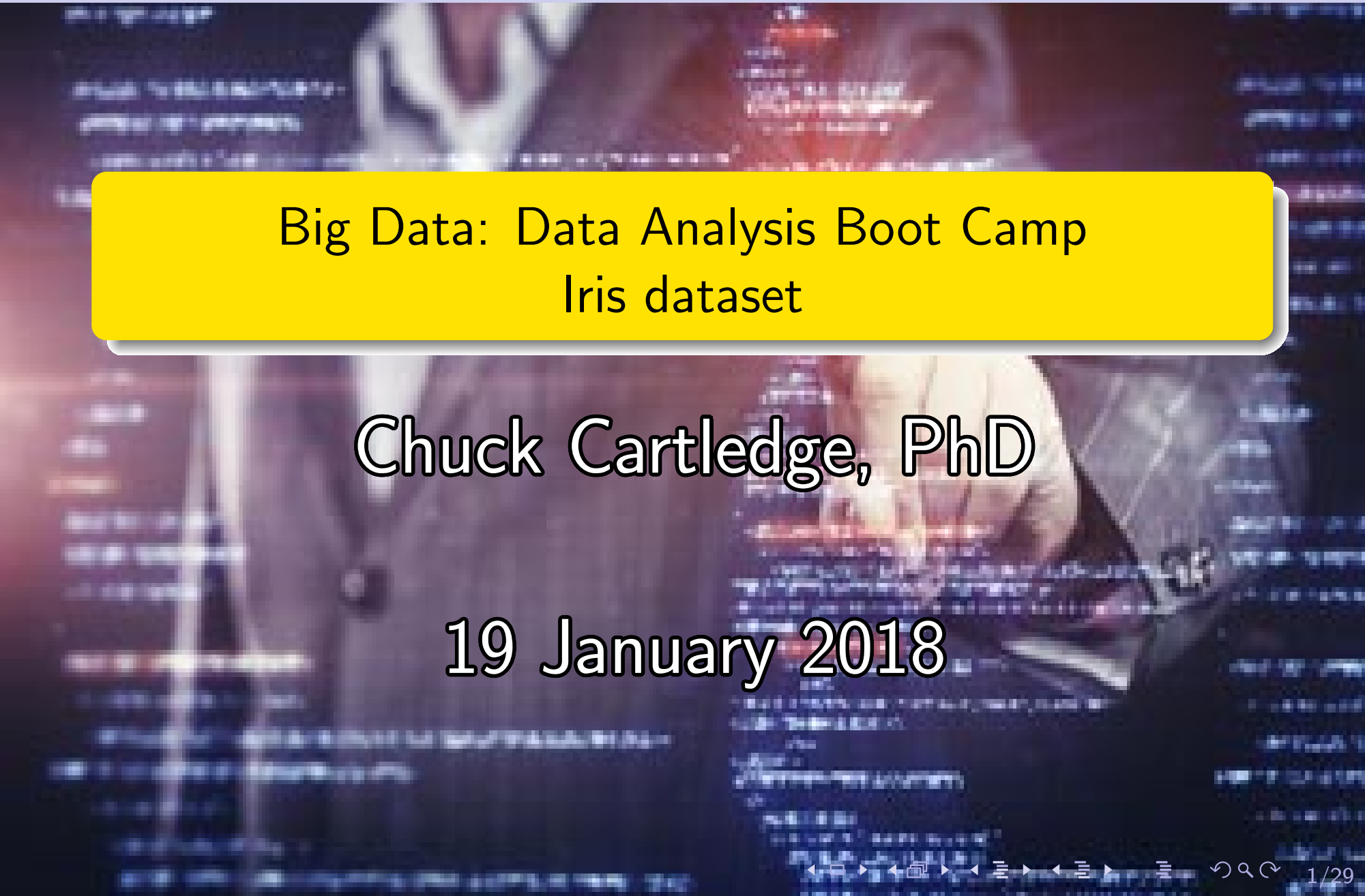


Big Data: Data Analysis Boot Camp
Iris dataset

Chuck Cartledge, PhD

19 January 2018

Table of contents (1 of 1)

1. Intro.
2. Built-in datasets
3. Iris dataset
   - Background
   - Iris dataset analysis
   - What can we learn from it?
4. Hands-on
5. Q & A
6. Conclusion
7. References
8. Files

What are we going to cover?

We're going to talk about:

- A few of the multitudes of R's built-in datasets.
- An overview of tools and techniques to look at the iris dataset.

R has over 120 built-in datasets

To see the currently installed ones:

```
data()
```

To see the `data()` function code:

```
data
```

To see over 2,000 available datasets:

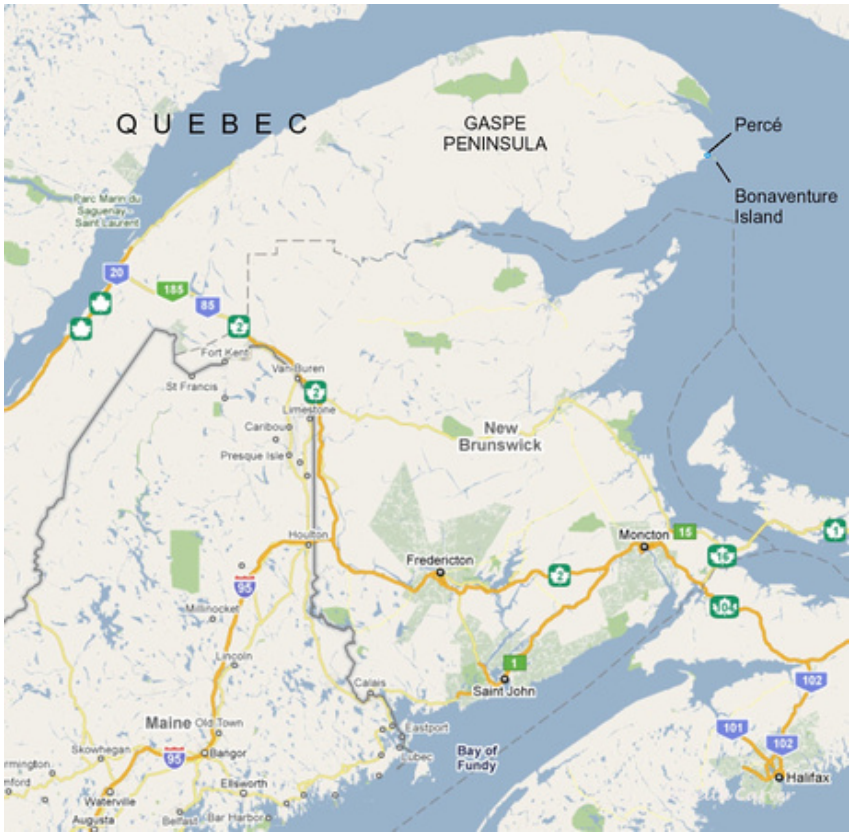
```
data(package = .packages(all.available = TRUE))
```

