Descriptive Statistics and Descriptive Multivariate Analysis

(1IK172 Introduction to Data Analytics)

Linnæus University

Summary

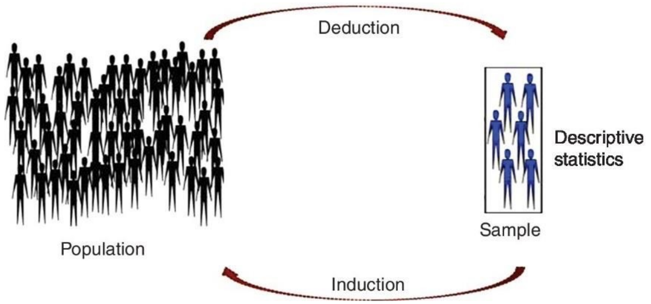
- Scale types
- Univariate analysis
  - Frequencies
  - Visualizations
  - Probability distributions
- Bivariate analysis
  - Relationship types
- Multivariate frequencies
- Multivariate data visualization
- Multivariate statistics
  - Location multivariate statistics
  - Dispersion multivariate statistics

Statistics

- **Population**
  - A set of similar instances/objects or events which is of interest for some question or experiment
  - E.g. all students of my school, all nails produced by a machine

- **Sample**
  - A set of a data collected and/or selected from a population by a defined procedure
  - E.g. a subset of the students of my school that answered to a survey, a subset of randomly selected nails produced by a machine

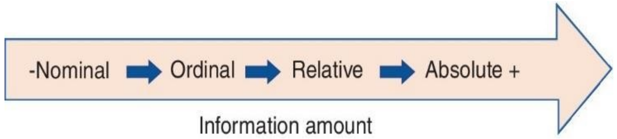
Statistics



- Deduction
  - Reasoning about the sample extracted from that population
  - Probabilities in about deduction
- Induction
  - Concerns reasoning about the population given a sample

- Descriptive statistics
  - Descriptive statistics are methods / techniques to describe or summarize samples in order to help humans to understand it

Scale Types



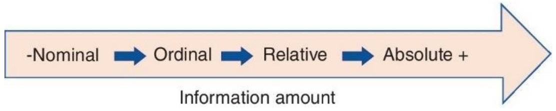
-Nominal ➡ Ordinal ➡ Relative ➡ Absolute +

Information amount

- **Qualitative** scales
  - **Nominal**: categorize data in a non-ordinal way
    - Operations: = and ≠
    - E.g. friend’s name and gender
  - **Ordinal**: categorize data in a ordinal way
    - Operations: =, ≠, <, >, ≤, and ≥
    - E.g. homestead, village, town, city, metropolis

- **Quantitative** scales
  - **Relative**: does not have an absolute zero
    - Operations: =, ≠, <, >, ≤, ≥, - and +
    - E.g. temperature
  - **Absolute**: has an absolute zero
    - Operations: =, ≠, <, >, ≤, ≥, -, +, / and ×
    - E.g. weight and height

Scale Types Example



-Nominal ➡ Ordinal ➡ Relative ➡ Absolute +

Information amount

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

Scales vs data types

- In software packages we must choose the data type for each attribute
  - Common types are text, character, factor, integer, real, float, timestamp, date or several others
  - A scale and a data type are different concepts despite related
  - For instance, a quantitative scale implies the use of numeric data types

- However, an attribute can be expressed as a number but the scale type can be qualitative
  - Think about an identity card you have with a numeric code
    - what kind of quantitative information does it have?
  - A code with letters could contain the same information

Descriptive Univariate Analysis: frequencies

- A frequency is basically a counter
- **Absolute frequency** counts how many times a value appears.
- **Relative frequency** counts the percentage of times that value appears.
- The **absolute cumulative frequency** is the number of occurrences less or equal than a given value
- The **relative cumulative frequency** is the percentage of occurrences less or equal than a given value

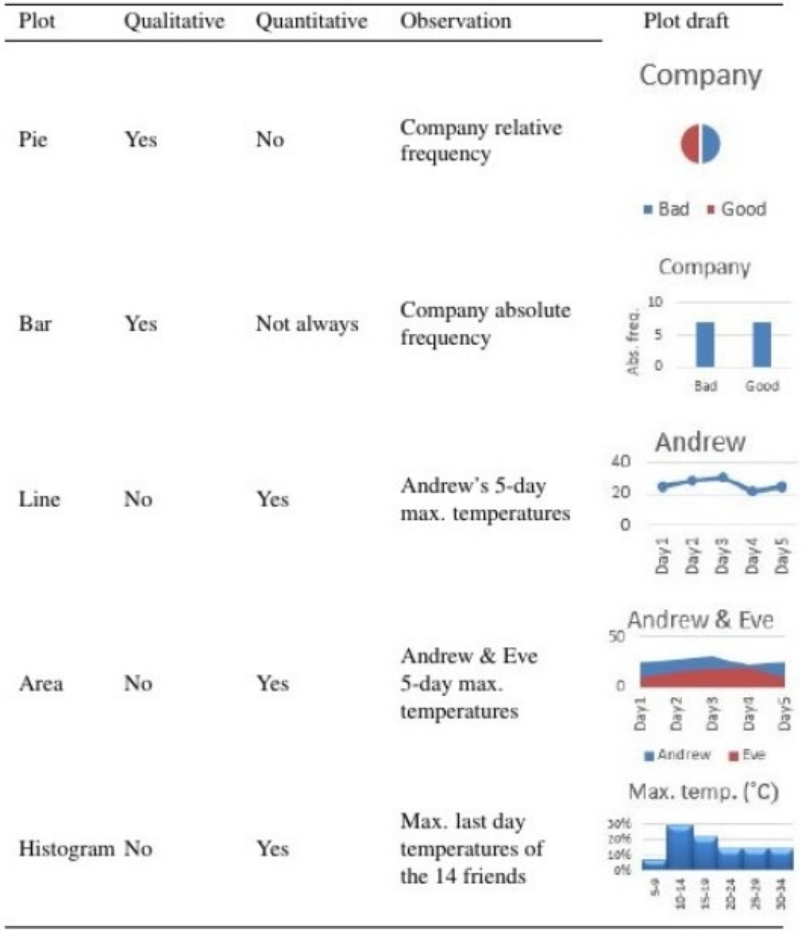
Descriptive Univariate Analysis: frequencies

Height	Abs. freq.	Rel. freq.	Abs. cum. freq.	Rel. cum. freq.
158	1	1/14=7.14%	1	1/14=7.14%
163	1	1/14=7.14%	2	2/14=14.29%
165	1	1/14=7.14%	3	3/14=21.43%
168	1	1/14=7.14%	4	4/14=28.57%
172	2	2/14=14.29%	6	6/14=42.86%
173	1	1/14=7.14%	7	7/14=50.00%
175	1	1/14=7.14%	8	8/14=57.14%
180	2	1/14=14.29%	10	10/14=71.43%
185	1	1/14=7.14%	11	11/14=78.57%
190	1	1/14=7.14%	12	12/14=85.70%
192	1	1/14=7.14%	13	13/14=92.86%
195	1	1/14=7.14%	14	14/14=100.00%

Descriptive Univariate Analysis: frequencies

- The relative frequencies define distribution functions, i.e., they describe how data are distributed
- Distribution functions are said empirical when they are obtained from a sample
- A discrete attribute, like one of the integer data type, has a probability mass (distribution) function
- A continuous attribute, like one of the real data type, has a density probability function

