Data Mining

Data Understanding

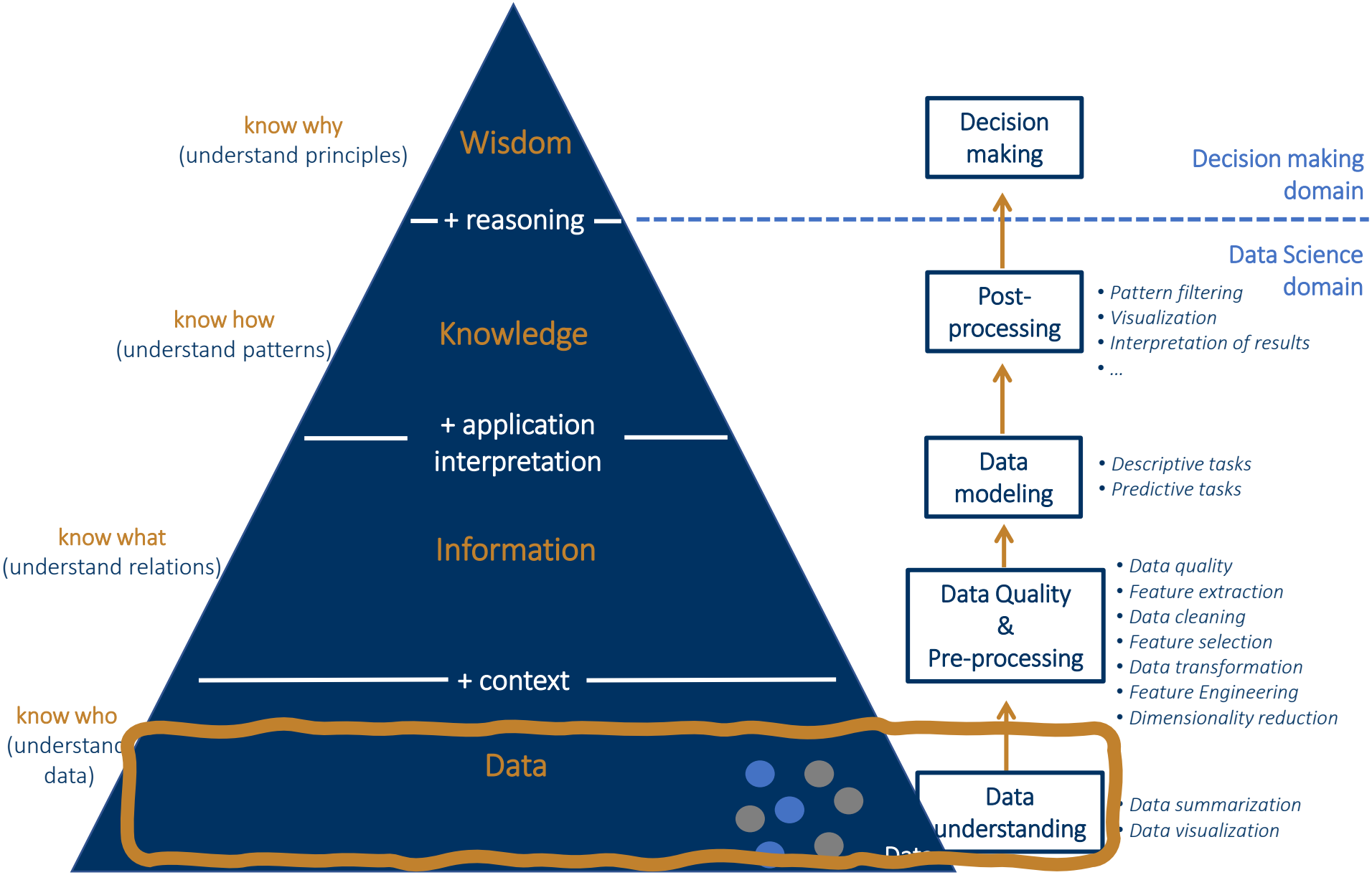
Raquel Sebastião

Departamento de Eletrónica, Telecomunicações e Informática

Universidade de Aveiro

raquel.sebastiao@ua.pt

2022/2023


know why
(understand principles)
Wisdom
+ reasoning
know how
(understand patterns)
Knowledge
+ application
interpretation
know what
(understand relations)
Information
+ context
know who
(understand
data)
Data
Decision
making
Decision making
domain
Data Science
domain
Post-
processing
• Pattern filtering
• Visualization
• Interpretation of results
• ...
Data
modeling
• Descriptive tasks
• Predictive tasks
Data Quality
&
Pre-processing
• Data quality
• Feature extraction
• Data cleaning
• Feature selection
• Data transformation
• Feature Engineering
• Dimensionality reduction
Data
understanding
• Data summarization
• Data visualization

Contents

- Attributes and Datasets
- Data Summarization
- Data Visualization
- Proximity measures
- Summary

Why get insights from our data?

What is data?

Collection of objects/examples described by attributes (taken from a domain)

- Object – entity
  - described by attributes
  - samples, examples, instances, data points, observations
- Attribute – data field
  - represent a characteristic/feature of a data object
  - dimension, feature, variable, characteristics

Attribute values are **numbers** or **symbols** assigned to an attribute for a particular object

Attribute values

Attribute values are **numbers** or **symbols** assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
  - But properties of attribute can be different than the properties of the values used to represent the attribute

Attributes

Type and scales of attributes

- Categorical
    - Nominal
    - Ordinal
- Numeric
    - Interval
    - Ratio

Scale of attributes

Categorical attributes

- finite number of symbols or names
- Sometimes, represented as integers (but they don't represent quantities)

Nominal

- there is no relationship between the values
- only equality is meaningful

Ordinal

- there is a meaningful order (ranking), but magnitude between successive symbols/values are unknown
- both equality and inequality is meaningful

Scale of attributes

Numeric attributes

- real-valued or integer-valued domain

Interval

- values vary within an interval
- equality, inequality and differences are meaningful
- the value 0 or scale origin, is defined arbitrarily
- no true **zero-point**

Ratio

- numbers have an absolute meaning
- equality, inequality, differences and ratios are meaningful
- inherent **zero-point**

Types and scales of attributes

- Categorical: set-valued domain composed of a set of symbols (qualitative)
  - **Nominal**: only equality (are two values the same?) is meaningful
    - domain(HairColor) = {brown, blond, black}, domain(Sex) = {M, F}
  - **Ordinal**: both equality and inequality (is one value less than another?) are meaningful
    - domain(Education) = {High School, BS, MS, PhD}, domain(size) = {S, L}
- Numeric: real-valued or integer-valued domain (quantitative)
  - **Interval-scale**: only differences are meaningful
    - temperature (Celsius or Fahrenheit), calendar years
  - **Ratio-scale**: differences and ratios are meaningful
    - Age, temperature (Kelvin), distance, income, counts, elapsed time

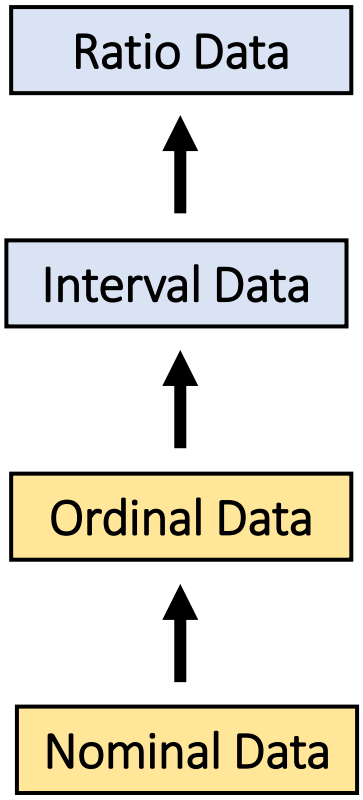
Types and scales of attributes

Differences between measurements, true zero exists

Differences between measurements but no true zero

Ordered Categories (rankings, order, or scaling)

Categories (no ordering or direction)



Quantitative Data

Qualitative Data

Types and scales of attributes

Amount of Information ↓

Attributes		distinctness	order	meaningful differences	operations
Type	Scale	(= , ≠)	(< , >)	(+, -)	(* , /)
Categorical	Nominal	✓			
	Ordinal	✓	✓		
Numeric	Interval	✓	✓	✓	
	Ratio	✓	✓	✓	✓

Types and scales of attributes

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ 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
Numeric Quantitative	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

This categorization of attributes is due to S. S. Stevens

Types and scales of attributes

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	*new_value = a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

This categorization of attributes is due to S. S. Stevens

Discrete vs. continuous attributes

- **Discrete Attribute**
  - finite or countably infinite set of values
  - it can take only distinct or separate values
    - zip codes, counts, or the set of words
      (binary attributes - special case of discrete attributes)

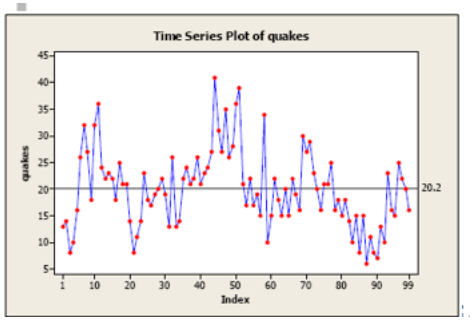
- **Continuous Attribute**
  - infinite set of values
  - real numbers
    - temperature, height, weight, distance

Special case: Binary attributes

- **Binary Attribute**
  - Nominal attribute with only 2 states (usually: 0 and 1)
  - Symmetric binary: both outcomes equally important
    - gender, sizes
  - Asymmetric binary: outcomes not equally important
    - Only presence (a non-zero attribute value) is regarded as important
    - **Convention: assign 1 to most important outcome (e.g., HIV positive)**
      - medical test (positive vs. negative), failure/not failure, words present in documents, items present in customer transactions

Types of datasets

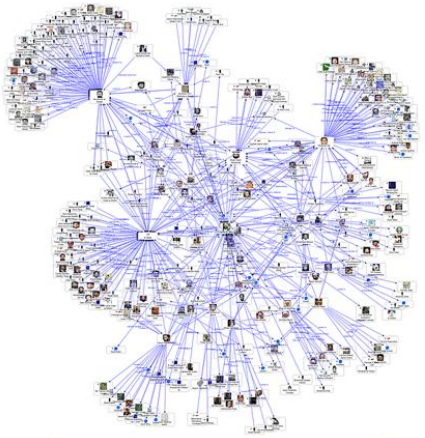
- Data tables
  - **Data matrix**, document data, transactional data
- Graphs and networks
  - Molecular structures, transportation networks, social networks
- Ordered data
  - temporal data (time-series), data streams genetic sequences
- Multimedia data
  - Maps, images, videos, audios
- …



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

Types of datasets

- Data tables
  - **Data matrix**, document data, transactional data
- Graphs and networks
  - Molecular structures, transportation networks, social networks
- Ordered data
  - temporal data (time-series), data streams genetic sequences
- Multimedia data
  - Maps, images, videos, audios
- …



A social network

Types of datasets

- Data tables
  - Data matrix, document data, transactional data
- Graphs and networks
  - Molecular structures, transportation networks, social networks
- Ordered data
  - temporal data (time-series), data streams genetic sequences
- Multimedia data
  - Maps, images, videos, audios
- …

Data matrix:

Attributes

Observations

Id	Age	Height	Sex	Marital Status	Ec. Status
1	56	182	masc	single	low inc.
2	43	176	masc	married	high inc.
3	53	164	fem	married	high inc.
4	47	171	fem	widowed	high inc.
5	52	170	fem	widowed	low inc.
6	49	169	fem	married	middle inc.
7	51	173	fem	married	middle inc.
8	48	173	fem	single	middle inc.
9	47	179	masc	married	low inc.
10	39	183	masc	single	low inc.
11	58	165	fem	married	low inc.
12	57	184	masc	married	middle inc.

Document data:

	course	department	university	class	student
Doc 1	3	6	4	0	8
Doc 2	0	5	3	2	3
Doc 3	0	1	0	1	0

Important characteristics of datasets

- Dimensionality
  - high dimensional data brings several challenges
    - Curse of dimensionality
- Size
  - type of analysis may depend on size of data
- Sparsity
  - only presence counts
- Resolution
  - patterns depend on the scale
    - *motivation for data preparation: next classes…*

Data matrix

Data represented as a *matrix* with *N* rows and *D* columns

$$\begin{array}{c|cccc} & X_1 & X_2 & \dots & X_D \\ \hline \mathbf{x}_1^T & x_{11} & x_{12} & \dots & x_{1D} \\ \mathbf{x}_2^T & x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_N^T & x_{N1} & x_{iN2} & \dots & x_{ND} \end{array}$$

- **Rows**: Also called *instances*, *examples*, *records*, *transactions*, *objects*, *points*, *feature-vectors*, etc. Given as a *D*-tuple [1]

$$\mathbf{x}_i^T = [x_{i1} \quad x_{i2} \quad \dots \quad x_{iD}]^T = (x_{i1} \quad x_{i2} \quad \dots \quad x_{iD})$$

- **Columns**: Also called *attributes*, *properties*, *features*, *dimensions*, *variables*, *fields*, etc. Given as an *N*-tuple

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{Nj} \end{bmatrix} = [x_{1j} \quad x_{2j} \quad \dots \quad x_{Nj}]$$

[1] the class label usually is excluded of this description ($\mathbf{x}_i$,label)

Data matrix

Data represented as a *matrix* with *N* rows and *D* columns

$$\begin{array}{c|cccc} & X_1 & X_2 & \dots & X_D \\ \hline \mathrm{x}_1^T & x_{11} & x_{12} & \dots & x_{1D} \\ \mathrm{x}_2^T & x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{x}_N^T & x_{N1} & x_{iN2} & \dots & x_{ND} \end{array}$$

Feature Vector **X**

- the *i – th* feature is X_i

Feature Space: dimension is equal to the number of features.

- with *i – th* coordinate related to *i – th* feature

Data: algebraic and geometric view

For numeric data matrix X:

- each row or point is a *D*-dimensional [2] row vector

$$\mathrm{x}_i^T = [x_{i1} \quad x_{i2} \quad \dots \quad x_{iD}]^T = (x_{i1} \; x_{i2} \; \dots \; x_{iD}) \in \mathbb{R}^D$$

- each column or attribute is a *N*-dimensional column vector

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{Nj} \end{bmatrix} = [x_{1j} \; x_{2j} \; \dots \; x_{Nj}] \in \mathbb{R}^N$$

[2] Classification problems: the class label is not included

Feature vector and feature space

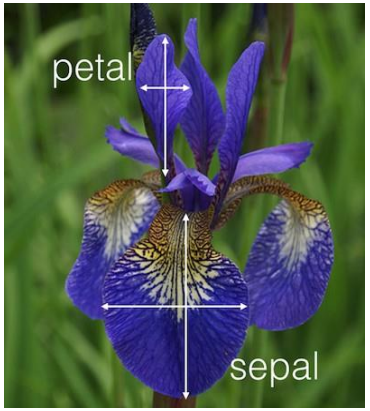
Feature vector

$$\mathrm{x} = [\mathrm{x}_1 \quad \mathrm{x}_2 \quad \dots \quad \mathrm{x}_D]^T = (\mathrm{x}_1 \quad \mathrm{x}_2 \quad \dots \quad \mathrm{x}_D)$$

- Each feature represents one dimension, and its values represent positions along one of the orthogonal coordinate axes in feature space

Feature space: set of all possible values for a chosen set of features from that data.

Example: Iris Dataset



	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
x_1	5.9	3.0	4.2	1.5	Iris-versicolor
x_2	6.9	3.1	4.9	1.5	Iris-versicolor
x_3	6.6	2.9	4.6	1.3	Iris-versicolor
x_4	4.6	3.2	1.4	0.2	Iris-setosa
x_5	6.0	2.2	4.0	1.0	Iris-versicolor
x_6	4.7	3.2	1.3	0.2	Iris-setosa
x_7	6.5	3.0	5.8	2.2	Iris-virginica
x_8	5.8	2.7	5.1	1.9	Iris-virginica
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
x_{149}	7.7	3.8	6.7	2.2	Iris-virginica
x_{150}	5.1	3.4	1.5	0.2	Iris-setosa

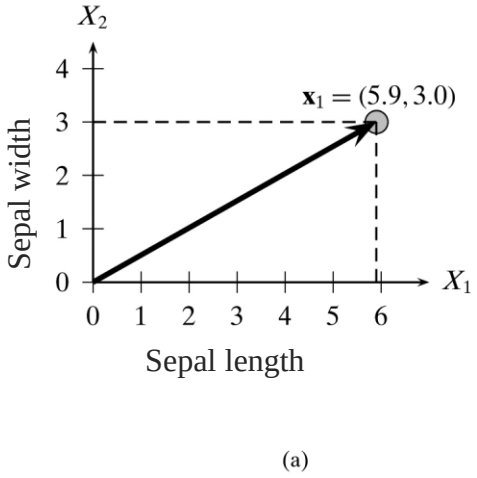
Example: Iris Dataset

Feature vector

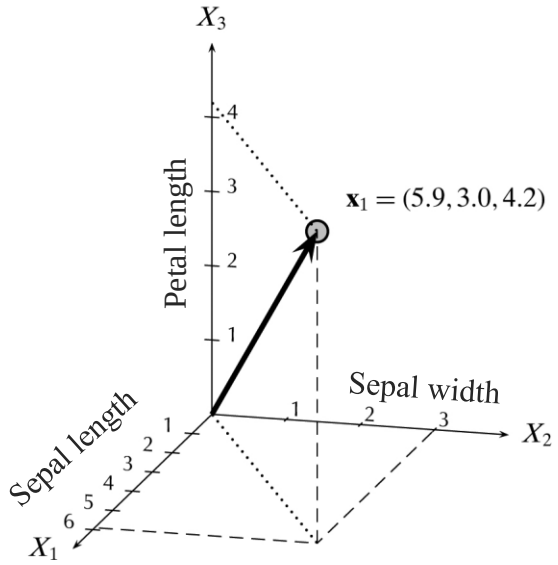
$$\mathbf{x} = [x_1 \quad x_2 \quad \dots \quad x_D]^T = (x_1 \quad x_2 \quad \dots \quad x_D)$$

Feature space: set of all possible values for a chosen set of features from that data

- **Visualization** : x_1 = (5.9; 3.0; 4.2; 1.5) in 2D and 3D



(a)



(b)

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
x_1	5.9	3.0	4.2	1.5	Iris-versicolor
x_2	6.9	3.1	4.9	1.5	Iris-versicolor
x_3	6.6	2.9	4.6	1.3	Iris-versicolor
x_4	4.6	3.2	1.4	0.2	Iris-setosa
x_5	6.0	2.2	4.0	1.0	Iris-versicolor
x_6	4.7	3.2	1.3	0.2	Iris-setosa
x_7	6.5	3.0	5.8	2.2	Iris-virginica
x_8	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
x_{149}	7.7	3.8	6.7	2.2	Iris-virginica
x_{150}	5.1	3.4	1.5	0.2	Iris-setosa

Homework…

- Assignment I (see eLearning)
  - Data Understanding:
    - Hands on: Datasets and attributes

Contents

- Attributes and Datasets
- Data Summarization
- Data Visualization
- Proximity measures
- Summary

Data summarization

Common questions in data analysis

- What is the most common value?
- What is the variance in the values?
- Are there strange values?

Choosing the appropriate data analysis depends on

- number of variables/attributes: univariate or multivariate
- type of variables/attributes: categorical or numeric

Data summarization

Motivation

- Big datasets: it's hard to understand data
- Data summaries: useful synopsis of key properties of data
- To better understand the data:
  - Frequency
  - Central tendency
  - Dispersion / Spread

Describe important properties of data distribution

Data summarization

Summary statistics for data exploration

better understand the data

- Frequency
- Central tendency
- Dispersion / Spread

Data summarization: univariate data

Frequency*

- Absolute (or relative) occurrence of each value
  - e.g. nr. of days by outlook (year)

sunny	overcast	rainy
198	65	102
54.2%	17.8%	27.9%

  - e.g. exam grades

8	10	13	14	17	19
1	2	9	8	3	1
4.2%	8.3%	37.5%	33.3%	12.5%	4.2%

*For both categorical and numeric variables/attributes (typically used for categorical data)

Data summarization

Central tendency

- Mean: the average value (sensitive to extremes)
- Quartiles
  - **Median**: the "middle" value (from a sorted list of values)
- Mode*: the most frequent value

*For both categorical and numeric variables/attributes

Data summarization: univariate data

Central tendency

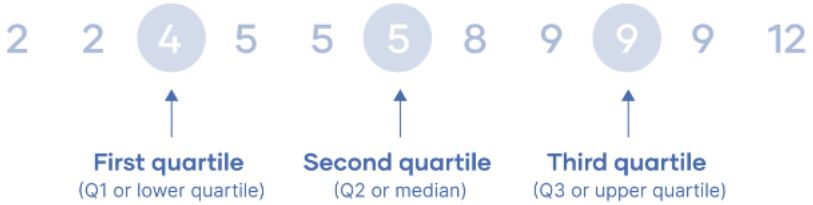
- Mean: the average value (sensitive to extremes)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Data summarization: univariate data

Central tendency: **Quartiles**

- Q_1 (1^{st} quartile or the lower quartile)
  - value halfway between the lowest value and the middle value
- Q_2 (2^{nd} quartile or the **median**)
  - value halfway between the lowest value and the highest value
- Q_3 (3^{rd} quartile or the upper quartile)
  - value halfway between the middle value and the highest value



Data summarization: univariate data

Central tendency

- Median: the "middle" value (from a sorted list of values)
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Approximate median

