

Basic Data Exploration using Iris

2024-01-19

```
data(iris)
dim(iris) #you can view the dimensions, which are also shown in the top-right pane
```

```
## [1] 150 5
```

```
summary(iris) #provides a brief summary of each variable
```

```
## 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
##
##
##
```

```
names(iris) #this allows you to view the column names easily
```

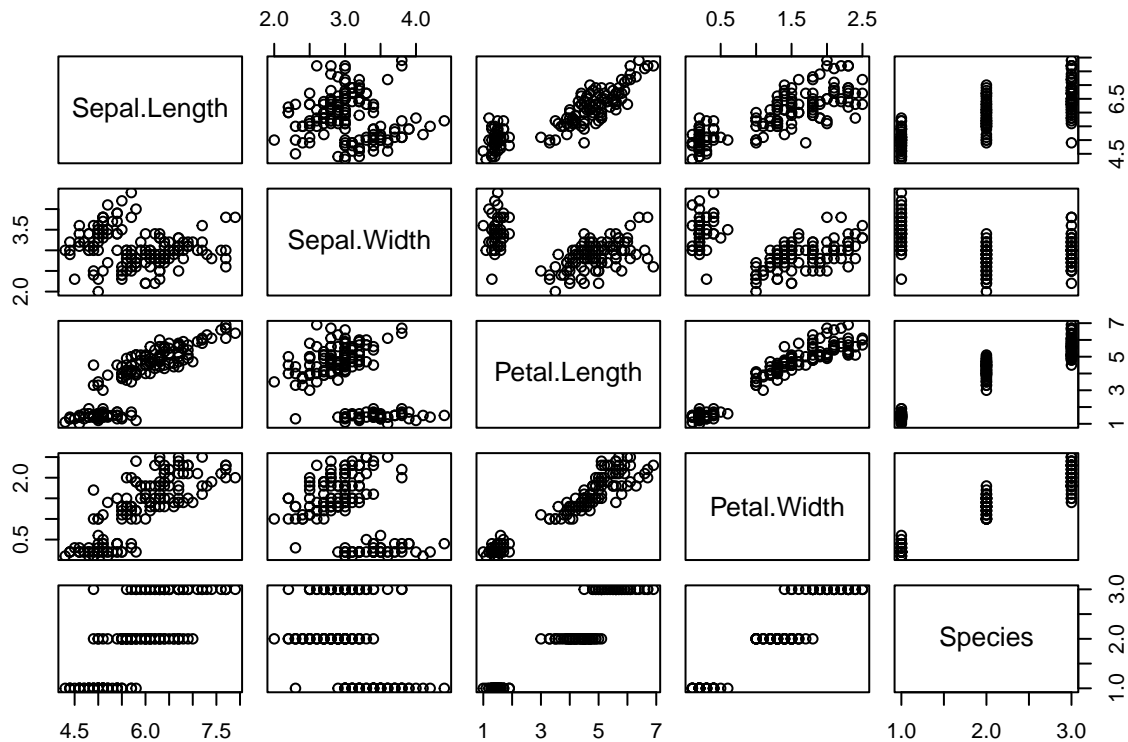
```
## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"
```

```
str(iris) #this provides a quicker alternative to looking at the data
```

