Can machine learning be used to identify species of Sorbus

PetraGuy

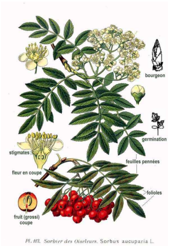
6 January 2018

Abstract.

Machine learning was used to separate six species of Sorbus within the subgenus Soraria based on morphological measurements of fruit and leaves. . . .

1. Introduction - The genus Sorbus.

Sorbus is a member of the Rosacea family, perhaps the best known species being Sorbus aucuparia, the Rowan or Mountain Ash.



However, there are over 50 species of Sorbus in the UK, 38 of these are vulnerable or critically endangered and most are endemic or native. There are four diploid species, but, as with many Rosaceae, Sorbus produce new apomictic polyploid species. These can also produce viable pollen and can therefore backcross with other diploid species. This results in the large number of genetically unique, stable, clonal communities, which can look very similar to each other. This presents a problem with recording and many Sorbus require expert knowledge to correctly identify to species level because much of the identification depends on comparitive knowledge. This tends to disuade recorders, or encourages records at aggregate level.

Sorbus are grouped into six subgenus, each of which are reasonably easy to identify by recorders with some knowledge, as the illustrations below show.

More difficulty arises when identifying plants within these subgeni, and this is where this work has concentrated. In this modelling only the subgenus Soraria has been trialled. This consists of seven species all similar in appearance to Sorbus intermedia, although only six species are considered based on the availability of data. These plants are distinguished from other subgeni by having leaves with rounded lobes which are tomentose beneath and the fruits having fewer lenticles. Perhaps the most noticeable difference between plants within the subgenus, are the larger fruits on S intermedia, the smaller leaves of S minima and the small fruits of S mougeotii.

2. Methods 2.1 Data.

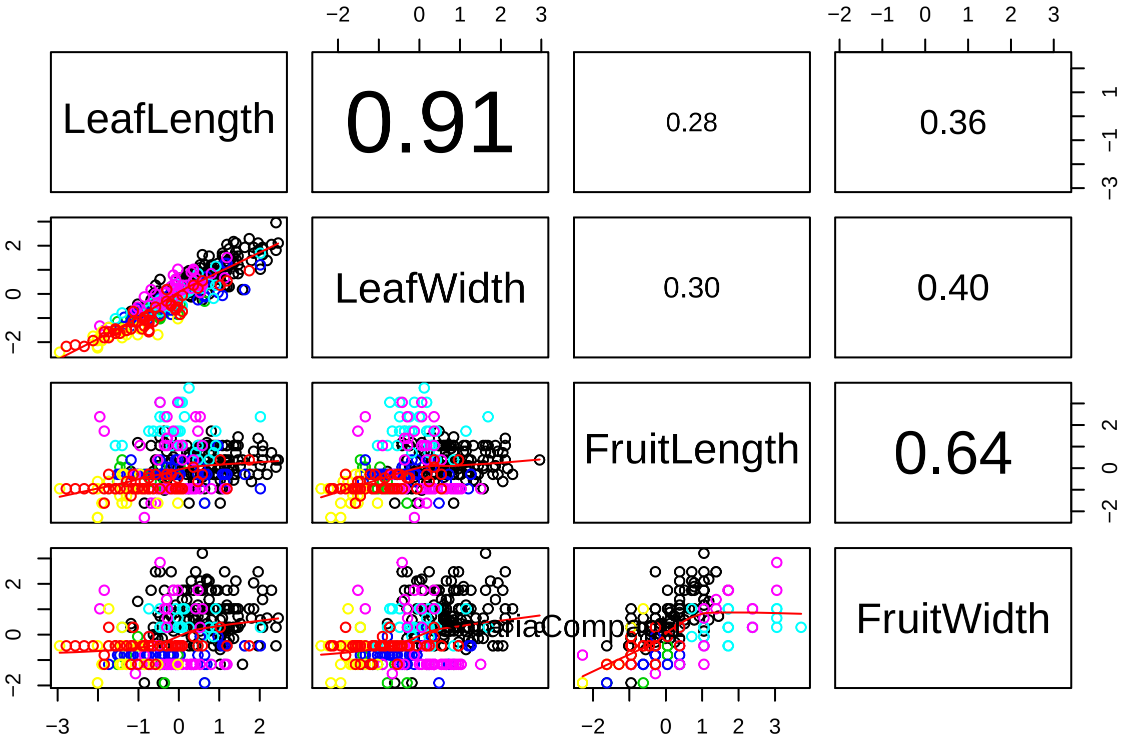
The data was provided by Dr T Rich, the British Botanical Society expert on Sorbus and consists of leaf and fruit measurements. For the leaves, the length, width, base angle, number of veins, depth of the lobes, vein angle and base angle have been recorded. For the fruit, the length and the width are used. Due to the variability in leaf size across one plant, the measurements were all carried out in a specific manner described by Rich [ ]. Essentially, repeated measurements of leaves on sterile spurs on the sunlight side of the tree are recorded and averaged over at least ten leaves.

The nature of collection means that the data was sparse. Every plant of every species did not have have complete set of measurements or the same number of measurements. For example, S intermedia had 126 observations but S leyana only had 39. This is due to the rarity S. leyana. S. intermedia is a common plant found throught the UK in easily accessible places, whilst S leyana is only found in two sites in South Wales, sometimes on the sides of cliffs. In addition, measurements cannot all be collected at the same time. Leaves must be measured when mature, around flowering time, and therefore cannot be measured in conjunction with fruit. Separate trips to re-measure fruit on the same trees may not be possible. This has lead to a sparse dataset in which not all morphological characteristics were availabe for every plant. S intermedia records are an example. Of 122 records, 72 are purely for fruit measurements and the remaining 50 purely for leaf measurements, and these occur on different plants If imputation was carried out, 59% of the leaf measurements would be imputed. This would reduce the effectiveness of some algorithsms. For example, in kNN, if you increase the frequency of the neighbours in the S. Intermedia group, it is more likely that a member of a different group will be close to that neighbour. Therefore, the sparsity was handled by reallocating measurements. For example, the 50 leaf measurements for S. intermedia were assigned to 50 fruit measurements and the excess 22 were not used. In some cases, where there were only a few additional rows of incomplete data, median imputation was carried out.

Although it seems dubious to assign records from one plant to another, in this analysis this was felt to be acceptible for two reasons. Firstly, this an exploration of a new technique for biological recording. It is not currently being proposed as a complete and accurate method for species identifiction at this stage. Secondly, the clonal nature of these plants implies that we would expect a great deal of similarity within a species. The variation within the species is more likely to come from the variety of leaf sizes which can be found on one plant, and these are controlled for, although they cannot be elimnated, when the data is collected. (The range of laef sizes witin each plant was not available, so this comparison has not been tested).

3.Data exploration. In order for machne learning algorithms to work accurately the groups to be separated should be clearly defined with very little overlap. Also, the data contains many variables, some of which may be unnecessary to the analysis. The following plots lok at group separation and correlations

Pairs Plots, Not all pairs plotted; there are so many the plot becomes too difficult to read and therefore uninformative.



