overview of The Data Set

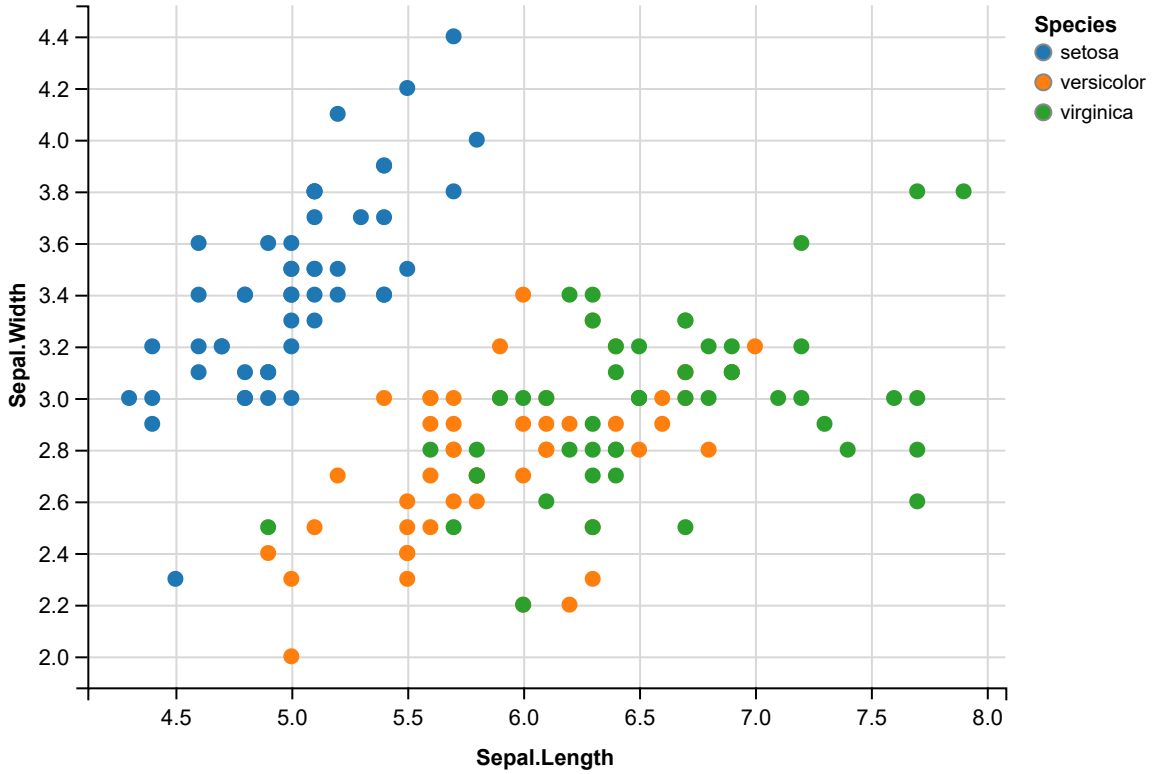
Hide

```
library(ggvis)
```

```
package 愀牶ggvis愀牷 was built under R version 4.0.5
```

Hide

```
iris %>% ggvis(~Sepal.Length, ~Sepal.Width, fill = ~Species) %>%
layer_points()
```



Hide

```
NA
```

QUESTION 1

What is the output? Based on the plot, which specie(s) has the highest correlation between the sepal length and the sepal width?

The output is above. Based on the plot Setosa has the highest correlation between the sepal length and width because the spread is closer together and it is distributed as a fairly linear trend. Setosa has the correlation on calculation as well shown below.