```
## 'data.frame': 150 obs. of 5 variables:
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
## $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

Let's try plotting

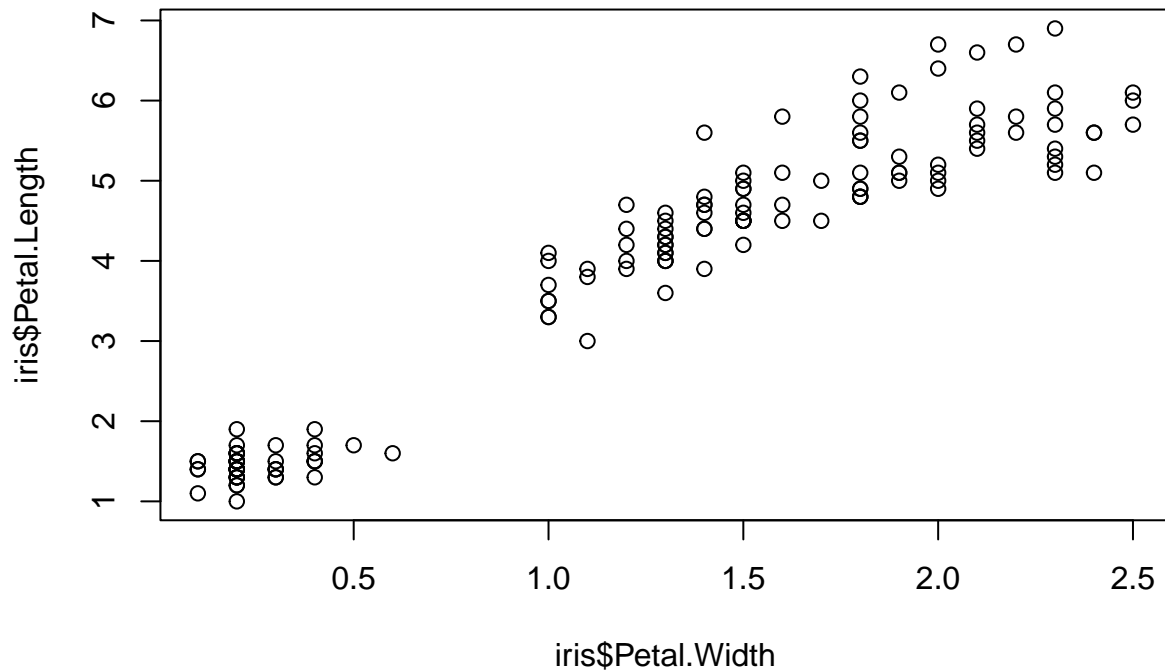
```
plot(iris)
```



Not very useful! It's important that we get more specific when we tell R to plot something. R is a wonderful tool to master, but it doesn't come with the intuitive training wheels that some other programs have. If we just say "plot this dataset", it's not going to give us some hodge podge of nice plots for each reasonable relationship in the dataset. We have to tell it exactly what we want it to do.

