LR_9

Marc

10/22/2024

add select in Final project instructions

1. Let's have an exercise in class: play with the 'iris' database
2. Let's think and talk about our final projects

Example of structure for the final project

Introduction

- Three species of flowers (setosa, versicolor, virginica) are currently at the heart of a debate. Some scholars claim that they belong to the same group since they have similar petals and sepals size while others argue that their sepals and petals differ so much that they should categorized as different species. Add a few references if possible...
- The following figure shows the location of petals and sepals on a flower.

Research question:

- Can we distinguish setosa, versicolor, virginica based on Sepal length/width and Petal length/width.

Loading the data and basic visualization

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

The repository of the data can be found at http://archive.ics.uci.edu/ml/datasets/Iris

In the following code, we load the required packages and the data.

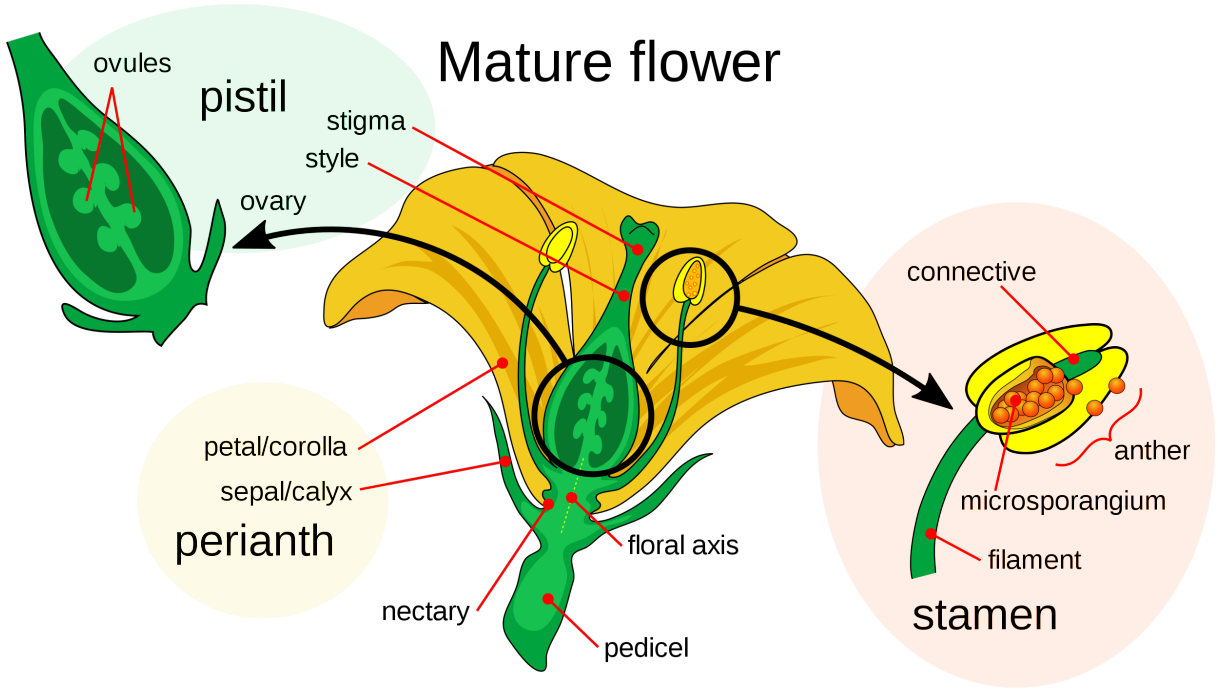


Figure 1: Sepal and Petal

```
#load the packages and suppress the annoying messages at the beginning
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(ggfortify))
suppressPackageStartupMessages(library(party))
# load the data
data(iris)
```

Then, we have a quick look at the data to check that it has been loaded correctly. We see that the data has five columns, sepal length/width, petal length/width and species.

```
data(iris)
str(iris)
```

```
## '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 ...
```