Sum before the median interval

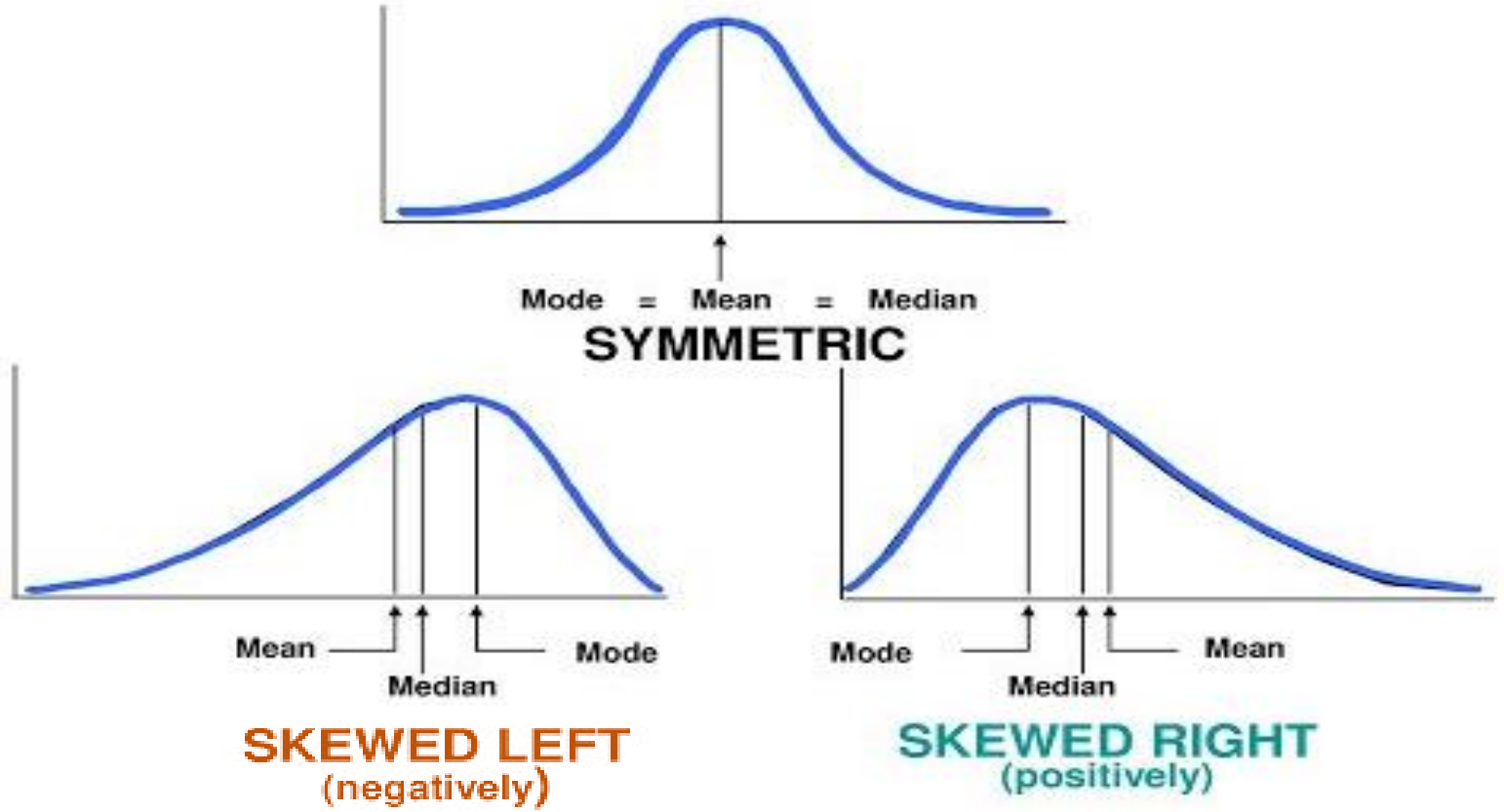
Interval width ($L_2 - L_1$)

$$median = L_1 + (\frac{n/2 - (\sum freq)_l}{freq_{median}})width$$

Low interval limit

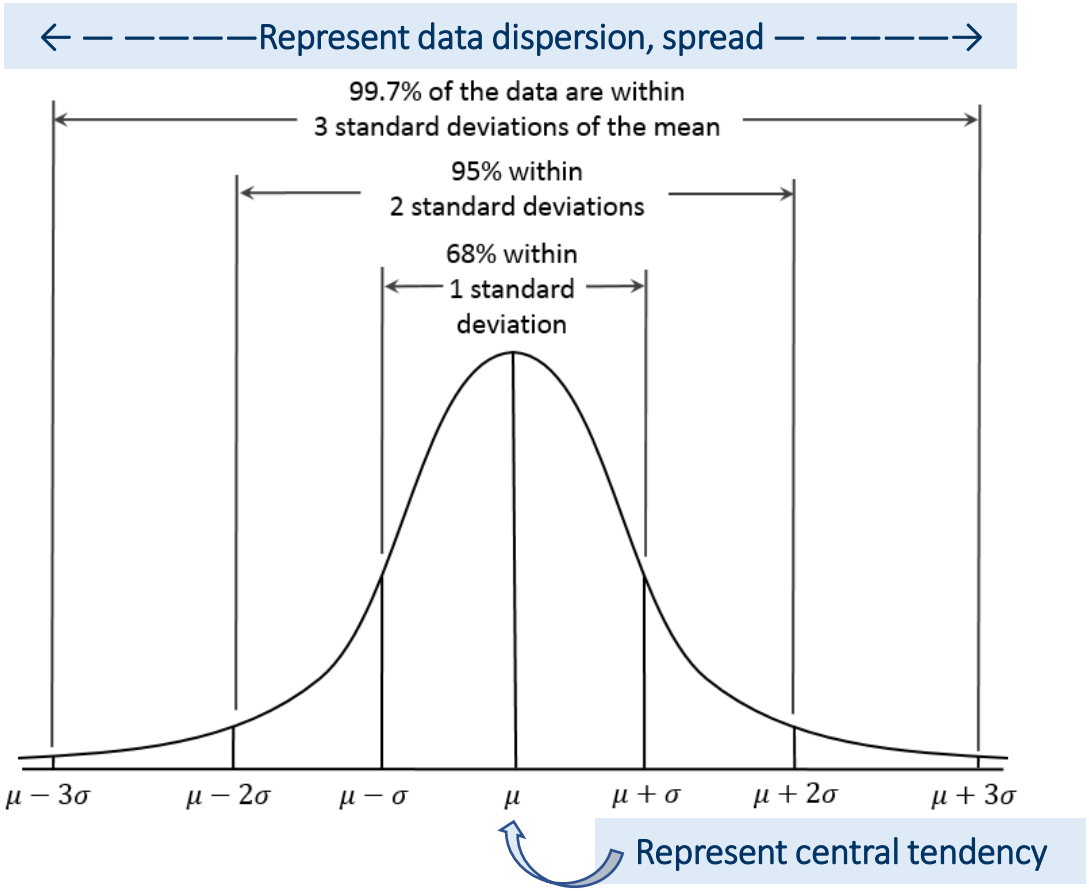
Data summarization: univariate data

Symmetric vs. Skewed data



Data summarization: univariate data

Properties of normal distribution curve



Data summarization: univariate data

Dispersion (spread)

- Inter-quartile range: IQR = Q_3 - Q_1
- Range: max_x – min_x
- Median/Average absolute deviation
- Standard deviation: sensitive to extreme values
- Variance: sensitive to extreme values

Data summarization: univariate data

Dispersion (spread)

- Median absolute deviation

$$\mathrm{MA}D = median(x_i - \bar{x})^2$$

- Average absolute deviation

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

Data summarization: univariate data

Dispersion (spread)

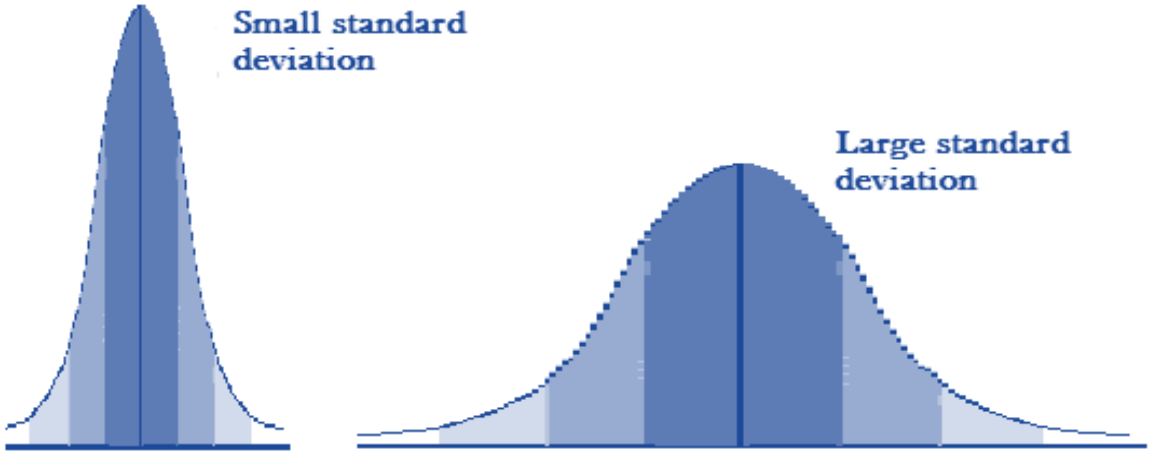
- Standard deviation

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

- Variance: s_X^2

Data summarization: univariate data

Dispersion (spread)



Data summarization: multivariate data

Frequency*

- Contingency tables: cross-frequency of values for two variables

  Season and outlook (year)

	winter	spring	summer	autumn
sunny	15	62	106	37
rainy	57	23	22	34

- Summarize the relationship between pairs of variables in a data set
- Gives the number of occurrences of $X = b_i$ and $Y = a_j$ in the data set.

*For both categorical and numeric variables/attributes (typically used for categorical data)

Data summarization: multivariate data

Dispersion (spread)

- **Covariance matrix**: variance between every pair of numeric variables
  - the value depends on the magnitude of the variable

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

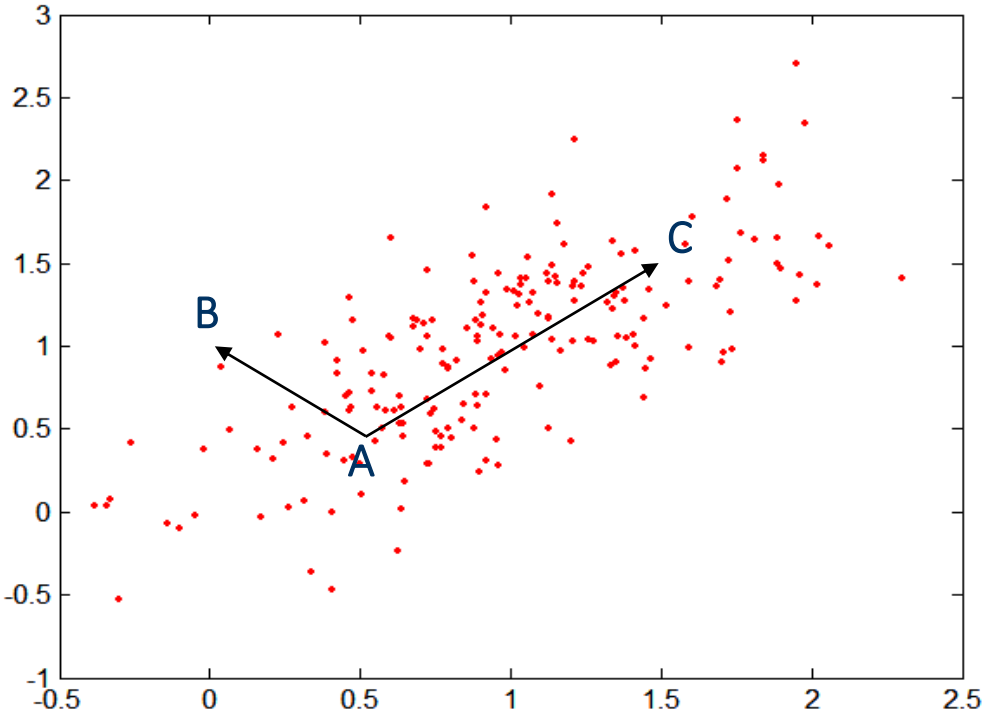
$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

- non-diagonal entries represent the covariance between pairs of variables
- diagonal is the variance of the variables
- matrix is symmetric

Data summarization: multivariate data

Dispersion (spread)

- Covariance matrix



$$\Sigma = \begin{bmatrix} 0.58 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Data summarization: multivariate data

Association (relationship)

- **Correlation matrix**: correlation between every pair of **numeric** variables (interval or scale)
  - the influence of the magnitude is removed

$$corr(x, y) = \frac{cov(x,y)}{std(x)std(y)}$$

$$-1 \leq corr(x, y) \leq 1$$

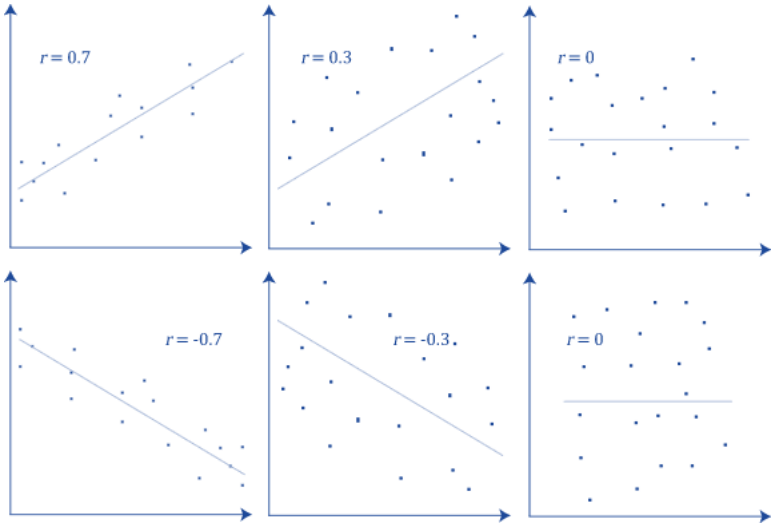
Data summarization: multivariate data

- Pearson's correlation coefficient (ρ_{xy})
  - Measures the linear correlation between a pair of **numeric** variables (interval or scale)

$$-1 \leq \rho_{xy} \leq 1$$

    - measures the **strength** and **direction** of the linear relationship between a pair of variables

$$\rho_{xy} = corr(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$



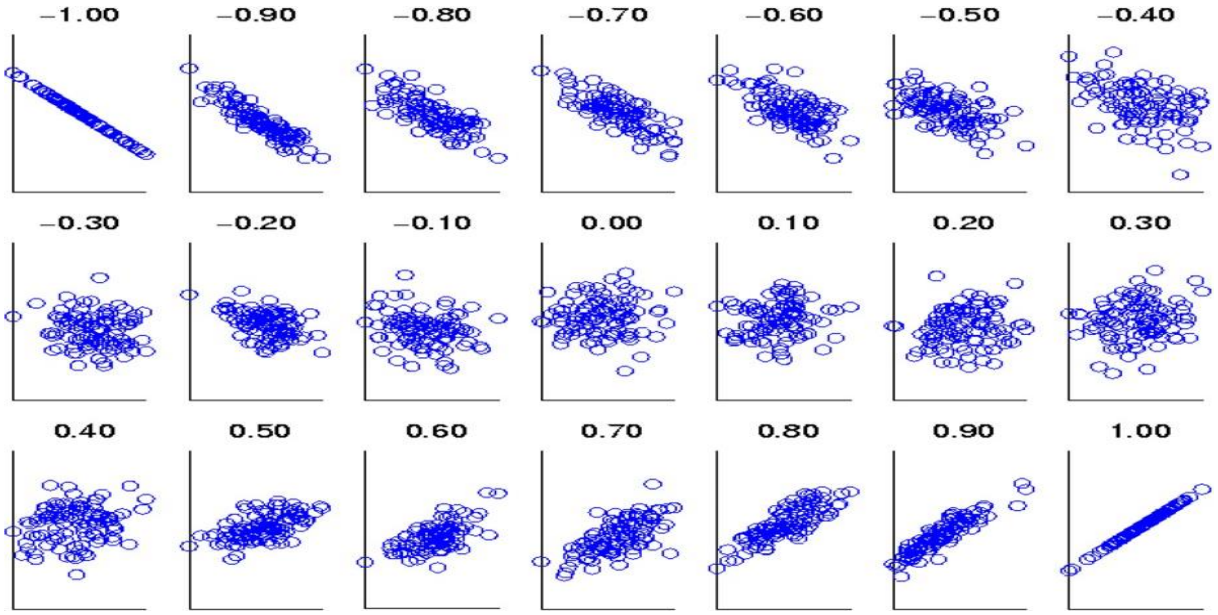
Data summarization: multivariate data

- Pearson's correlation coefficient (ρ_{xy})

  - Measures the linear correlation between a pair of **numeric** variables (interval or scale)

    $$-1 \leq \rho_{xy} \leq 1$$

    Scatter plots of pair of variables / features / attributes



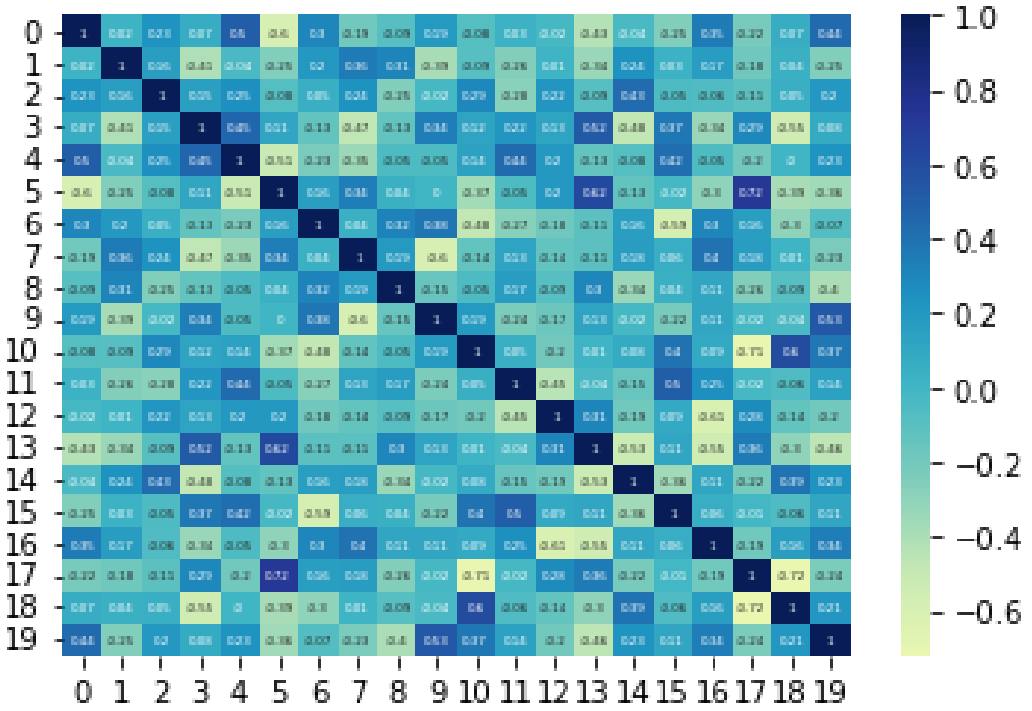
Data summarization: multivariate data

- Pearson's correlation coefficient (ρ_{xy})

  - Measures the linear correlation between a pair of **numeric** variables (interval or scale)

$$-1 \leq \rho_{xy} \leq 1$$

Heatmaps



Data summarization: multivariate data

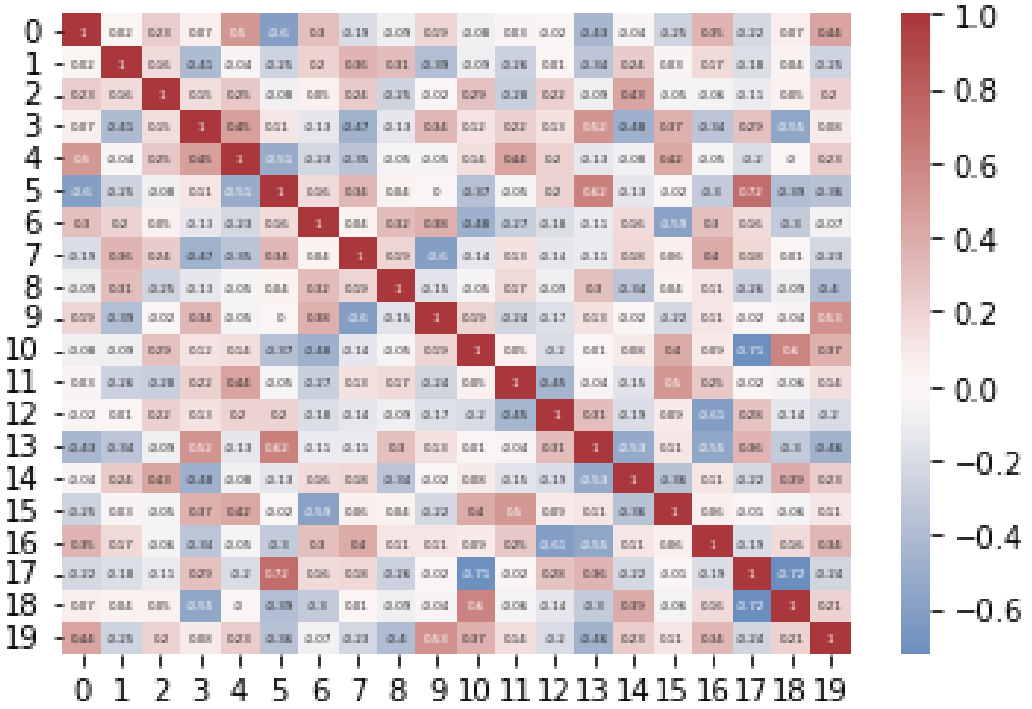
- Pearson's correlation coefficient (ρ_{xy})

  - Measures the linear correlation between a pair of **numeric** variables (interval or scale)

$$-1 \leq \rho_{xy} \leq 1$$

Heatmaps

Easier interpretation



Data summarization: multivariate data

- Spearman's rank-order correlation coefficient (rs_{xy})
  - Measures the monotonic association between a pair of variables

    $-1 \leq rs_{xy} \leq 1$

    - two variables can be related according to a type of non-linear but still monotonic relationship
    - measures the **strength** and **direction** of the monotonic relationship between a pair of variables

Data summarization: multivariate data

- Spearman's rank-order correlation coefficient (rs_{xy})
  - Measures the monotonic association between a pair of variables

    $-1 \leq rs_{xy} \leq 1$

    - rank-based and non-parametric version of Pearson correlation coefficient
    - Ordinal variables
    - Numerical variables (when the assumptions for Pearson coefficient are violated)

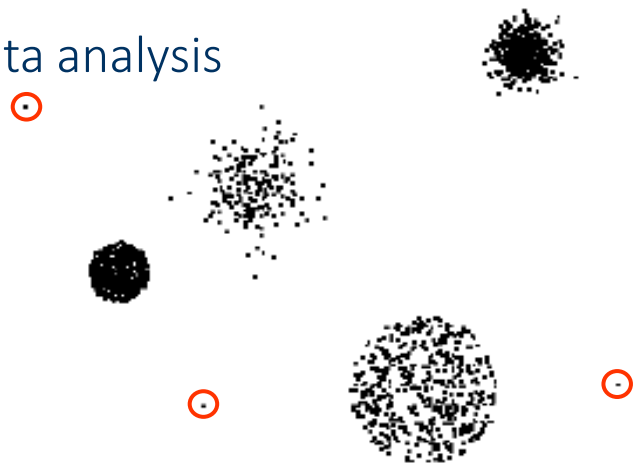
$$rs_{xy} = \rho_{rank_x rank_y}$$

Data summarization: outliers

"An outlier is a point that deviates so much from the other data points as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980)"
Hawkins, 1980

- Outliers can be univariate or multivariate
- 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



Data summarization: outliers

"An outlier is a point that deviates so much from the other data points as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980)"
Hawkins, 1980

- Statistical Parametric Techniques:
  - univariate case: boxplot definition (Tukey, 1977) is the most used:

$$any\ value\ outside\ the\ interval\ [Q_1 - 1.5\ IQR, Q_3 + 1.5\ IQR]$$

  - multivariate case: Mahalanobis distance (Mahalanobis, 1936).
- Statistical Non-parametric Techniques
  - Kernel functions
  - (…)

Homework…

- Assignment I (see eLearning)
  - Data Understanding:
    - Hands on: Summarization
  - Notes with exercises: Ex. 1.1 and 1.2

Contents

- Attributes and Datasets
- Data Summarization
- Data Visualization: Amounts, Distributions, Associations, Trends
- Proximity measures
- Summary

Data visualization: what?

Visualization is the conversion of data into a visual format so that the characteristics of the data and the relationships among data items or variables can be analyzed or reported.

Main types of visualization

- Amounts
- Distributions
- Associations
- Trends
- Geospatial data
- Uncertainty

Main types of visualization techniques

- Pie charts
- Bar plots
- Histograms
- Density plots
- QQ plots
- Boxplots
- Violin plots
- Scatter plots
- Heatmaps
- Parallel coordinates
- Correlograms
- (…)

Data visualization: why?

