

Applied Statistics

Iris Data Set



**Data exploration
Visualization and
Analysis using R**



Fisher's iris data set

- 150 rows
- 3 species
- 4 columns



Setosa



Virginica



Versicolor



Data exploration

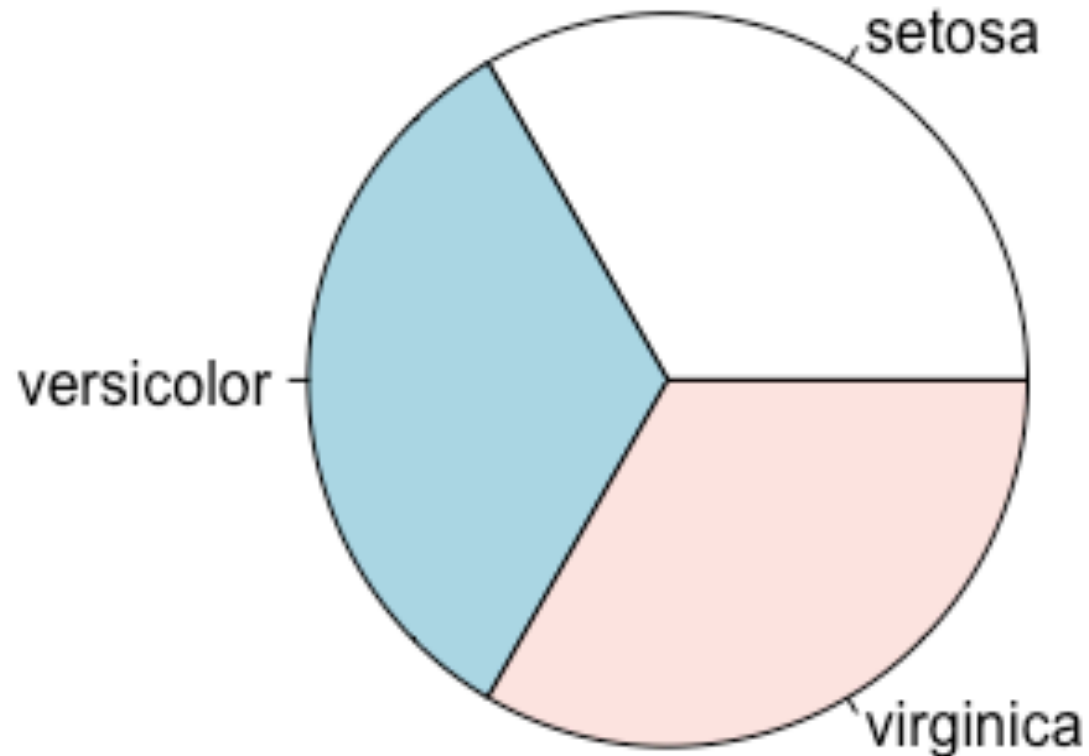
- 4 numerical data and 1 categorical data

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
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	