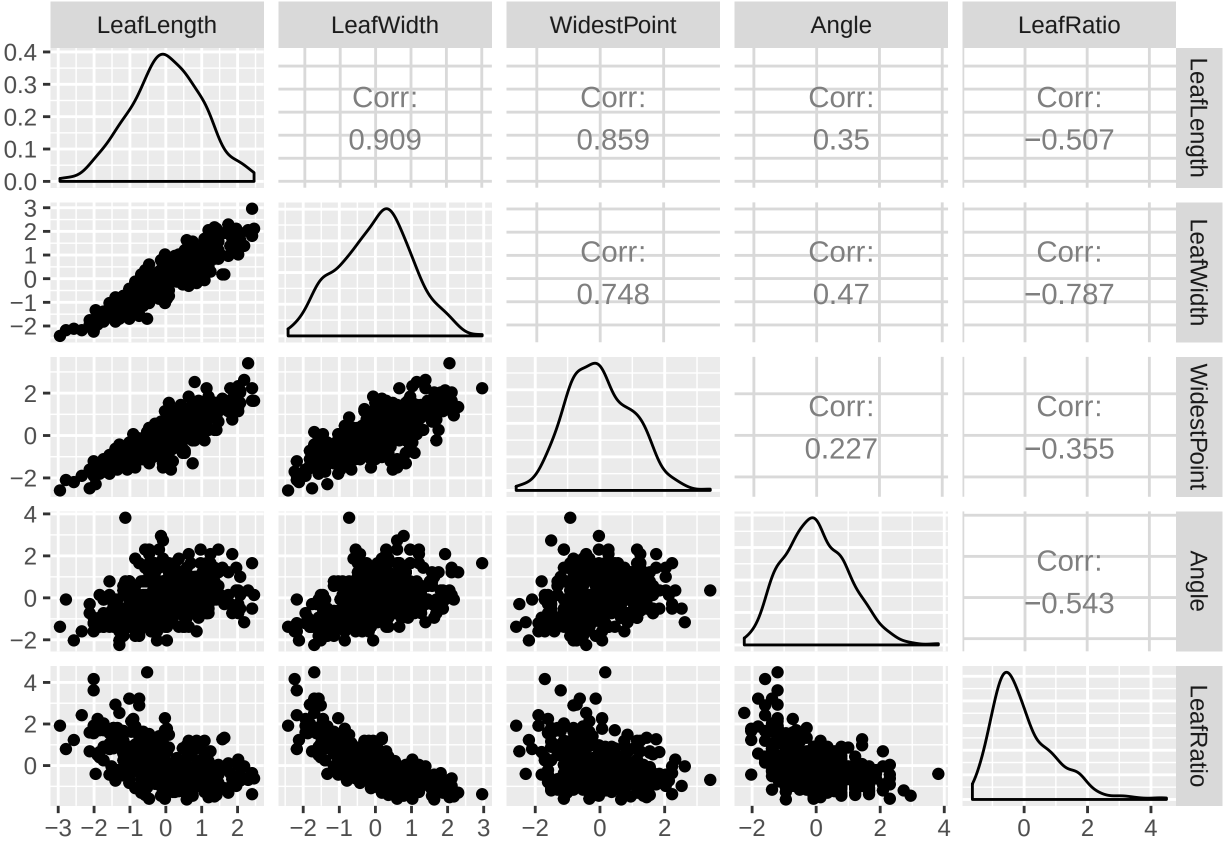
Repeat the pairs plots using ggplot

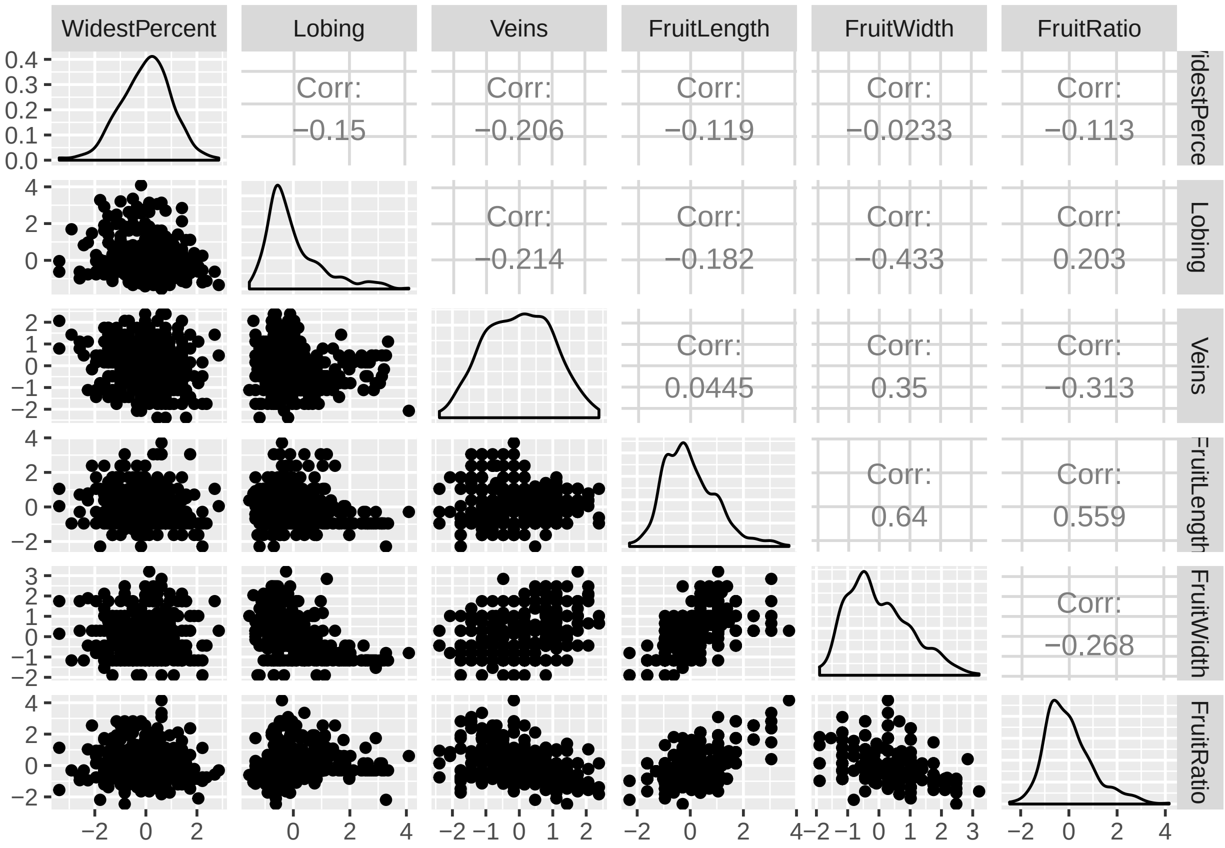
```
## Warning: replacing previous import by 'utils::capture.output' when loading
## 'GGally'
```

```
## Warning: replacing previous import by 'utils::head' when loading 'GGally'
```

```
## Warning: replacing previous import by 'utils::installed.packages' when
## loading 'GGally'
```

```
## Warning: replacing previous import by 'utils::str' when loading 'GGally'
```





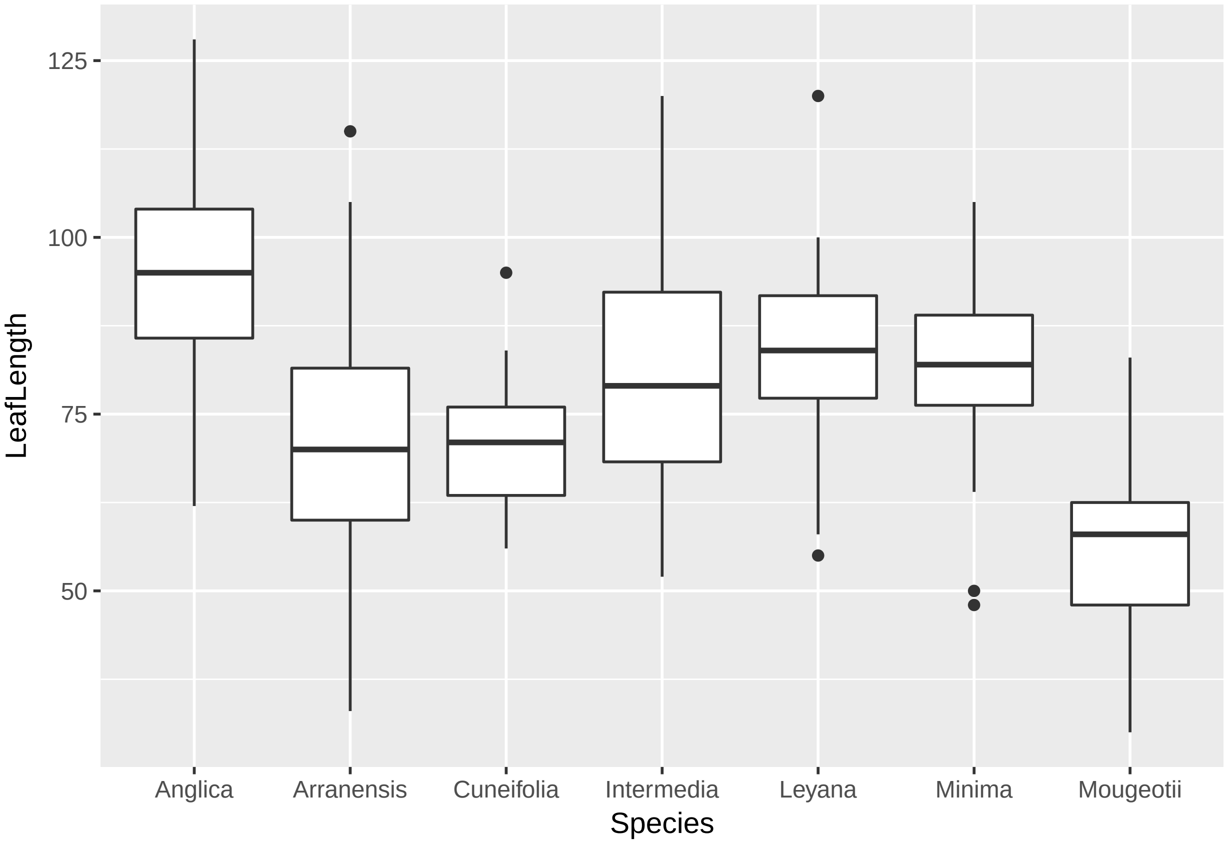
The pairs plot show a strong correlation between leaf width and length, leaf width andlength and the widest point, and fruit width and length.

