Questions in Data Understanding

Goal

Gain insight in your data

1. with respect to your project goals
2. and general

Find answers to the questions

1. What kind of attributes do we have?
2. How is the data quality?
3. Does a visualization helps?
4. Are attributes correlated?
5. What about outliers?
6. How are missing values handled?

Attribute understanding

We (often) assume that the data set is provided in the form of a simple table.

	attribute_1	...	attribute_m
record_1			
⋮			
record_n			

- The rows of the table are called **instances**, **records** or **data objects**.
- The columns of the table are called **attributes**, **features** or **variables**.

Types of attributes

categorical (nominal): finite domain
The values of a categorical attribute are often called classes or categories.
Examples: {female,male}, {ordered,sent,received}

ordinal: finite domain with a linear ordering on the domain.
Examples: {B.Sc.,M.Sc.,Ph.D.}

numerical: values are numbers.

discrete: categorical attribute or numerical attribute whose domain is a subset of the integer number.

continuous: numerical attribute with values in the real numbers or in an interval

Data quality

Low data quality makes it impossible to trust analysis results: "Garbage in, garbage out"

Accuracy: Closeness between the value in the data and the true value.

- Reason of low accuracy of numerical attributes: noisy measurements, limited precision, wrong measurements, transposition of digits (when entered manually).
- Reason of low accuracy of categorical attributes: erroneous entries, typos.

Data quality

Syntactic accuracy : Entry is not in the domain.
Examples: fmale in gender, text in numerical attributes, ...
Can be checked quite easy.

Semantic accuracy : Entry is in the domain but not correct.
Example: John Smith is female
Needs more information to be checked (e.g. "business rules").

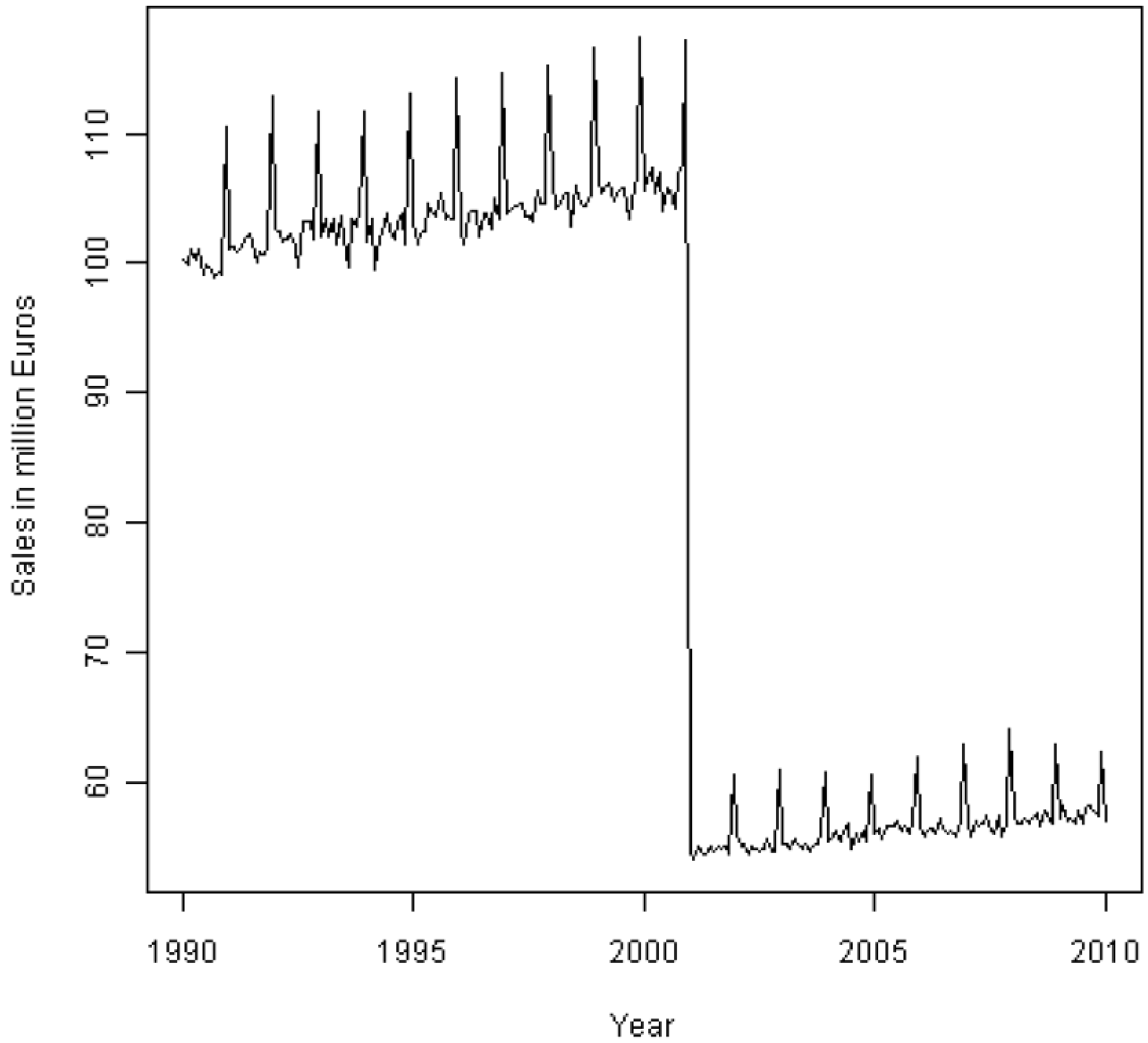
Completeness : is violated if an entry is not correct although it belongs to the domain of the attribute.
Example: Complete records are missing, the data is biased (A bank has rejected customers with low income.)

Unbalanced data: The data set might be biased extremely to one type of records.
Example: Defective goods are a very small fraction of all.