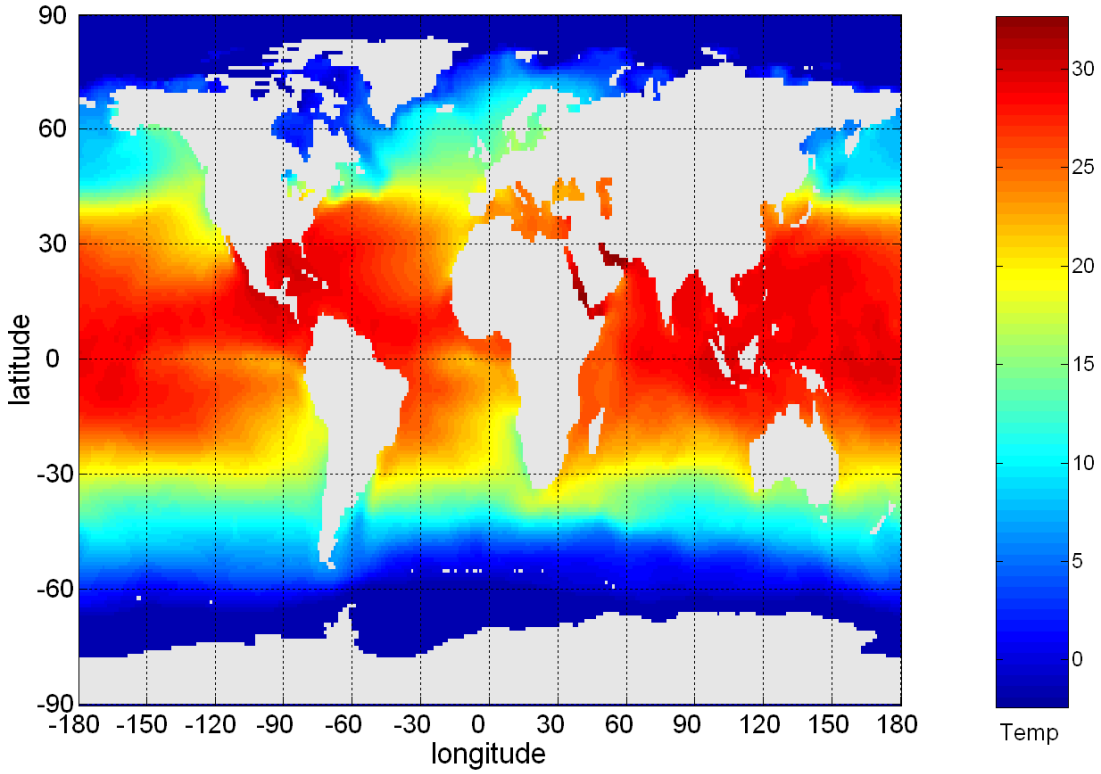
Visualization of data is one of the most powerful and appealing techniques for **data exploration**:

- Provide graphical display of basic description and qualitative overview of large data sets
- Humans have a well developed ability to analyze large amounts of information that is presented visually
- Gain insight into an information space
- Help detecting patterns, trends, structure, irregularities, relationships among data
- Help detecting unsual patterns, outliers, irregularities
- Help find interesting regions and suitable parameters for further quantitative analysis

Data visualization: example

The following map shows the Sea Surface Temperature (SST)

- Thousands of data points are summarized in a single figure

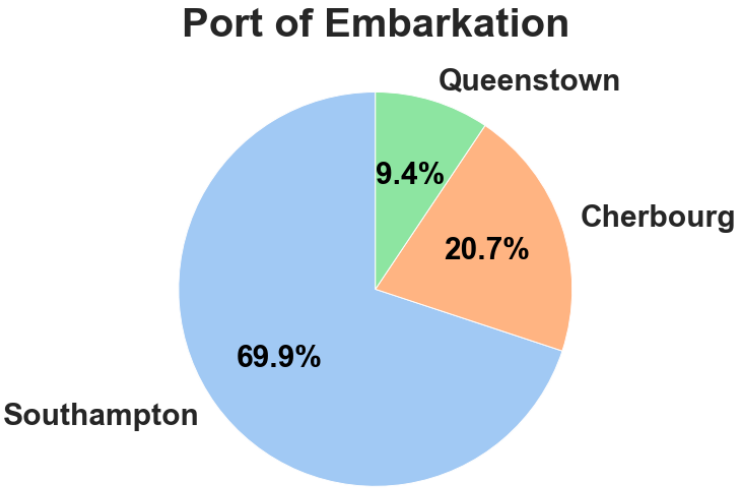


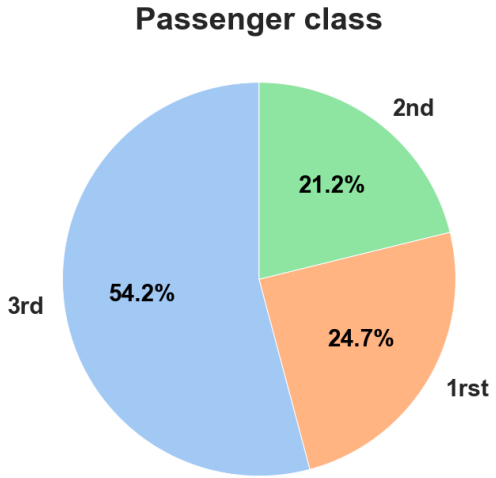
Data visualization: amounts

Pie charts

Typically used with categorical variables / attributes

- The slices correspond to categories:
  - each slice of the pie chart represents a value of the categorical variable
  - each slice of the pie chart display the relative frequency of that value





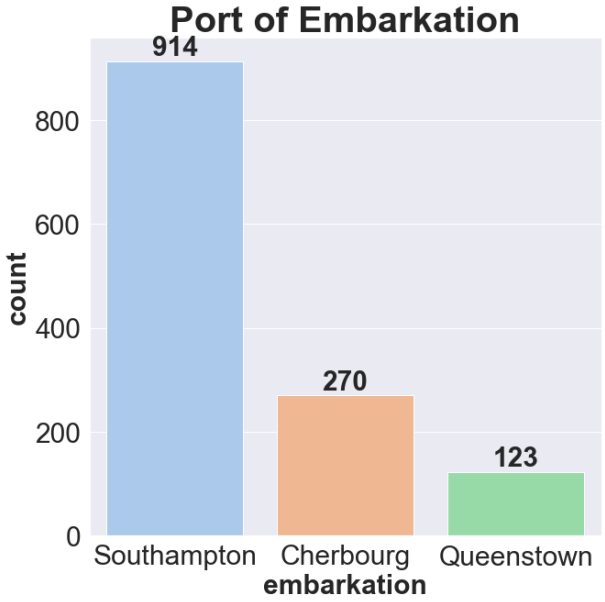
- Pie charts effectively illustrate the different sizes for parts of the whole
- Not a good option for **comparative** purposes.

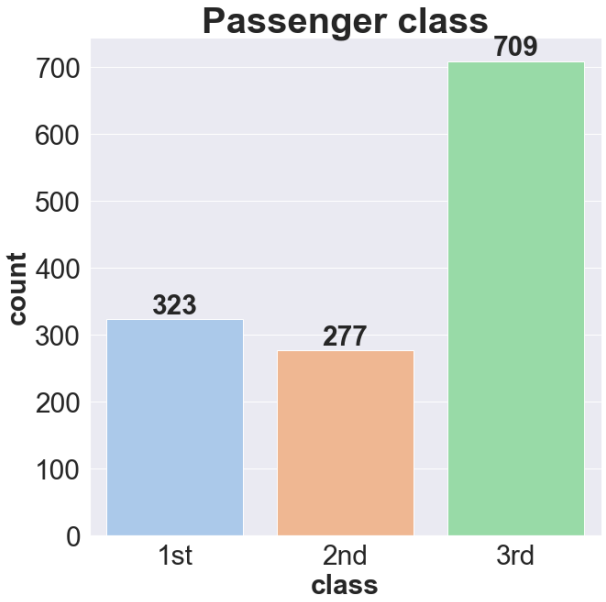
Data visualization: amounts

Bar plots

Typically used with categorical variables / attributes

- The bars correspond to categories:
  - each bar represents a value of the categorical variable
  - The height of each bar display the relative frequency of that value

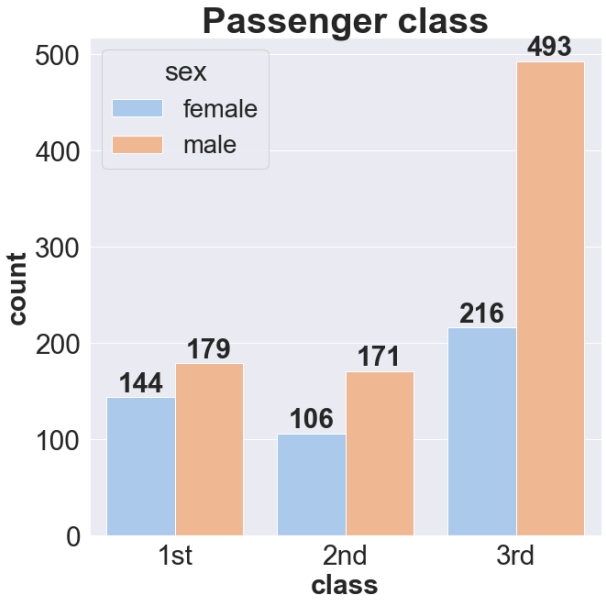


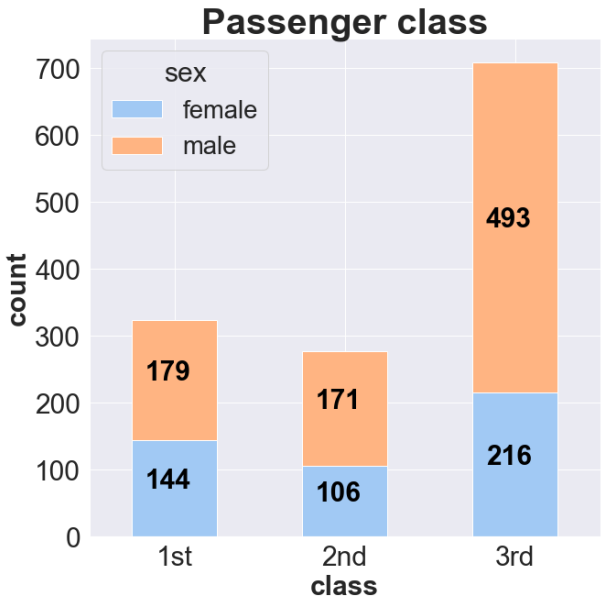


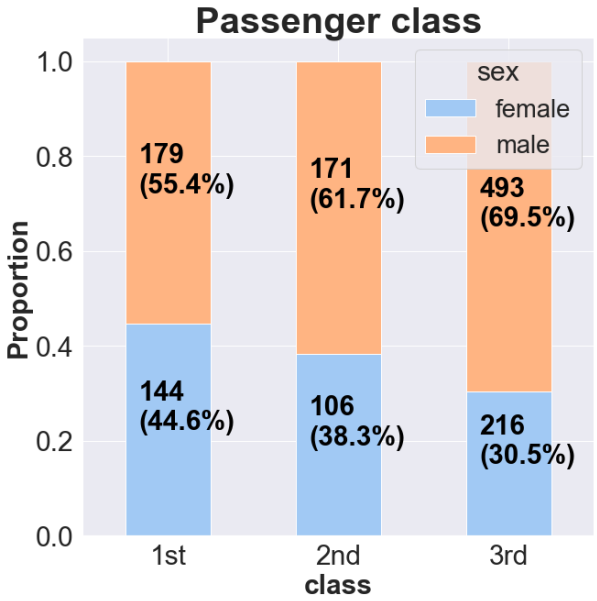
Data visualization: amounts

Bar plots with two variables

- Grouped
- Stacked
- Percent stacked



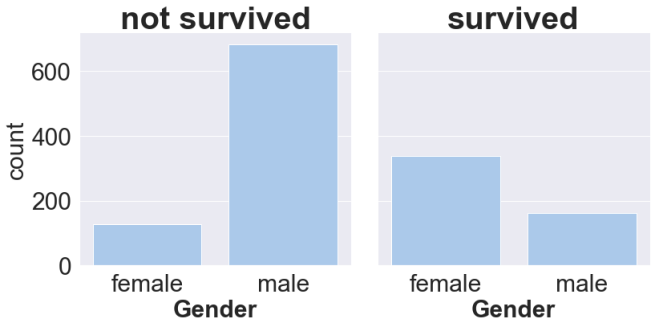


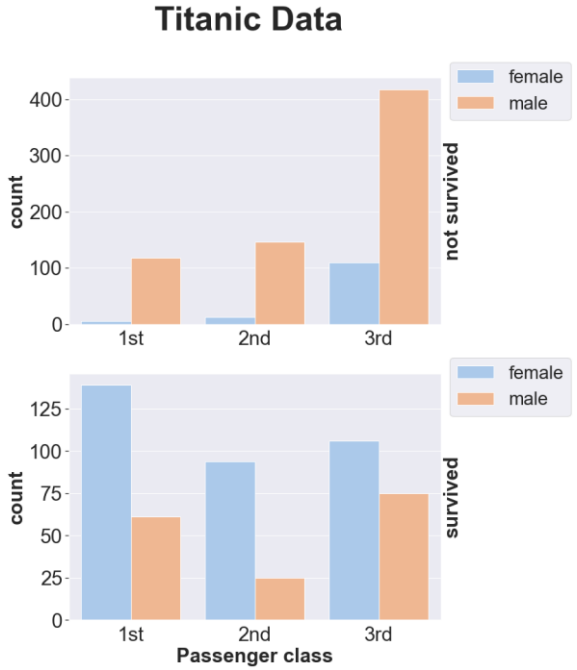


Data visualization: amounts

Bar plots w.r.t. other variables

- conditional plots
  - display the distribution of a variable (or the relationship between multiple variables) separately within sub-groups of the data set
  - allow finding eventual differences between the sub-groups
- can be drawn with up to three dimensions:
  - row, col
  - hue (third dimension)
    - different levels are plotted with different colors)

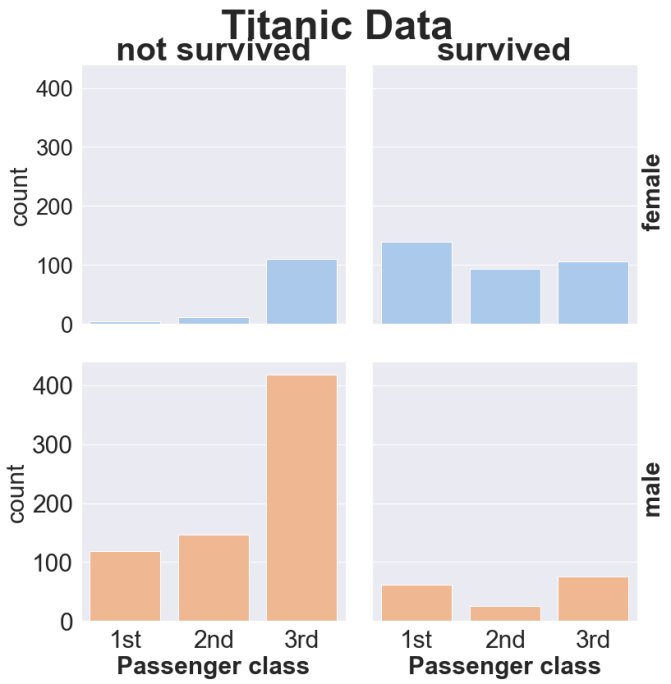


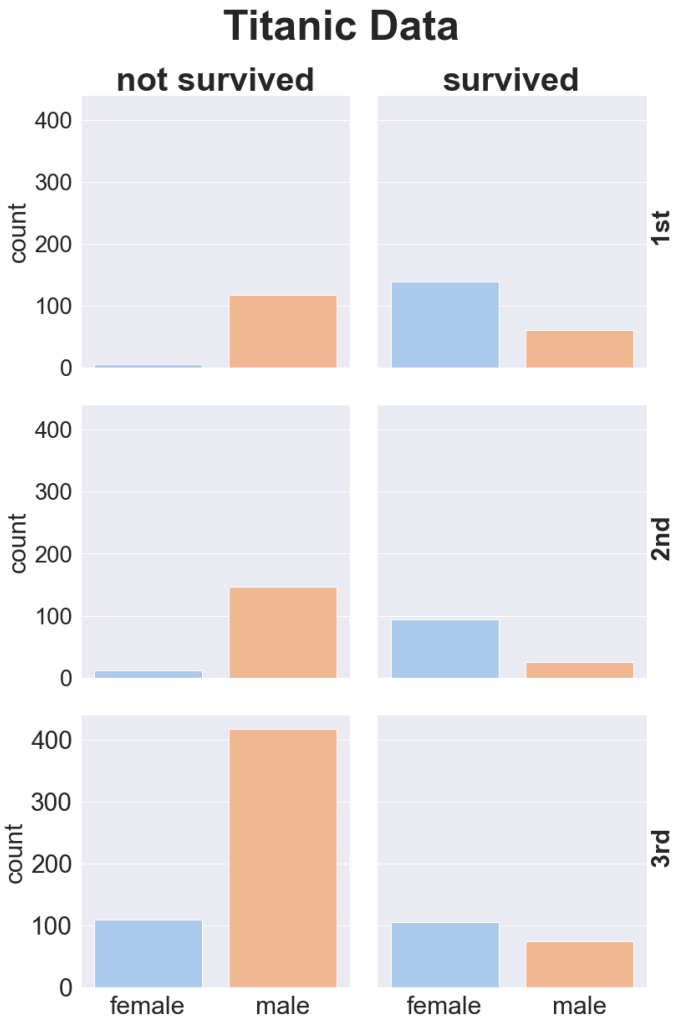


Data visualization: amounts

Bar plots w.r.t. other variables

- conditional plots





Data visualization: distributions

Histograms

Show how the values of **continuous** variables are distributed

- Usually shows the distribution of values of a single variable
- Give an idea of the shape of the data

Constructing a histograms

- The range of the variable is divided into a set of *bins* (intervals of values)
- The number of occurrences of values in each bin is counted
- A bar with this number is plotted
  - The height of each bar indicates the number of occurrences

- The shape of histogram depends on the number of bins

Data visualization: distributions

Histograms

- Relative (frequency) histograms: counts are replaced by the relative frequency
- Equal-width: all bins have the same width
- Equal-frequency: all bins have the same frequency

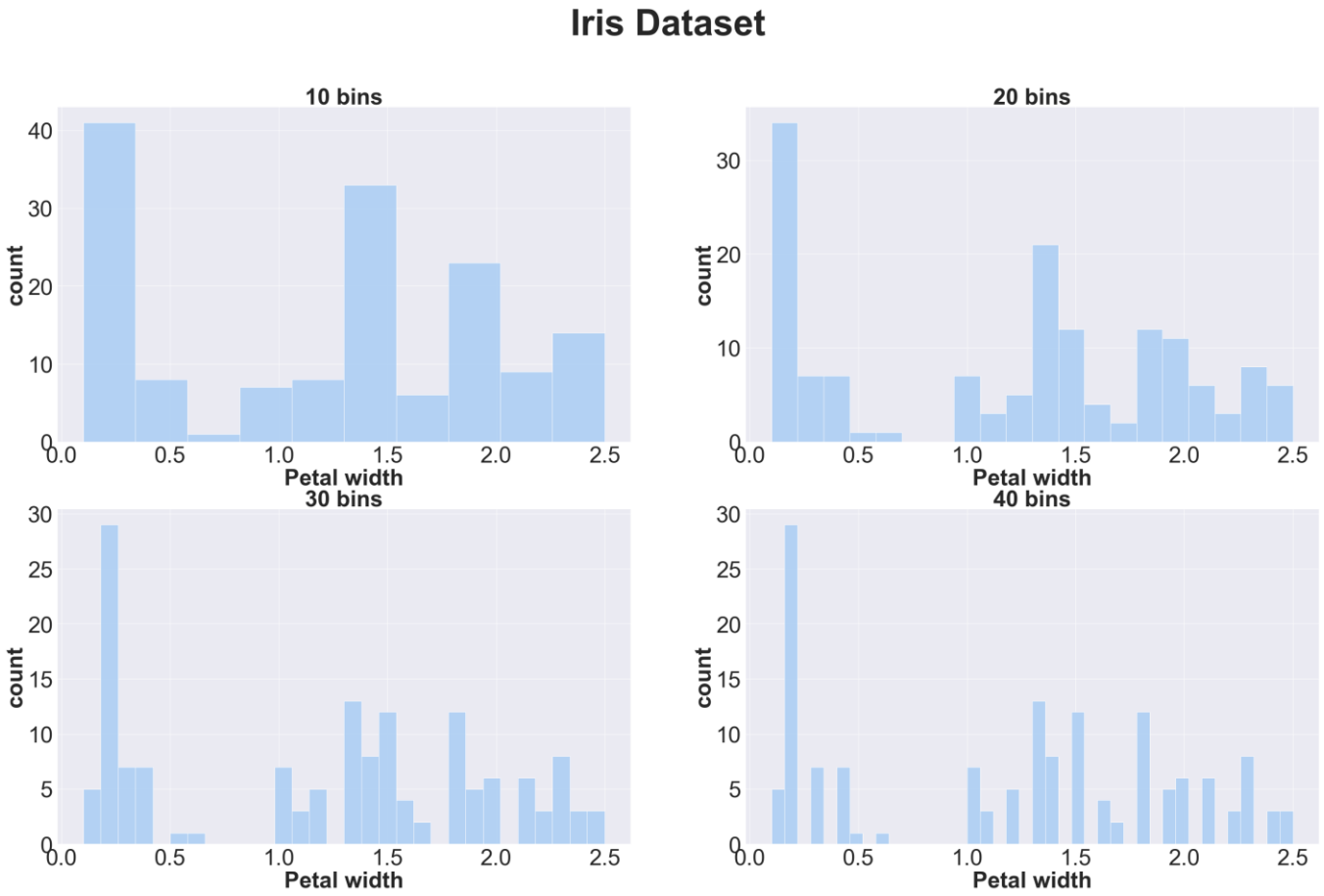
Disadvantages of histograms

- Requires to choose the number of bins (several algorithms for that)
- Shape depends on the number of bins
- Can be misleading
  - Small data sets
  - Intervals can hide individual values (difficult to detect relevant values)
- Hard to compare several distributions

