Plan:

* KMeans mit allen drei Arten: Laevigata, Monogyna, Rhipidophylla
* Feature Importance der drei Arten
* Neue Methoden?
* PCA mit allen drei Arten
* Welche Merkmale teilen sie in Gruppen ein? Passen die Loadings dazu? Mit Literatur vergleichen?
* Center finden und so genaue Merkmale feststellen
* Kombinieren von PCA und KMeans
* Laevigata von den anderen Arten trennen

04.10.2021

Code-File: KMeans\_Crataegus\_Laevigata\_Rhipidophylla\_Monogyna.py

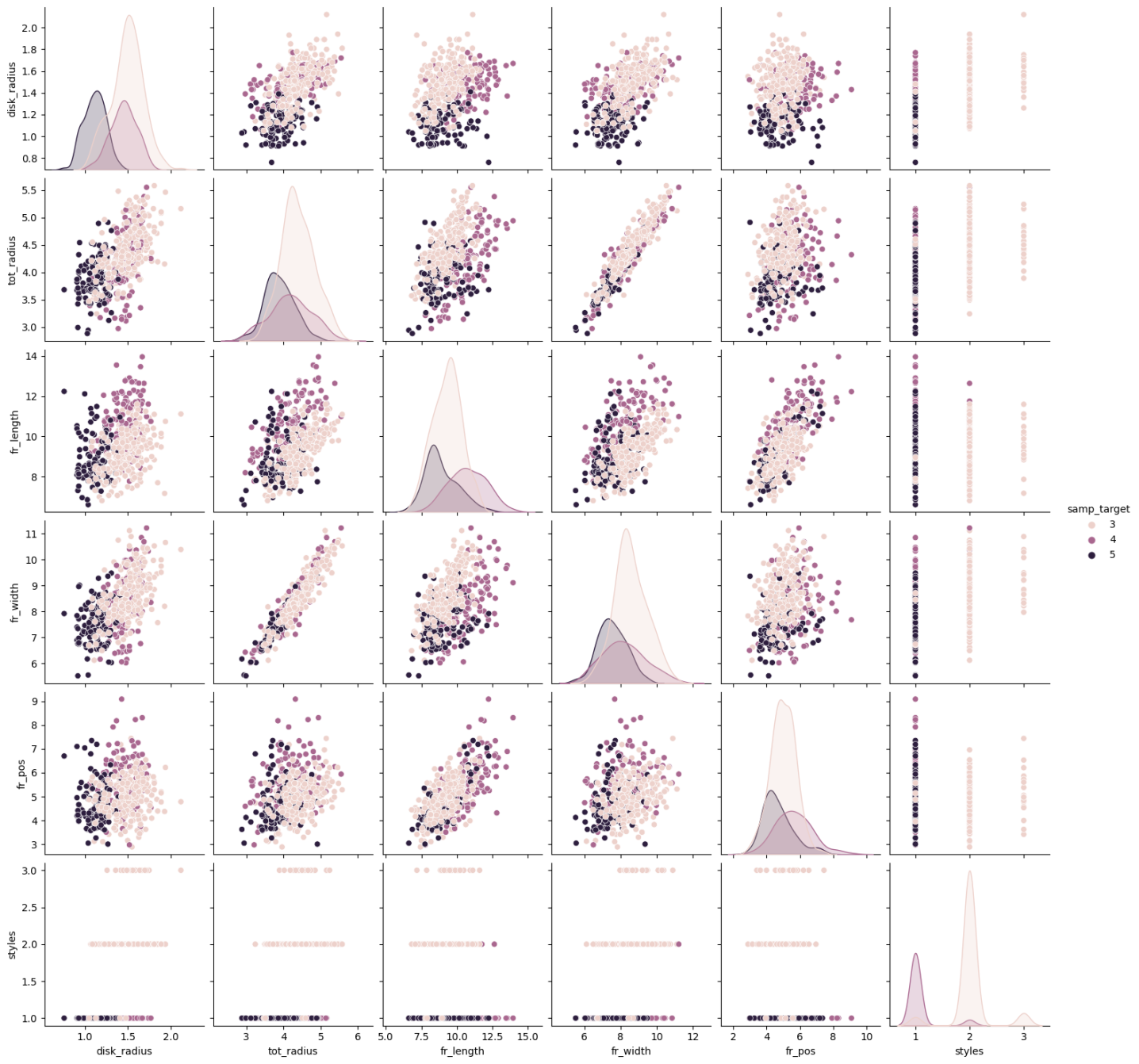
Loading dataset “RearrangedTable.csv” and kept only data regarding Laevigata, Monogyna and Rhipidophylla. Species are encoded:

|  |  |
| --- | --- |
| 3 | Laevigata |
| 4 | Rhipidophylla |
| 5 | Monogyna |

Scatterplot with Seaborn:

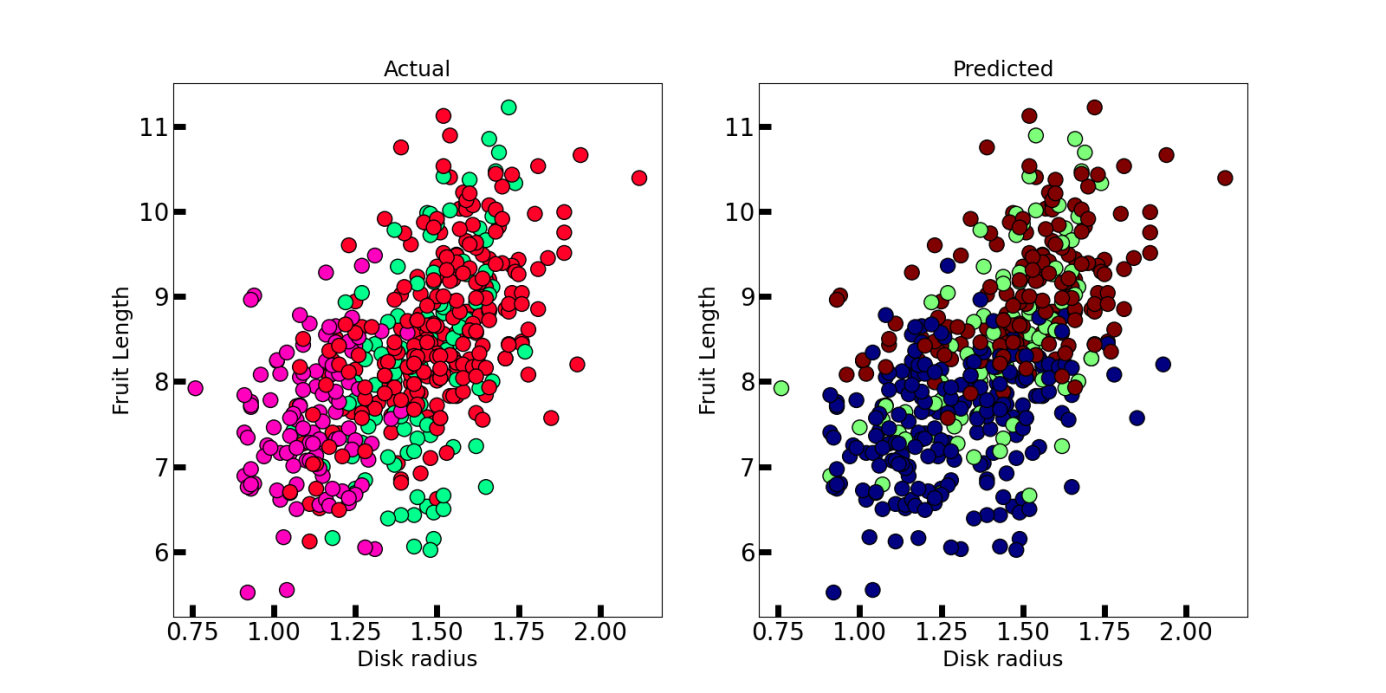
sns\_plot = sns.pairplot(df\_LRM, hue='samp\_target', height=2.5)

sns\_plot.savefig("Scatterplot\_CombinationallFeatures\_Laev\_Mon\_Rhip.png")