Descriptive Univariate Analysis: data visualization

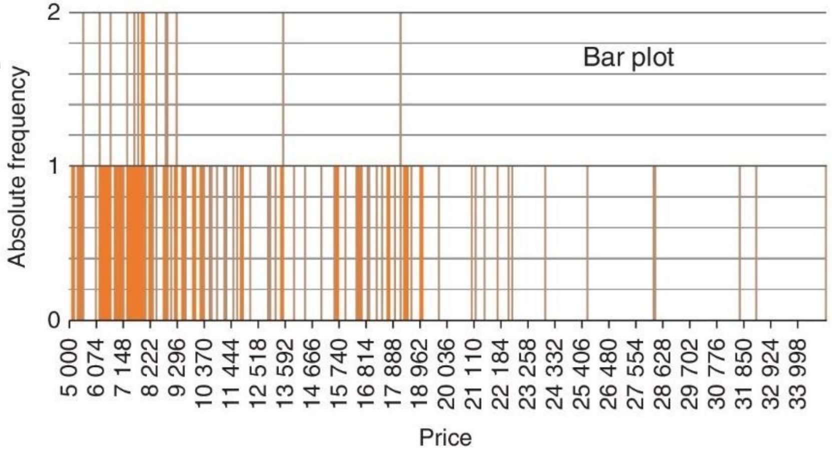


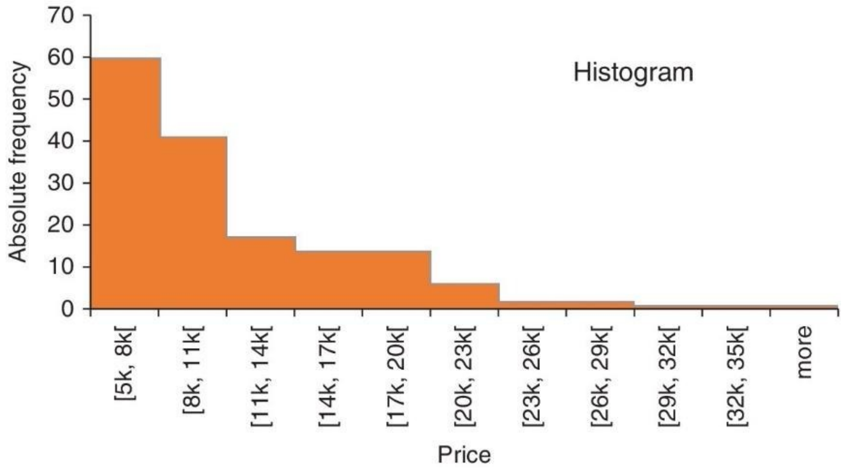
Plot	Qualitative	Quantitative	Observation	Plot draft
Pie	Yes	No	Company relative frequency	Company ■ Bad ■ Good
Bar	Yes	Not always	Company absolute frequency	Company Abs. freq. 10 5 0 Bad Good
Line	No	Yes	Andrew's 5-day max. temperatures	Andrew 40 20 0 Day1 Day2 Day3 Day4 Day5
Area	No	Yes	Andrew & Eve 5-day max. temperatures	Andrew & Eve 50 0 Day1 Day2 Day3 Day4 Day5 ■ Andrew ■ Eve
Histogram	No	Yes	Max. last day temperatures of the 14 friends	Max. temp. (˚C) 30% 20% 10% 0% 5-9 10-14 15-19 20-24 25-29 30-34

- Pie chart: it is used typically for nominal scales
- Bar chart: It is used typically for qualitative scales or quantitative scales with a limited number of values
- Line chart: they are specially used to deal with the notion of time
- Area charts: are specially used to compare time series and distribution functions
- Histograms: are used to represent empirical distributions for attributes with a quantitative scale

Descriptive Univariate Analysis: data visualization

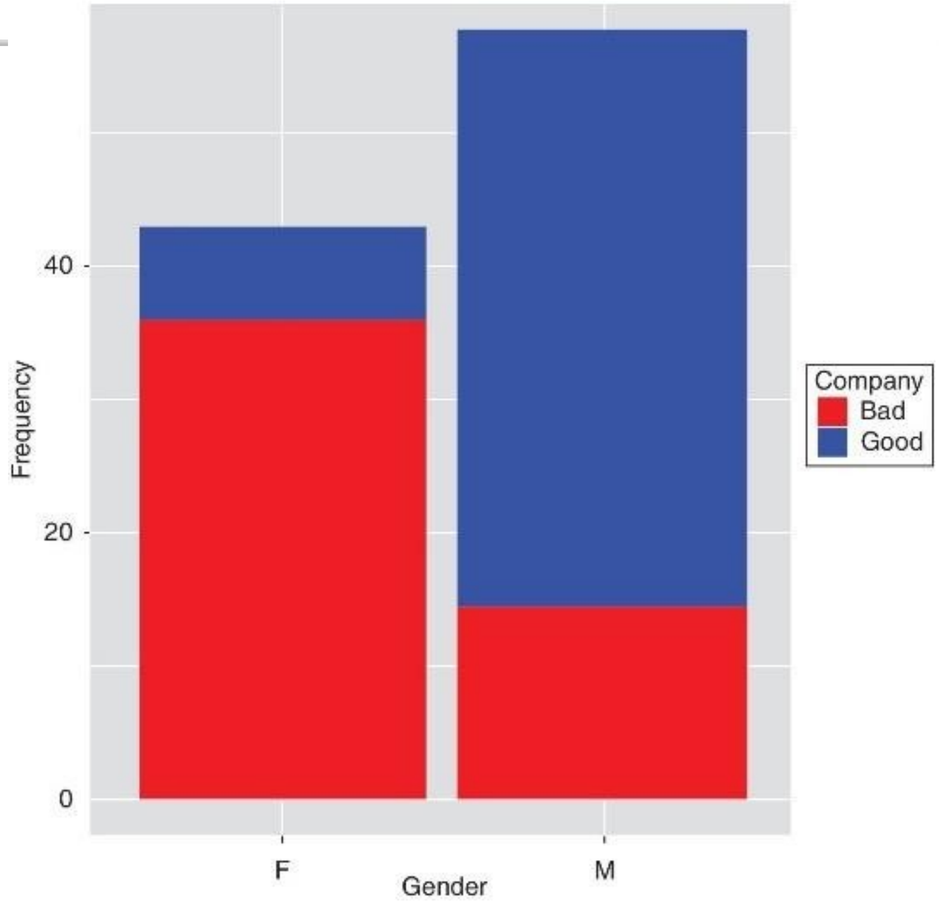
- An important decision to draw a histogram is to define the number of cells
- The most advisable value is problem dependent
- As rule of thumb you can use a number around the square root of the number of values





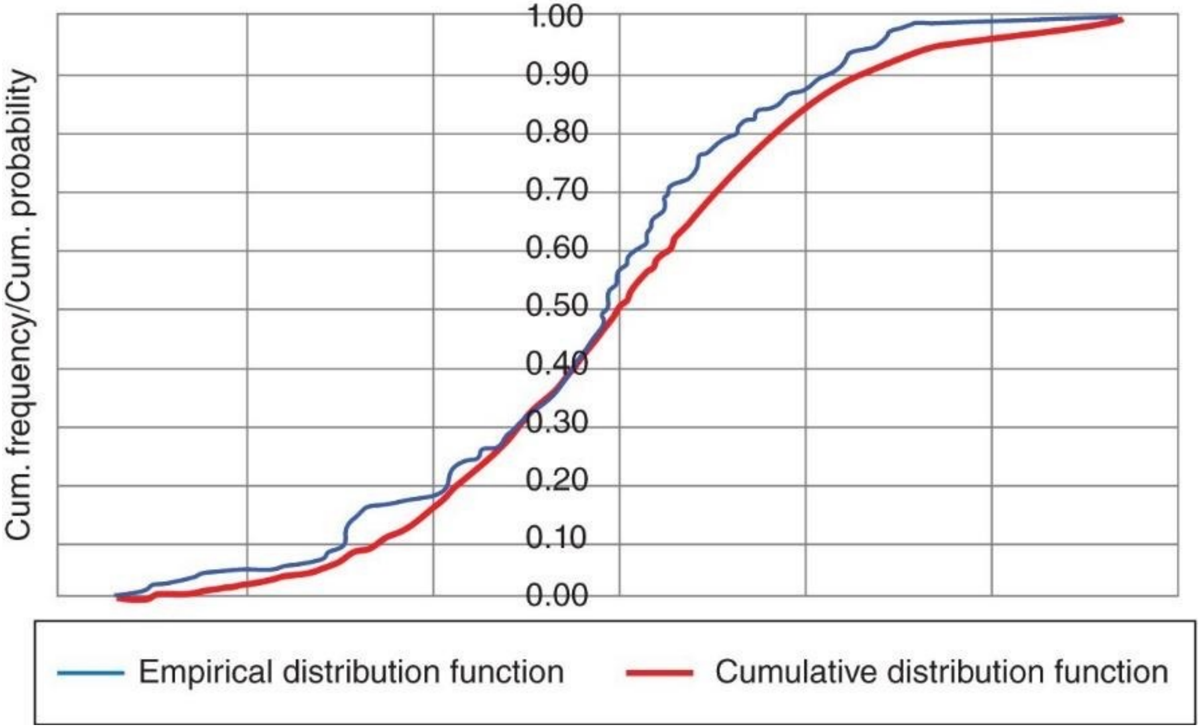
Descriptive Univariate Analysis: data visualization

