Chapter 3

Descriptive multivariate analysis

Printer-friendly slides to the book

A General Introduction to Data Analytics

João Mendes Moreira, André C. P. L. F. de Carvalho and Tomáš Horváth

Descriptive Multivariate Analysis

Friend	Max temp	Weight	Height	Years	Gender	Company
Andrew	25	77	175	10	M	Good
Bernhard	31	110	195	12	M	Good
Carolina	15	70	172	2	F	Bad
Dennis	20	85	180	16	M	Good
Eve	10	65	168	0	F	Bad
Fred	12	75	173	6	M	Good
Gwyneth	16	75	180	3	F	Bad
Hayden	26	63	165	2	F	Bad
Irene	15	55	158	5	F	Bad
James	21	66	163	14	M	Good
Kevin	30	95	190	1	M	Bad
Lea	13	72	172	11	F	Good
Marcus	8	83	185	3	F	Bad
Nigel	12	115	192	15	M	Good

Summary

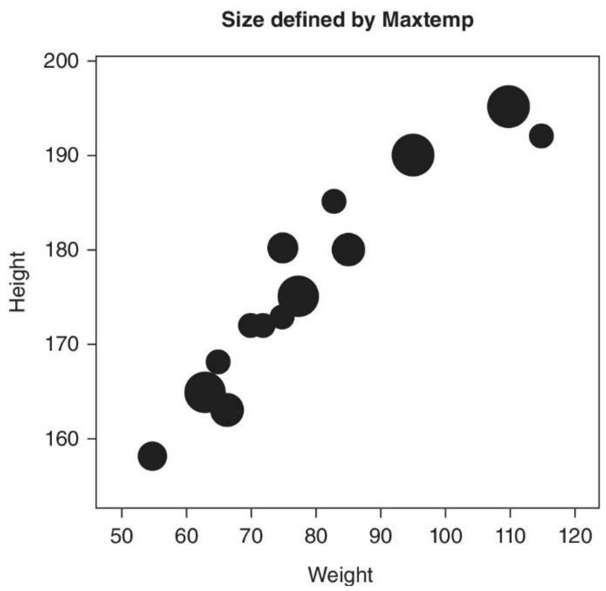
- Multivariate frequencies
- Multivariate data visualization
- Multivariate statistics
  - Location multivariate statistics
  - Dispersion multivariate statistics
- Final remarks

Multivariate frequencies

- The multivariate frequency values can be computed independently for each attribute
  - Thus, we can represent the frequency values for each attribute presenting them in a matrix like structure

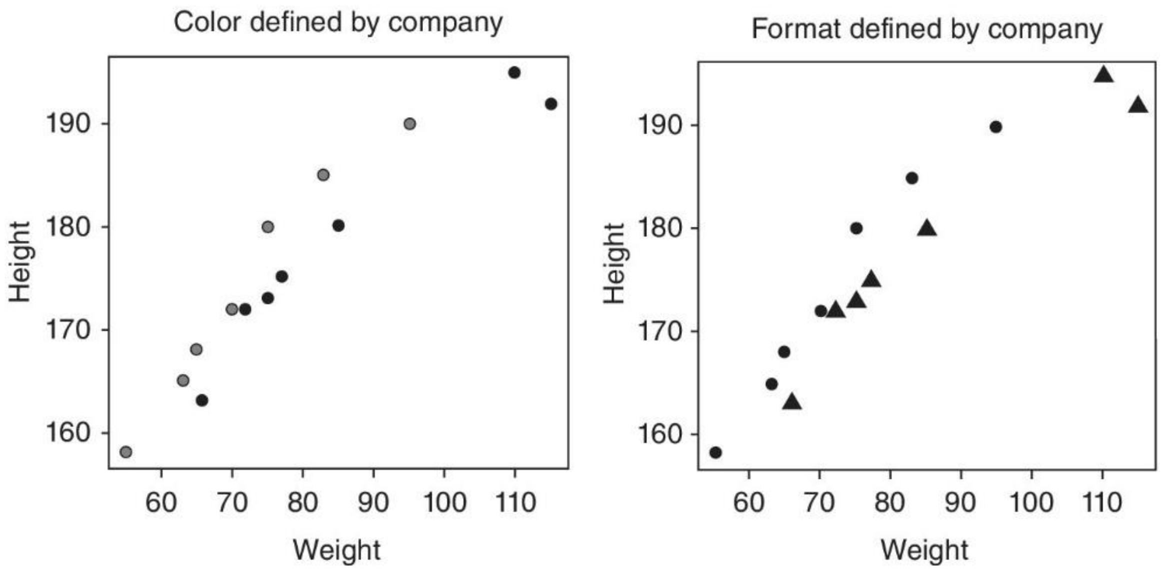
Multivariate data visualization

- When the multivariate data has **three attributes**, at least two of them quantitative, the data can still be visualized by a bivariate plot
  - This is done by associating the scale types of the values of the third attribute to how each data object is represented in the plot