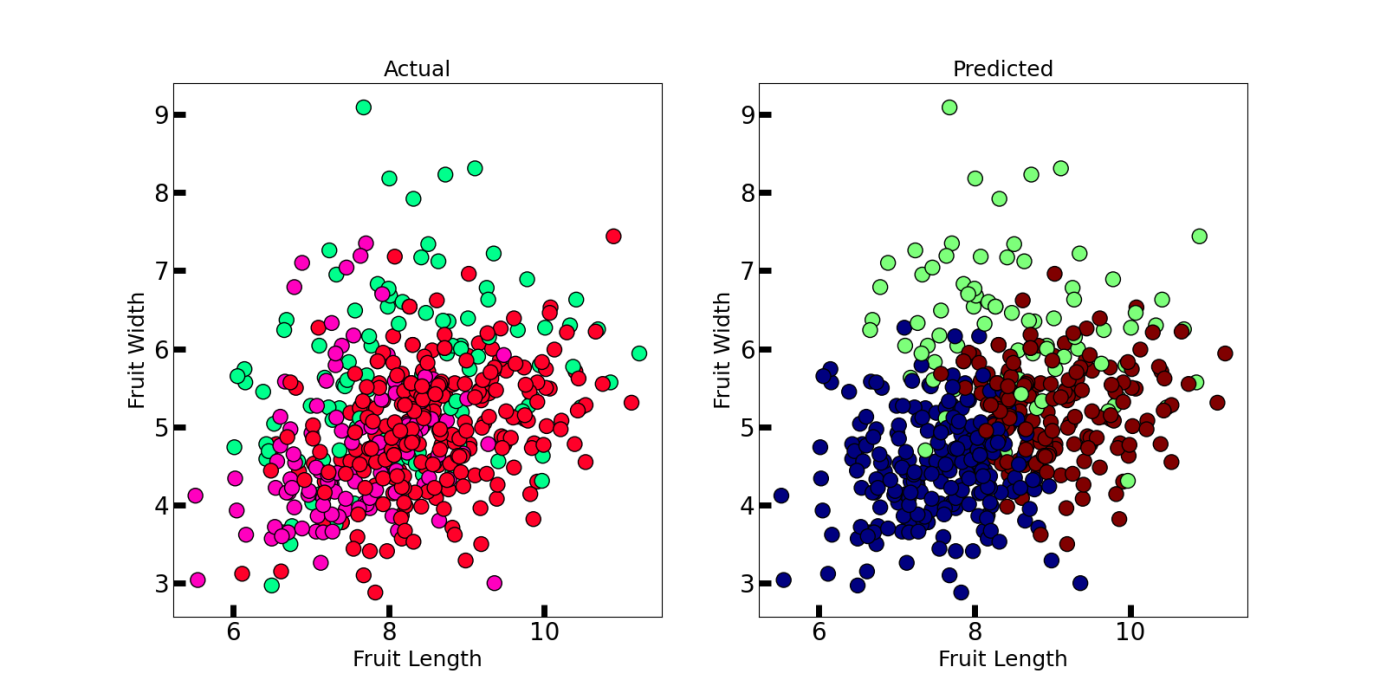


Scatterplots show that the species cannot really be separated only by the features of the fruits. Maybe disk radius and fruit length or fruit length and fruit width. Also styles indicate a lot but only for Monogyna. The distribution graphs however show that Laevigata is quite different compared to the other ones. It’s probably easier to make a separation only with two of the species, where one of them is Laevigata.

KMeans with disk radius – fruit length and fruit length – fruit width:



It looks more like there are only two clusters which does make sense according to the distribution graphs.



KMeans\_Crataegus\_Laevigata\_Rhipidophylla\_Monogyna\_ver2.py is a different version, but I didn’t use the output.

06.10.2021

Feature Importance:

Code: FeatureSelection\_Laev\_Mon\_Rhyp.py

Tutorial: <https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b>

Loading dataset “RearrangedTable.csv” and kept only data regarding Laevigata, Monogyna and Rhipidophylla. Species are encoded:

|  |  |
| --- | --- |
| 3 | Laevigata |
| 4 | Rhipidophylla |
| 5 | Monogyna |

Correlation:

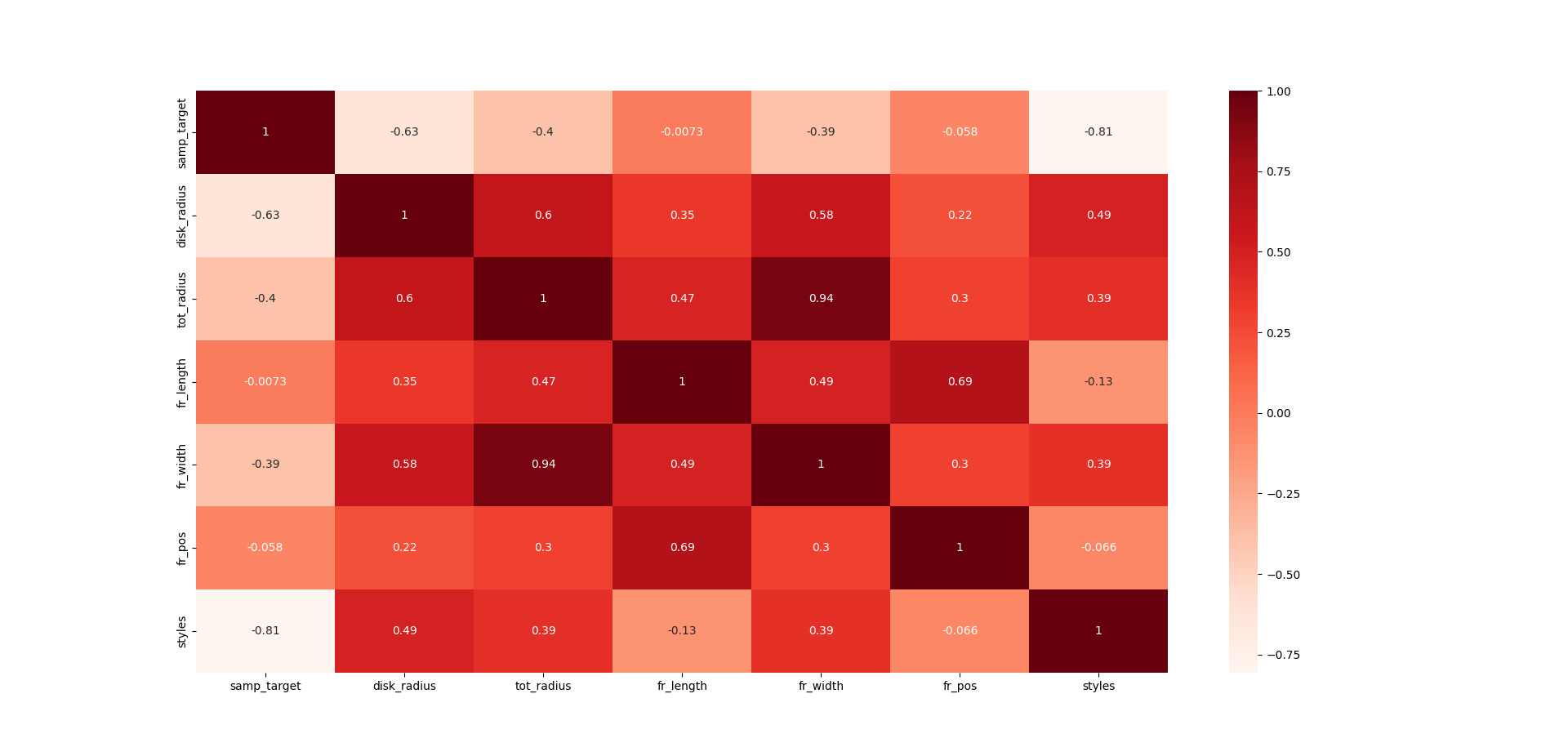
The correlation coefficient has values between -1 to 1

— A value closer to 0 implies weaker correlation (exact 0 implying no correlation)

— A value closer to 1 implies stronger positive correlation

— A value closer to -1 implies stronger negative correlation

Using Pearson Correlation:



High correlation (above 0.5):

|  |  |  |
| --- | --- | --- |
| First feature | Second feature | Corr. coefficient |
| Disk radius | Fruit width | 0.58 |
| Disk radius | Total radius | 0.6 |
| Total radius | Fruit width | **0.94** |
| Fruit length | Fruit position | 0.69 |

Correlation with output variable:

Output:

samp\_target 1.000000

disk\_radius 0.626925

styles 0.806804

Name: samp\_target, dtype: float64

Disk radius and styles are correlated with the output variable “samp\_target”

Does it make sense to use? Used samp target for the tutorial’s “MEDV” which is a float.

disk\_radius styles

disk\_radius 1.000000 0.470509

styles 0.470509 1.000000

Not above 0.5, I would keep both.

Wrapper Method:

The performance metric used here to evaluate feature performance is pvalue. If the pvalue is above 0.05 then we remove the feature, else we keep it.