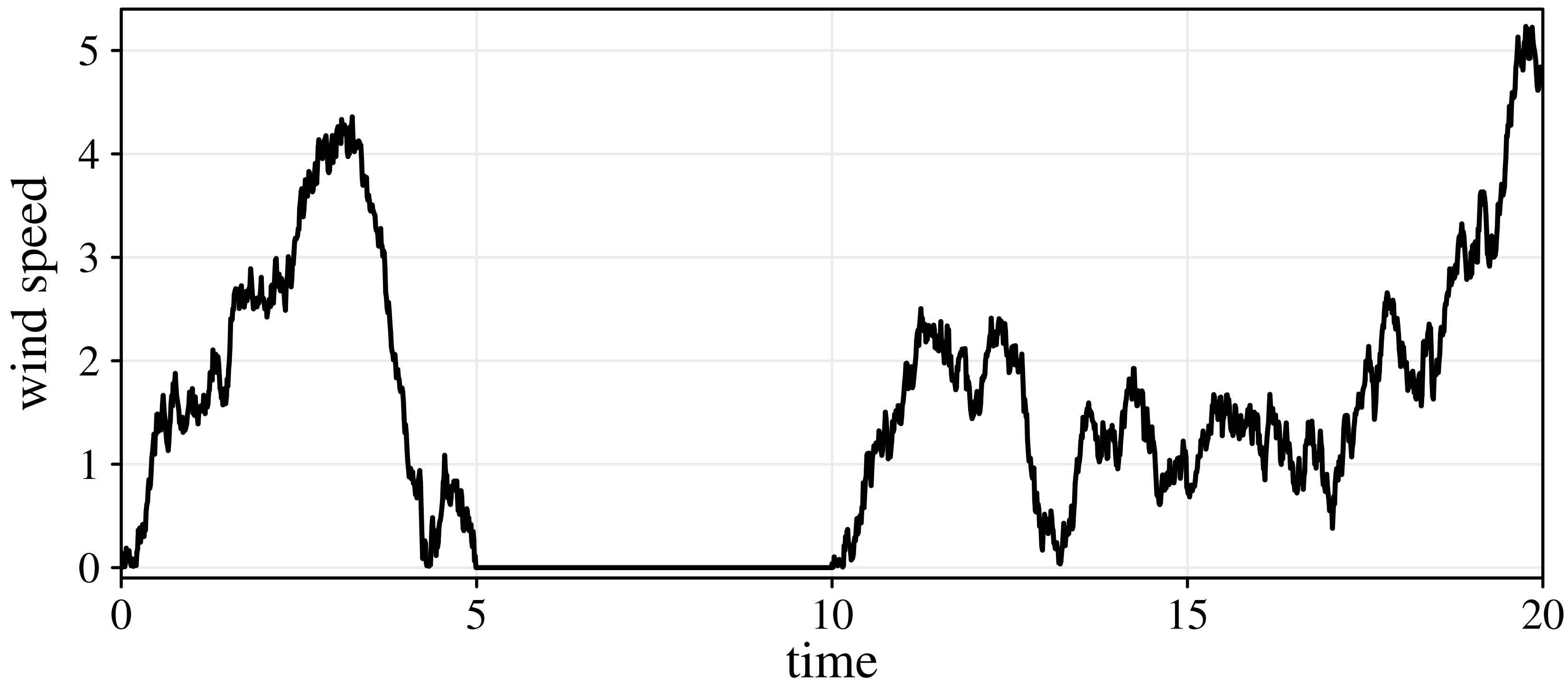
Timeliness: Is the available data up to date?

Data visualisation

Tukey: There is no excuse for failing to plot and look.



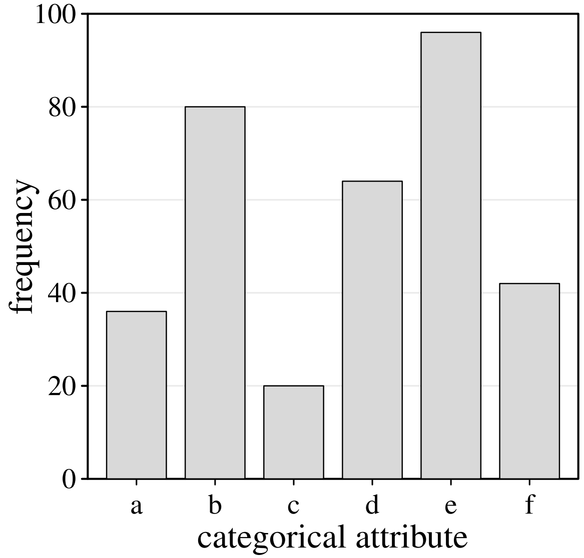
Hidden missing values



The zero values might come from a broken or blocked sensor and might be consider as missing values.

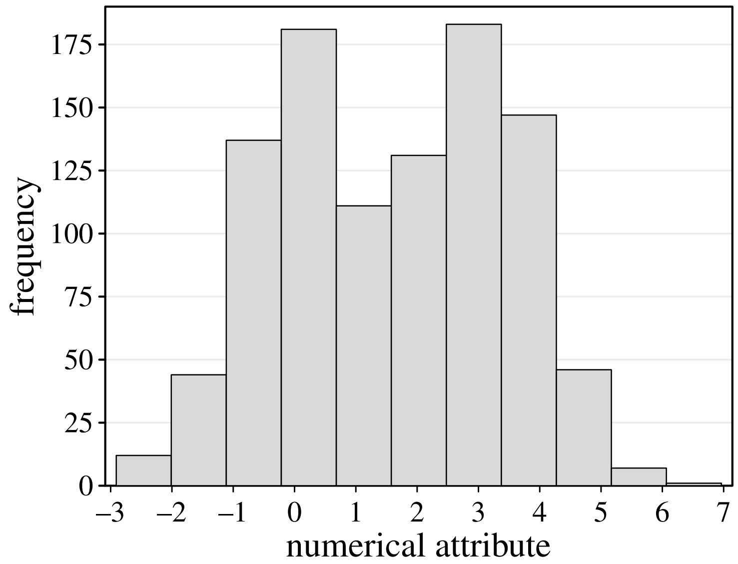
Bar charts

A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.