Multivariate data visualization

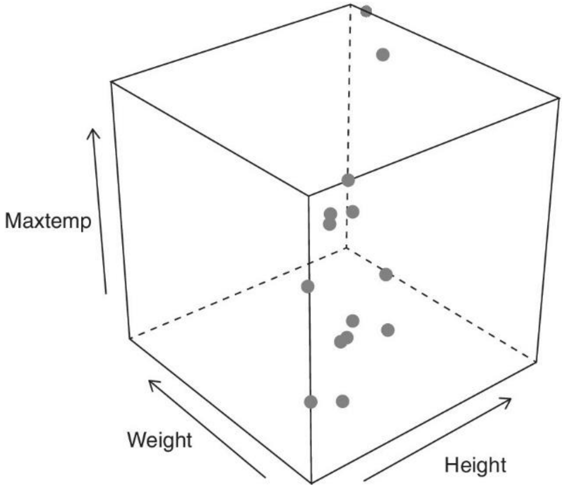
- If the third attribute is qualitative, its value can be represented in the plot by either the color or by the shape of the object in the plot
    - The number of colors or shapes will be the number of values the attribute can assume

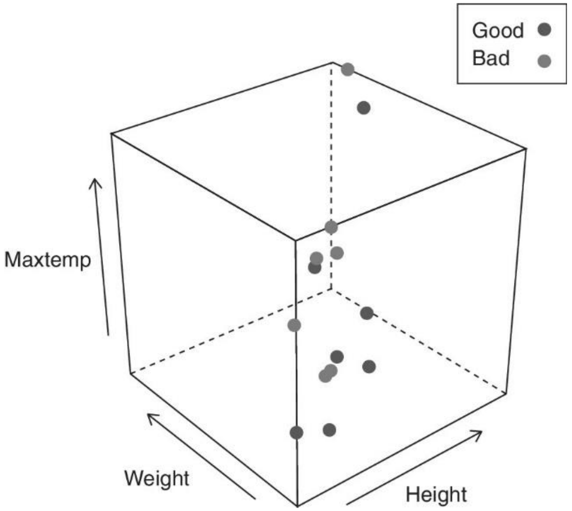


Multivariate data visualization

- Another approach to represent three attributes is to use a 3-dimensional plot

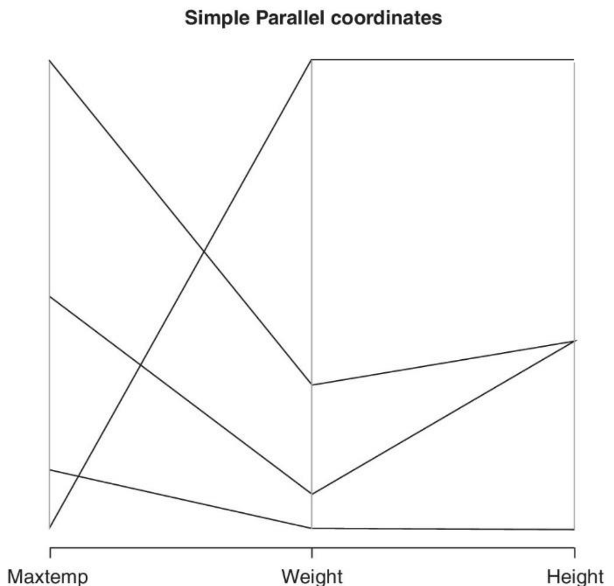
- A fourth attribute can e represented the same way a third attribute was represented in a bi-dimensional space

