

Data Understanding

Anna Monreale
Computer Science Department

Introduction to Data Mining, 2nd Edition
Chapter I & Data Exploration (Additional Resources)

Getting To Know Your Data

- For preparing data for data mining task it is essential to have an overall picture of your data
- Gain insight in your data
 - with respect to your project goals
 - and general to understand properties
- Find answers to the questions
 - What kind of attributes do we have?
 - How is the data quality?
 - Does a visualization helps?
 - Are attributes correlated?
 - What about outliers?
 - How are missing values handled?
 - Do we need to extract other attributes

Which is the type of data?

Types of data sets

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

What is Data?

- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
 - Object is also known as record, point, case, sample, entity, or instance

Objects

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a ‘term’ vector
 - Each term is a component (attribute) of the vector
 - The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

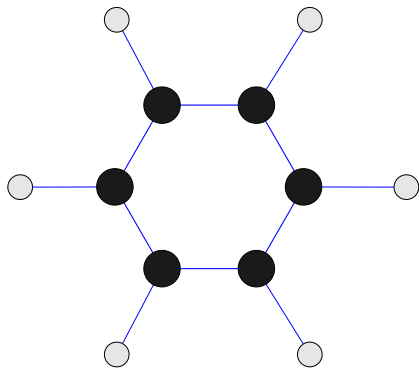
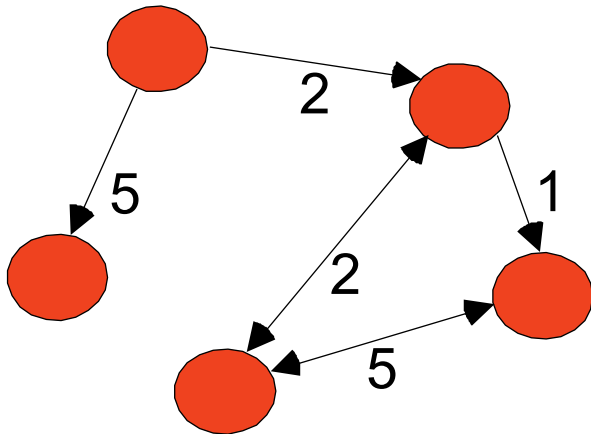
Transaction Data

- A special type of record data, where
 - Each record (transaction) involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- Examples: Generic graph, a molecule, and webpages



Useful Links:

- [Bibliography](#)
- Other Useful Web sites
 - [ACM SIGKDD](#)
 - [KDnuggets](#)
 - [The Data Mine](#)

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Benzene Molecule: C₆H₆

Ordered Data

- Sequences of transactions

Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)



An element of
the sequence

Retail data



Ordered Data

- Genomic sequence data

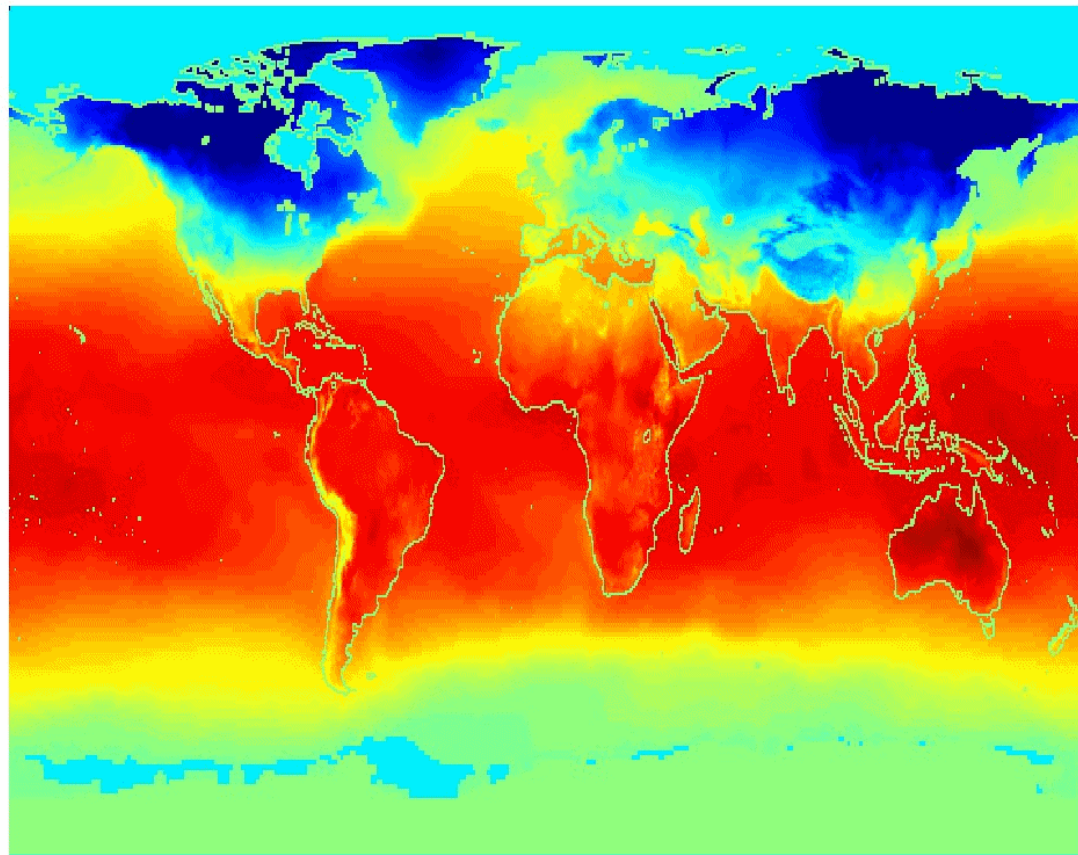
**GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCCGCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

Ordered Data

- Spatio-Temporal Data

**Average Monthly
Temperature of
land and ocean**

Jan



Types of Attributes

- There are different types of attributes
 - **Nominal/Categorical:** attribute values in a finite domain, categories, “name of things”
 - **Examples:** ID numbers, eye color, zip codes
 - **Binary:** Nominal attribute with only 2 states (0 and 1)
 - **Symmetric binary:** both outcomes equally important (e.g., gender)
 - **Asymmetric binary:** outcomes not equally important. (e.g., medical test positive vs. negative) The convention is to assign 1 to most important outcome (e.g., having cancer)
 - **Ordinal:** finite domain with a meaningful ordering on the domain
 - **Examples:** rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}

Types of Attributes

- **Numeric:** quantity (integer or real-valued)
 - **Interval-Scaled**
 - Measured on a scale of equal-sized units
 - Values have order
 - **Examples:** calendar dates, temperatures in Celsius
- **Ratio-Scaled:** We can speak of values as being an order of magnitude larger than the unit of measurement
 - **Examples:** length, counts, elapsed time (e.g., time to run a race)
 - A baseball game lasting 3 hours is **50% longer** than a game lasting **2 hours**.

Discrete and Continuous Attributes

- **Discrete Attribute**

- Has only a finite or countably infinite set of values
- **Examples:** zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: **binary attributes** are a special case of discrete attributes

- **Continuous Attribute**

- **Has real numbers** as attribute values
- **Examples:** temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

Properties of Attribute Values

The type of an attribute depends on which of the following properties/operations it possesses:

- Distinctness: $= \neq$
- Order: $< >$
- Differences are meaningful : $+ -$
- Ratios are meaningful $* /$
- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & meaningful differences
- Ratio attribute: all 4 properties/operations

Categorical
QualitativeNumeric
Quantitative

Attribute Type	Description	Examples	Operations
Nominal	Nominal attribute values only distinguish. ($=$, $\neq$)	zip codes, employee ID numbers, eye color, sex: $\{male, female\}$	mode, entropy, contingency correlation, χ^2 test
Ordinal	Ordinal attribute values also order objects. ($<$, $>$)	hardness of minerals, $\{good, better, best\}$, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, differences between values are meaningful. ($+$, $-$)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	For ratio variables, both differences and ratios are meaningful. ($*$, $/$)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Data Quality

- Poor data quality negatively affects many data processing efforts
- “The most important point is that poor data quality is an unfolding disaster.
- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”

Thomas C. Redman, DM Review, August 2004

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
 - Some credit-worthy candidates are denied loans
 - More loans are given to individuals that default

Data Quality ...

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
 - Wrong data
 - Duplicate data
 - Noise and outliers
 - Missing values

Data Quality issues ...

- **Syntactic accuracy:** Entry is not in the domain.
 - **Examples:** **f**male 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?

Duplicate Data

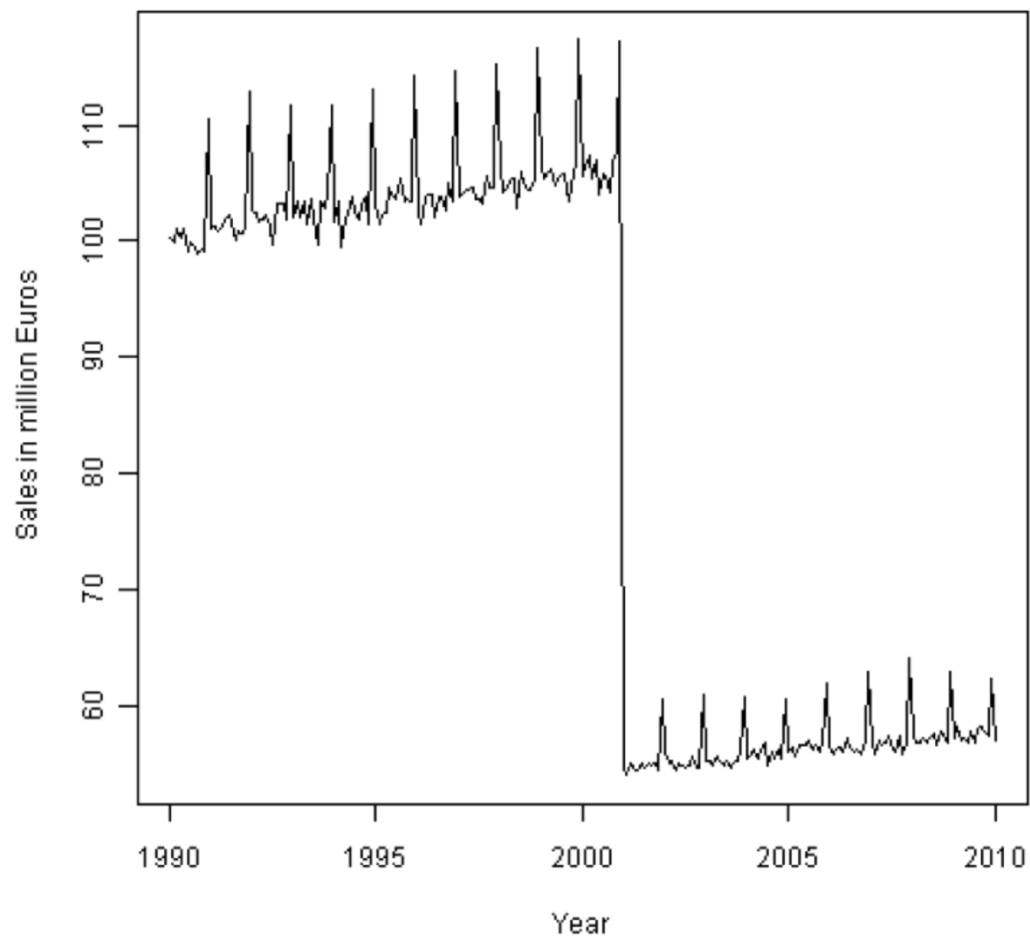
- Data set may include data objects that are duplicates, or almost duplicates of one another
 - Major issue when merging data from heterogeneous sources
- Examples:
 - Same person with multiple email addresses
- Data cleaning
 - Process of dealing with duplicate data issues