Data visualization: distributions

Histograms

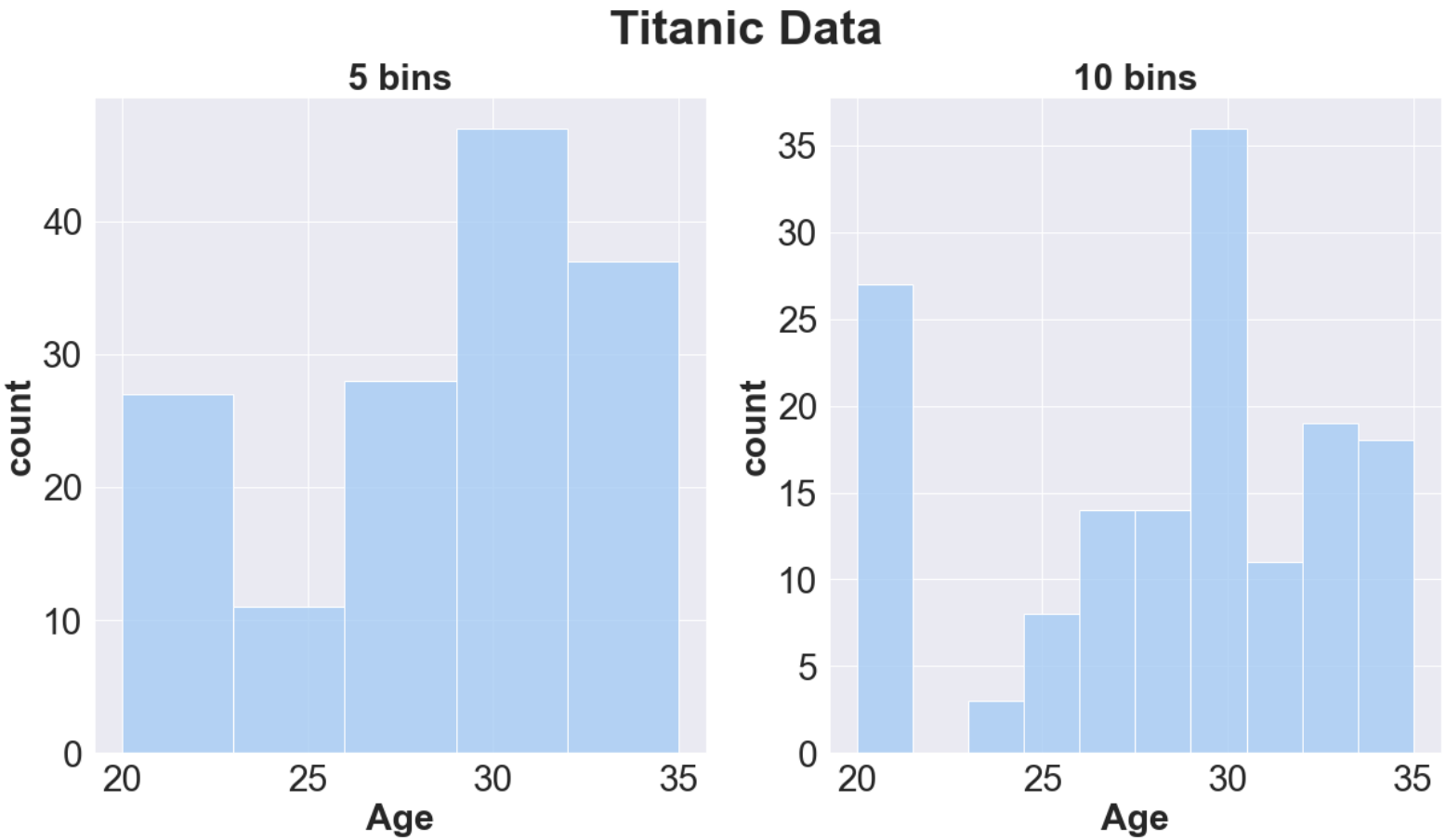
- Shape depends on the number of bins



Data visualization: distributions

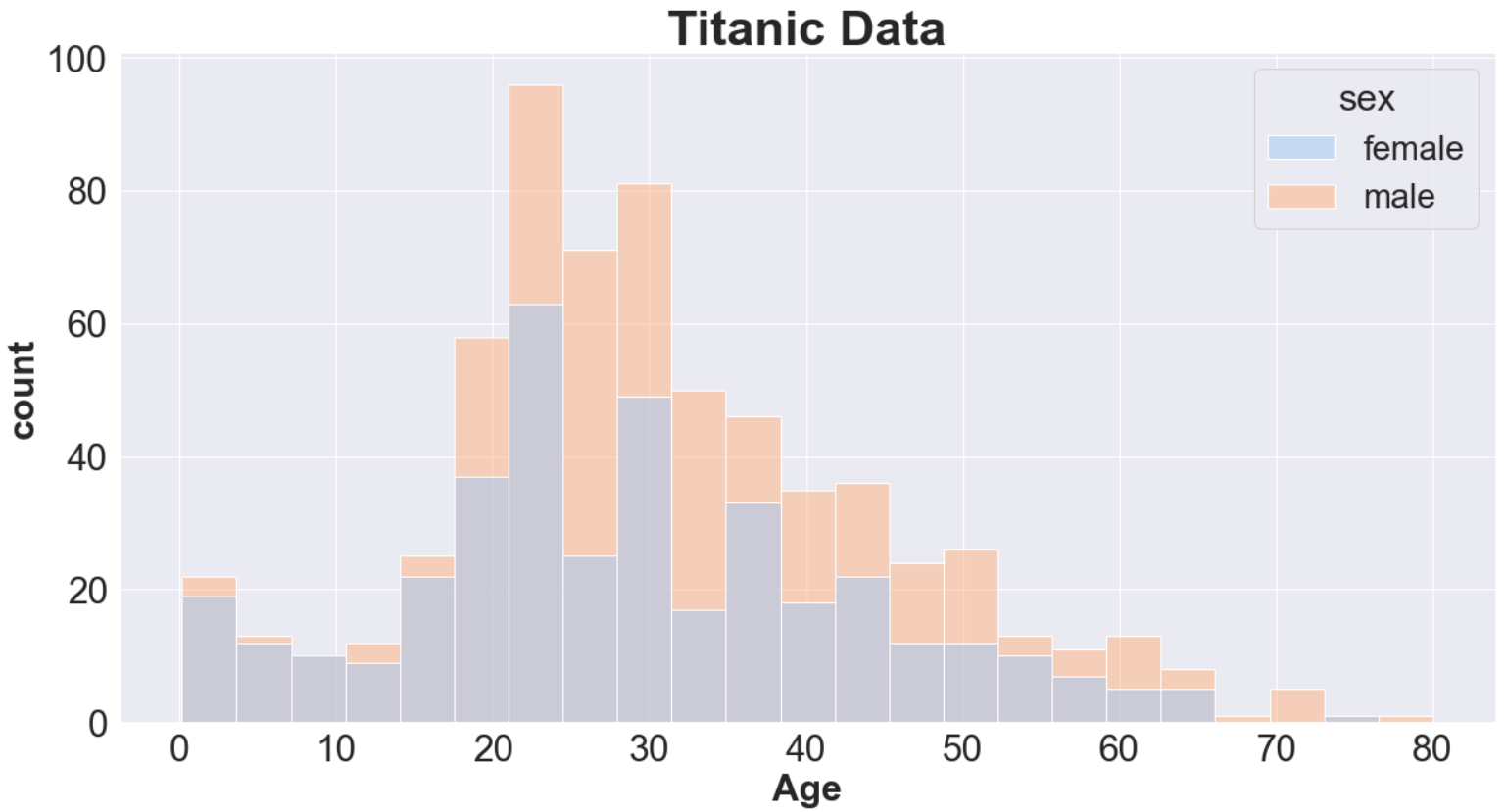
Histograms

- Intervals can hide individual values (difficult to detect relevant values)



Data visualization: distributions

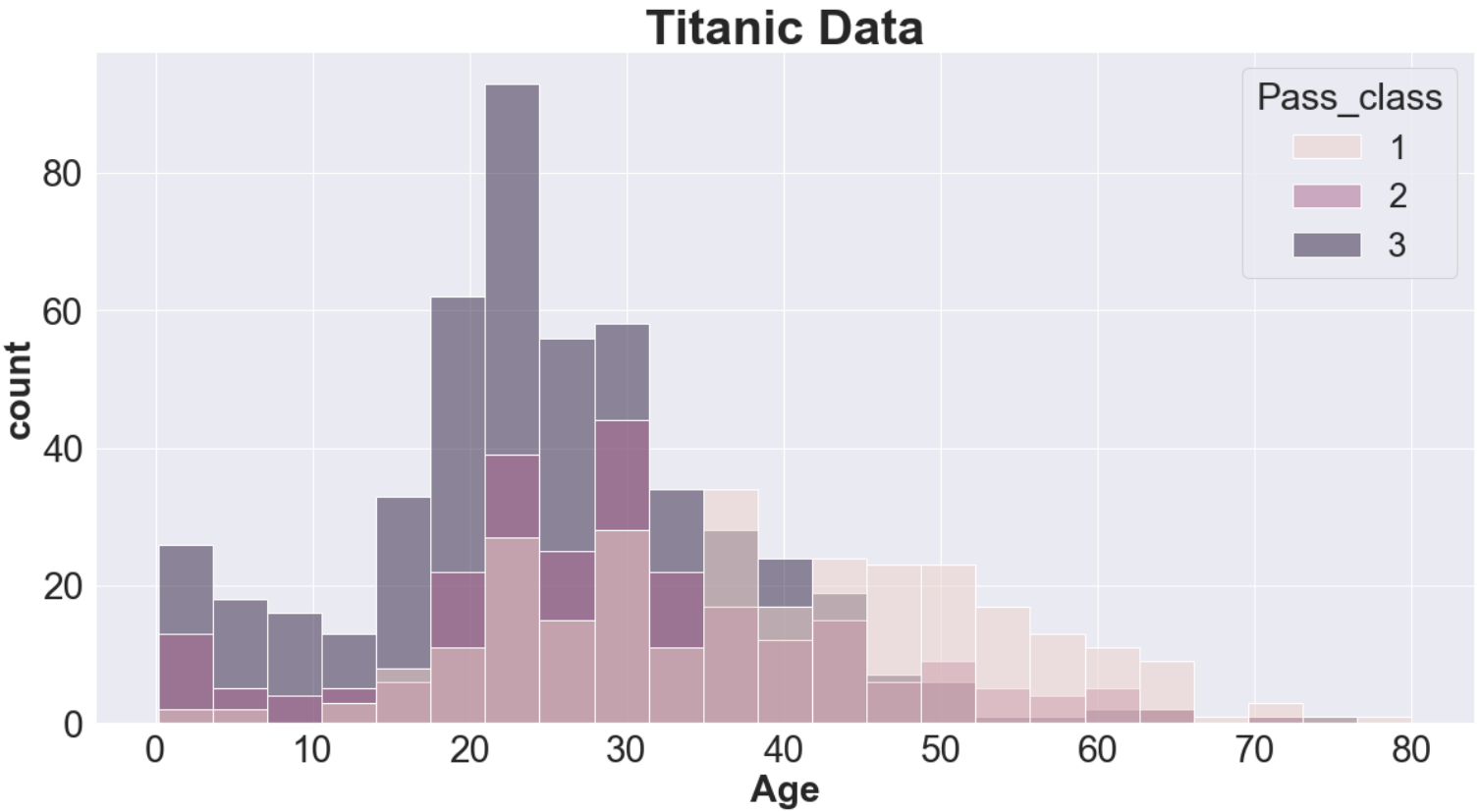
Histograms w.r.t. a categorical variable



Data visualization: distributions

Histograms w.r.t. a categorical variable

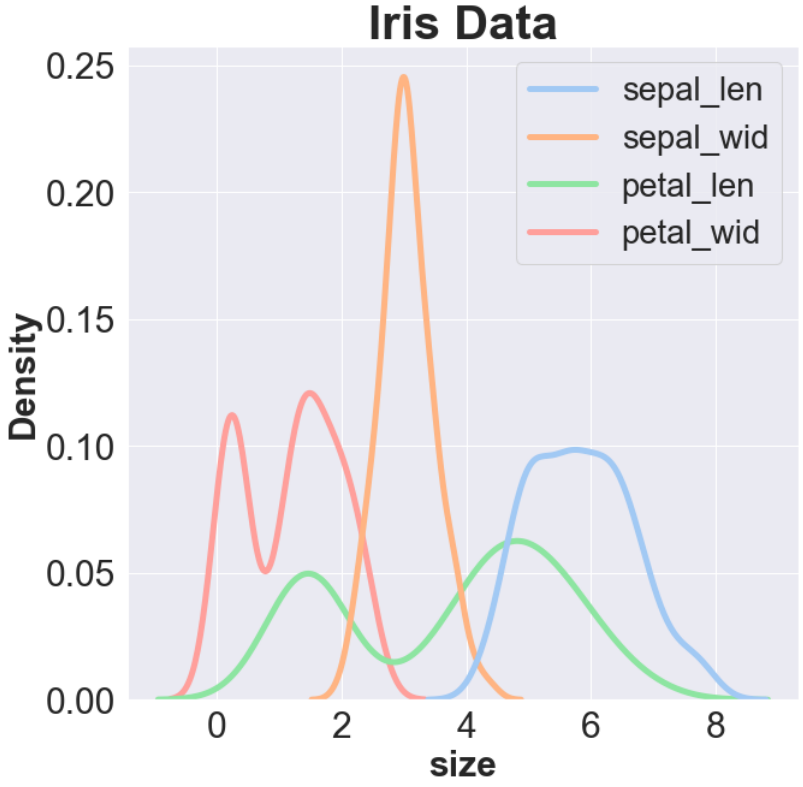
- Can be hard to compare several distributions

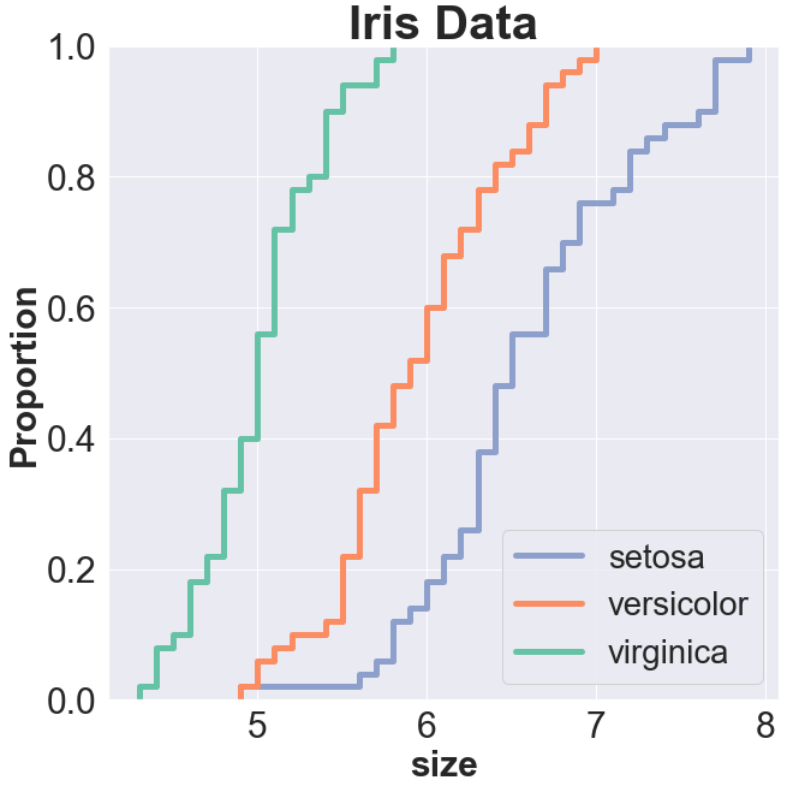


Data visualization: distributions

Density plots: overcoming some drawbacks of histograms

- Kernel Density Estimate (**KDE**)
- Cumulative Distribution Function (**CDF**)





Data visualization: distributions

Kernel Density Estimate (KDE)

- Smooth the estimates of the distribution of the values
- Kernel estimates compute the estimate of the distribution at a certain point by smoothly averaging over the neighboring points
- The density is estimated by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i}^{n} K\left(\frac{x - x_i}{h}\right)$$

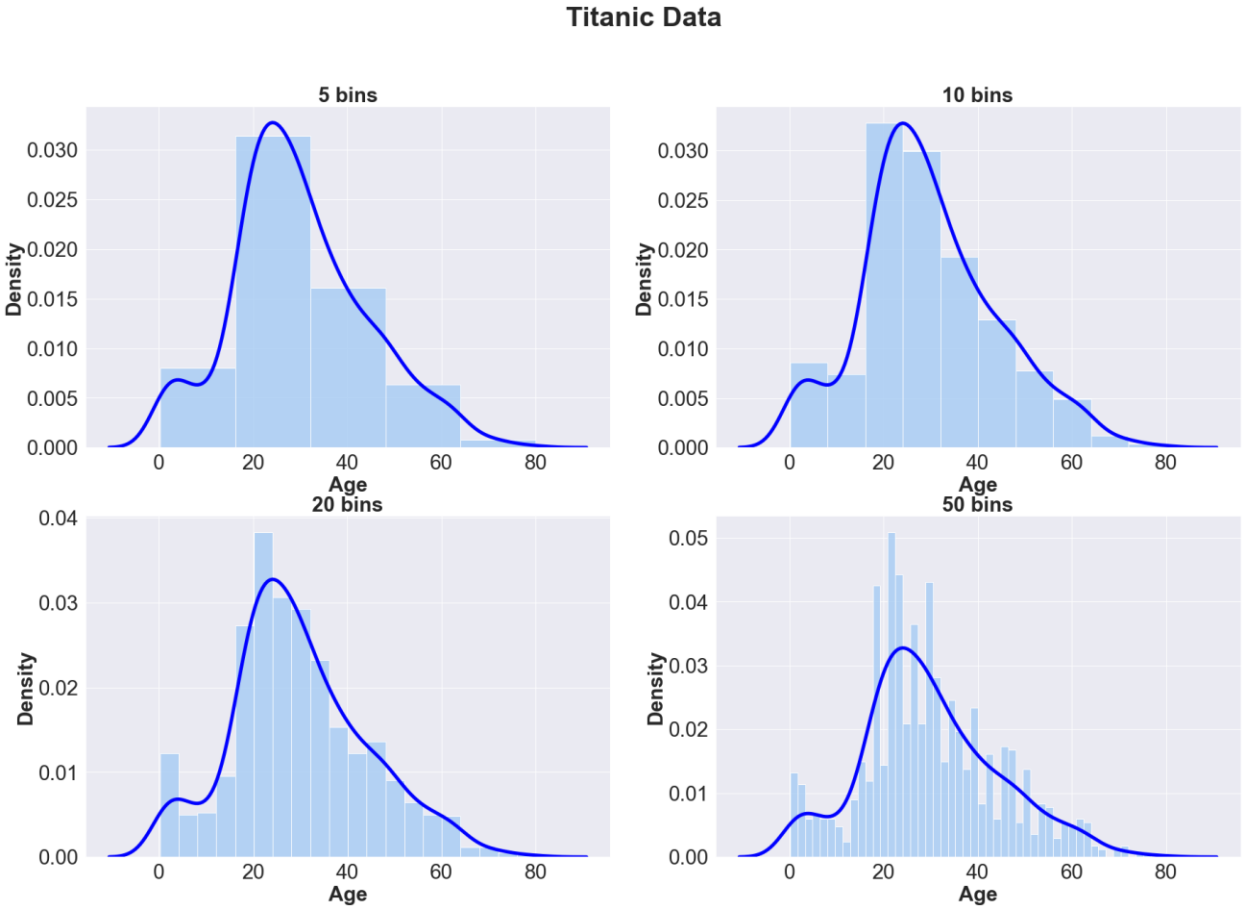
where

- $K(.)$ is the kernel — a non-negative function
- $h > 0$ is a smoothing parameter called the bandwidth.
- $x \in [x_i - h, x_i + h]$

Data visualization: distributions

Kernel Density Estimate (KDE)

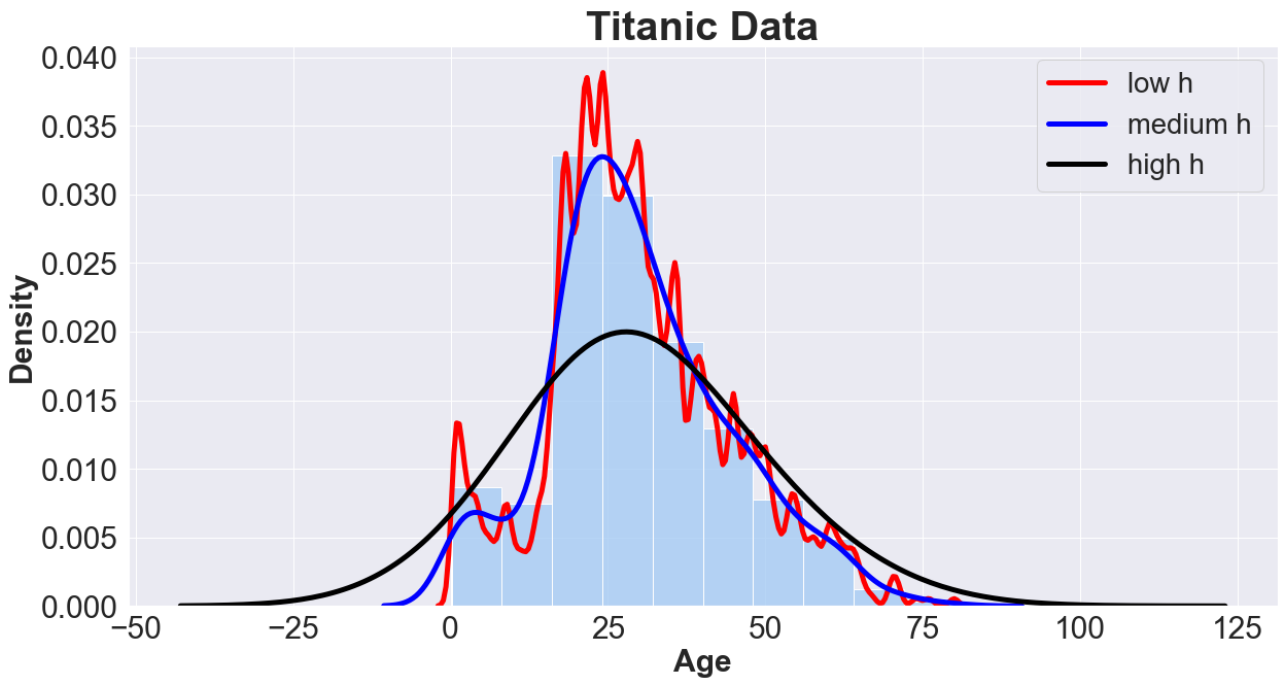
- Bandwidth (h) can be estimated with several methods



Data visualization: distributions

Kernel Density Estimate (KDE)