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

```
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

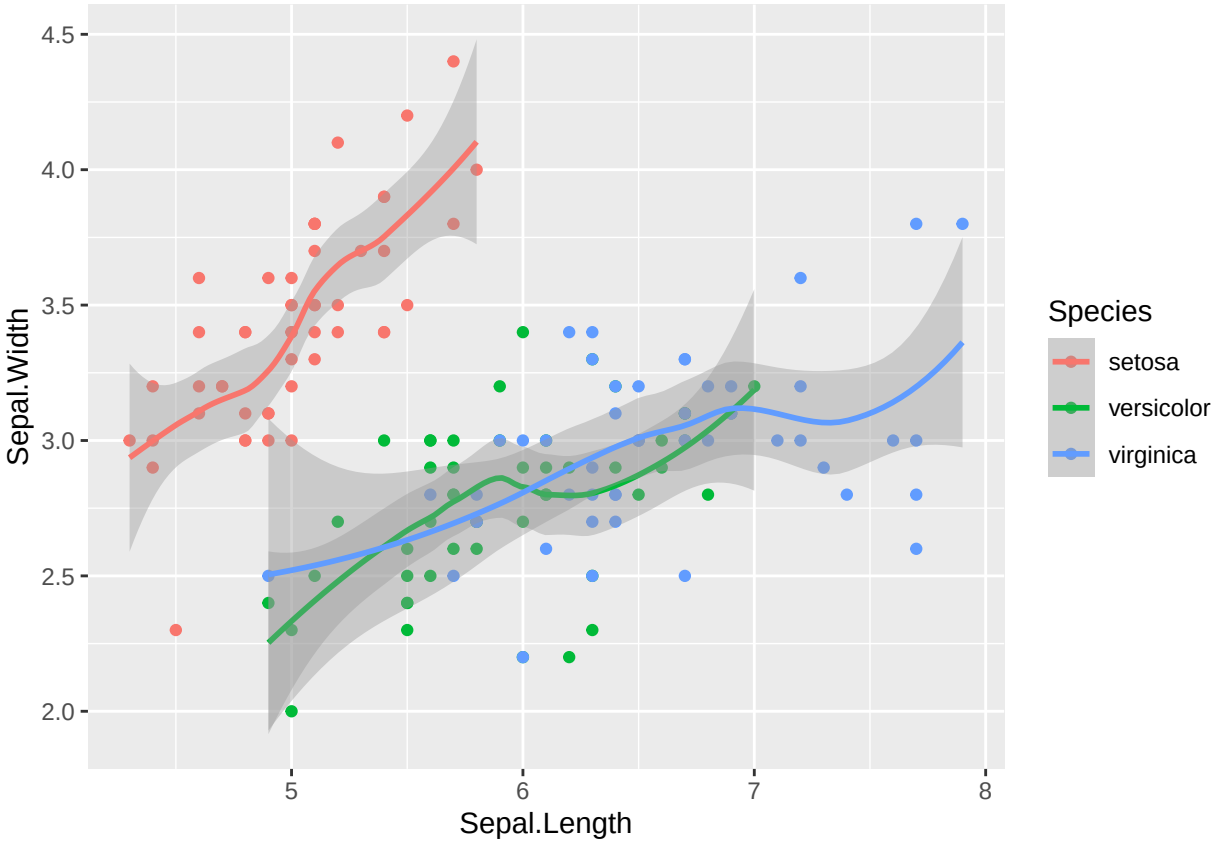
Before we do any analysis, let's have a visualization of the data. For instance, we see that petal length and width are positively correlated. Sepal length and width seem to be positively correlated too, but the effect is stronger in setosa.

```
iris %>%
  ggplot(aes(x=Petal.Length,y=Petal.Width, colour = Species)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