- In a histogram, we can also separate the distributions for the values of some other attribute
- This is illustrated in the figure where the frequencies for the target value of "company" is split by gender



Descriptive Univariate Analysis: data visualization

- **Empirical distributions** are based in samples
- **Probability distributions** are about populations



Descriptive Univariate Analysis: statistics

- A statistic is a descriptor
- It describes numerically a characteristic of the sample or the population
- There are two main groups of univariate statistics:
  - **Location statistics**
  - **Dispersion statistics**

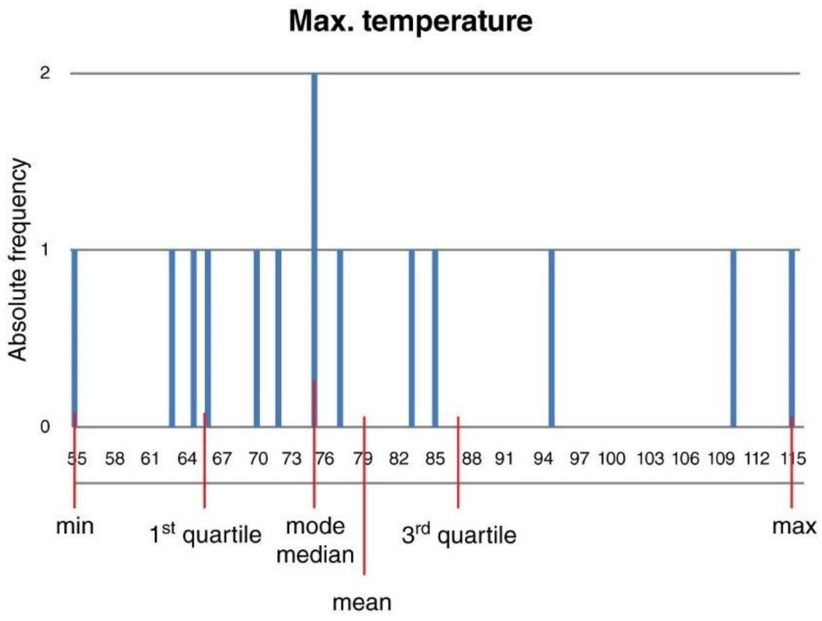
- **Location statistics**:
  - Minimum: is the lowest value
  - Maximum: is the largest value
  - Mean: is the average value
  - Mode: is the most frequent value
  - The value that is larger than:
    - 25% of all values is the 1st quartile
    - 50% of all values is the median or 2nd quartile (Median)
    - 75% of all values is the 3rd quartile

Example Location statistics

- Let us use as example the attribute weight from our data set

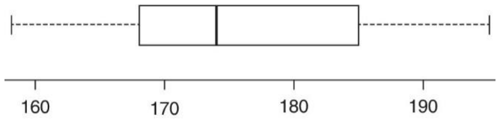
Location statistic	Weight (kg)
Min	55.00
Max	115.00
Mean or average	79.00
Mode	75.00
1st quartile	65.75
2nd quartile or mode	75.00
3rd quartile	87.50

- Graphical representation of the statistics



Descriptive Univariate Analysis: statistics

- **Box-plots** present the minimum, the 1st quartile, the median, the 3rd quartile and the maximum statistics, by this order, bottom-up or from left to right
  - The attribute height

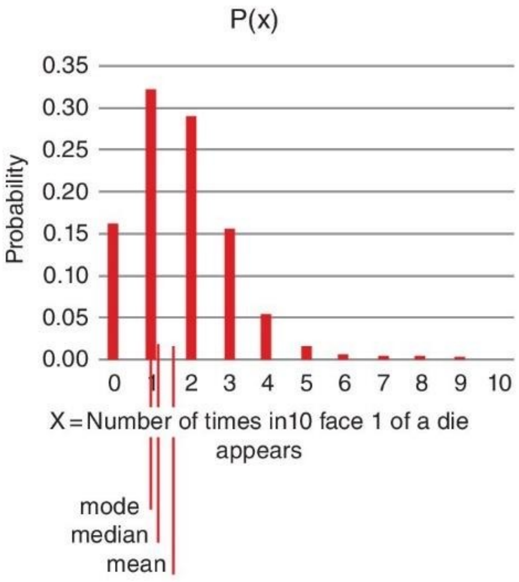


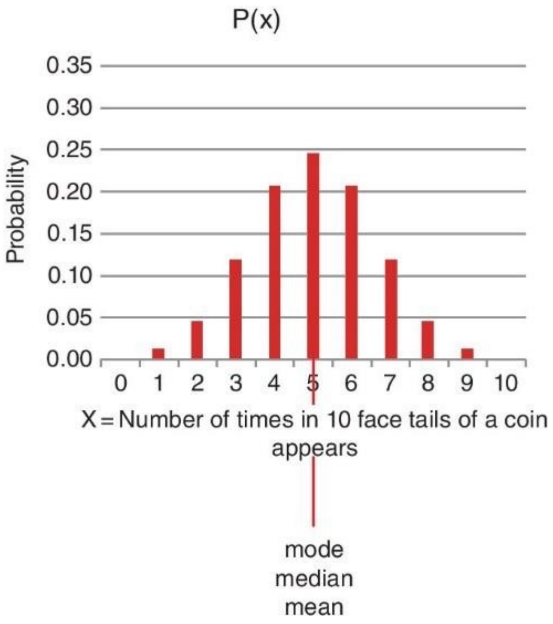
- Mean (or average), median and mode are known as **measures of central tendency**, because return a central value from a set of values

Location statistic	Nominal	Ordinal	Quantitative
Mean	No	Eventually	Yes
Median	No	Yes	Yes
Mode	Yes	Yes	Yes

Descriptive Univariate Analysis: statistics

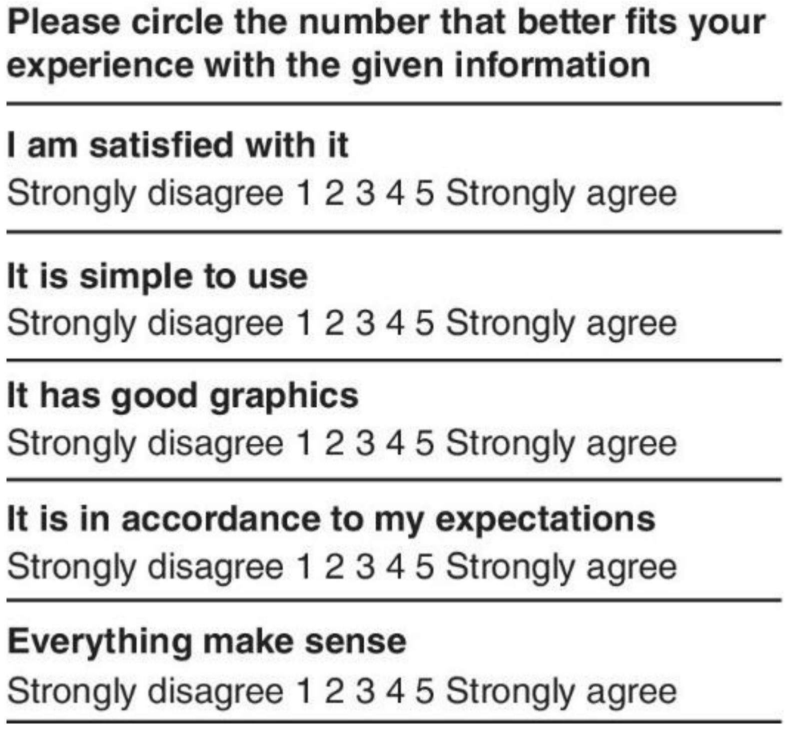
- Box-plots can also be used to describe the symmetry/ skewness of an attribute
- The median or the mode are more robust as a central tendency statistic than the mean in the presence of extreme values or strongly skewed distributions