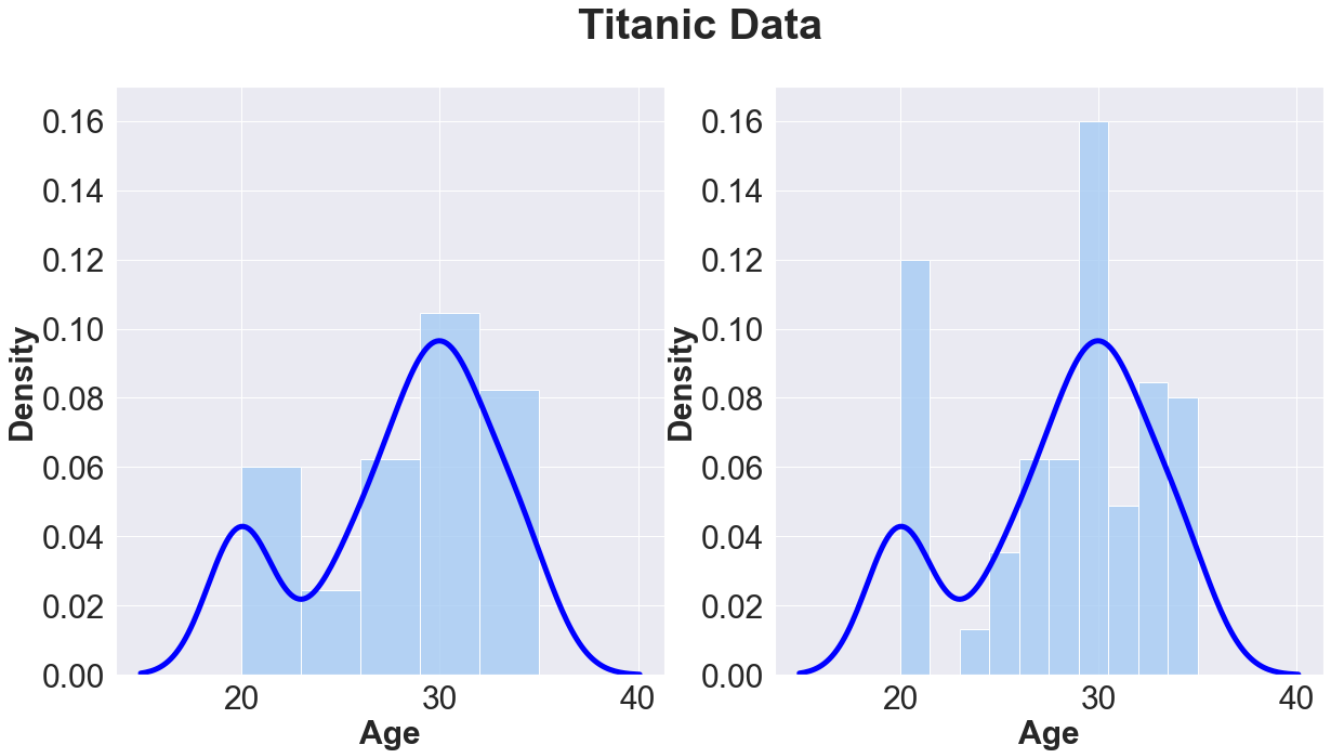
- Bandwidth (h) can be estimated with several methods
- Different values of the bandwidth (h) affect the shape of the KDE



Data visualization: distributions

Kernel Density Estimate (KDE)

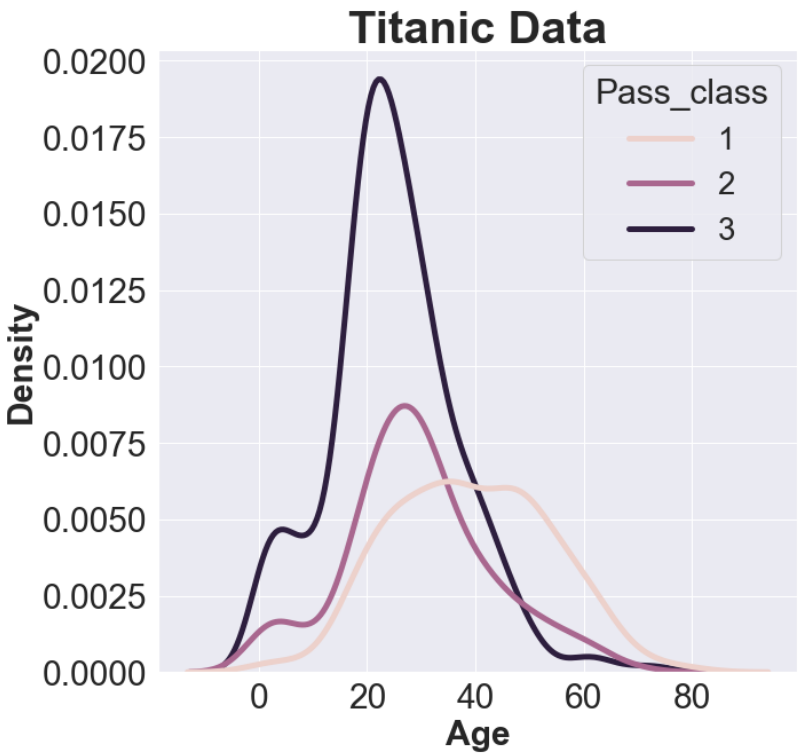
- Identify relevant values is easier (than with histograms)

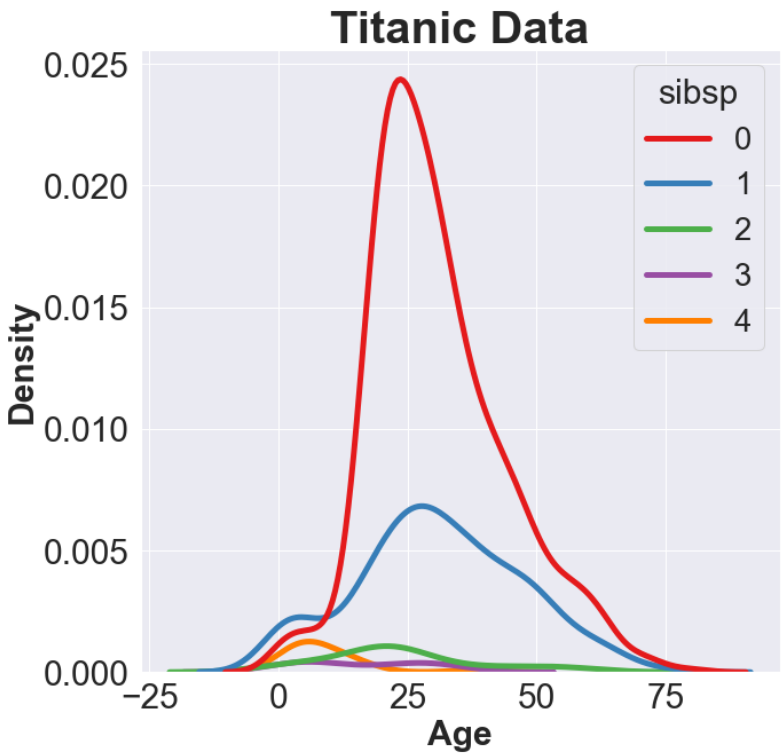


Data visualization: distributions

Kernel Density Estimate (KDE) w.r.t. a categorical variable

- Compare distributions is easier (than with histograms)





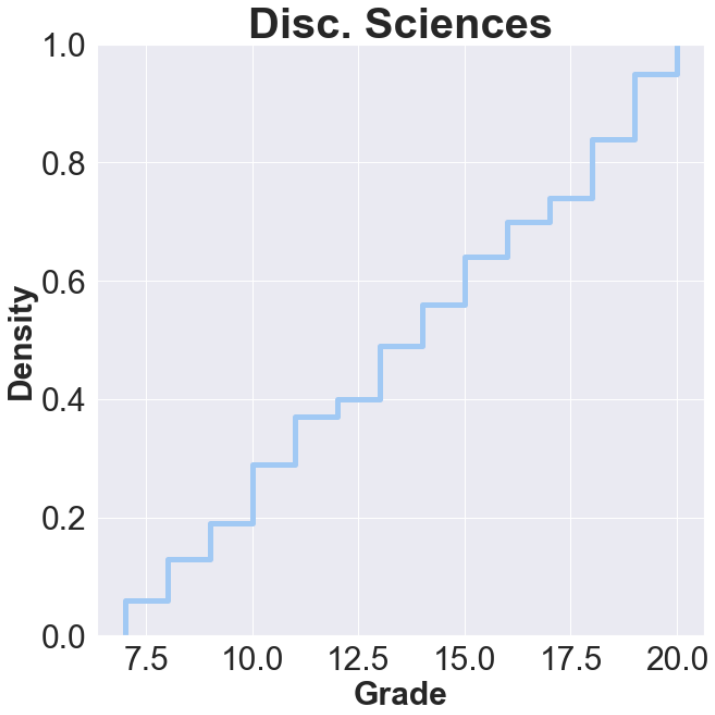
Data visualization: distributions

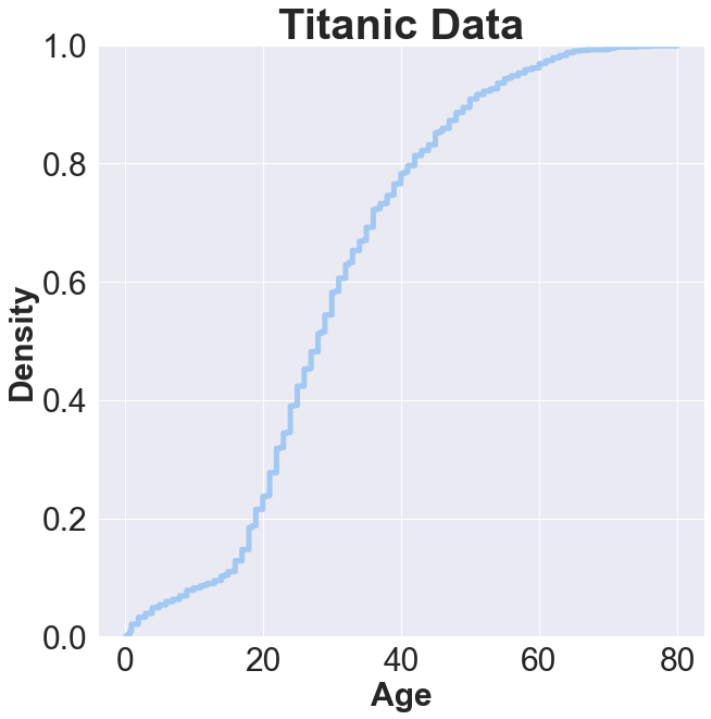
Cumulative Distribution Function (CDF)

- Gives the probability that the variable takes a value less than or equal to x:

$$F_X(x) = P(X \leq x)$$

- It allows to recognize a discrete variable at first glance

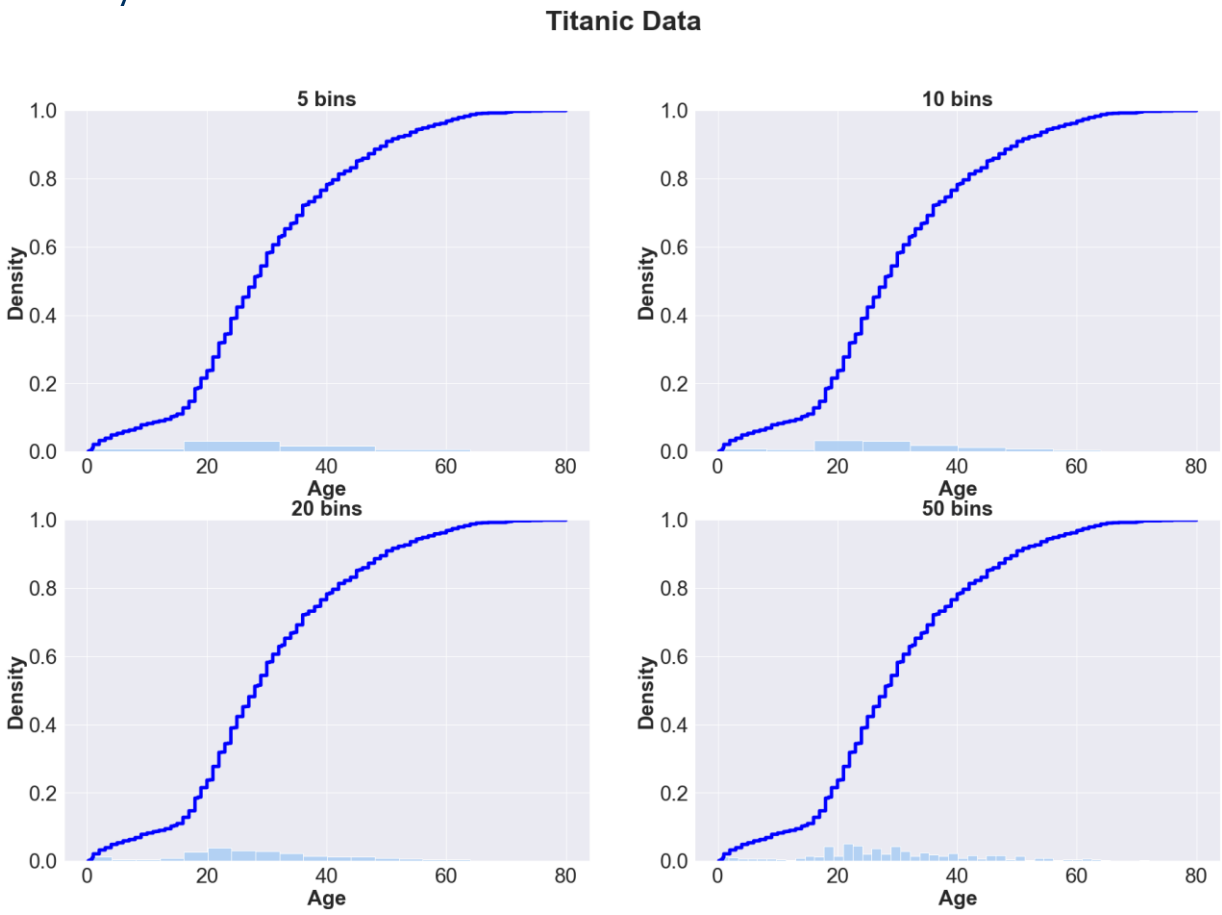




Data visualization: distributions

Cumulative Distribution Function (CDF)

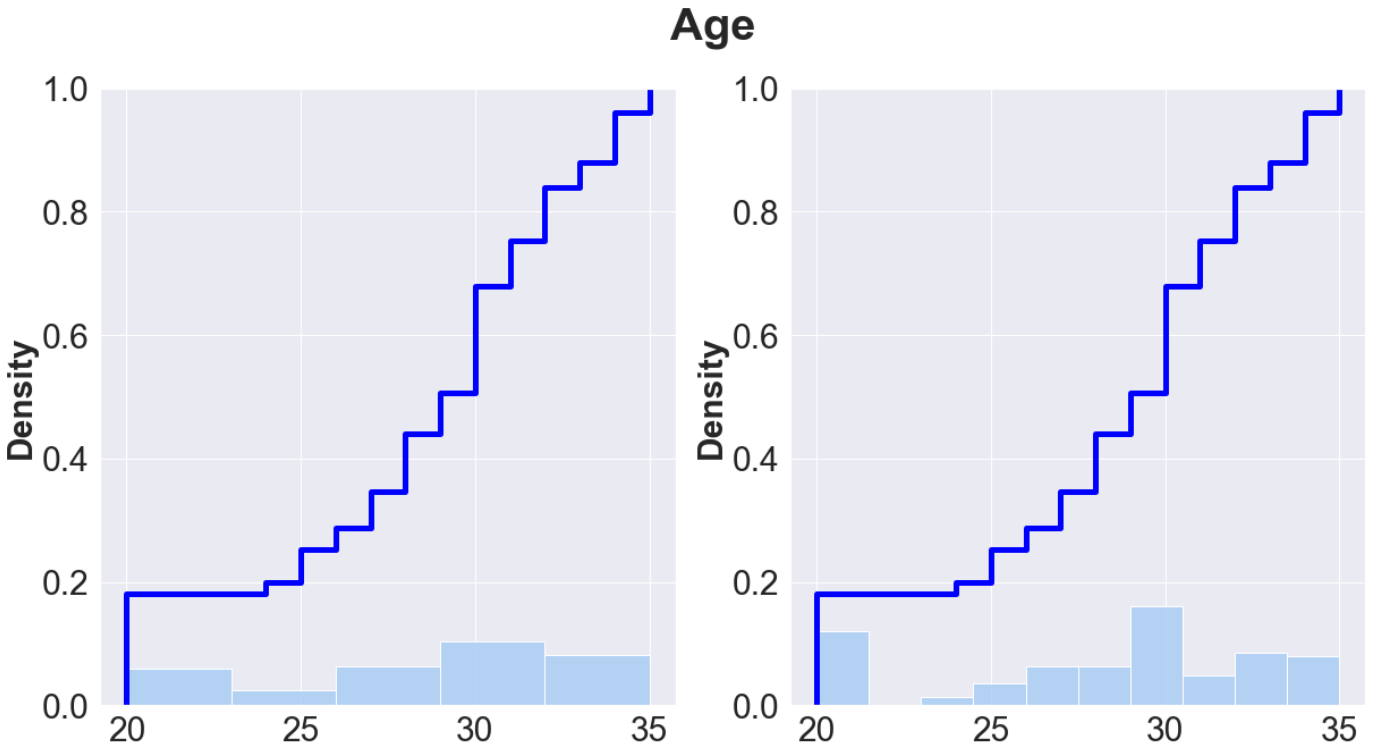
- It doesn't require any user choice



Data visualization: distributions

Cumulative Distribution Function (CDF)

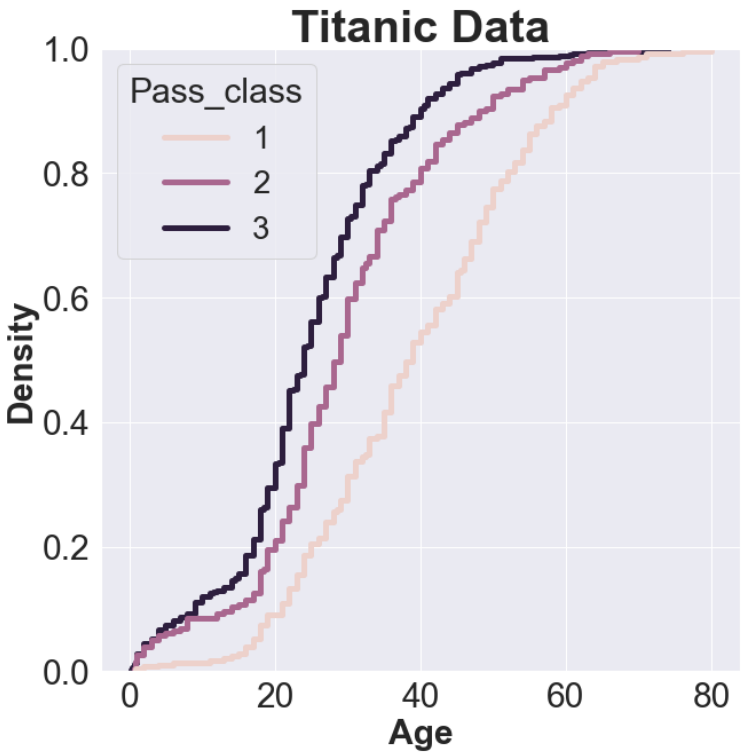
- Identify relevant values is easier (than with histograms)

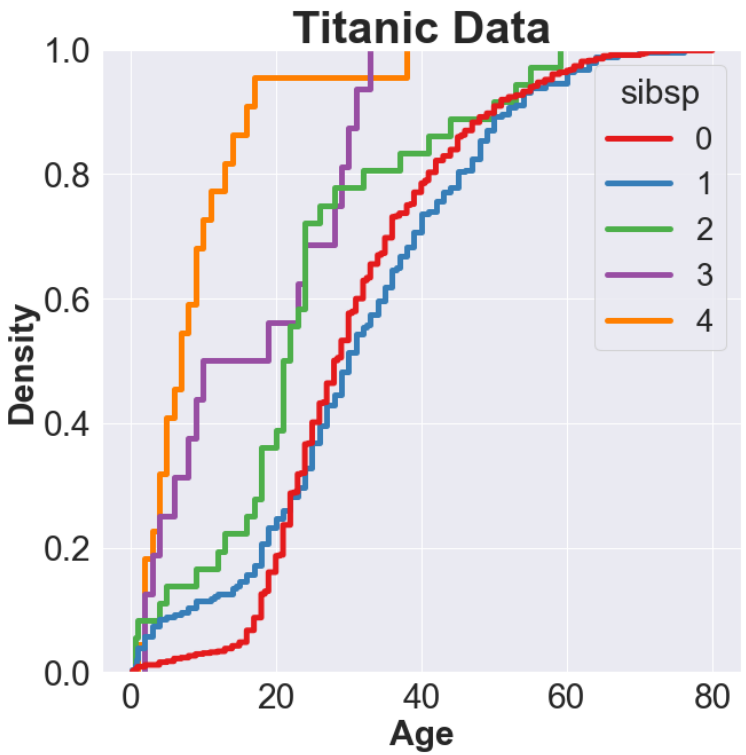


Data visualization: distributions

Cumulative Distribution Function (CDF) w.r.t. a categorical variable

- Compare distributions is easier (than with histograms)





Data visualization: distributions

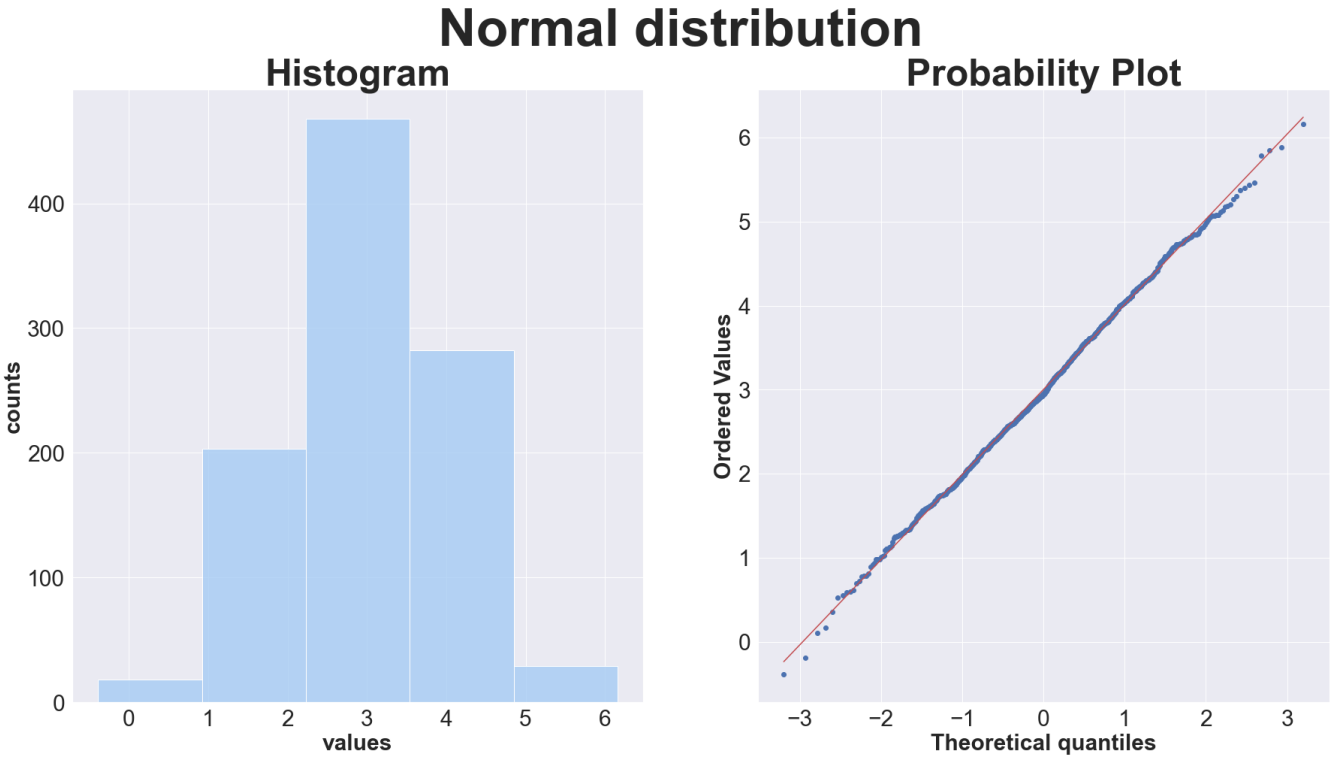
Quantile-Quantile plots (QQ plots)

- Graphs that can be used to compare the observed distribution against the Normal distribution
- Can be used to visually check the hypothesis that the variable under study follows a normal distribution
- Obviously, more formal tests also exist