Hide

```
vir <- subset(iris, iris$Species == "virginica")
s <- subset(iris, iris$Species == "setosa")
color <- subset(iris, iris$Species == "versicolor")

cor(s$Sepal.Length, s$Sepal.Width)
```

```
[1] 0.7425467
```

Hide

```
cor(color$Sepal.Length, color$Sepal.Width)
```

```
[1] 0.5259107
```

Hide

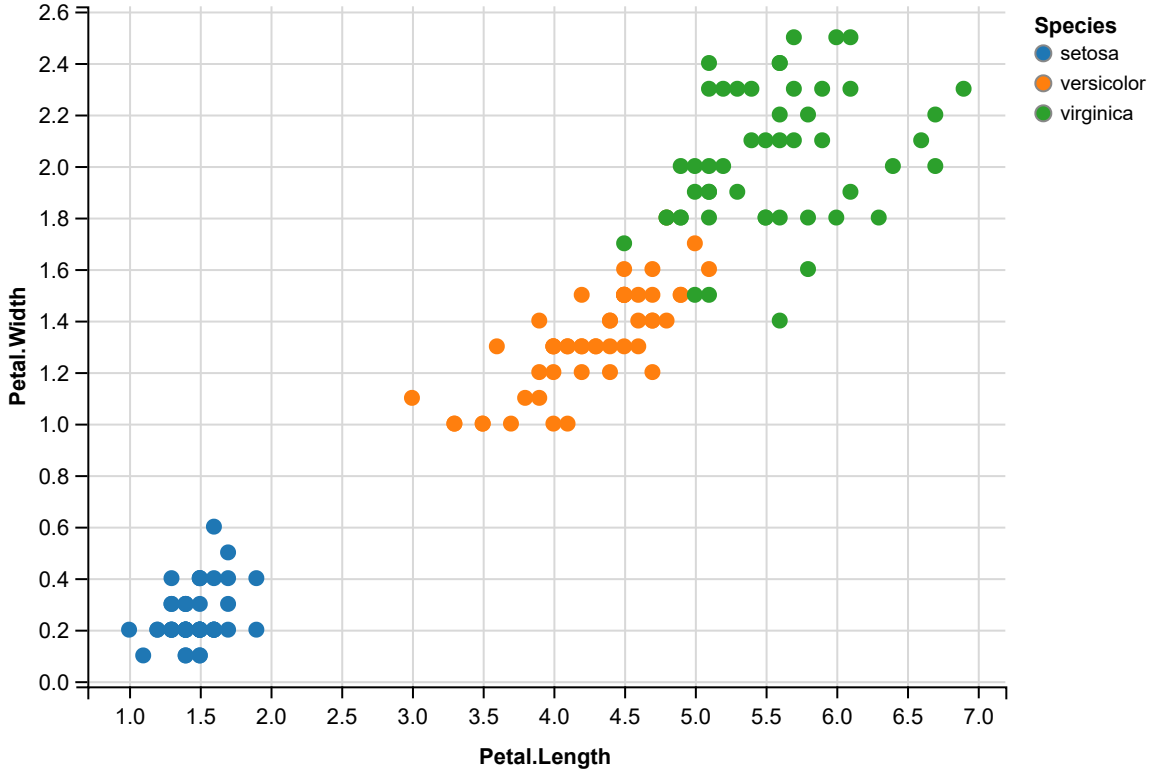
```
cor(vir$Sepal.Length, vir$Sepal.Width)
```

```
[1] 0.4572278
```

Now, let's try the scatter plot that maps the petal length and the petal:

Hide

```
iris %>% ggvis(~Petal.Length, ~Petal.Width, fill = ~Species) %>%
layer_points()
```



QUESTION 2

What is the output? Based on the plot, which specie(s) has the highest correlation between the petal length and the petal width? Is this result consistent with Question 1? Why or why not?

The output is above. Based on the plot versicolor seems to have the highest correlation between the petal length and the petal width, and it seems to have a linear relationship. It is not consistent with the results of question one because you can't attribute different parts of the flower with each other, difference species are expected to have different lengths and widths for petals and sepals. We will run the calculations below.It shows true that versicolor as the highest correlation on petal length and width.

Hide

```
vir <- subset(iris, iris$Species == "virginica")
s <- subset(iris, iris$Species == "setosa")
color <- subset(iris, iris$Species == "versicolor")

cor(s$Petal.Length, s$Petal.Width)
```

```
[1] 0.33163
```

Hide

```
cor(color$Petal.Length, color$Petal.Width)
```

```
[1] 0.7866681
```

Hide

```
cor(vir$Petal.Length, vir$Petal.Width)
```

```
[1] 0.3221082
```

Setting training and labels

Hide

```
set.seed(1234)

ind <- sample(2, nrow(iris), replace=TRUE, prob=c(0.67, 0.33))
iris.training <- iris[ind==1, 1:4]
iris.test <- iris[ind==2, 1:4]

iris.trainLabels <- iris[ind==1, 5]
iris.testLabels <- iris[ind==2, 5]
```

Training KNN and Prediction.