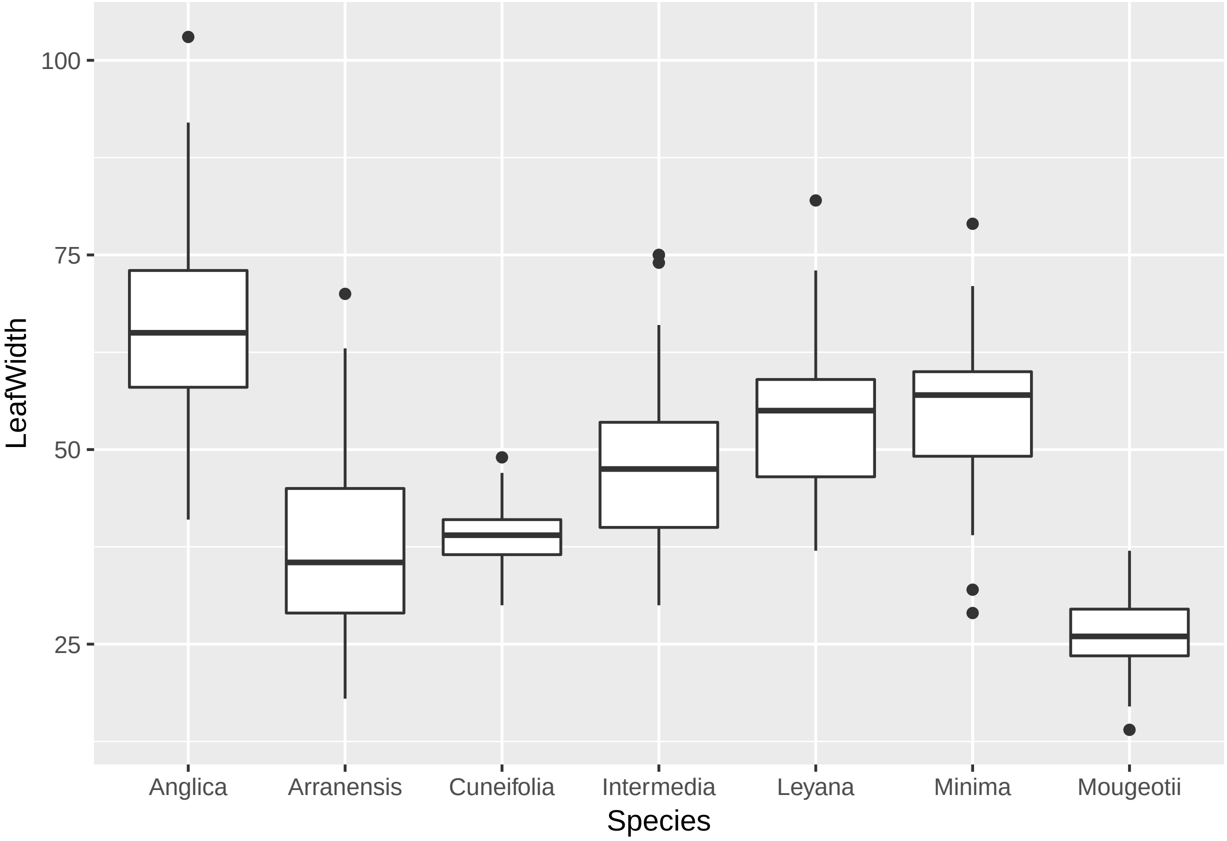
Look at scatter plots by species of main characters.

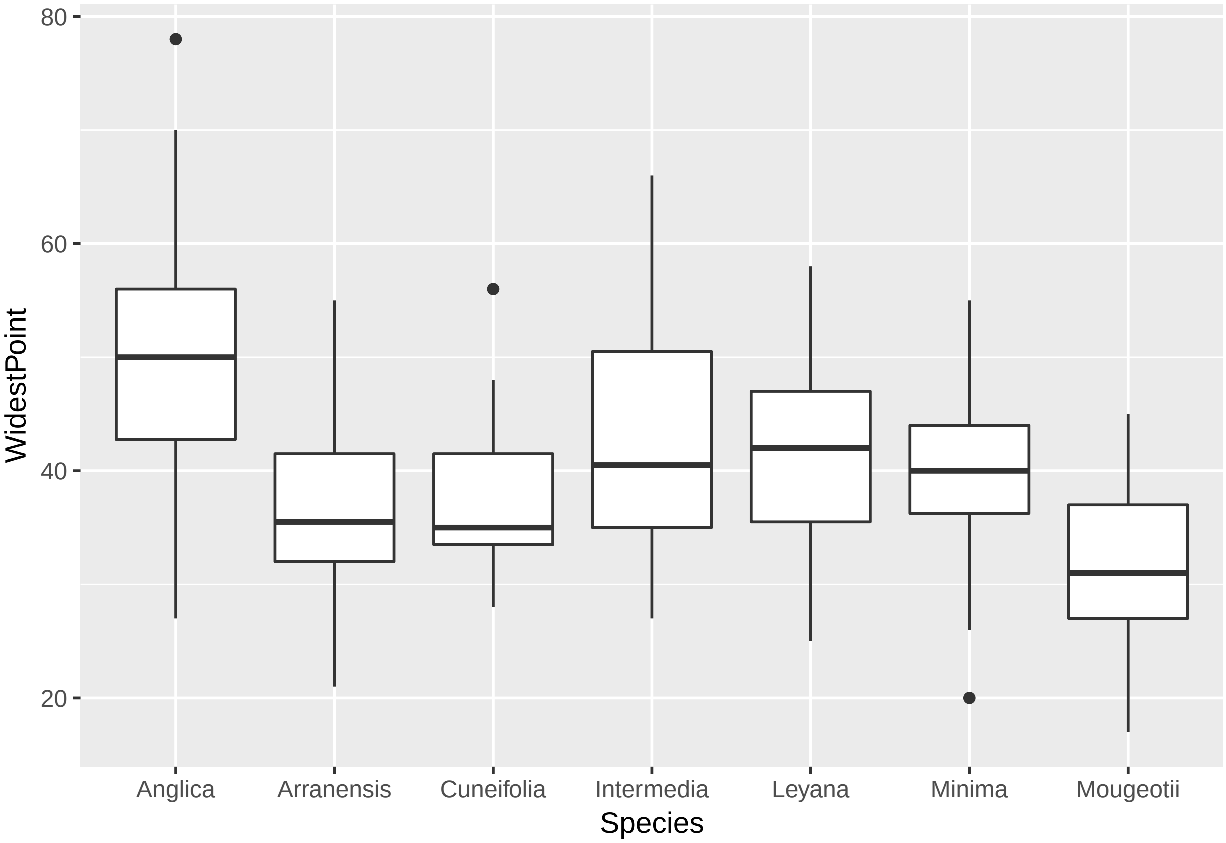
Scatter plots of the most correlated features

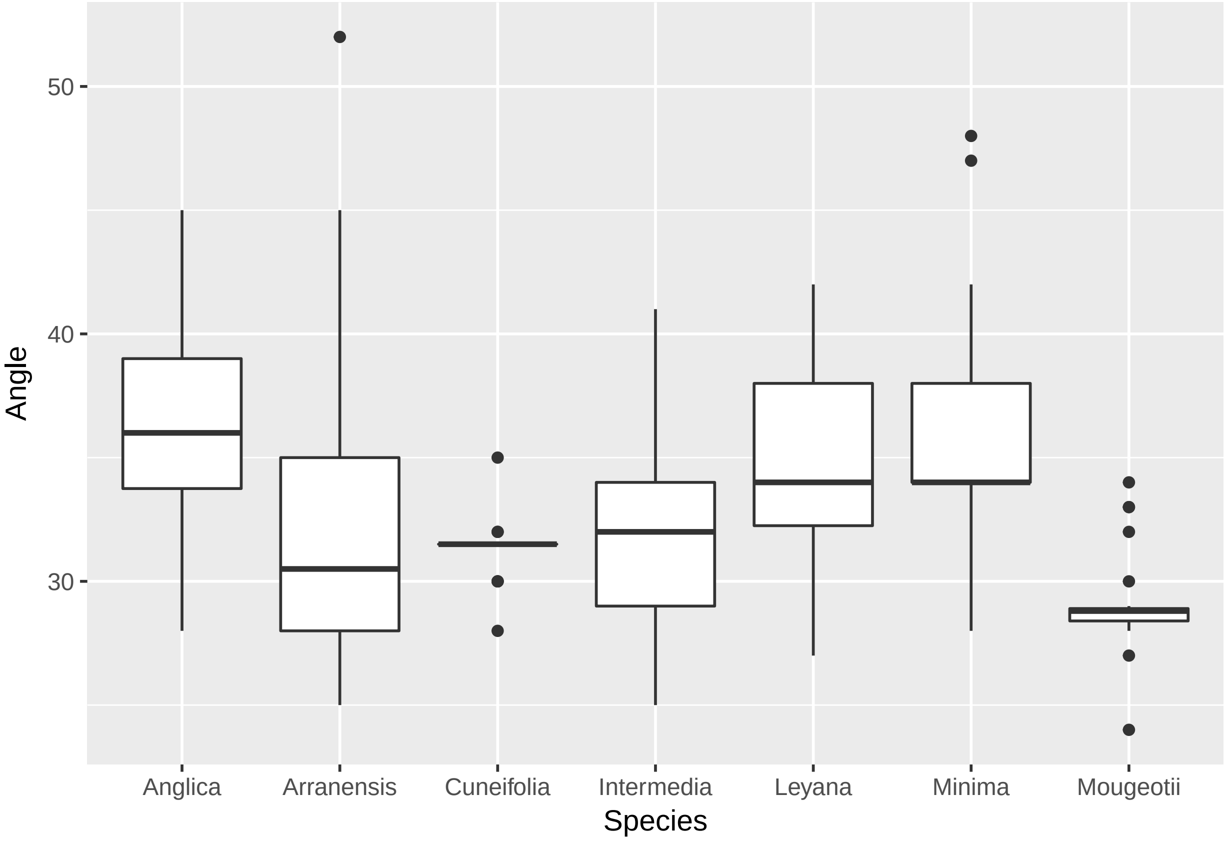
The scatter plots do not demostrate a clear separation of features between species, there is considerable overlap. S. anglica appears to have some separation with the longest and widest leaves, but the range of leaf measurements plotted shows that it overlaps witj most others, perhaps excepting S. mougeotii. S.Anglica stands out in the second plot as having fruits wider than long.