Data visualization: distributions

Quantile-Quantile plots (QQ plots)

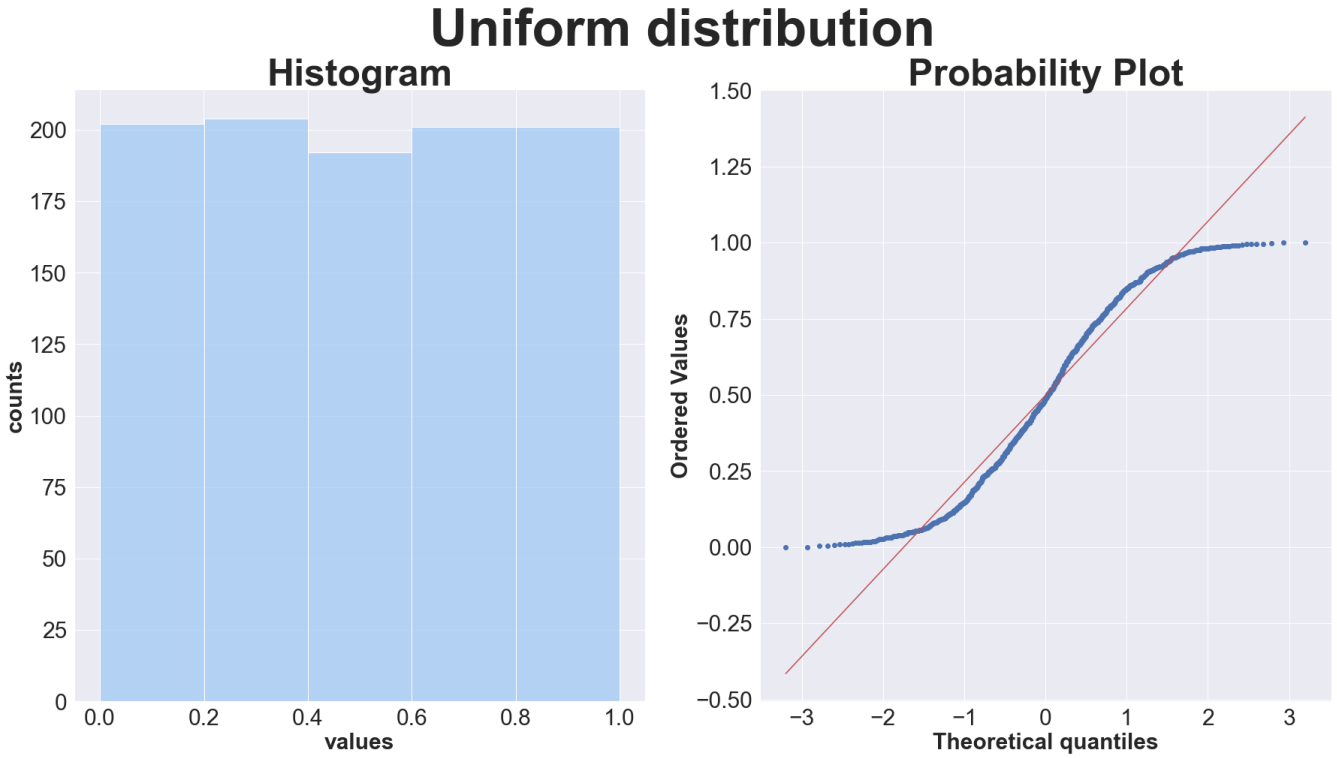
- It doesn't require any user choice
- Lesser intuitive than histograms or density plots



Data visualization: distributions

Quantile-Quantile plots (QQ plots)

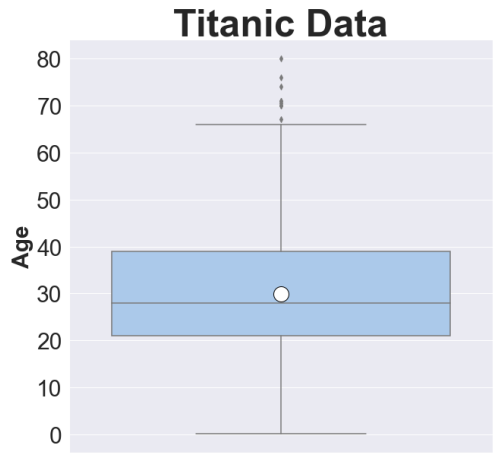
- It doesn't require any user choice
- Lesser intuitive than histograms or density plots

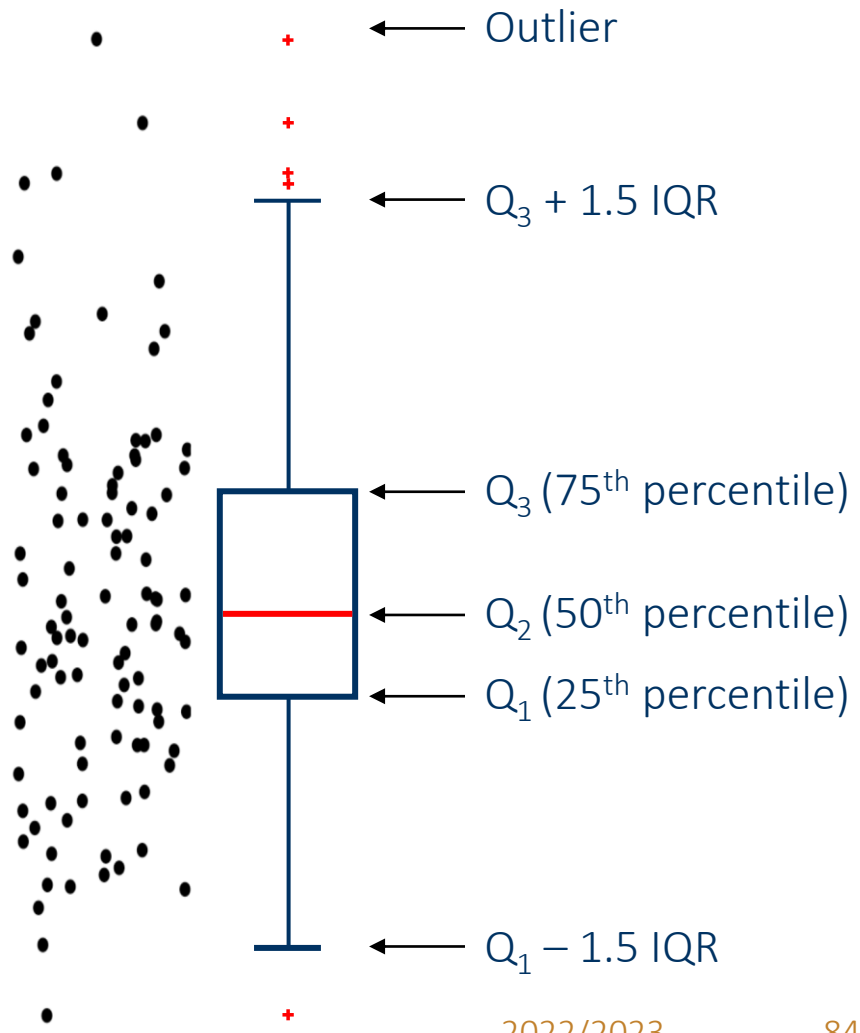


Data visualization: distributions

Boxplots (Tukey, 1977)

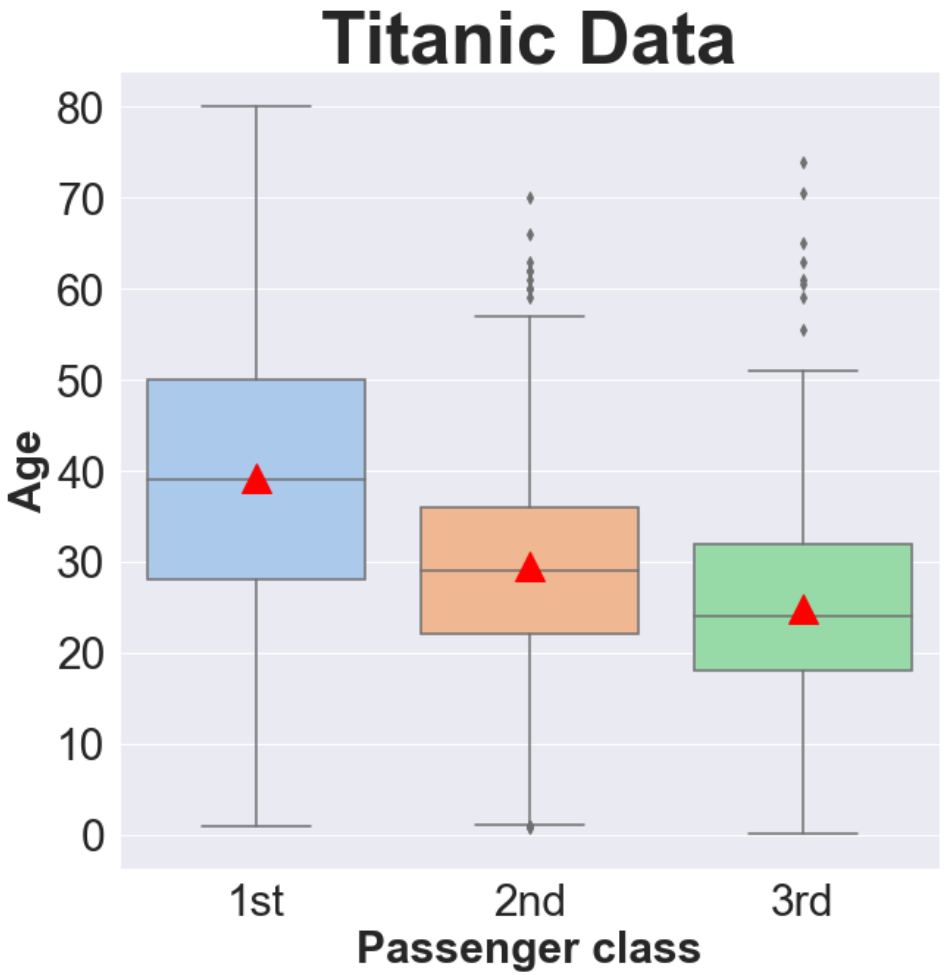
- Summary of a variable distribution
- Understand the spread of data
- Information on the interquartile range
- Determine if data is skewed
- Identify outliers (if any)





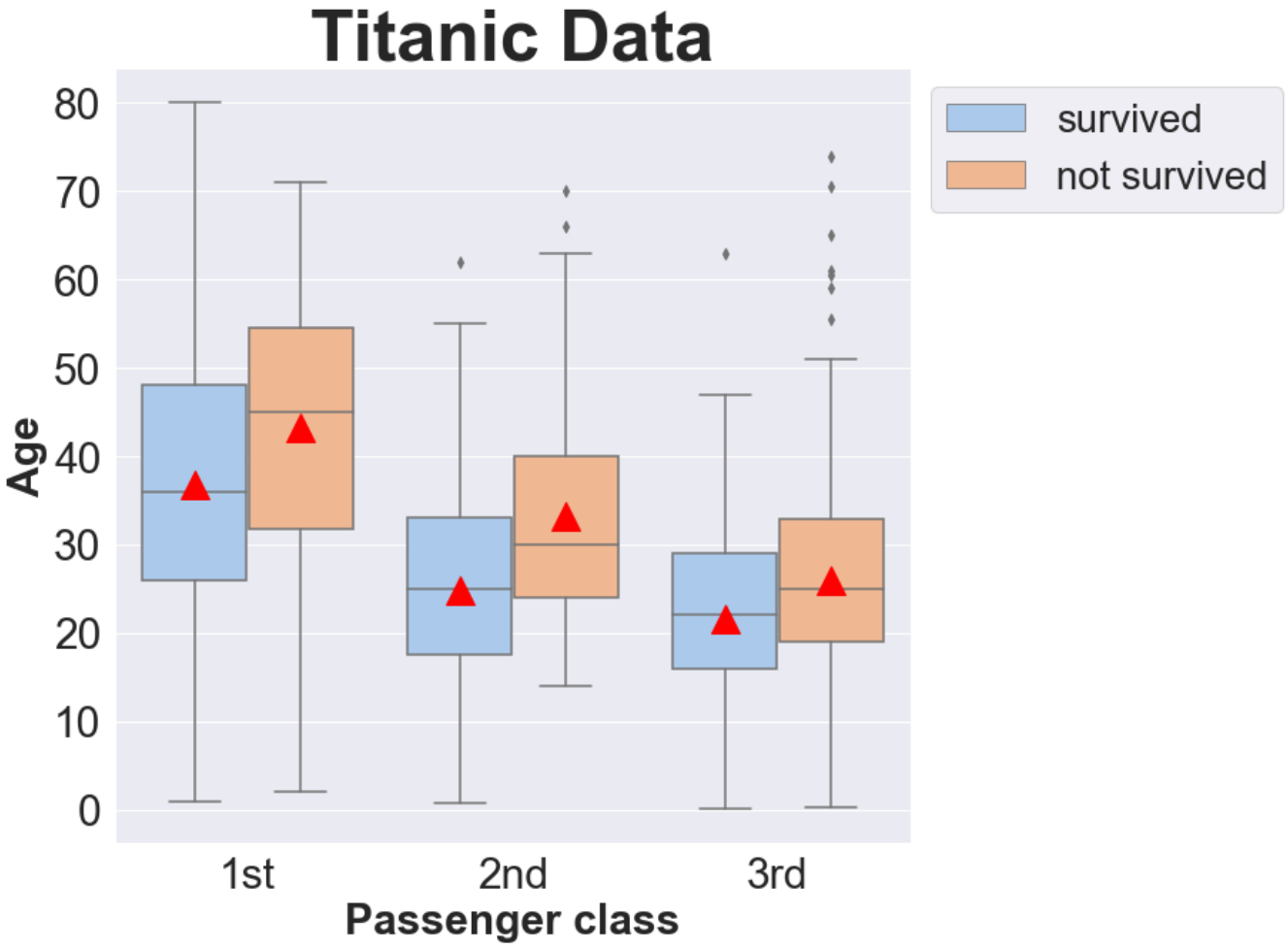
Data visualization: distributions

Boxplots w.r.t. a categorical variable



Data visualization: distributions

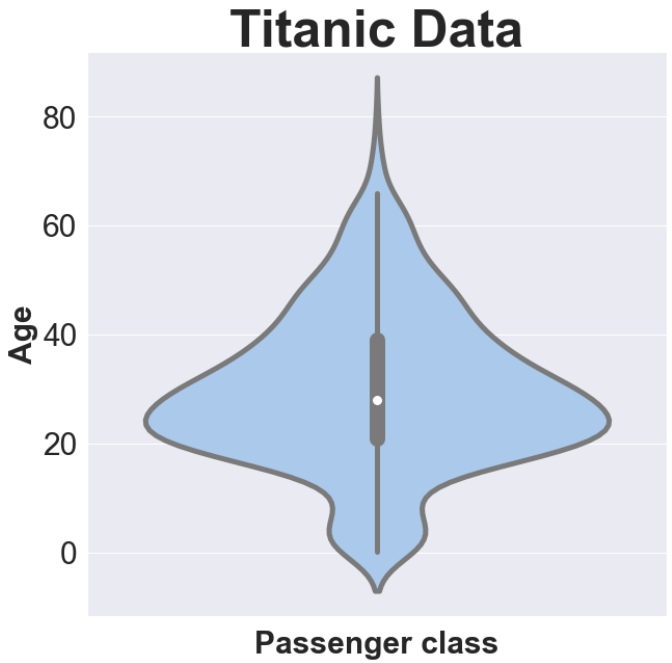
Boxplots w.r.t. two categorical variables



Data visualization: distributions

Violinplots*

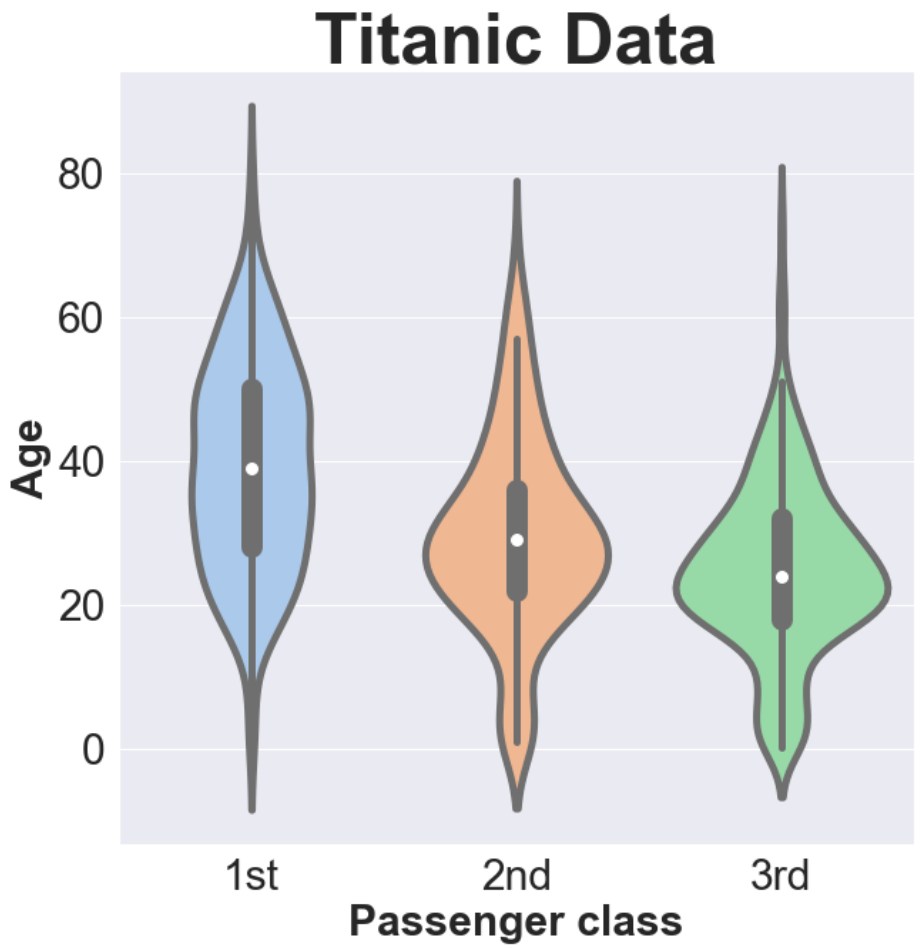
- Similar to boxplots
- Shows the probability density of the data at different value
  - can capture the distribution better compared to boxplot
  - can capture the structure in the data



*Need enough data to estimate the density

Data visualization: distributions

Violinplots w.r.t. a categorical variable



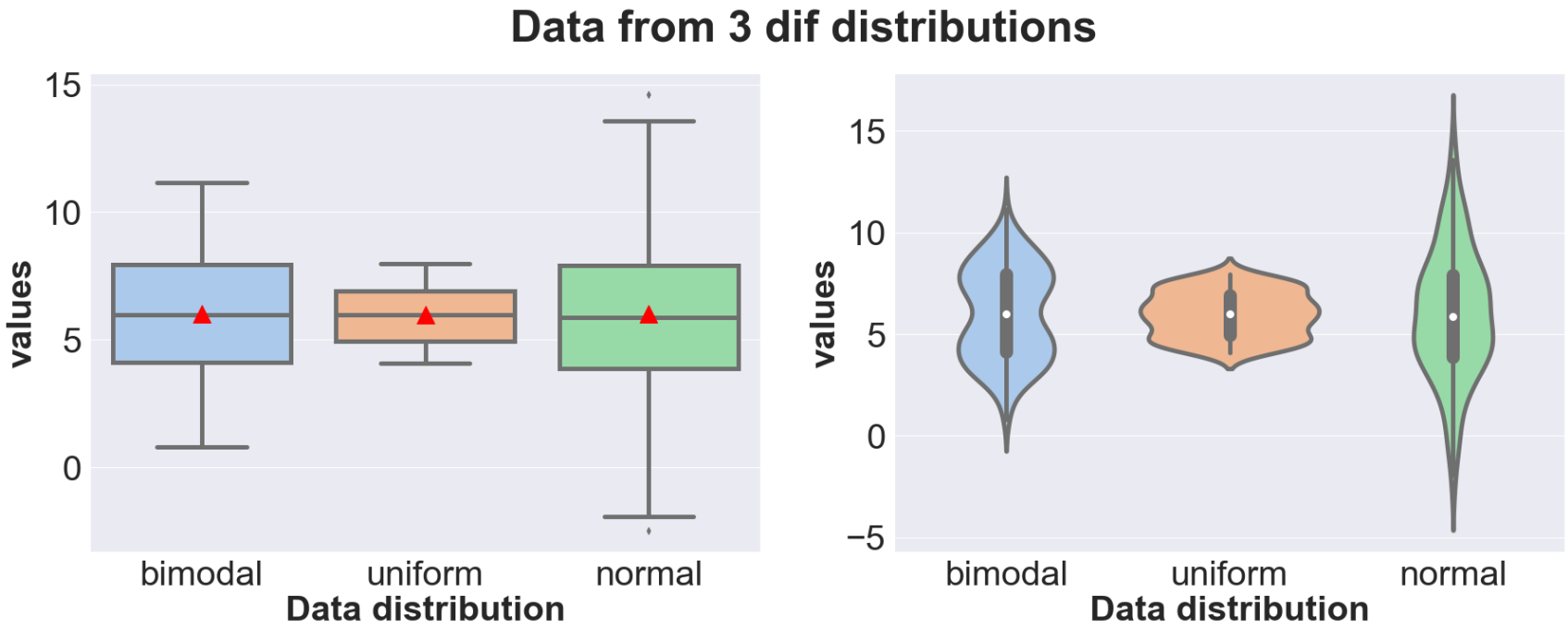
Data visualization: distributions

Violinplots w.r.t. a categorical variable

- Similar median values
- distributed differently

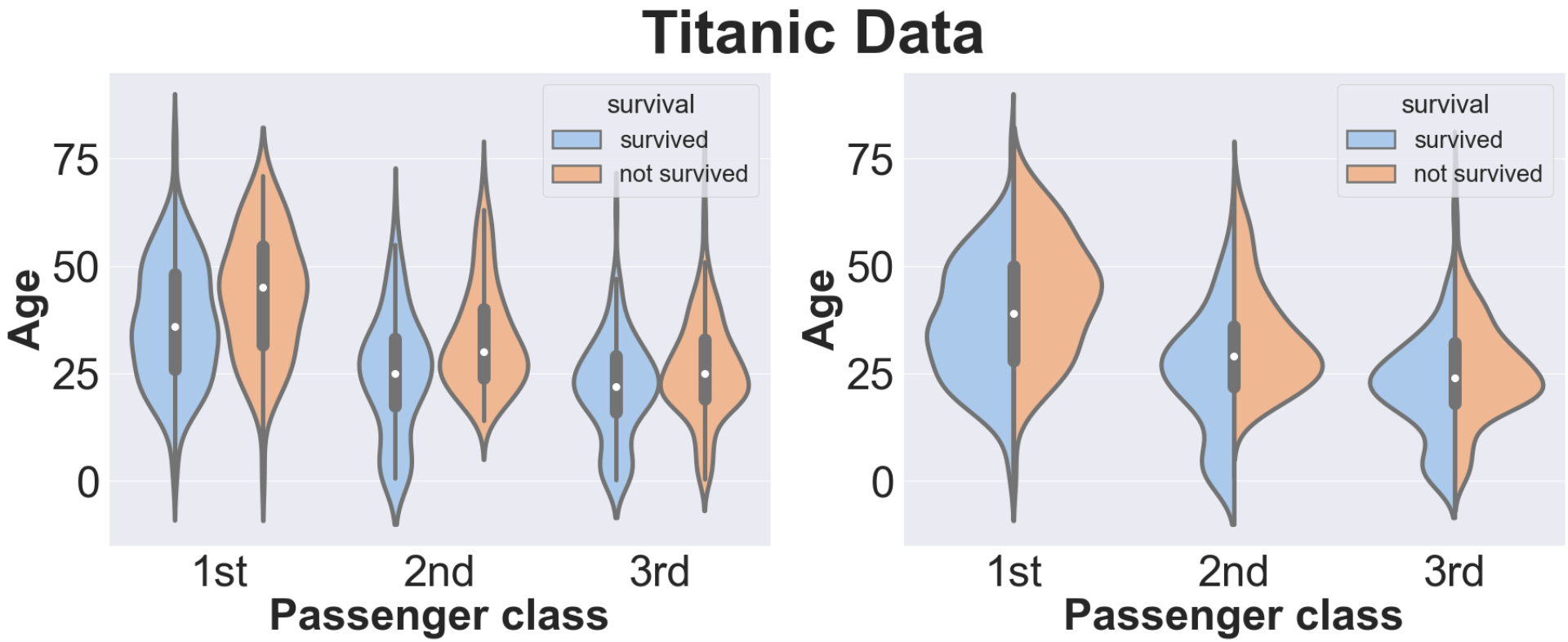
Violinplot: adds **density** information