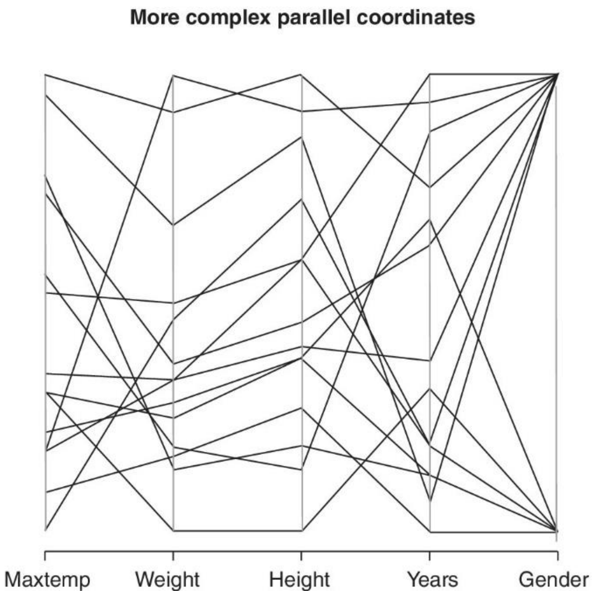


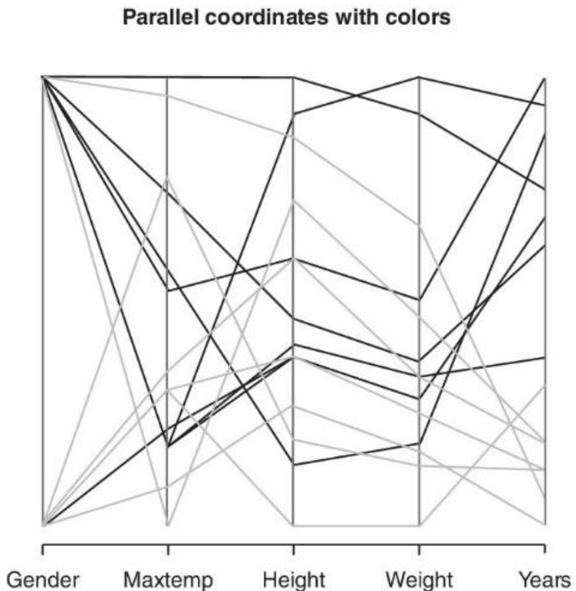


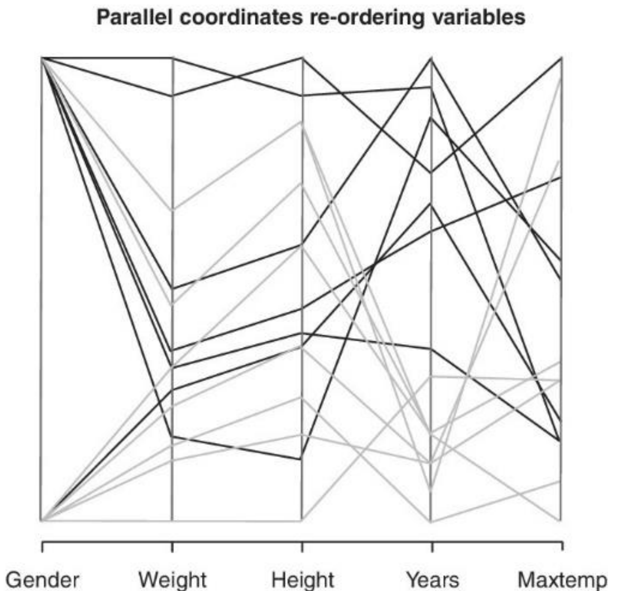
Multivariate data visualization

- Methods to visualize more than four attributes
  - One of the most popular of them is the **parallel coordinates**, also known as **profile plots**