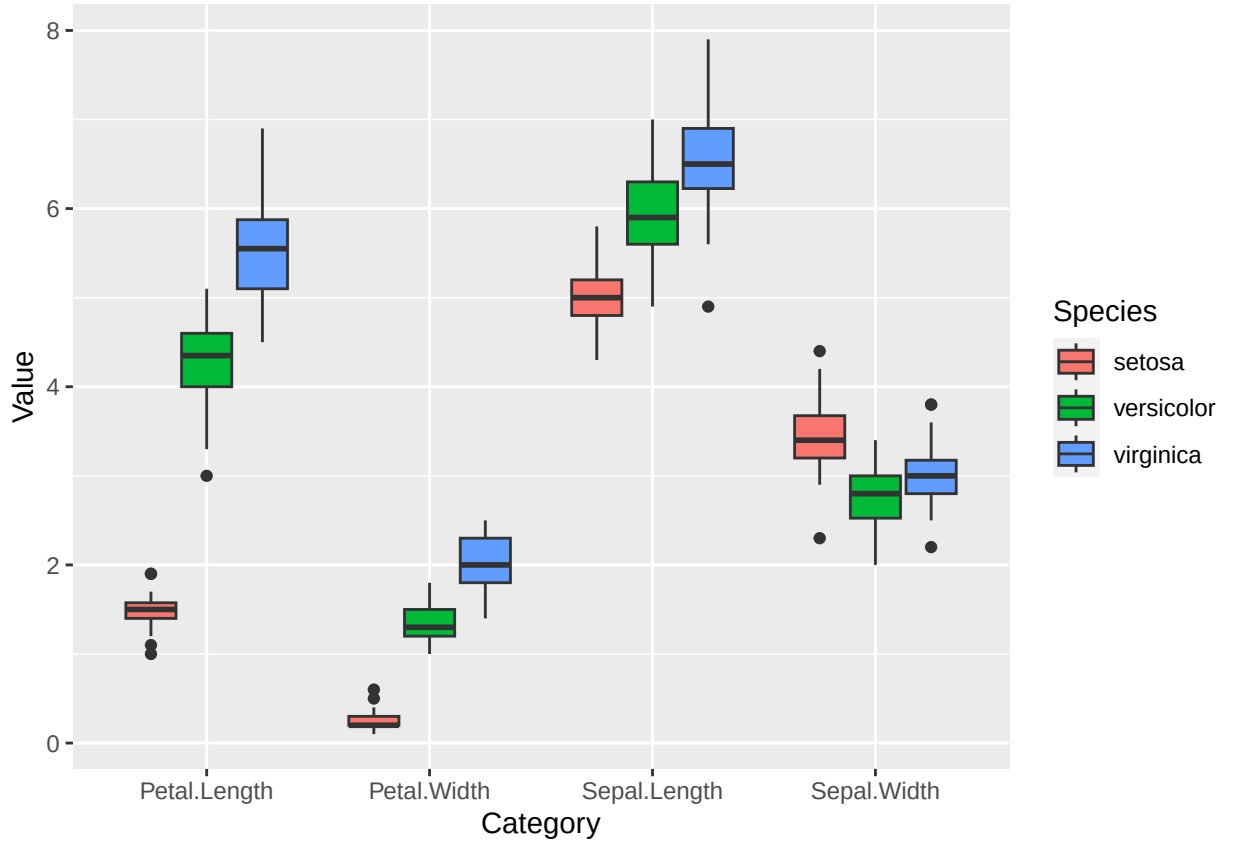

```
iris %>%
  ggplot(aes(x=Sepal.Length,y=Sepal.Width, colour = Species)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



When distinguishing the three species in the graph below. It seems that virginica has a larger size for almost each feature, whereas setosa has the smallest values, except for sepal width. Versicolor is located in the middle for almost every feature. From this visualization, we can expect that the three species are indeed quite different according to sepals and petals length/width. However, further investigation is needed to confirm this point.