Descriptive Univariate Analysis: statistics

- Can the mean be used in ordinal scales?
- This is strongly arguable but there are examples of its use with numeric ordinal scales such as the Likert scale
- The Likert uses an ordered scale, e.g., integers from 1 (highest disagreement) to 5 (highest agreement)



Please circle the number that better fits your experience with the given information

I am satisfied with it
Strongly disagree 1 2 3 4 5 Strongly agree

It is simple to use
Strongly disagree 1 2 3 4 5 Strongly agree

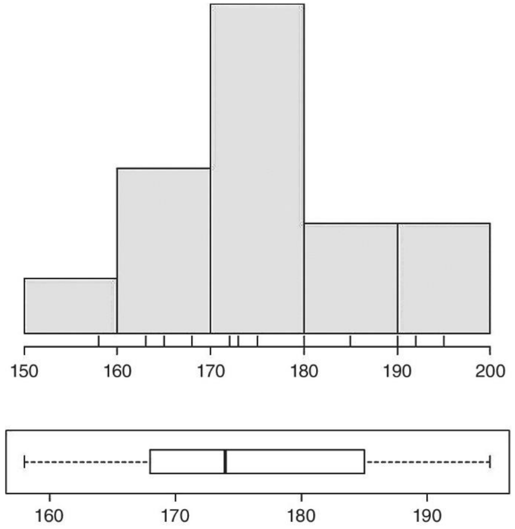
It has good graphics
Strongly disagree 1 2 3 4 5 Strongly agree

It is in accordance to my expectations
Strongly disagree 1 2 3 4 5 Strongly agree

Everything make sense
Strongly disagree 1 2 3 4 5 Strongly agree

Descriptive Univariate Analysis: statistics

- Plots can also be combined
  - An example with the attribute length



- There is only one value for the mean of a population
- There is only one value for the mean of a sample but can exist several samples from a single population
- The population mean and the sample mean are calculated in the same way but are differently represented:
  - $\mu_x$is the mean population of x
  - $\bar{x}$ is a mean sample of x

Descriptive Univariate Analysis: statistics

- Dispersion statistic measures how distant the different values are
- **Dispersion statistics**:
  - **Amplitude**: is the difference between the maximum and the minimum values
  - **Interquartile range**: is the difference between the values of the 3rd and 1st quartiles
- Dispersion statistics (cont.):
  - **Mean absolute deviation** is a measure for the mean absolute distance between the observations and the mean
    - Its math formula for the population is:

$$MAD_x = \frac{\sum_{i=1}^{n} |x_i - \mu_x|}{n}$$

    - Its math formula for a sample is:

$$\overline{MAD}_x = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n-1},$$

Descriptive Univariate Analysis: statistics

- Dispersion statistics (cont):
  - Standard deviation: is another measure for the typical distance between the observations and their mean
    - Its math formula for the population is: $\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{n}}$
    - Its math formula for a sample is:

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

      - The square of the standard deviation is named variance

- Using again as example the weight attribute, dispersion statistics are as shown in the table

Dispersion statistic	**Weight (kg)**
Amplitude	60.00
Interquartile range	21.75
$\overline{MAD}$	14.31
s	17.38

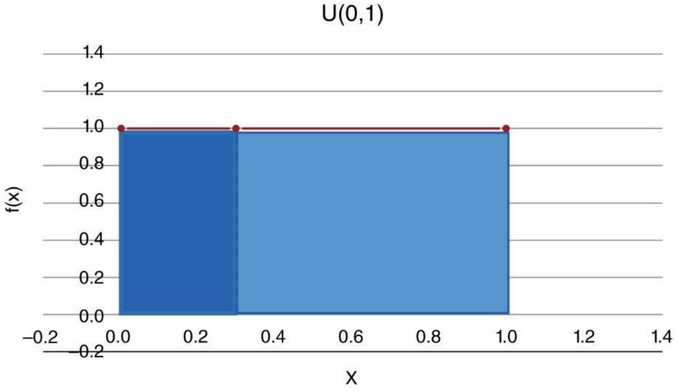
Descriptive Univariate Analysis: common univariate probability distributions

- Different events of our life follow already studied distributions
- E.g. the height of adult men, the value of a random number, or the number of cars passing in a given highway toll

- We present two of these distributions:
  - The Uniform distribution
  - The Normal distribution, also known as the Gaussian
- Both are continuous distributions and have known probability density functions

Descriptive Univariate Analysis: common univariate probability distributions

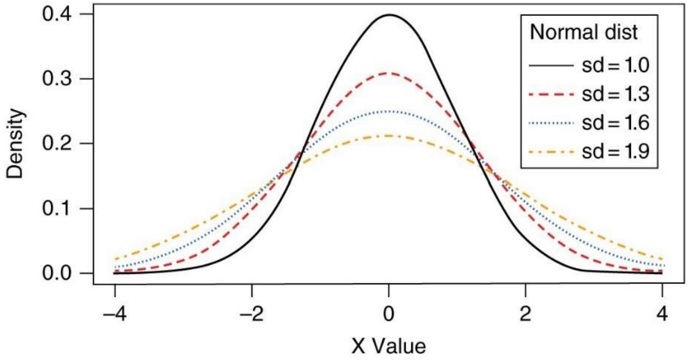
- An attribute x that follows the **uniform distribution** with parameters a and b, has equal frequency of occurrence of values in any interval of a given size



- $x \sim U(a, b)$
- $P(x < x_0) = \begin{cases} 0, if\ x_0 < a \\ \frac{x_0 - a}{b - a}, if\ a \le x_0 \le b \\ 1, if x_0 > b \end{cases}$
- $\mu_x = \frac{a+b}{2}$ $\sigma_x^2 = \frac{(b-a)^2}{12}$

Descriptive Univariate Analysis: common univariate probability distributions

- The Normal distribution
  - Physical quantities that are expected to be the sum of many independent factors (e.g., the men' height or the perimeter of 30 years old Quercus Rubra) typically have approximately Normal distributions



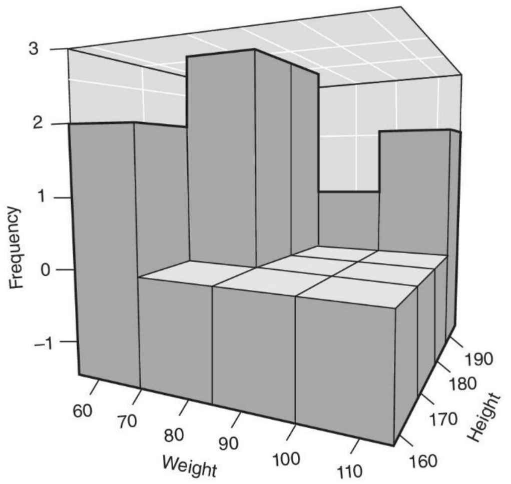
- The Normal distribution is a symmetric and continuous distribution with two parameters:
  - The mean localizes the highest point of the bell like distribution
  - The standard deviation defines how thin or larger the bell form of the distribution is
- $x \sim N(\mu_x, \sigma_x)$

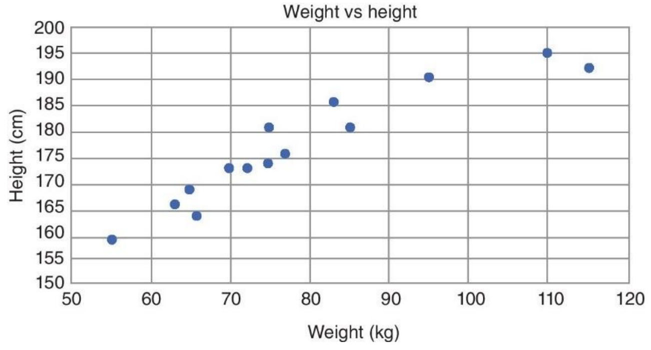
Descriptive bivariate analysis

- Talking about pairs of attributes, the relative behaviour between them
- Cases according to the scale types of the attributes:
  1. When the two attributes are quantitative
  2. When one of the attributes is qualitative and the other is quantitative
  3. When the two attributes are qualitative, at least one of them nominal
  4. When the two attributes are ordinal