Simple Parallel coordinates

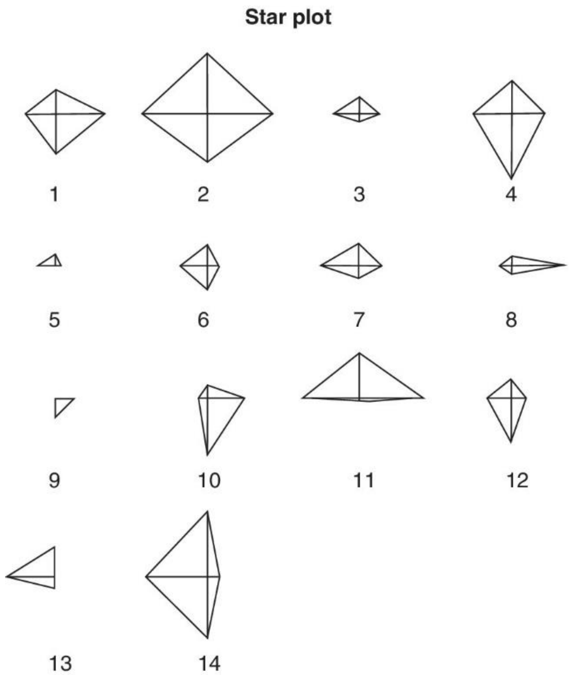

More complex parallel coordinates

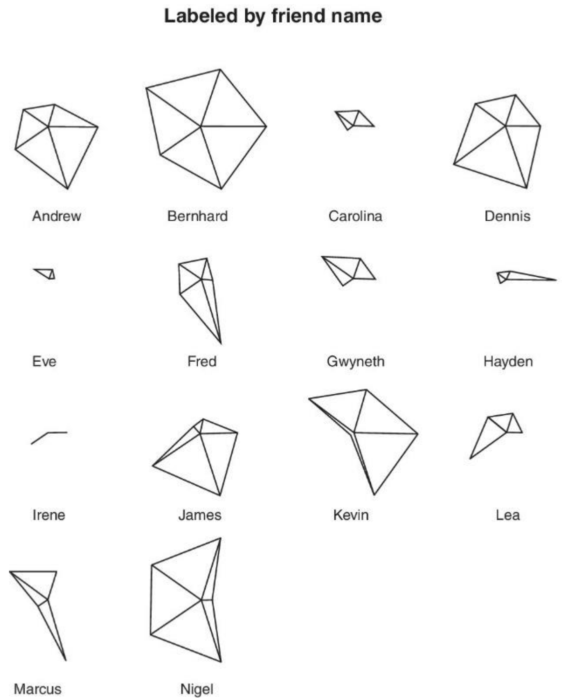

Parallel coordinates with colors

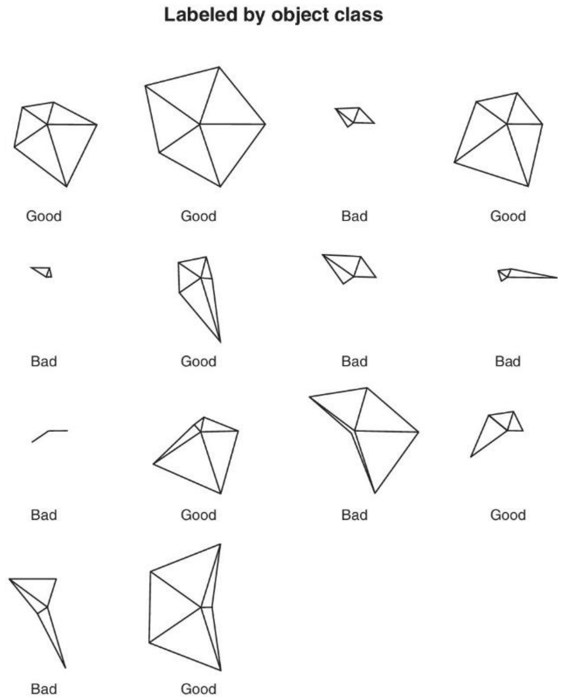

Parallel coordinates re-ordering variables

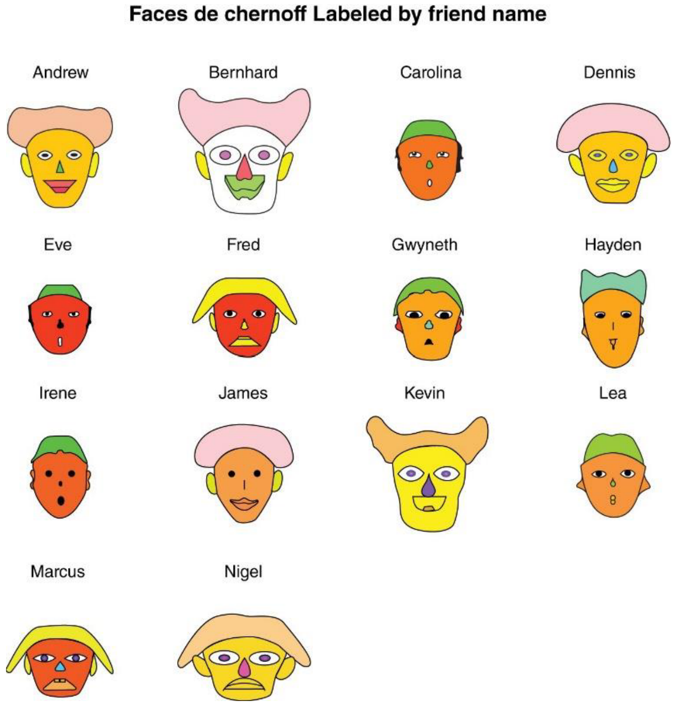
Multivariate data visualization

- Methods to visualize more than four attributes
  - **Star plots** and **Chernoff faces**









