PROGRAMA DE VERÃO DO LNCC
JORNADA DE CIÊNCIA DE DADOS

Exploratory analysis

CEFET/RJ

Eduardo Ogasawara
http://eic.cefet-rj.br/~eogasawara

Types of Data Sets

- Record
  - Relational datasets
- Matrix
  - numerical matrix, crosstabs
- Documents
  - texts, term-frequency vector
- Transactions
- Graph and network
  - World Wide Web
  - Social or information networks
- Ordered
  - Temporal data: time-series
  - Sequential data: transaction sequences
- Spatial, image, and multimedia
  - Spatial data: maps
  - Images
  - Videos

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Documents	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

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Data	PIB - R$ (milhões)
1990.01	0.2
1990.02	0.4
1990.03	0.8
1990.04	0.7
1990.05	0.8

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

Relational data

- Data sets are made up of data objects
- A data object represents an entity
  - sales database: customers, store items, sales
  - medical database: patients, treatments, illness
  - university database: students, professors, courses
- Attributes describe data objects
- Database
  - rows: data objects (tuples)
  - columns: attributes

Attributes

- Attribute (or dimensions, features, variables)
  - a data field, representing a characteristic or feature of a data object
  - E.g., customer _ID, name, address
- Types
  - Nominal
  - Binary
  - Ordinal
  - Numeric

Attribute Types

- Nominal: categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- Binary
  - Attribute with only two 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)
    - Convention: assign 1 to the most important outcome (e.g., HIV positive)
- Ordinal
  - Values have a meaningful order (ranking), but magnitude between successive values is not known
  - Size = {small, medium, large}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- Interval
  - Measured on a scale of equal-sized units
  - Values have order
    - E.g., the temperature in C˚or F˚, calendar dates
  - No true zero-point
- Ratio
  - Inherent zero-point
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., the temperature in Kelvin, length, counts, monetary quantities

Discrete vs. Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Sometimes, represented as integer variables
- Continuous Attribute
  - Has real numbers as attribute values
    - E.g., 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

Iris Dataset

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
numeric	numeric	numeric	numeric	factor

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data:
    - central tendency, variation and spread
- Data centrality and dispersion characteristics
  - median, max, min, quantiles, outliers, variance
- Numerical dimensions correspond to sorted intervals
  - Boxplot or quantile analysis on sorted intervals

Descriptive Measures

- Centrality
  - Mean (algebraic measure)
    - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - Median
    - Middle value if an odd number of values, or weighted average of the middle two values otherwise
  - Mode
    - The value that occurs most frequently in the data
    - Unimodal, bimodal, trimodal
    - Empirical formula:
      - $mean - mode = 3 \cdot (mean - median)$
- Dispersion
  - Variance and standard deviation
    - Variance: (algebraic, scalable computation)
    - Standard deviation (σ): square root of the variance (σ^2)
      - $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} = \frac{\sum_{i=1}^{n} {x_i}^2}{n} - \mu^2$

Measuring the Dispersion of Data

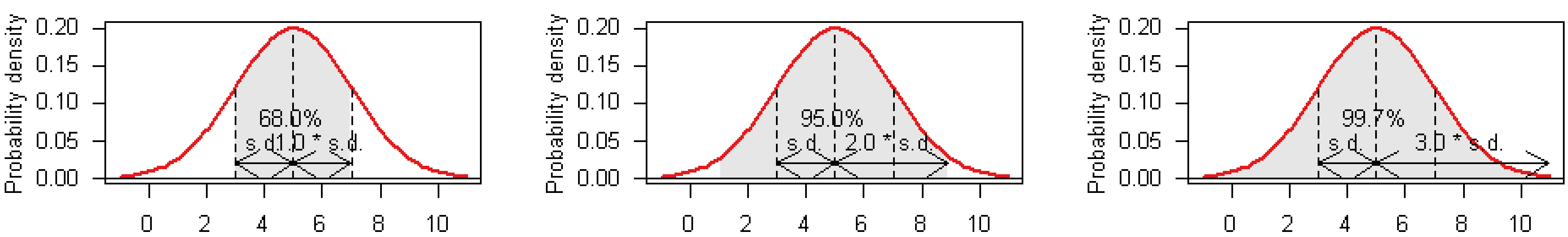
- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: $IQR = Q_3 - Q_1$
  - Five number summary: min, Q_1, median, Q_3, max
  - Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually

Statistics	Freq
Min.	4.300000
1st Qu.	5.100000
Median	5.800000
Mean	5.843333
3rd Qu.	6.400000
Max.	7.900000

```
[1] "IQR=1.3"
```

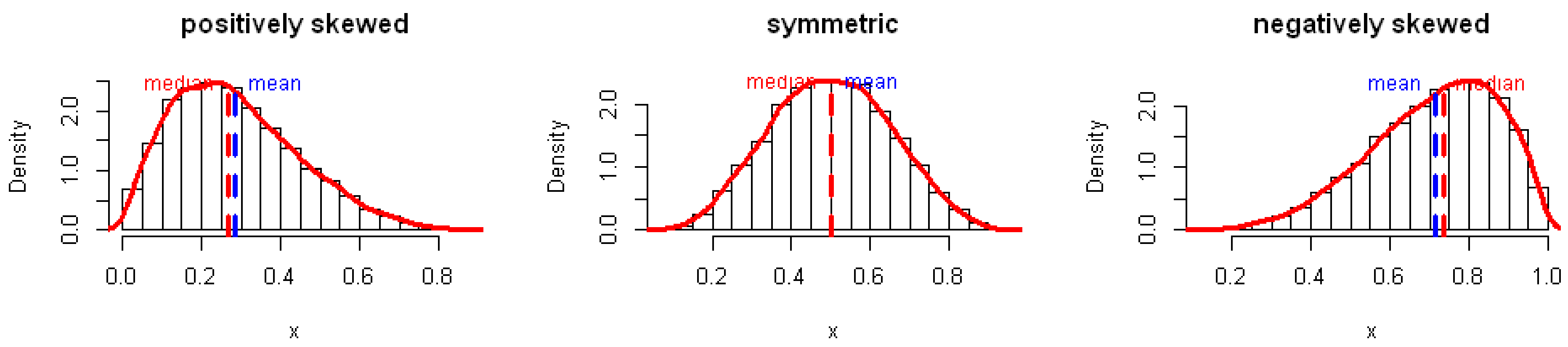
Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

