SILESIAN UNIVERSITY OF TECHNOLOGY

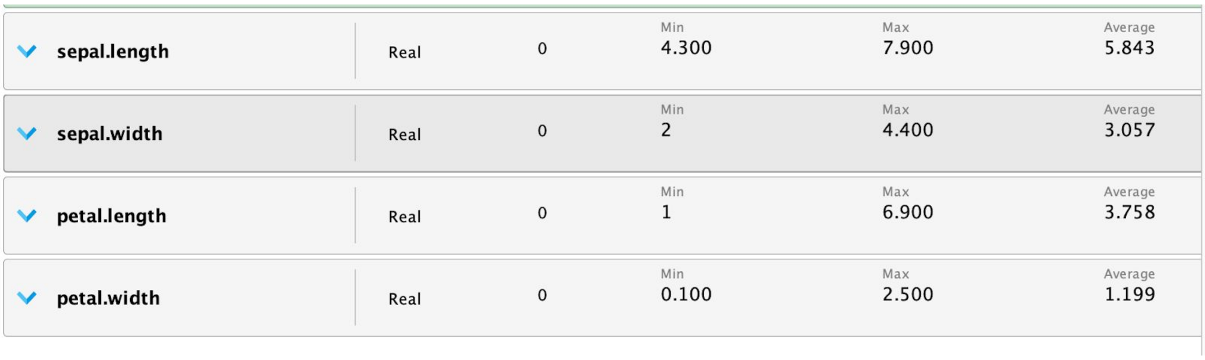
Faculty of Automatic Control, Electronics and Computer Science

DATA SCIENCE
Semester II

DATA MINING IN PRACTICE

.

1. Dataset statistics



			Min	Max	Average
sepal.length	Real	0	4.300	7.900	5.843
sepal.width	Real	0	2	4.400	3.057
petal.length	Real	0	1	6.900	3.758
petal.width	Real	0	0.100	2.500	1.199

a) How many objects does the data set contain? 150 objects.

b) How many attributes do these objects describe, and what do the individual objects mean?

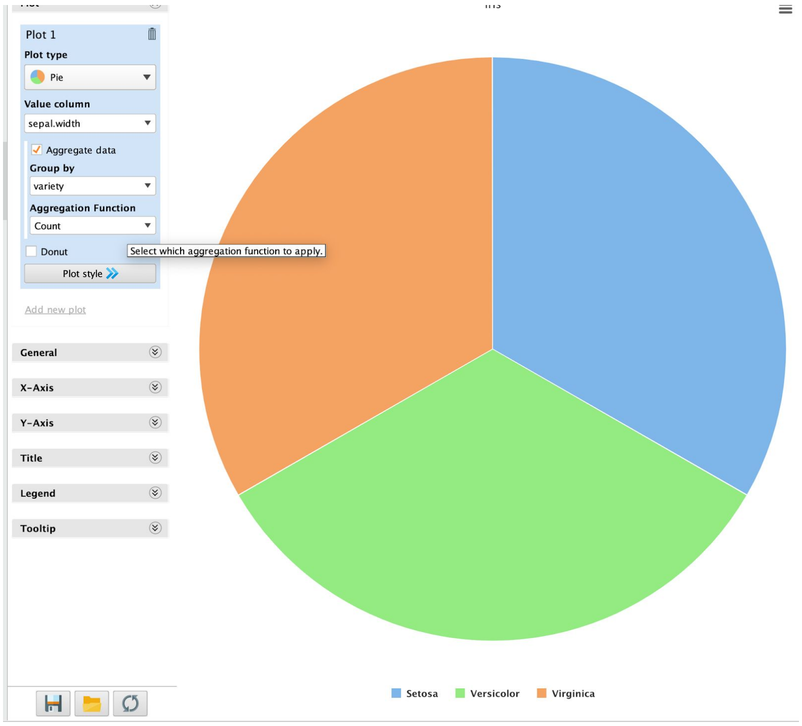
The dataset is containing 5 attributes:

- sepal length
- sepal width
- petal length
- petal width
- class (label) - setosa, versicolor and virginica

a) How many decision classes can be distinguished in this set - 3 decision classes:

- setosa,
- versicolor
- virginica

2. Data visualisation



Plot 1

Plot type: Pie

Value column: sepal.width

☑ Aggregate data

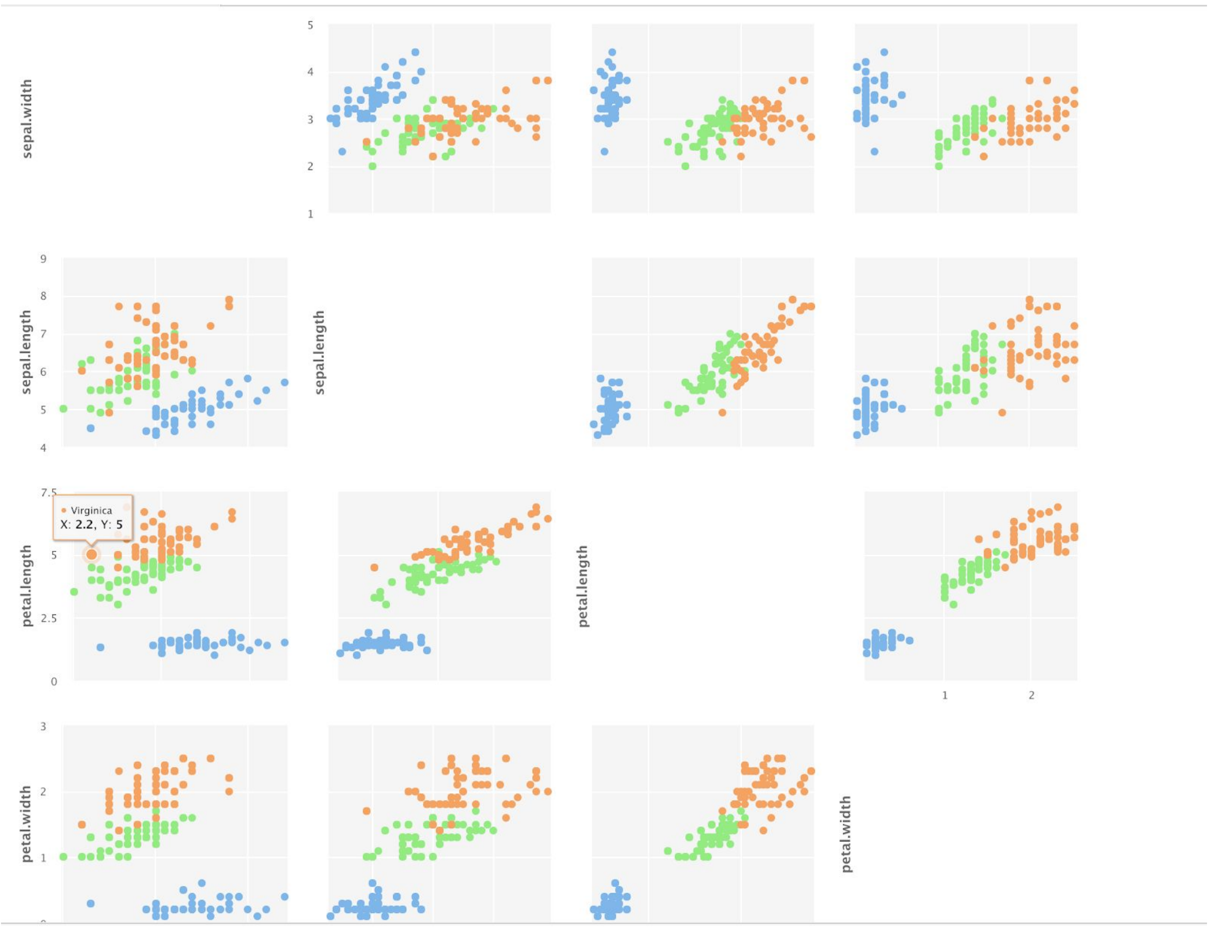
Group by: variety

Aggregation Function: Count

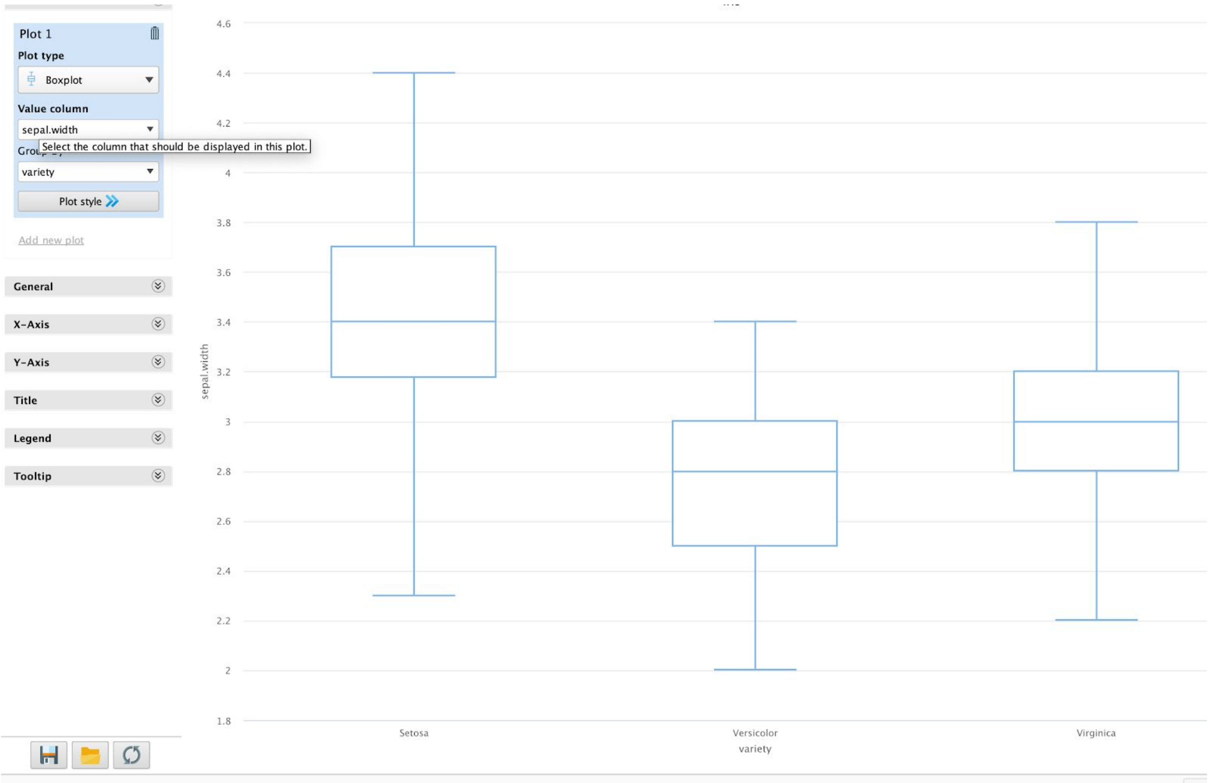
□ Donut

Select which aggregation function to apply.

This charts demonstrates that the dataset is balanced (equal number of records for each class)