To get detailed information about the iris dataset:

```
?iris
```

Some background

- Edgar Anderson collected data on 3 different iris species on the Gaspé Peninsula, Quebec, Canada[1]
- Ronald Fisher used Anderson's data to see if linear regression could be used to could be used to "maximize the ratio of the difference between the specific means to the standard deviations within species." [3]



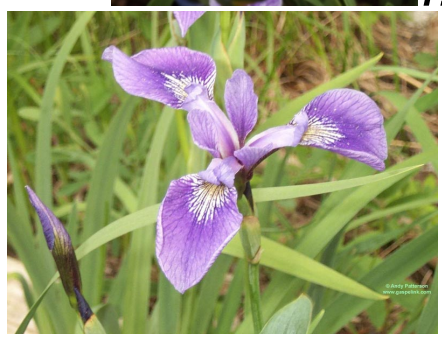
Anderson classified 50 examples of 3 different species

Each specimen was:

- Collected on the same day
- Collected by the same person
- Measured using the same instruments

I. virginica

I. setosa



I. versicolor

Data collected from each specimen:

- Sepal length,
- Sepal width,
- Petal length, and
- Petal width

Sepals enclose the flower bud. They fold over and protect the closed bud from weather or injuries while developing. Petals attract hummingbirds and insects so that pollination may occur. Petals also protect the stamen and pistil, the parts of the plants needed for reproduction.[6]

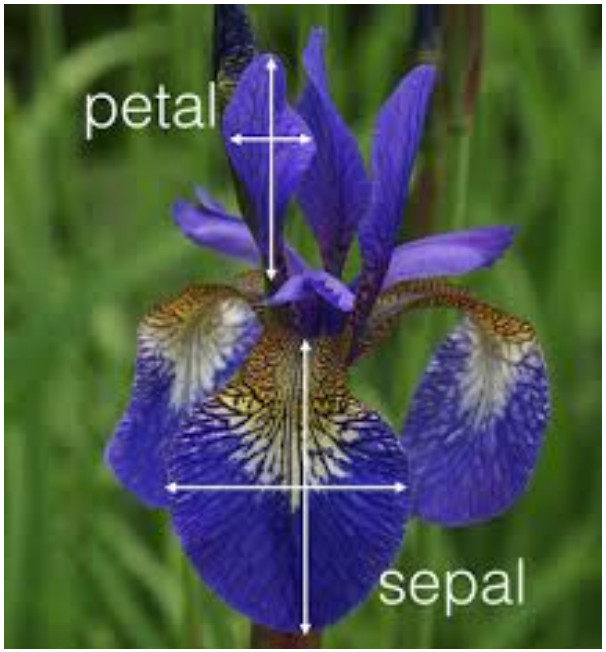


Image from [2].

What does his data look like?

To see a few rows:

```
1 head(iris)
```

To see how many rows:

```
1 nrow(iris)
```

To see simple summary information:

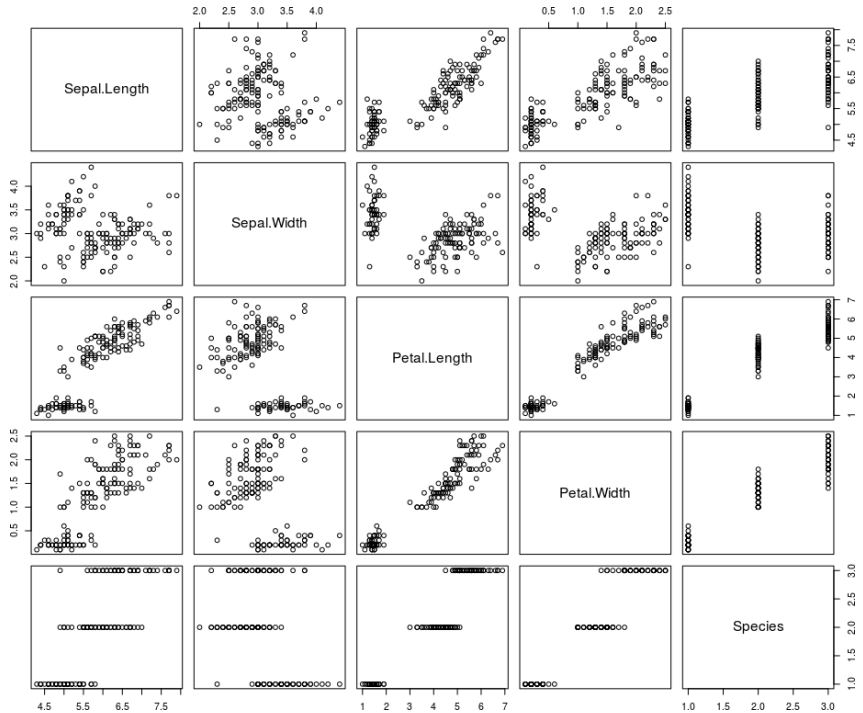
```
1 str(iris)
```

which returns:

```
'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 ...
```