Statistics & Visualization

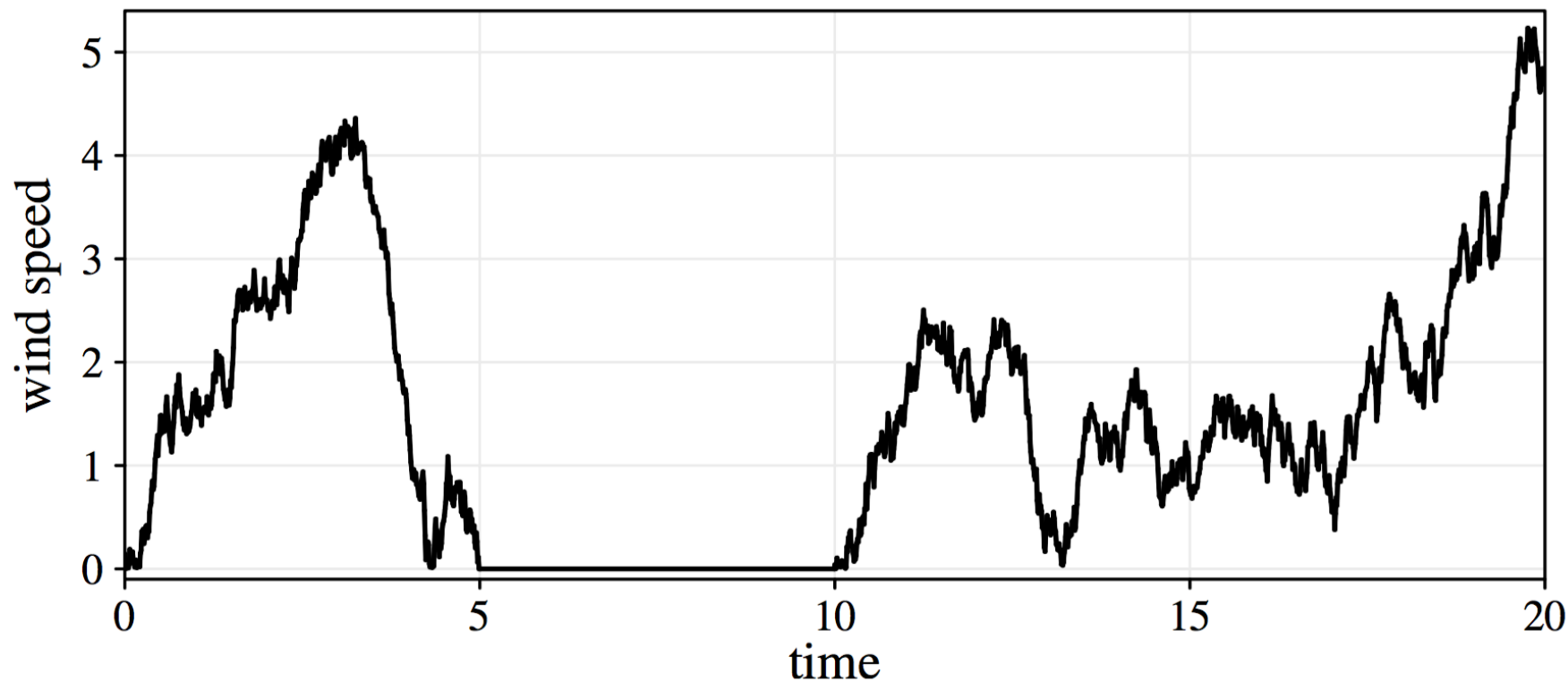
In order to know our data and discovery quality issues we need:

- Use descriptive statistics for getting a global picture and summarize properties of data
- Compare statistics with the expected behaviour
- Exploit visualization techniques that can help in detecting
 - general patterns and trends
 - outliers and unusual patterns

Data Visualization



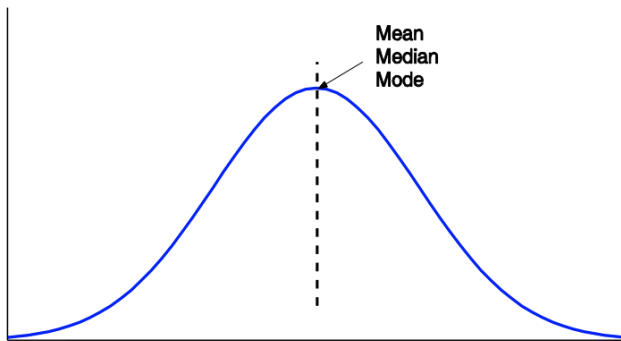
Data Visualization



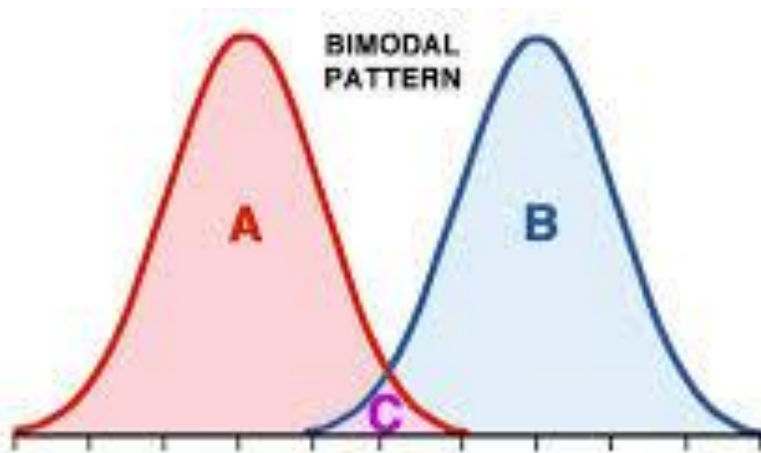
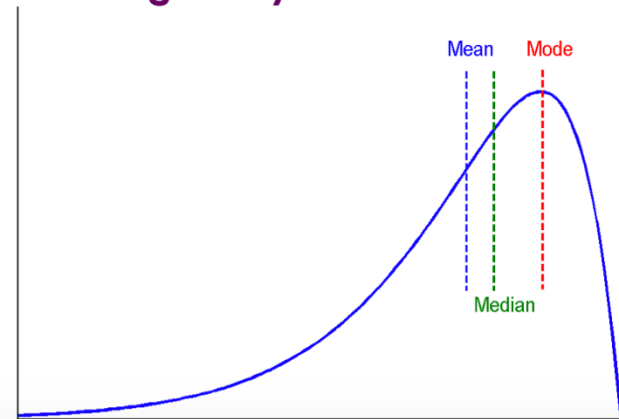
The zero values might come from a broken or blocked sensor and might be considered as missing values.

Observing Data Distribution

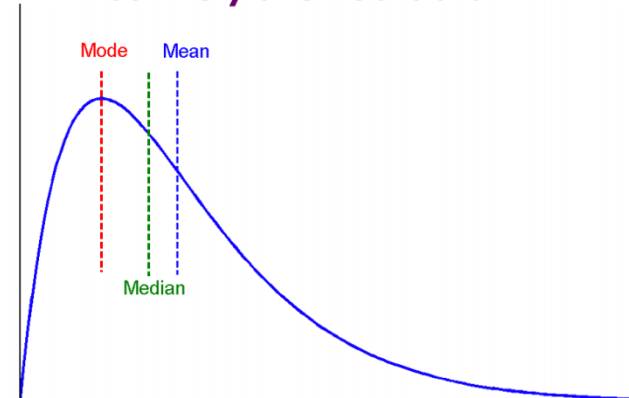
Symmetric data



Negatively skewed data



Positively skewed data



Example

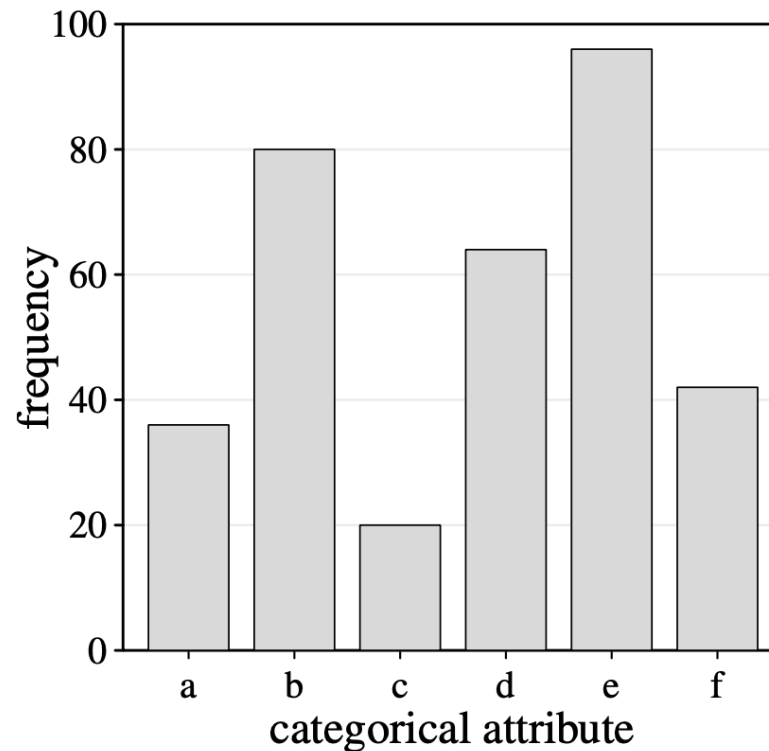
Give an example of something having a positively skewed distribution

- **income** is a good example of a positively skewed variable: there will be a few people with extremely high incomes, but most people will have incomes bunched together below the mean.