```
iris %>%
  gather("Category","Value",-c(Species)) %>%
  ggplot(aes(x=Category,y=Value,fill=Species)) +
  geom_boxplot()
```



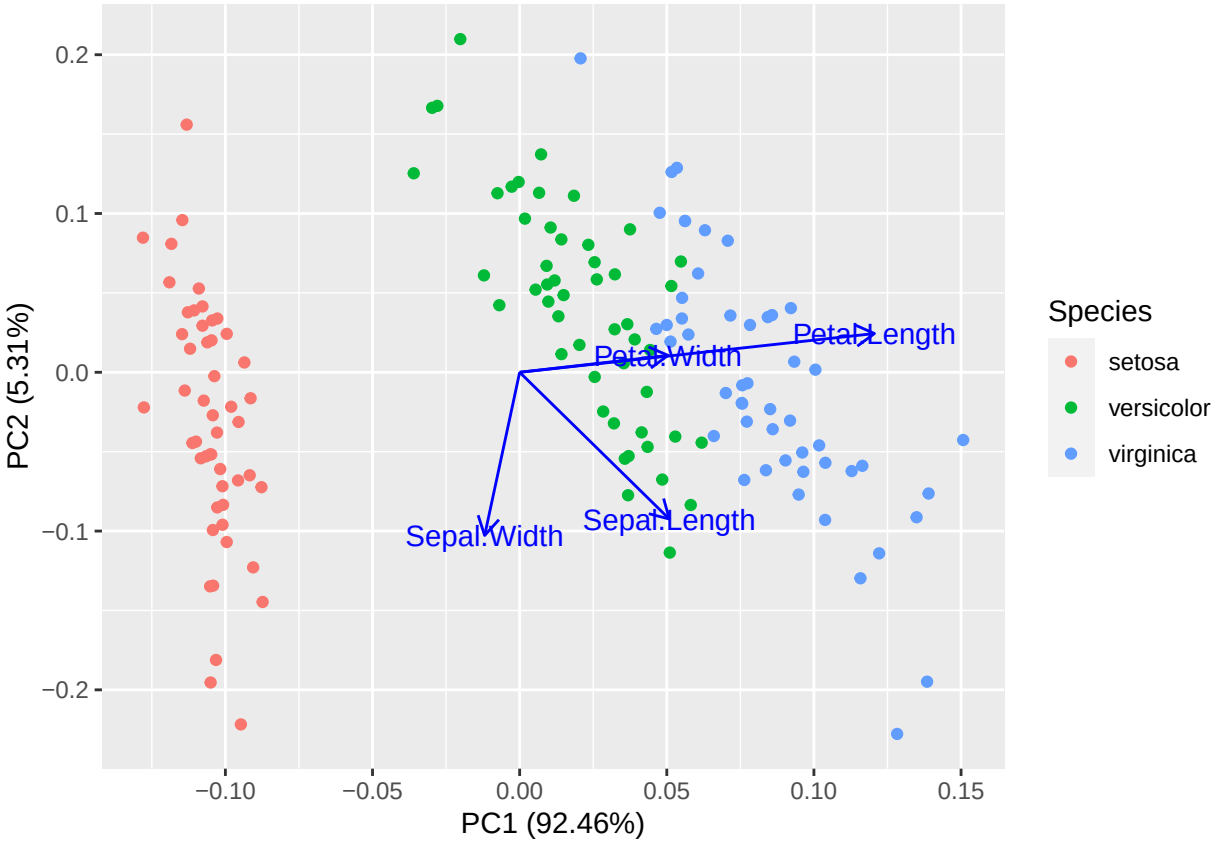
Analysis

In this section, we carry out PCA and classification with decision trees to check if information on petals and sepals can actually help to differentiate the three species.

First, we run the PCA. PCA is a method of dimensionality reduction. In the previous plots with ggplot, we were only able to compare two variables (two columns) of the data at a time. PCA compresses the information of all the columns and allows us to visualize how similar/different are the flowers in our data. In the plot below, each point represents a surveyed flower. The closer the two points the more similar the two flowers are in terms of sepal and petal length/width. The PCA indicates that virginica tends to have bigger petal length and sepal length, since most of the blue points are following the arrows of these two features. Versicolor is located at the starting point of the arrows, which means that most flowers of this species are close to the average values. Finally, setosa is expected to have small petal length and width since most of the red points are found in the opposite direction of the arrows. These observations match with our findings in the visualization section.