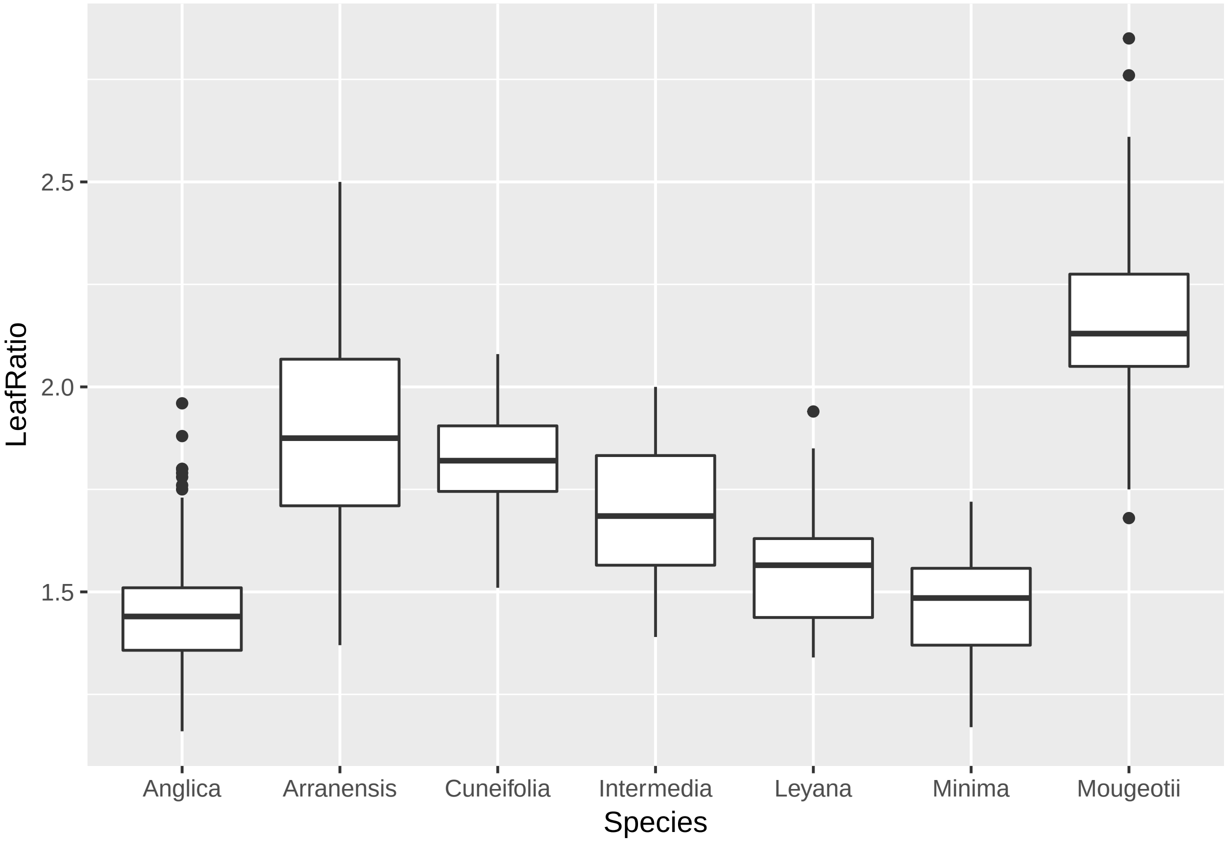
Box plots

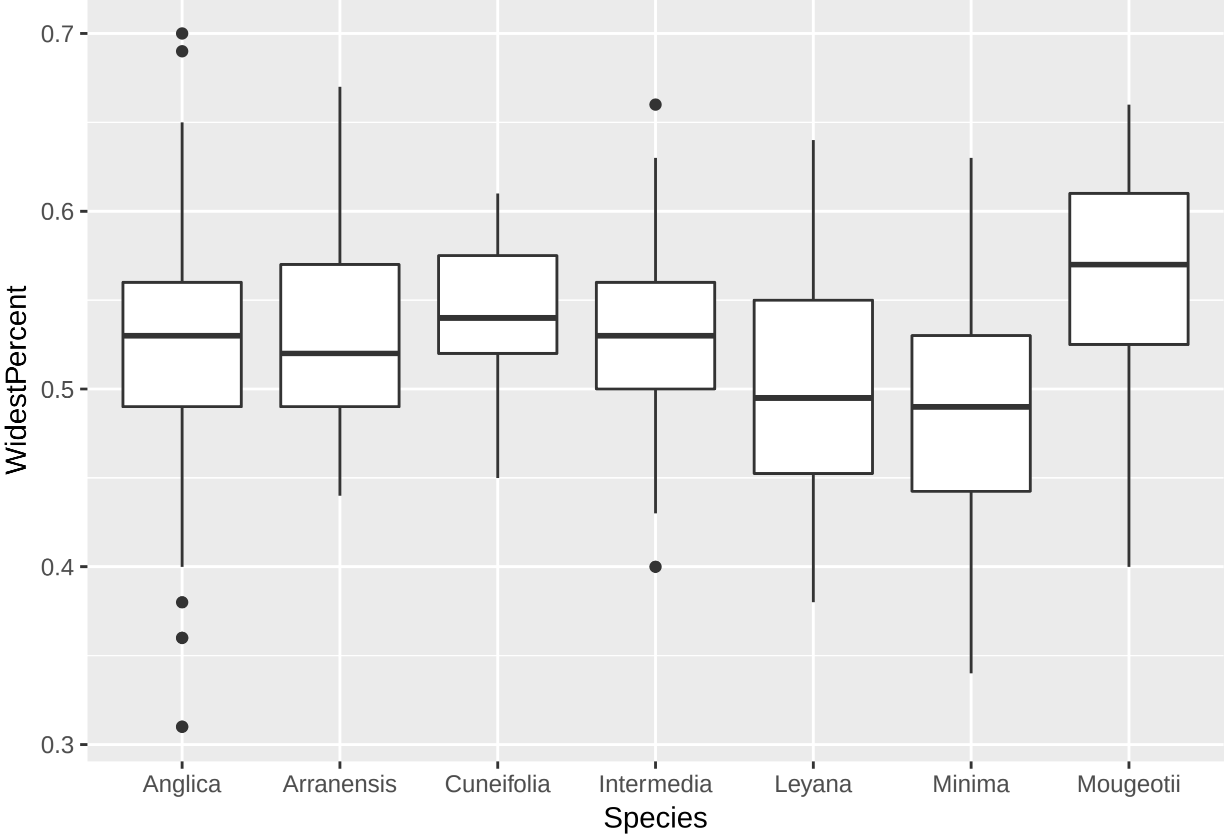

LeafLength
125
100
75
50
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species