Give an example of something having a bimodal distribution

- bimodal distribution has some kind of underlying binary variable that will result in a separate mean for each value of this variable.
- One example can be **human weight** – the gender is binary and is a statistically significant indicator of how heavy a person is.

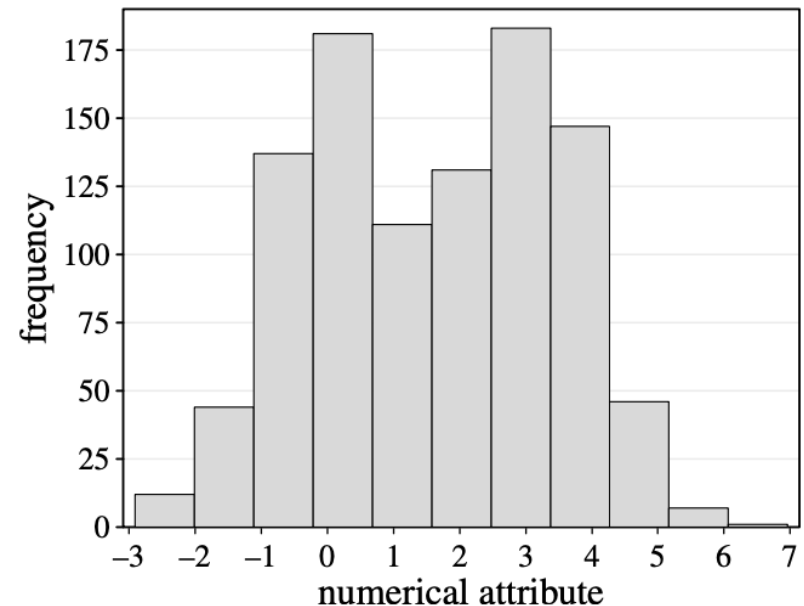
Bar Chart for Categorical Attributes



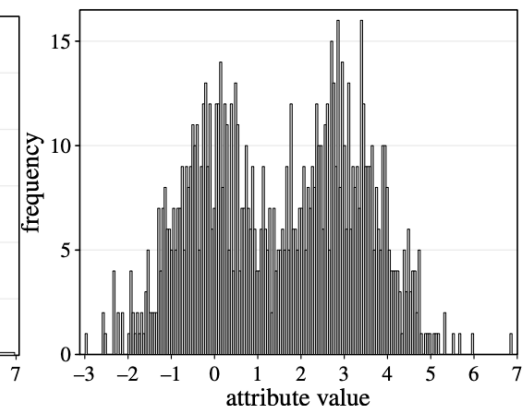
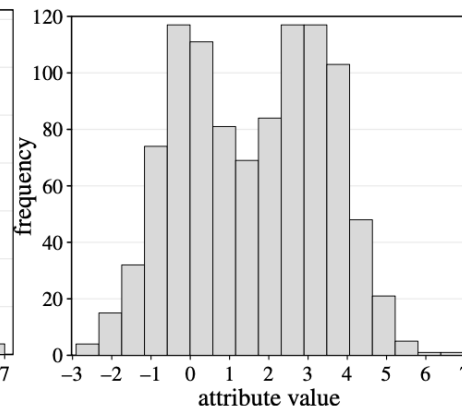
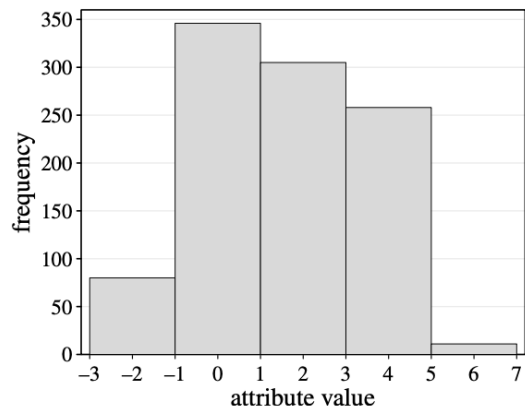
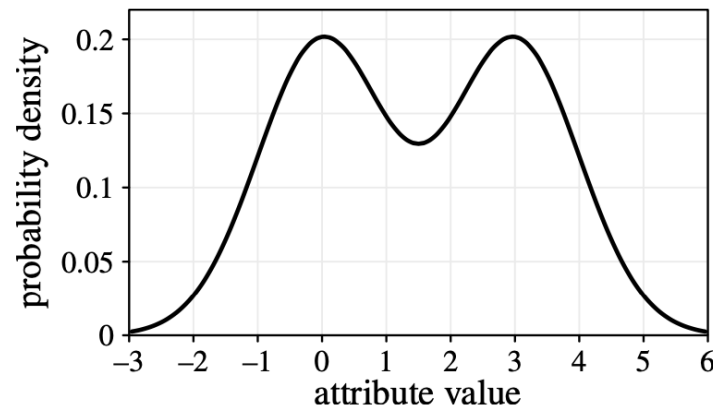
A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.

Histograms for Numerical Attributes

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



Histograms: Number of bins



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

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.

Measuring the Central Tendency

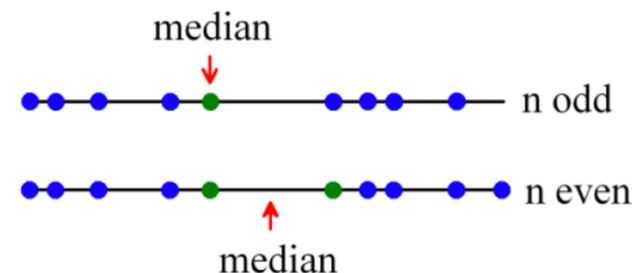
- Mean

- m is the sample size
- A distributive measure can be computed by partitioning the data into smaller subsets
- However, the mean is very sensitive to outliers
- The median or a trimmed mean are also commonly used

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

- Median

- Middle value if odd number of values, or average of the middle two values otherwise



- Mode

- Value that occurs **most frequently** in the data
- It is possible that several different values have the greatest frequency: Unimodal, bimodal, trimodal, multimodal
- If each data value occurs only once then **there is no mode**

Measuring the Dispersion of Data

- The degree in which data tend to spread is called the **dispersion**, or **variance** of the data
- The most common measures for data dispersion are **range**, **standard deviation**, the **five-number summary** (based on quartiles), and the **inter-quartile range**
- **Range**: The distance between the largest and the smallest values

Measuring the Dispersion of Data

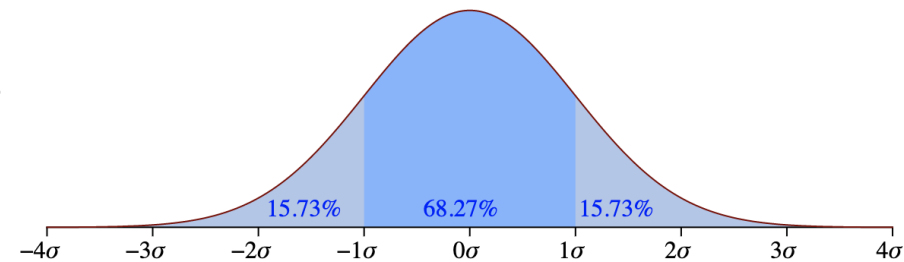
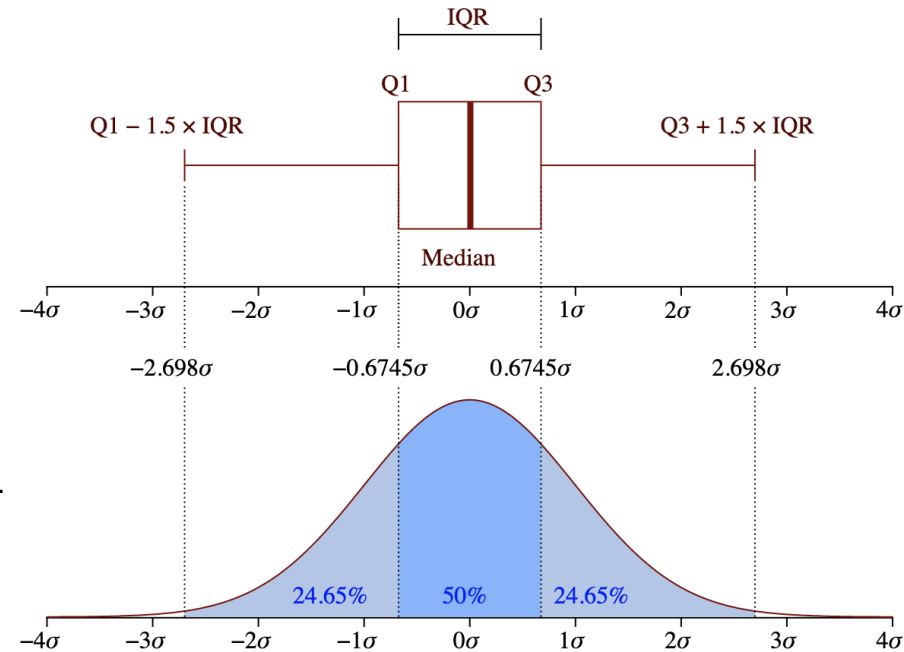
- **Variance** $\text{variance}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$
- **Standard deviation** σ is the square root of variance σ^2
 - σ measures spread about the mean and should be used only when the mean is chosen as the measure of the center
 - $\sigma=0$ only when there is no spread, that is, when all observations have the same value. Otherwise $\sigma>0$
- **Because of outliers**, other measures are often used:
 - **absolute average deviation (AAD)**
 - **median average deviation (MAD)**

$$\text{AAD}(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}|$$

$$\text{MAD}(x) = \text{median}\left(\{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\}\right)$$