Graphic data representation

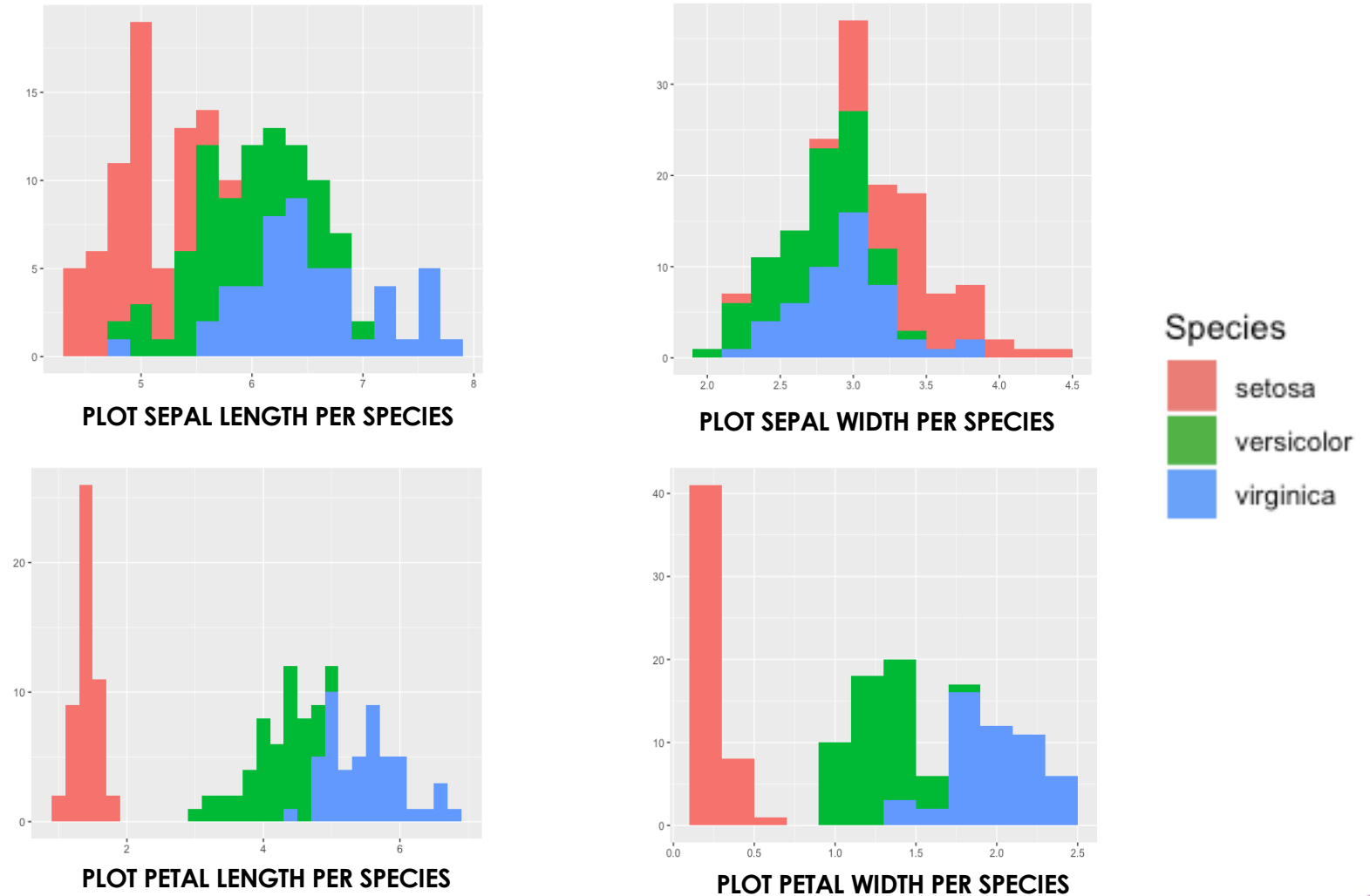
- Species repartition piechart





Graphic data representation

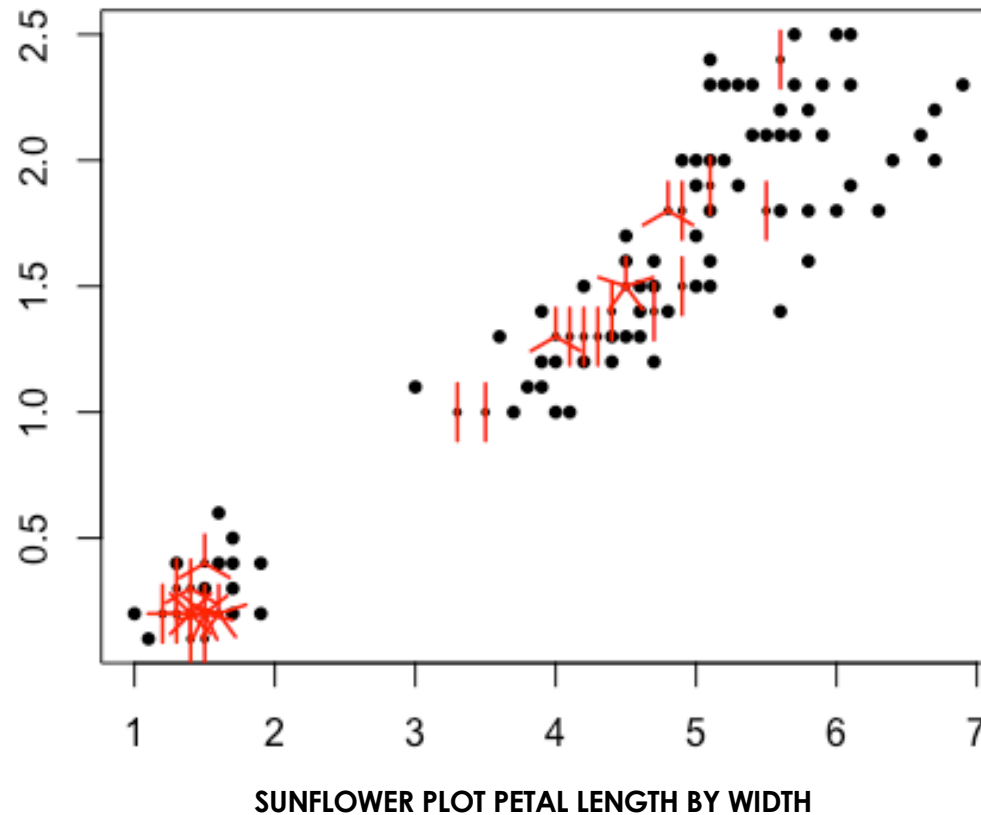
- Species repartition depending on their attributes