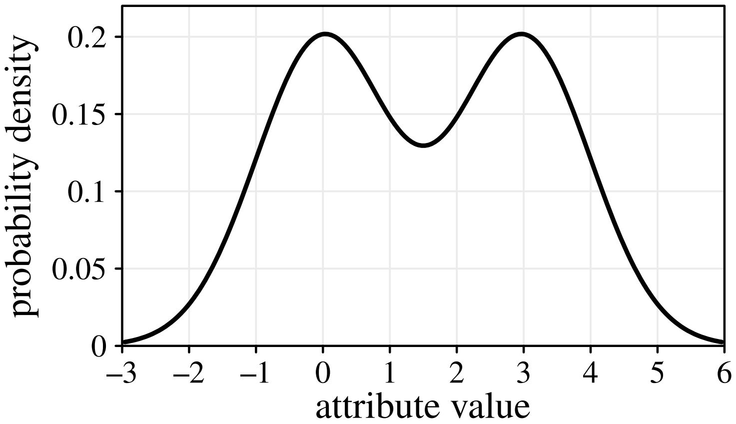


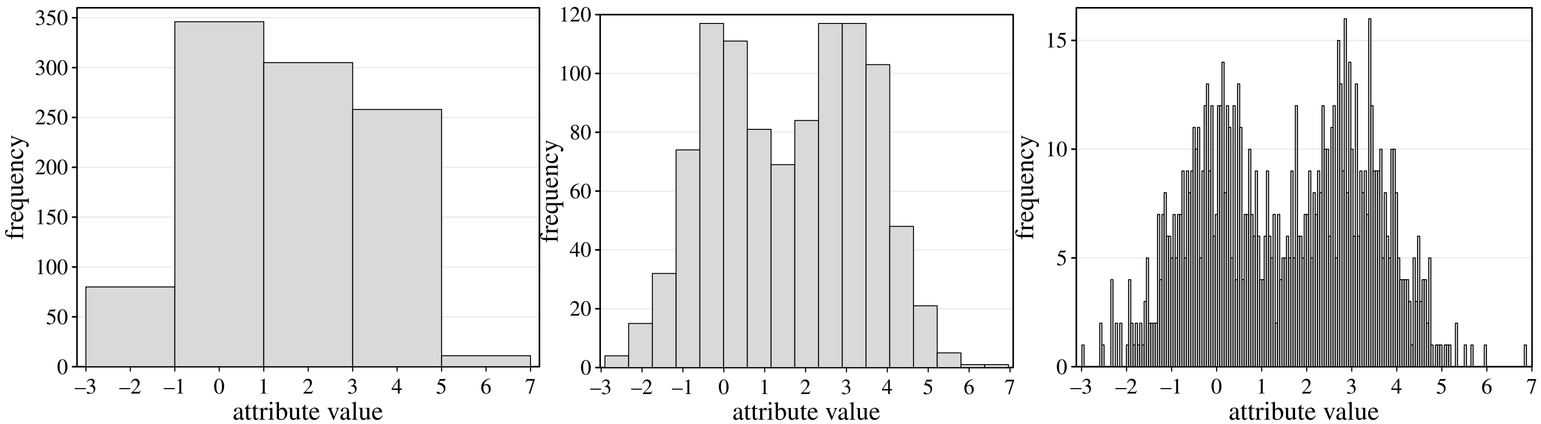
Histograms

A histogram shows the frequency distribution for a numerical attribute. The range of the numerical attribute is discretized into a fixed number of intervals (called bins), usually of equal length. For each interval the (absolute) frequency of values falling into it is indicated by the height of a bar.



Histograms: Number of bins





Three histograms with 5, 17 and 200 bins for a sample from the same bimodal distribution.

Histograms: Number of bins

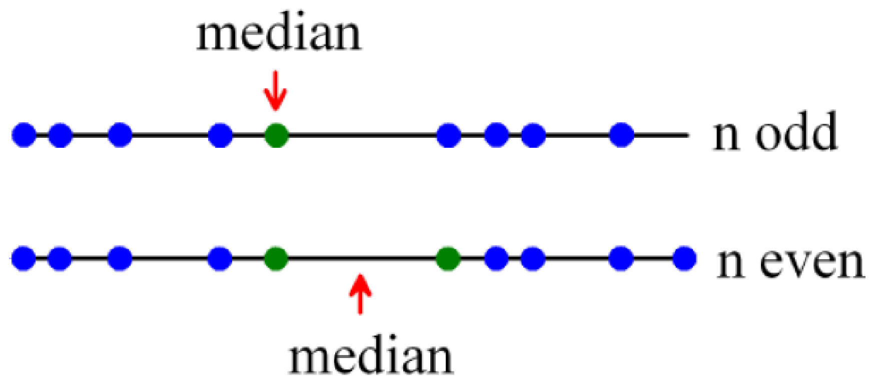
Number of bins according to Sturges' rule:

$$k = \lceil \log_2(n) + 1 \rceil$$

where n is the sample size.

(Sturges' rule is suitable for data from normal distributions and from data sets of moderate size.)

Reminder: Median, quantiles, quartiles, interquartile range



Median: The value in the middle (for the values given in increasing order).

q%-quantile $(0 < q < 100)$: The value for which q% of the values are smaller and 100-q% are larger.
The median is the 50%-quantile.

Quartiles: 25%-quantile (1st quartile), median (2nd quantile), 75%-quantile (3rd quartile).

Interquartile range (IQR): 3rd quantile - 1st quantile.