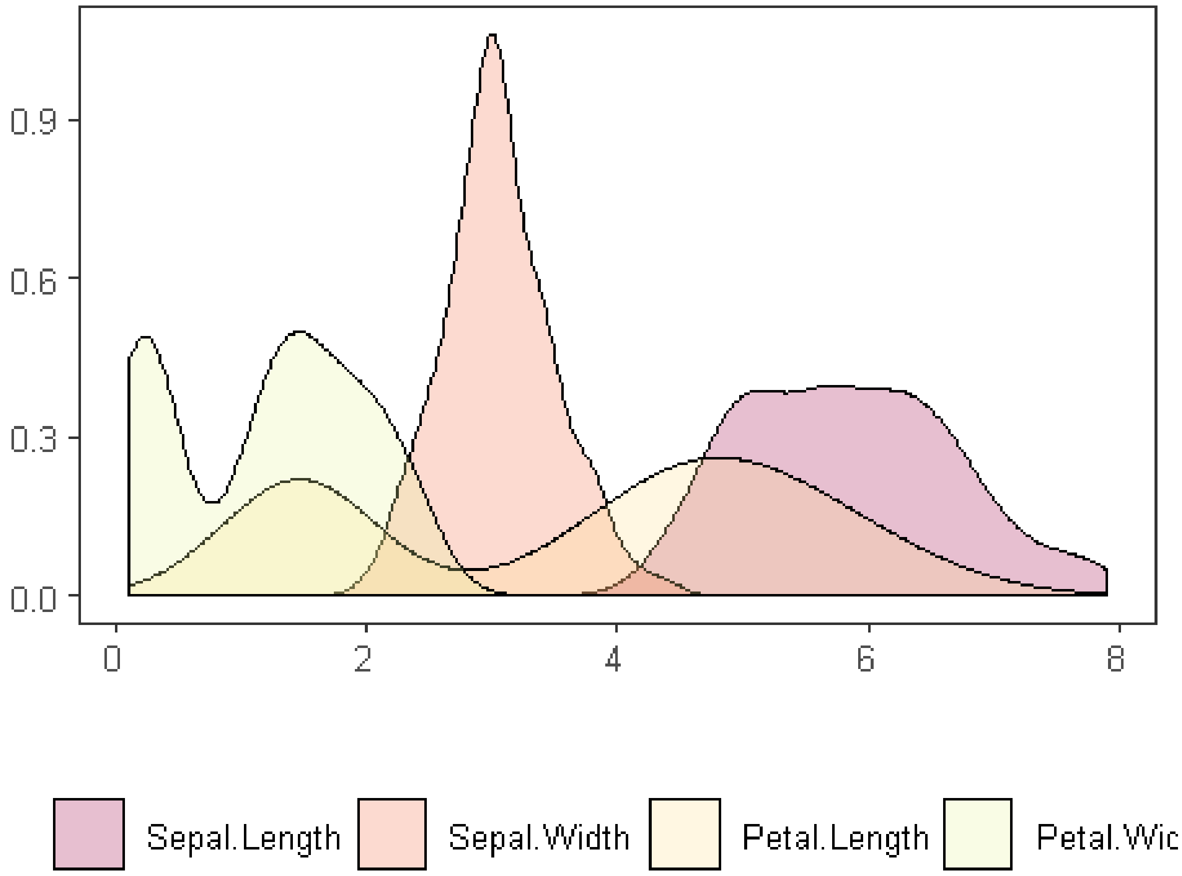


Symmetric vs. Skewed Data

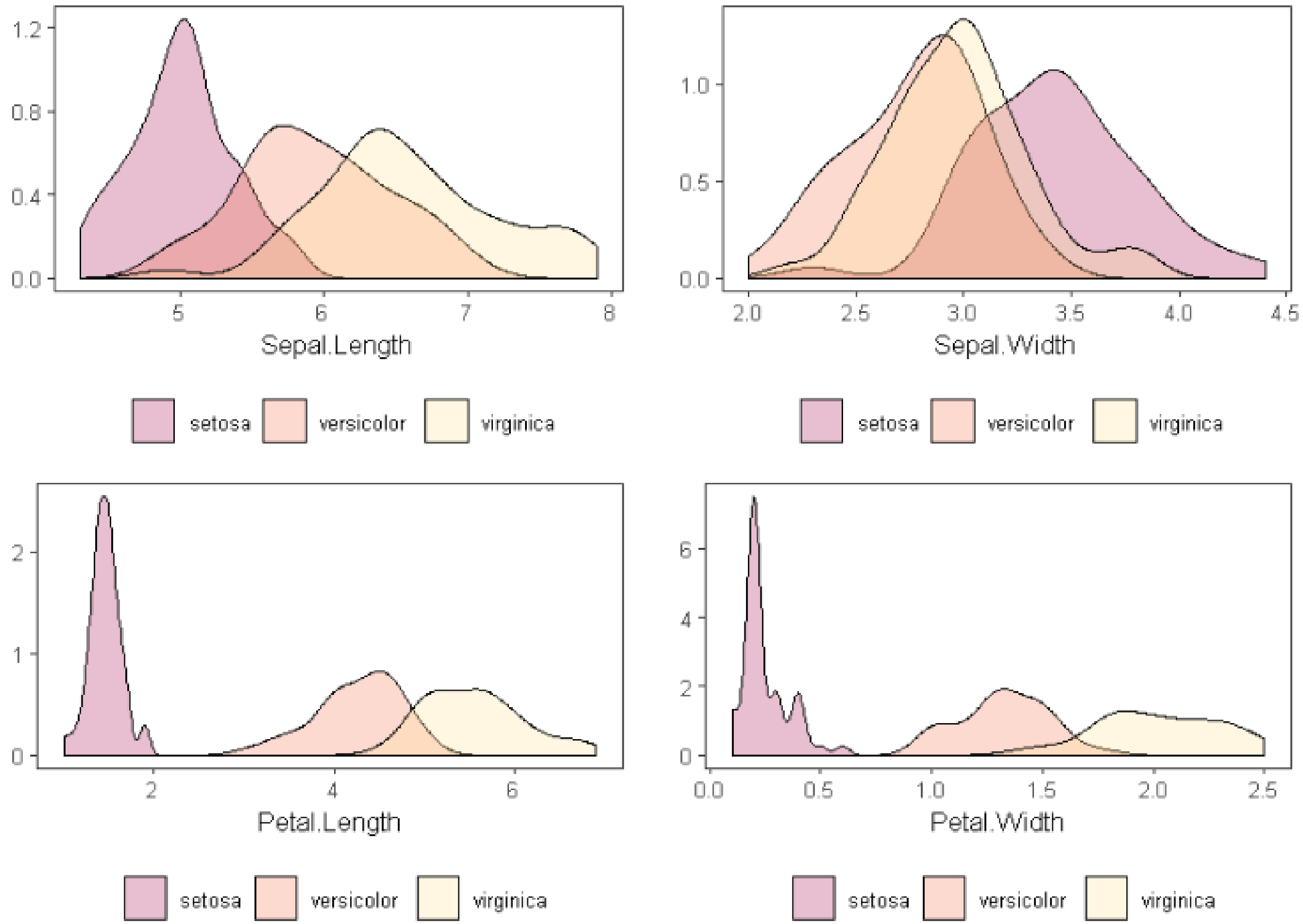
- Median and mean for:
  - positive, symmetric, and negatively skewed data



Probability density function



Density distributions per class label

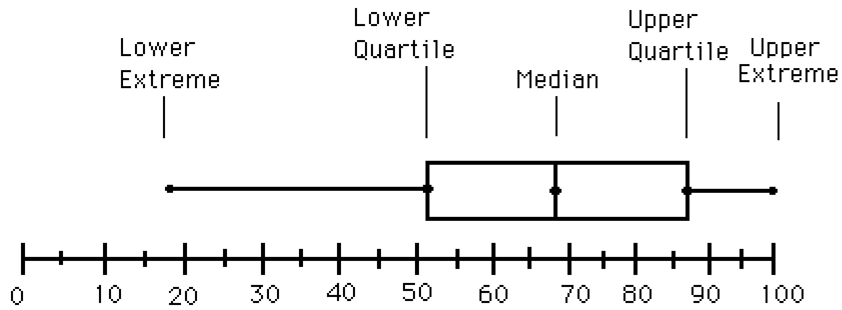


Graphic Displays of Basic Statistical Descriptions

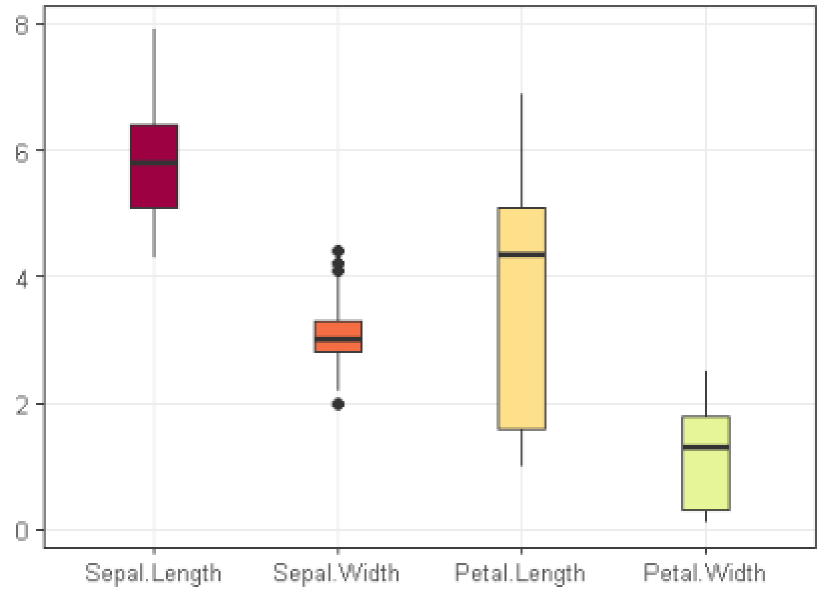
- Boxplot
- Histogram
- Quantile-quantile (q-q) plot
- Scatter plot