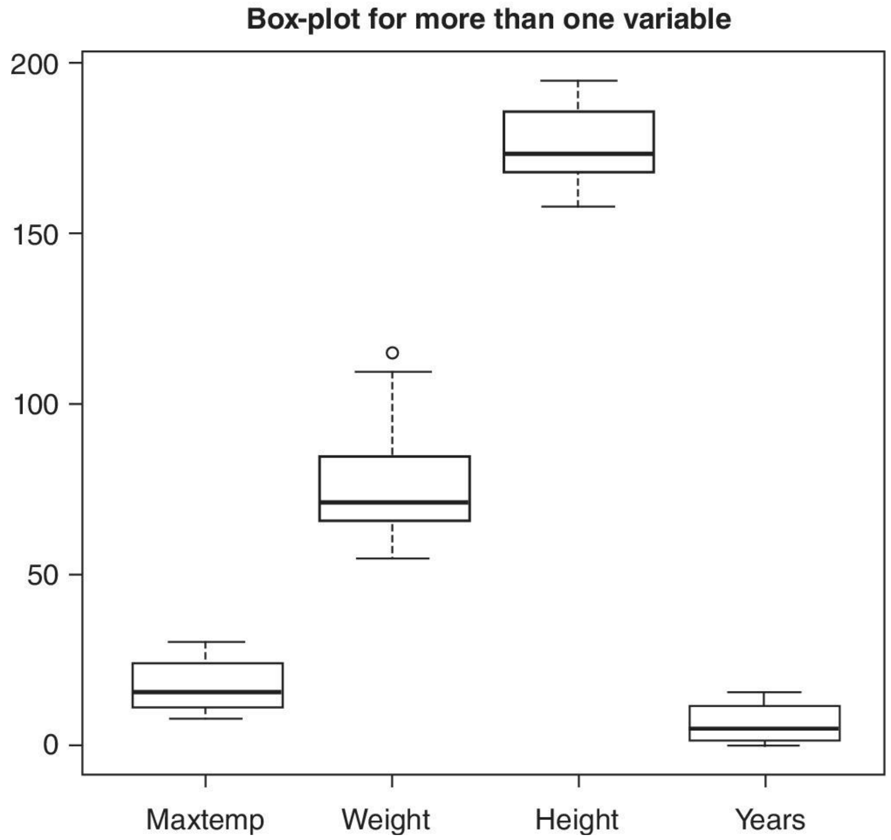
Location multivariate statistics

- To measure the location statistics of several attributes we just measure the location value for each attribute
    - Thus, we can represent the frequency values for each attribute presenting them in a matrix like structure

Location statistics	Max temp	Weight	Height	Years
min	8.00	55.00	158.00	0.00
max	31.00	115.00	195.00	16.00
average	18.14	79.00	176.29	7.14
mode	15.00	75.00	172.00	2.00
1st quartile	12.25	67.00	169.00	2.25
Median or 2nd quartile	15.50	75.00	174.00	5.50
3rd quartile	24.00	84.50	183.75	11.75

Location multivariate statistics

- To measure the location statistics of several attributes we just measure the location value for each attribute
  - If the number of attributes is not too large, a set of **box-plots**, one for each attribute, can be used



Dispersion multivariate statistics

- The extraction of some of the dispersion values for multivariate statistics, like amplitude, interquartile range, mean absolute deviation and standard deviation, can be also independently performed for each attribute

Dispersion statistics	**Max temp**	**Weight**	**Height**	**Years**
Amplitude	23.00	60.00	37.00	16.00
Interquartile range	11.75	17.50	14.75	9.50
$\overline{MAD}$	7.41	14.09	11.12	6.67
s	7.45	17.38	11.25	5.66