Box Plot: Five-number summary of a distribution

- Data represented with a **box**
- **Whiskers**: two lines outside the box extended to **Minimum** and **Maximum**
- The ends of the box are at the
 - **1st quartiles** (25%-quantile)
 - **3rd quartiles** (75%-quantile)
- **Median**: value in the middle (values in increasing order) is the **2nd quartile** (50%-quantile)
- The height of the box is **Interquartile range (IQR)**: 3rd quartile - 1st quartile
- **p%-quantile** ($0 < p < 100$): The value x s.t. $p\%$ of the values are smaller and $100-p\%$ are larger.
- **Outliers**: points beyond whiskers



Example data set: Iris data



iris setosa



iris versicolor



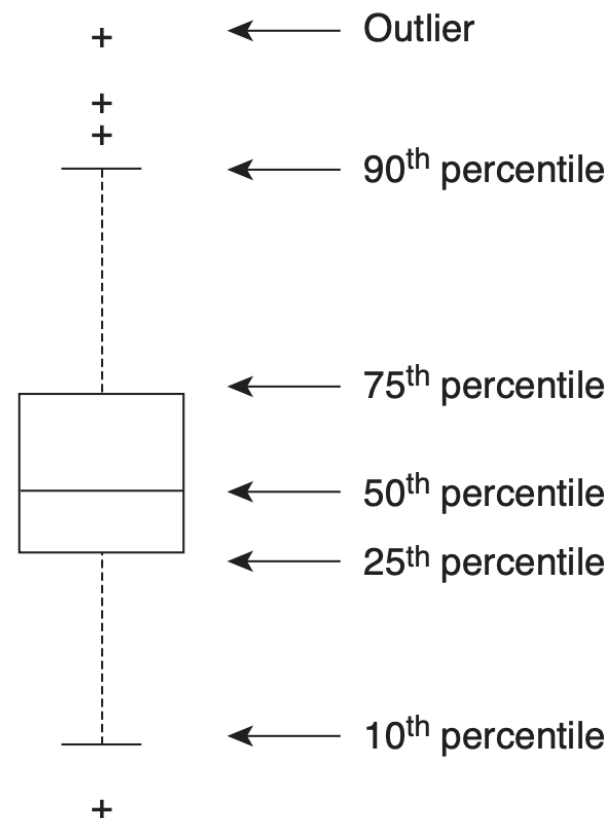
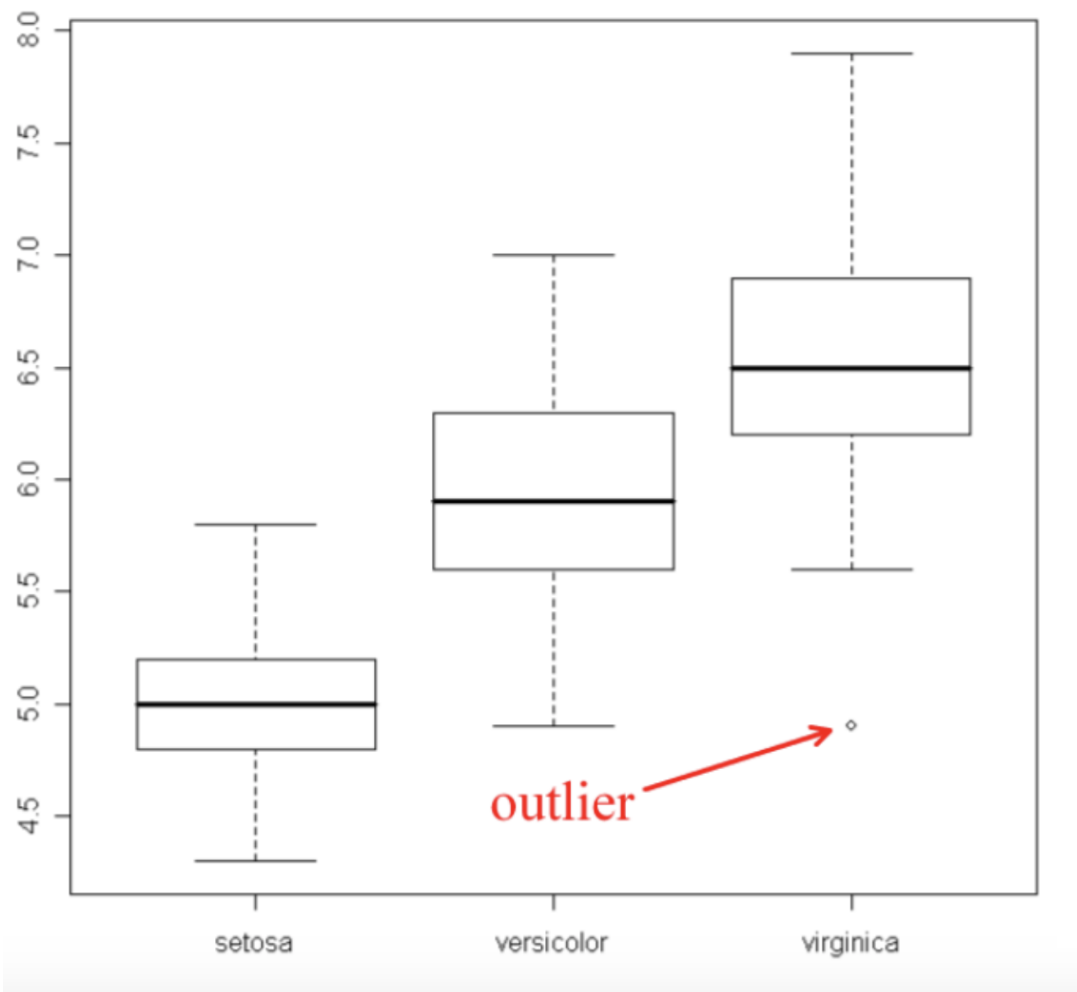
iris virginica

- 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

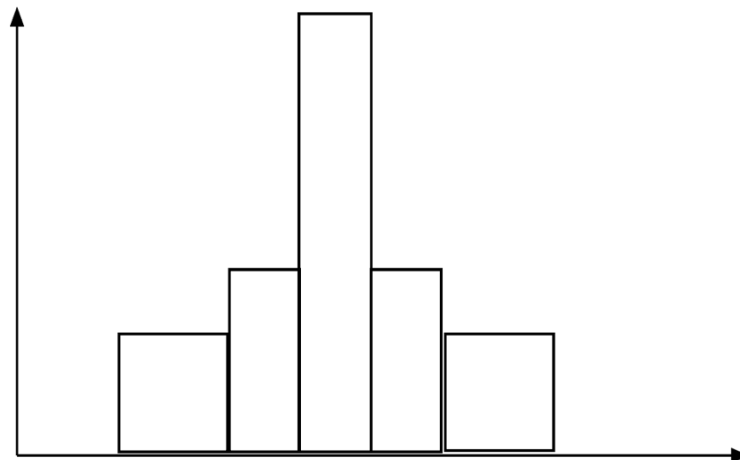
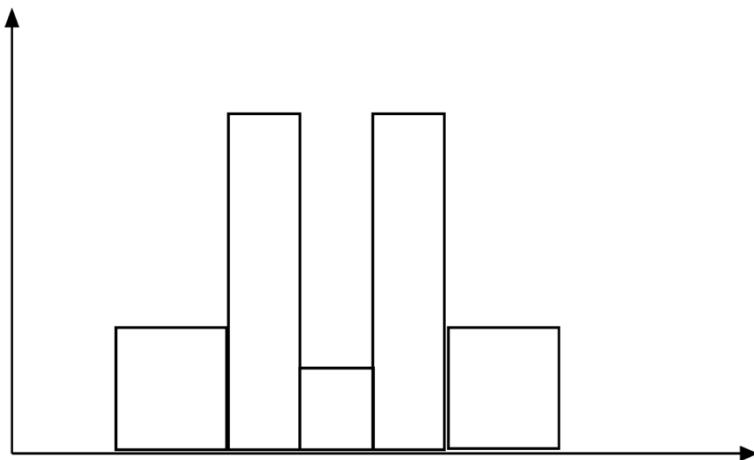
Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	Iris-setosa
...				
...				
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
...				
...				
5.1	2.5	3.0	1.1	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
...				
...				
5.9	3.0	5.1	1.8	Iris-virginica

Example of Conditional Box Plot



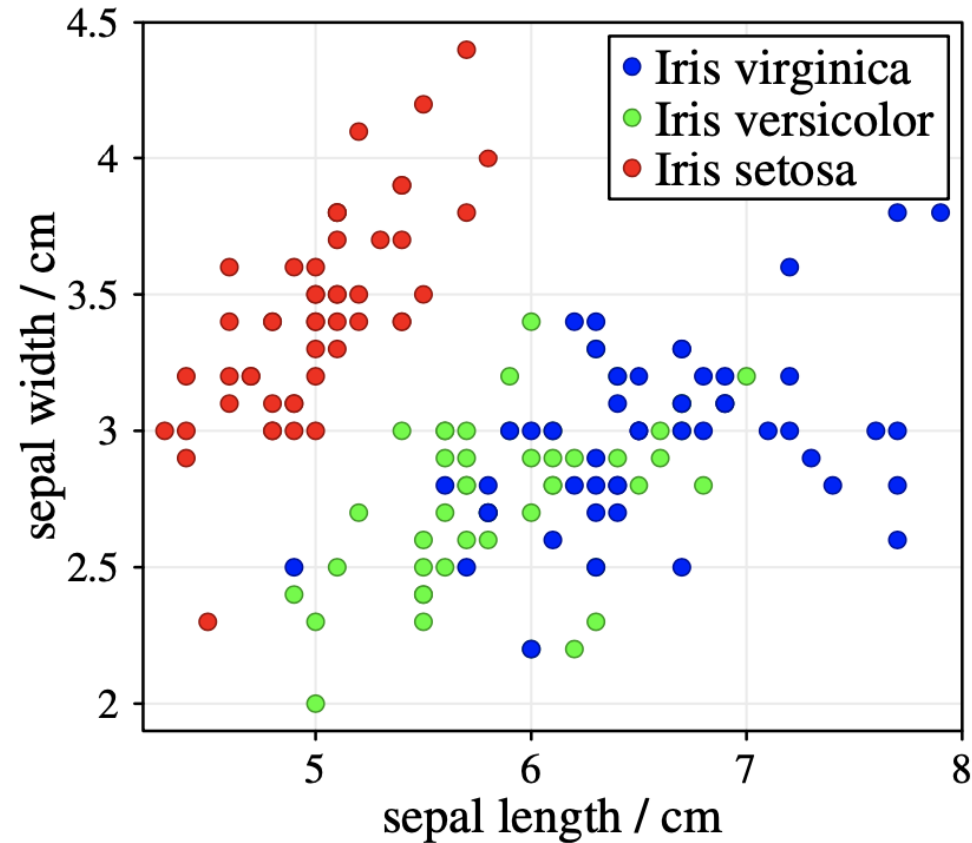
Histograms Often Tell More than Boxplots