Example data set: Iris data

iris setosa

iris versicolor

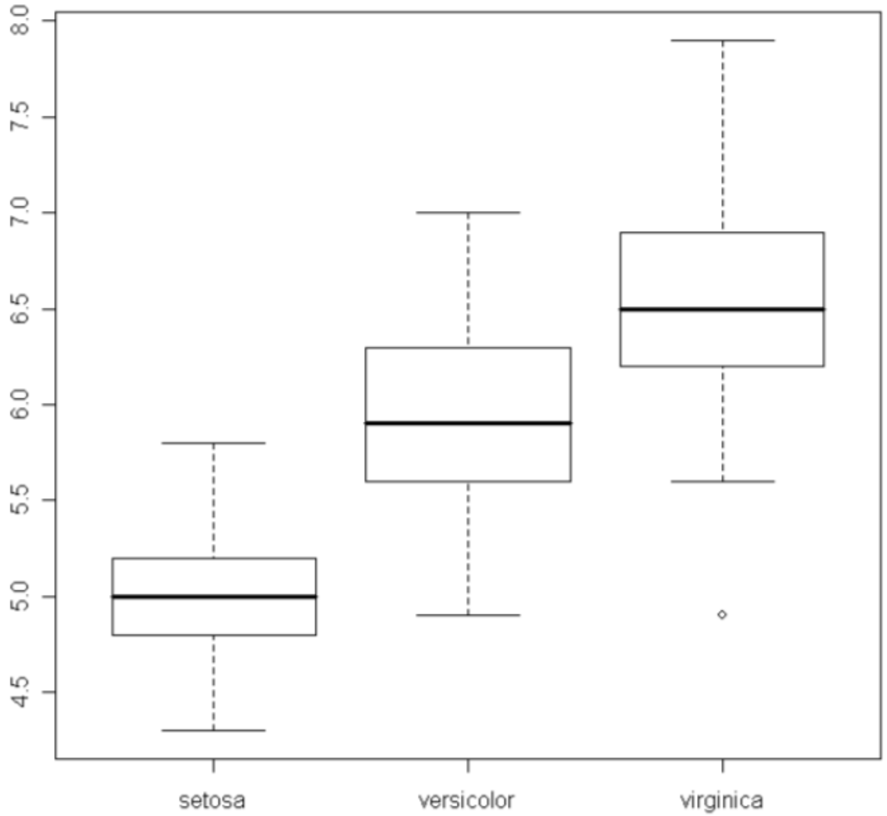
iris virginica

- collected by E. Anderson in 1935
- contains measurements of four real-valued variables:
- sepal length, sepal widths, petal lengths and petal width of 150 iris flowers of types Iris Setosa, Iris Versicolor, Iris Virginica (50 each)
- The fifth attribute is the name of the flower type.

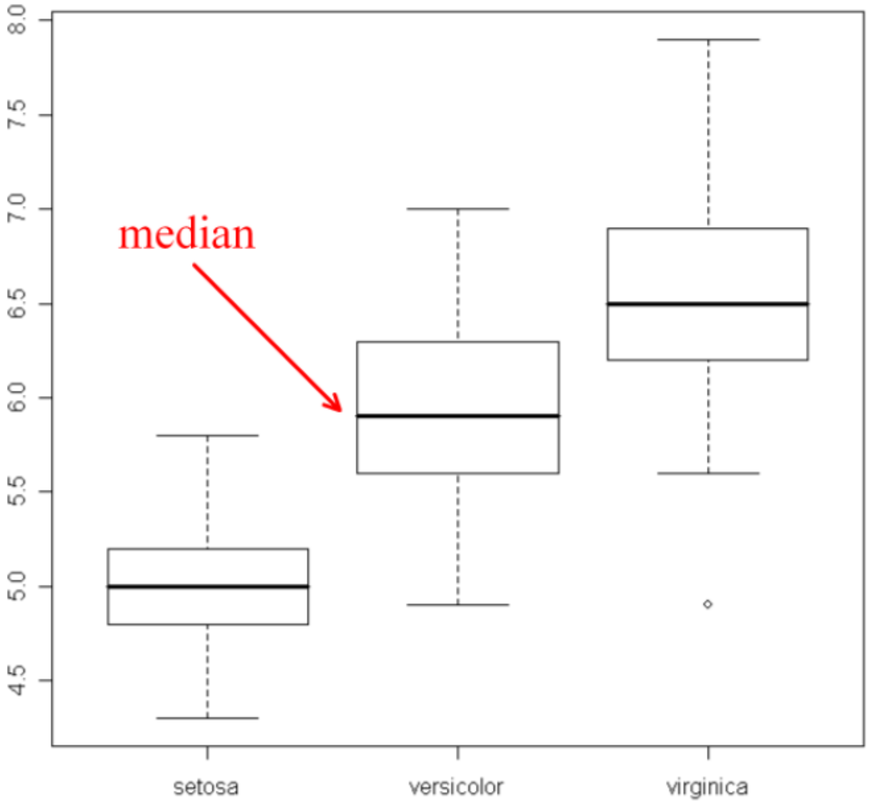
Example data set: Iris data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	Iris-setosa
...				
...				
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
...				
...				
5.1	2.5	3.0	1.1	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
...				
...				
5.9	3.0	5.1	1.8	Iris-virginica