Distribution of each feature based on class



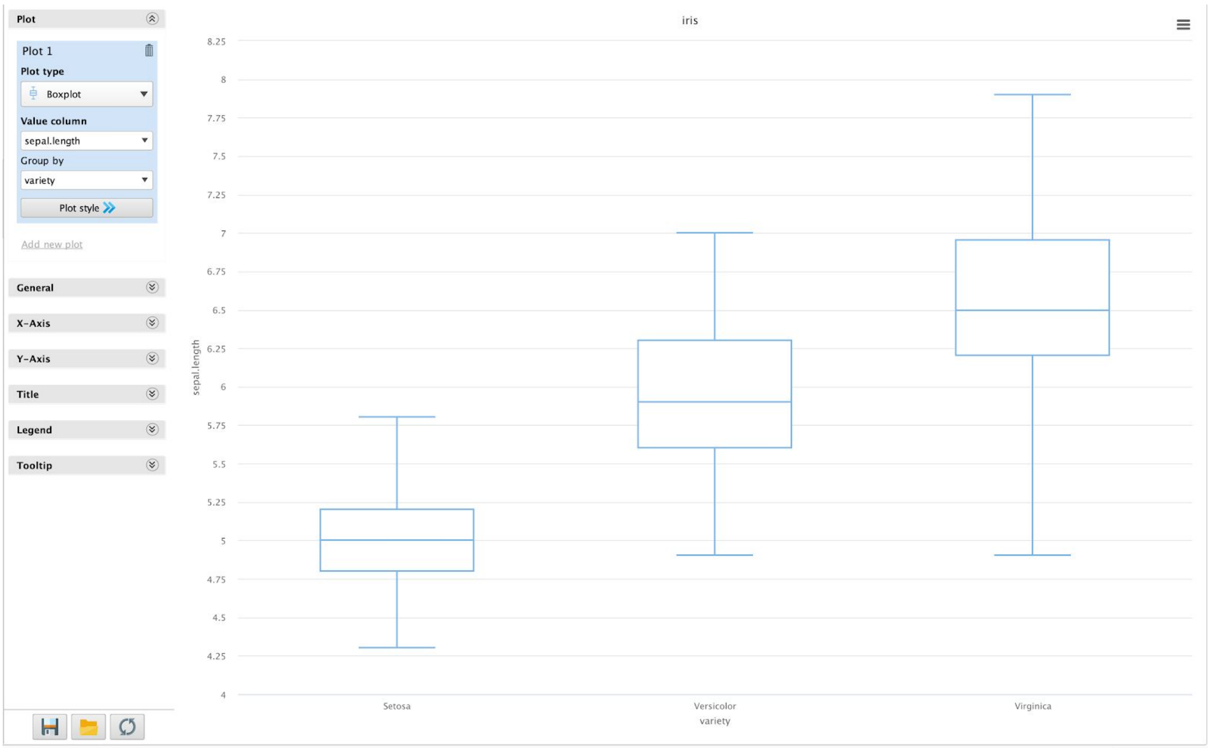
Plot 1

Plot type: Boxplot

Value column: sepal.width

Select the column that should be displayed in this plot.