Model 1: const 1.435213e-136

disk\_radius 1.523325e-22

tot\_radius 2.983396e-01

fr\_length 3.799661e-01

fr\_width 9.991064e-01

fr\_pos 2.635867e-02

styles 2.592749e-71

Backward Elimination:

['disk\_radius', 'tot\_radius', 'fr\_pos', 'styles']

Fruit length and width are both greater than 0.05, therefore the features should be removed according to this analysis

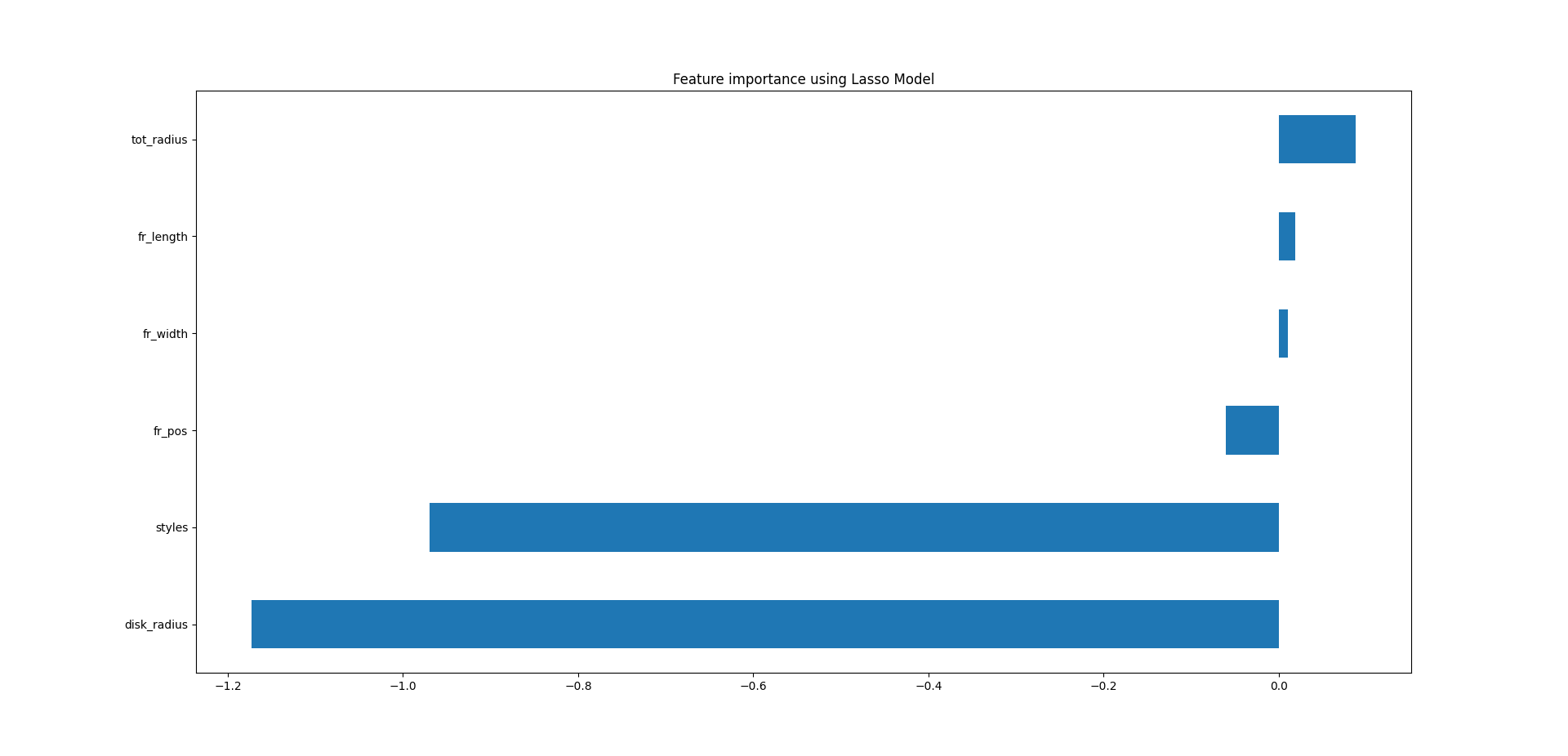
Embedded Method

Embedded methods are iterative in a sense that takes care of each iteration of the model training process and carefully extract those features which contribute the most to the training for a particular iteration. Regularization methods are the most commonly used embedded methods which penalize a feature given a coefficient threshold.

Here we will do feature selection using Lasso regularization. If the feature is irrelevant, lasso penalizes it’s coefficient and make it 0. Hence the features with coefficient = 0 are removed and the rest are taken.

Output:

Lasso picked 6 variables and eliminated the other 0 variables

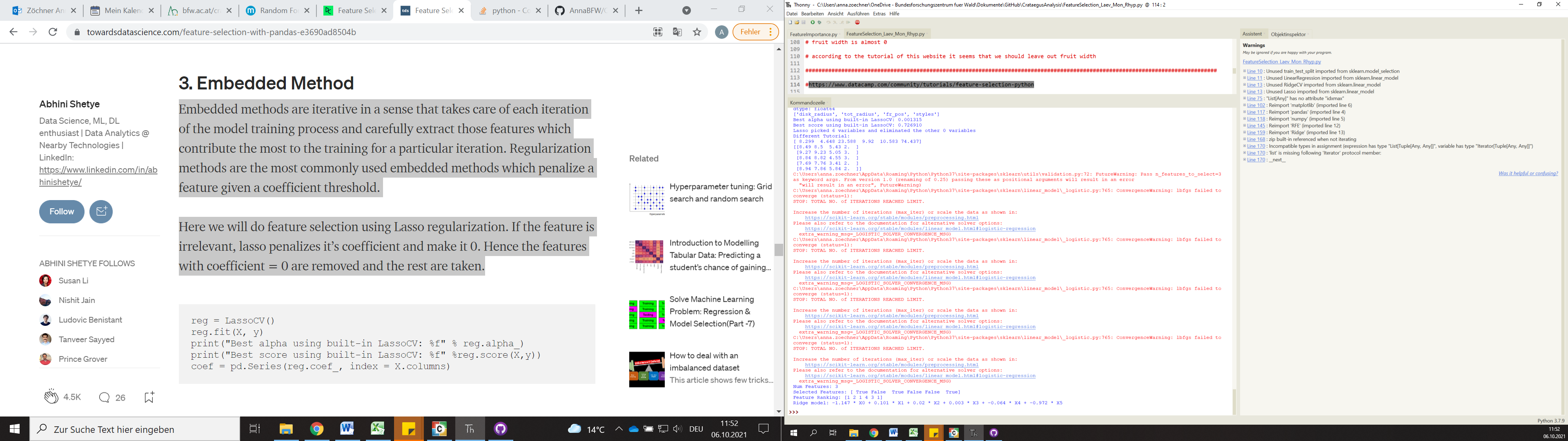


According to Lasso we should keep all features. However, the fruit width is really close to 0 which is why I would suggest leaving it as it also fits to the analysis before.

Different Tutorial:

<https://www.datacamp.com/community/tutorials/feature-selection-python>

I also tried a different tutorial but I got the following warning:



As it was just a future warning, I still got results.

Output:

Num Features: 3

Selected Features: [ True False True False False True]

Feature Ranking: [1 2 1 4 3 1]

Ridge model: -1.147 \* X0 + 0.101 \* X1 + 0.02 \* X2 + 0.003 \* X3 + -0.064 \* X4 + -0.972 \* X5

According to this we should only keep: disk radius, fruit length, styles

Which makes sense as total radius and fruit width are highly correlated and don’t provide “additional” information.