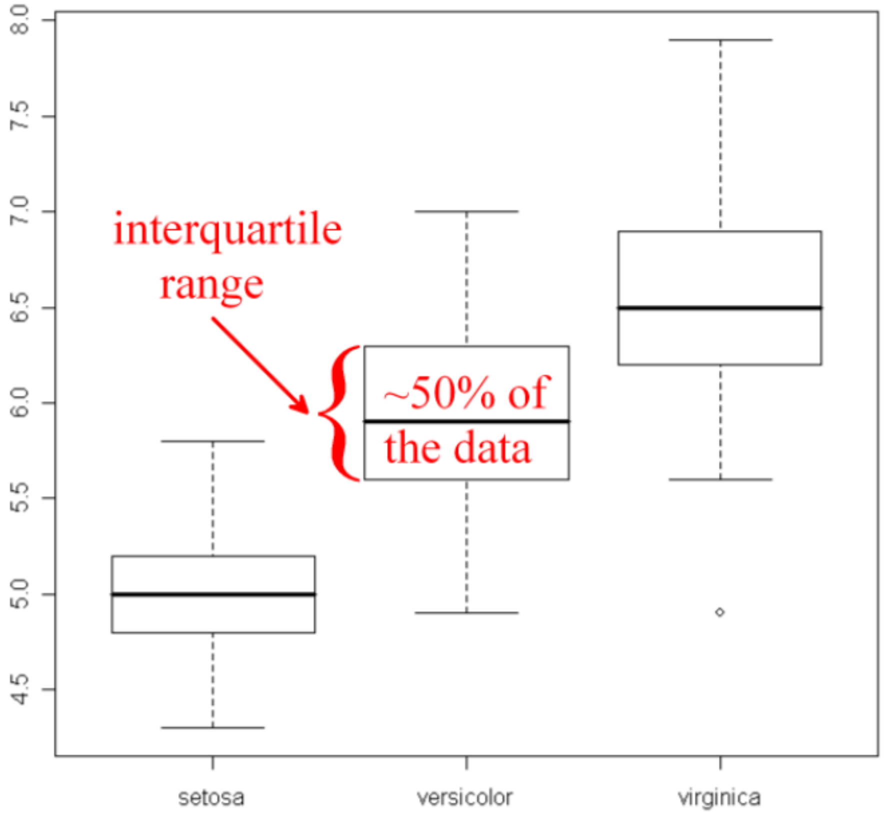
Boxplots



Boxplots

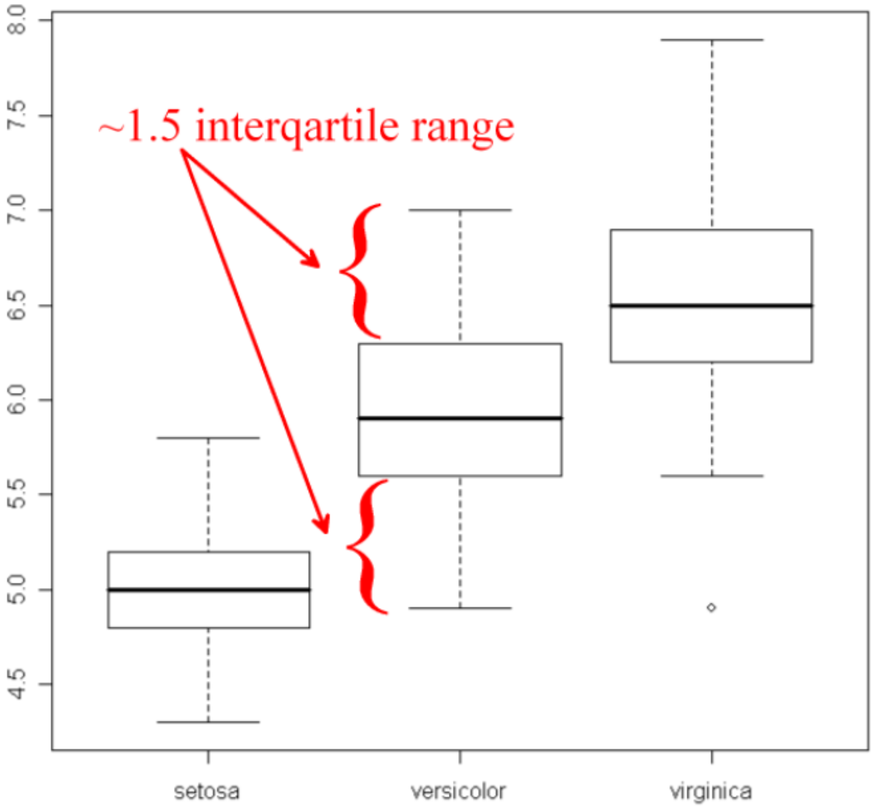


Boxplots



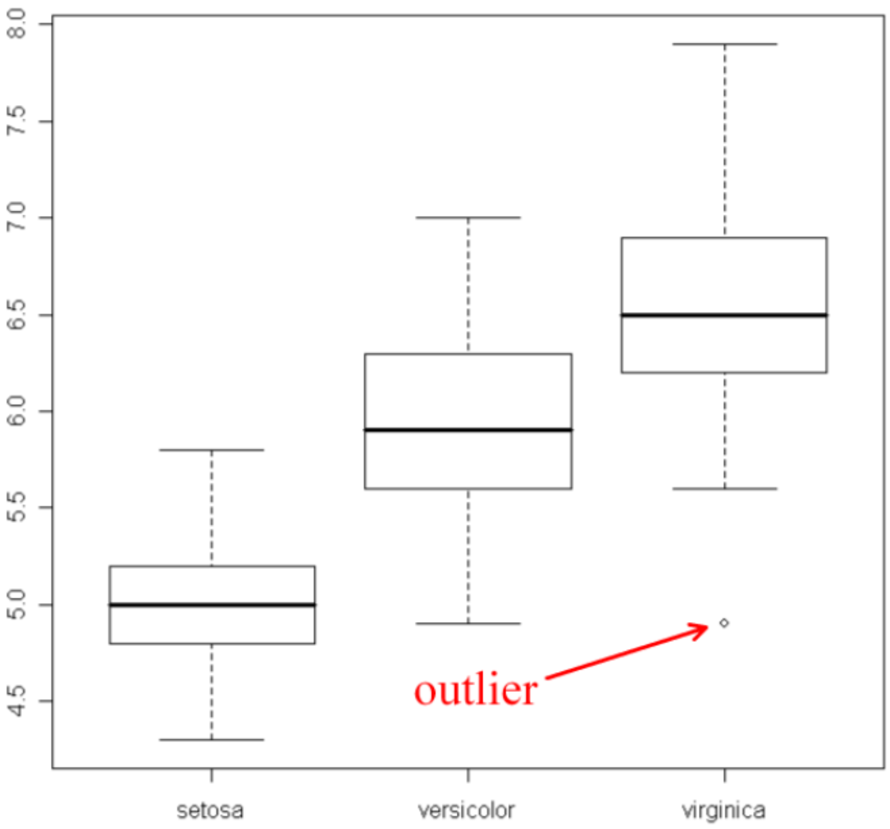
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
interquartile range
~50% of the data
setosa
versicolor
virginica