Dispersion multivariate statistics

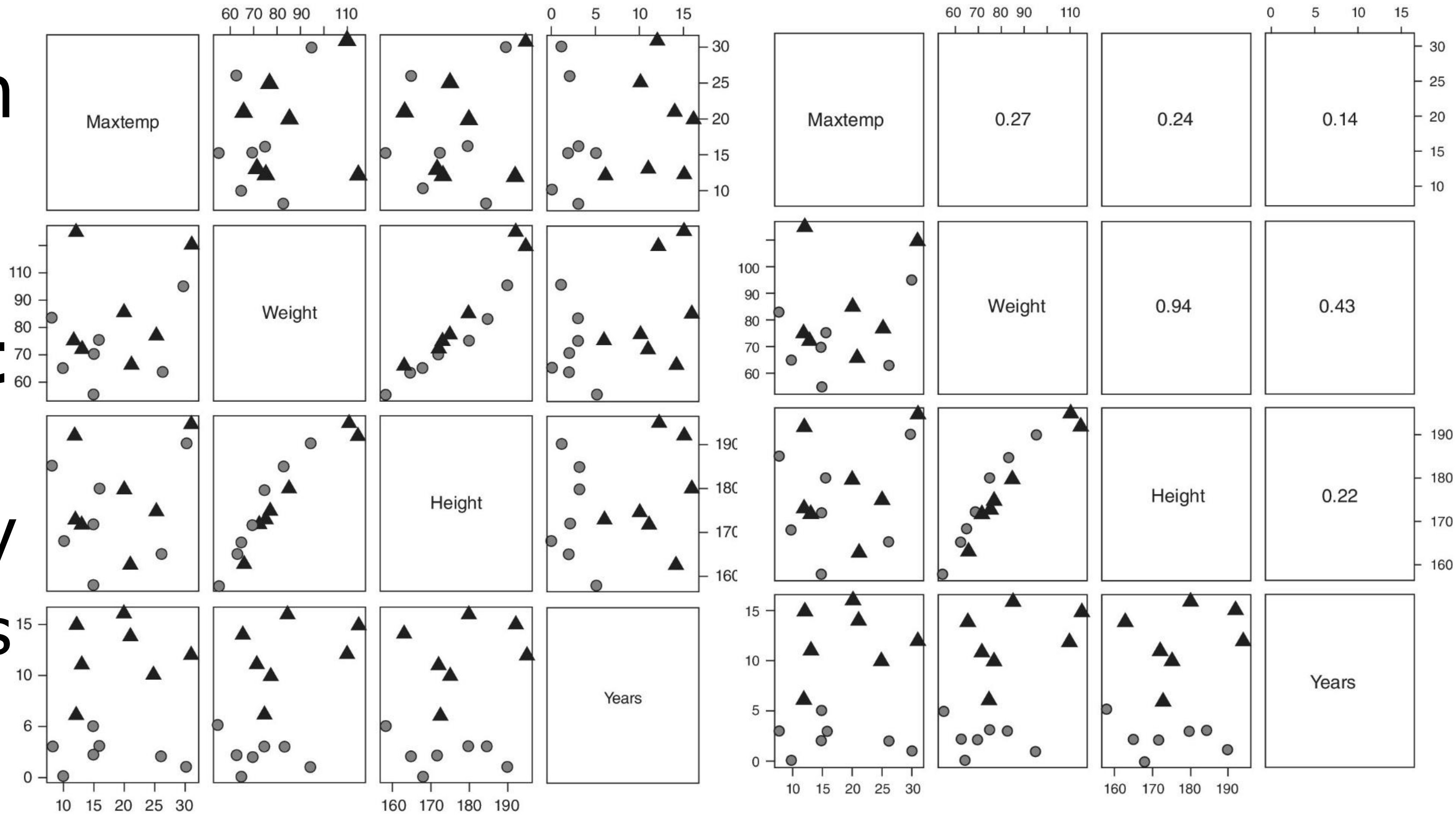
- The relation between two attributes is evaluated using covariance or correlation measures
  - The main diagonal of the **covariance matrix** shows the variance of each attribute
  - The matrices are symmetric: the values above the main diagonal are the same as the value below the main diagonal

Covariance	**Max temp**	**Weight**	**Height**	**Years**
Max temp	55.52	34.46	20.19	5.82
Weight	34.46	302.15	184.62	42.39
Height	20.19	184.62	126.53	14.03
Years	5.82	42.39	14.03	31.98

Pearson correlation	**Max temp**	**Weight**	**Height**	**Years**
Max temp	1.00	0.27	0.24	0.14
Weight	0.27	1.00	0.94	0.43
Height	0.24	0.94	1.00	0.22
Years	0.14	0.43	0.22	1.00

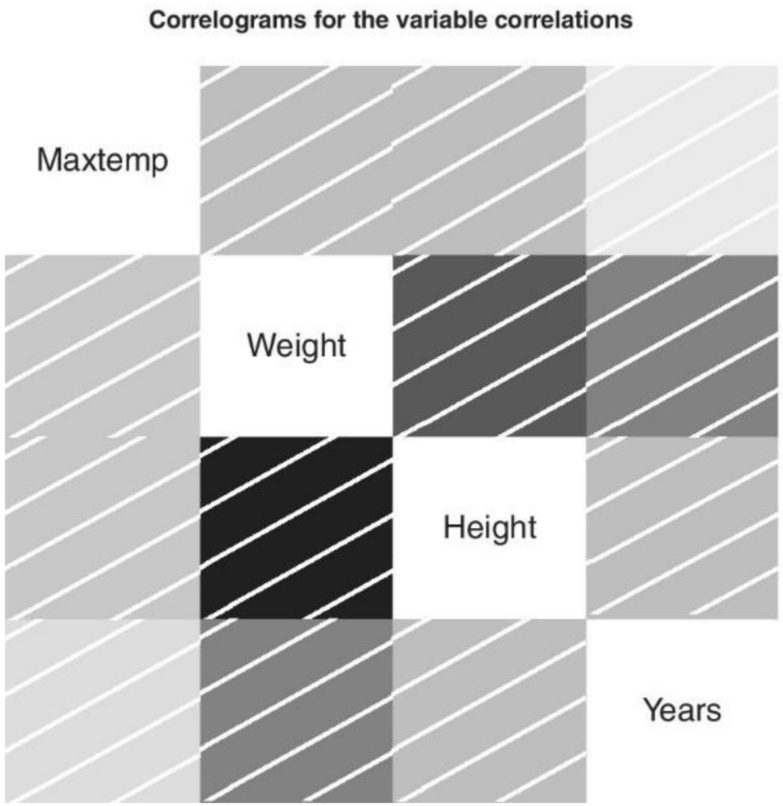
Dispersion multivariate statistics

- The relation between several attributes can be evaluated using a **scatter plot** matrix aka **Draftsman's** display
  - The shape represents the class label



