Data Analysis and Knowledge Discovery

Data Understanding I

Antti Airola

University of Turku
Department of Computing

Antti.Airola@utu.fi

Goals of data understanding

- Gain general insight on the data (independent of the project goal).
- Checking the assumptions made during the project understanding phase.
  (representativeness, informativeness, data quality, presence/absence of external factors, dependencies, ...)
- Checking the specified domain knowledge.
- Suitability of the data for the project goals.

Rule of thumb: never trust any data before some plausibility tests

Attribute understanding

We (often) assume that the data set is provided in the form of one or more simple tables.

	attribute_1	...	attribute_m
record_1			
$\vdots$			
record_n			

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

Data matrix

This table is often encoded as a $\mathbb{R}^{n \times d}$ data matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix}$$

where the data set consist of n instances and d features. Many data analysis methods are defined in terms of this matrix, or its transpose $\mathbf{X}^T$.

Transforming table to data often requires using preprocessing techniques to convert non-numerical features (e.g. colour, product type) to numerical form, and imputation of missing values if any. $m \neq d$ because many standard transformations results in additional features being generated.

Types of attributes

categorical (nominal): finite domain
The values of a categorical attribute are often called classes or categories.
Examples: {Finnish, Swedish, English}, {pine,spruce,birch}, {blue, green, yeallow, violet}

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

Scales for numerical attributes

interval scale: The definition of the value 0 is arbitrary. Ratios are meaningless.
Examples: date, temperature measured in Celsius or Fahrenheit degrees.

ratio scale: 0 has a canonical meaning. Ratios make sense.
Examples: distance, duration

absolute scale: Domain with a unique measurement unit.
Examples: any kind of counting process (number of children, number of visits to the doctor)

Specific problems of categorical attributes

- Different levels of granularity might be definable.
  Examples: Product categories/types:
  - General category: drinks, food, clothes,...
  - More refined categories for drinks: water, beer wine,...
  - Further refinement for water based on the producer.
  - Further refinement of the water of each producer based on the bottle size (0.33 l, 0.5 l, 1 l, 1.5 l).

  The most refined level provides the most detailed information, but might not help to discover general associations like *Wine and cheese are often bought together*.

  It is crucial to choose an appropriate level of granularity to support the analysis goals.

Specific problems of categorical attributes

- dynamic domains: The possible values of the domain might change over time.
  Example: Certain product categories or products might not be sold anymore.
  New product categories or products are introduced.

The analysis of such data can cause problems, for instance:

- Products that have just entered the market will not show significant (accumulated) sales numbers compared to products that have been sold for years.

Encoding binary attributes

Binary attributes: categorical variables with two possible values

- {Pass, Fail}, {False, True}, {Sick, Healthy}
- 0/1 binary encoding
  - Pass=0, Fail=1
  - False=0, True=1
  - Sick=0, Healthy=1
- Or vice versa (Pass=1, Fail=0...), order does not matter

Encoding categorical attributes

One hot encoding

- {Finland, Sweden, Denmark, Norway}, {Blue, Green, Yellow}
- Wrong: ordinal Finland=1, Sweden=2... encoding not useful, since this implicitly assumes ordered categories
- Right: transform each possible value into binary variable
  - isFin=0/1, isSwed=0/1, isDen=0/1, isNor=0/1
  - isBlue=0/1, isGreen=0/1, isYellow=0/1
- categorical feature with k possible values transformed into k binary variables
- Finland = 1000, Sweden=0100, Denmark=0010, Norway=0001

One hot encoding

Name	Finnish	Swedish	Danish	Norwegian
Sauli	1	0	0	0
Margrethe	0	0	1	0
Carl	0	1	0	0
Harald	0	0	0	1

Extending one hot encoding

Name	Finnish	Swedish	Danish	Norwegian
Sauli	1	0	0	0
Margrethe	0	0	1	0
Carl	0	1	0	0
Harald	0	0	0	1
Björn	?	?	?	?
Charles	?	?	?	?

- Björn had dual nationality: Finnish and Swedish
- Charles is from the United Kingdom

Extending one hot encoding

Name	Finnish	Swedish	Danish	Norwegian
Sauli	1	0	0	0
Margrethe	0	0	1	0
Carl	0	1	0	0
Harald	0	0	0	1
Björn	1	1	0	0
Charles	0	0	0	0

- Could also add new possible value to domain of the nationality variable (e.g. "British" or "Other")

Encoding ordinal attributes

Ordinal variables are ordered categorical variables

- {B.Sc.,M.Sc.,Ph.D.}, {$\star$, $\star\star$, $\star\star\star$, $\star\star\star\star$, $\star\star\star\star\star$}
- Order gives natural numerical representation:
  - B.Sc.=1, M.Sc.=2, Ph.D. =3
  - $\star = 1$, $\star\star = 2$, $\star\star\star = 3$, $\star\star\star\star = 4$ $\star\star\star\star\star = 5$
- meaningful to say that average education was 1.7, or that 'goodness' distance from $\star\star\star$ movie to $\star$ movie is same as from $\star\star\star\star\star$ to $\star\star\star$?
- not really, but often works good enough
- alternatively, treat ordinal variable as categorical

Data quality

- Low data quality makes it impossible to trust analysis results: "Garbage in, garbage out"
- Mistakes made in the data are most often very difficult to recover by computational methods.

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.

Syntactic and semantic accuracy

Data quality: syntactic accuracy

Syntactic accuracy is violated if an entry does not belong to the domain of the attribute.

Examples:

- The entry *fmale* for the categorical attribute *sex* violates syntactic accuracy.
- Text entries for numerical attributes violate syntactic accuracy.
- Values out of the range for numerical attributes violate syntactic accuracy (negative numbers for weight, distance, counting processes,...).

Syntactic accuracy can be checked quite easily.

Data quality: semantic accuracy

Semantic accuracy is violated if an entry is not correct although it belongs to the domain of the attribute.

Example:

- The entry *PhD* for the attribute *education level* and entry *10* for the attribute *age* are both within the domains of the attributes. However, having both these values for the same record would almost certainly mean a semantic error.

Semantic accuracy is more difficult to check than syntactic accuracy, and sometimes even impossible to check.

Semantic accuracy can only be checked based on "business rules" (e.g. no one lives over 120 years) and plausibility checks.

Data quality: completeness

Completeness is violated if an entry is missing.

- w.r.t. attribute values: Fraction of null entries for an attribute. Note that missing values are not always marked explicitly as missing, for instance in the case of default entries.
- w.r.t. records: Complete records might be missing because
  - three years ago SAP was introduced and not all customer data were transferred to the new system.
  - the data set is biased and non-representative. (A bank might have rejected customers with no income.)

Data quality: unbiased and representative

The data should always be unbiased and representative, i.e. it should contain all information about the inherent patterns and rules in the data.
In many applications we do not have that sort of data

- Machine condition monitoring: A lot of examples when machine in running normally. Sometimes not possible to obtain interesting data, such as in the case of nuclear power plant.
- Natural disasters: E.g. no earthquake data from a certain area where future earthquake probability should be estimated.
- Mortgage/insurance etc. analysis: Certain types of customers are totally missing, e.g. we may only have information about customers who have been granted a loan.

Data quality: unbiased and representative

The biasness and non-representativeness of the data is one of the most difficult problems in data analysis.
Sometimes it is not even known what is needed, e.g.

- Electricity load prediction: What information should be used to estimate the electricity load for tomorrow.
- Pattern recognition tasks: Often impossible to describe where the recognition should be based.

Data quality: unbalanceness and timeliness

Unbalanced data: The data set might be biased extremely to one type of records.

Example: Production line for goods including quality control. Defective goods will be a very small fraction of all records.

Timeliness: Are the available data up to date to be considered to be representative? This is related to the non-stationarity of the domain where only the recently collected data provide relevant information.

Example: Many industrial processes change dynamically and are non-stationary in nature. Hence too old data may not be much of use for analysing current and future states of the process.

Missing values

For some instances values of single attributes might be missing.

Causes for missing values:

- broken sensors
- refusal to answer a question
- medical test not done for this person
- combined data from different data bases with different variables recorded
- irrelevant attribute for the corresponding object (gas consumption in km/L for electric car)

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

Types of missing values

Consider the attribute X_{obs}. A missing value is denoted by ?.

X is the true value of the considered attribute, i.e. we have

$$X_{\text{obs}} = X, \qquad \text{if } X_{\text{obs}} \neq ?$$

Let Y be the (multivariate) (random) variable denoting the other attributes apart from X.

Types of missing values

Missing completely at random (MCAR): The probability that a value for X is missing does neither depend on the true value of X nor on other variables.

$$P(X_{\text{obs}} = ?) \;=\; P(X_{\text{obs}} = ? \mid X, Y)$$

Examples of MCAR:

- Each survey respondent answers randomly chosen subset of questions
- Fields of data base corrupted randomly
- Battery of a sensor runs out and replaced at random times

MCAR is also called Observed At Random (OAR).

Implications of MCAR

- probability of missing value same for each instance
- missing values follow same distribution as observed
- ignoring missing cases or using simple imputation methods will not systematically bias your analysis
- example: data base with yearly income recorded for random sample of Finns. Hard disk breaks, and data recovered only for 50 % of the people. Assuming records were lost randomly, mean income estimated from this sample should not be biased by the missing values.
- often MCAR is not a realistic assumption

Types of missing values

Missing at random (MAR): The probability that a value for X is missing depends on Y, but conditionally independent of X given Y.

$$P(X_{\text{obs}} = ? \mid Y) = P(X_{\text{obs}} = ? \mid X, Y)$$

Examples of MAR (assuming relevant attributes recorded in data):

- Battery of a sensor runs out, when it is replaced depends on time of day or weather
- Older people are more likely to have their blood pressure measured and recorded as part of care
- High earning people in certain demographics (age, education, job...) less likely to report their income

Implications of MCAR

- missing values do not follow same distribution as observed
- if other variables not controlled for, analysis can be biased
- example: can estimate temperature given time of day, even if more missing values at night
- problem: when can we be sure that Y provides all required information about missingness mechanism?

Types of missing values

Nonignorable: The probability that a value for X is missing depends on the true value of X and is not conditionally independent of X given Y.

Examples of Nonignorable:

- Temperature sensor does not work at below zero temperatures
- No person with income over 1000 000 € will report their income

For MCAR and MAR, the missing values can be estimated – at least in principle with enough data – based on the values of the other attributes.

In these extreme cases, it is impossible to provide any statements concerning temperatures below 0°C or incomes over 1000 000 €.