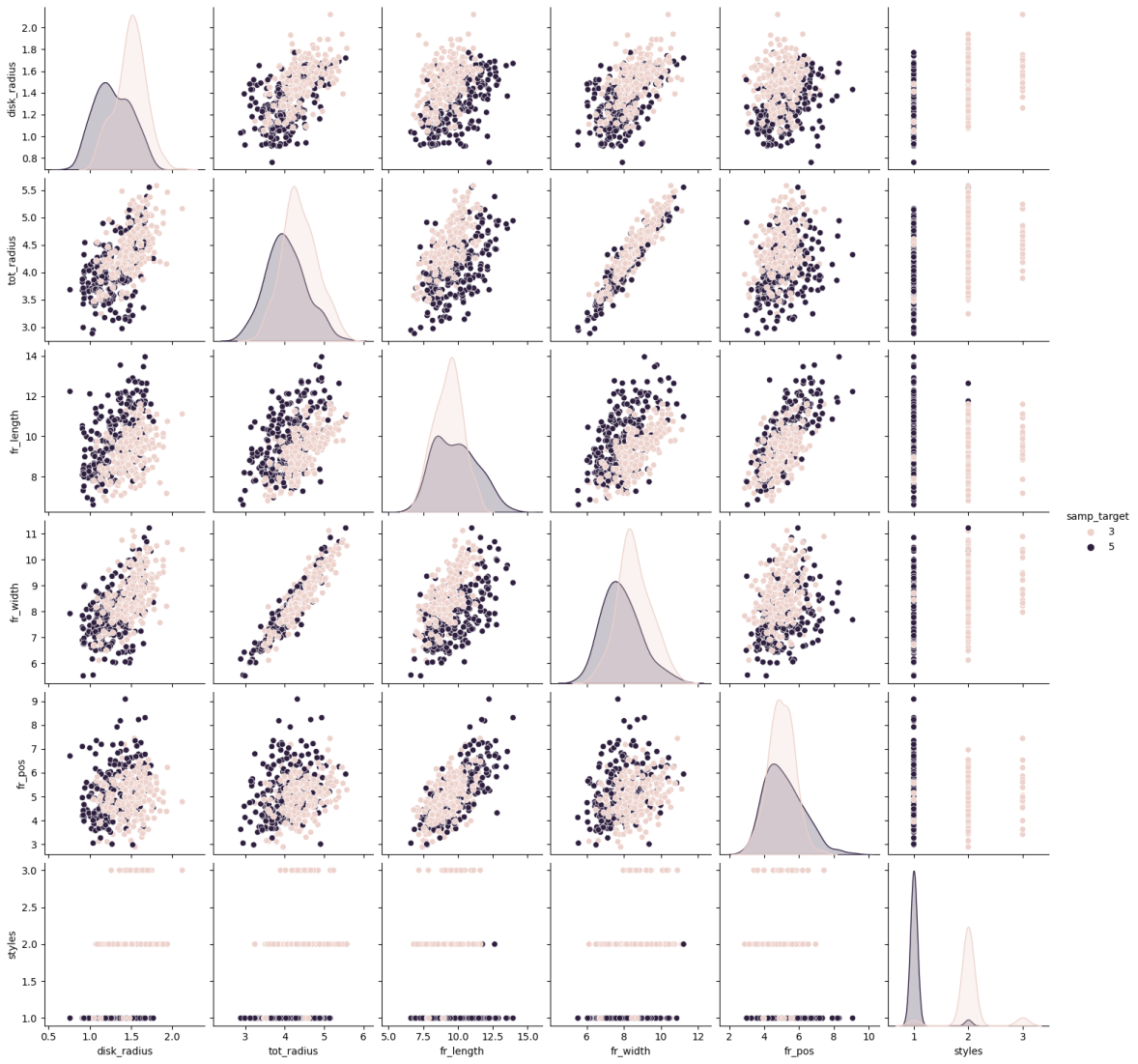
07.10.2021

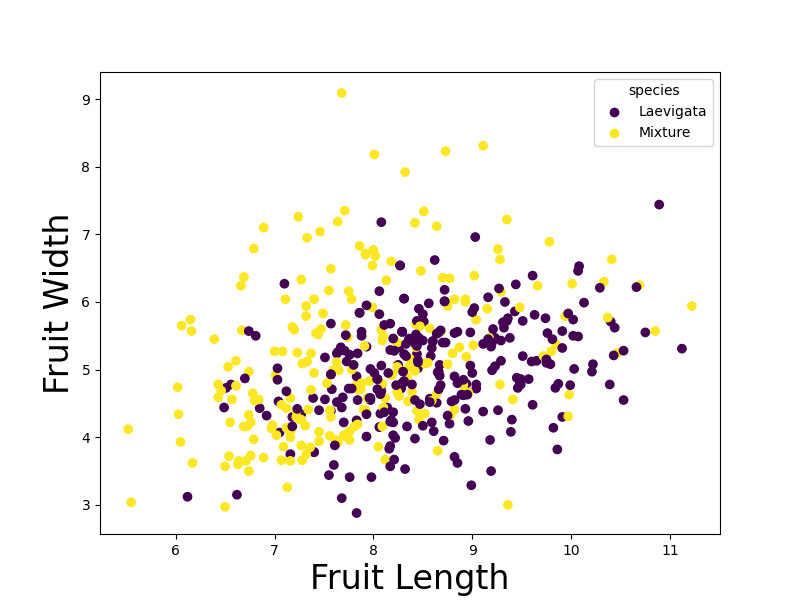
Laevigata seems to differentiate from the two other species a lot.

Idea: Put Monogyna and Rhipidophylla together and see if there are two clusters

Same Code as before

* KMeans\_Laevigata\_against\_MonogynaRhipidophylla.py





Did not improve anything

Looked again at Feature Selection – Analysis:

Every method says something a bit different. However, fruit width and total radius are highly correlated. Those measurements are actually the same. If we keep both of them it is redundant. Not sure what is the best to do with the fruit position. I would like to keep it as it says something about the shape.

Next Step: PCA with disk radius, fruit length, fruit position, styles

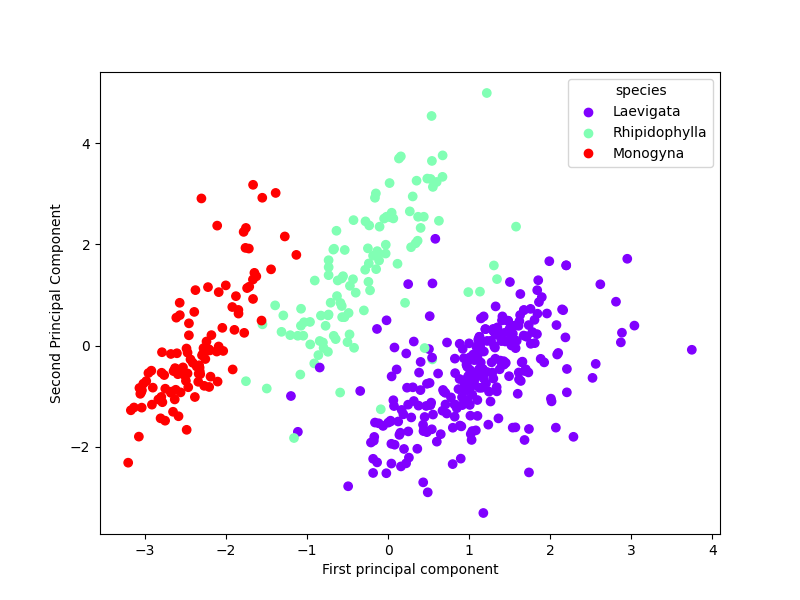
PCA with all species

Features according to previous analysis:

Left out fruit width and total radius because they are highly correlated

Dropped fruit width because of feature selection analysis

Left sepals as always



Best output so far. Groups are definitely visible. However some Rhipidophylla and Laevigata fruits are overlapping.

Next: find the centre of each cluster, make circle around it and detect the “perfect” fruits for each species

11.10.2021

Other Literature:

Topodeme differentiation in Ontario taxa of Crataegus (Rosaceae: Maloideae): leaf morphometric evidence

What measurements of the fruits?

* Styles, Fruit length, fruit width

13.10.2021

Exploratory Data Analysis:

Tutorial: <http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.html>

Overview:

> summary(df)