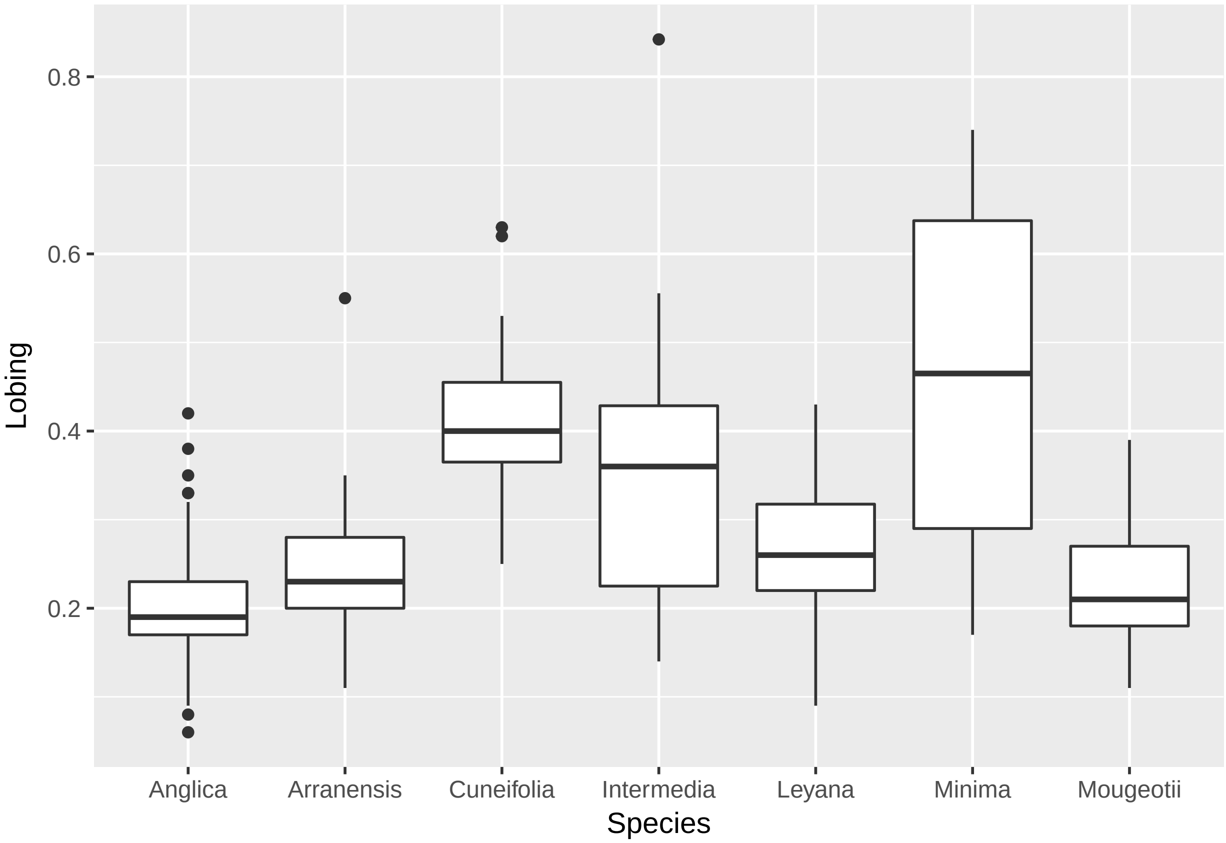

LeafWidth
100
75
50
25
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species

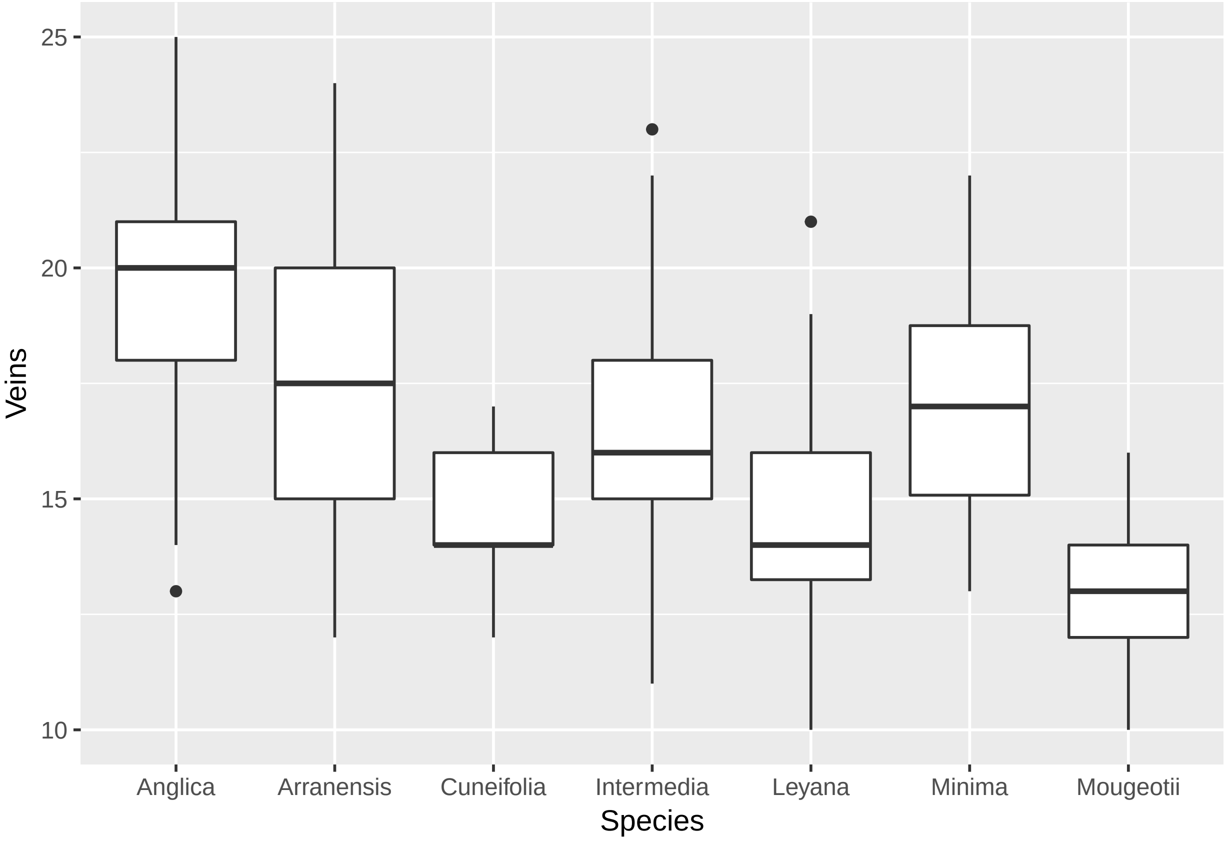

80
60
40
20
WidestPoint
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species