- revealing the structure in the data



Data visualization: distributions

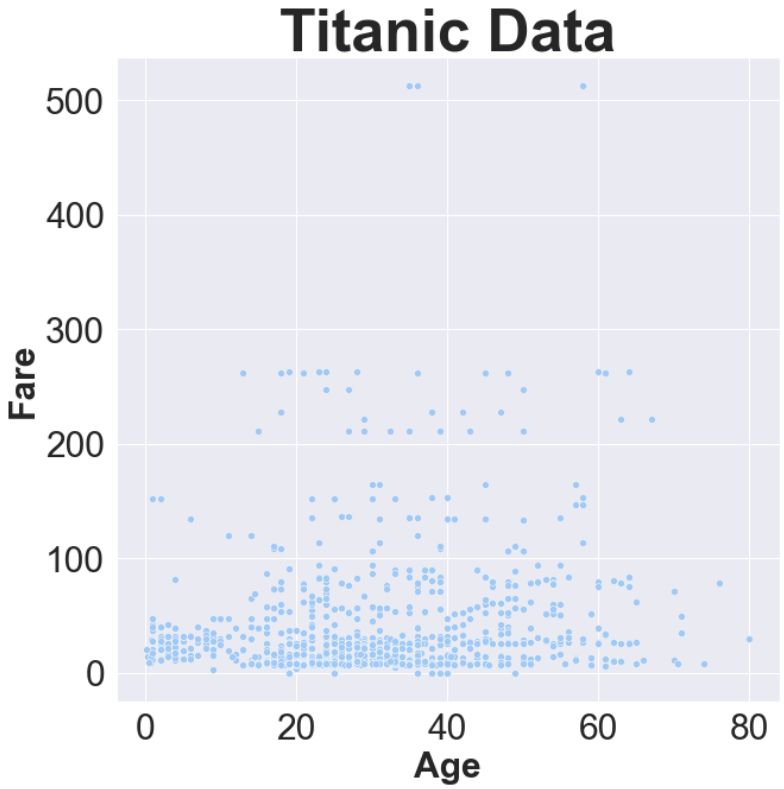
Violinplots w.r.t. two categorical variables



Data visualization: associations

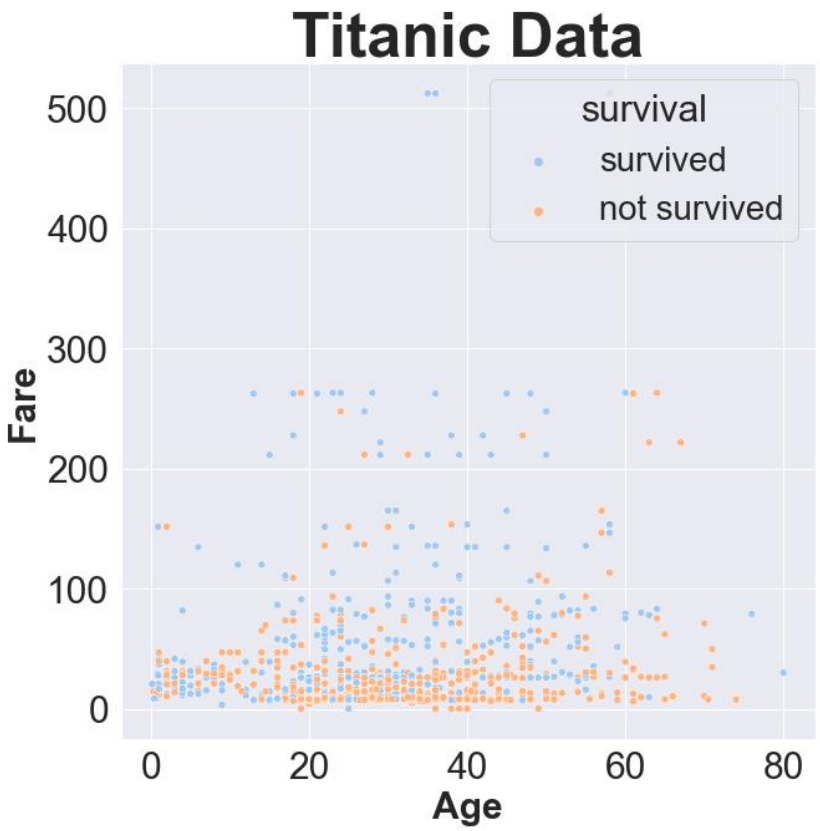
Scatter plots

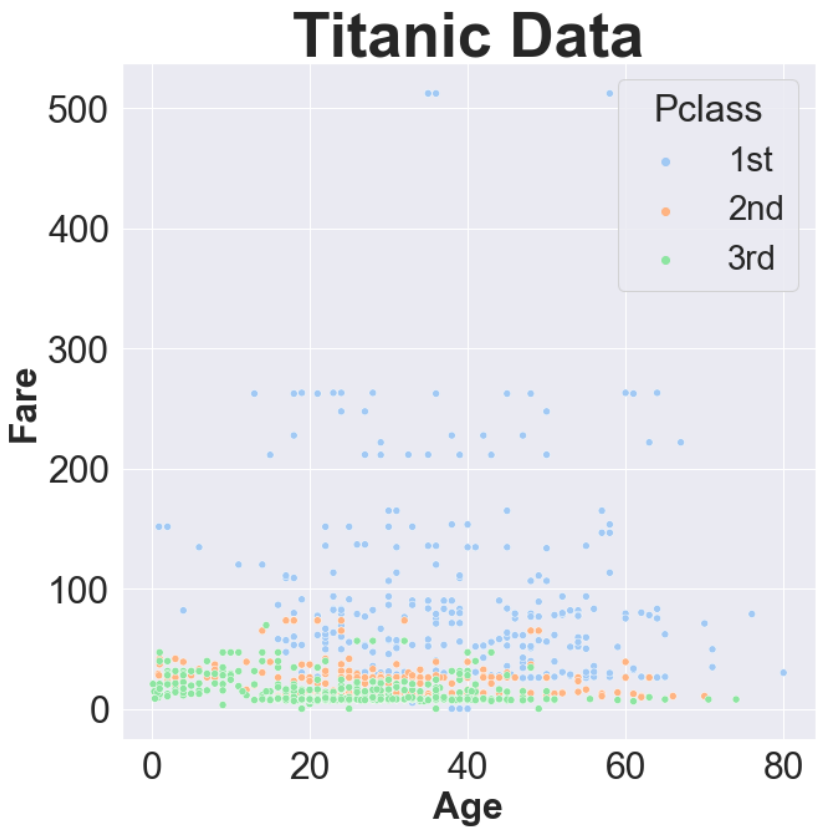
- Show one quantitative variable relative to another
- Data points are shown as dispersed cloud



Data visualization: associations

Scatter plots w.r.t categorical variable

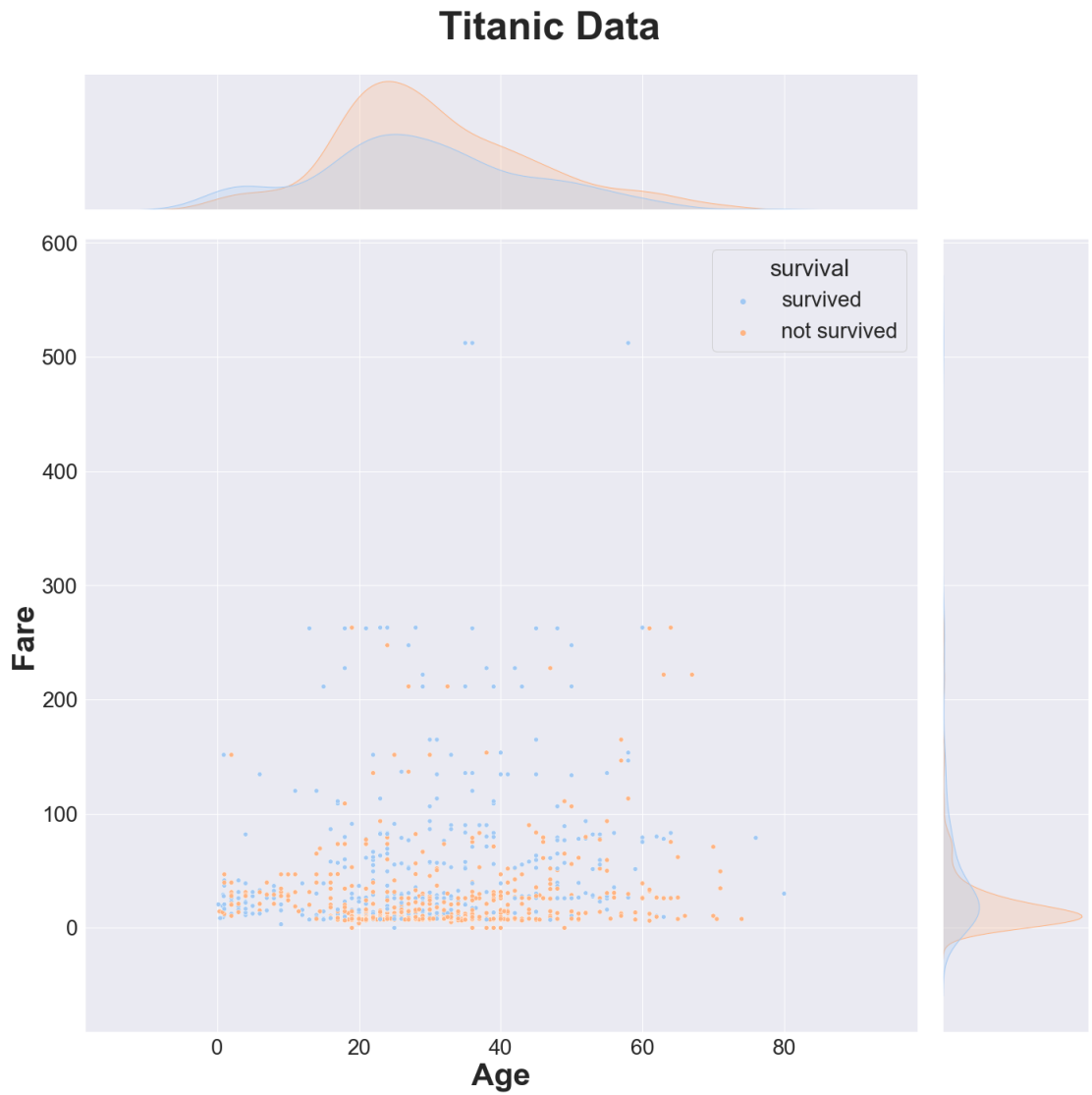




Data visualization: associations

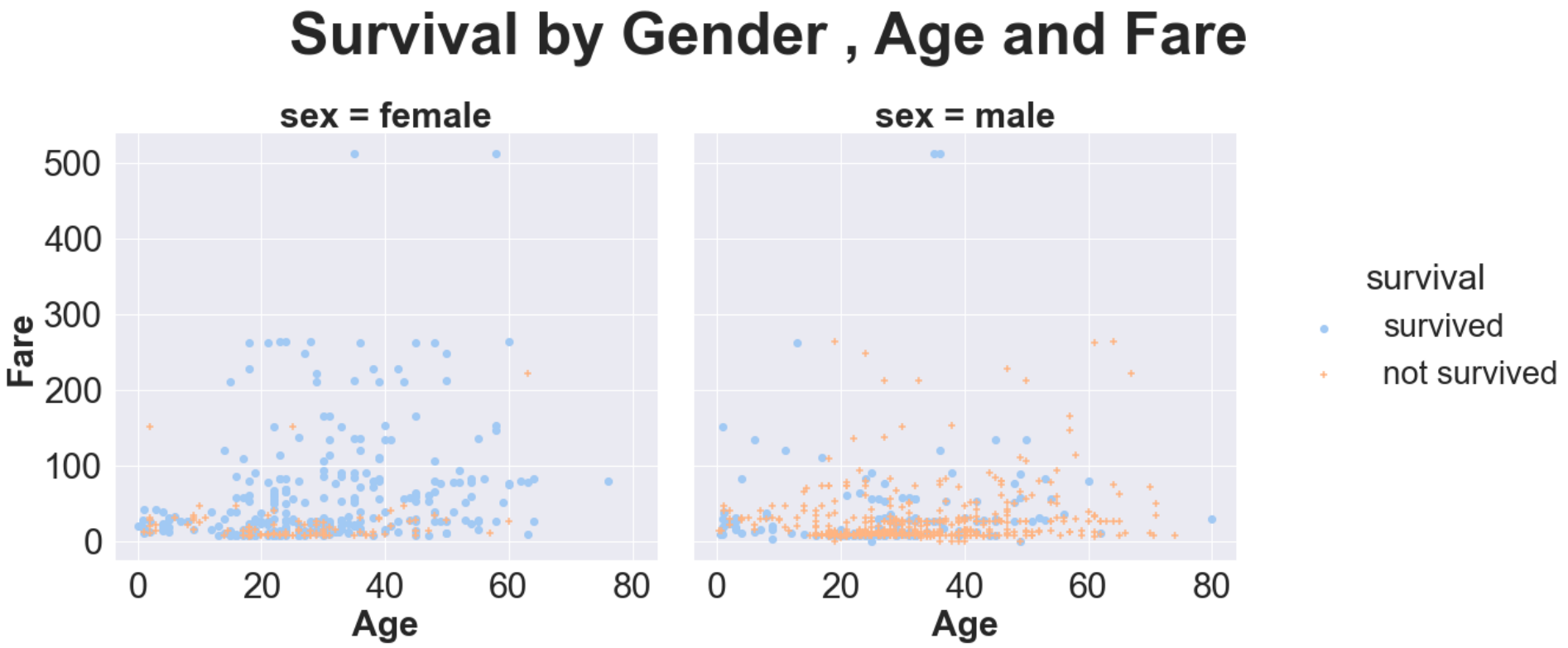
Scatter plots

Multiple views on the data



Data visualization: associations

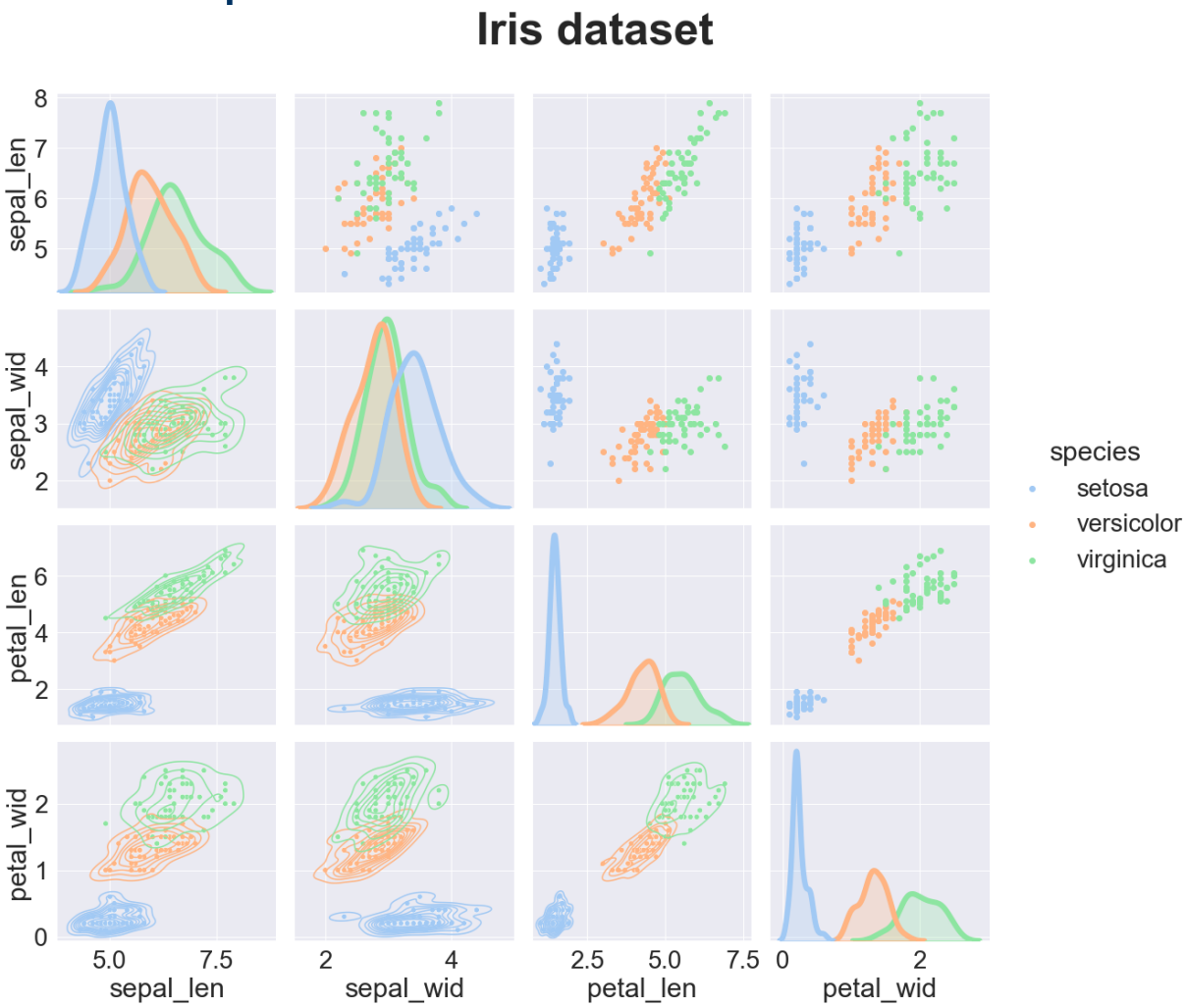
Scatter plots w.r.t. two categorical variables