**samp\_target disk\_radius tot\_radius fr\_length**

Min. :0.000 Min. :0.760 Min. :2.880 Min. : 6.000

1st Qu.:2.000 1st Qu.:1.230 1st Qu.:3.890 1st Qu.: 8.637

Median :3.000 Median :1.430 Median :4.220 Median : 9.495

Mean :2.845 Mean :1.401 Mean :4.228 Mean : 9.579

3rd Qu.:4.000 3rd Qu.:1.560 3rd Qu.:4.550 3rd Qu.:10.380

Max. :5.000 Max. :2.120 Max. :5.850 Max. :13.960

**fr\_width fr\_pos styles**

Min. : 5.510 Min. :2.700 Min. :1.000

1st Qu.: 7.560 1st Qu.:4.465 1st Qu.:1.000

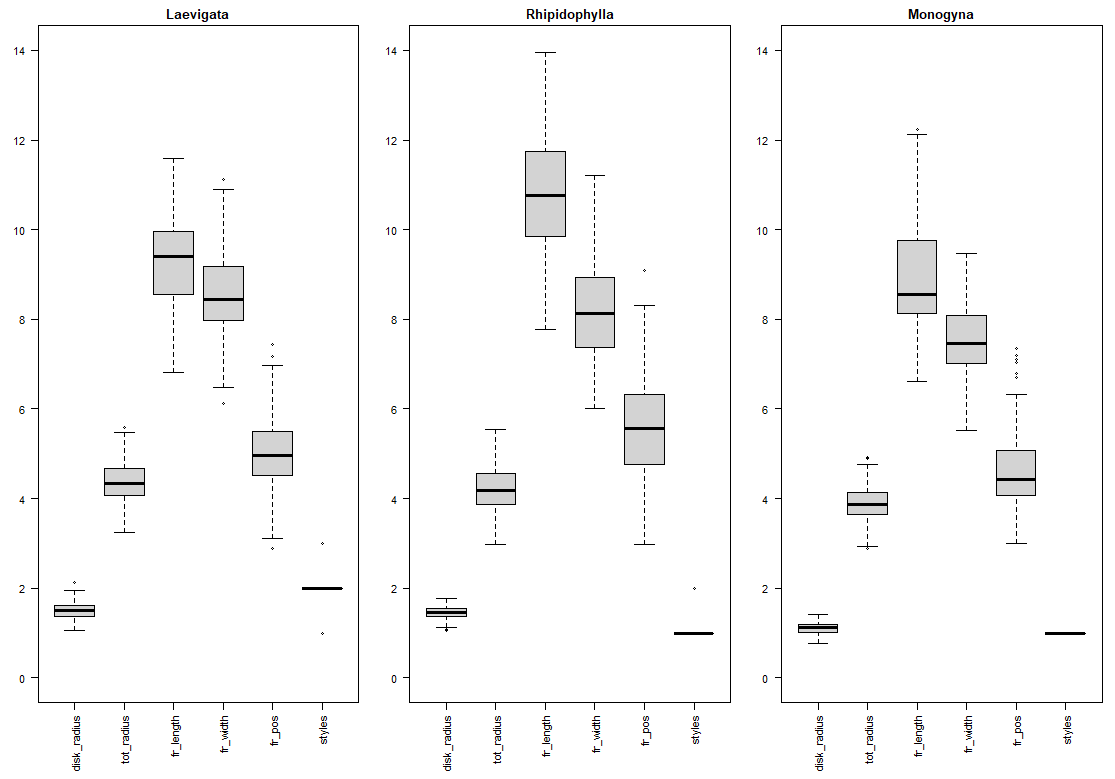
Median : 8.210 Median :5.035 Median :2.000

Mean : 8.250 Mean :5.120 Mean :1.548

3rd Qu.: 8.875 3rd Qu.:5.662 3rd Qu.:2.000

Max. :11.610 Max. :9.090 Max. :3.000

Boxplot:



Disk Radius: slightly smaller for Monogyna, Laevigata and Rhipidophylla quite the same

Total Radius: slightly smaller for Monogyna, Laevigata and Rhipidophylla quite the same

Fruit Length: Each of them different, Rhipidophylla apparently larger

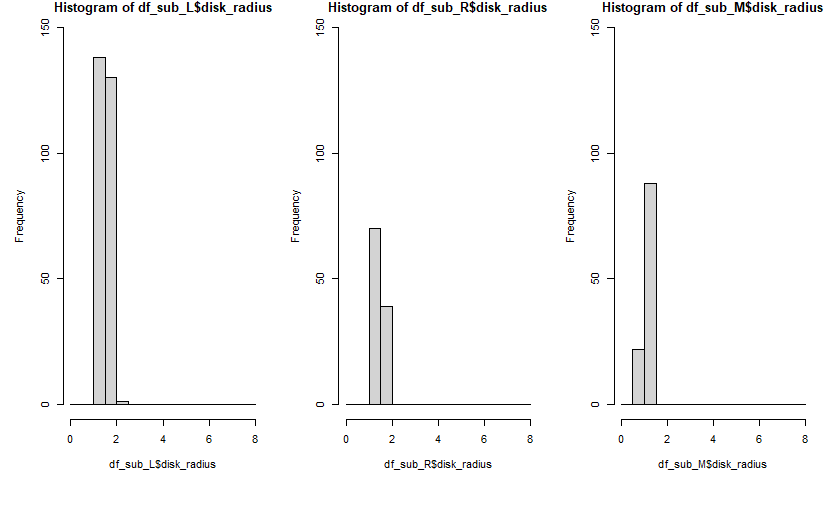
Fruit Width: Not exactly like total radius, few differences visible. Monogyna again slightly smaller. Range differs for each species.

Fruit position: Laevigata and Monogyna seem to have similar shape. Rhipidophylla is higher, means the widest part of the fruit is closer to sepals.

Styles: Laevigata has mostly two styles.

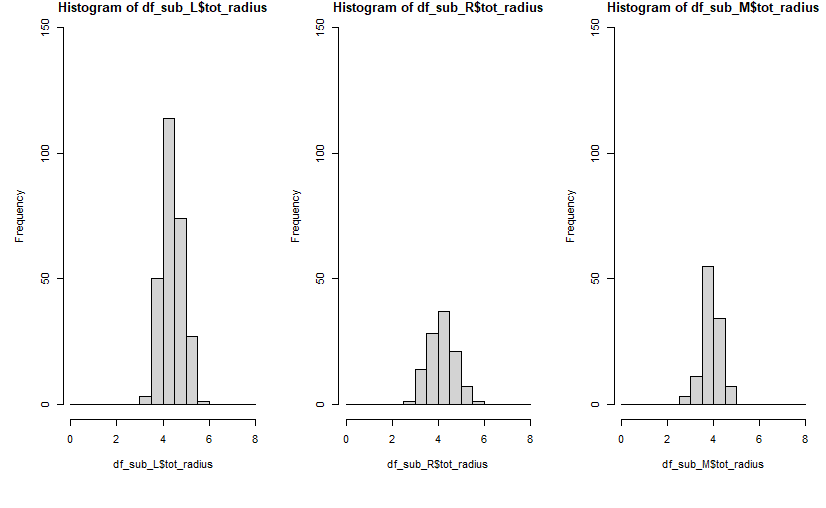
Histogram for each feature:

1. Disk Radius



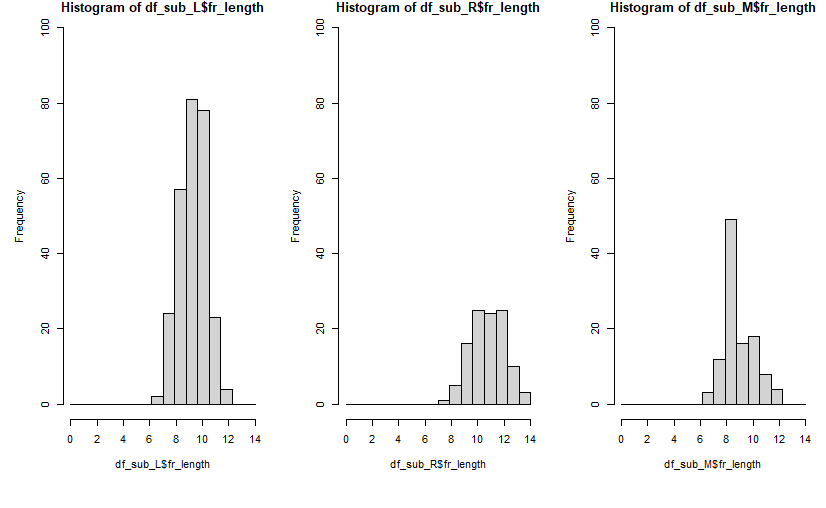
Range and distribution similar.

1. Total Radius



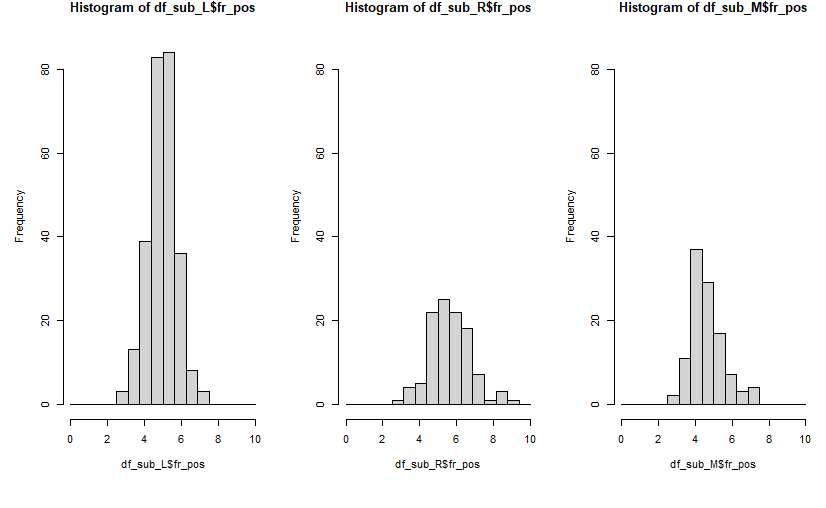
Range and distribution similar.