Now, I don't know much about flowers, but I would assume that petal length and width are correlated.

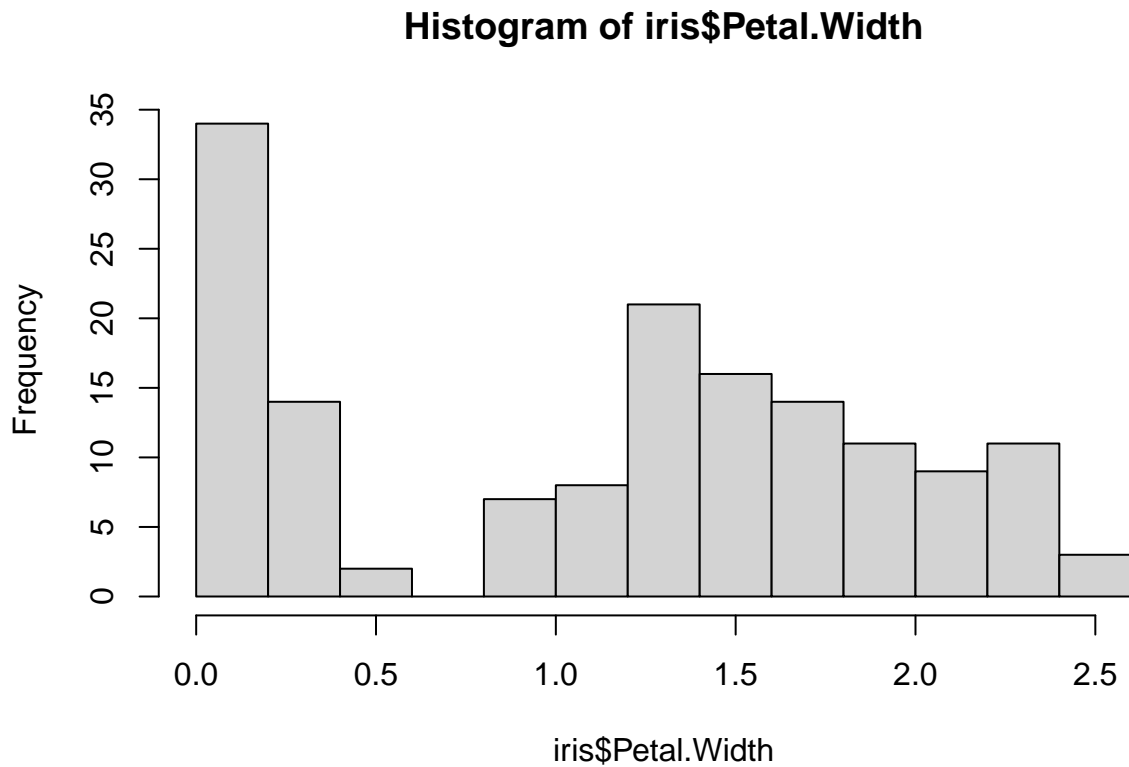
```
plot(iris$Petal.Width, iris$Petal.Length)
```



We can clearly see a positive correlation between the two.

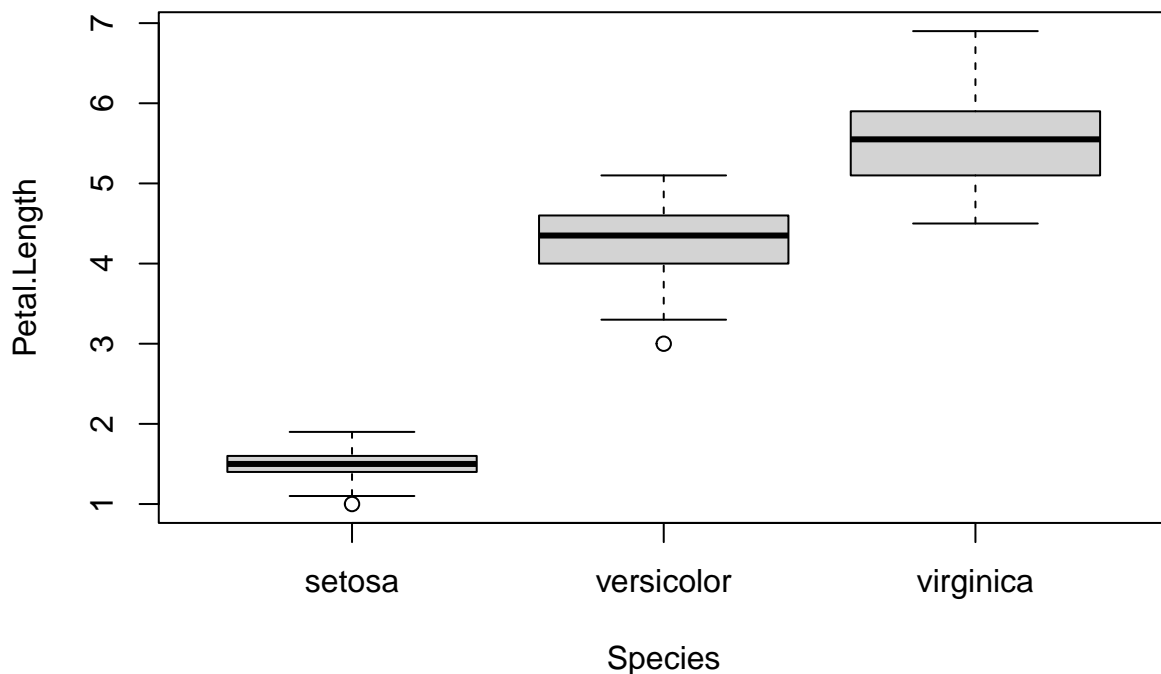
Let's see how our variables are distributed. You may note that when you knit a chunk to create multiple figures, it will print them out individually with each line above the exact figure it creates.