QUESTION 3

What is the output?

Output is below.

Hide

```
library('class')

iris_pred <- knn(train = iris.training, test = iris.test, cl =
iris.trainLabels, k=3)

iris_pred
```

```
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [8] setosa     setosa     setosa     setosa     setosa     versicolor versicolor
[15] versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[22] versicolor versicolor versicolor virginica  virginica  virginica  virginica
[29] versicolor virginica  virginica  virginica  virginica  virginica  virginica
[36] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

QUESTION 4

What is the command that may produce above output?

Command should be data.frame to see all. ALthough I included an extra table of the counts and errors.

Hide

```
Predicted_Species <- iris_pred
Observed_Species <- iris.testLabels

tab <- data.frame(Predicted_Species, Observed_Species)
print.data.frame(tab)
```

```
   Predicted_Species Observed_Species
1             setosa           setosa
2             setosa           setosa
3             setosa           setosa
4             setosa           setosa
5             setosa           setosa
6             setosa           setosa
7             setosa           setosa
8             setosa           setosa
9             setosa           setosa
10            setosa           setosa
11            setosa           setosa
12            setosa           setosa
13        versicolor       versicolor
14        versicolor       versicolor
15        versicolor       versicolor
16        versicolor       versicolor
17        versicolor       versicolor
18        versicolor       versicolor
19        versicolor       versicolor
20        versicolor       versicolor
21        versicolor       versicolor
22        versicolor       versicolor
23        versicolor       versicolor
24        versicolor       versicolor
25         virginica        virginica
26         virginica        virginica
27         virginica        virginica
28         virginica        virginica
29        versicolor        virginica
30         virginica        virginica
31         virginica        virginica
32         virginica        virginica
33         virginica        virginica
34         virginica        virginica
35         virginica        virginica
36         virginica        virginica
37         virginica        virginica
38         virginica        virginica
39         virginica        virginica
40         virginica        virginica
```

Hide

```
table(Predicted_Species, Observed_Species)
```

```
                Observed_Species
Predicted_Species setosa versicolor virginica
       setosa         12          0         0
       versicolor      0         12         1
       virginica       0          0        15
```

QUESTION 5

What is the output?

The output is listed below.

Correct Classification = (12 + 12 + 15) / 40 = 39/40 Error Rate = 1 - (39/40) = 1/40 = 0.025

Hide

```
library(gmodels)
```

```
package 愀牶gmodels愀牛 was built under R version 4.0.5
```

Hide

```
CrossTable(x = iris.testLabels, y = iris_pred, prop.chisq=FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40


                | iris_pred
iris.testLabels |     setosa | versicolor |  virginica |  Row Total |
----------------|------------|------------|------------|------------|
         setosa |         12 |          0 |          0 |         12 |
                |      1.000 |      0.000 |      0.000 |      0.300 |
                |      1.000 |      0.000 |      0.000 |            |
                |      0.300 |      0.000 |      0.000 |            |
----------------|------------|------------|------------|------------|
     versicolor |          0 |         12 |          0 |         12 |
                |      0.000 |      1.000 |      0.000 |      0.300 |
                |      0.000 |      0.923 |      0.000 |            |
                |      0.000 |      0.300 |      0.000 |            |
----------------|------------|------------|------------|------------|
      virginica |          0 |          1 |         15 |         16 |
                |      0.000 |      0.062 |      0.938 |      0.400 |
                |      0.000 |      0.077 |      1.000 |            |
                |      0.000 |      0.025 |      0.375 |            |
----------------|------------|------------|------------|------------|
   Column Total |         12 |         13 |         15 |         40 |
                |      0.300 |      0.325 |      0.375 |            |
----------------|------------|------------|------------|------------|
```