1. Fruit Length



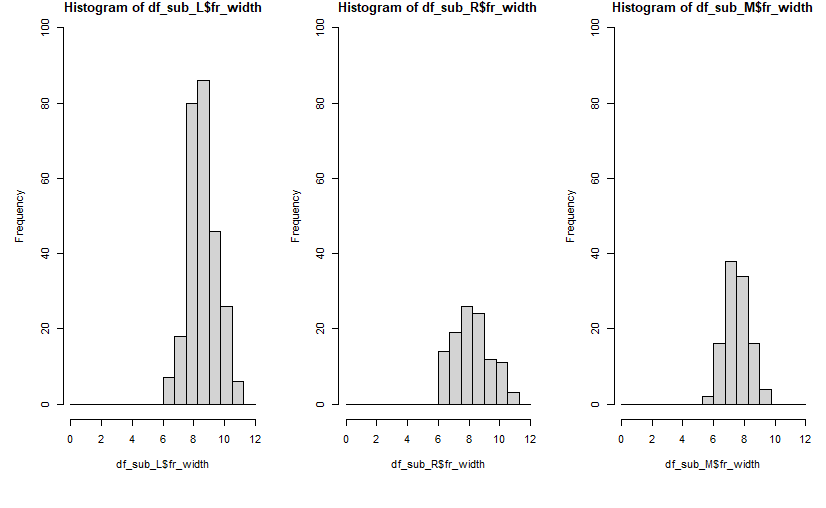
Range for fruit length of Rhipidophylla is shifted to the right. Fruits are apparently larger.

1. Fruit position



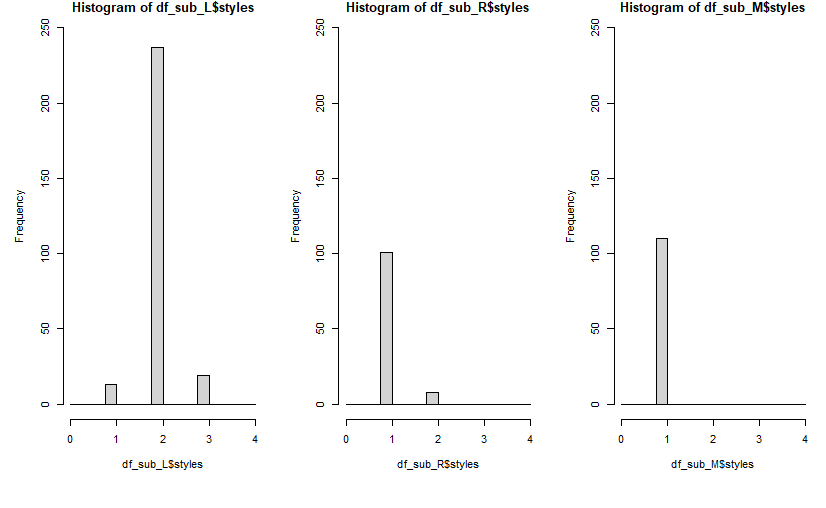
Range of fruit position for Rhipidophylla larger. Variance bigger.

1. Fruit width



Similar.

1. Styles



Laevigata has different number of styles. Rhipidophylla mostly 1. Monogyna just 1.

18.10.2021

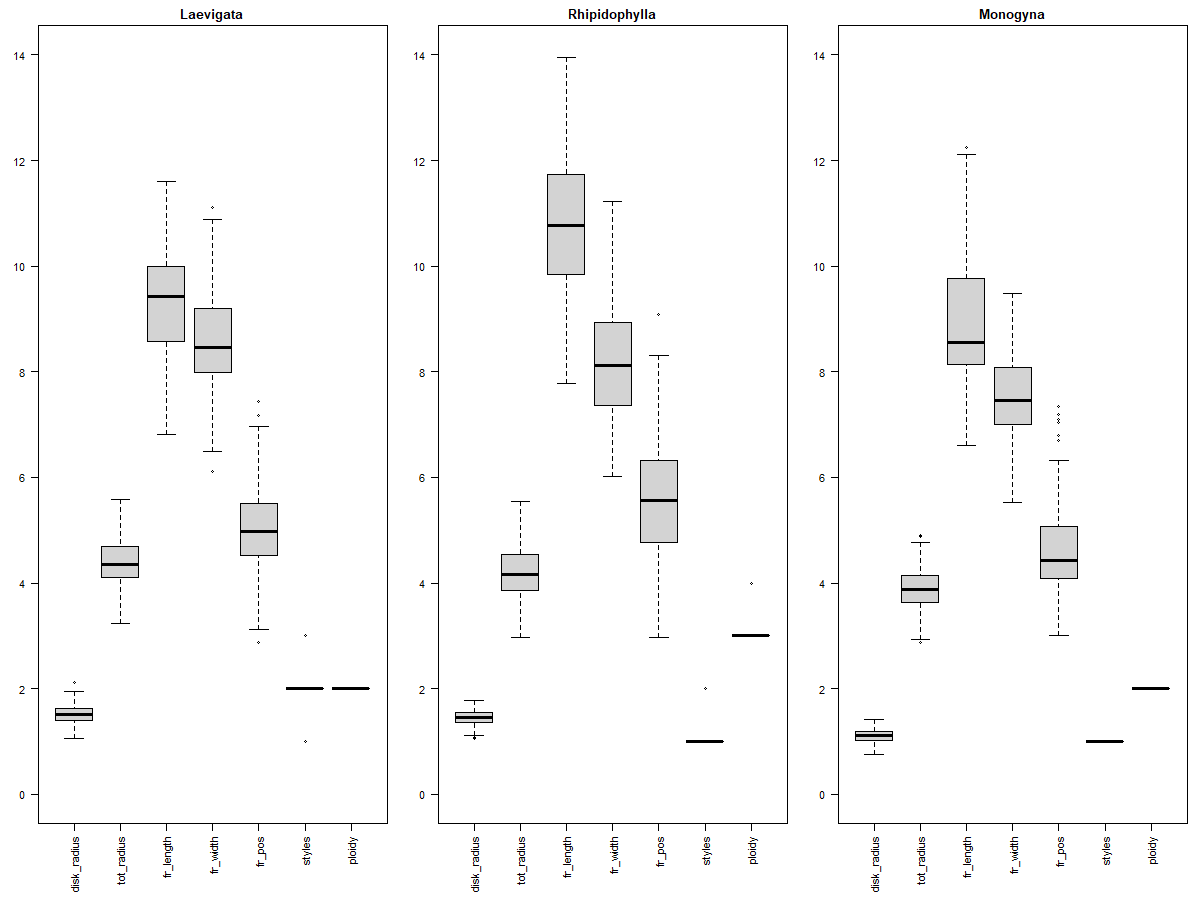
Ploidy data:

I merged ploidy data and morphometric data on the samp\_name in order to analyse the whole dataset.

First I read it in. Headers had to be added, the x in the ploidy column had to be deleted and then the column was converted to numeric.

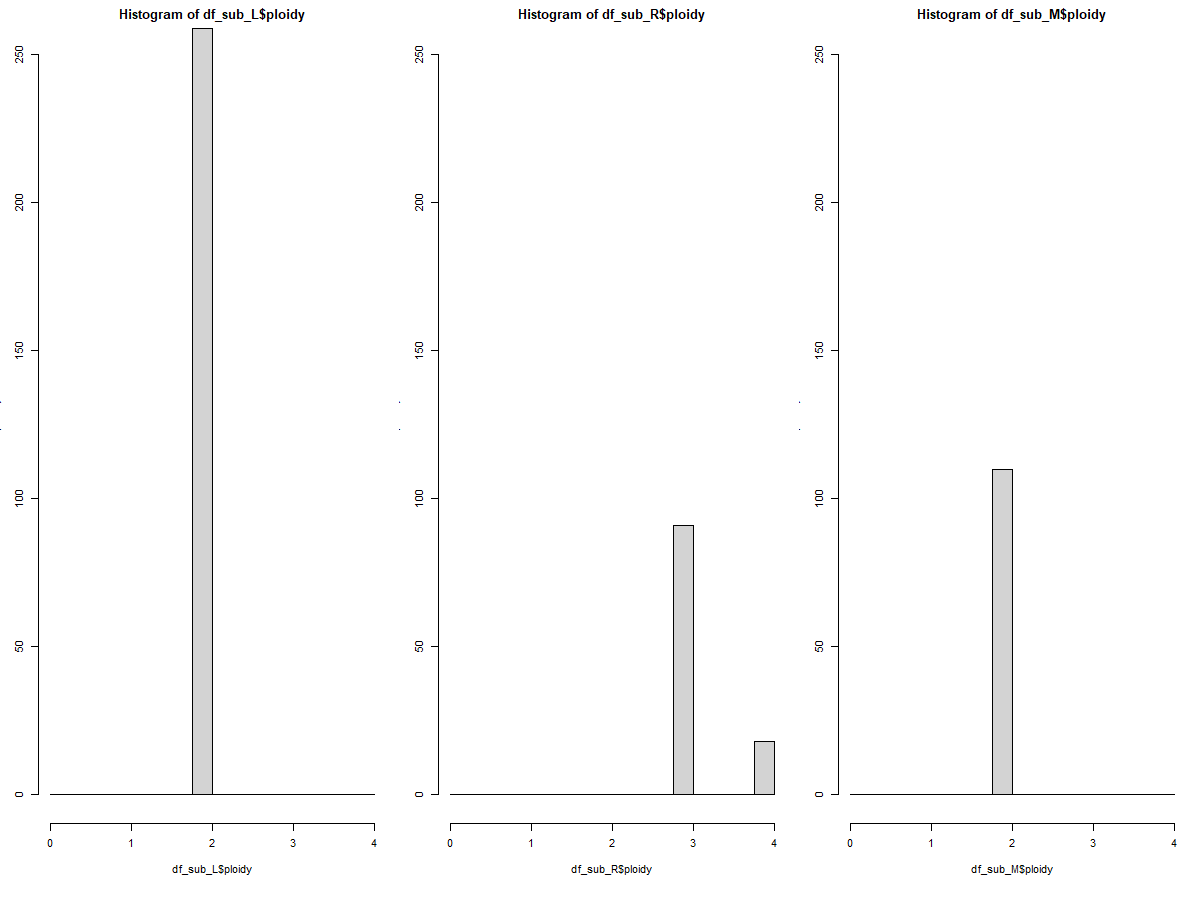
|  |
| --- |
| # Read in table with ploidy data  crataegus\_ploidy <- read.csv("Ploidy\_Crataegus.csv", na="NA", sep=";", header = FALSE)  colnames(crataegus\_ploidy) = c("samp\_name","place","ploidy") # give dataframe header  crataegus\_ploidy$ploidy = substr(crataegus\_ploidy$ploidy,1,nchar(crataegus\_ploidy$ploidy)-1) # deleting x after ploidy level  crataegus\_ploidy[, c(3)] <- sapply(crataegus\_ploidy[, c(3)], as.numeric) # convert ploidy to numeric |