```
hist(iris$Petal.Width)
```

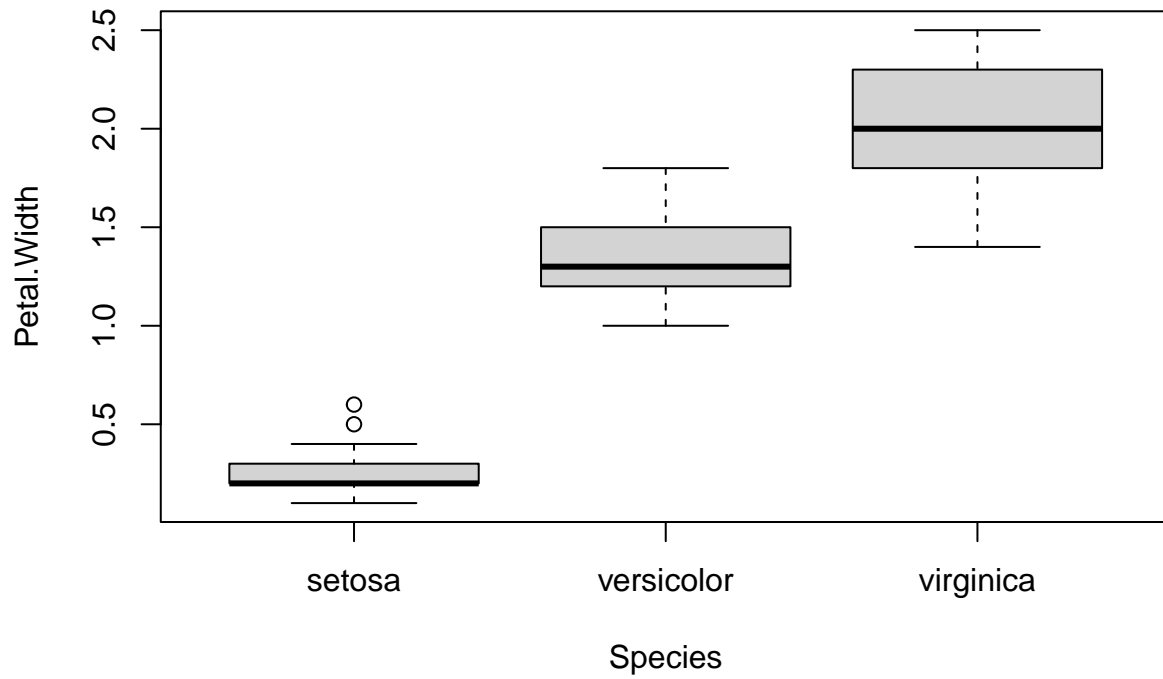


The species of Iris are categories, to compare categorical variables to something numeric, we use boxplot. We will tell boxplot to show us petal and sepal length and width as a function of the species.

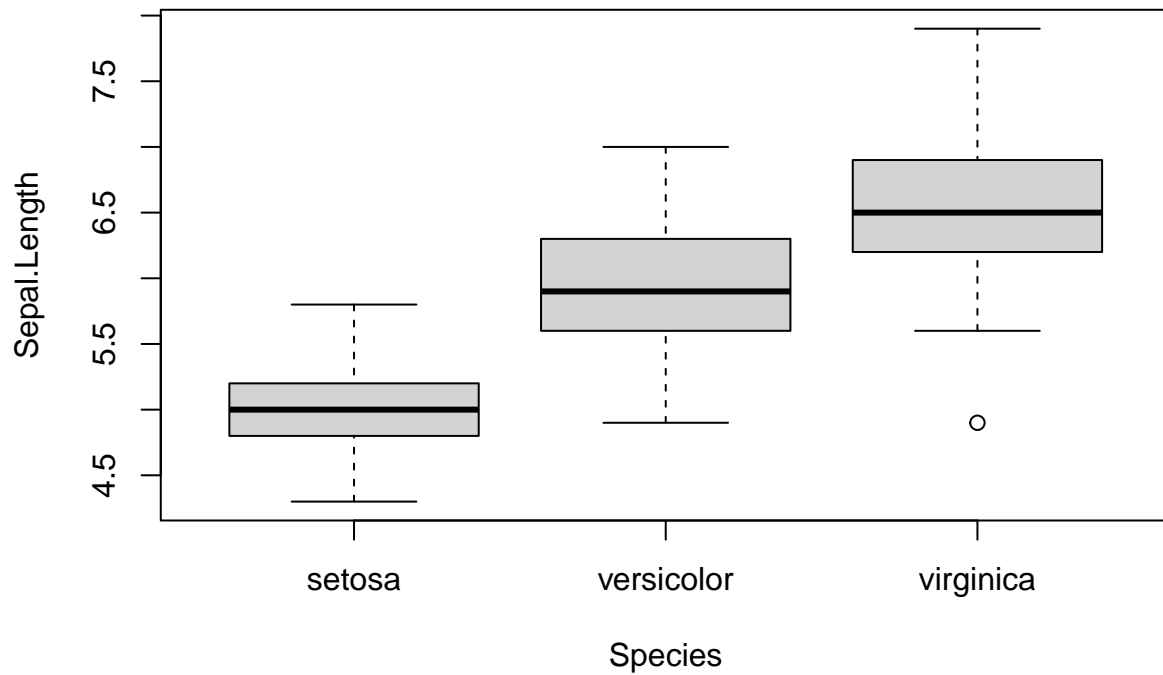
```
boxplot(Petal.Length ~ Species, data=iris)
```



```
boxplot(Petal.Width ~ Species, data=iris)
```



```
boxplot(Sepal.Length ~ Species, data=iris)
```



```
boxplot(Sepal.Width ~ Species, data=iris)
```



Our data analyses must be informed by the actual context of the “data generating process”.

Our data tells us about the various dimensions of various types of irises, but how is that data generated? It's generated by a process of people actually measuring real irises. Let's see what an iris looks like.



iris setosa



petal sepal

iris versicolor



petal sepal

iris virginica



petal sepal

BY PRAMOD SAHU

Is our data analysis congruent with what we see here? What are some observations we can make about irises from the picture? Do these align with what we found?

Exercise.

In the empty chunk below, I want you all to come up with **four unique lines of code** that involve checking conditions or logic or whatever you want regarding the irises. One example is something like `iris$Sepal.Width > iris$Petal.Width`". What do we expect to get from this command and why? Make sure the dimensions of the irises align with what you expect from the pictures (they generally should).

```
#fill this in with stuff
```