Group by: variety



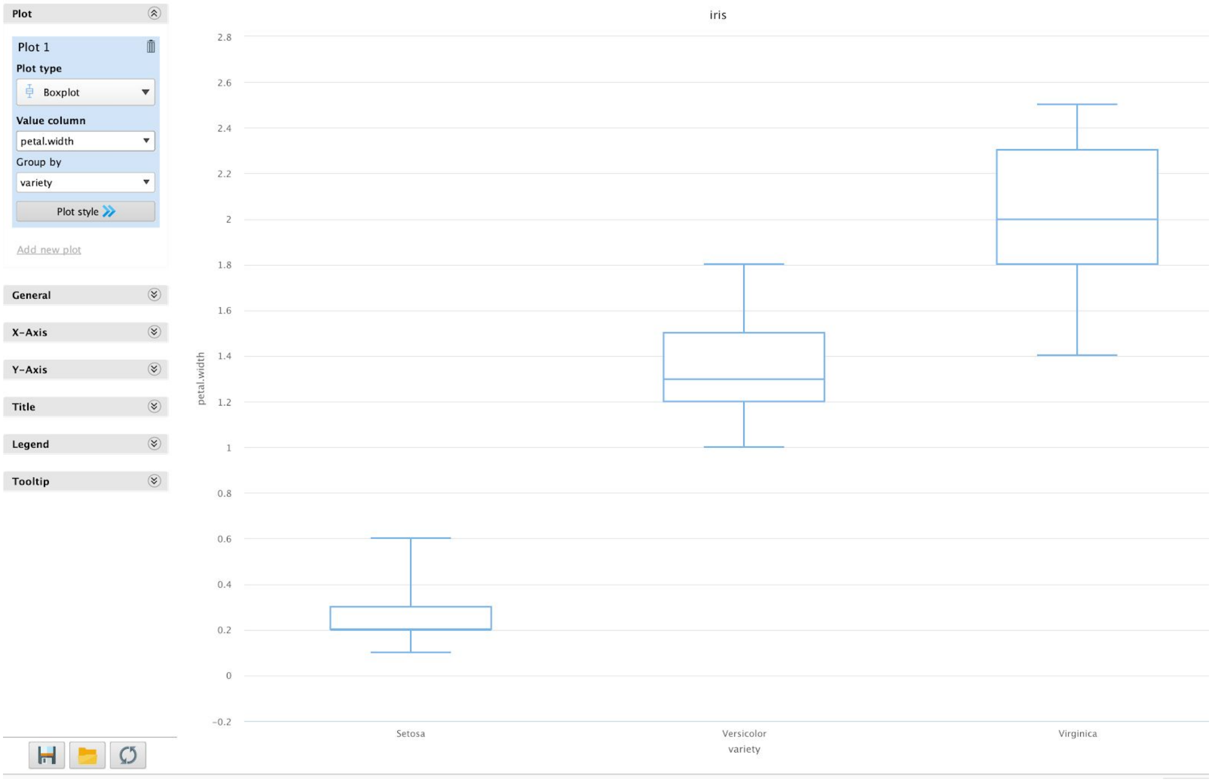
Plot

Plot 1

Plot type: Boxplot

Value column: sepal.length

Group by: variety



Plot

Plot 1

Plot type: Boxplot

Value column: petal.width

Group by: variety

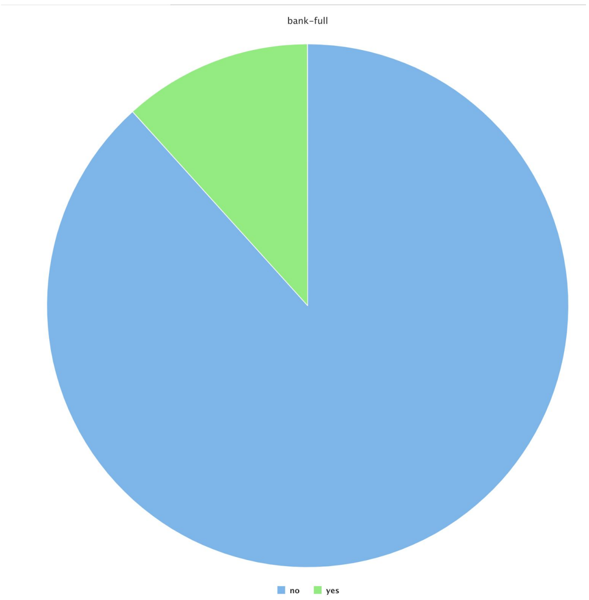
3. Second dataset

3.1. How many objects are in the data set - 41 188

3.2. How many attributes describe these objects - 20

3.3. How many decision classes are there and what is the distribution of objects in each decision class? - 2

3.4. Is the data set balanced?



No it’s not. There many more records in class “no”.

4. Missing values