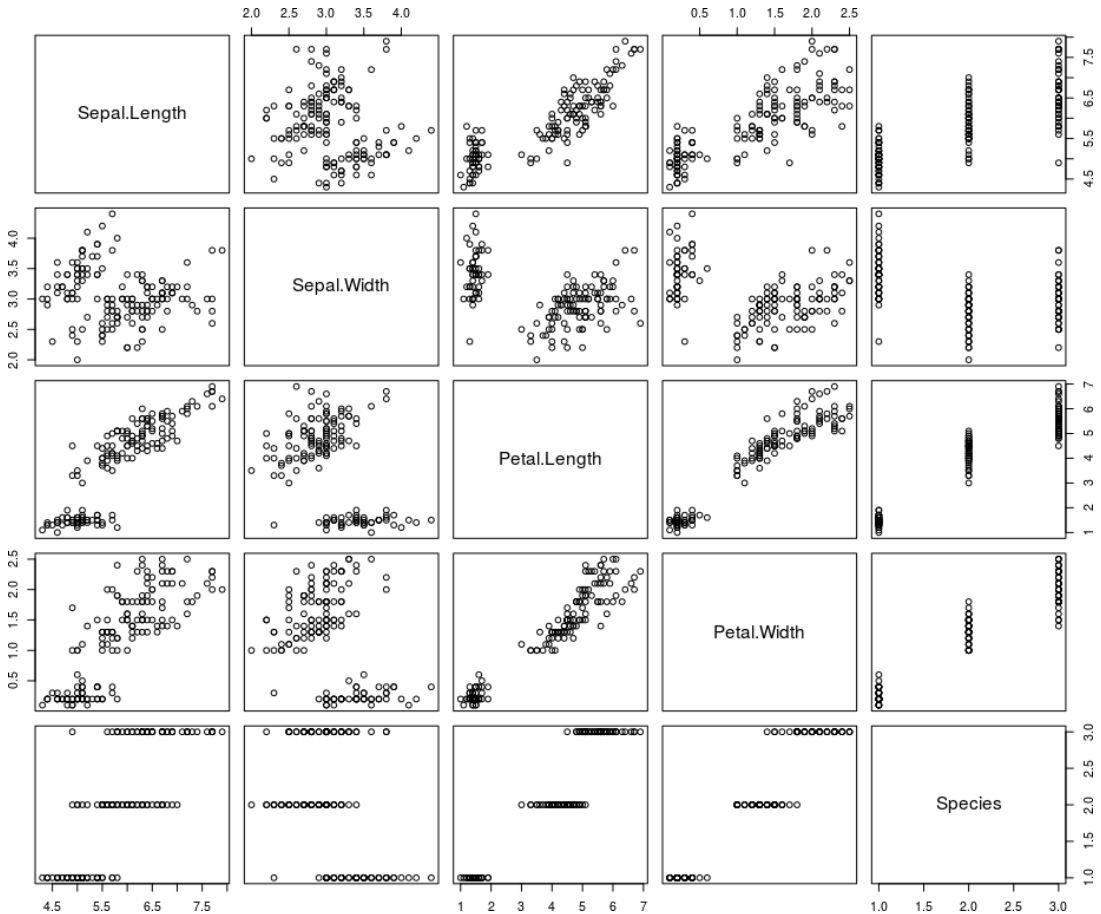
Other ways

Quick and dirty:
`plot(iris)`



Pairwise plotting of all numerical columns. Missing species (factor) classification.

Same image.

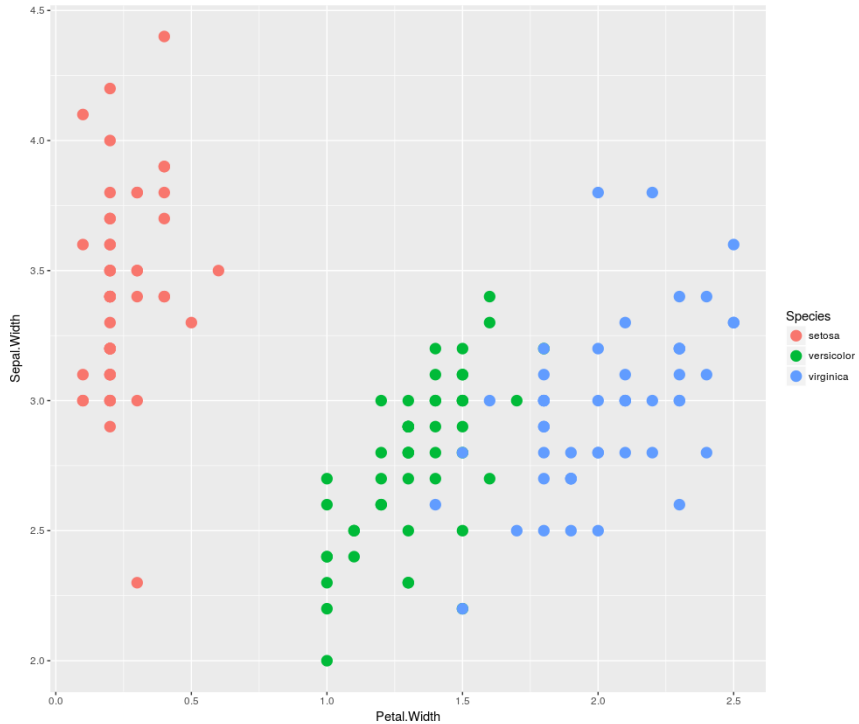


Pairwise plotting of all numerical columns. Missing species (factor) classification.

As a 3D-ish plot

Another view:

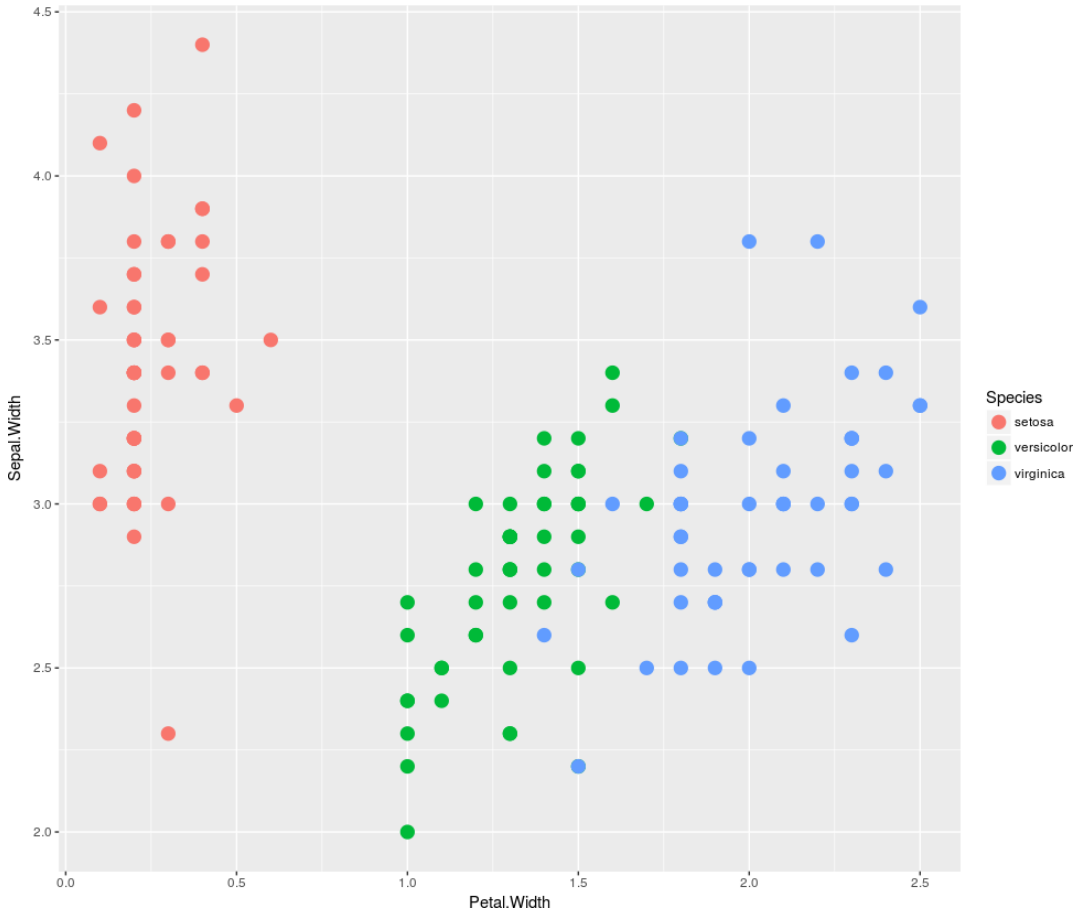
```
library(ggplot2)
qplot(Petal.Width,
Sepal.Width, data=iris,
colour=Species, size=I(4))
```



Ideas taken from [8].