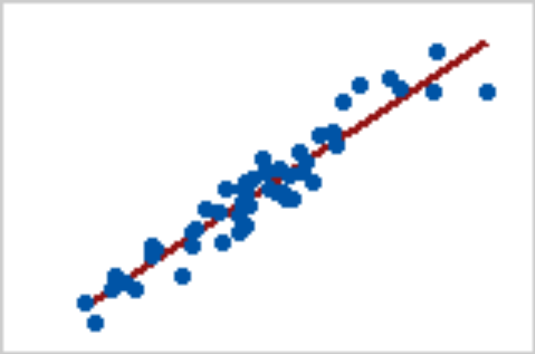
Descriptive bivariate analysis

- **When the two attributes of the pair are quantitative**
- There are several visualization techniques able to visually show the distribution of points with two quantitative attributes
  - One of these techniques is an extension of histograms, named 3-dimensional histograms
  - Another are the scatter plots

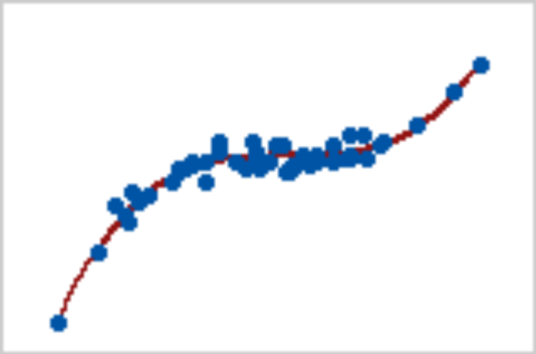




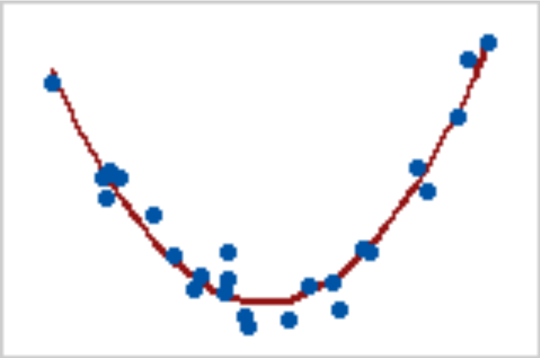
Bivariate relationship types



Linear



Monotonic



Non-Linear

Source: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/linear-nonlinear-and-monotonic-relationships/

Descriptive bivariate analysis

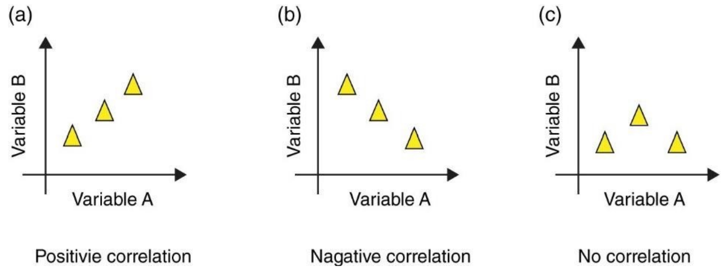
- **Covariance**
  - Measures the degree of presence of linear relation between two attributes
  - Sample covariance:

$$s_{ij} = \operatorname{cov}(x_i, x_j) = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

  - The scale of the attributes influence the covariance values obtained

Descriptive bivariate analysis

- **Pearson correlation**
  - Sample Pearson correlation
    - $r_{ij} = cor\left(x_i, x_j\right) = \frac{cov(x_i,x_j)}{s_i \times s_j}$



Positivie correlation

Nagative correlation

No correlation

  - Is scale independent: values always between [-1, 1]
  - If the points form:
    - an increasing line, the Pearson correlation coefficient will be 1
    - a decreasing line, its value will be -1
    - a horizontal line or a cloud without increasing or decreasing tendency, its value will be 0

Descriptive bivariate analysis

- The **Spearman's rank correlation**, as the name suggests, is based on rankings
- Compares how similar are the ranking positions of the values of the two attributes

$$r_{x,y} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[(rx_i - \overline{rx}) \times (ry_i - \overline{ry})]}{s_{rx} \times s_{ry}},$$

- Ranking columns are calculated for both variable. For resolving ties we can use mean, min, or max value

Example

- Pearson correlation
  - $r_{weight,height} = 0.94$
- Spearman's rank correlation
  - $\rho_{weight,height} = 0.96$
- Note the tie breaker ranked values. Mean is used here.

Friend	Weight (kg)	Height (cm)	Ranked Weight	Ranked Height
Andrew	77	175	9.0	8.0
Bernhard	110	195	13.0	14.0
Carolina	70	172	5.0	5.5
Dennis	85	180	11.0	9.5
Eve	65	168	3.0	4.0
Fred	75	173	7.5	7.0
Gwyneth	75	180	7.5	9.5
Hayden	63	165	2.0	3.0
Irene	55	158	1.0	1.0
James	66	163	4.0	2.0
Kevin	95	190	12.0	12.0
Lea	72	172	6.0	5.5
Marcus	83	185	10.0	11.0
Nigel	115	192	14.0	13.0

Descriptive bivariate analysis

- **When one of the attributes is qualitative and the other is quantitative**
  - Box-plots can be used as previously discussed using one box plot for the values of the quantitative attribute per each different value of the qualitative attribute

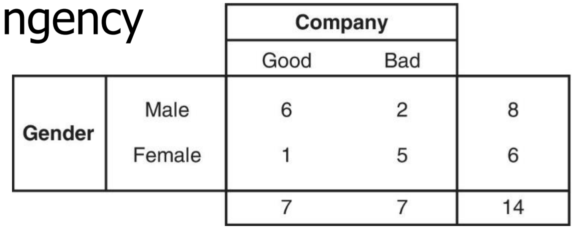
Descriptive bivariate analysis

- **When one of the attributes is qualitative and the other is quantitative**
  - **Contingency tables**
    - They have a matrix like format, i.e., cells in a square with labels in the left and in the top
    - In the right most column are the totals per row while in the bottom most row are the totals per column
    - The bottom right corner has the total number of values

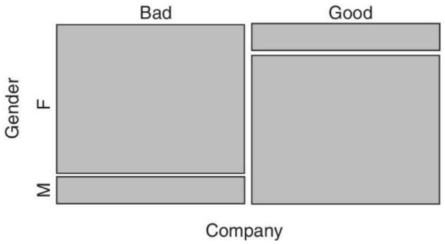
Descriptive bivariate analysis

- Two qualitative attributes, at least one of them nominal
  - **Mosaic plots**
    - Show the same information than contingency tables but in a more appealing visual way
    - The areas displayed are proportional to their relative frequency

Contingency



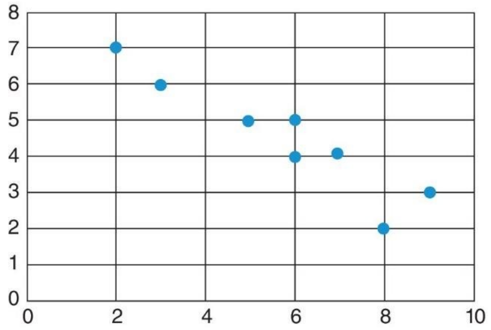
		Company		
		Good	Bad	
Gender	Male	6	2	8
	Female	1	5	6
		7	7	14



Mosaic

Descriptive bivariate analysis

- **When the two attributes are ordinal**
  - Any of the methods previously described to bivariate analysis can also be used in the presence of two ordinal attributes:



- The Spearman's rank correlation should be used instead of the Pearson correlation
- Scatter plots with ordinal attributes
  - Use the jitter effect, which add a random deviation to the values, in order to avoid that all points with the same values are represented as a unique point
- Contingency tables can be used and mosaic plots too
  - The values should be in increasing order