```
iris %>%
  # remove information of species to avoid overclassification
  select(-Species) %>%
  # run PCA
  prcomp() %>%
  autoplot(data = iris,
           # add colour to the points based on species
           colour = 'Species',
           # add arrows showing the effect of each variable
```

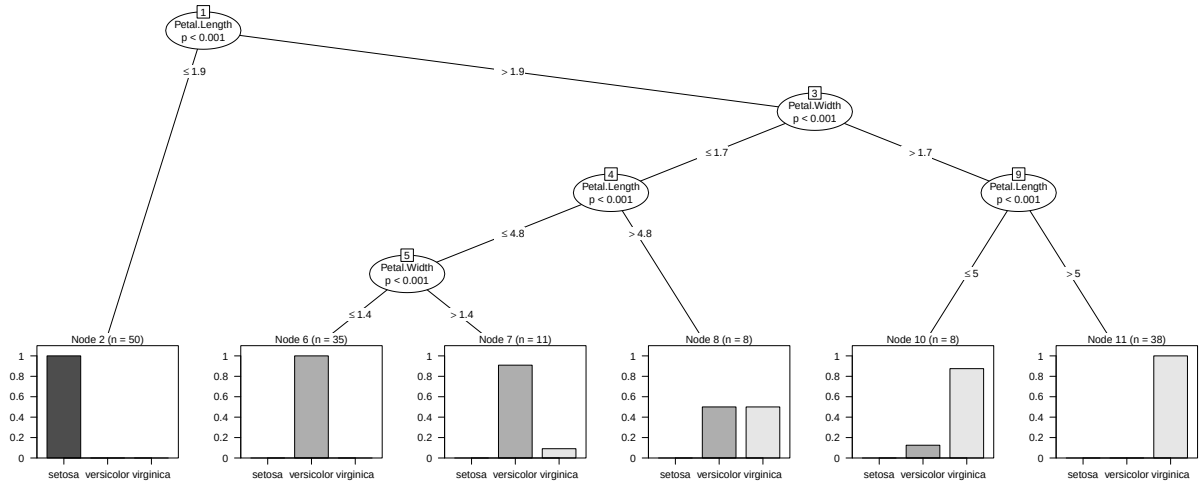
```
          loadings = TRUE,
          loadings.colour = "blue",
          # add the names of the arrows
          loadings.label = TRUE,
          loadings.label.colour = "blue")
```



If you want to compare the clusters with the original families

Still, we need to show if these tendencies are actually significant. To do so, we run a classification task based on conditional inference trees. The results show that petal length is mostly sufficient to identify the three species. Petal length below 1.9 cm (Node 1 to Node 2) indicates setosa, whereas above 1.9 (Node 1 to Node 3) and with petal width above 1.7 (Node 3 to Node 9) mostly refers to virginica.

```
ctree(Species ~. ,
      data = iris,
      controls = ctree_control(testtype = "MonteCarlo")) -> flower_tree
plot(flower_tree, gp = gpar(fontsize = 8))
```



Yet, we need to measure measure how good is the decision tree shown above. To do so, we measure its accuracy, i.e., how precise can we get when using this tree to predict the species of the flowers in our dataset. We compare this accuracy with the majority baseline, that is to say, what could we get by guessing randomly that every flower belongs to the species with the most data points. In our case, the baseline is 33.33% (50/150) since every species had 30 samples.

```
# we extract the responses from the tree
pred_flower<-predict(flower_tree)
# and use them to predict the value of "To" in every row
accuracy_table<-table(pred_flower,iris$Species)
# the columns are the right output, the rows are the predictions
accuracy_table
```

```
##
## pred_flower  setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         49         5
##   virginica       0          1        45
```

```
# finally, we convert the table into a percentage
sum(diag(accuracy_table))/sum(accuracy_table)
```