4.1. How many missing values are there and for which attribute (write down the column name).

- there are 3 missing values for duration column

4.2. • What are the identifiers of the objects for which the value is missing?

- duration

4.3. • Which function was used to fill in the missing values?

- average

4.4. • Enter the numeric value that was used to replace the values of missing attributes.

- 258

5. Data reduction

5.1. What is the distribution of objects between classes for non-reduced data? Write down the ratio.

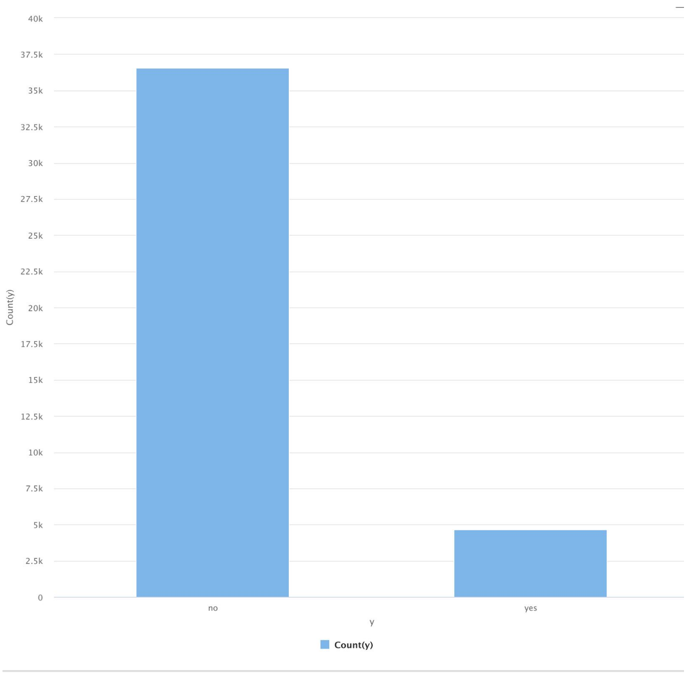
- normal distribution

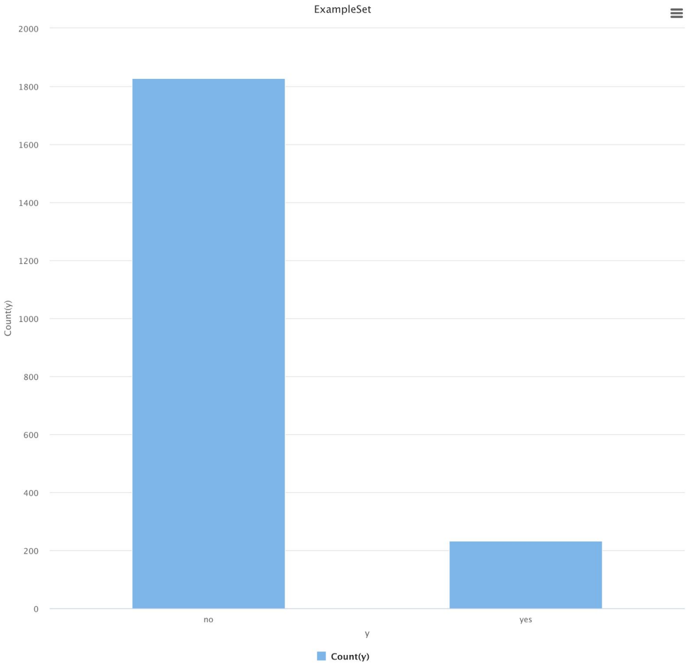
5.2. How many objects are left in the dataset after reduction?