Boxplot Analysis

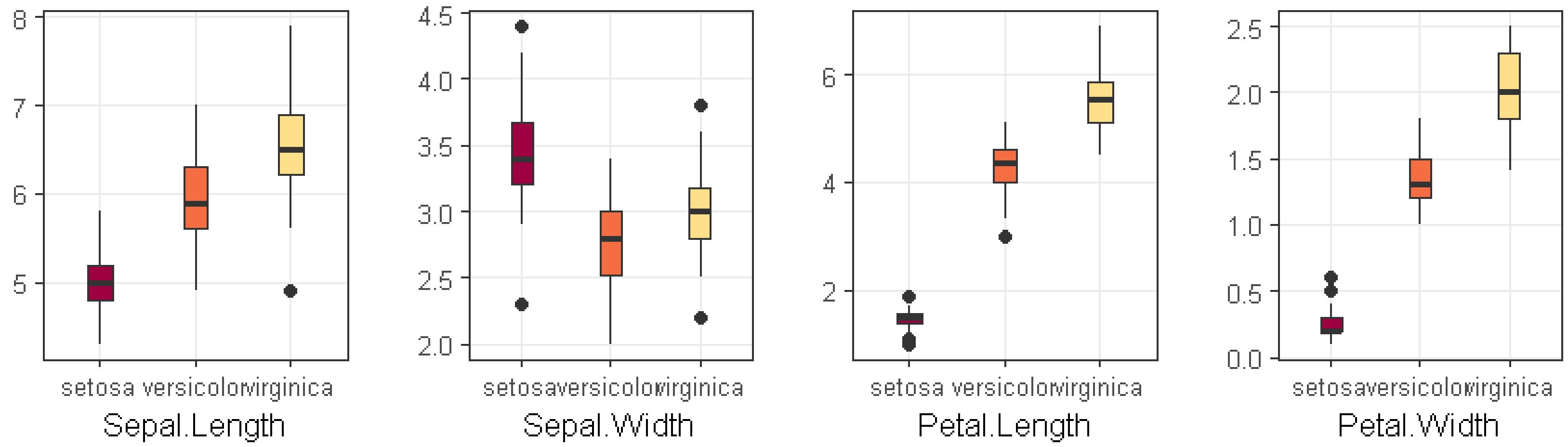
- Five-number summary of a distribution
  - Min., Q1, Median, Q3, Max.
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - A line within the box marks the median
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers are values:
    - higher than Q3 + 1.5 x IQR
    - lower than Q1 - 1.5 x IQR



Boxplot for all variables

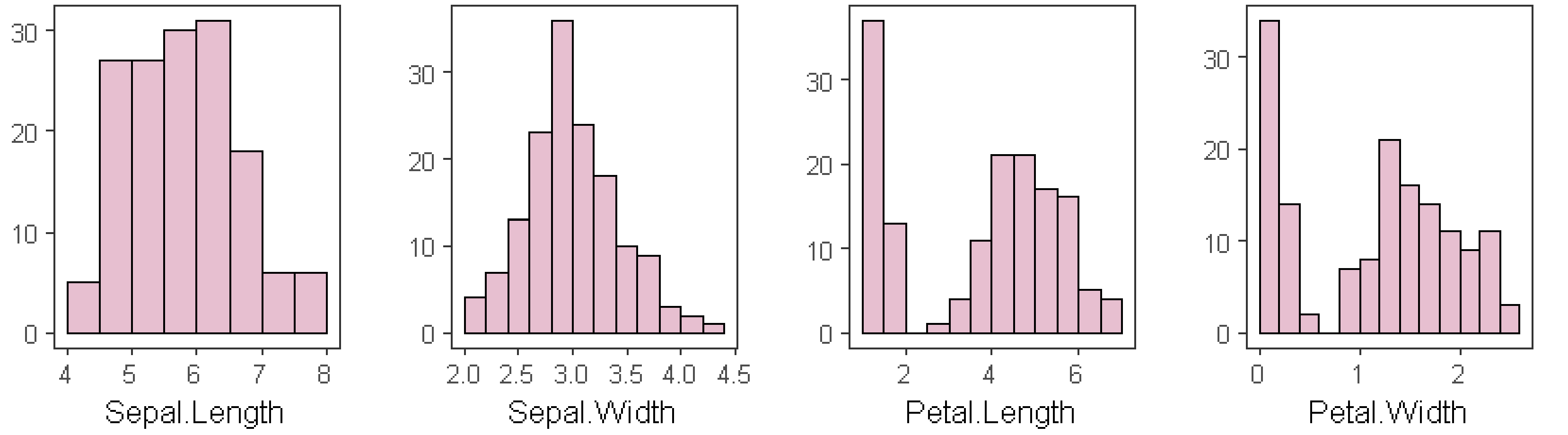


Boxplot per class label



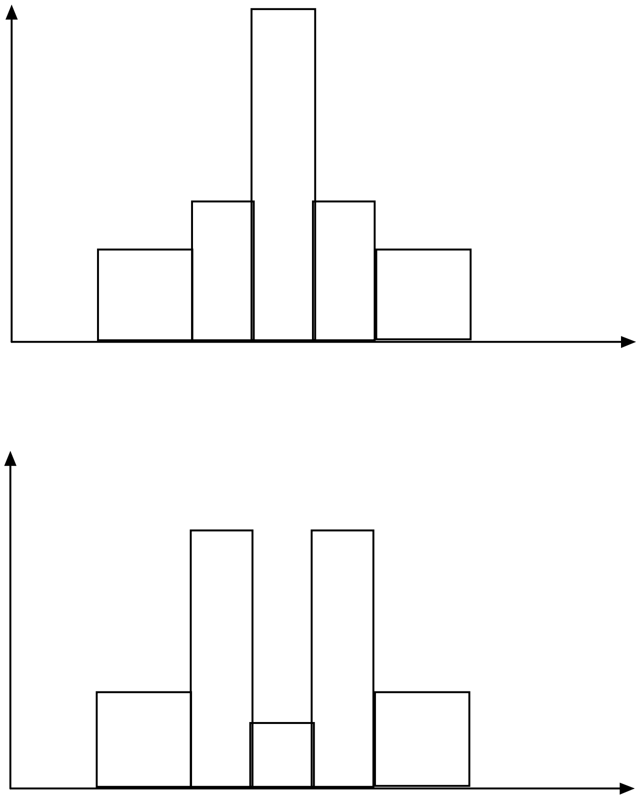
Histogram Analysis

- The histogram displays values of tabulated frequencies
- It shows what proportion of cases into each category
- The area of the bar that denotes the value
  - It is a crucial property when the categories are not of uniform width
- The categories specify non-overlapping intervals of some variable
- The categories (bars) must be adjacent



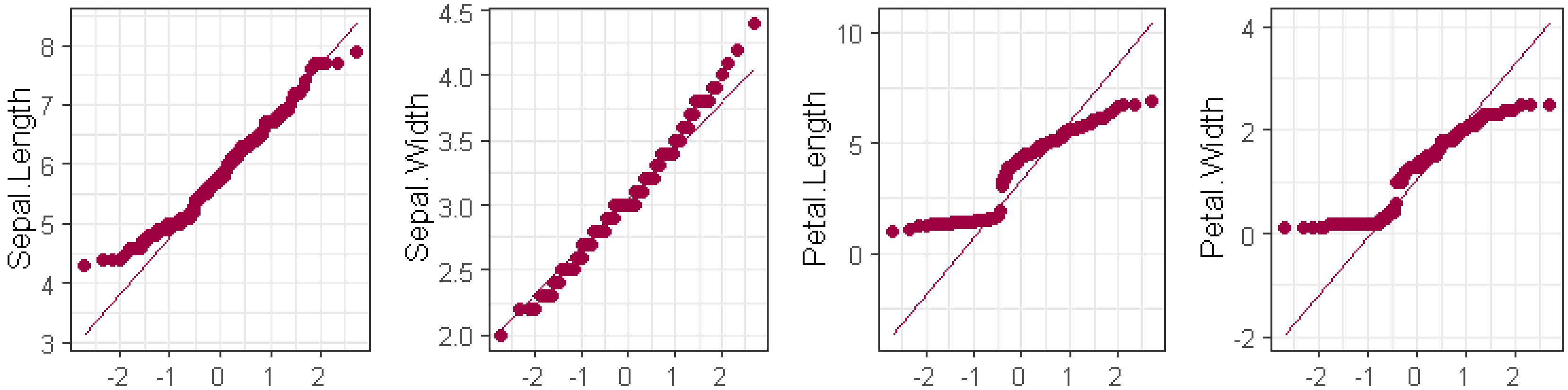
Histograms may tell more than Boxplots

- The two histograms shown in the left may have the same boxplot representation
  - The same values for min, Q1, median, Q3, max
- However, they have rather different data distributions