Types of missing values

- In the case of MCAR, it can be assumed that the missing values follow the same distribution as the observed values of X.
- In the case of MAR, the missing values might not follow the distribution of X. But by taking the other attributes into account, it is possible to derive reasonable imputations for the missing values.
- In the case of nonignorable missing values it is impossible to provide sensible estimations for the missing values.

Types of missing values

If it is not known based on domain knowledge which kind of missing values can be expected, the following strategy can be applied.

1. Turn the considered attribute X into a binary attribute:
   - Replace all measured values by the values *yes* and all missing values by the value *no*.
2. Build a classifier with binary attribute X as the target attribute and use all other attributes for the prediction of the class values *yes* and *no*.
3. Determine the misclassification rate. The misclassification rate is the proportion of data objects that are not it assigned to the correct class by the classifier.

Types of missing values

- In the case of MCAR, the other attributes should not provide any information, whether X has a missing value or not.
  - Therefore, the misclassification rate of the classifier should not differ significantly from pure guessing, i.e. if there 10% missing values for the attribute X, the misclassification rate of the classifier should not be much smaller than 10%.
- If, however, the misclassification rate of the classifier is significantly better than pure guessing, this is an indicator that there is a correlation between missing values for X and the values of the other attributes. The missing values are not MCAR.
- MCAR and MAR cannot be distinguished from nonignorable this way.

Data visualisation

Data visualisation is one of the most important steps for

- Data understanding and preliminary data quality evaluation
- Learning from the domain

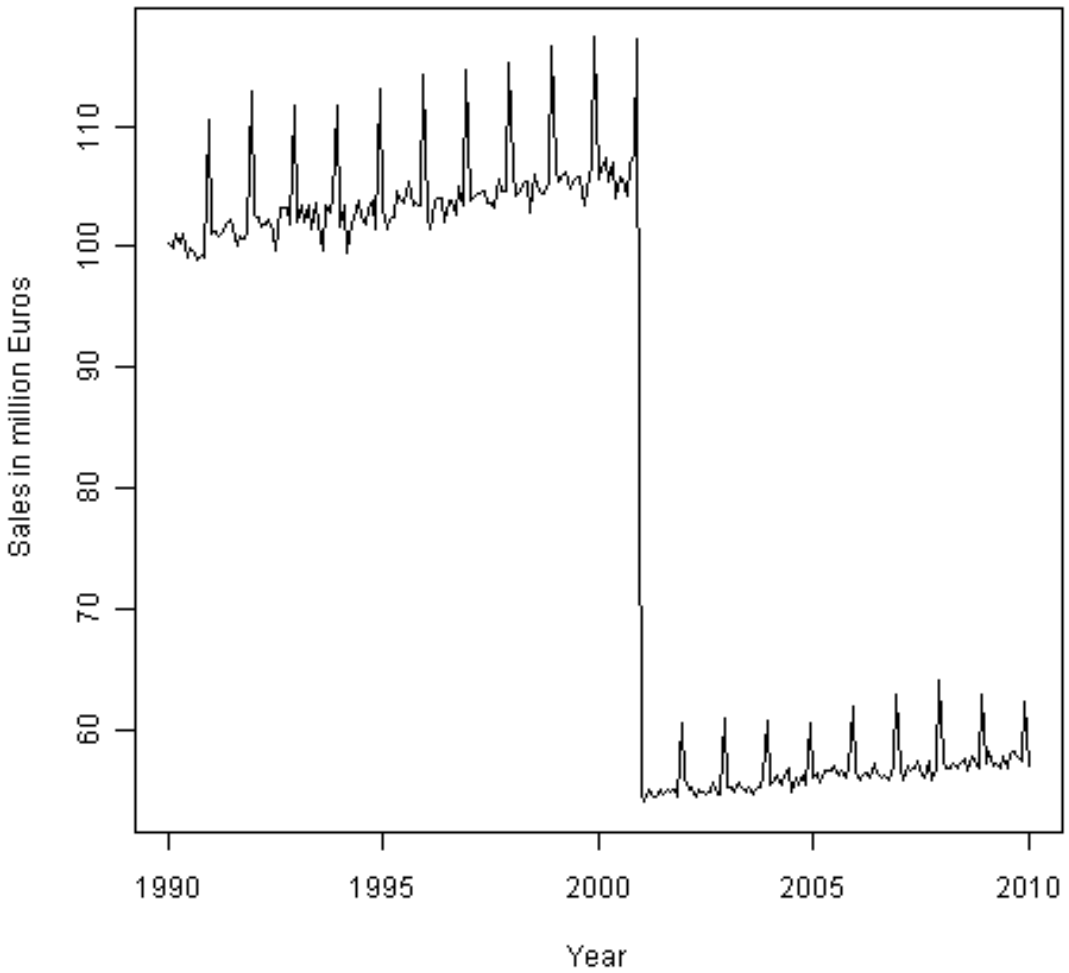
Visualisation as a test

- When visualisations reveal patterns or exceptions, then there is "something" in the data set.
- When visualisations do not indicate anything specific, there might still be patterns or structures in the data that cannot be revealed by the corresponding (simple) visualisation techniques.

There are infinitely many possibilities for visualisation
Here we cover some very useful data visualisation techniques

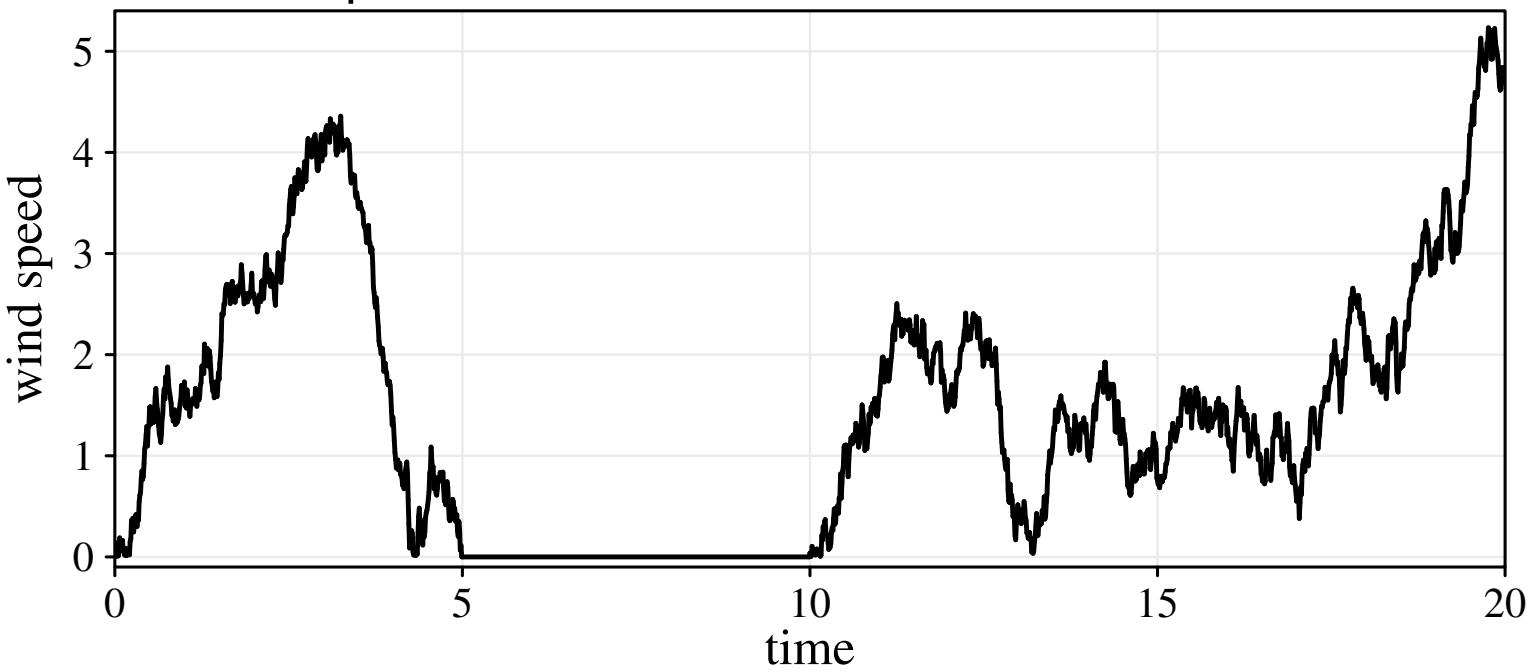
Data visualisation

There is no excuse for failing to plot and look.



Data visualisation: hidden missing values

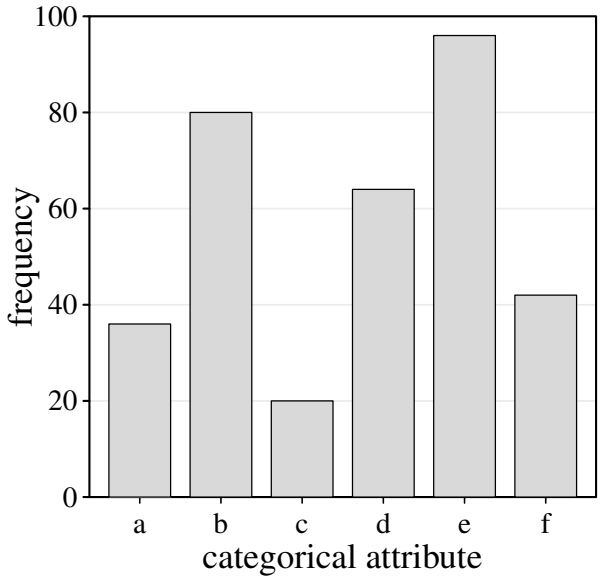
Measured wind speeds.