- 2059

5.3. What is the distribution of objects between classes for data after reduction?

- normal distribution





6. Outlier detection

6.1. What is the name of the additional attribute that was created after the outlier search process

- outlier

6.2. Check how many outliers were found and compare the results with the next section. Is the number of objects found different? If yes, why? If no, why?

- 10

6.3. What function was used to determine the distance between objects?

- euclidean distance

6.4. Determine Euclidean distance between objects 1 and 3 and 5 and 6 described in the table below. Which pair of objects is more similar? Write down the results of distance calculations and provide the formula for the Euclidean distance.

Formula for calculating euclidean distance value

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

Results:

- D13 = 0,5099
- D56 = 0,6164

Points 1 and 3 are more similar than 5 and 6, because the distance between them is smaller.

7. Data Normalization

7.1. What was the minimum and maximum value of the euribor3m attribute before normalization

- min: 0.635 max : 4.970

7.2. What was the minimum and maximum value of the euribor3m attribute after normalization

- min: 0 max: 1

7.3. Which attributes were normalized after Changing the operator parameter attribute filter type to all

- all numerical attributes were changed: age, duration, campaign, pday, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor.3m, nr.employed

8. Feature reduction

8.1. How many attributes the result set contains if the selection criterion is numeric attributes

- ten attributes, all numerical

8.2. How many attributes does the result set contain if the selection criterion is random

- four attributes

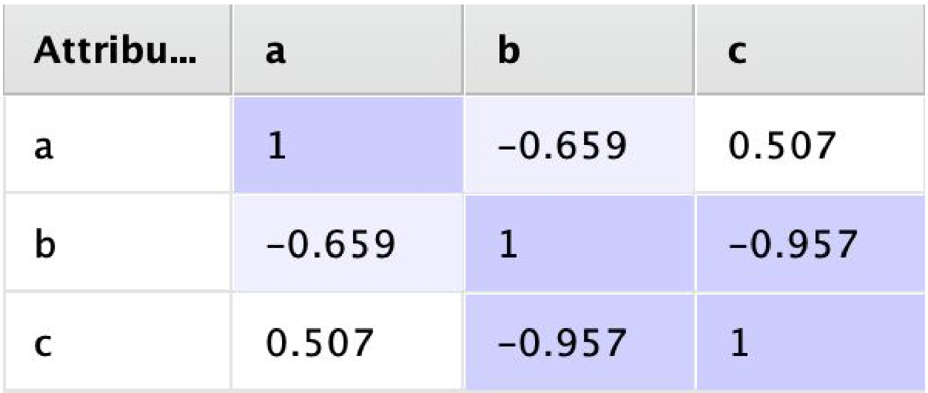
8.3. How many attributes the result set contains if the selection criterion is correlation. Describe how the Remove Correlated Attributes operator works and why we use it for data analysis.

- two were removed

8.4. Check which attributes were removed by Remove Correlated Attributes operator and find out their meaning. What do you think about the results of using this operator?

- euribor3m, nr.employed,

9. Correlation



Attribu...	a	b	c
a	1	-0.659	0.507
b	-0.659	1	-0.957
c	0.507	-0.957	1

10. Division of the set into test and training. Saving the resulting files to a csv file

10.1. How many objects contains the result set of 30%?

- **618**

10.2. How many objects contains the 70% result set?

- **1441**

10.3. What is the distribution of objects between classes in the test and training set - enter the number of objects belonging to individual classes in both sets?

- 548 / 70
- 133 /1308

10.4. Has the original distribution of objects between classes been preserved in both collections? If not, which parameter of the SplitData operator should be used (and what value should it have), so that the distribution objects between classes is as in the original data set.

- stratified sampling

10.5. What is the default column separator for the Write CSV operator?

- ‘;’