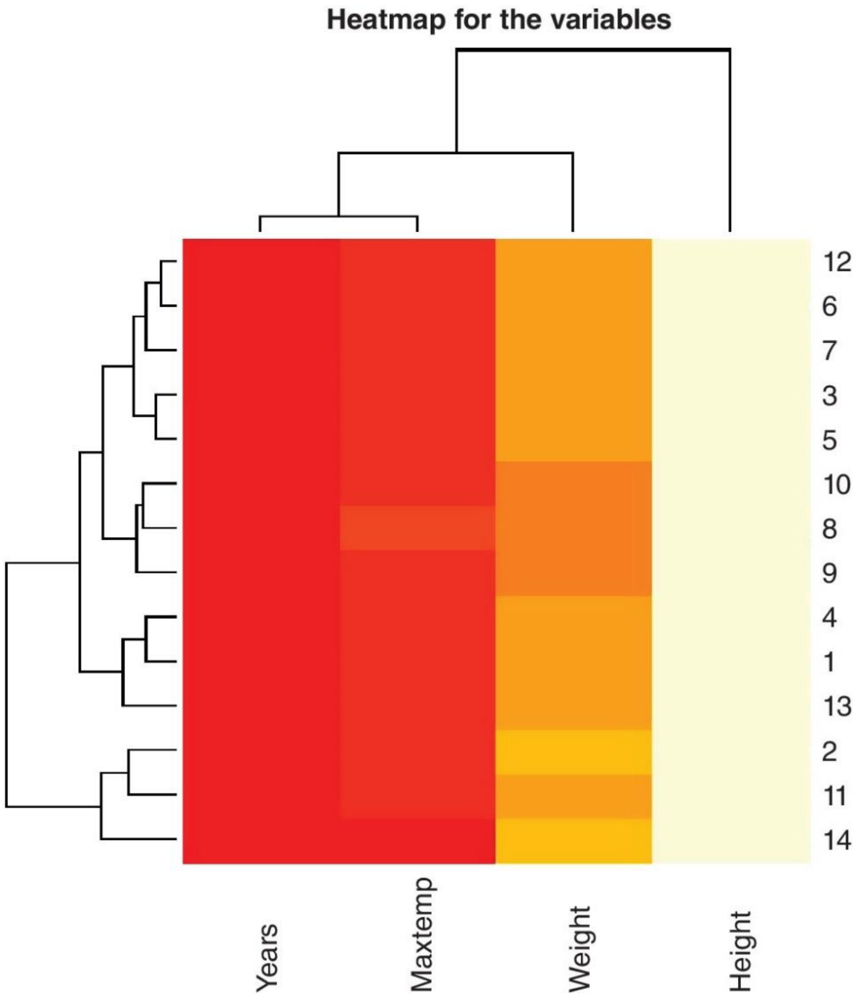
Dispersion multivariate statistics

- We can use a simpler plot to illustrate a summary of the information present in the scatter plot matrix
- For such, we can plot the linear correlation matrix using a **correlogram**
    - In this figure, the darker the square associated with two attributes, the more correlated they are



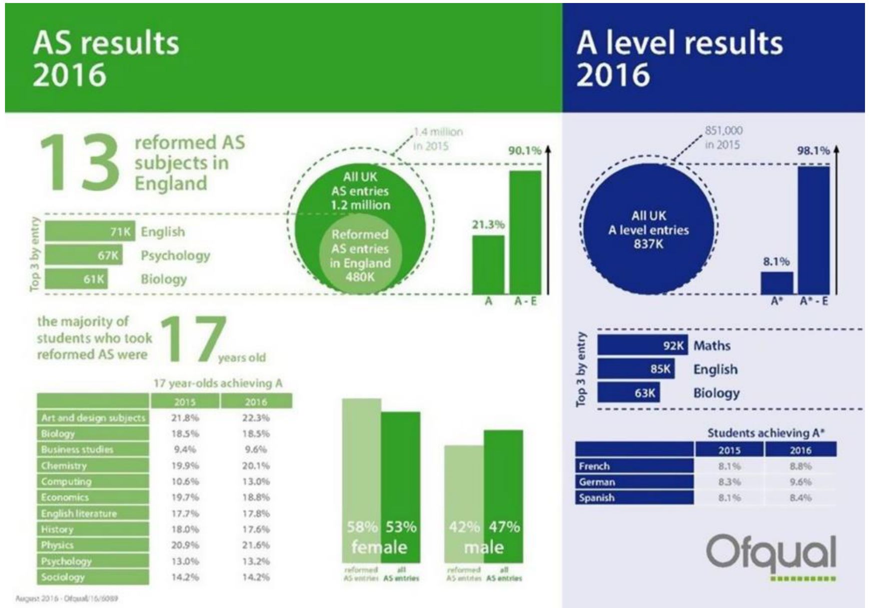
Dispersion multivariate statistics

- Another common plot for multivariate data is the **heatmap**, which can show how close the objects of a data set are considering the predictive attribute values
  - Each column (attribute) has boxes of a particular color. Although the boxes in each column have the same color, they can have different color tones
  - The tone of the color in a box is defined by the value of its corresponding predictive attribute in the object associated with its row