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

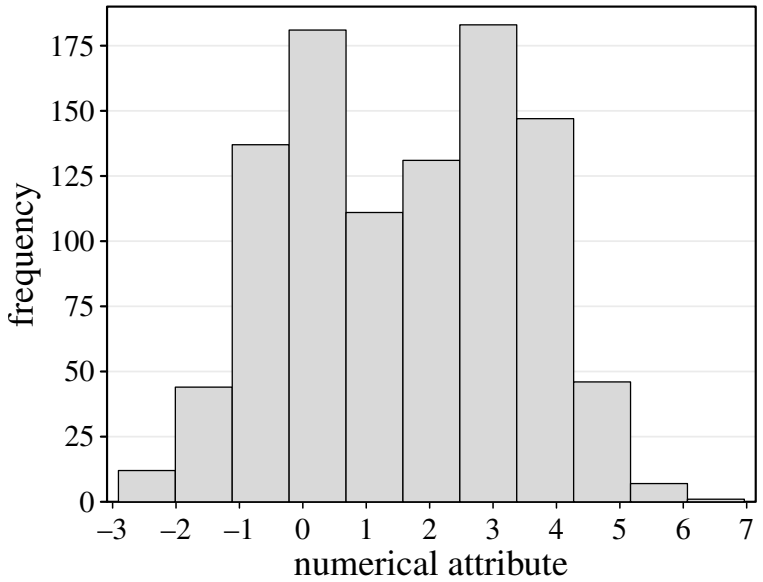
Bar charts

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

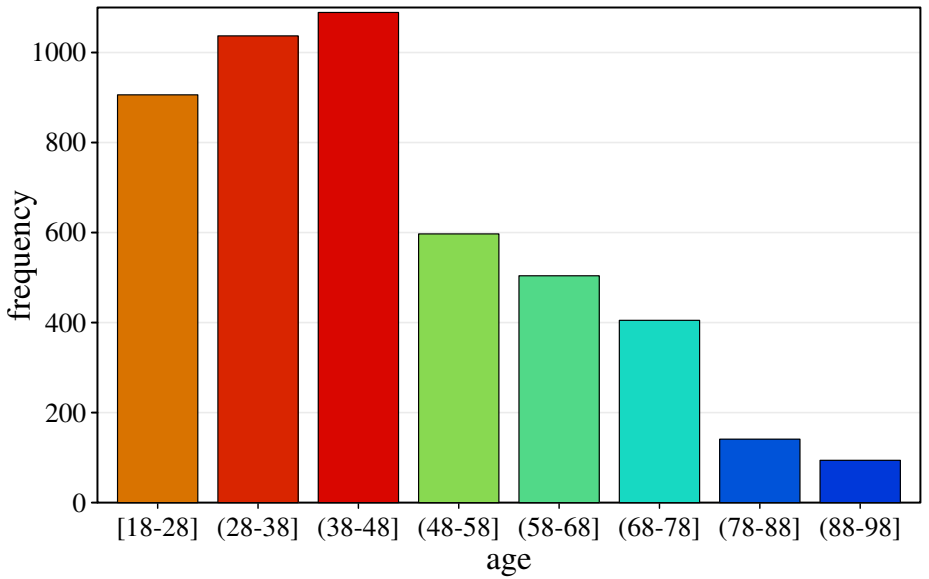


Histograms

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.

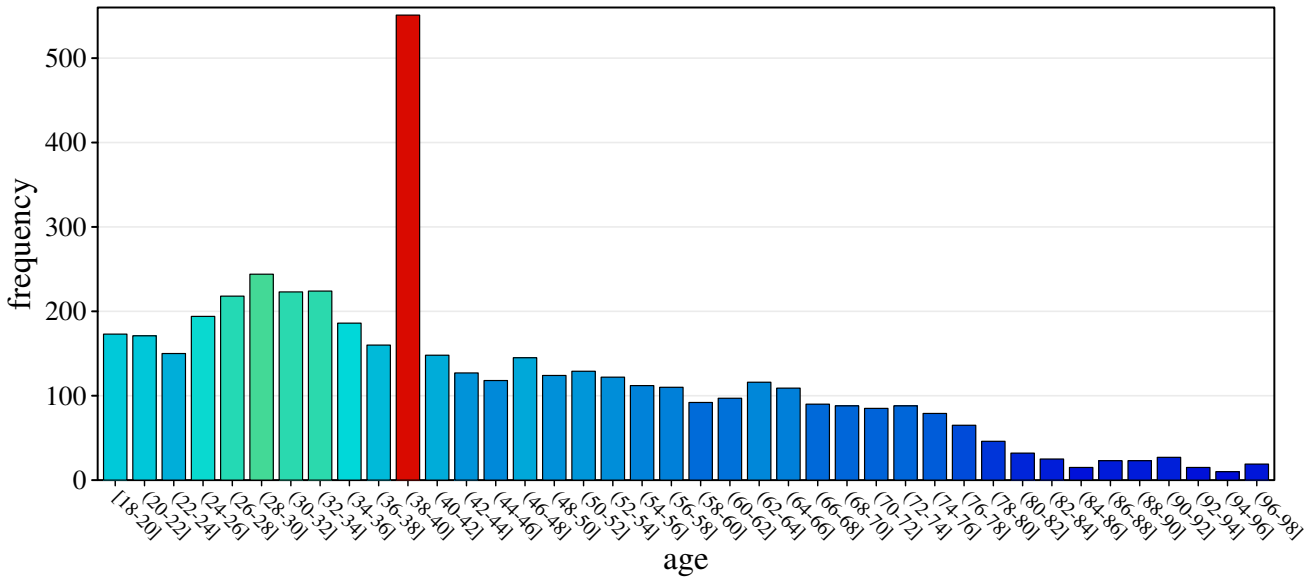


Histograms: number of bins



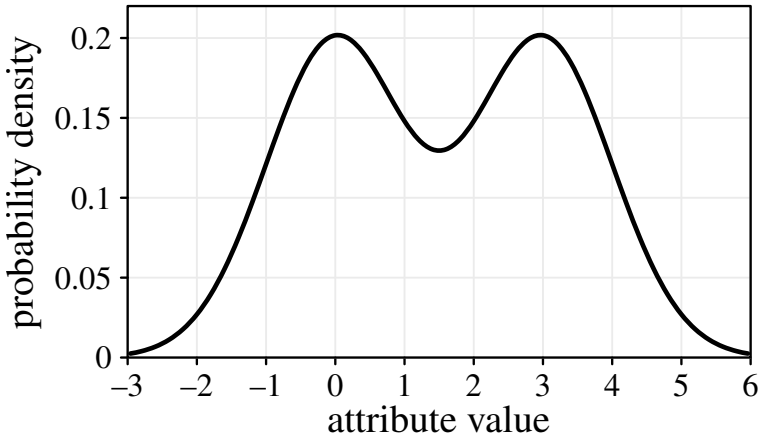
Distribution of the age of the customers in 2010

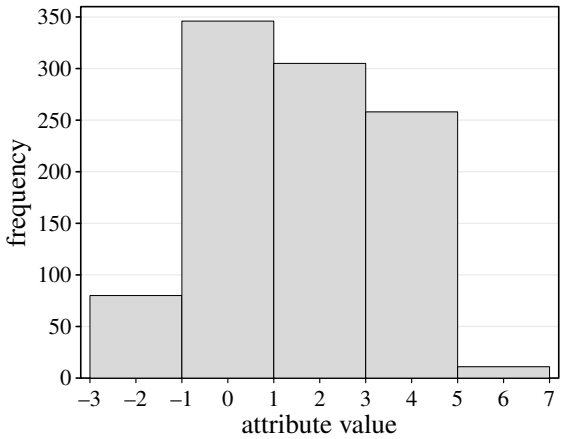
Histograms: number of bins

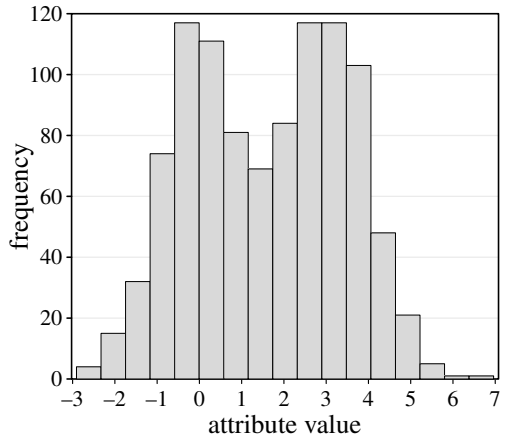


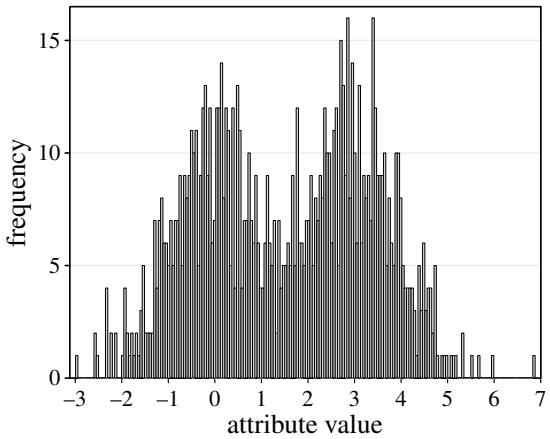
Distribution of the age of the customers in 2010

Histograms: number of bins









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

Histograms: number of bins

The choice of the number of bins clearly influences the results.

- There is no *best* number of bins
- Different bin sizes can reveal different features of the data.
- Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution.
- Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate
- Some default values with total number of observations n:
  - Excel: $\sqrt{n}$
  - R hist(): $\lceil \log_2(n) + 1 \rceil$ (Sturges' rule)
  - Python Pandas / numpy: 10 bins

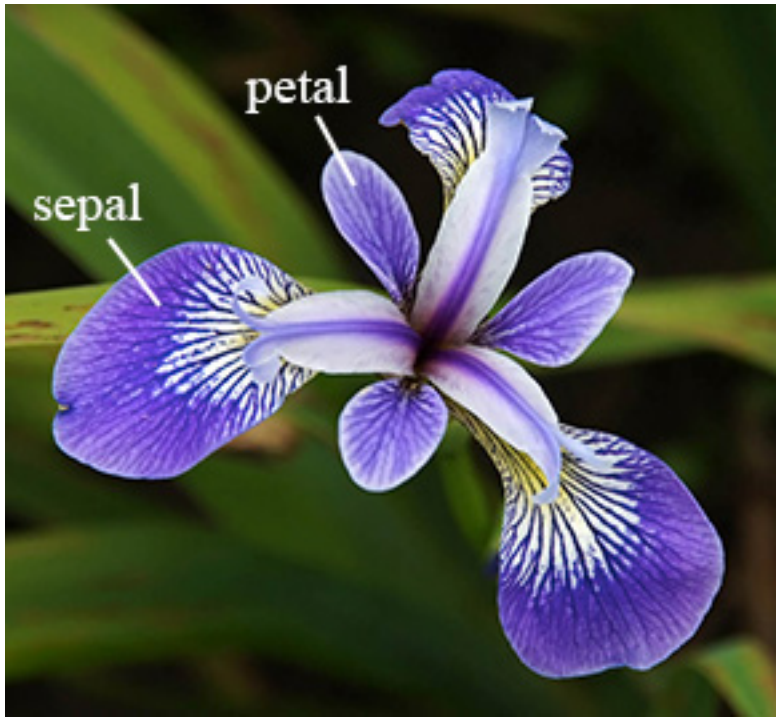
Experimentation is usually needed to determine an appropriate width.