```
## [1] 0.96
```

```
# for instance, if the accuracy is 75%, it means that on 100 random cases, the model can guess correctl
# But how good or bad is that? let's have a look at the Majority baseline
max( c(sum(accuracy_table[,1]),sum(accuracy_table[,2]))) /sum(accuracy_table)
```

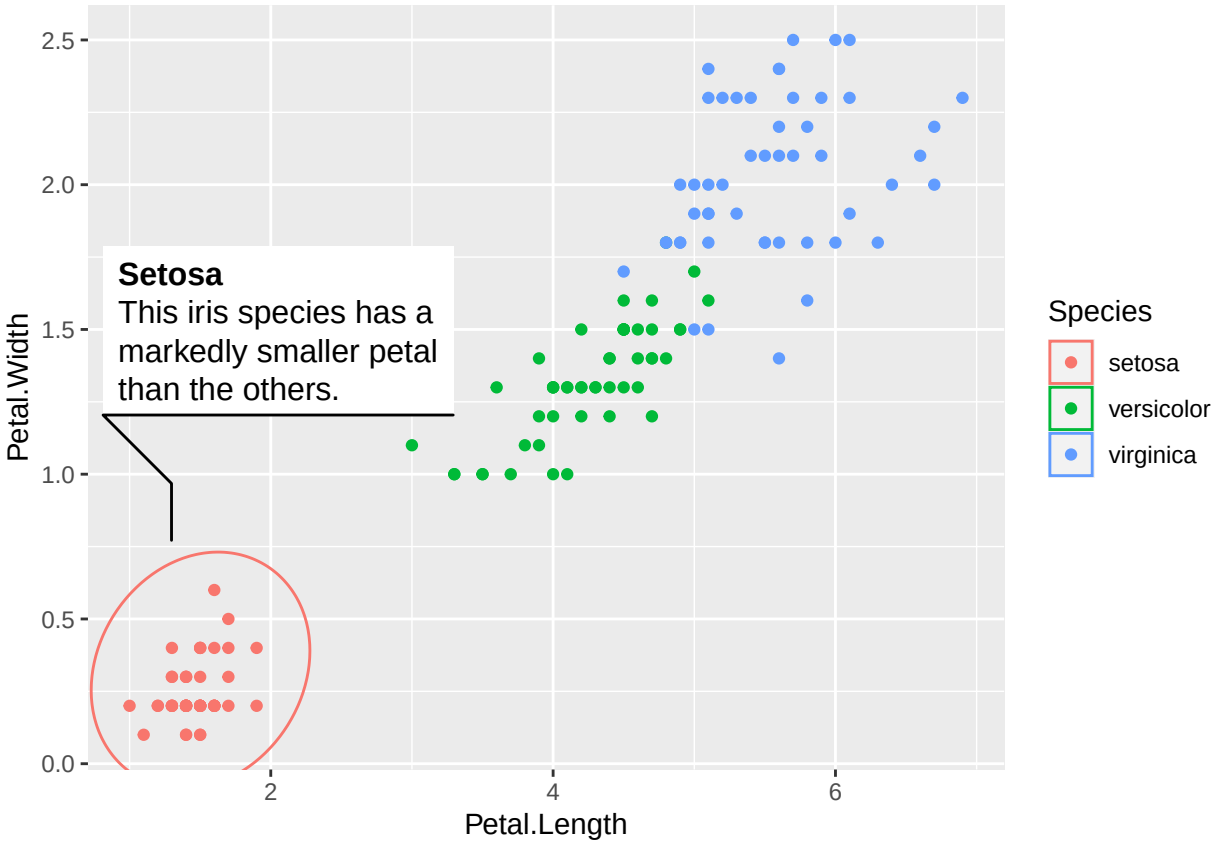
```
## [1] 0.3333333
```

The results show that our tree can predict 96% of the data correctly, which is far above the baseline of 33.33%. We can thus infer that information on sepal and petal length/width is useful for distinguishing the three species.

Summary

We can answer our research question as follows: We can distinguish setosa, versicolor, virginica based on Sepal length/width and Petal length/width. As shown in the graph below, the petals of setosa are generally smaller than the two other species, whereas virginica has the largest petal length and width.

```
suppressPackageStartupMessages(library(ggforce))
# specify the content to be added on the plot
desc <- 'This iris species has a markedly smaller petal than the others.'
# make the plot
ggplot(iris,
       aes(x=Petal.Length,y=Petal.Width,colour=Species )) +
  geom_mark_ellipse(aes(filter = Species == 'setosa', # only add the ellipse for setosa
                        label = 'Setosa',
                        # link to the content to be added on the plot
                        description = desc)) +
  geom_point()
```



References

If you have any reference, put it here.