Boxplots



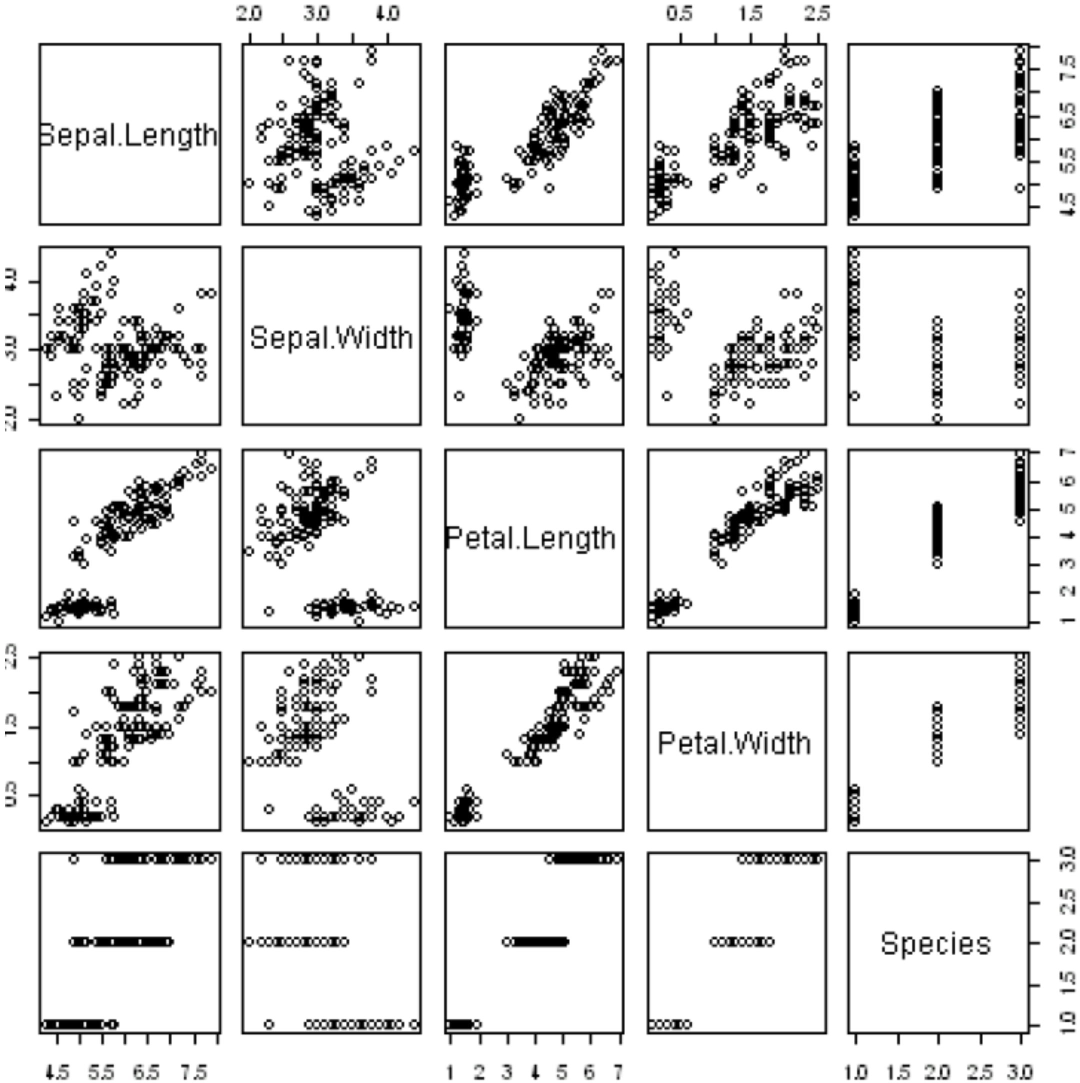
~1.5 interqartile range
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
setosa
versicolor
virginica

Boxplots

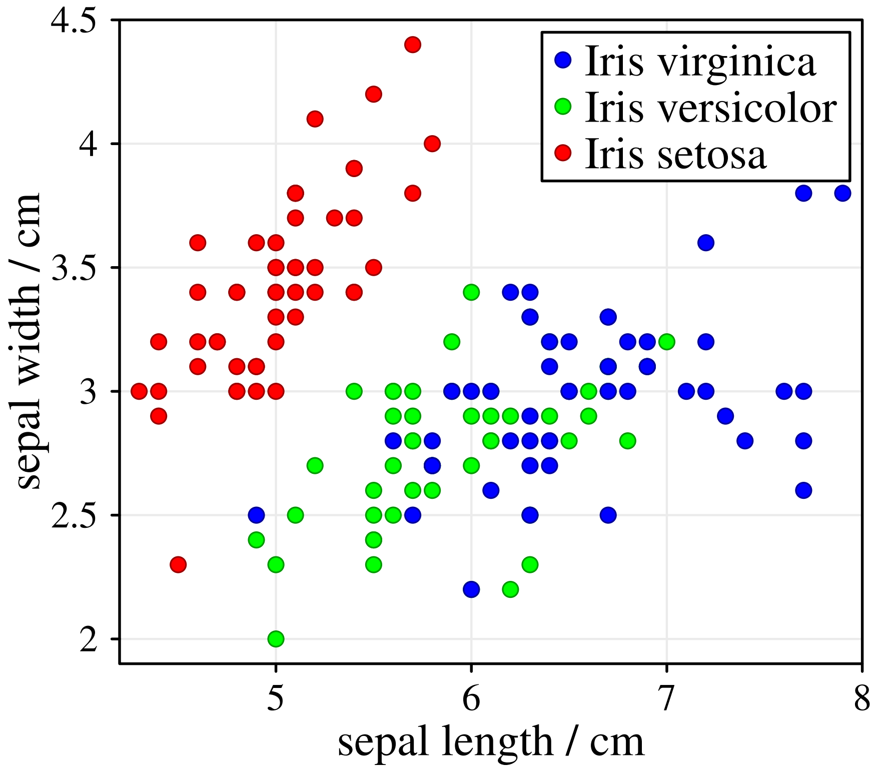


Scatter plots

Scatter plots visualize two variables in a two-dimensional plot.
Each axes corresponds to one variable.