Example data set: Iris data

Collected by E. Anderson in 1935 and 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

iris setosa

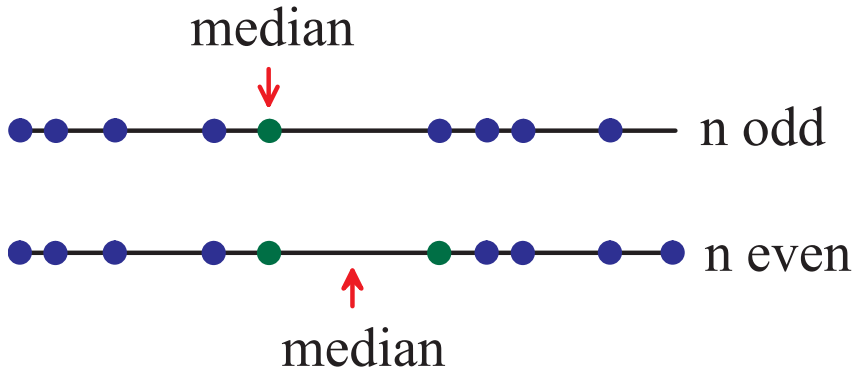
iris versicolor

iris virginica

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

Reminder: Median, quantiles, quartiles, interquartile range



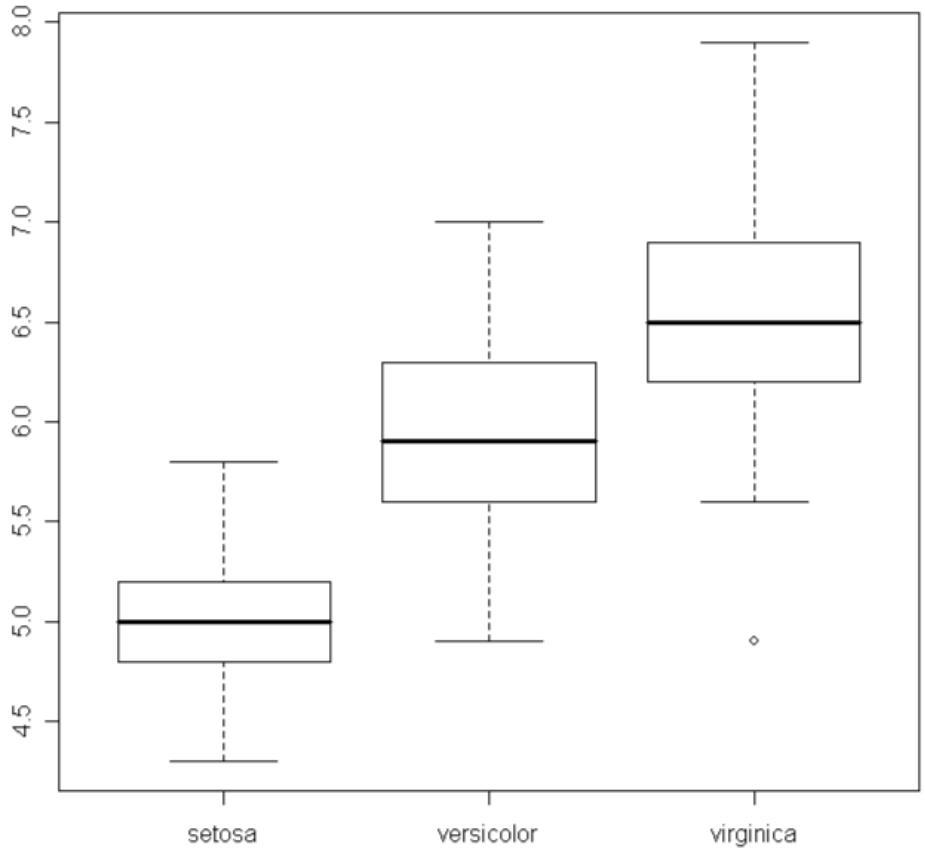
Median: The value in the middle (for the values given in increasing order).

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.

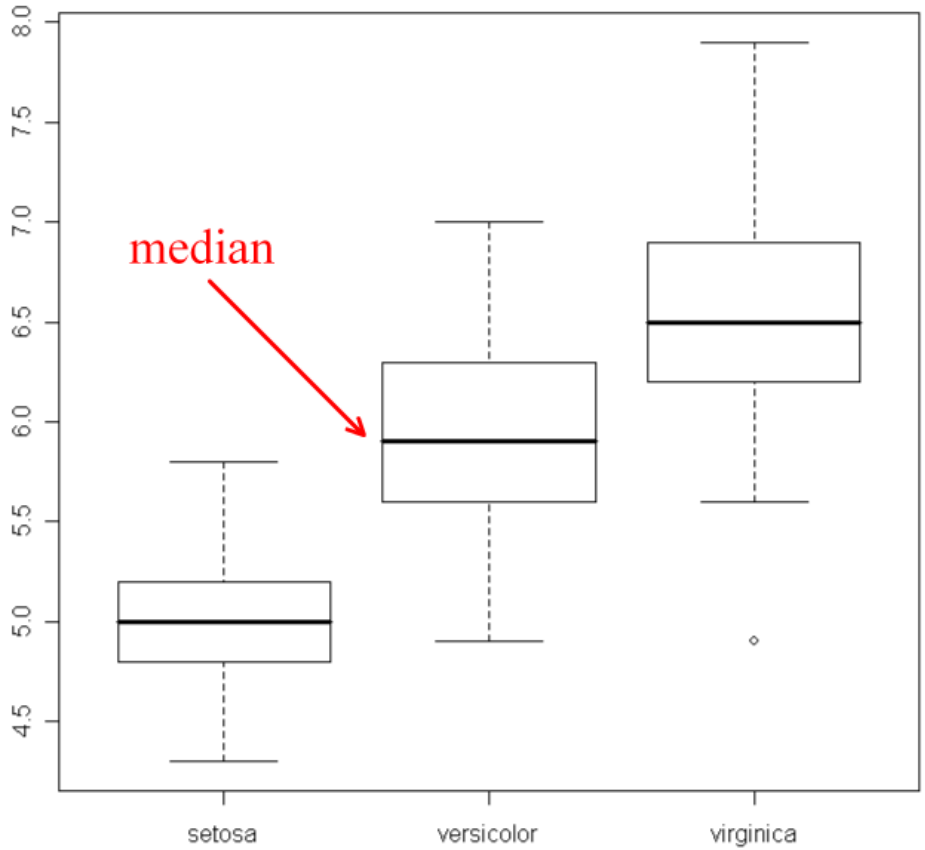
Quartiles: 25%-quantile (1st or lower quartile), median (2nd quartile), 75%-quantile (3rd or upper quartile).