- The two histograms may have the same boxplot representation
 - **The same values** for: **min, Q1, median, Q3, max**
 - But they have rather **different data distributions**



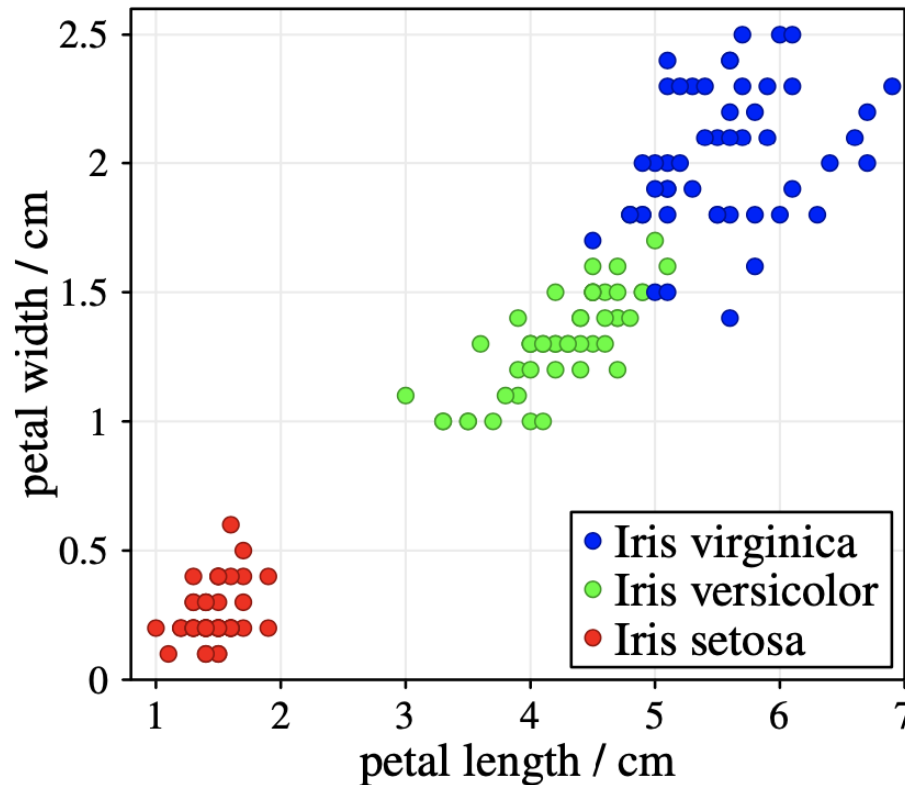
Scatter Plot

- Provides a first look at **bivariate data** to see **clusters** of points, **outliers**, **correlations**
- Each pair of values is treated as a **pair of coordinates** and plotted as points in the plane



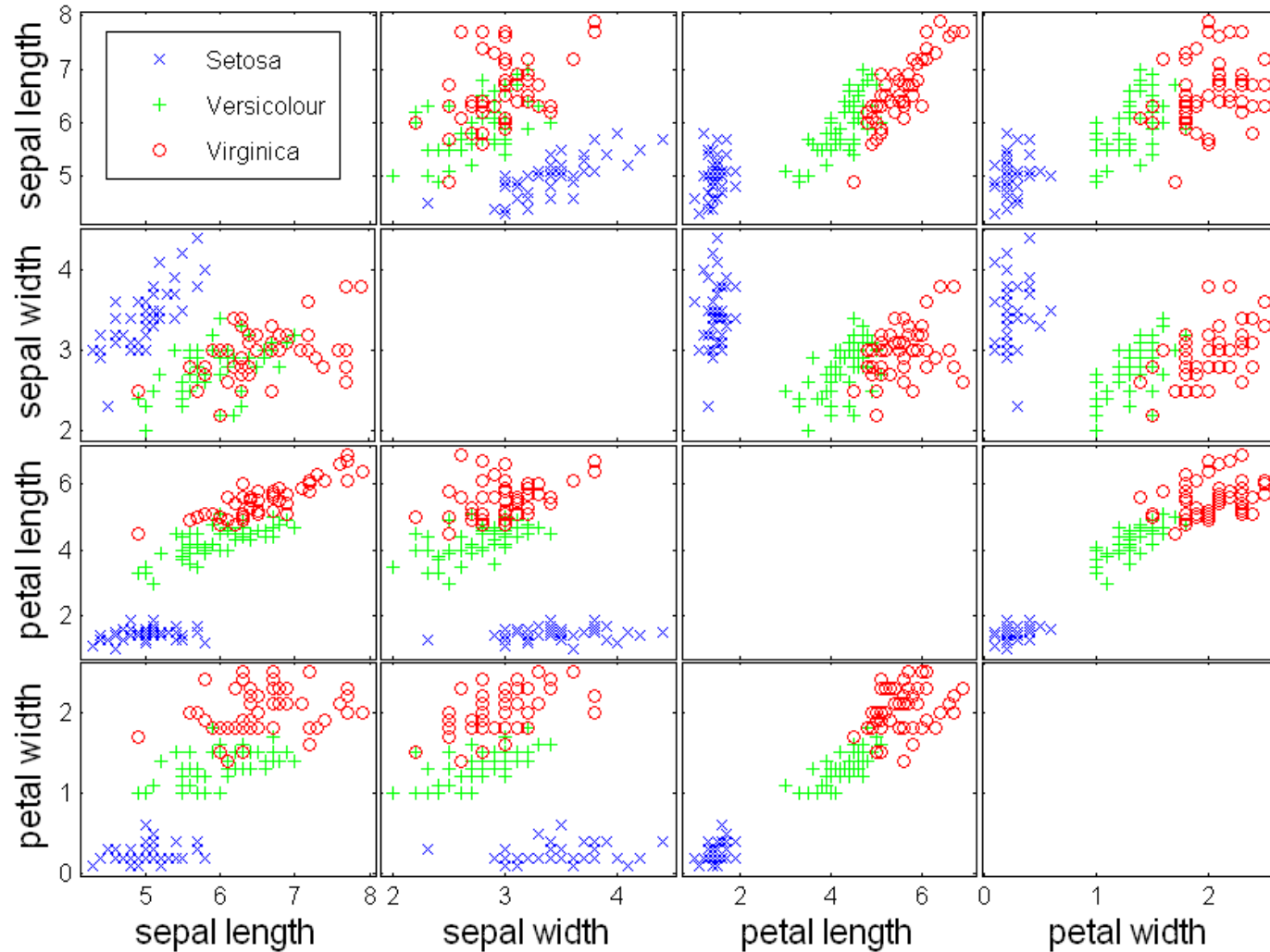
Scatter plots can be enriched with additional information: **Colour** or **different symbols** to **incorporate a third attribute** in the scatter plot.

Scatter Plot



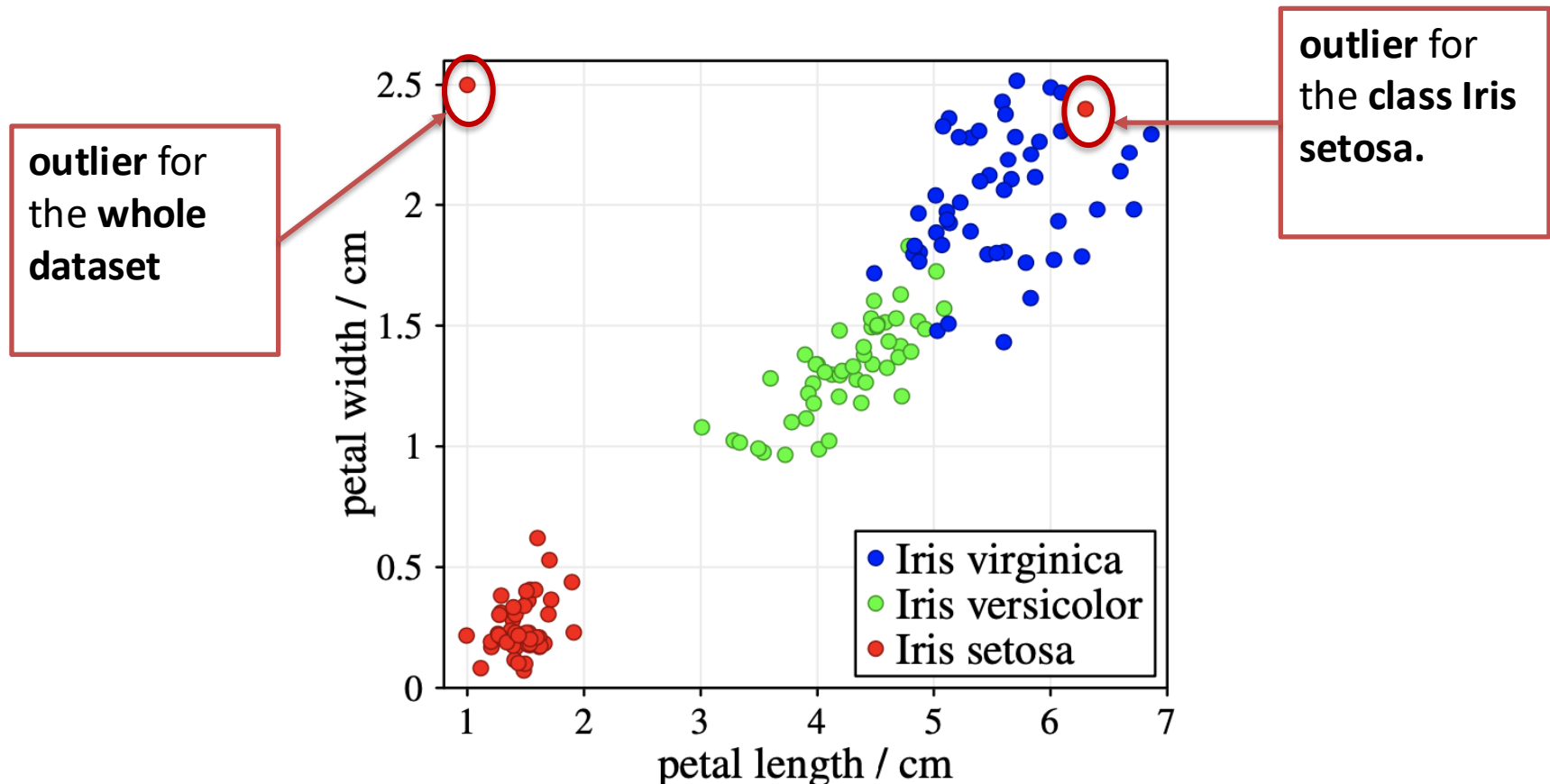
The two attributes petal length and width provide a **better separation of the classes** Iris versicolor and Iris virginica than the sepal length and width.