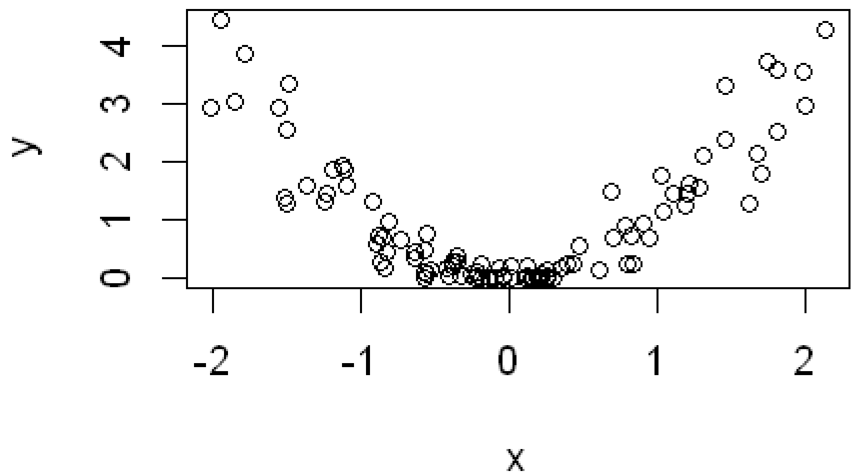
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another (theoretical distribution)
- A good approach to visual inspect if the distribution is similar to a standard normal

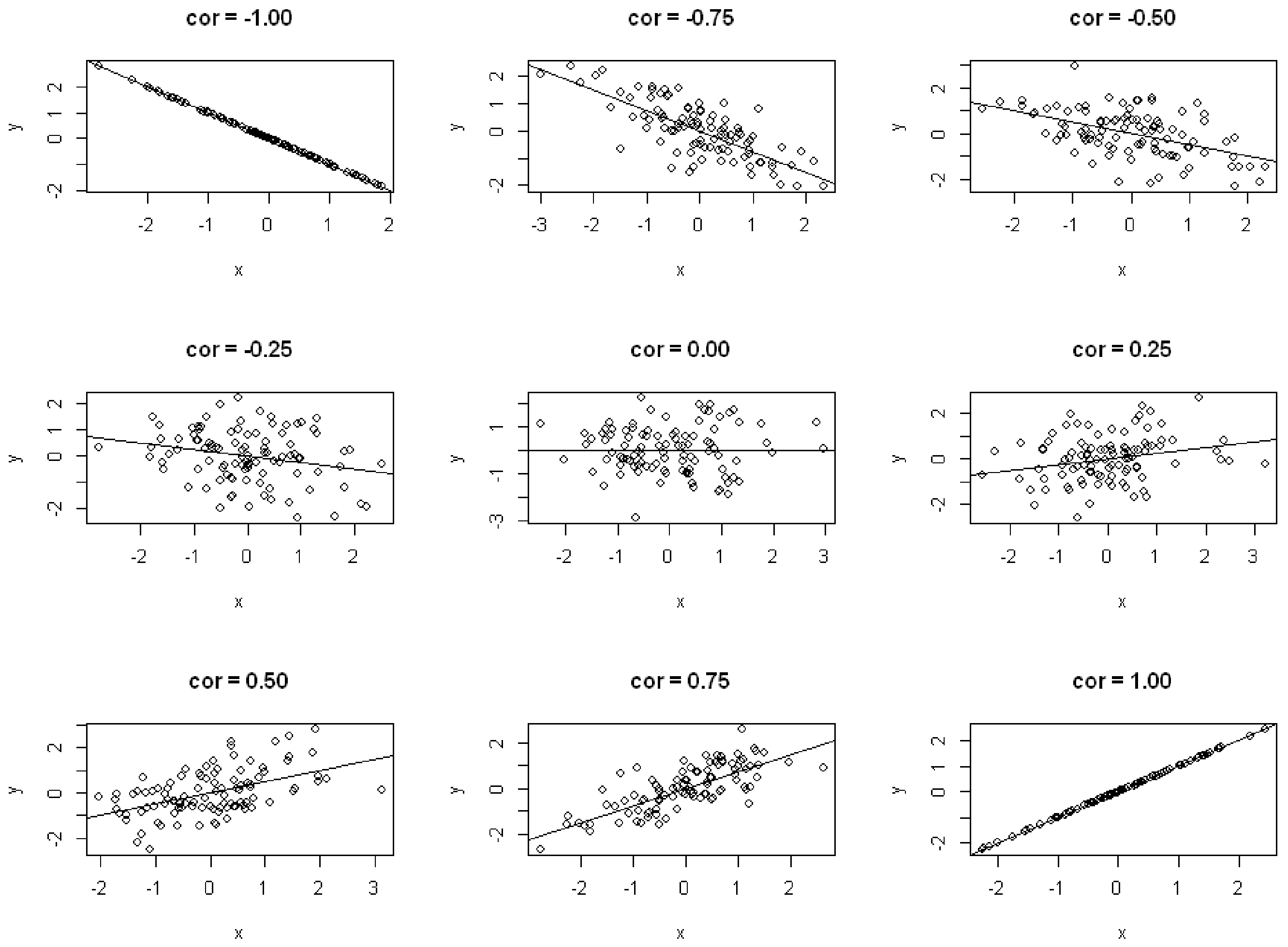


Scatter plot

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



Data correlation



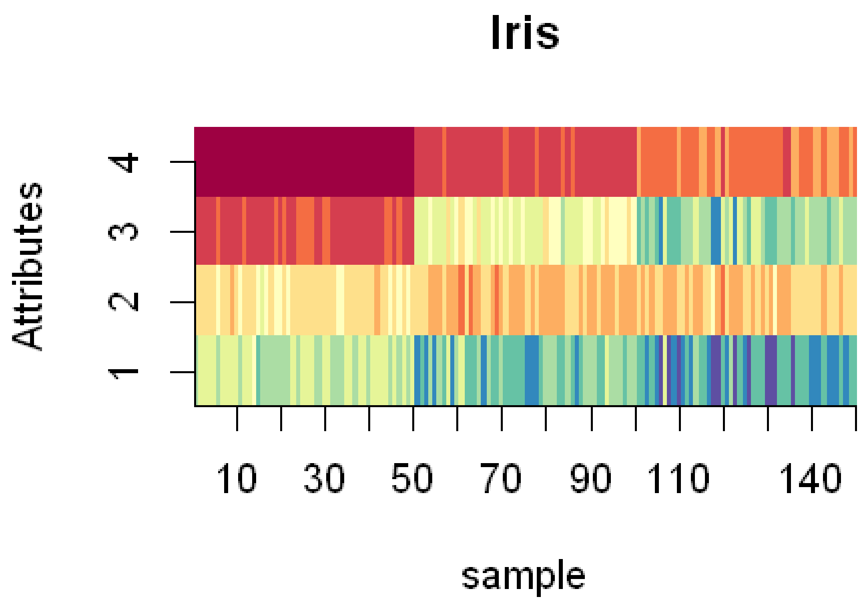
The first row presents negatively correlated data
The second row presents uncorrelated data
The third row presents positively correlated data

Data Visualization

- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide a qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