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

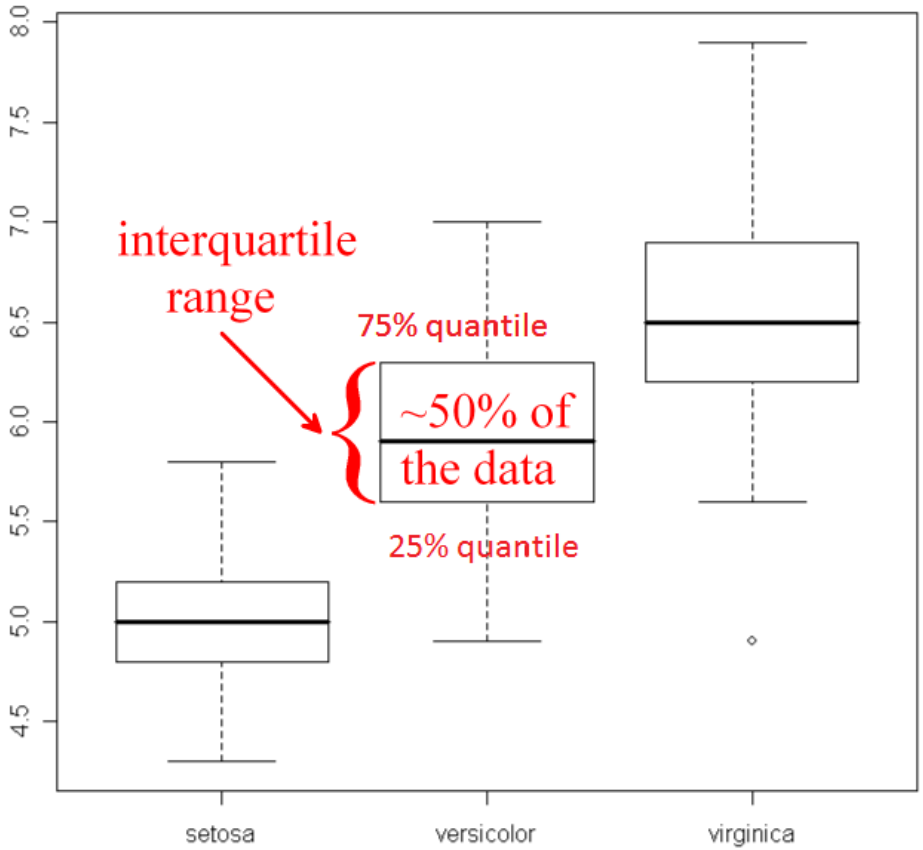
Boxplots



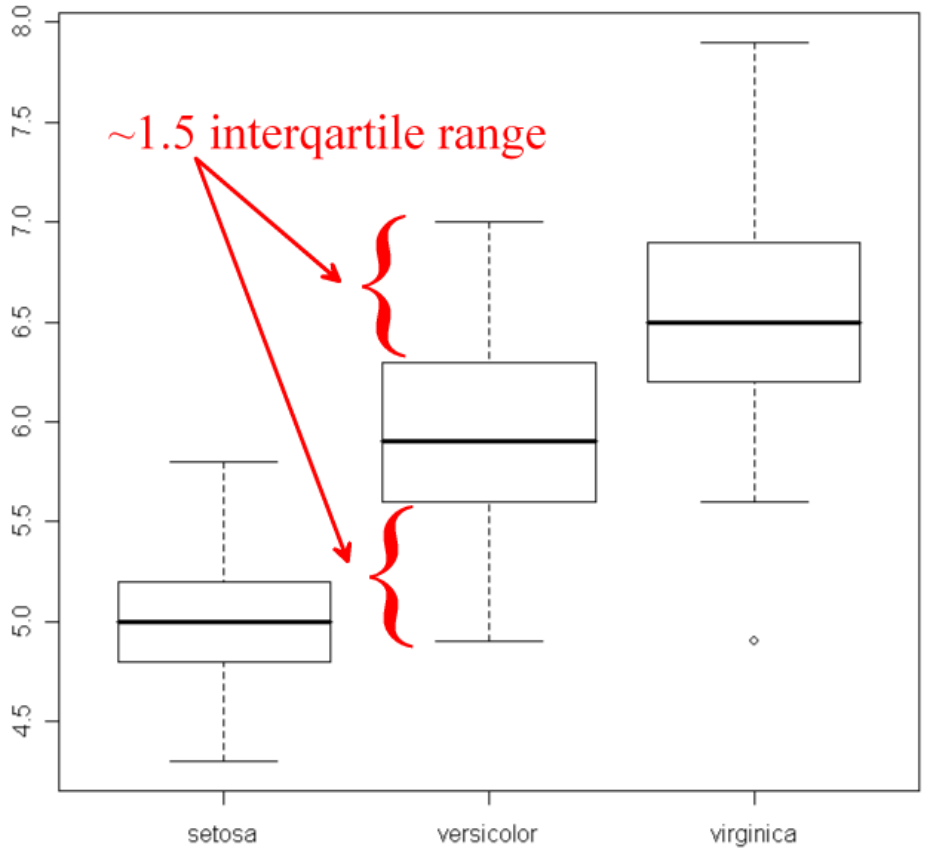
Boxplots



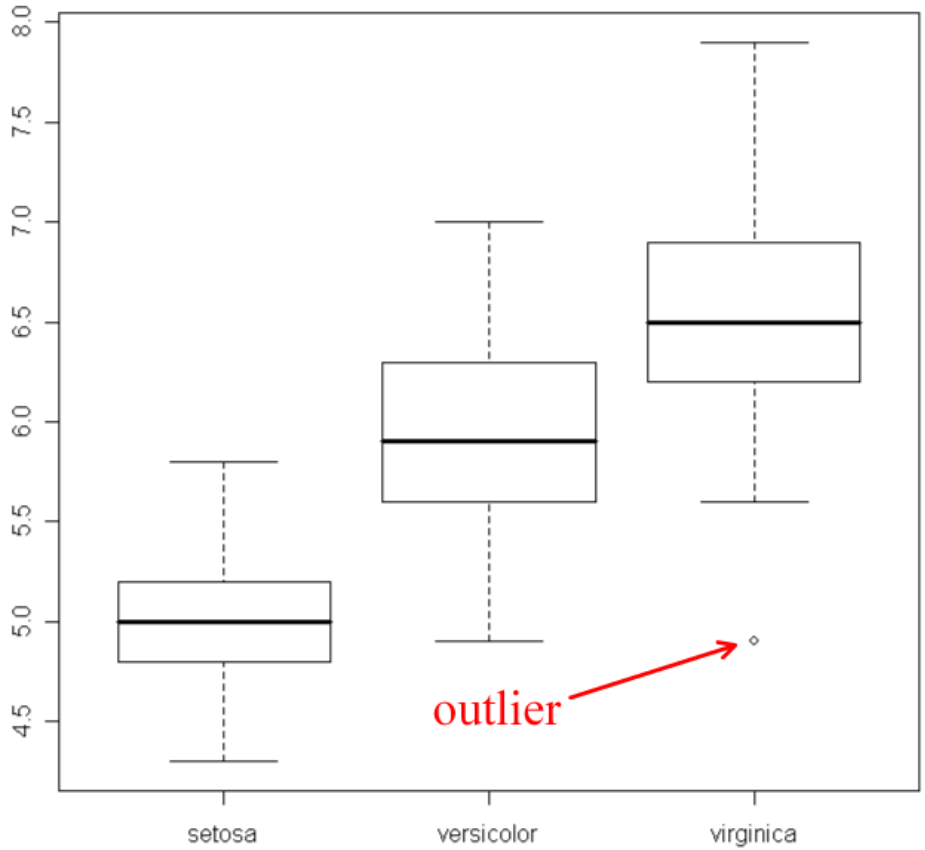
Boxplots


8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
interquartile
range
75% quantile
~50% of
the data
25% quantile
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

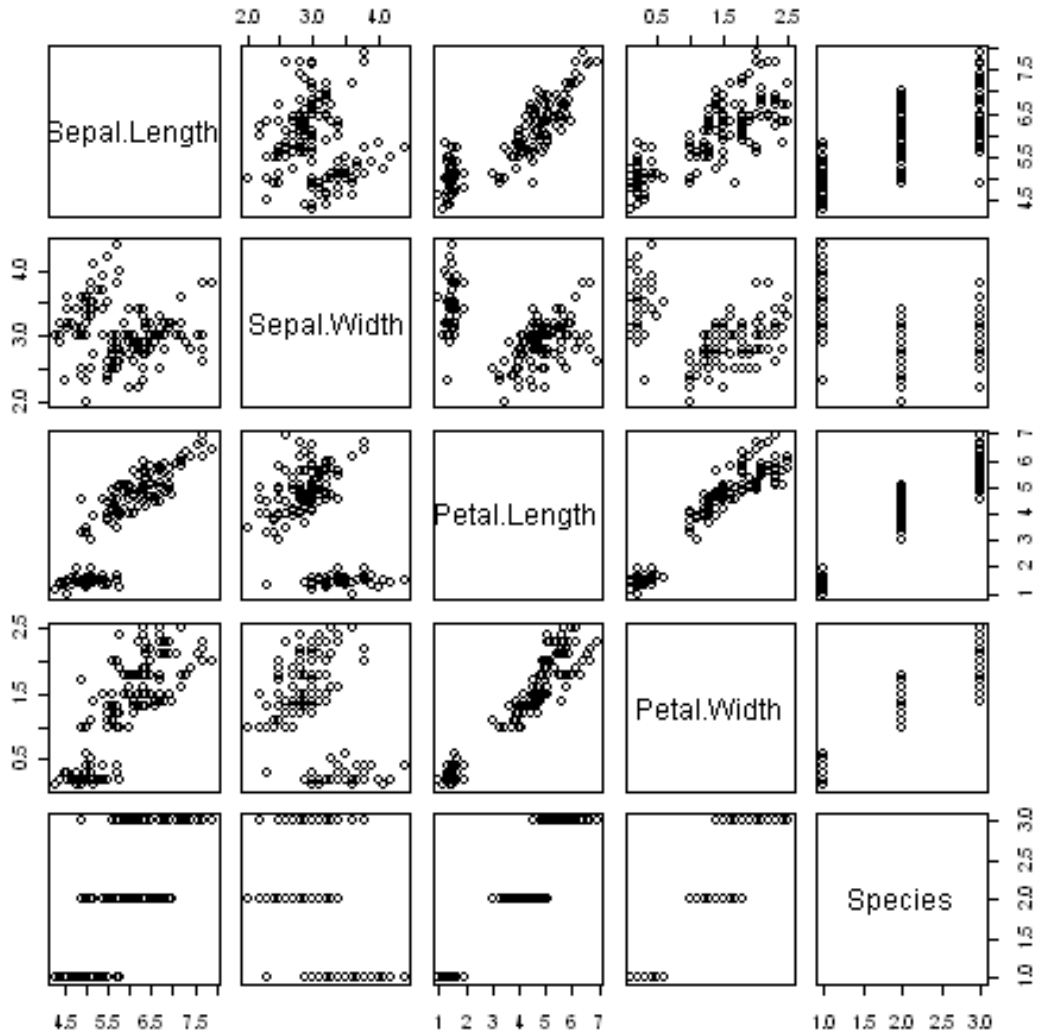


Boxplot construction

1. Determine the median. Draw a thick line at the position of the median.
2. Determine the 25%- and the 75%-quartiles q_1 and q_3 for the sample. Draw a box limited by these two quartiles. The other dimension of the box can be chosen arbitrarily.
3. iqr$= q_3 - q_1$ is the interquartile range. The inner fence is defined by the two values $f_1 = q_1 - 1.5 \cdot \text{iqr}$ and $f_3 = q_3 + 1.5 \cdot \text{iqr}$.
4. Find the smallest data point greater than f_1 and the largest data point smaller than f_3. Add "whiskers" to the box extending to these two data points.
5. Data points lying outside the box and the whiskers are called outliers. Enter these data points in the diagram, for instance by circles.
6. Sometimes, extreme outliers (out of the outer fence defined by $F_1 = q_1 - 2 \cdot 1.5 \cdot \text{iqr}$ and $F_3 = q_3 - 2 \cdot 1.5 \cdot \text{iqr}$) are drawn in a different way than mild outliers outside the whiskers, but within the inner fence.