Iris sepal and petal widths, showing species classification. Errors?

Same image.



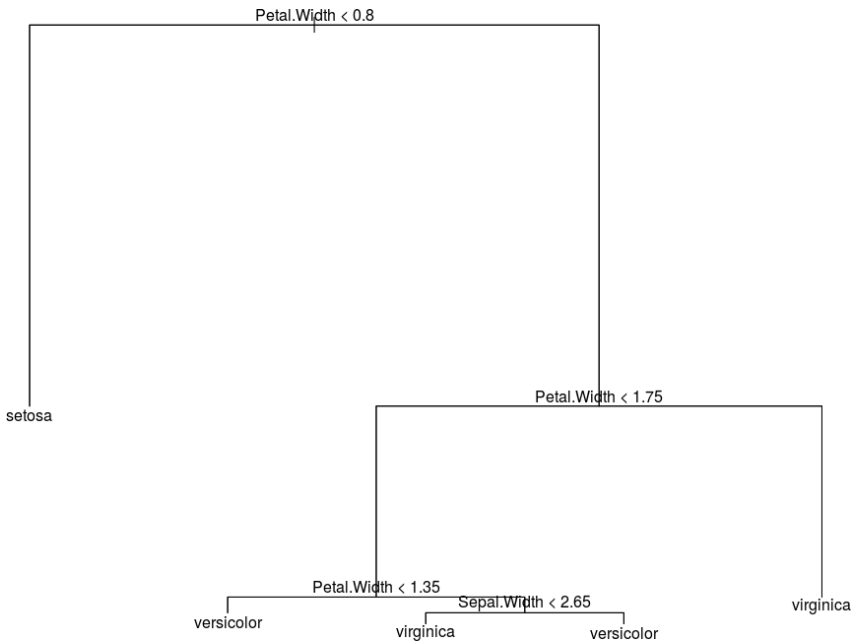
Ideas taken from [8].

Iris sepal and petal widths, showing species classification. Errors?

As a decision tree

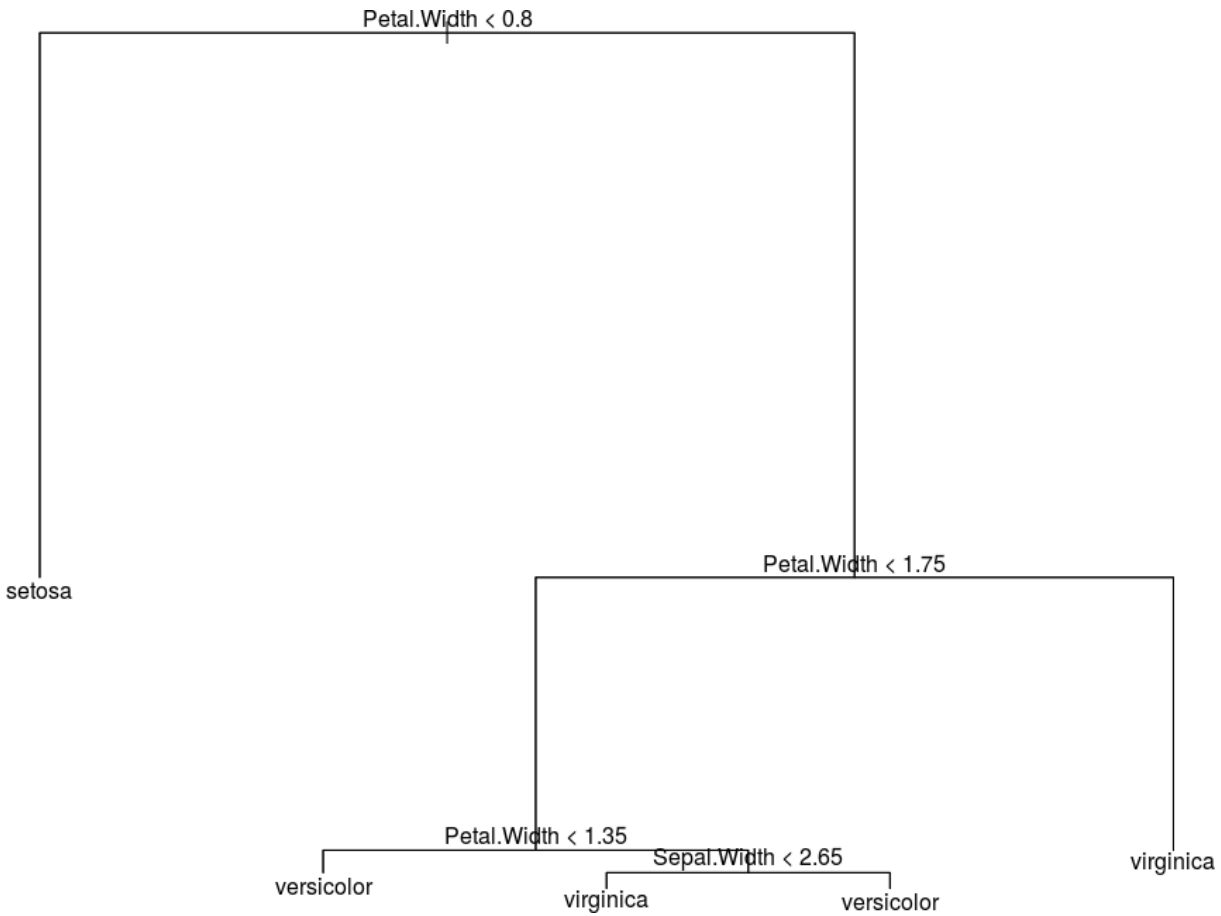
More informative:

```
library(tree)
tree1 <- tree(Species
~Sepal.Width +
Petal.Width, data = iris)
plot(tree1)
text(tree1)
```



An iris species classification decision tree.