Scatter Matrix of Iris Attributes



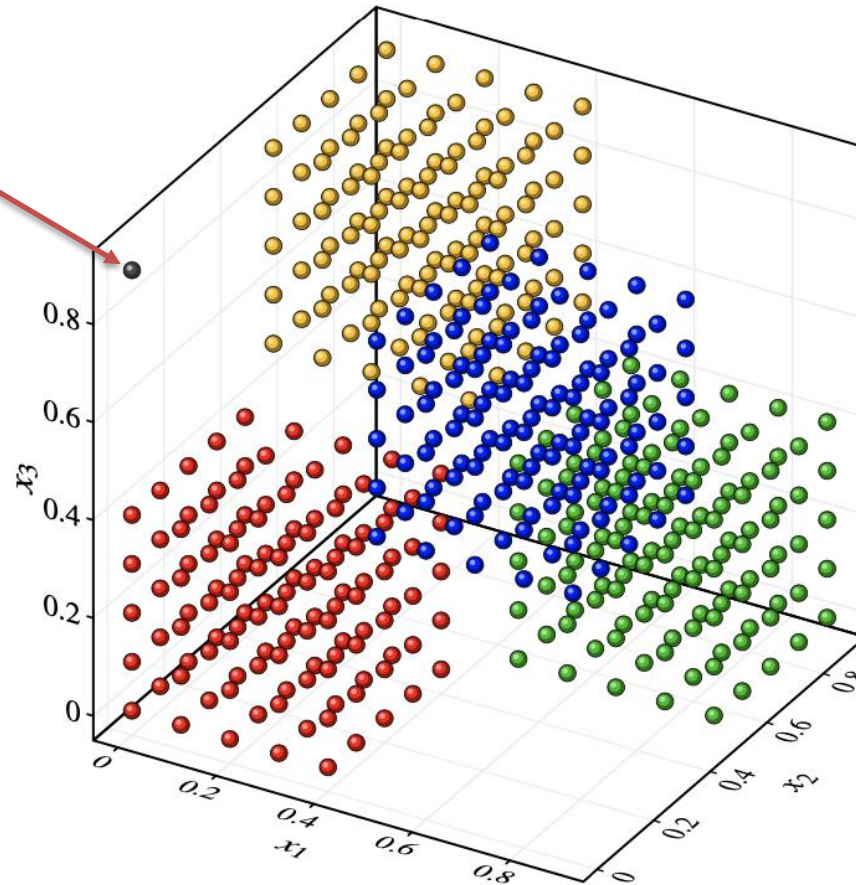
Scatter Plot & Outliers

The Iris data set with two (additional artificial) outliers



3D Scatter Plot

Outlier



Visualization 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.

Parallel Coordinates

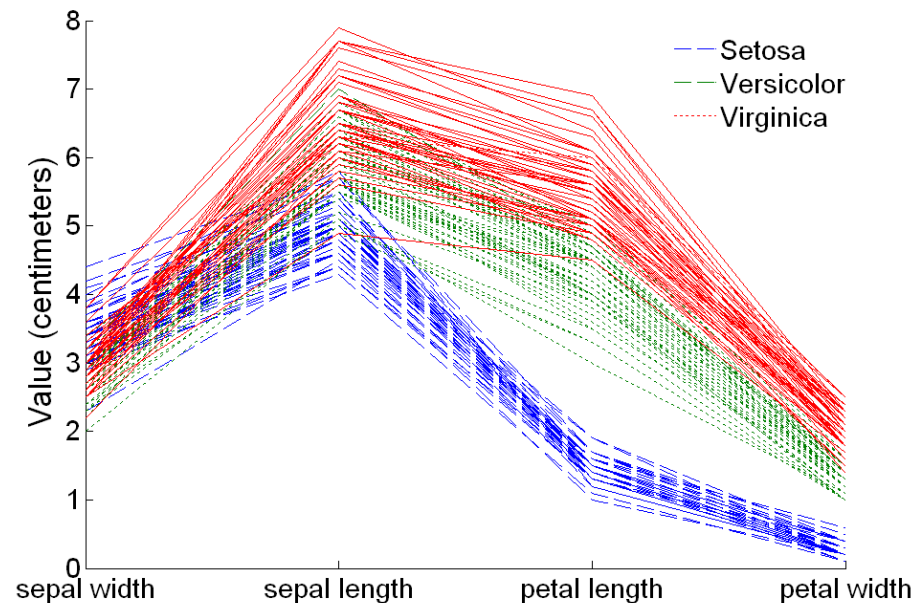
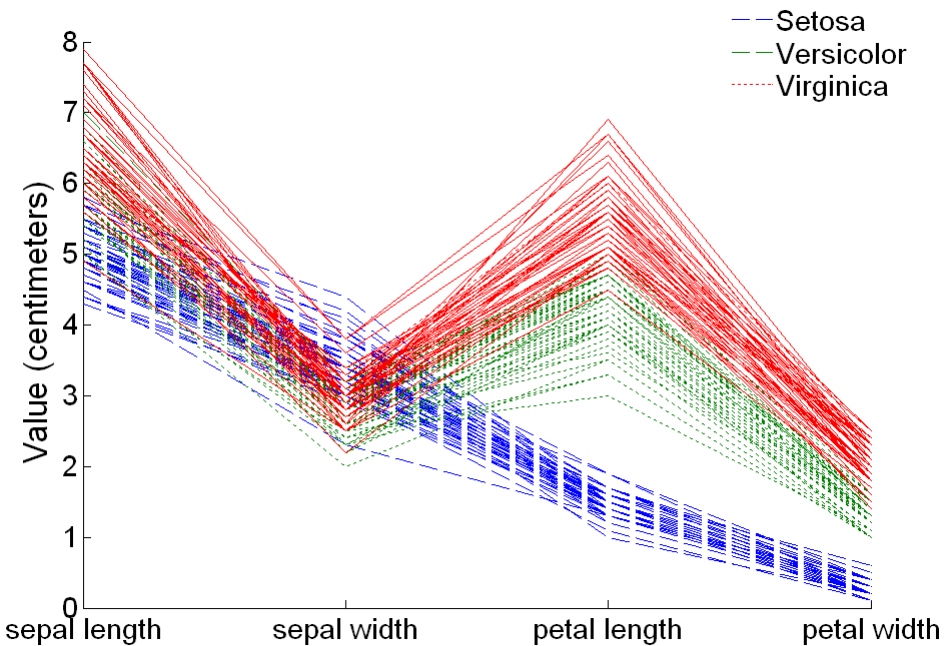
- **Parallel Coordinates**

- Used to plot the attribute values of high-dimensional data
- Instead of using perpendicular axes, use a set of parallel axes

The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line

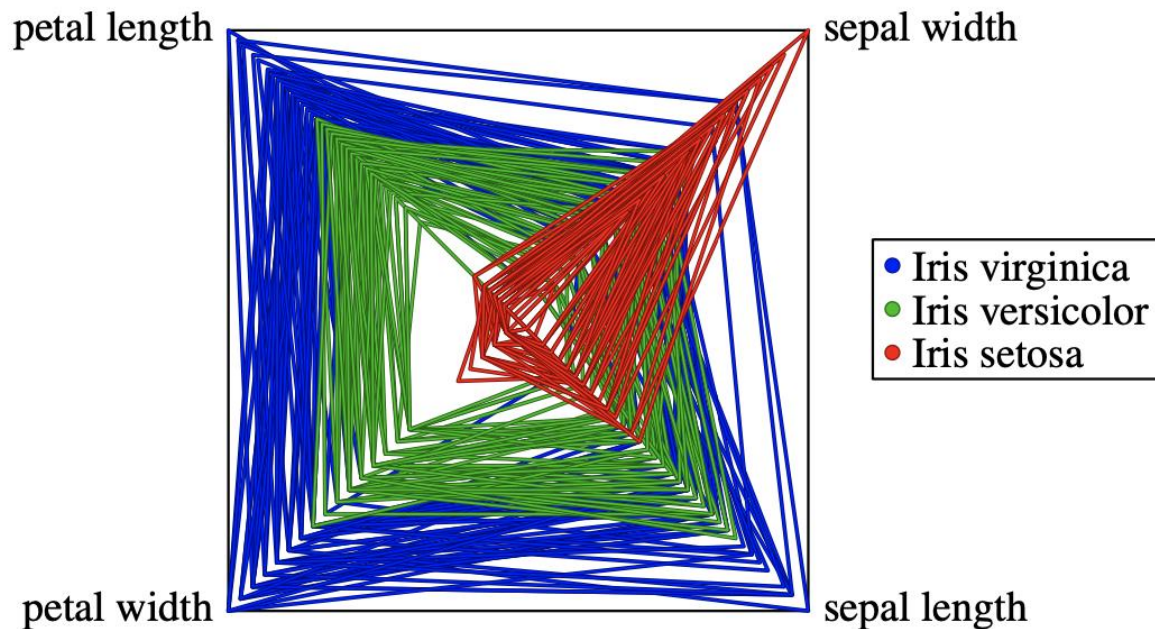
- Thus, each object is represented as a line
- Often, the lines representing a distinct class of objects group together, at least for some attributes
- Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data