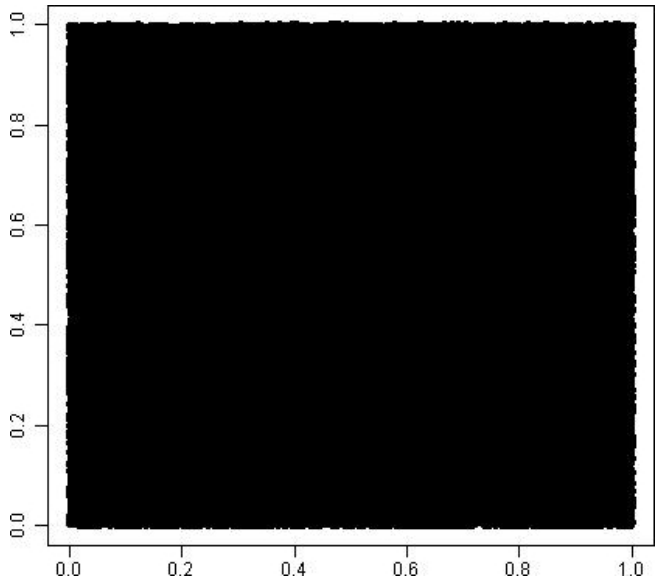
Scatter plots

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



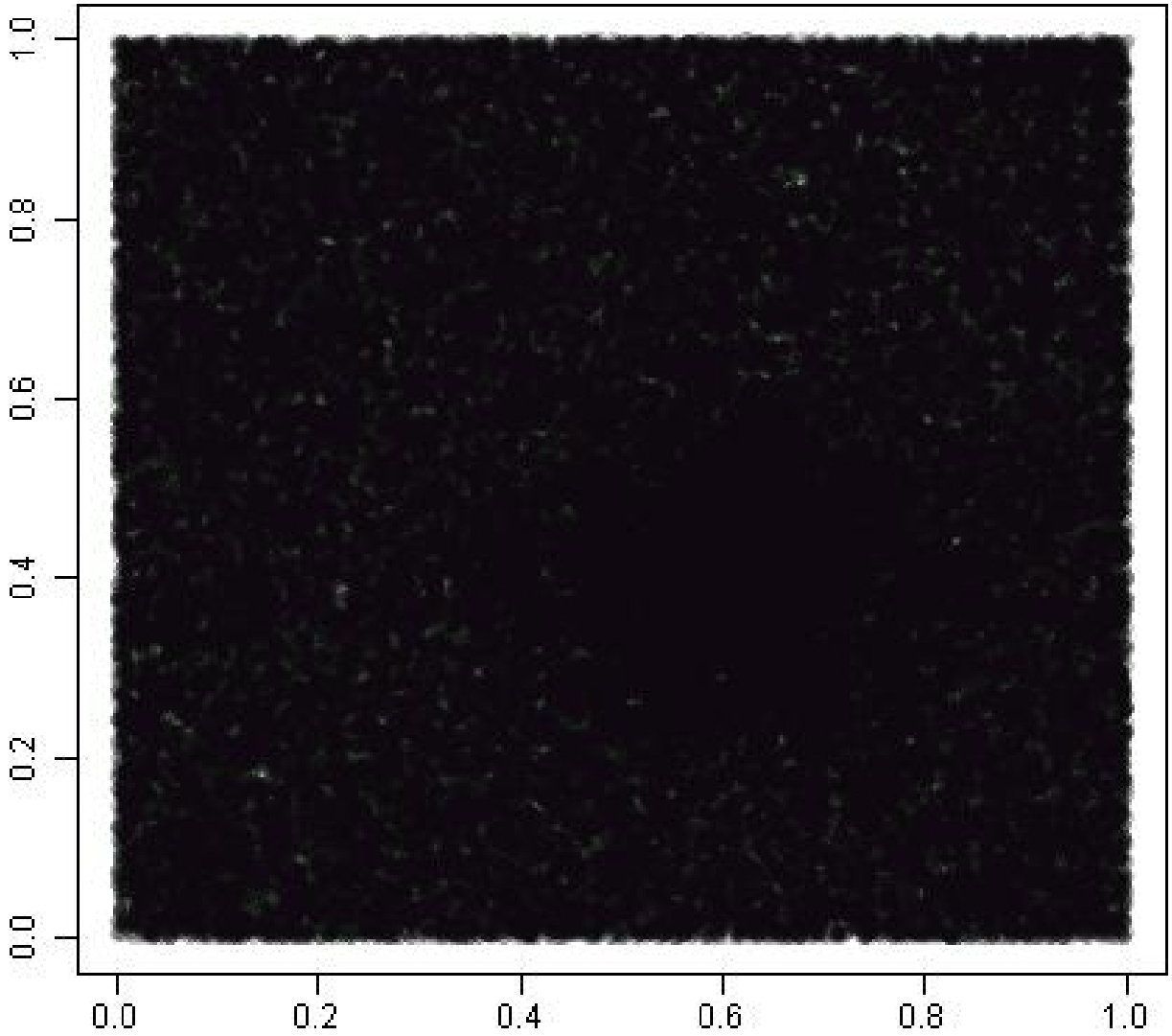
Scatter plots

For large data sets, points are plotted over each other.
Density information is lost.
In the worst case, all information is lost.



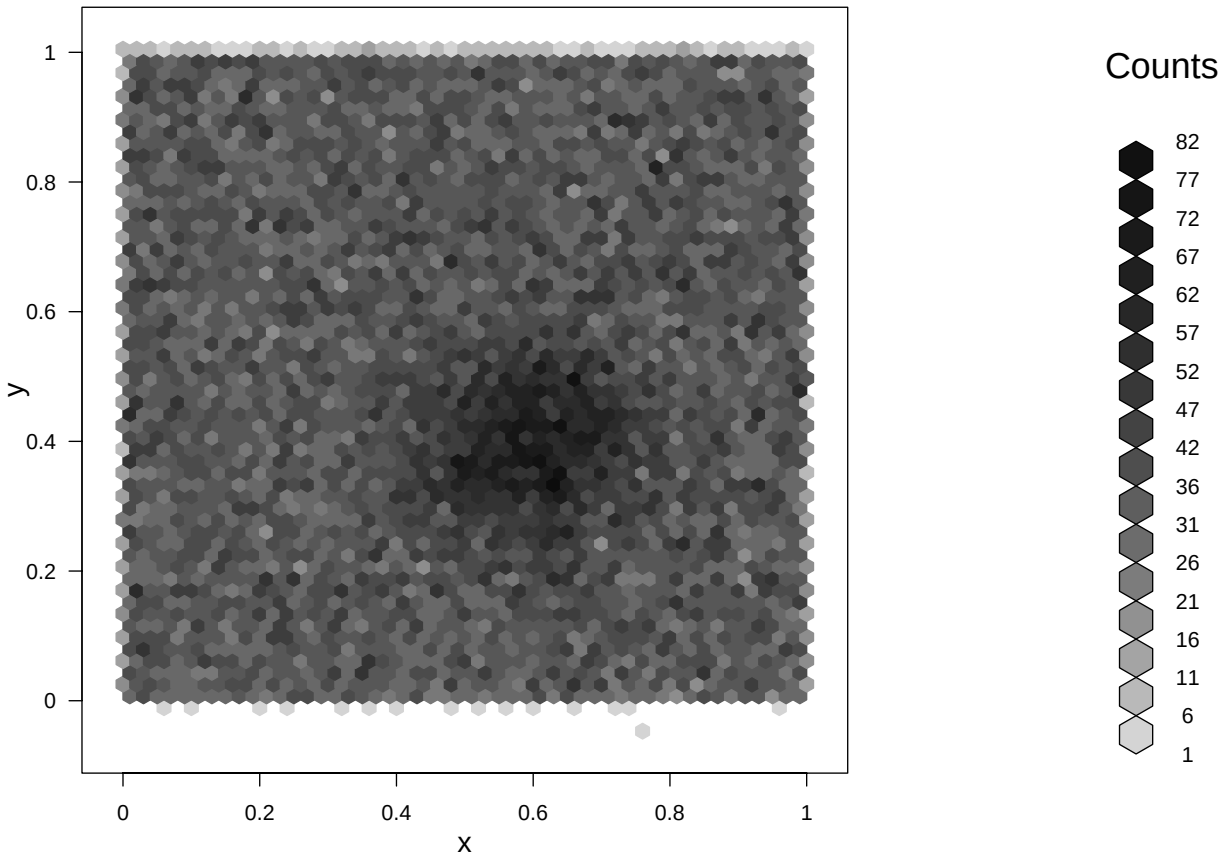
A scatter plot for a data set with 100000 objects.

Scatter plots



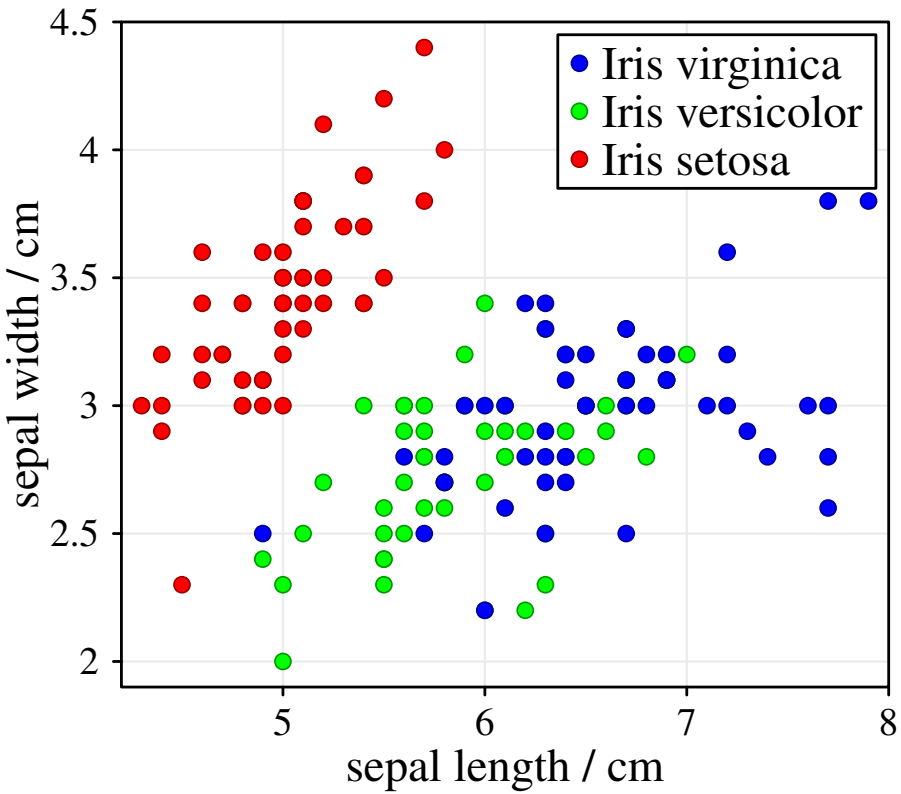
A density plot. Instead of solid points, semitransparent points are plotted.

Scatter plots



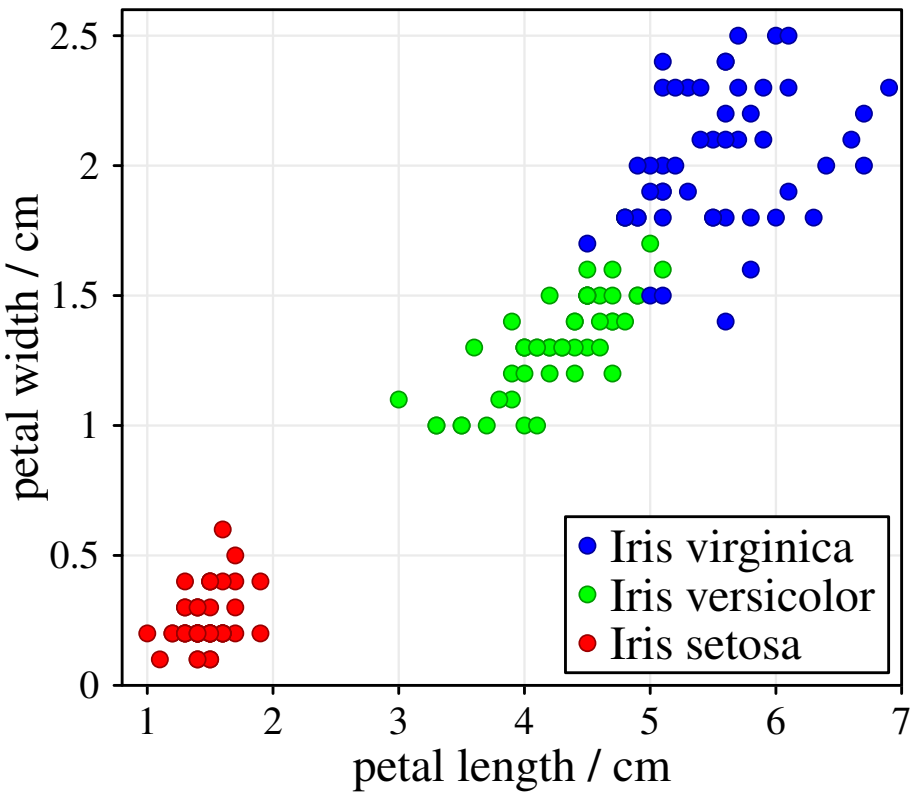
A scatter plot based on hexagonal binning. The grey intensity in each bin indicates the number of points falling into the bin.