Same image.



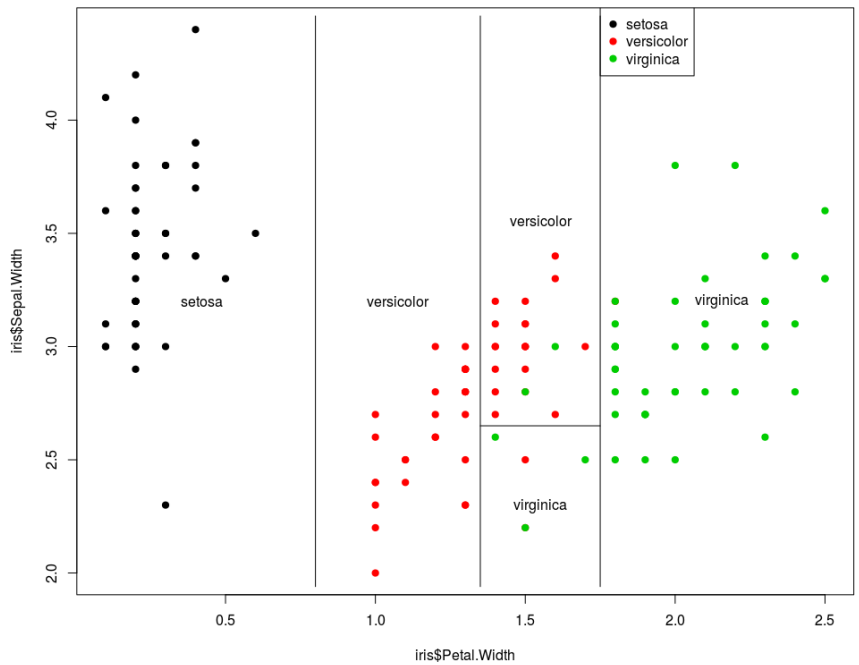
An iris species classification decision tree.

Combining decision tree and 3D-ish plot

A slightly more complex example:

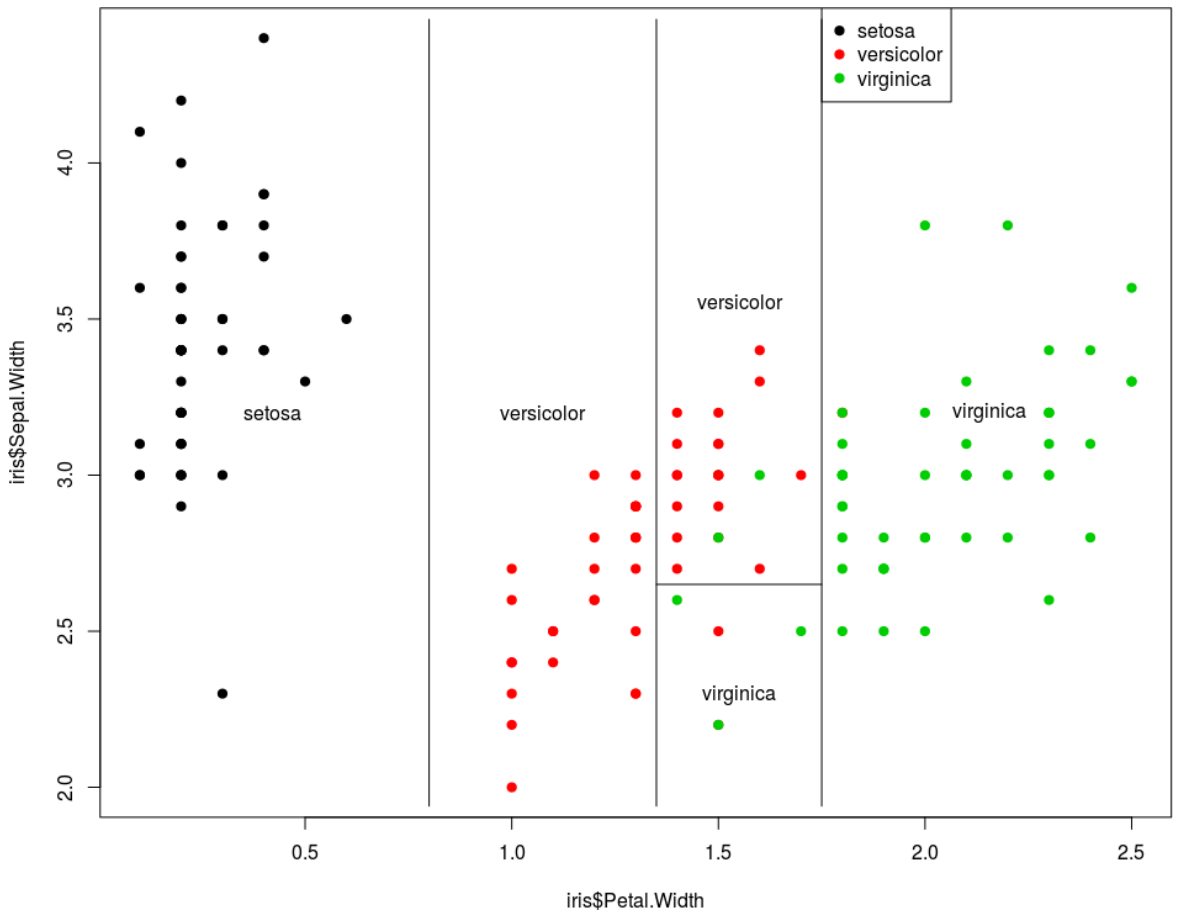
```
library(tree)
tree1 <- tree(Species ~Sepal.Width +
Petal.Width, data = iris)
plot(iris$Petal.Width, iris$Sepal.Width, pch=19,
col=as.numeric(iris$Species))
partition.tree(tree1, label="Species", add=TRUE)
legend(1.75,4.5, legend=unique(iris$Species),
col=unique( as.numeric(iris$Species)), pch=19)
```

(Lines broken for readability.)



Why are there misclassifications?

Same image.