Descriptive Multivariate Analysis

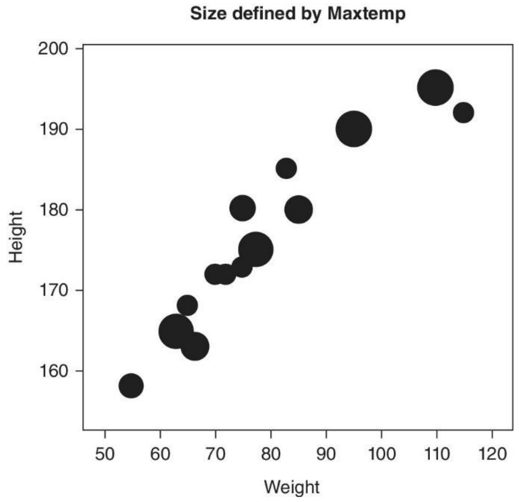
Friend	Max temp	Weight	Height	Years	Gender	Company
Andrew	25	77	175	10	M	Good
Bernhard	31	110	195	12	M	Good
Carolina	15	70	172	2	F	Bad
Dennis	20	85	180	16	M	Good
Eve	10	65	168	0	F	Bad
Fred	12	75	173	6	M	Good
Gwyneth	16	75	180	3	F	Bad
Hayden	26	63	165	2	F	Bad
Irene	15	55	158	5	F	Bad
James	21	66	163	14	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

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

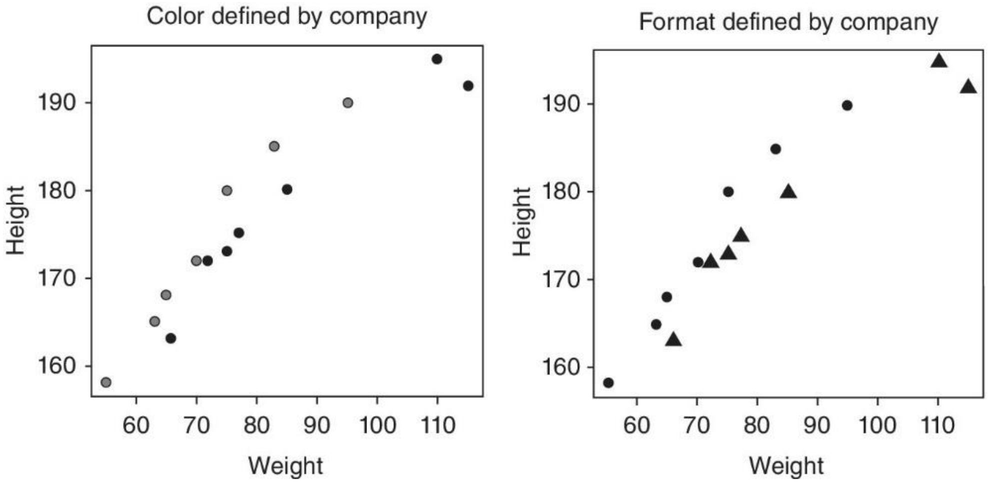
Multivariate data visualization

- 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

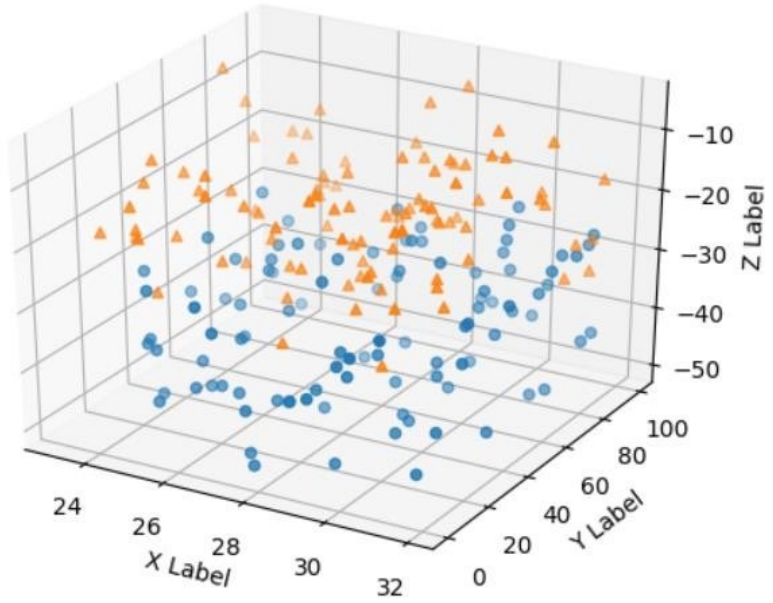


Multivariate data visualization

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



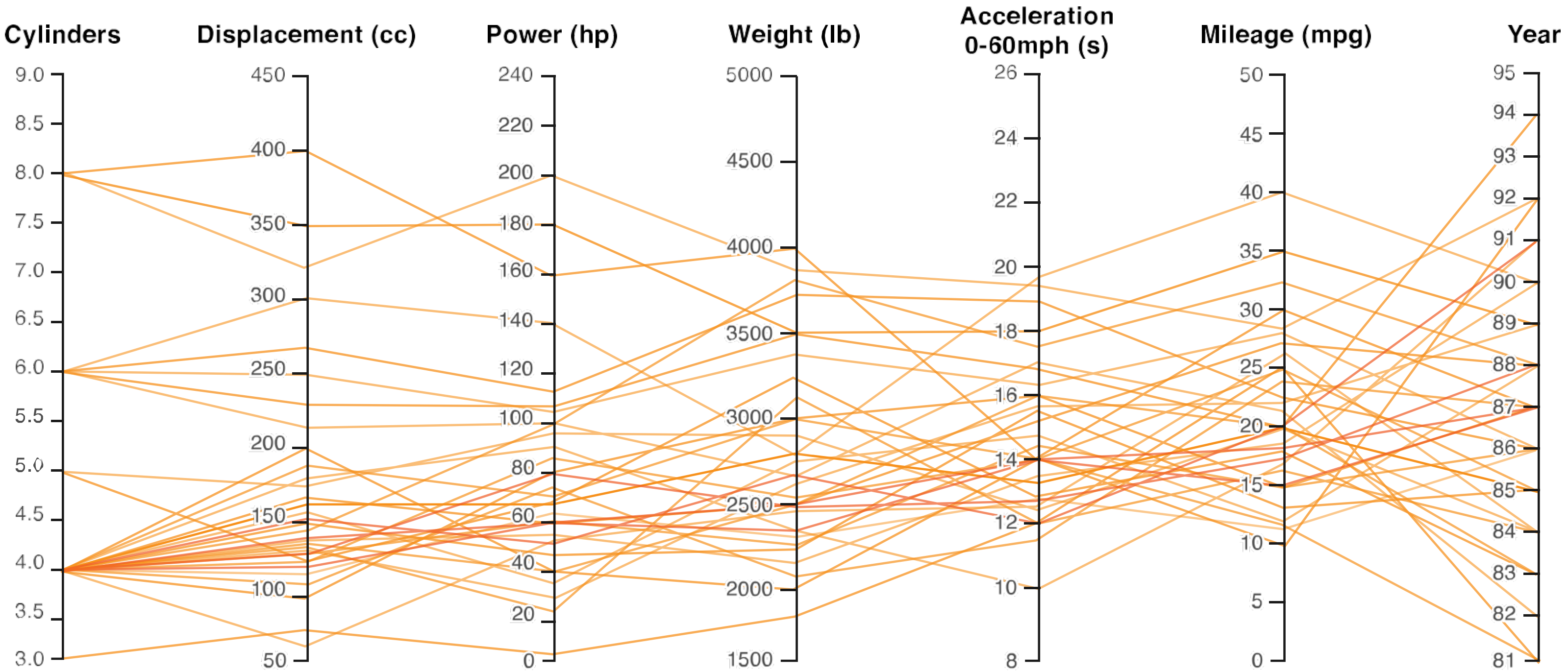
Multivariate data visualization



- Another approach to represent three attributes is to use a 3-dimensional plot
- A fourth attribute can be represented the same way a third attribute was represented in a bi-dimensional space
- We can also map a surface or wireframe on the points

Multivariate data visualization

- **Parallel Coordinates**
  - Each attribute is a vector, the vectors are of the same length and ranges between min-max values of the attribute