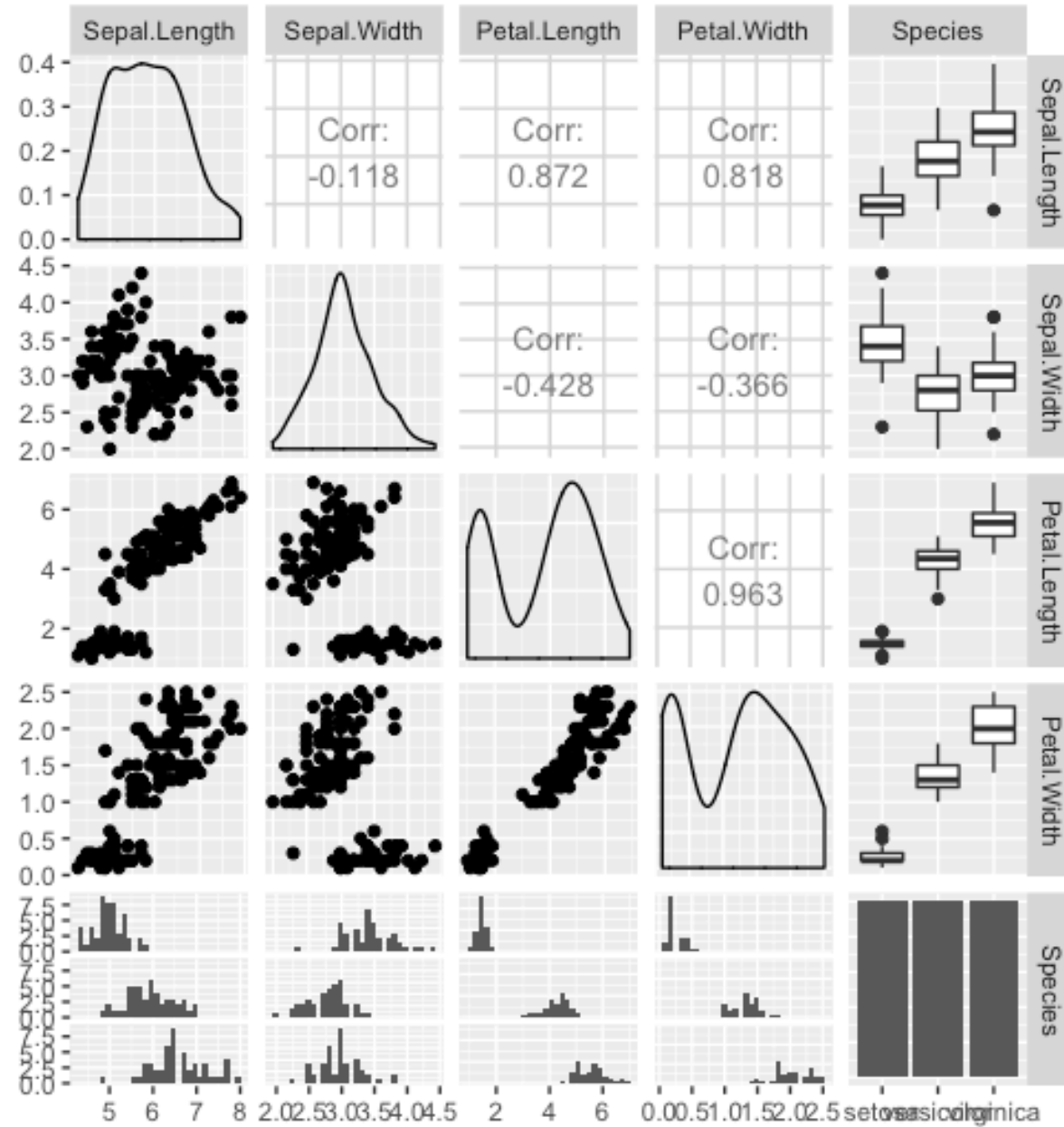
Graphic data representation

- Sunflower






- Pair plot





Data prediction and classification

- Chi-square on each variable

	Sepal.Width	Petal.Length	Petal.Width	Species
Sepal.Length	0.17	3.7e-17	3.5e-02	6.6e-09
Sepal.Width	-	4.2e-02	1.2e-04	6.0e-05
Petal.Length	-	-	5.1e-09	1.1e-21
Petal.Width	-	-	-	2.1e-35



Data prediction and classification

- Value splitting
 - 120 test
 - 30 train

```
120 samples
4 predictor
3 classes: 'setosa', 'versicolor', 'virginica'
```