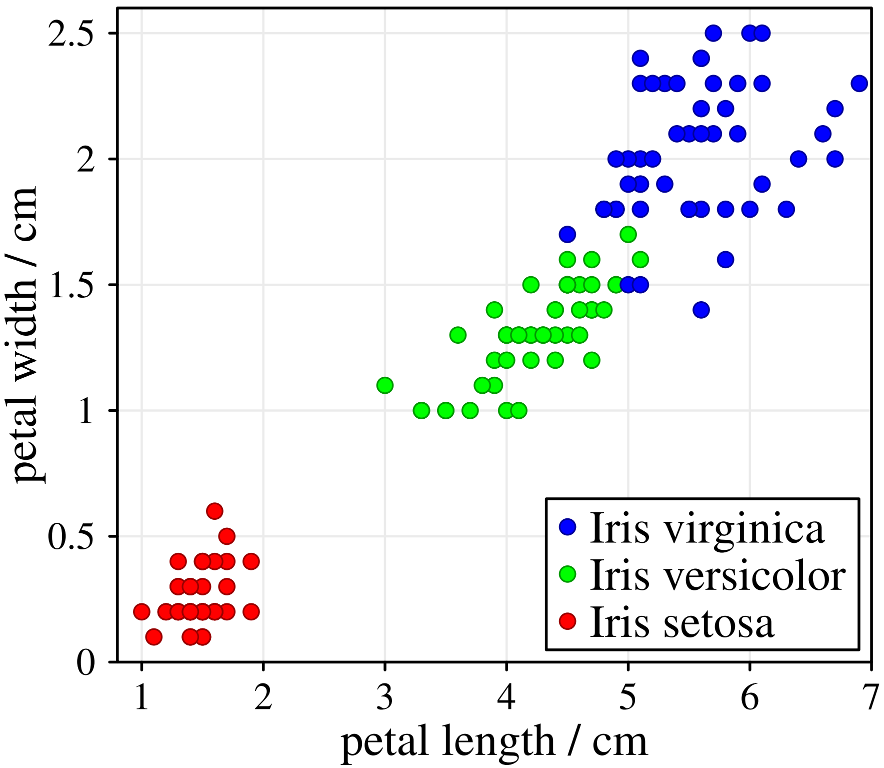


Scatter plots



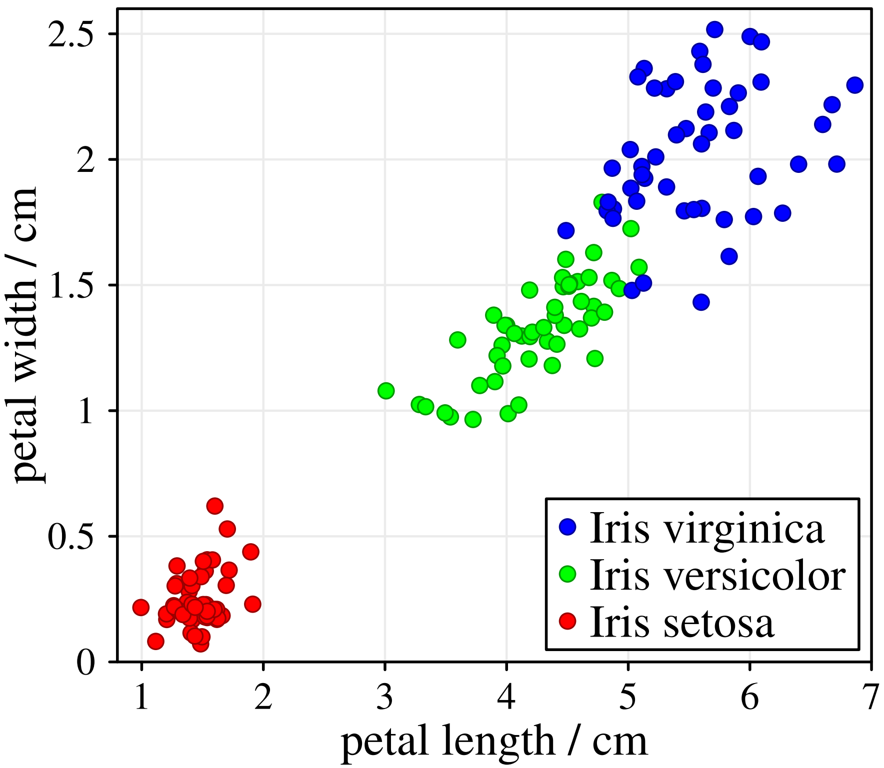
Scatter plots can be enriched with additional information: Colour or different symbols to incorporate a third attribute in the scatter plot.

Scatter plots



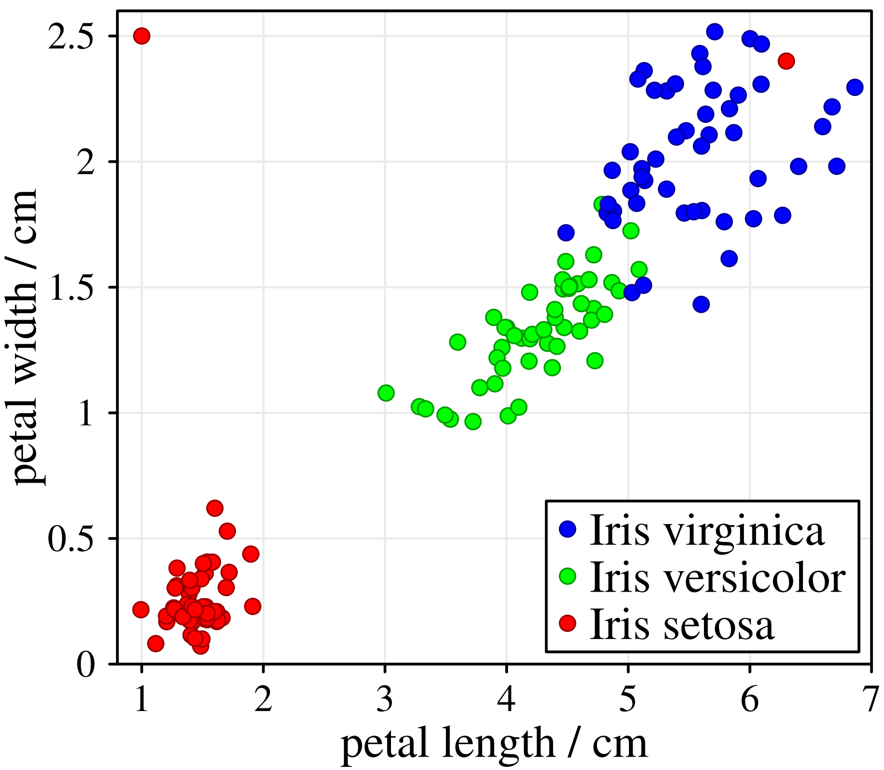
The two attributes petal length and width provide a better separation of the classes Iris versicolor and Iris virginica than the sepal length and width.

Scatter plots



Data objects with the same values cannot be distinguished in a scatter plot. To avoid this effect, jitter is used, i.e. before plotting the points, small random values are added to the coordinates. Jitter is essential for categorical attributes.