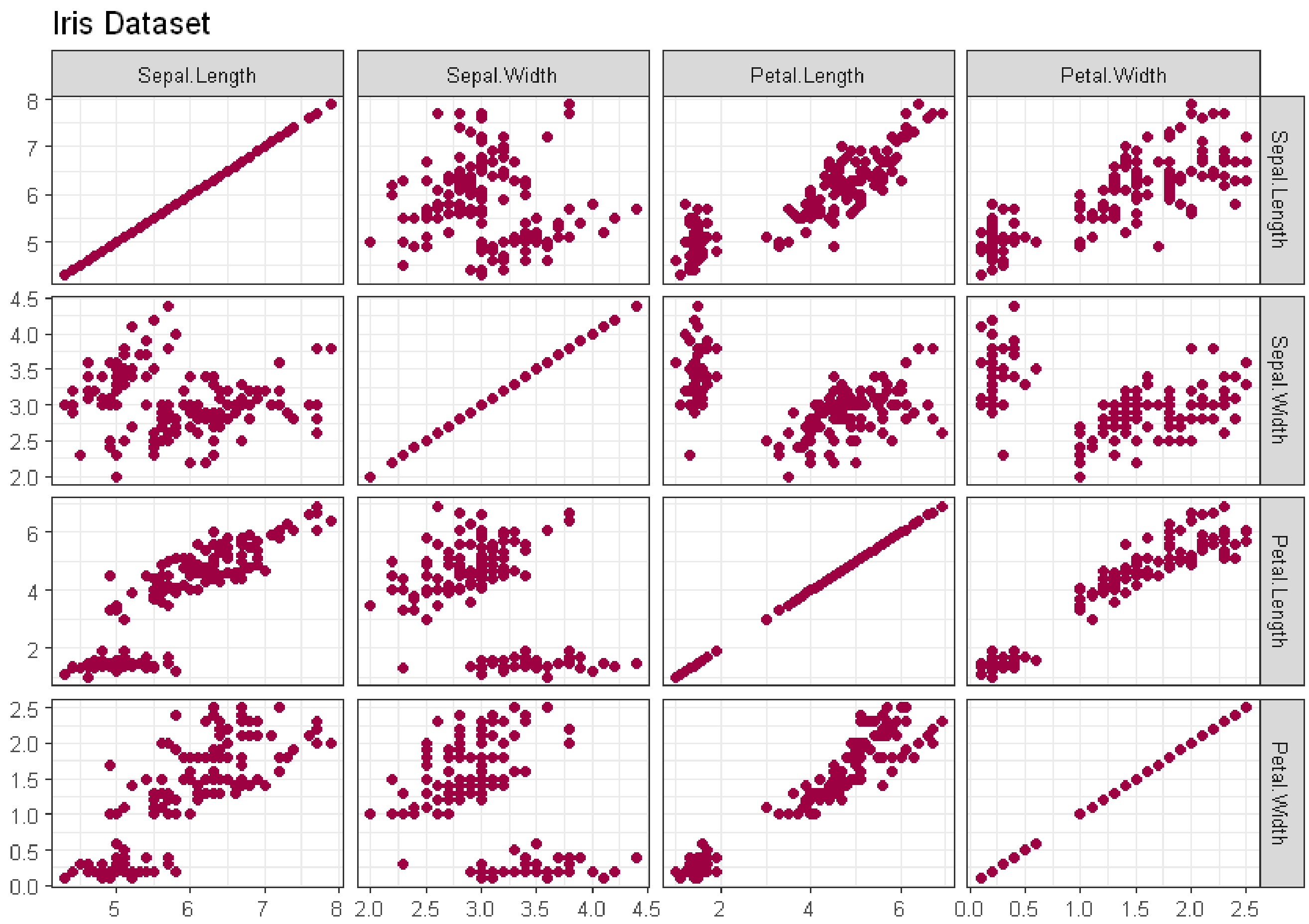
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



Geometric Projection Visualization Techniques

- Visualization of geometric transformations and projections of the data
- Methods
  - Direct visualization
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Parallel coordinates

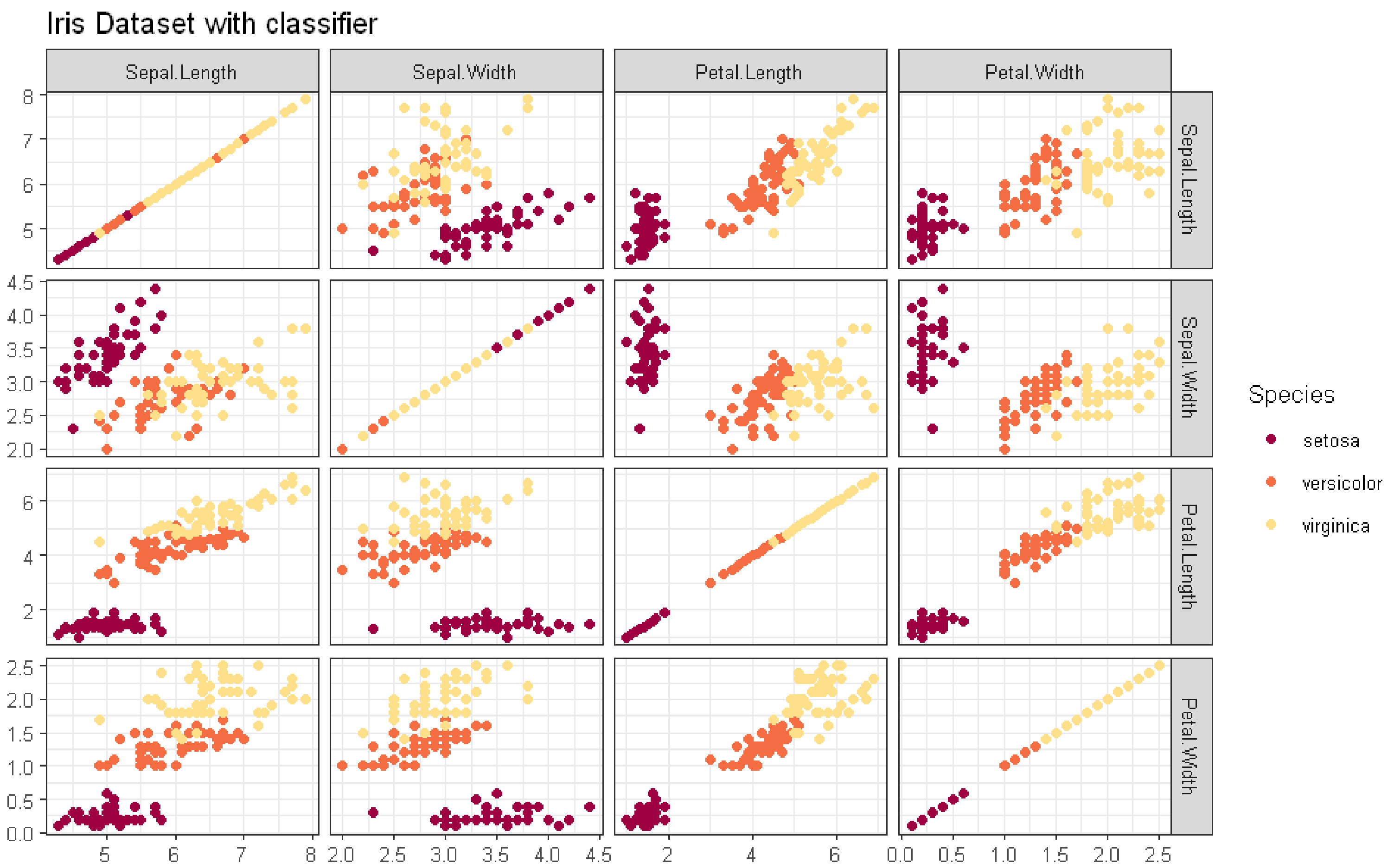
Scatterplot Matrices



A matrix of scatterplots (x-y-diagrams)

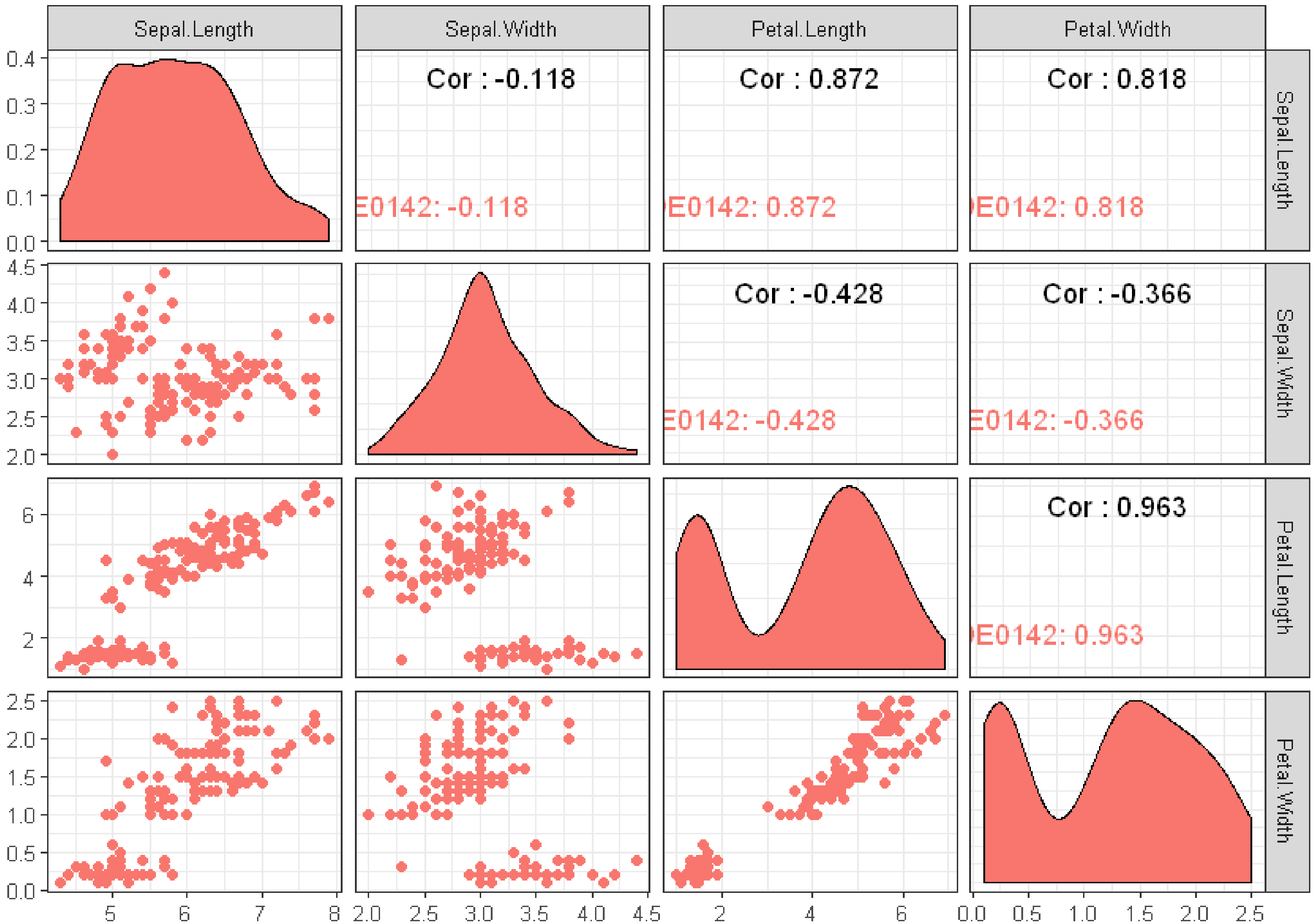
k-dimensional data: total of ($k^2/2$-k) scatterplots]