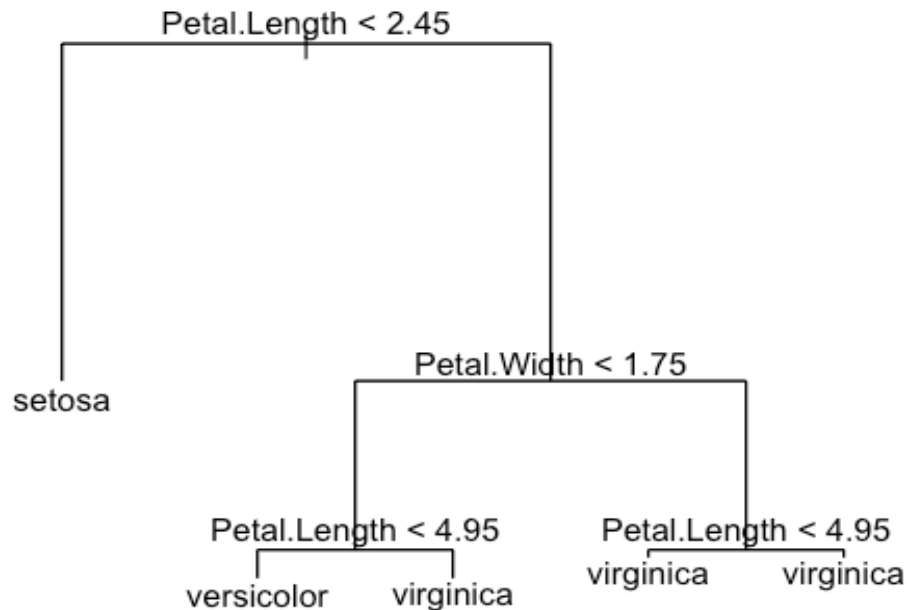
- Predictions

	Reference		
Prediction	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	9	1
virginica	0	1	9

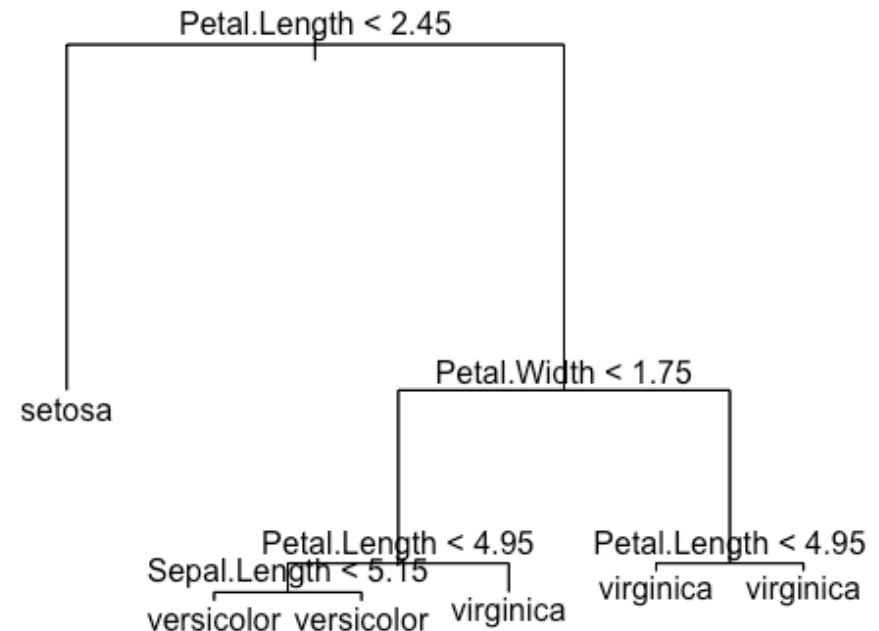


Data prediction and classification

- Classification trees



CLASSIFICATION TREE USING PETAL LENGTH AND PETAL WIDTH



CLASSIFICATION TREE OF SPECIES WITH ALL FEATURES



Models validation

- Linear Discriminant Analysis

	Class: setosa	Class: versicolor	Class: virginica
Sensitivity	1.0000	0.9000	0.9000
Specificity	1.0000	0.9500	0.9500
Pos Pred Value	1.0000	0.9000	0.9000
Neg Pred Value	1.0000	0.9500	0.9500
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.3333	0.3000	0.3000
Detection Prevalence	0.3333	0.3333	0.3333
Balanced Accuracy	1.0000	0.9250	0.9250



Models validation

- Classification trees

Classification tree:

```
tree(formula = Species ~ Sepal.Width + Sepal.Length + Petal.Length +  
      Petal.Width, data = iris)
```

Variables actually used in tree construction:

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

Number of terminal nodes: 6

Residual mean deviance: 0.1253 = 18.05 / 144

Misclassification error rate: 0.02667 = 4 / 150



Conclusion

- Clean data set
 - Good data repartition
- Good classification models