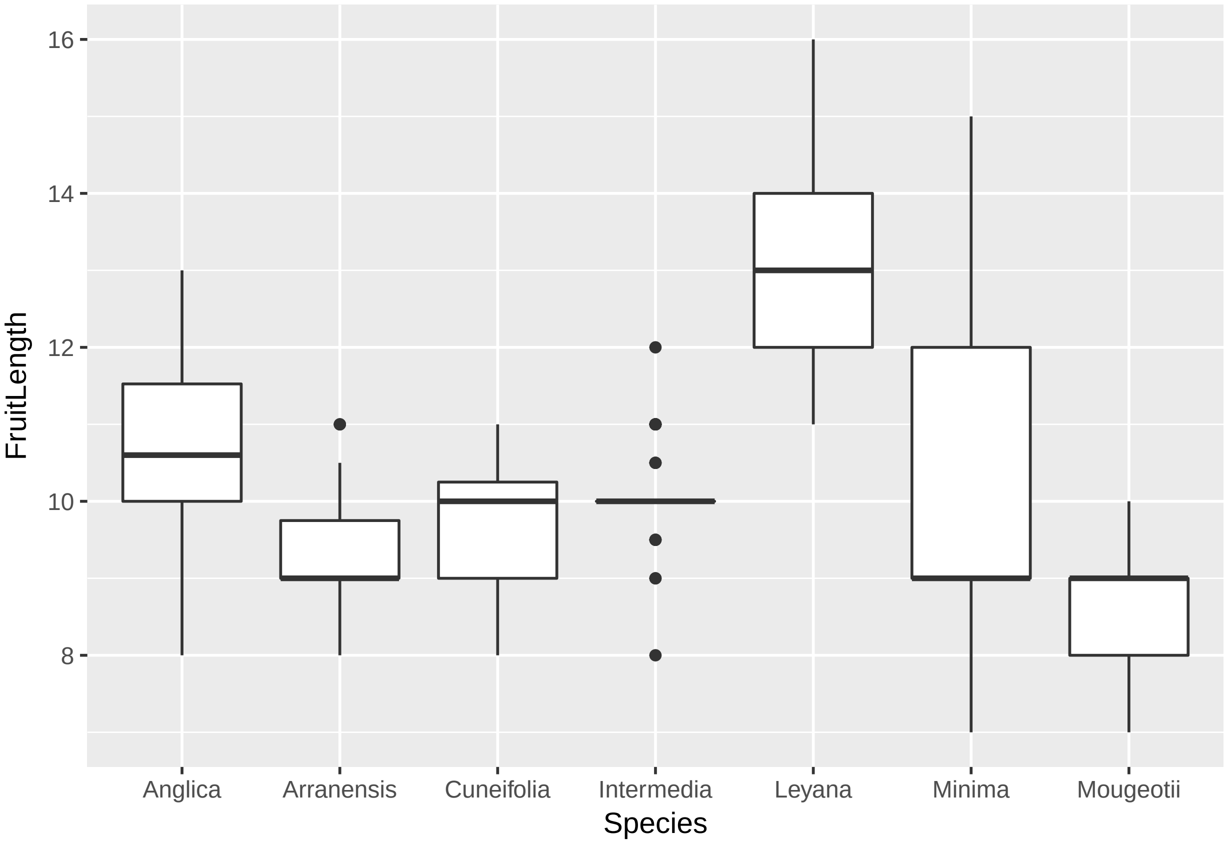

50
40
30
Angle
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species

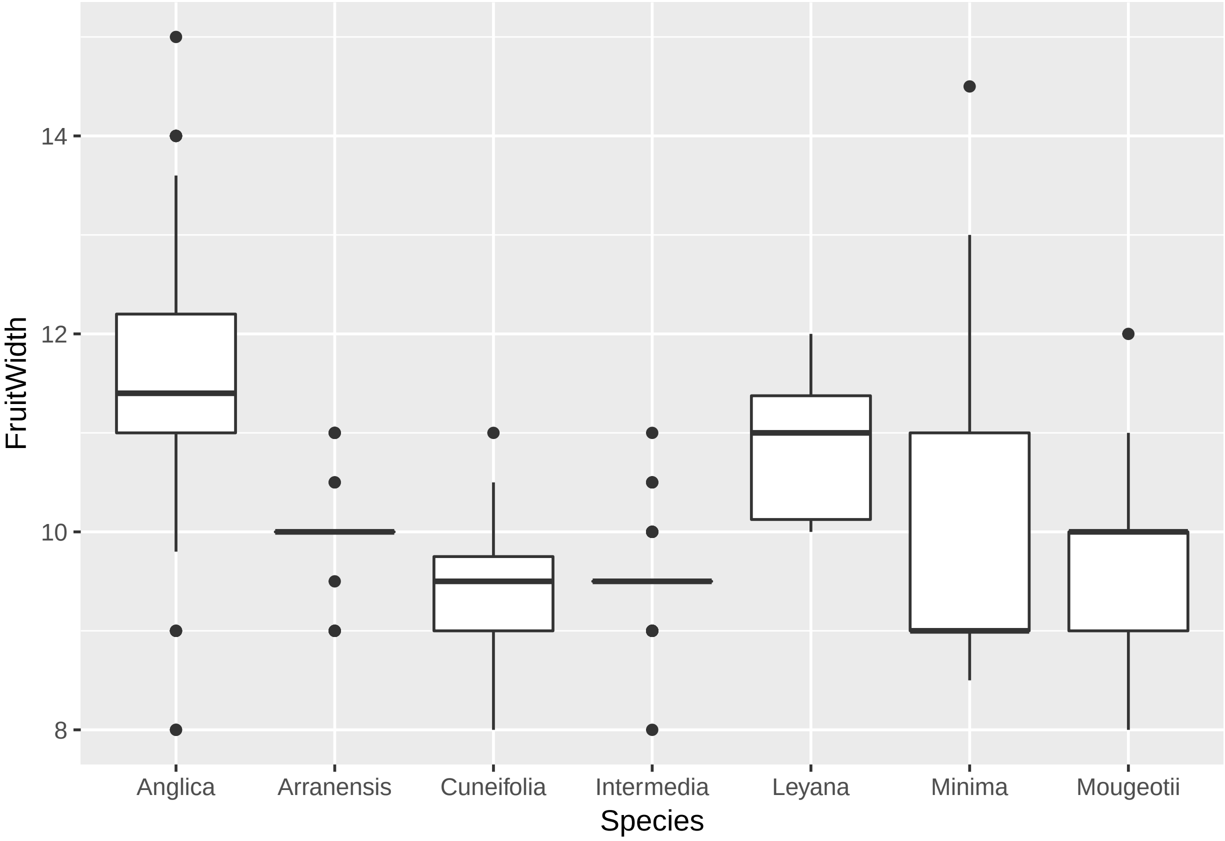

LeafRatio
2.5
2.0
1.5
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species

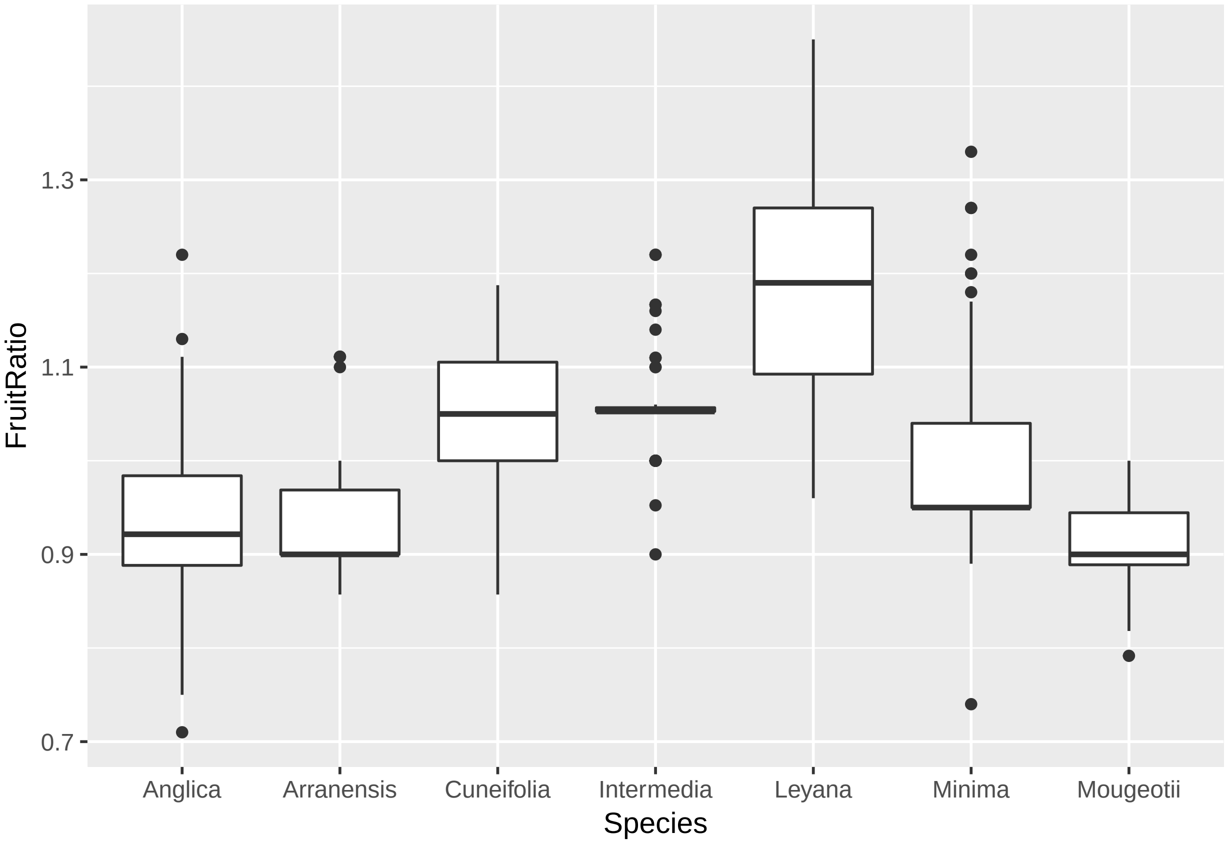

WidestPercent
0.7
0.6
0.5
0.4
0.3
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species


Lobing
0.8
0.6
0.4
0.2
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species