Dispersion multivariate statistics

- Currently, several studies graphically highlight important facts by using **infographics**
  - While data visualization is objective, automatically produced and can be applied to several data sets
  - Infographics are subjective, manually produced and customized for a particular data set



17 year-olds achieving A

	2015	2016
Art and design subjects	21.8%	22.3%
Biology	18.5%	18.5%
Business studies	9.4%	9.6%
Chemistry	19.9%	20.1%
Computing	10.6%	13.0%
Economics	19.7%	18.8%
English literature	17.7%	17.8%
History	18.0%	17.6%
Physics	20.9%	21.6%
Psychology	13.0%	13.2%
Sociology	14.2%	14.2%

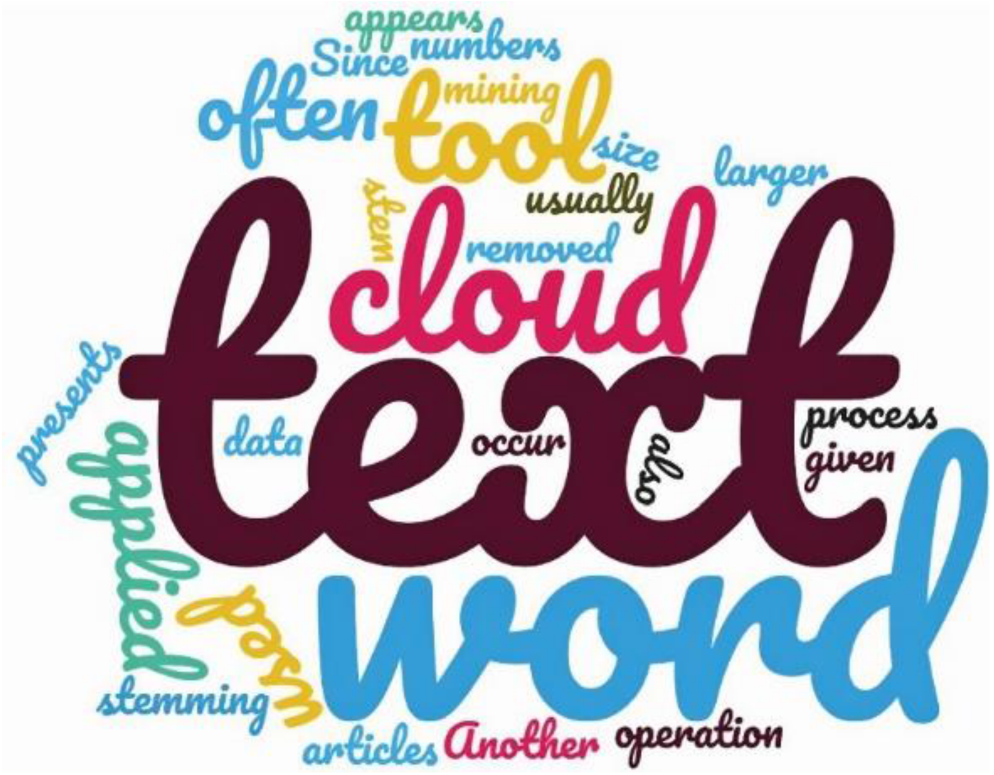
Students achieving A*

	2015	2016
French	8.1%	8.8%
German	8.3%	9.6%
Spanish	8.1%	8.4%

Infographic illustrating AS and A level results in England in 2016 (Contains public sector information licensed under the Open Government License v3.0.)

Dispersion multivariate statistics

- A visualization tool frequently used in text mining to illustrate text data is the **word cloud**, which presents how often each word appears in a given text
  - The higher the frequency of a word in the text, the larger its size in the word cloud



Dispersion multivariate statistics

- Since articles and prepositions occur very often in a text, and numbers are not text, they are usually removed before the word cloud tool is applied to a text
- Another text process operation, stemming, which substitutes a word in a text by its stem, is also applied to the text before the word cloud tool is used

Final remarks

- Descriptive multivariate analysis is more complex as the number of attributes increases
- It extends naturally from univariate and bivariate descriptive statistics
- The area of multivariate data visualization is an active research area

Questions?