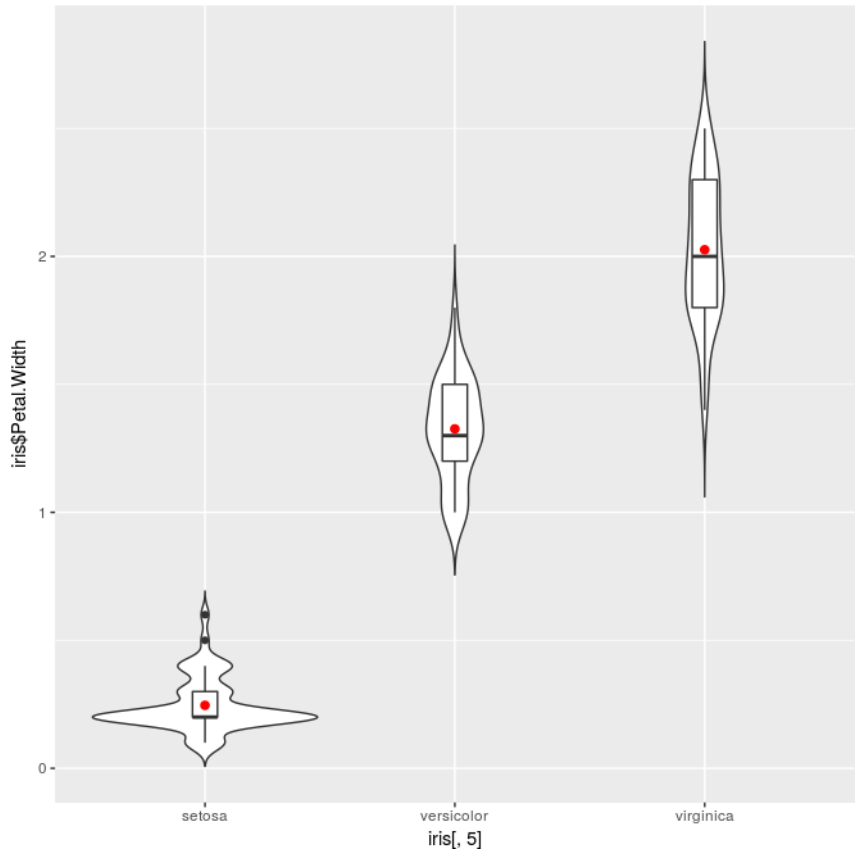
Why are there misclassifications?

How far should we go?

A even more slightly complex example:

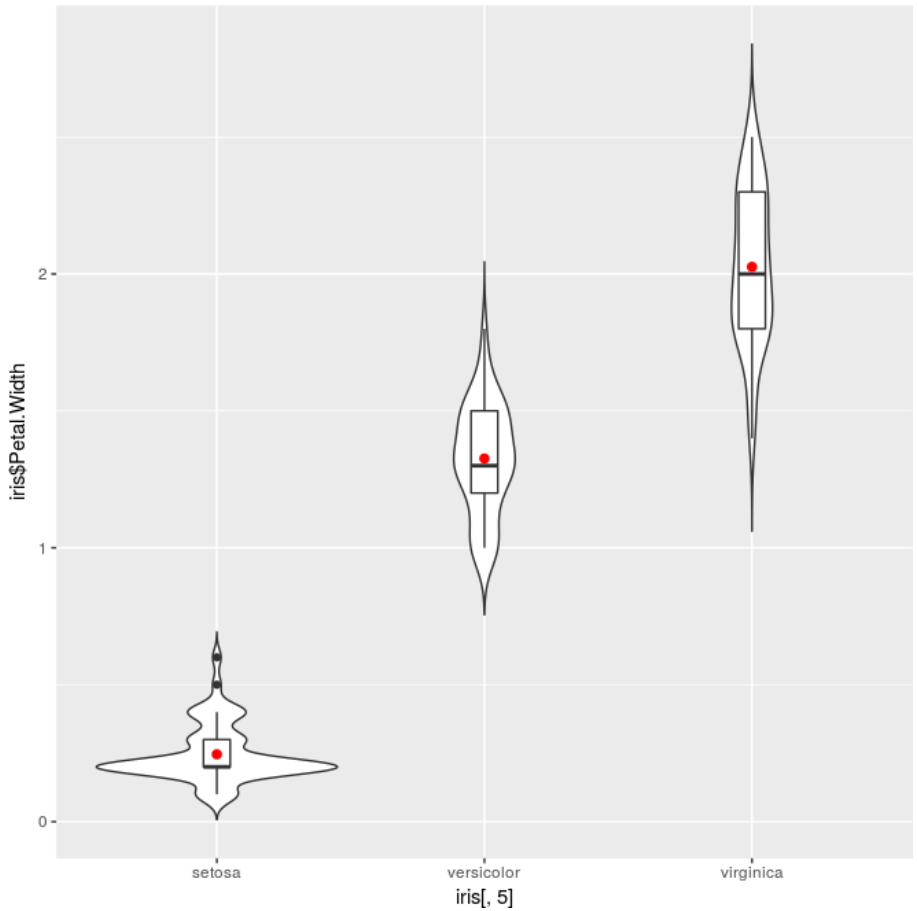
```
library(ggplot2)
p <- ggplot(iris[,-5],
aes(x=iris[,5],
y=iris$Petal.Width)) +
geom_violin(trim=FALSE)
p + geom_boxplot(width=0.1)
+ stat_summary(fun.y=mean,
geom="point", size=2,
color="red")
```

(Lines broken for readability.)



Perhaps there are attributes that aren't being captured.

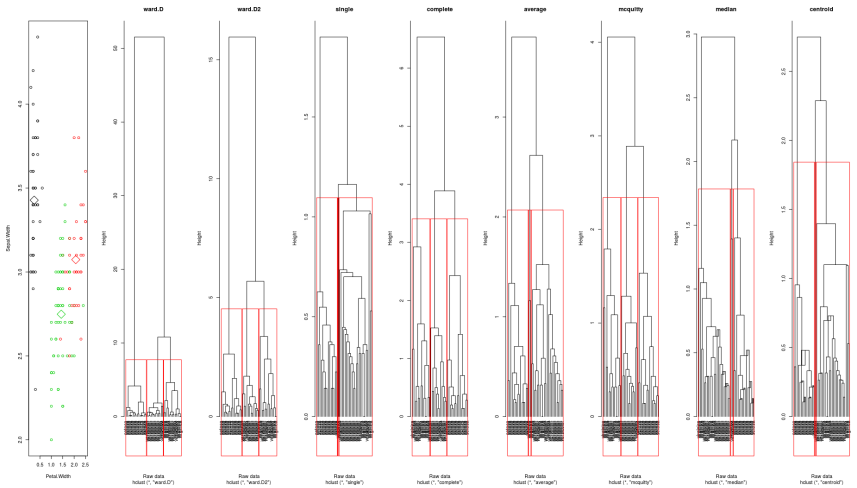
Same image.



Perhaps there are attributes that aren't being captured.