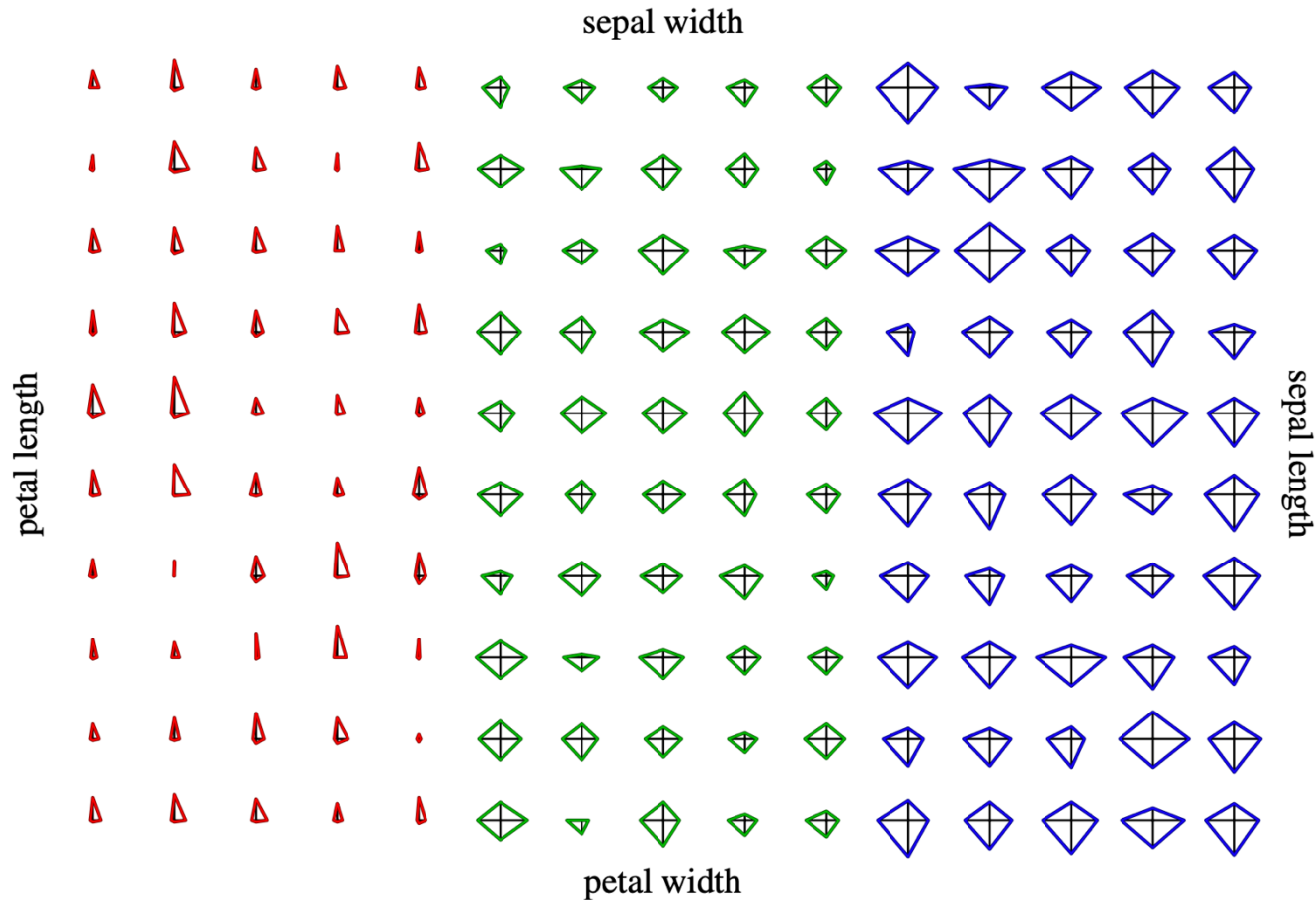
Radar Plot for Iris Data

- Similar idea as parallel coordinates
- **Coordinate axes are drawn as parallel lines, but in a star-like fashion intersecting in one point**
- Axes radiate from a central point
- The line connecting the values of an object is a polygon



Star Plots for Iris Data

Star plots are the same as radar plots where **each data object is drawn separately**.



Correlation Analysis

- Identifies the **linear relationship** between variables
- Captures similar behaviour of two attributes thus **detects redundant features**;
- Prepares data for regression, clustering and dimensionality reduction;
- Detects possible multicollinearity occurring when two or more predictor variables in a regression model are strongly linearly related
 - E.g., suppose we **predict house prices** using: surface area in square metres and surface area in square feet.
 - These variables contain almost the same information. The regression model then struggles to distinguish their individual contributions.

From variance to covariance

Variance

How one variable differs from its mean

$$s^2_x = 1/(n-1) \text{ Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad 1) \quad \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

$$s^2_x = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Covariance

How two variables differ from their means together

..... is covariance of a variable with itself

$$\text{Cov}(X, X) = \text{Var}(X)$$

From Variance to Covariance

- Variance measures how much the values of a variable differ from their mean

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Variance concerns one variable but in data mining, we are often interested in relationships between variables.
- Example:
 - Do customers with higher incomes spend more?
 - Does the price of a product affect its sales?
 - Do students who study more hours obtain higher marks?
 - Do two sensors provide largely redundant information?
- We need to examine how two variables vary together: **Covariance!**

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

From Variance to Covariance

- Given two variables (X) and (Y), their sample covariance is:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Variance is covariance of a variable with itself!

$$\text{Cov}(X, X) = \text{Var}(X)$$

The sign comes from paired deviations

$$x_i > \bar{x} \quad \text{and} \quad y_i > \bar{y}$$

positive $\times$ positive

Positive

$$x_i < \bar{x} \quad \text{and} \quad y_i < \bar{y}$$

negative $\times$ negative

Positive

one above, one below

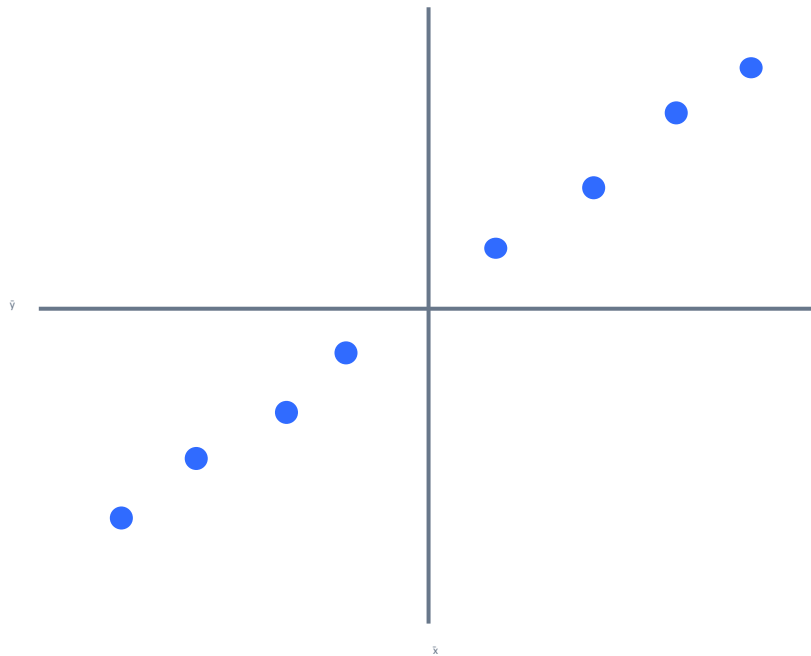
positive $\times$ negative

Positive

Question: What sign do you expect for study hours and examination marks?

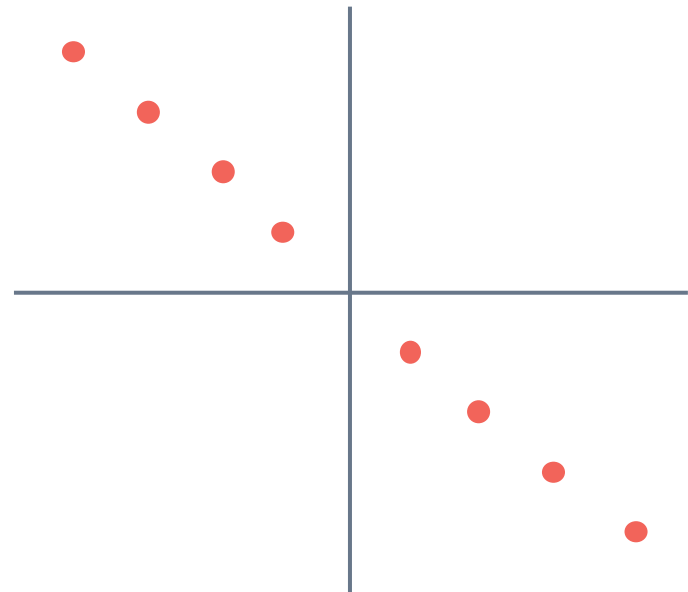
Geometric interpretation

Positive covariance



same-side quadrants dominate

Negative covariance



opposite-side quadrants dominate

Limitation:

Covariance depends on measurement units

Income in euros

$\text{Cov}(\text{income}, \text{expenditure})$

units: euro²

Change only Income in thousands of euros

The numerical covariance changes

The underlying relationship does not

**Covariance gives the direction of linear relationship
but its magnitude cannot be compared easily across different scales**