Scatter plots



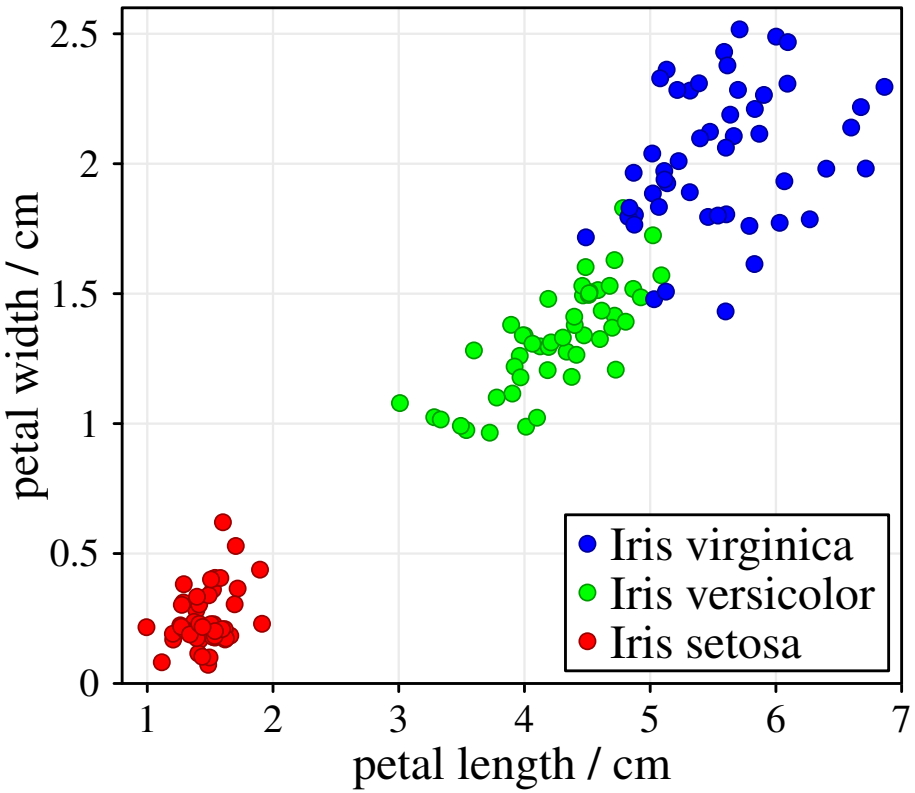
Scatter plots can be enriched with additional information: Colour or different symbols to incorporate a third attribute in the scatter plot.

Scatter plots



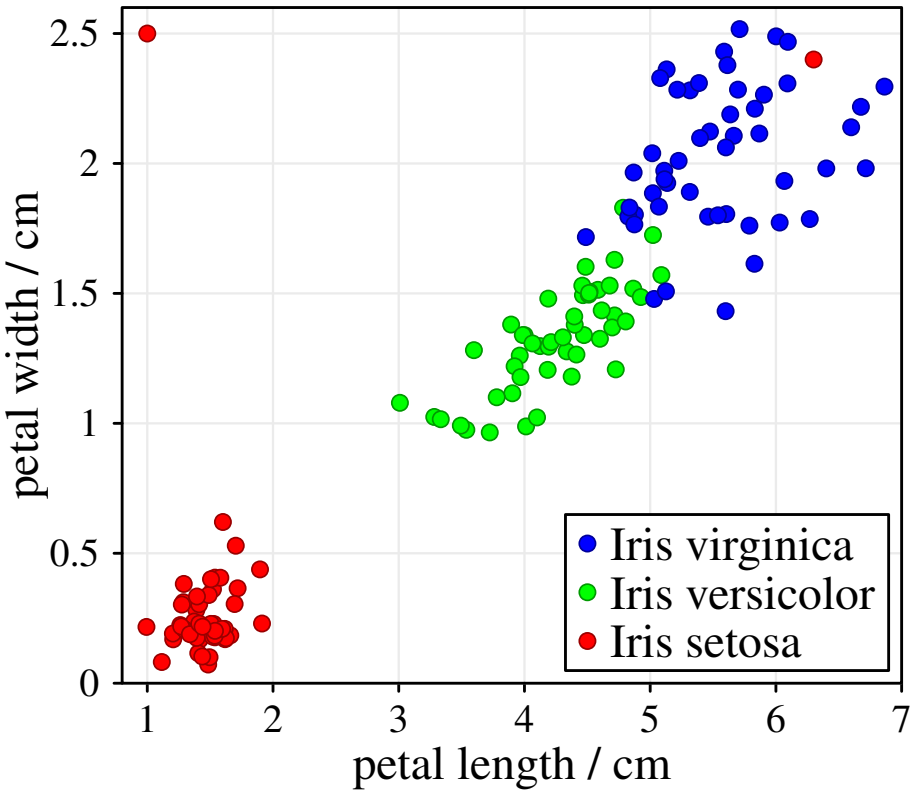
The two attributes petal length and width provide a better separation of the classes Iris versicolor and Iris virginica than the sepal length and width.

Scatter plots: jitter



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.

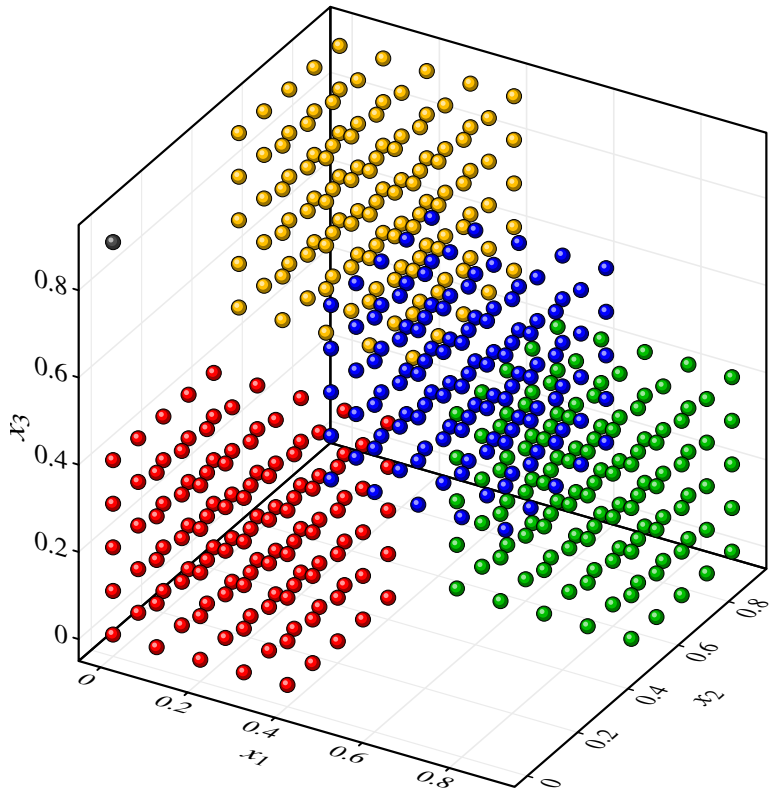
Scatter plots



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.

Outliers

In statistics, an outlier is an value or data object that is far away or very different from all or most of the other data.

Grubbs, F. E.: Procedures for detecting outlying observations in samples. Technometrics 11, 1-21, 1969:

- An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.

Outliers

Outliers may be indicative of data points that belong to a different population than the rest of the sample set (example from Wikipedia):

- One is calculating the average temperature of 10 objects in a room, and most are between 20 and 25 degrees Celsius, but an oven is at 175 °C.
- The median of the data may be 23 °C but the mean temperature will be about 40 °C.
- In this case, the median better reflects the temperature of a randomly sampled object than the mean.
- However, naively interpreting the median as "a typical sample", is incorrect.

Outlier detection

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
- On the other hand, instances that appear to be outliers might be even the most valuable part of your data
  - Mobile game analytics: "whales" rare players who spend much more time and money on a game, game companies are most interested in them

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.