Boxplot for each species:



I did the same boxplots as in the ExploratoryDataAnalysis.R but added the feature “ploidy”. First we can see that the ploidy is different for Rhipidophylla comparing to Laevigata and Monogyna. The styles and ploidy in Laevigata are in average the same. In the other species the two features are different.

Histogram for ploidy in each species:

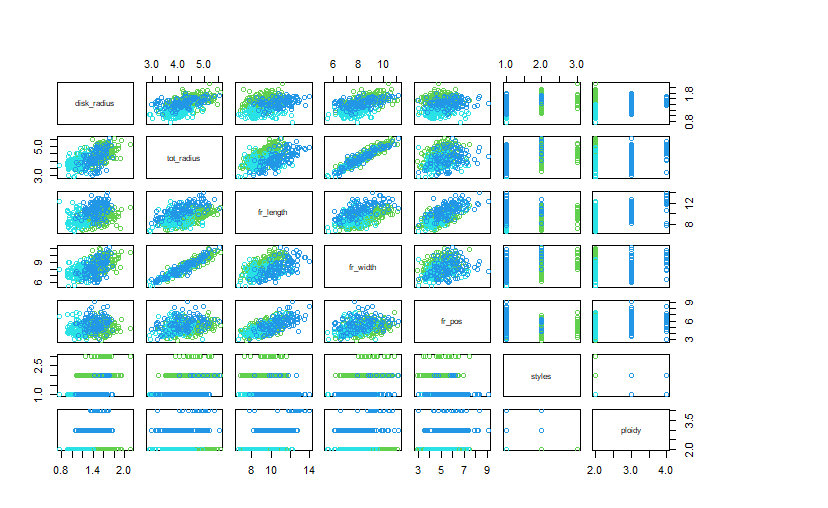


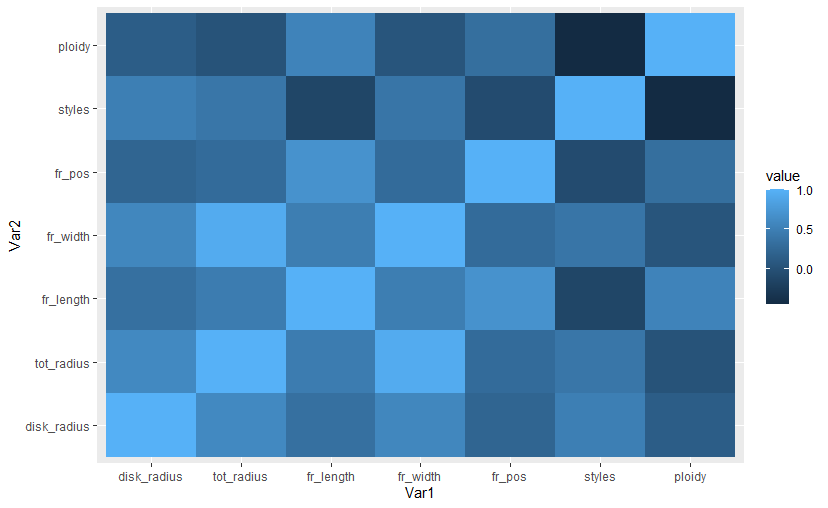
In Laevigata and Monogyna the ploidy is mostly 2. Rhipidophylla has either 3 or 4.

Correlation:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **disk\_radius** | **tot\_radius** | **fr\_length** | **fr\_width** | **fr\_pos** | **styles** | **ploidy** |
| **Disk\_radius** | 1 | 0.596 | 0.346 | 0.577 | 0.219 | 0.502 | 0.14 |
| **Tot\_radius** | 0.596 | 1 | 0.468 | 0.944 | 0.292 | 0.407 | 0.024 |
| **Fr\_length** | 0.346 | 0.468 | 1 | 0.485 | 0.691 | -0.125 | 0.544 |
| **Fr\_width** | 0.577 | 0.944 | 0.485 | 1 | 0.291 | 0.397 | 0.05 |
| **Fr\_pos** | 0.219 | 0.292 | 0.691 | 0.291 | 1 | -0.059 | 0.327 |
| **Styles** | 0.502 | 0.407 | -0.125 | 0.397 | -0.059 | 1 | -0.442 |
| **Ploidy** | 0.14 | 0.024 | 0.544 | 0.05 | 0.327 | -0.442 | 1 |

As we already know fruit width and total radius correlate strongly. Ploidy has no high correlation with other data. Only with fruit length it is above 0.5.

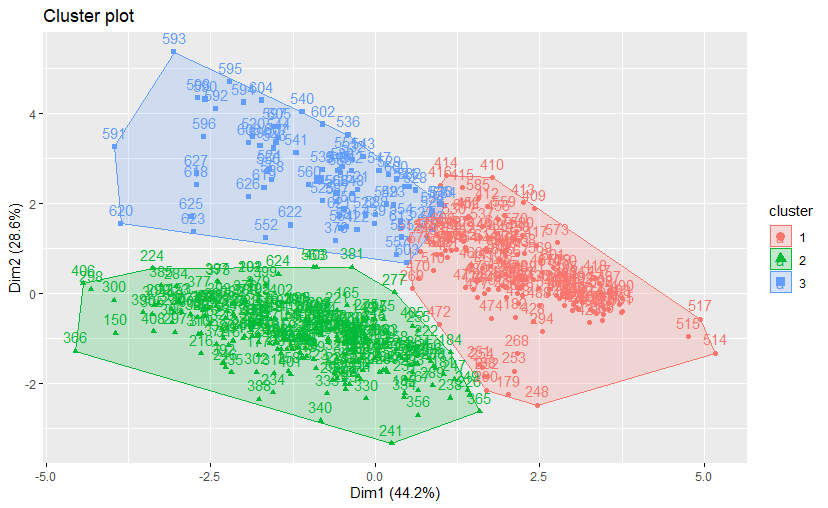




This is also visible in the two pictures.

KMeans:

First you have to scale the data.