$\text{Cov}(\text{age}, \text{income})$

$\text{Cov}(\text{height}, \text{weight})$

Hard to compare without standardization

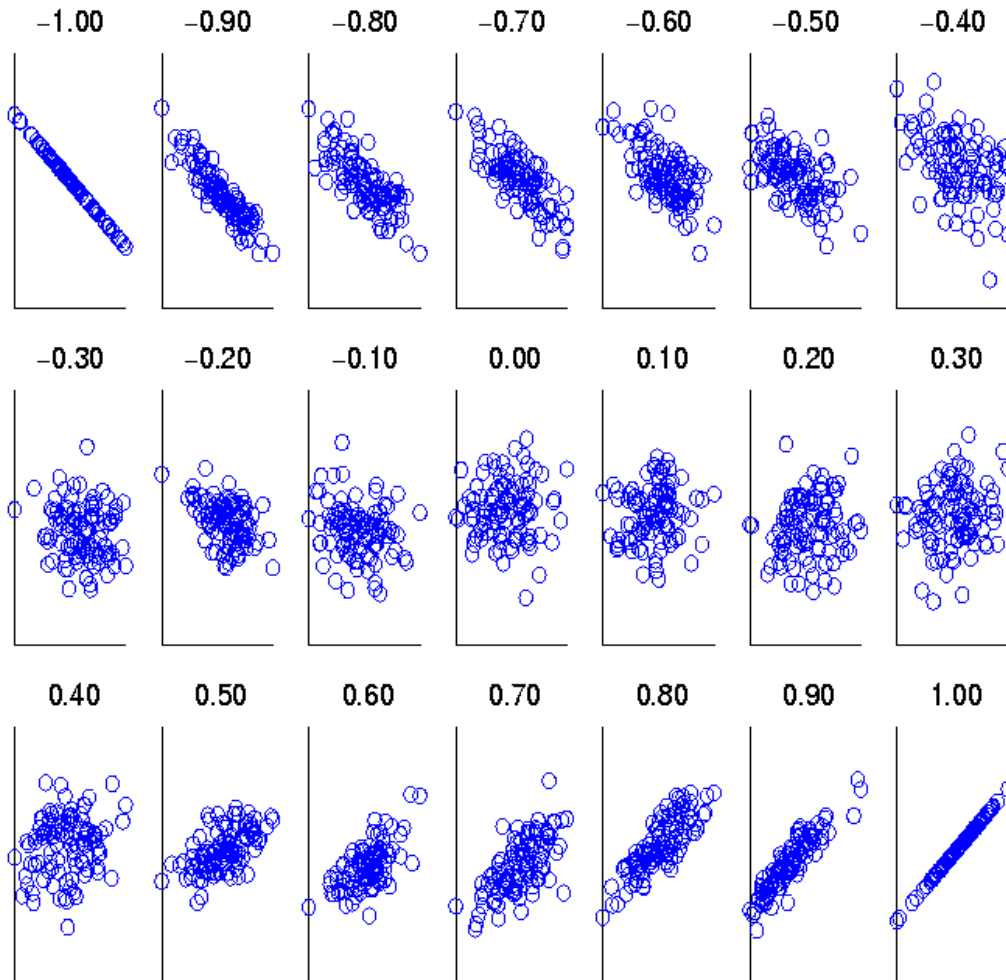
Pearson's correlation coefficient

The (sample) **Pearson's correlation coefficient** standardizes covariance by dividing it by the standard deviations of the two variables

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

- where $\bar{x}$ and $\bar{y}$ are the mean values of X and Y, respectively. s_x and s_y are the corresponding (sample) standard deviations.
- **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.
- Around zero indicates little or **no linear relationship**.

Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1 .

Let's reason together on the correlation!

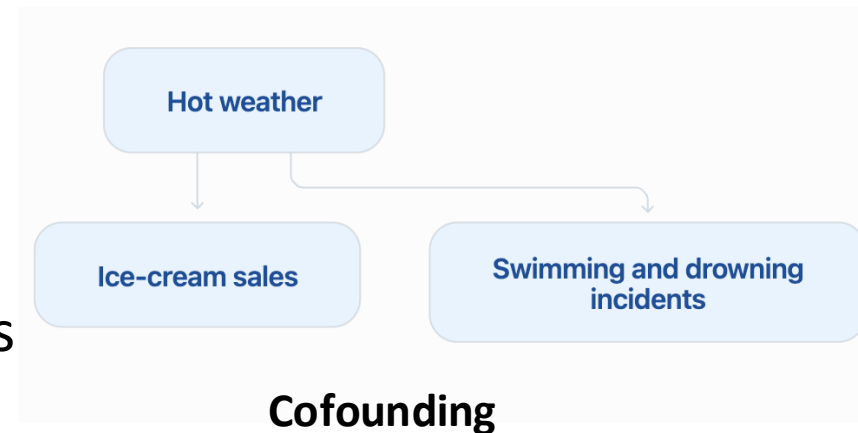
- Students with fewer **study hours** tend to have **marks** below the mean, while students with more study hours tend to have marks above the mean.
- Suppose a dataset where Pearson Correlation is 0.98
 - This indicates a very strong positive linear association
- **Question:** Can we conclude that 2 additional study hours necessarily **causes** a particular increase in marks?
 - Answer: NO!!!

Correlation describes association, not causation!

- Other variables may influence both quantities

Correlation does not identify the causal direction

- Correlation tells us that two variables change together:
statistical association
- It does not tell us that changing one variable causes the other to change, i.e., $X \rightarrow Y$ or $Y \rightarrow X$
- For example
 - ice-cream sales and drowning incidents are positively correlated.
 - However, buying ice cream does not cause drowning. A third variable, **hot weather**, increases both.



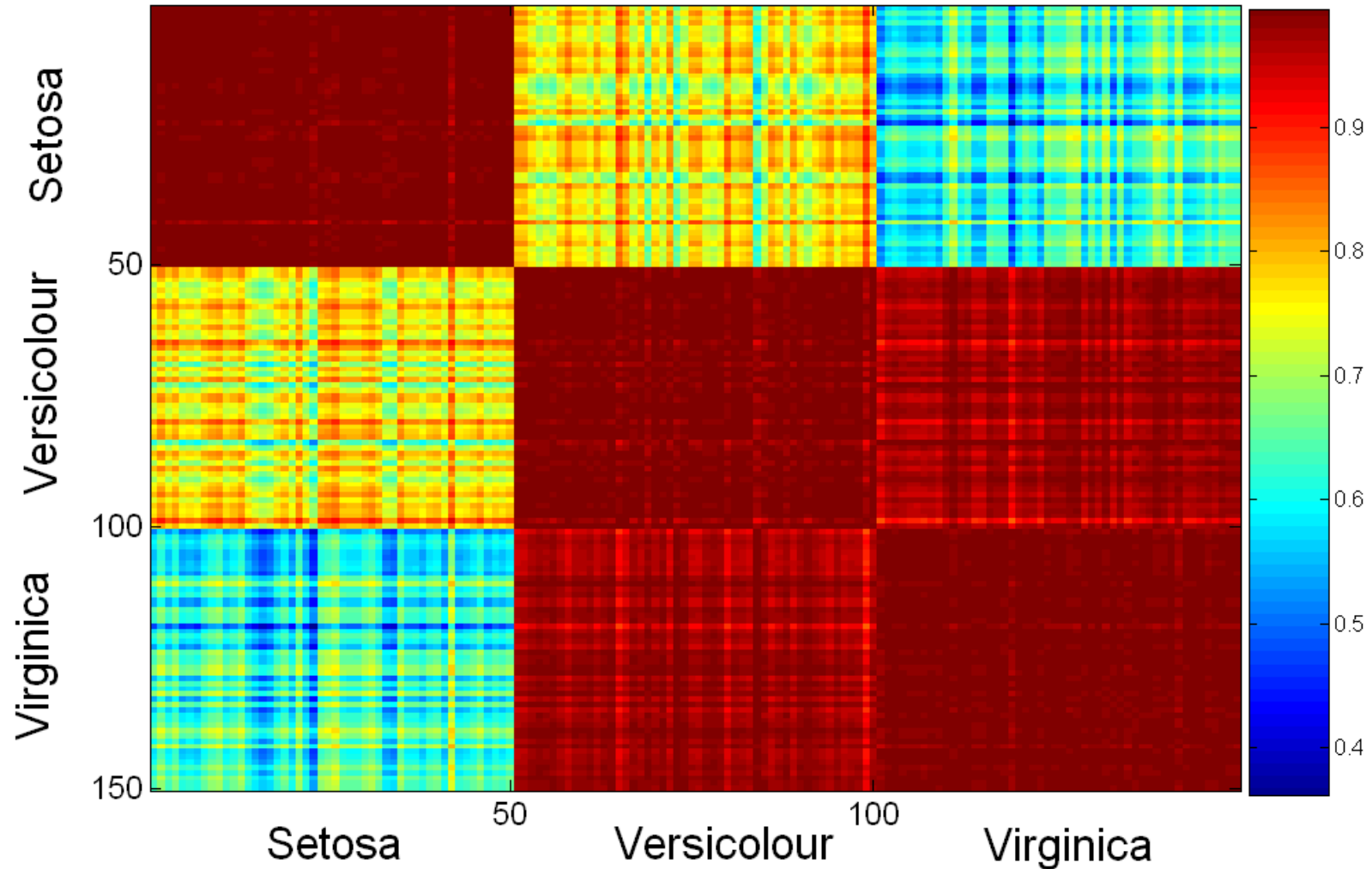
Spearman's rank correlation

$$\rho = 1 - 6 \frac{\sum_{i=1}^n (r(x_i) - r(y_i))^2}{n(n^2 - 1)},$$

where $r(x_i)$ is the rank of value x_i when we sort the list $(x_1, \dots, x_n)$ in increasing order. $r(y_i)$ is defined analogously.

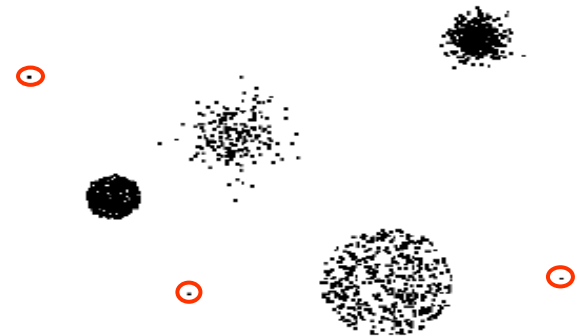
- When the rankings of the x- and y-values are exactly in the same order, Spearman's rho will yield the value 1.
- If they are in reverse order, we will obtain the value -1.
- It does not compare the original values directly. It replaces each value with its position in the ordered data.
- It measures whether the variables tend to move in the same or opposite direction. The relationship does not need to be linear, but it should be monotonic.

Visualization of the Iris Correlation Matrix



Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set
 - **Case 1:** Outliers are noise that interferes with data analysis
 - **Case 2:** Outliers are the goal of our analysis
 - Credit card fraud
 - Intrusion detection
- **Causes:**
 - Data quality problems (erroneous data coming from wrong measurements or typing mistakes)
 - Exceptional or unusual situations/data objects.



Outliers as noise

- 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 attribute:**
 - **Categorical** attributes: An outlier is a value that occurs with a frequency extremely lower than the frequency of all other values.
 - **Numerical** attributes: box plots
- **Multidimensional attribute:**
 - Scatter plots for (visually detecting) outliers w.r.t. two attributes
 - PCA 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
- Reasons for missing values
 - **Information is not collected**
(e.g., people decline to give their age and weight)
 - **Attributes may not be applicable to all cases**
(e.g., annual income is not applicable to children)
 - **Broken sensors**
 - **Refusal to answer a question**
- Missing value might not necessarily be indicated as missing (instead: zero or default values).

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 behaviour.