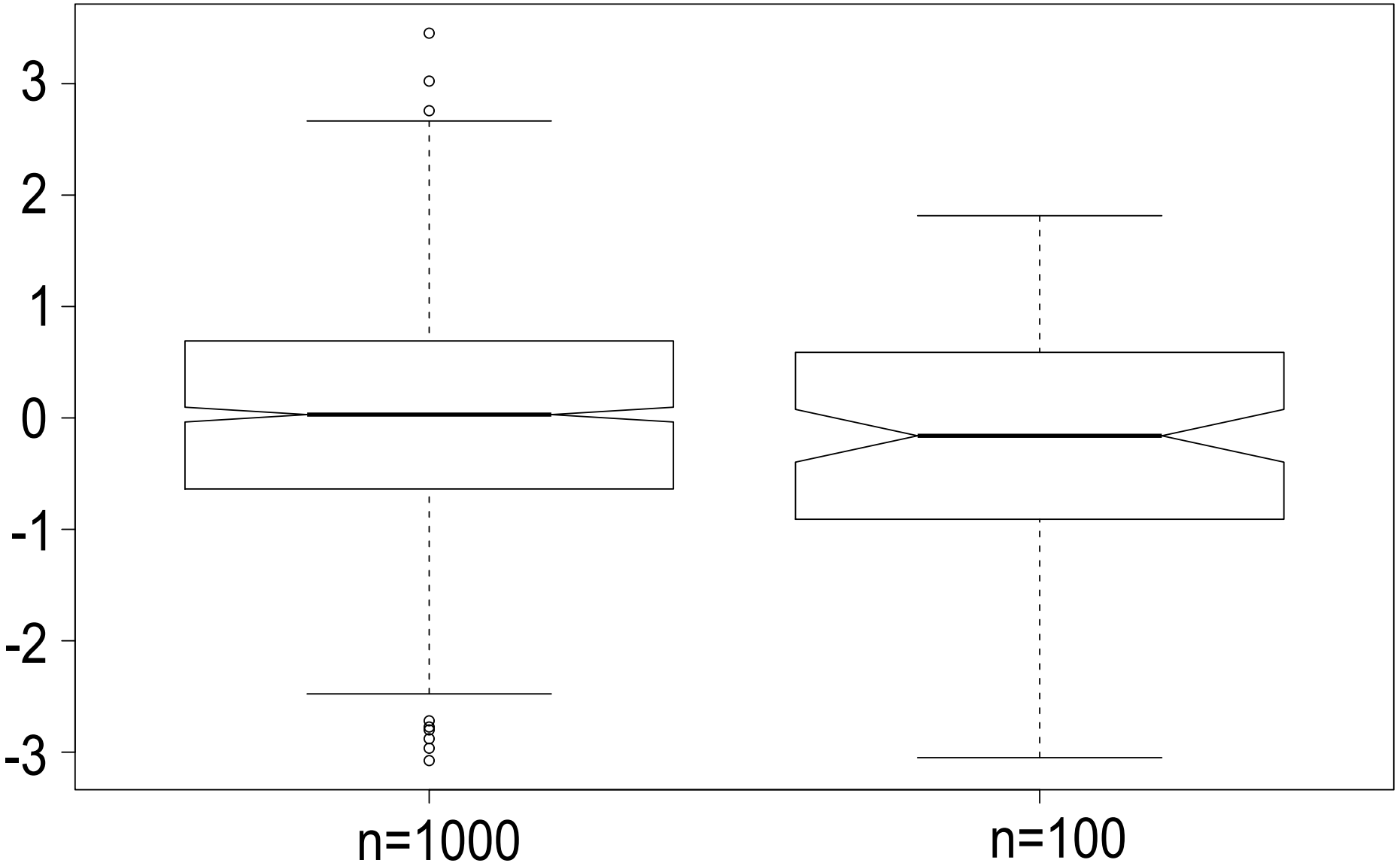
In some cases, the outliers can even be the target objects of the analysis.

Example: Automatic quality control system
Goal: Train a classifier, classifying the parts as correct or with failures based on measurements of the produced parts.
The frequency of the correct parts will be so high that the parts with failure might be considered as outliers.

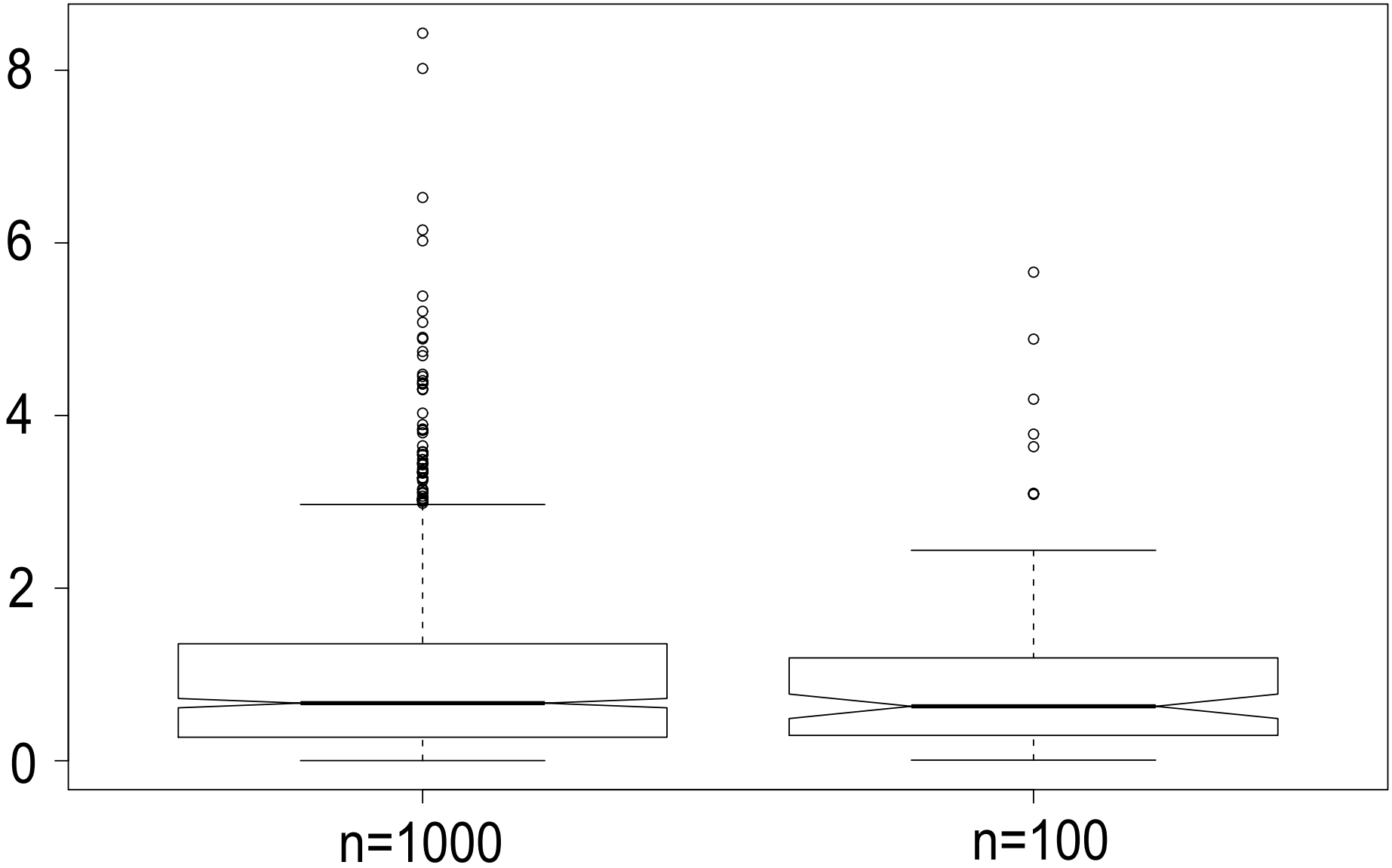
Outlier detection in single dimension

- Boxplots for (visually detecting) outliers
- Statistical tests like Grubb's test (see book).
- Tests tend to make assumptions about the underlying distribution, which may not hold
- As sample size increases the more you will see uncommon cases, that can appear as outliers

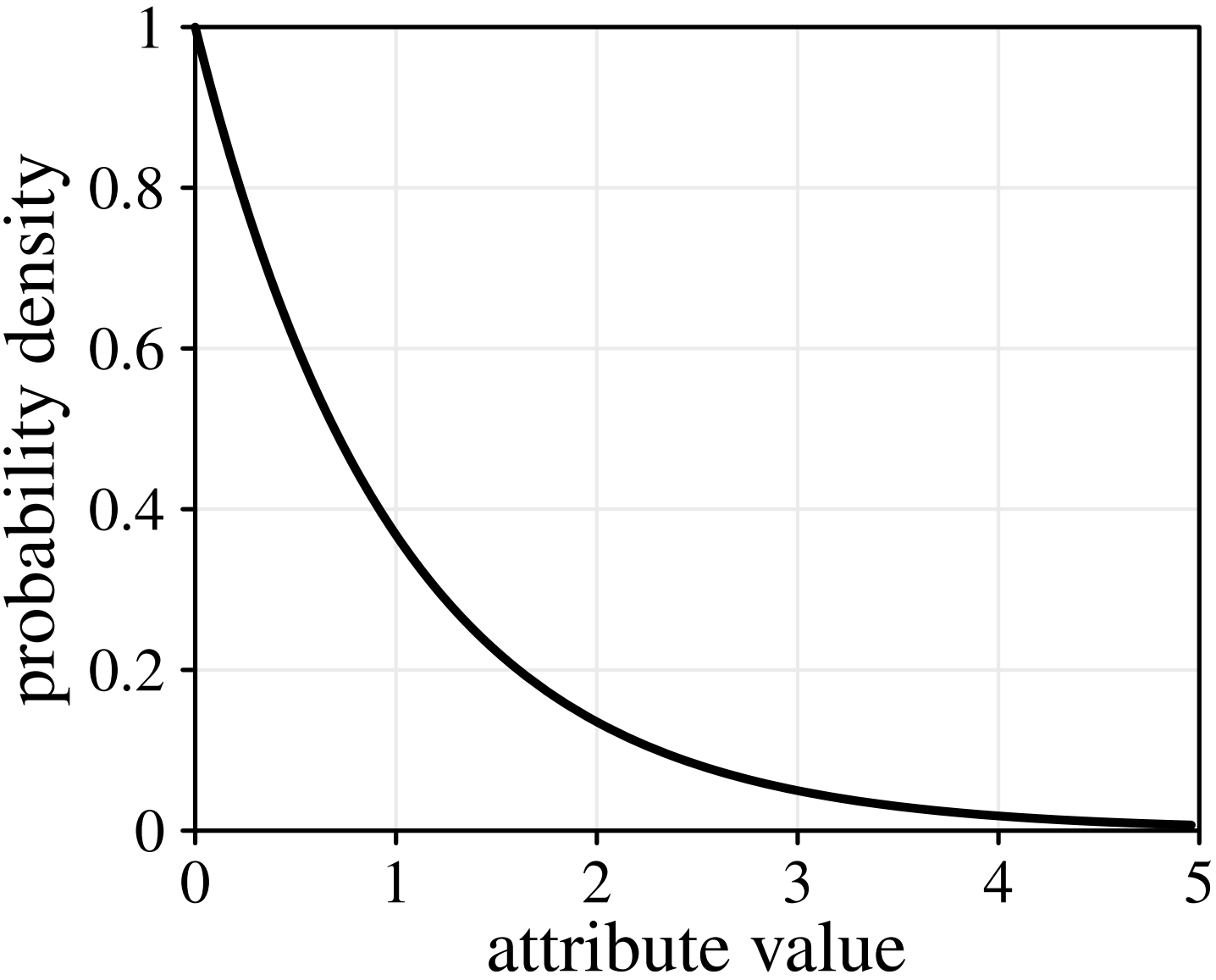
Outlier detection in single dimension



Outlier detection in single dimension



Outlier detection in single dimension



Outlier detection for multidimensional data

- Scatter plots for (visually detecting) outliers w.r.t. two attributes.
- PCA plots for (visually detecting) outliers.
- Cluster analysis techniques