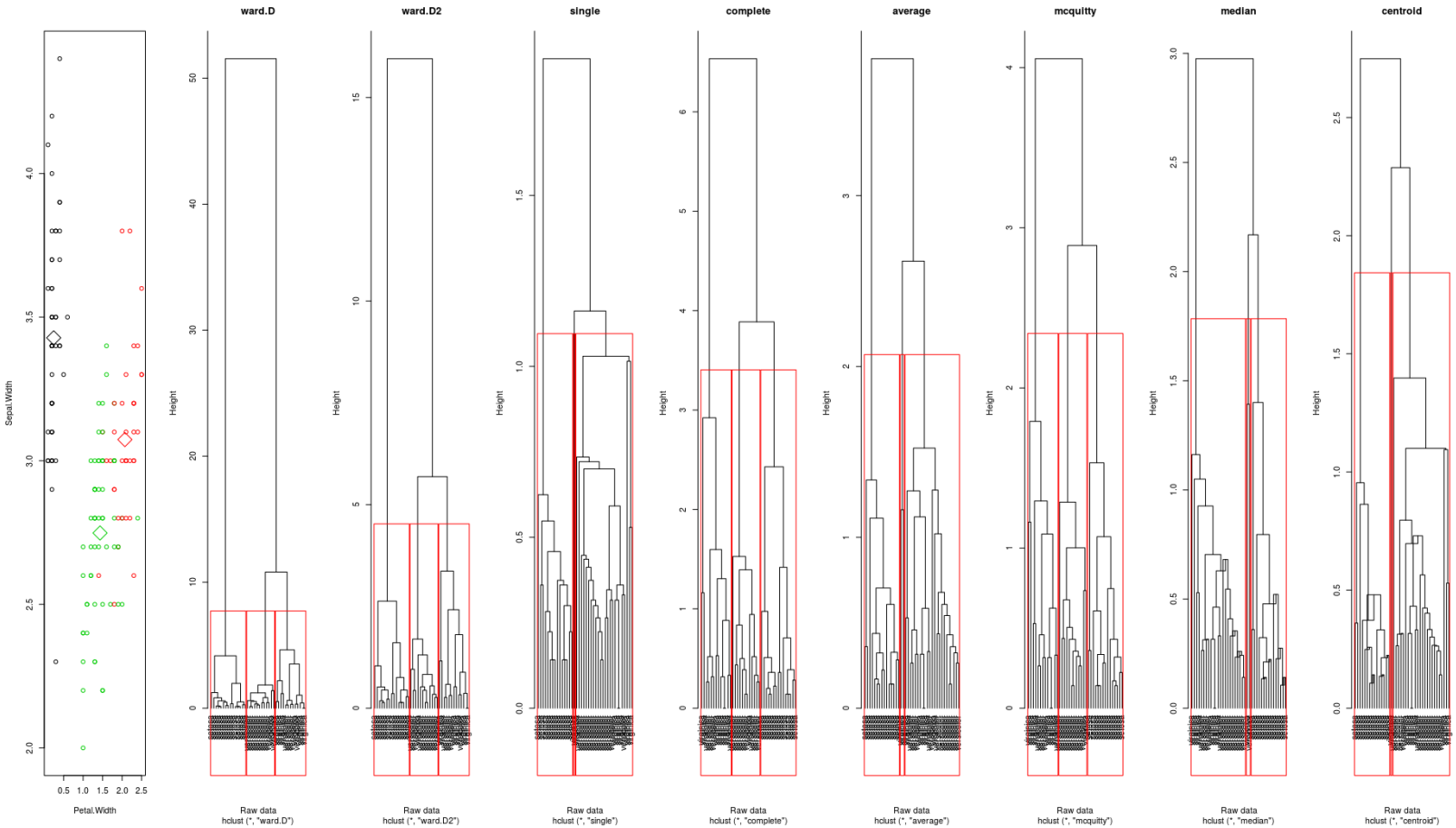
Different tools give different views

The default *kmeans* clustering is applied with different k values.
(A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering[5].)
R script is attached[7].



Different clusterings yield different dendograms.

Same image.

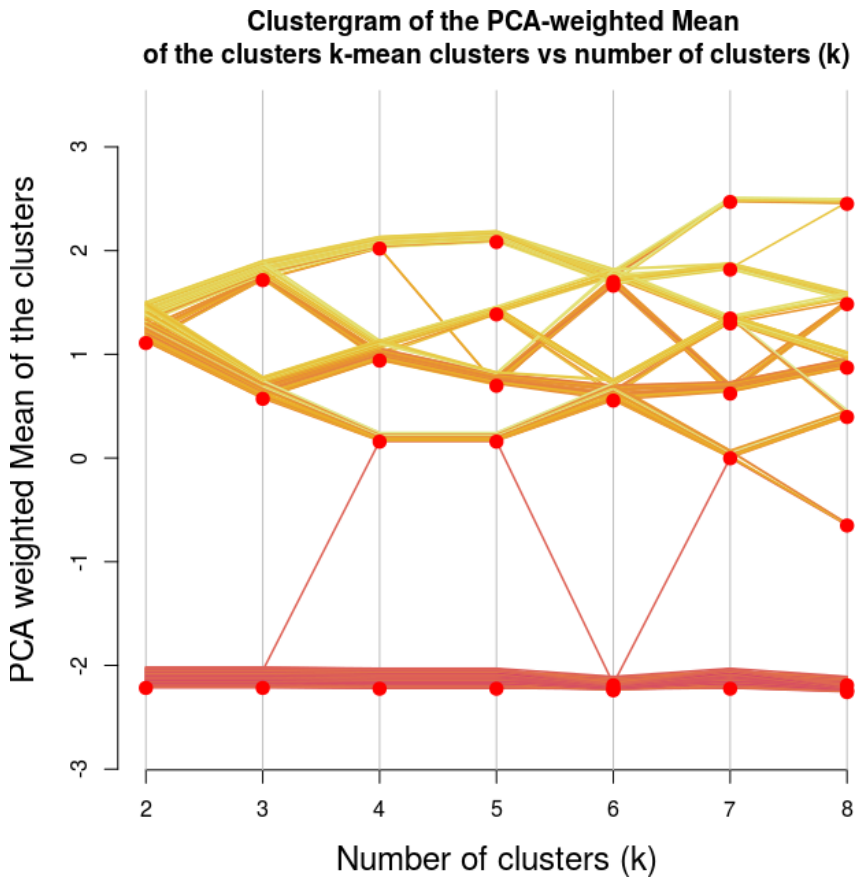


Different clusterings yield different dendograms.