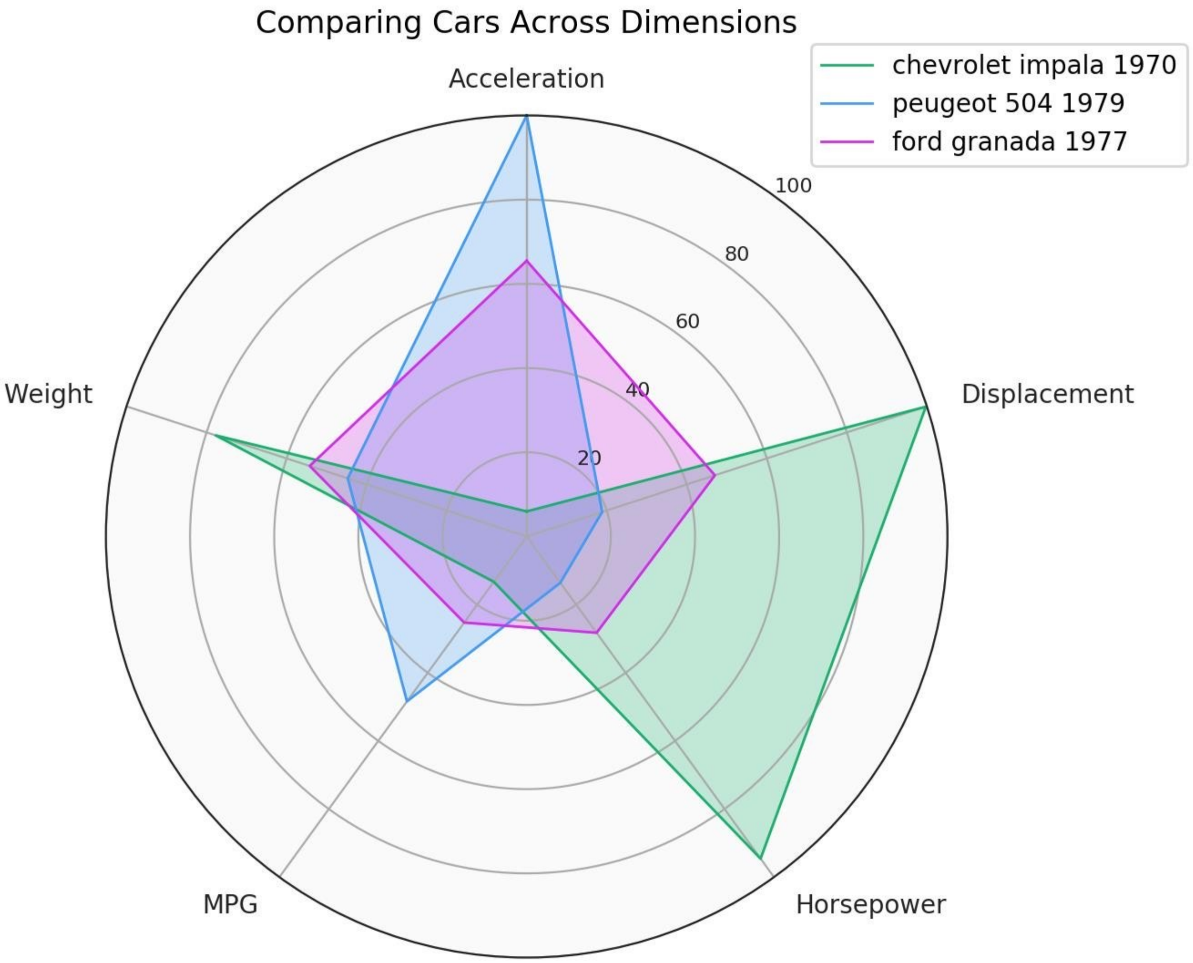


visualization

Multivariate data

Radar Chart (Spider Plot)

Same vectorization as with parallel coordinates, but instead of columns we organize the lines into a circle or polygon. Good for showing trade-offs.



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 location statistical 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

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
Standard deviation	7.45	17.38	11.25	5.66