Veins
25
20
15
10
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species

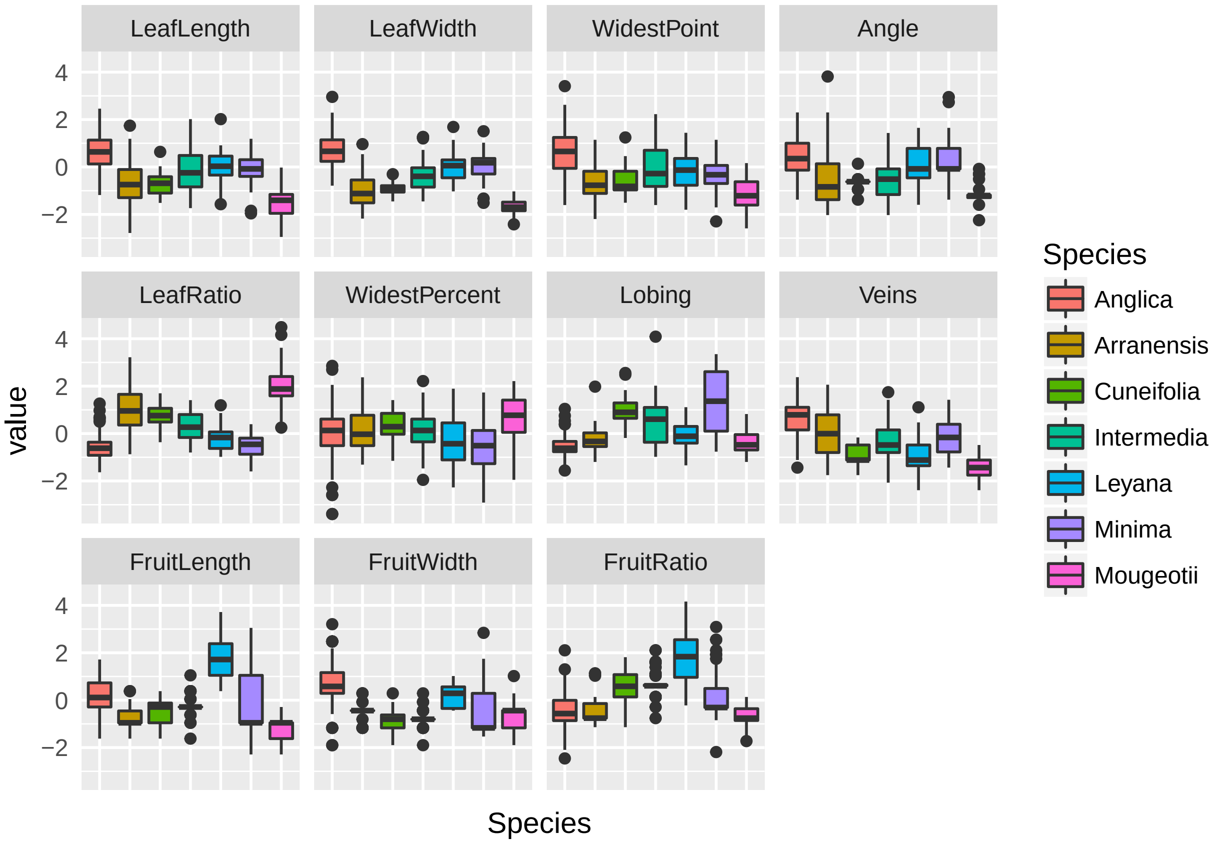

FruitLength
16
14
12
10
8
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species


FruitWidth
14
12
10
8
Anglica
Arranensis
Cuneifolia
Intermedia
Leyana
Minima
Mougeotii
Species



Try ggplot option for boxplots, but they'll need scaling because facet wrap will use same scale for all

```
## Using Species as id variables
```



Box plots clearly show the amount of overlap for different features across the species. Some features clearly separate one or two species from the others. For example the fruit length differentiates S. leyana and S. mougeotti and the fruit width for S.anglica does not overlap with other species. There is also considerable overlap of many features.

Analysis of varaince for each feature quantifies this.

```
##          Feature         F            p         r2
## 1     LeafLength 42.493932 1.905751e-39 0.40601652
## 2      LeafWidth 83.487163 6.401483e-66 0.57318893
## 3    WidestPoint 30.326764 1.279428e-29 0.32788020
## 4          Angle 22.566110 1.065977e-22 0.26632091
## 5      LeafRatio 99.022478 4.543084e-74 0.61432473
## 6  WidestPercent  5.294727 2.956507e-05 0.07848528
## 7         Lobing 58.156090 1.362569e-50 0.48333409
## 8          Veins 42.262660 2.860631e-39 0.40470107
## 9    FruitLength 45.222753 1.685044e-41 0.42110995
## 10    FruitWidth 42.083353 3.921765e-39 0.40367717
## 11    FruitRatio 57.052329 7.451557e-50 0.47855066
```

But what happens if we need to scale the data, repeated anovas for scaled dataframe

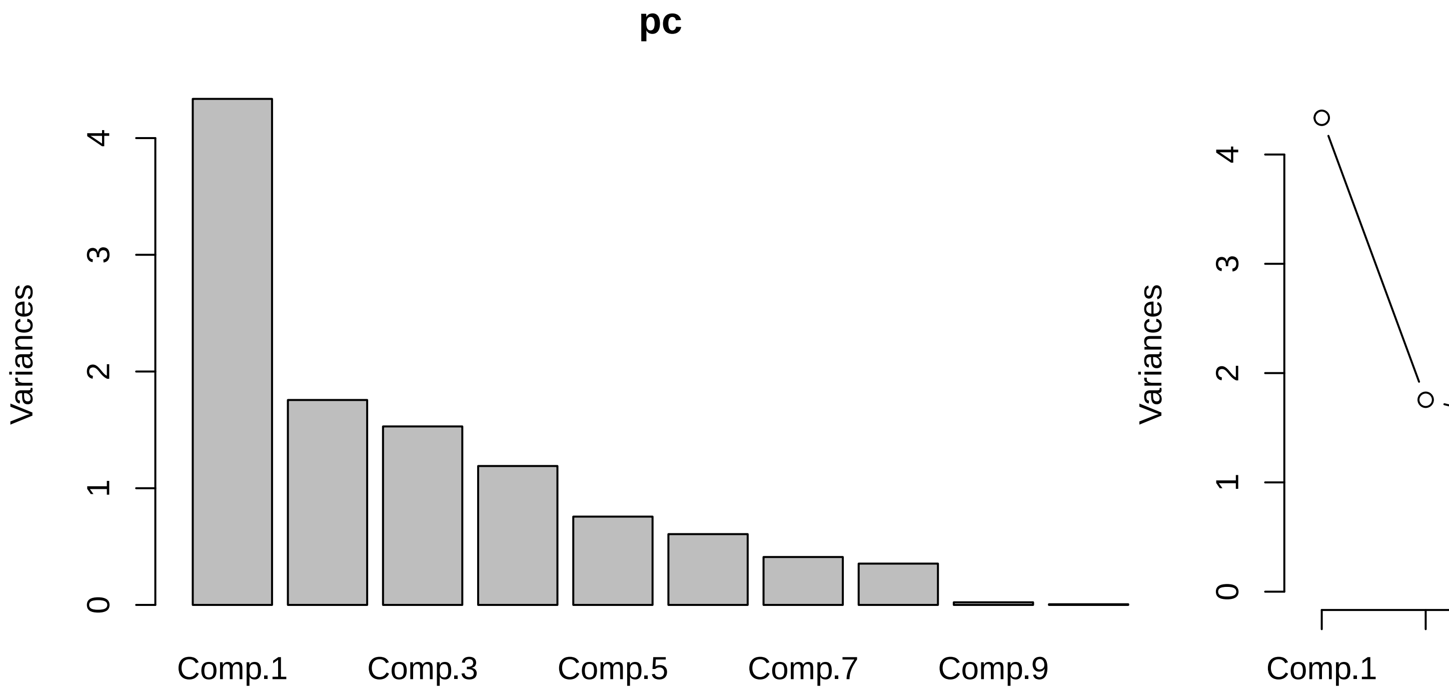
```
##          Feature         F            p         r2
## 1     LeafLength 42.493932 1.905751e-39 0.40601652
## 2      LeafWidth 83.487163 6.401483e-66 0.57318893
## 3    WidestPoint 30.326764 1.279428e-29 0.32788020
## 4          Angle 22.566110 1.065977e-22 0.26632091
## 5      LeafRatio 99.022478 4.543084e-74 0.61432473
## 6  WidestPercent  5.294727 2.956507e-05 0.07848528
## 7         Lobing 58.156090 1.362569e-50 0.48333409
## 8          Veins 42.262660 2.860631e-39 0.40470107
## 9    FruitLength 45.222753 1.685044e-41 0.42110995
## 10    FruitWidth 42.083353 3.921765e-39 0.40367717
## 11    FruitRatio 57.052329 7.451557e-50 0.47855066
```

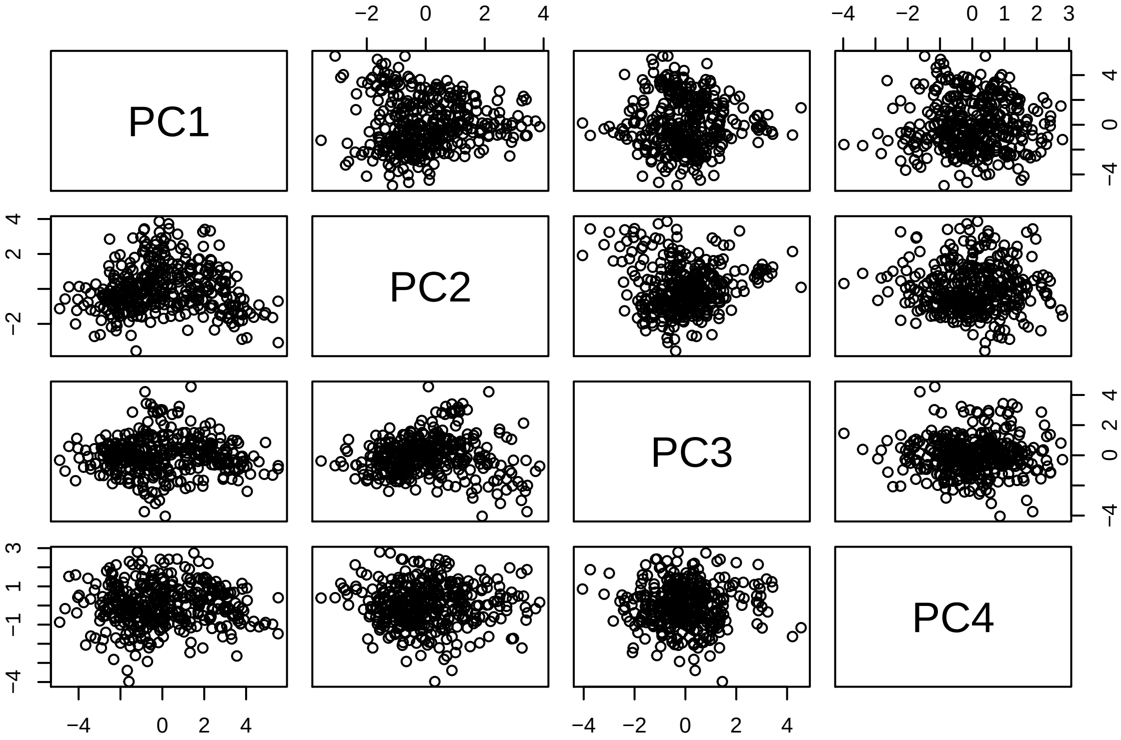
There is a difference with within group means and between group means, as we can see from the large F values, except in WidestPercent. LeafWidth, LeafLength and WidestPoint are highly correlated and may not all be requried in a model