Mean values for each feature and species:

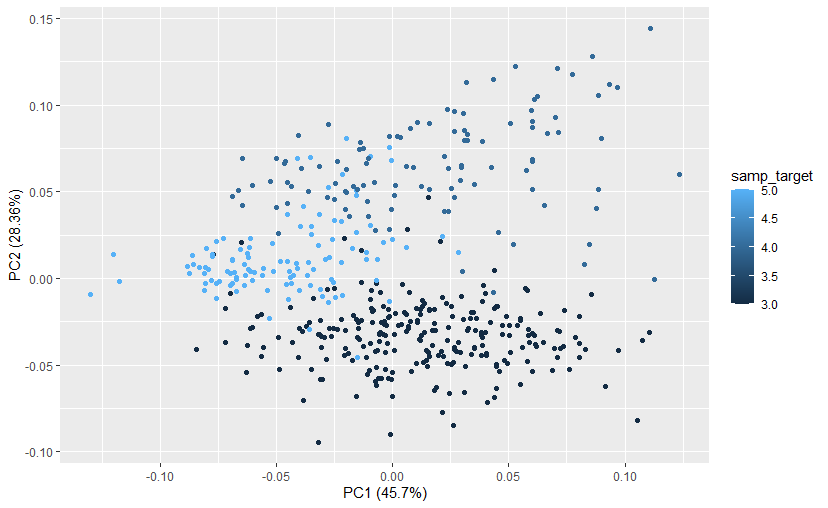
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| cluster | disk\_radius | tot\_radius | fr\_length | fr\_width | fr\_pos | styles | ploidy |
| 1 | 1.164 | 3.787 | 8.916 | 7.355 | 4.662 | 1.053 | 2.217 |
| 2 | 1.514 | 4.440 | 9.382 | 8.663 | 5.001 | 2.040 | 2.008 |
| 3 | 1.483 | 4.433 | 11.322 | 8.699 | 6.034 | 1.077 | 3.179 |

Remarkable differences:

Laevigata is less wide than the others (see disk radius, total radius and fruit length). Monogyna is larger and a higher ploidy level.

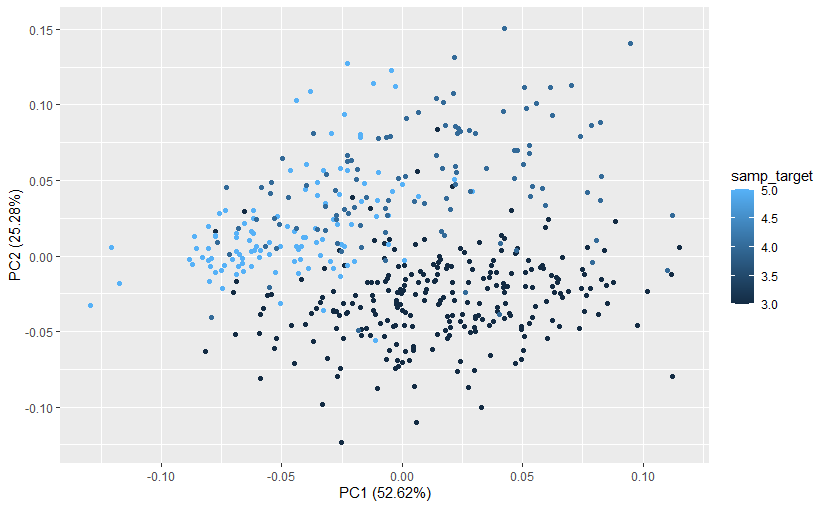
PCA:

1. PCA with all features (morphometric and ploidy data)



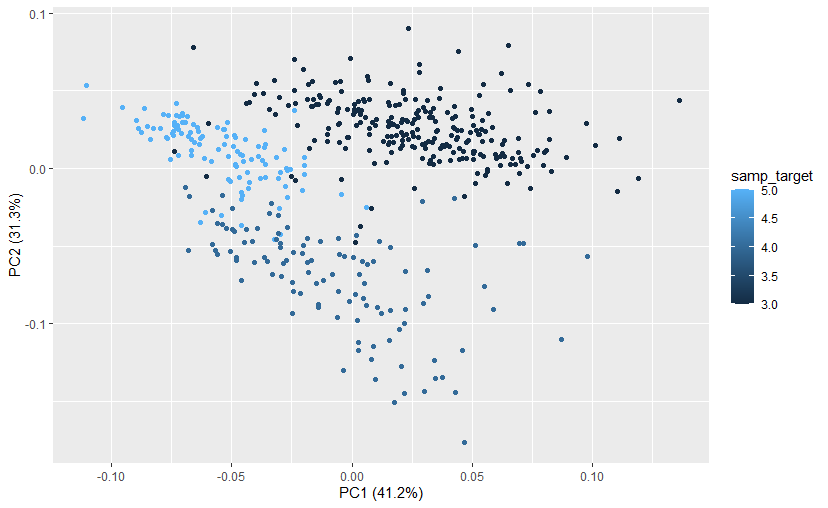
With PC1 and PC2 74.06 % can be explained. There a 3 cluster visible but not really clear. Laevigata differentiates quite strong from the others. Monogyna itself forms a clear cluster but it overlaps with Rhipidophylla as it scatters a lot.

1. All features



For comparison I again plotted the PCA with all features. No clear clusters are visible.

1. PCA with disk radius, fruit length, fruit position, styles and ploidy

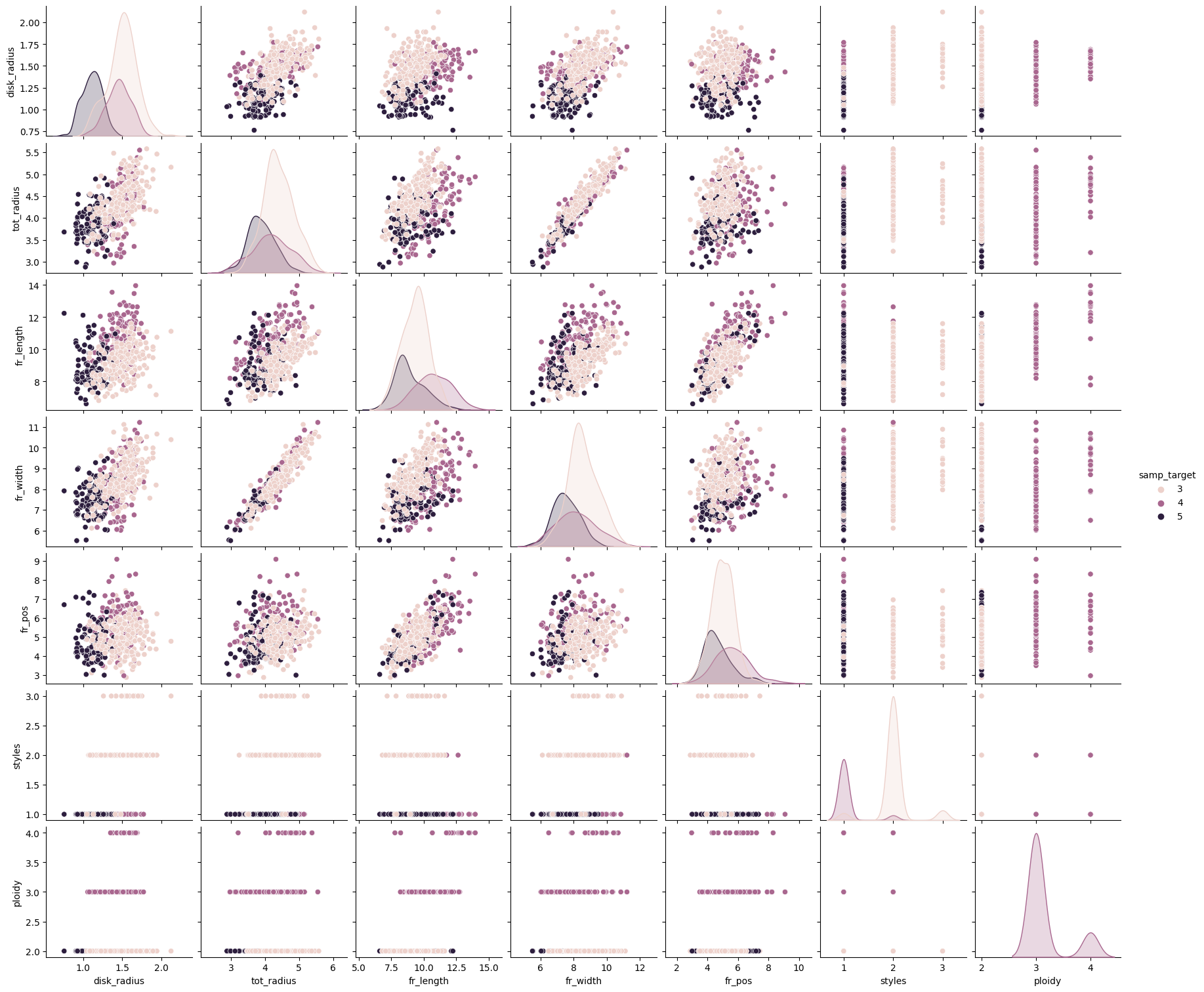


Good separation between the cluster is possible. 72.5 % can be explained with PC1 and PC2.

25.10.2021

Comparison of the 3 species:

Plan: include all data, seaborn package



I did the same as on the 4th of October. It seems not to include all ploidy data, or shows the wrong distribution. Most of the ploidy levels were 2 but there is no peak at two.

* Find logical mistake!