Scatterplot matrices with a class label

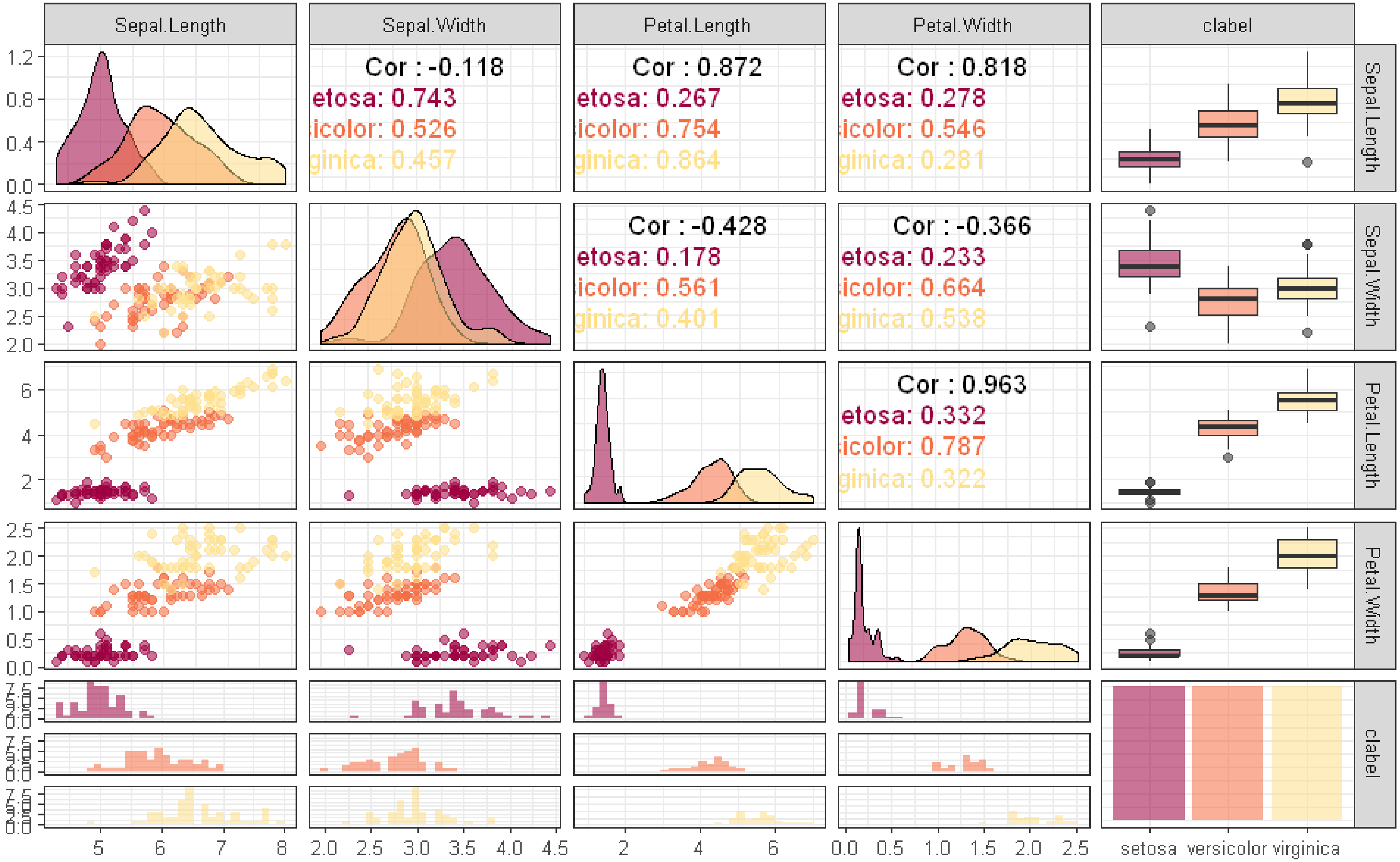


Advanced Matrices Plot

- The matrix of optimized plots of the k-dim. data



Advanced Matrices Plot with a class label



Landscapes

- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

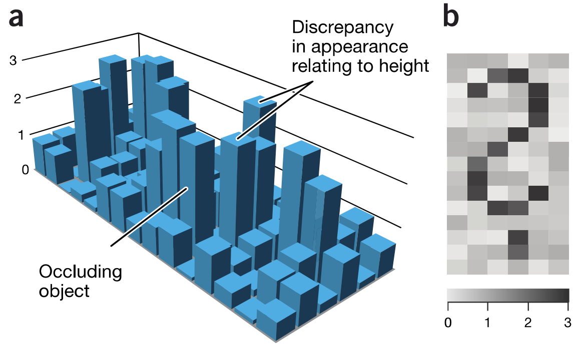
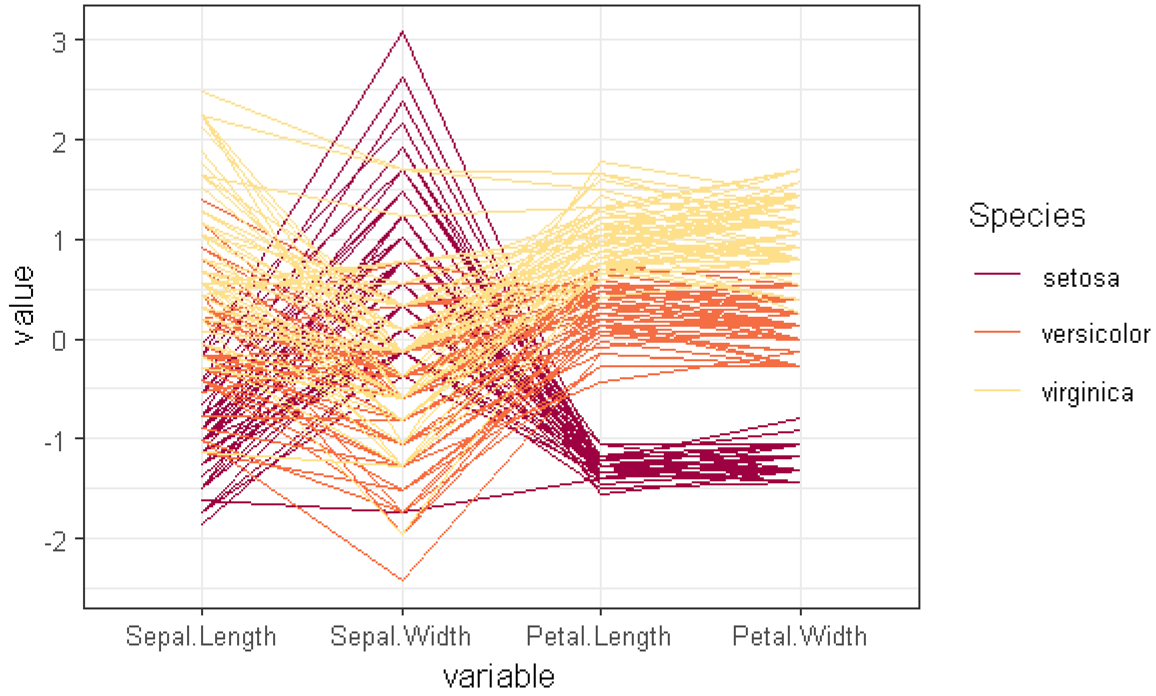


Figure 2 | Three-dimensional representation of abstract data. (**a**) Data occlusion and interference of visual encodings with depth cues can be problematic in three-dimensional space. (**b**) The same data as in **a** plotted as a two-dimensional heat map.