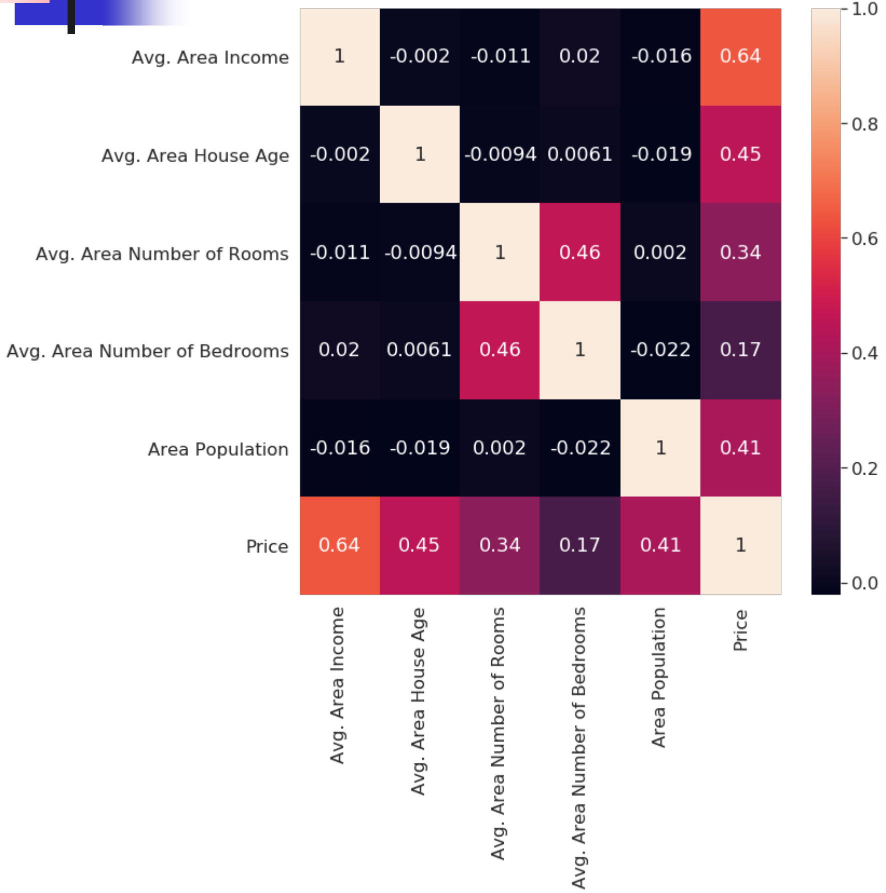
Dispersion multivariate statistics

- 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

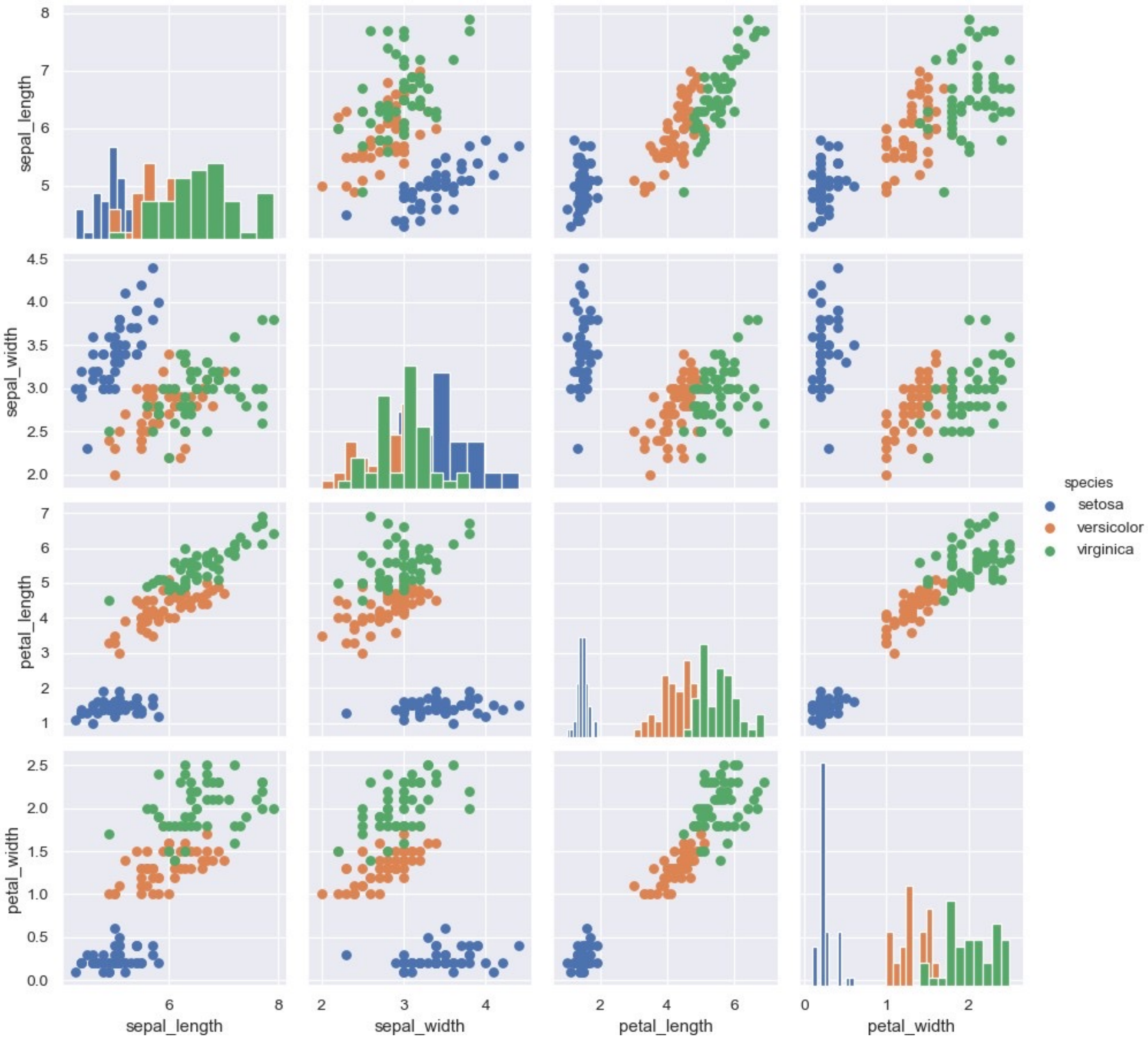


	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1	-0.002	-0.011	0.02	-0.016	0.64
Avg. Area House Age	-0.002	1	-0.0094	0.0061	-0.019	0.45
Avg. Area Number of Rooms	-0.011	-0.0094	1	0.46	0.002	0.34
Avg. Area Number of Bedrooms	0.02	0.0061	0.46	1	-0.022	0.17
Area Population	-0.016	-0.019	0.002	-0.022	1	0.41
Price	0.64	0.45	0.34	0.17	0.41	1

- **Correlation Heatmap** is used to show the pairwise correlation between variables

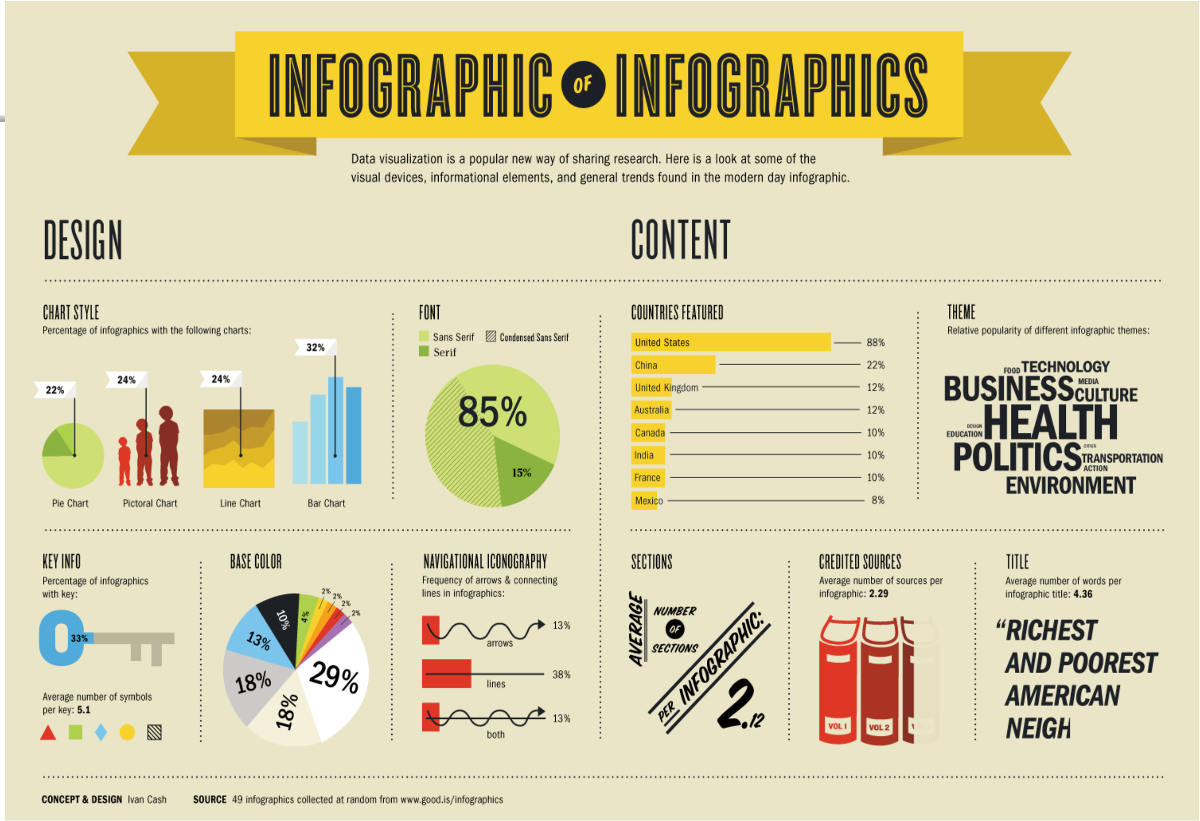
Dispersion multivariate statistics

- **Pair Plot** is used to visualize pairwise relationships between variables
  - The diagonal shows the distribution of the variables
  - As with other plots, we can increase dimensionality with colours and shapes



Dispersion multivariate statistics

- An **infographic** is a collection of imagery, charts, and minimal text that gives an easy-to-understand overview of a topic.
  - 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



https://www.wired.com/2012/07/you-suck-at-infographics/

Dispersion multivariate statistics

BIG
DATA
INFORMATION
PETABYTES
STORAGE
LARGE
SOFTWARE
SETS
MILLION
SCIENCE
FUTURE
DEVELOPMENT
TECHNOLOGY
BILLION
NEW
PARALLEL
VOLUME
PROCESS
GOVERNMENT
TECHNOLOGIES
MANAGEMENT
DATABASES
PROCESSING
SIZE
EXABYTES
ANALYSIS
RESEARCH
LEARNING
WORLD
BASED
SEARCH
INTERNET
SAN
NAS
SOCIAL
YEARS
RESULTS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
ANNOUNCED
PAST
HAND
CAPACITY
NETWORKS
INCLUDE
HYPOTHESIS
END
MOST
NEARLY
MAPREDUCE
CRITIQUES
STATE
CRITIQUE
PEOPLE
TRAFFIC
USE
CURRENTLY
TERABYTES
RELATED
MARKET
TIME
SENSOR
WOULD
NOW
USED
ANNUAL
VARIETY
RELEVANT
DISTRIBUTED
PROJECT
ECONOMIC
INITIATIVE
APPROACH
MANY
OUT
ONE
SECOND
NATIONAL
FLOW
INTELLIGENCE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
FACTOR
SYSTEMS
MANAGE
ONLINE
DAY
WORLDWIDE
MAKING
INTEGRATION
MASSIVE
FRAMEWORK
BUSINESS
BETWEEN
EFFECTIVE
DEFINITION
USING
STATISTICS
VISUALIZATION
ANALYTICS
SIMULATIONS
DETERMINE
WELL
SET
DEPARTMENT
QUERIES
HIGH
USERS
AMOUNT
RATE
REAL
STRUCTURE
WORK
APPLICATIONS
INFRASTRUCTURE
HIGHER
ALGORITHMS
COMPANIES
TERMS
YEAR
PRIVATE
LESS
HUGE
CAPABILITIES
SUCCESSFUL
NEEDED
USES
COMPARED
EVERY
VARIOUS
TIMES
SHARED
APPROACHES
PROCESSES
ARCHITECTURE
PARADIGM
FUNDING
INSIGHT
TOOLS
CENTER
RECORDS
COST
PER

- 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
  - 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. For example: a, the, is
  - 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. For example:connection, connected, connections, connects -> connect

Thank You!