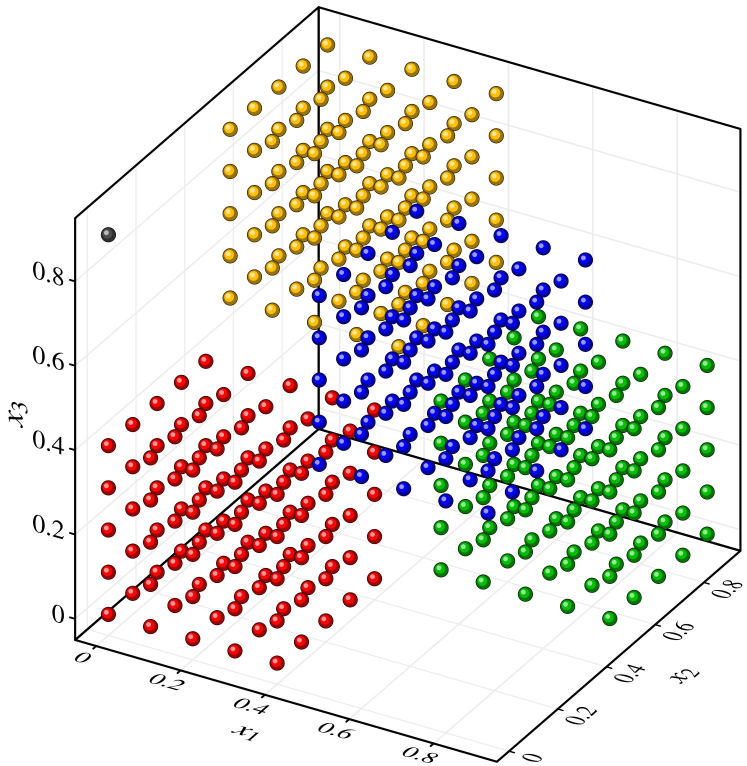
Scatter plots



The Iris data set with two (additional artificial) outliers. One is an outlier for the whole data set, one for the class Iris setosa.

3D scatter plots

For data sets of moderate size, scatter plots can be extended to three dimensions.



A 3D scatter plot of an artificial data set filling a cube in a chessboard-like manner with one outlier.

Correlation Analysis

How can the similiar behaviour of two attributes be proved?

- Pearson's correlation coefficient
- Spearman's rank correlation coefficient (Spearman's rho)
- more in the book ...

Pearson's correlation coefficient

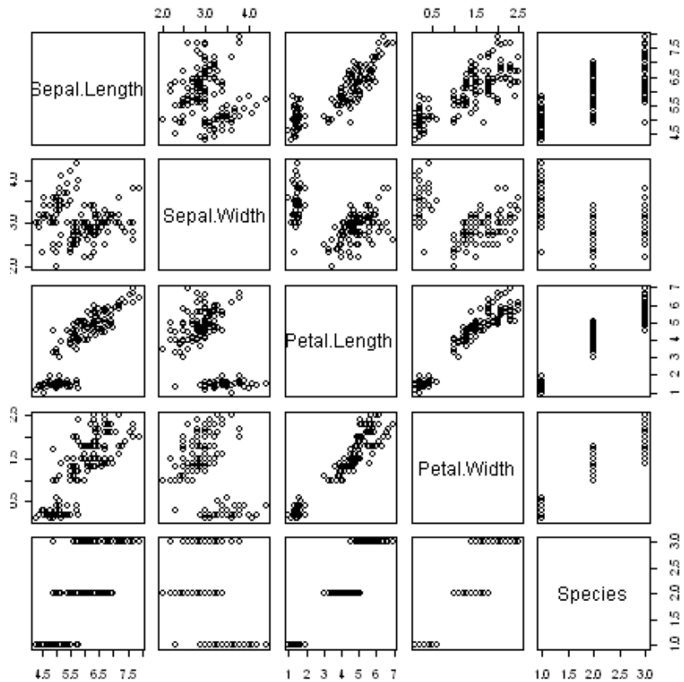
The (sample) Pearson's correlation coefficient is a measure for a **linear relationship** between two **numerical** attributes X and Y and is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the attributes X and Y, respectively. s_x and s_y are the corresponding (sample) standard deviations.

- $-1 \leq r_{xy} \leq 1$
- The larger the absolute value of the Pearson correlation coefficient, the stronger the linear relationship between the two attributes. For $|r_{xy}| = 1$ the values of X and Y lie exactly on a line.
- Positive (negative) correlation indicates a line with positive (negative) slope.

Pearson's correlation coefficient: Iris data set



	sepal length	sepal width	petal length	petal width
sepal length	1.000	-0.118	0.872	0.818
sepal width	-0.118	1.000	-0.428	-0.366
petal length	0.872	-0.428	1.000	0.963
petal width	0.818	-0.366	0.963	1.000

Outlier detection

An outlier is a value or data object that is far away or very different from all or most of the other data.

Causes for outliers:

- Data quality problems (erroneous data coming from wrong measurements or typing mistakes)
- Exceptional or unusual situations/data objects.

- Outliers coming from erroneous data should be excluded from the analysis.
- Even if the outliers are correct (exceptional data), it is sometime useful to exclude them from the analysis.
  For example, a single extremely large outlier can lead to completely misleading values for the mean value.

Outlier detection: Single attributes

Categorical attributes: An outlier is a value that occurs with a frequency extremely lower than the frequency of all other values.

Numerical attributes:

- Outliers in boxplots.
- Statistical tests, for example Grubb's test: ... (see the exercise)

Outlier detection for multidimensional data

- Scatter plots for (visually detecting) outliers w.r.t. two attributes.
- PCA or MDS plots for (visually detecting) outliers.
- Cluster analysis techniques: Outliers are those points which cannot be assigned to any cluster.

Missing values

For some instances values of single attributes might be missing.

Causes for missing values:

- broken sensors
- refusal to answer a question
- irrelevant attribute for the corresponding object (pregnant (yes/no) for men)

Missing value might not necessarily be indicated as missing (instead: zero or default values).

A checklist for data understanding

- Determine the quality of the data. (e.g. syntactic accuracy)
- Find outliers. (e.g. using visualization techniques)
- Detect and examine missing values. Possible hidden by default values.
- Discover new or confirm expected dependencies or correlations between attributes.
- Check specific application dependent assumptions (e.g. the attribute follows a normal distribution)
- Compare statistics with the expected behavior.

A checklist for data understanding: Must Do

- Check the **distributions for each attribute**
  (unexpected properties like outliers, correct domains, correct medians)

- Check **correlations or dependencies** between pairs of attributes