NATURE METHODS | VOL.9 NO.9 | SEPTEMBER 2012 | 851

Parallel Coordinates of a Data Set



Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Salience
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

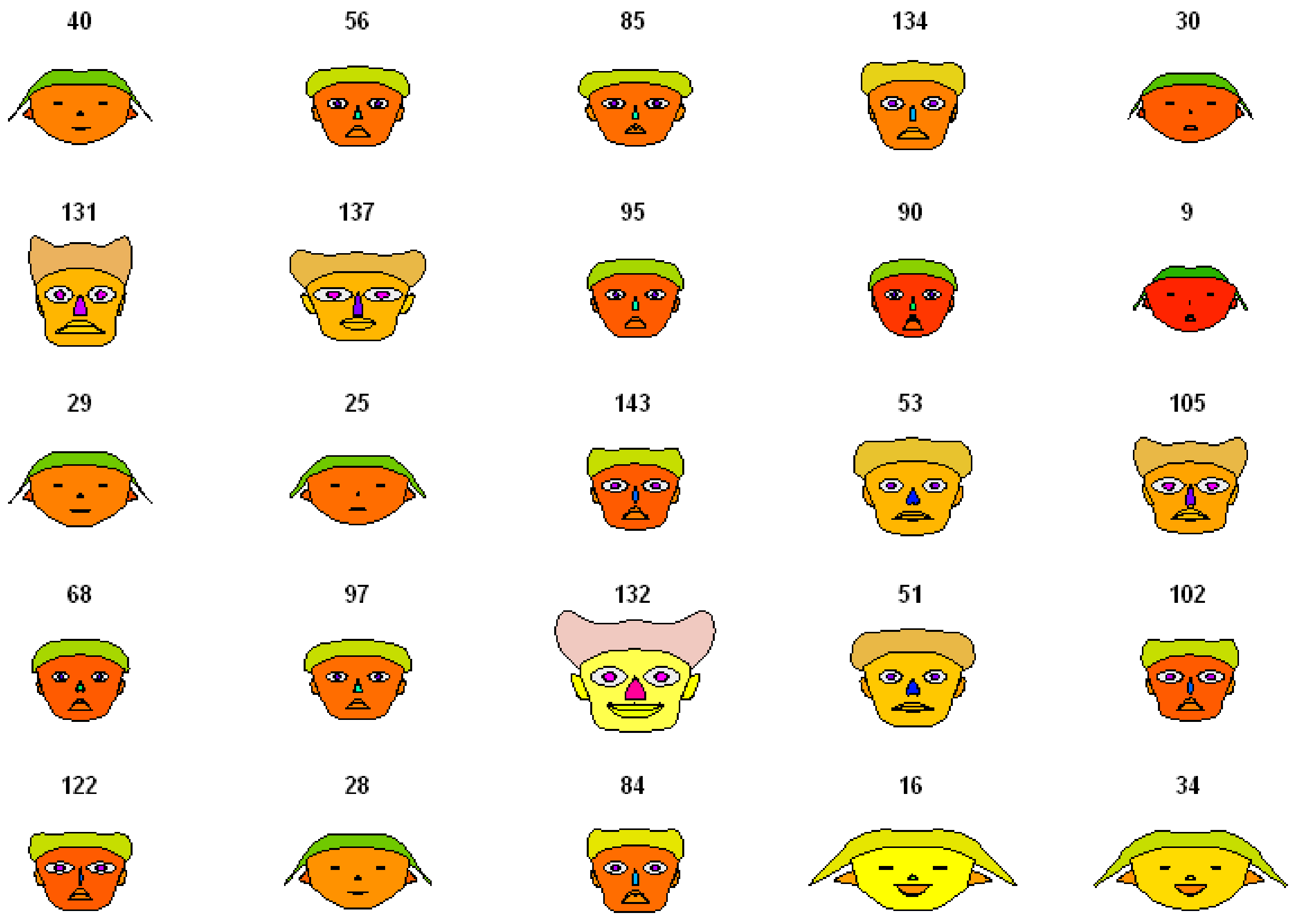
Chernoff Faces

- A way to display variables on a two-dimensional surface
  - Let x be eyebrow slant, y be eye size, z be nose length
- The figure shows faces produced using ten characteristics: head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening):
  - Each assigned one of 10 possible values

Gonick, L. and Smith, W. The Cartoon Guide to Statistics. New York: Harper Perennial, p. 212, 1993

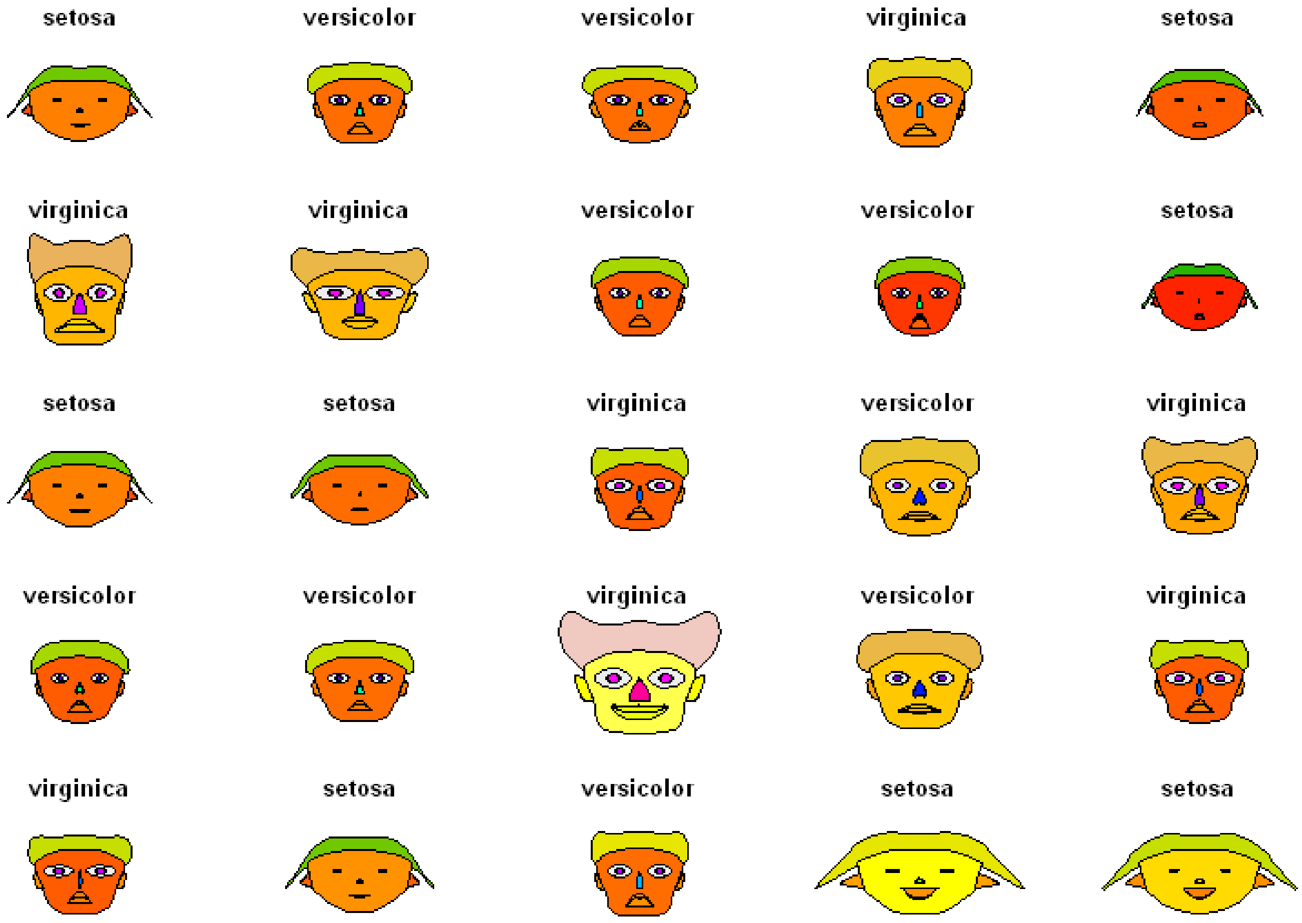
Weisstein, Eric W. "Chernoff Face." From MathWorld -A Wolfram Web Resource. mathworld.wolfram.com/ChernoffFace.html

Chernoff Faces example with the Iris dataset



Can you see any pattern?