Some simple cluster analysis

```
##
##                1  2  3  4  5  6  7
##   Anglica     28 35 40  0 35  5 17
##   Arranensis   4  1  2 15  1 17 10
##   Cuneifolia   0  0  1  3  0 10  5
##   Intermedia   6  1  8  3  3 15 12
##   Leyana      10  4  3  1  1  2  9
##   Minima      18  8  3  2  1  3 15
##   Mougeotii    0  0  0 17  0  5  1
```

The first plot shows the modelled clusters and the second is the scatterplot of the real data for comparison. As we suspected, S. anglica has been allocated to most clusters, S. cuniefolia and S. mougeotti are the most differentiated. What about principle components?





```
##                [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## LeafLength     0.18 0.00 0.00 0.07 0.08 0.00 0.03 0.10 0.03  0.01  0.49
## LeafWidth      0.21 0.00 0.02 0.02 0.00 0.00 0.06 0.04 0.61  0.01  0.04
## WidestPoint    0.13 0.06 0.01 0.23 0.00 0.00 0.02 0.02 0.17  0.00  0.36
## Angle          0.07 0.04 0.04 0.03 0.48 0.15 0.19 0.00 0.00  0.00  0.00
## LeafRatio      0.13 0.05 0.04 0.00 0.08 0.02 0.51 0.03 0.13  0.00  0.01
## WidestPercent 0.01 0.17 0.13 0.20 0.25 0.01 0.01 0.08 0.05  0.00  0.10
## Lobing         0.02 0.10 0.18 0.10 0.02 0.50 0.09 0.00 0.00  0.00  0.00
## Veins          0.12 0.04 0.05 0.01 0.07 0.00 0.06 0.65 0.00  0.00  0.00
## FruitLength    0.05 0.18 0.29 0.00 0.00 0.03 0.02 0.01 0.01  0.41  0.00
## FruitWidth     0.08 0.00 0.17 0.19 0.00 0.20 0.02 0.01 0.00  0.31  0.00
## FruitRatio     0.00 0.35 0.06 0.15 0.02 0.09 0.00 0.06 0.00  0.26  0.00
```

If leaf length is correlated with width and widest point and leaf width, AND pc shows that leafwidth, veins, leaf ratio, fruit length and widest percent together explain 80% of the variance, perhaps a model using only these is sufficient.

```
##
##                1  2  3  4  5  6  7
##   Anglica     18  4 37 43 33  0 25
##   Arranensis   6 16  1  2  5 20  0
##   Cuneifolia   3 13  0  0  0  3  0
##   Intermedia  14 16  3  3  7  5  0
##   Leyana      10  4  1  1 13  0  1
##   Minima      13  2  4  7 21  2  1
##   Mougeotii    0  1  0  0  0 22  0
```

Again, S. cuneifolia, S mougeotti are differentiated, there does not appear to be much improvement in the other species.

Modelling. Supervised Learning Create training and test sets. Cross validation will be used to train the model. The caret package is need to do this within the model fitting algorithm, otherwise you would need to write your own for loops for each fold of cross validation.

Classification Random forest - many decision trees fittied by repeated bootstrapping.

```
## Loading required package: lattice
```

```
## Warning: replacing previous import by 'plyr::ddply' when loading 'caret'
```

```
## Warning: replacing previous import by 'rlang::!!' when loading 'recipes'
```

```
## Warning: replacing previous import by 'rlang::expr' when loading 'recipes'
```

```
## Warning: replacing previous import by 'rlang::f_lhs' when loading 'recipes'
```

```
## Warning: replacing previous import by 'rlang::f_rhs' when loading 'recipes'
```

```
## Warning: replacing previous import by 'rlang::is_empty' when loading
## 'recipes'
```

```
## Warning: replacing previous import by 'rlang::lang' when loading 'recipes'
```

```
## Warning: replacing previous import by 'rlang::na_dbl' when loading
## 'recipes'
```

```
## Warning: replacing previous import by 'rlang::names2' when loading
## 'recipes'
```

```
## Warning: replacing previous import by 'rlang::quos' when loading 'recipes'
```

```
## Warning: replacing previous import by 'rlang::sym' when loading 'recipes'
```

```
## Warning: replacing previous import by 'rlang::syms' when loading 'recipes'
```

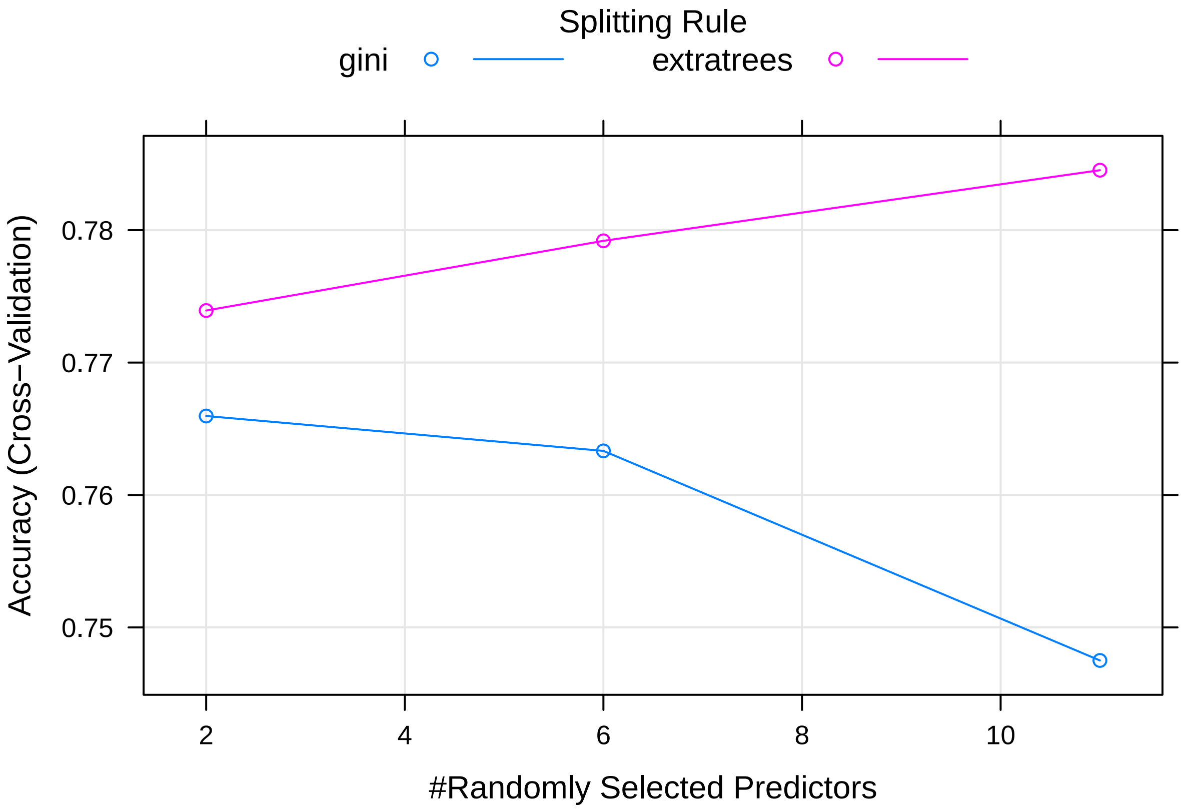
```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following object is masked from 'package:GGally':
##
##     nasa
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



```
## Random Forest
##
## 380 samples
##  11 predictors
##   7 classes: 'Anglica', 'Arranensis', 'Cuneifolia', 'Intermedia', 'Leyana', 'Minima', 'Mougeotii'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 304, 305, 305, 303, 303
## Resampling results across tuning parameters:
##
##   mtry  splitrule   Accuracy   Kappa
##    2    gini        0.7659645  0.6825944
##    2    extratrees  0.7739294  0.6908270
##    6    gini        0.7633320  0.6817737
##    6    extratrees  0.7791925  0.7004835
##   11    gini        0.7475056  0.6621390
##   11    extratrees  0.7845259  0.7090933
##
## Accuracy was used to select the optimal model using  the largest value.
## The final values used for the model were mtry = 11 and splitrule
##  = extratrees.
```