Viewing the results of many clusters

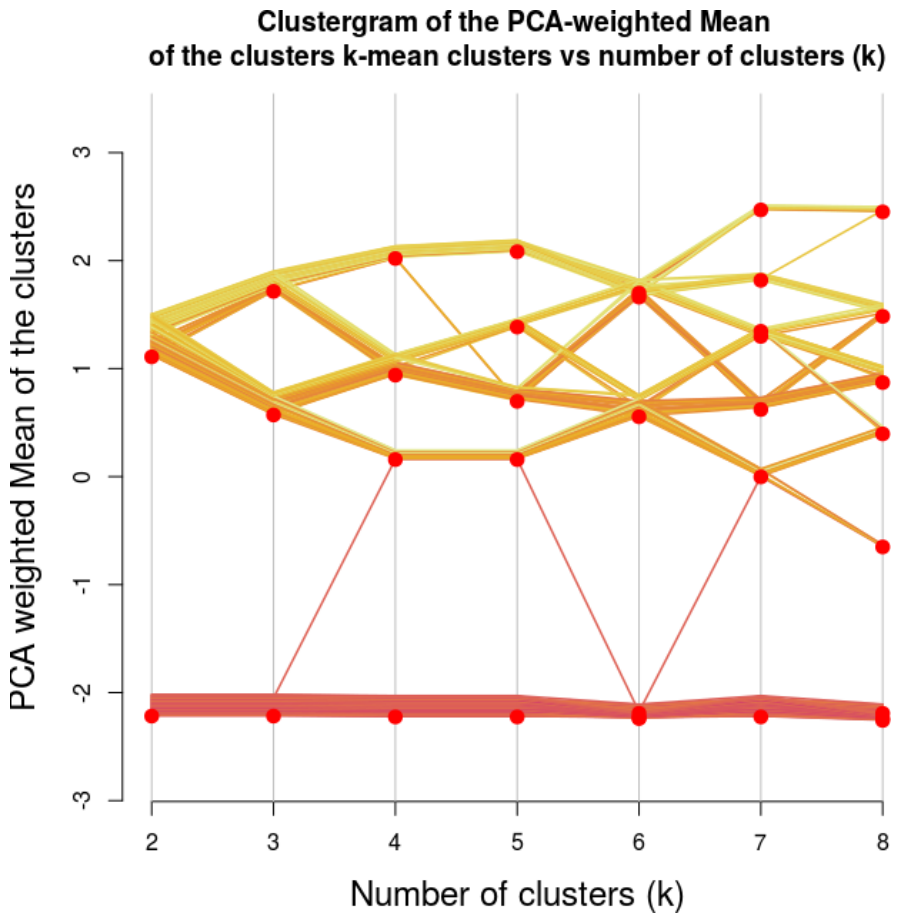
> *"Principal component analysis (PCA) refers to the process by which principal components are computed, . . . PCA is an unsupervised approach, since it involves only a set of features . . . , and no associated response Y."*
>
> *James, et al. [5]*

The default *kmeans* clustering is applied with different *k* values. PCA is plotted for each solution. R script is attached[4].



Some specimens move from one cluster to another.

Same image.



Some specimens move from one cluster to another.

Some simple exercises to get familiar with data analysis

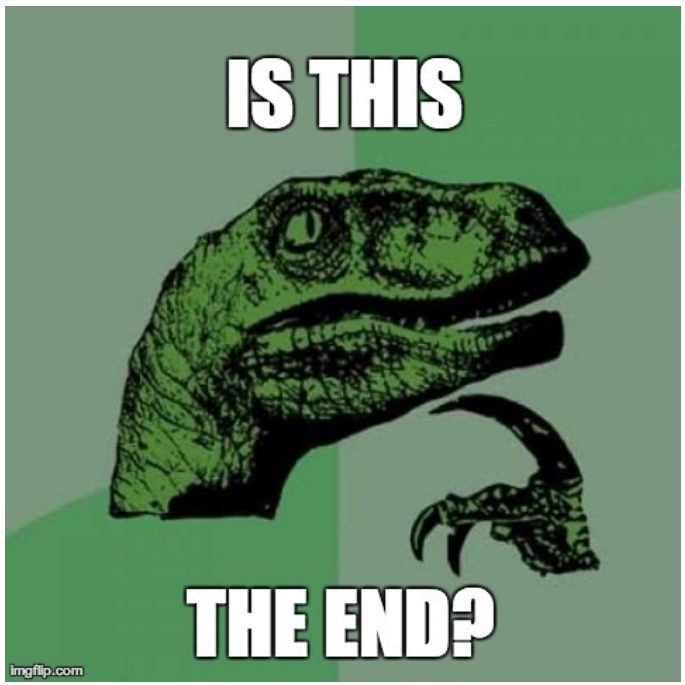
1. Build an iris classification tree using only sepal data
2. Build an iris classification tree using all data without specifying each element
3. Build a 3D-ish decision tree using only sepal data
4. Looking at the image on page 17, what can be said about using petal width as a decision attribute?

Q & A time.

Q: Why was Stonehenge abandoned?
A: It wasn't IBM compatible.

What have we covered?

- R has a multitude of built-in datasets
- About the iris dataset:
  1. It isn't too large (only 150 rows)
  2. It lends it self reasonably to linear regression[3]
  3. There appear to be some "errors" (as in misclassifications), so it isn't "pure" data
  4. Decision trees aren't too large, nor too complex
  5. It is an easy place to start
- Different tools and techniques give different insights into the dataset



Next: Look at R's built-in Titanic dataset

References (1 of 3)

[1] Edgar Anderson, The irises of the Gaspé Peninsula, Bulletin of the American Iris society **59** (1935), 2–5.

[2] Sarthak Dasadia, Machine Learning with Iris Dataset, `https://rstudio-pubs-static.s3.amazonaws.com/202738_7cad2477d76b4acc82b44244f94ccfa8.html#/`, 2016.

[3] Ronald A Fisher, The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics **7** (1936), no. 2, 179–188.

References (2 of 3)

[4] Tal Galili,
Clustergram: visualization and diagnostics for cluster analysis (R code
https://www.r-bloggers.com/clustergram-visualization-and-diagnostics-for-cluster-analysis-r-code/, 2010.

[5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning, vol. 6, Springer, 2013.

[6] Kimberly Napier,
What Is the Difference Between Sepals & Petals?,
https://www.hunker.com/13426267/what-is-the-difference-between-sepals-petals, 2017.

References (3 of 3)

[7] RDM Staff, k-means Clustering, http://www.rdatamining.com/examples/kmeans-clustering, 2017.

[8] Dave Tang, Building a classification tree in R, http://davetang.org/muse/2013/03/12/building-a-classification-tree-in-r, 2013.

Files of interest

1. iris dendogram R script
2. iris clustergram R script
3. R library script file