Chernoff Faces example with the Iris dataset



Salience

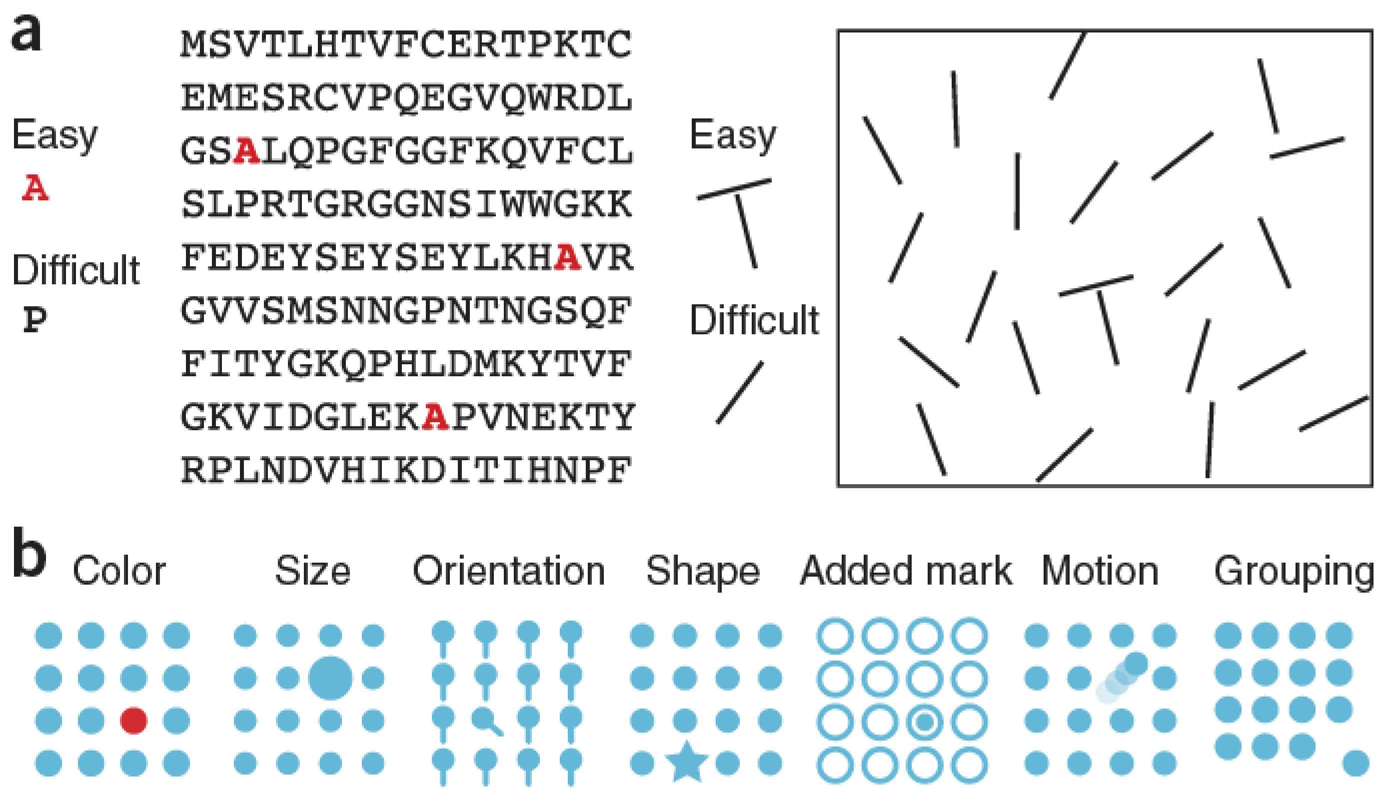


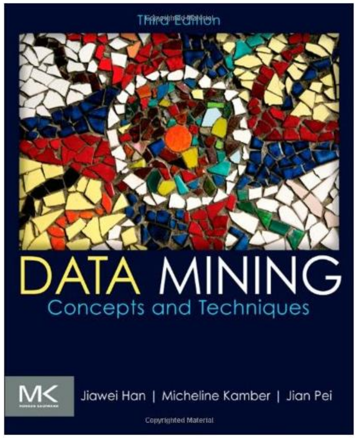
Figure 1 | Salience through visual features. (**a**) Certain elements can be seen in a single glance, whereas others are difficult to find. (**b**) Examples of visual features that make objects distinct.

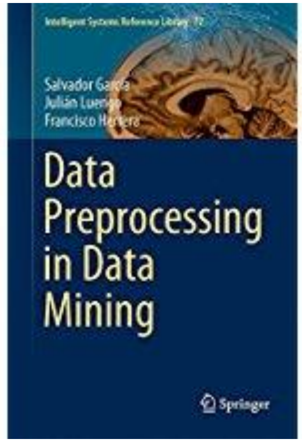
NATURE METHODS | VOL.7 NO.10 | OCTOBER 2010 | 773

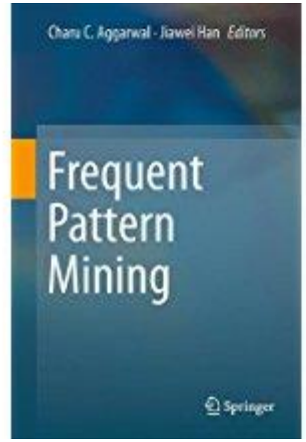
Practicing

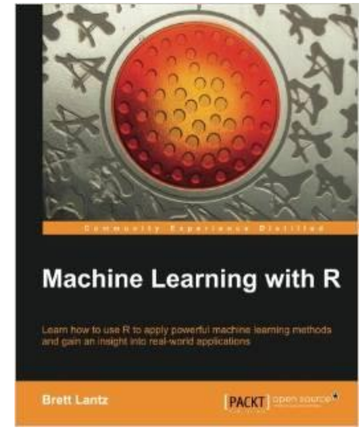
- Take some time to practice the examples
  - https://nbviewer.jupyter.org/github/eogasawara/mylibrary/blob/master/myExploratoryAnalysis.ipynb
- Learn to use Jupyter with R
  - http://jupyter.org

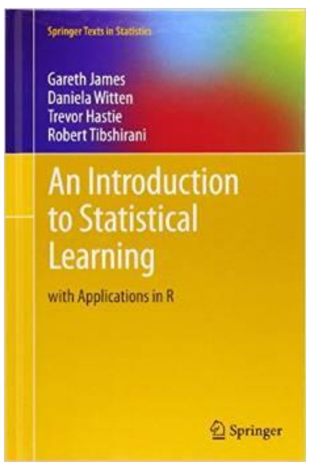
Main References

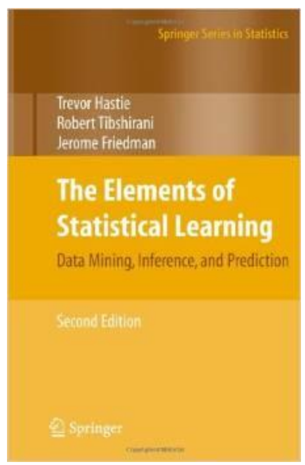


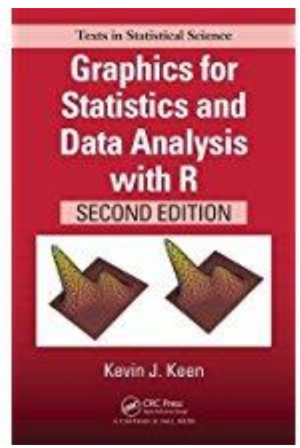


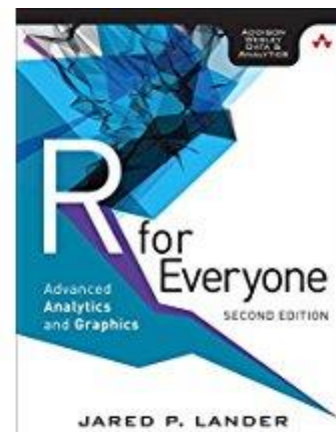












Most of the slides were extracted from
Data Mining Concepts and Techniques