Data visualization: associations

Scatter matrix: pairwise scatter plots

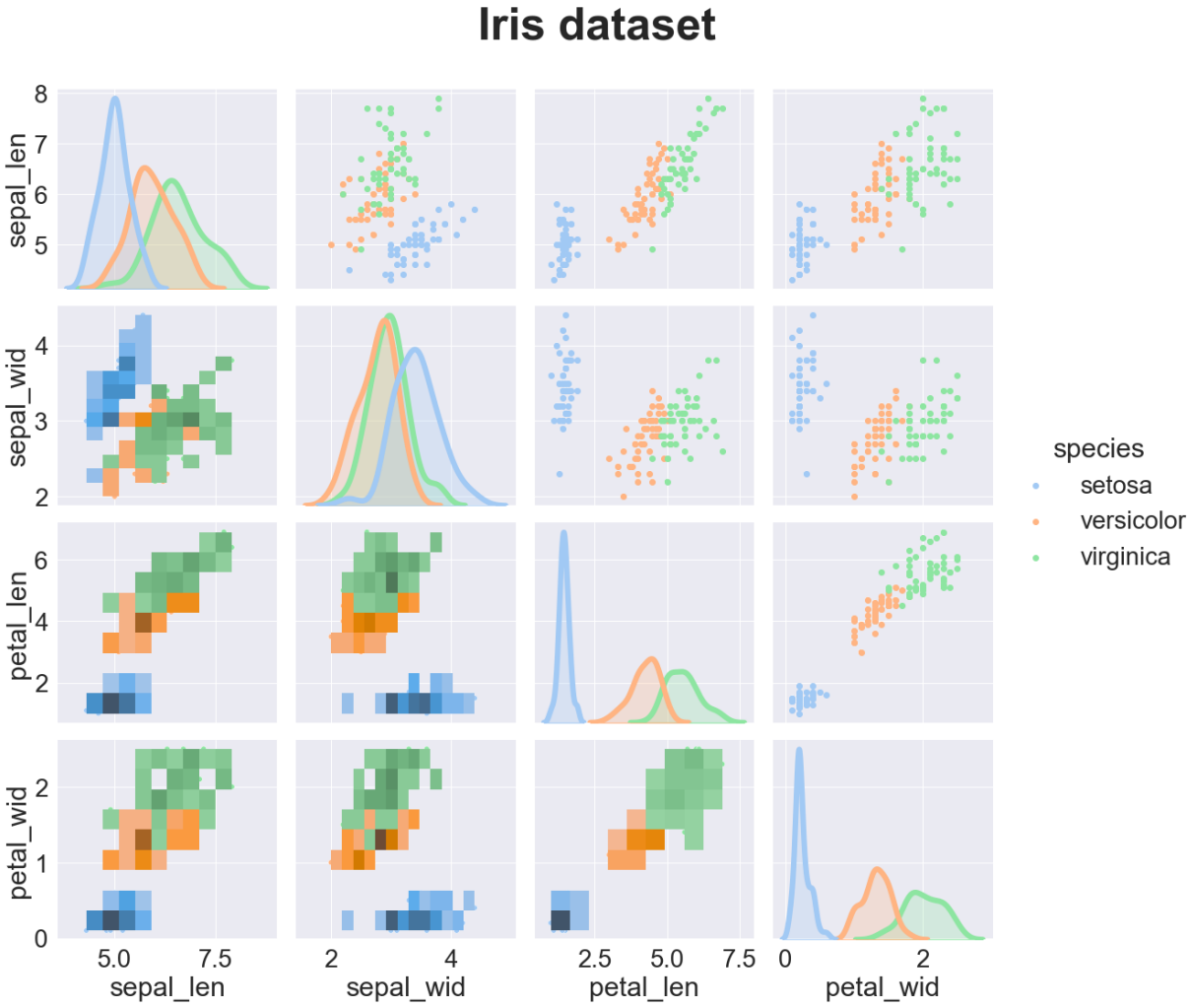
- all-against-all



Data visualization: associations

Scatter matrix: pairwise scatter plots

- all-against-all



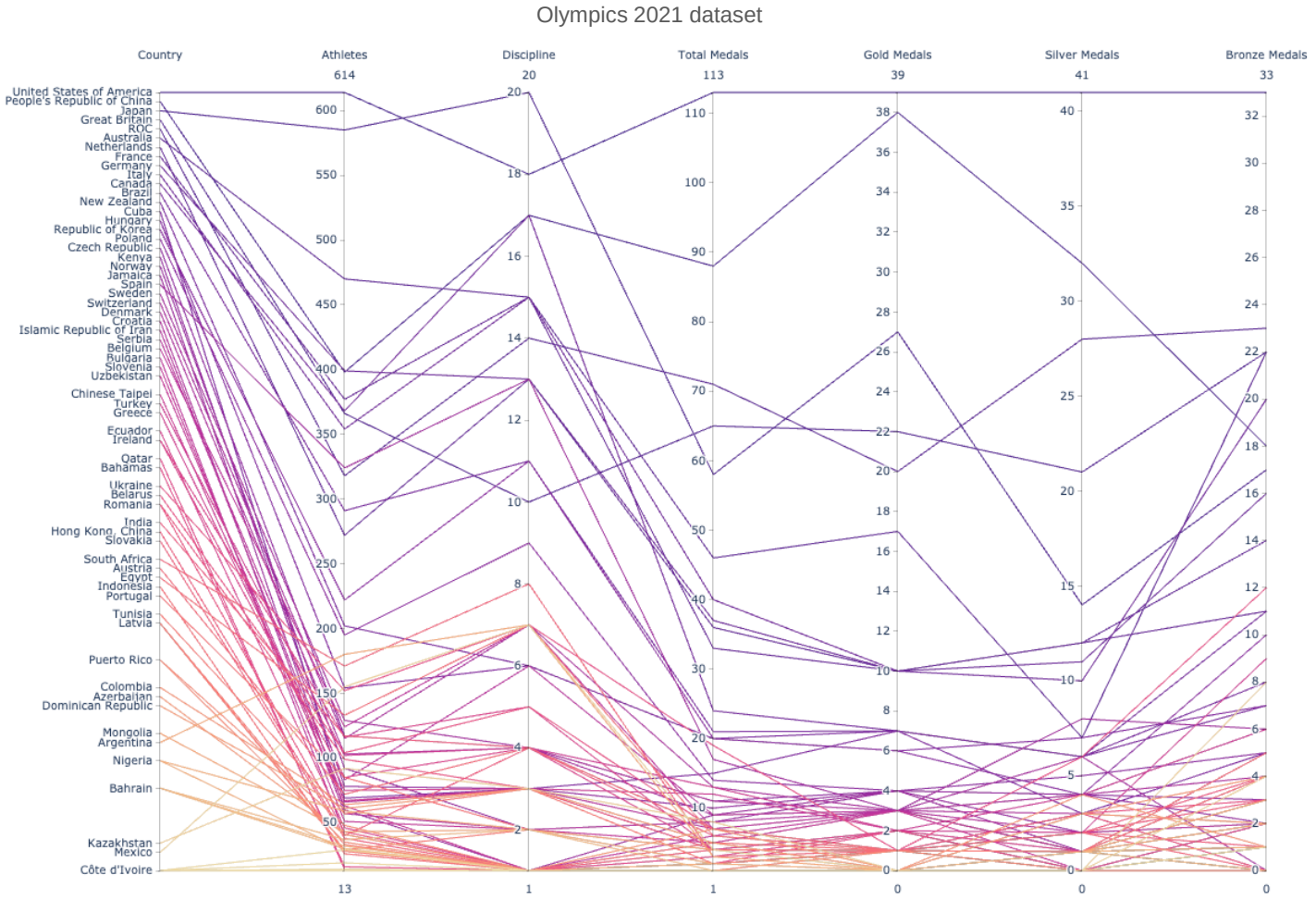
Data visualization: associations

Parallel coordinates plots

- Map n-dimensional relations into 2D patterns
- Allows a comparison of the samples across multiple numerical variables
- Used for high dimensional numerical
- Each variable is represented by a separate axis (equidistants)
  - All the axes are equally spaced and parallel to each other
  - Each axis can have a different scale and unit of measurement: scaled to the [minimum, maximum] of the corresponding variable
- Each sample/observation corresponds to a polygonal line, intersecting the axis of each numerical variable at the point which corresponds to the value for that variable
- The order of variables is important to show groups of observations

Data visualization: associations

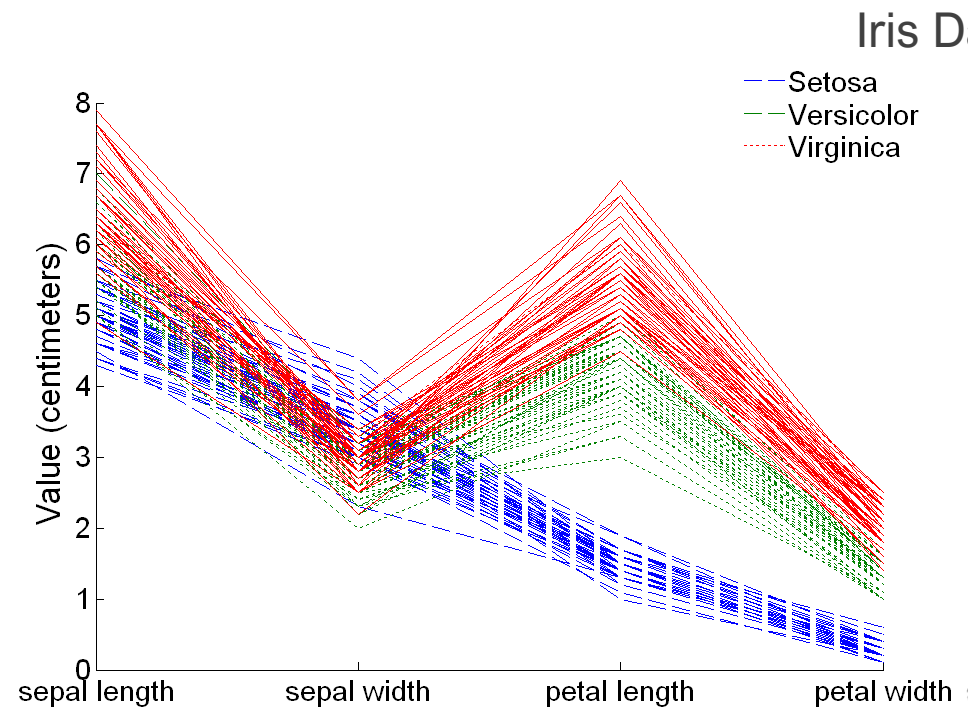
Parallel coordinates plots

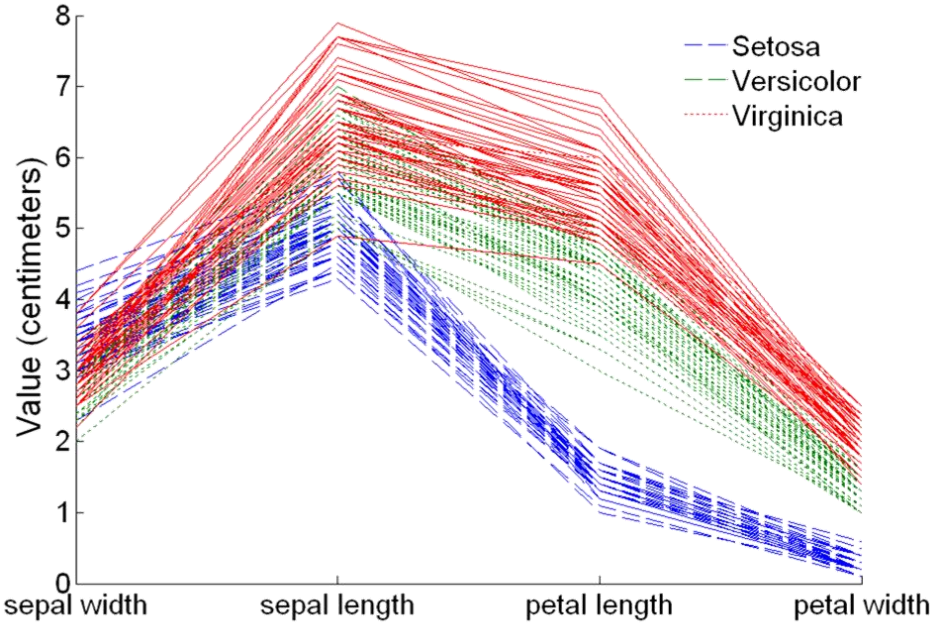


Data visualization: associations

Parallel coordinates plots

- The order of variables is important to show groups of observations

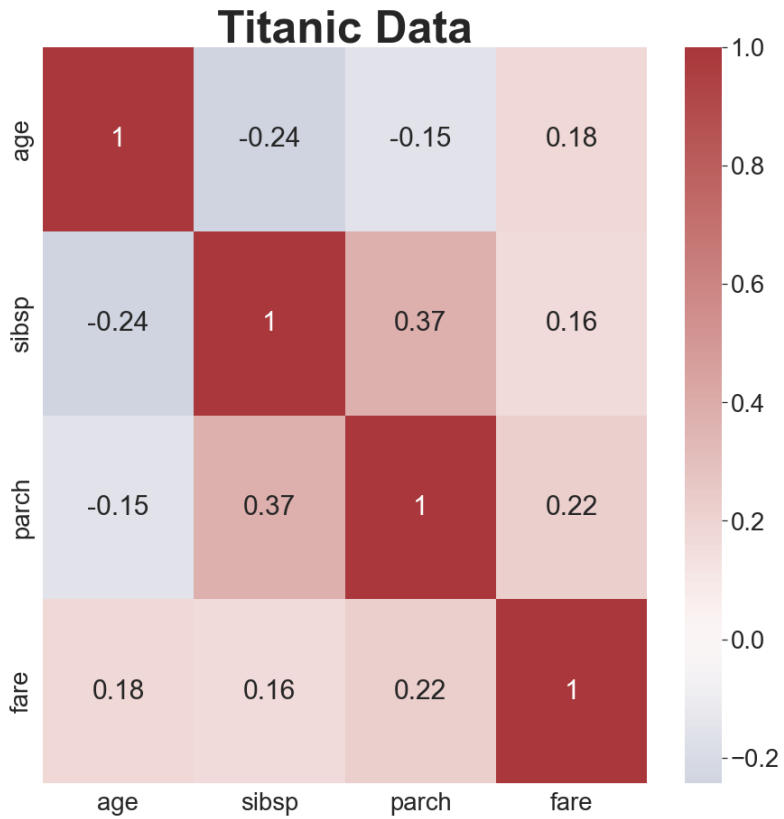




Data visualization: associations

Correlograms

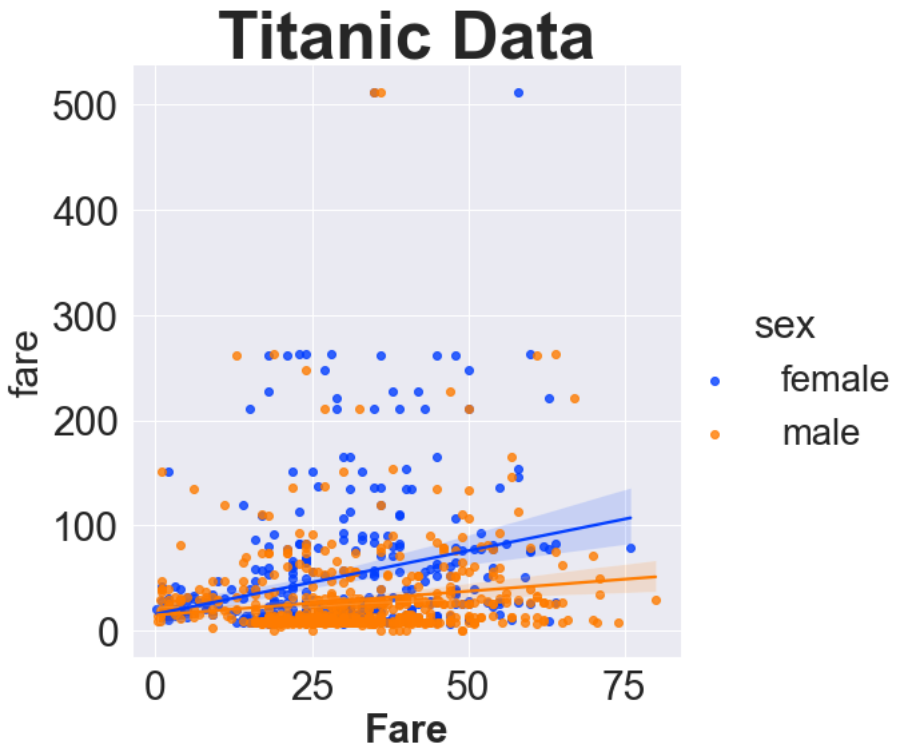
- Display the correlation between every pair of **numeric** variables



Data visualization: trends

Scatterplots

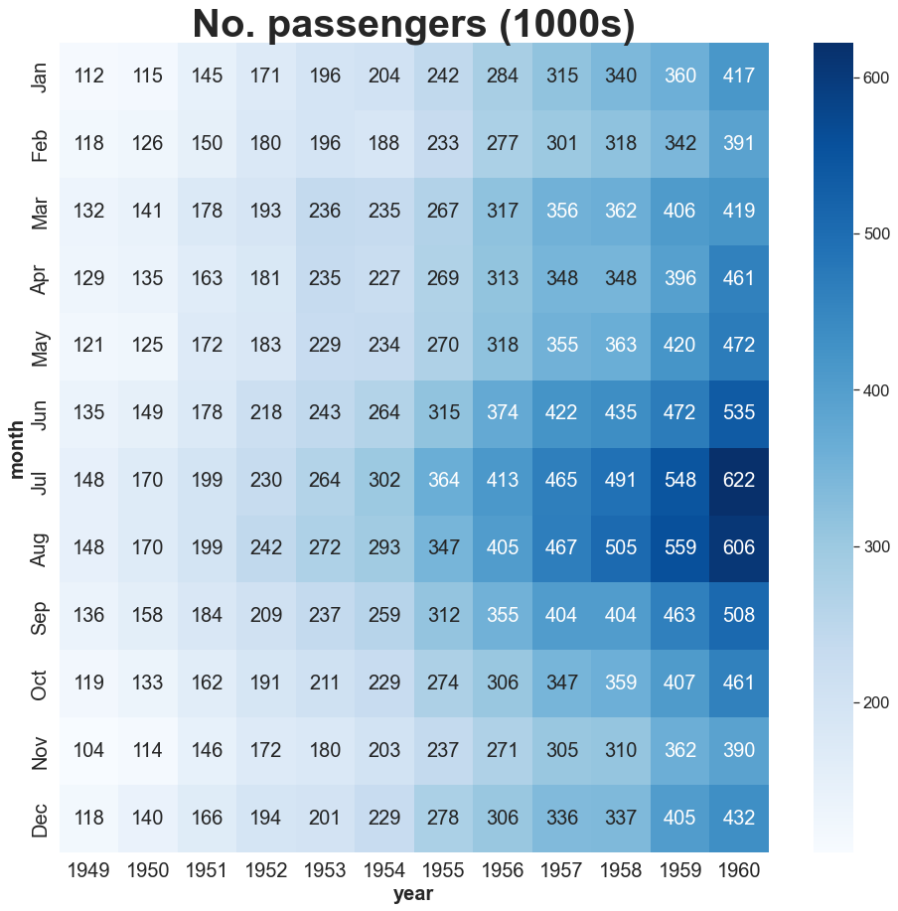
- Several functions allows to approximate the relationship between two numeric variables
- Scatter plot helps to perceive the trends



Data visualization: trends

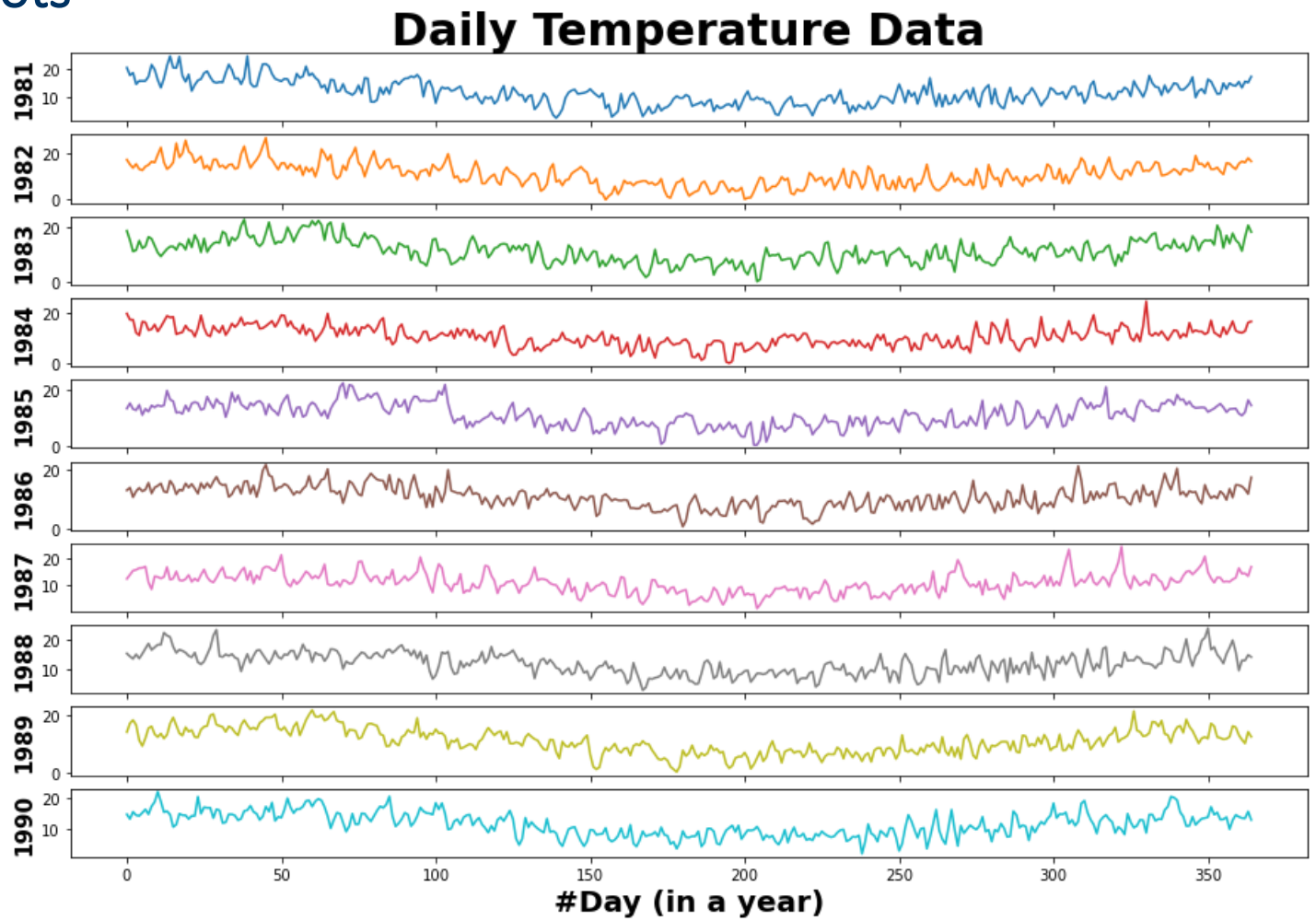
Heatmaps

- Display the cross tabulation of two categorical variables
- Allow to see gradual trends in data



Data visualization: trends

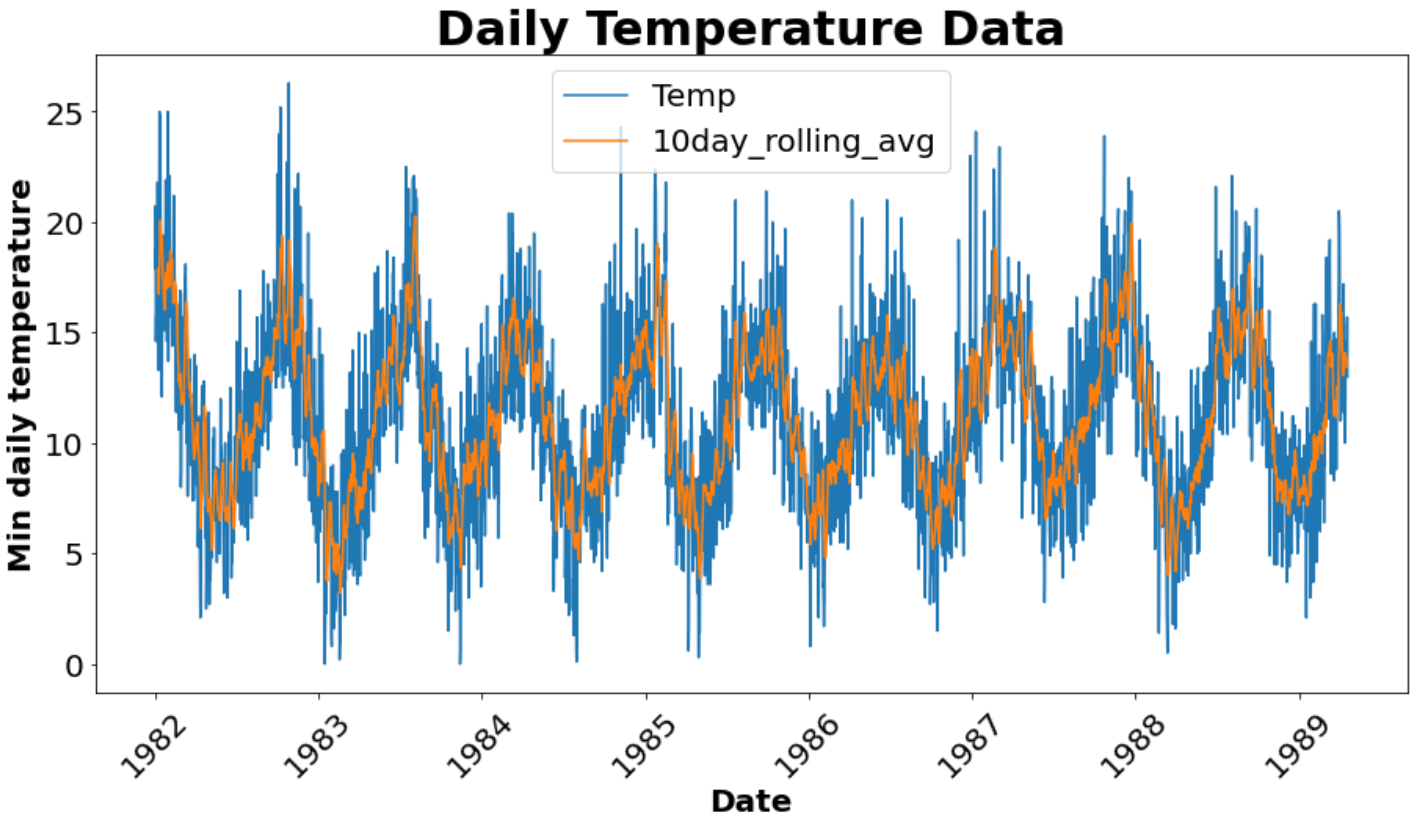
Time Series Plots



Data visualization: trends

Time Series Plots

- moving average and other smoothing functions can be drawn on top of the original time series to perceive trends



Data visualization: remarks

Data visualization must accurately convey the data

- Color scales
  - Distinguish groups of data from each other
  - Represent data values
  - Highlight data/information
- Right context with appropriate
  - title
  - axis labels
  - legends
  - other annotations

Data visualization in Python

Matplotlib

https://matplotlib.org/

Seaborn

https://seaborn.pydata.org/

Pandas.DataFrame.plot

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.html

Homework…

- Assignment I (see eLearning)
  - Data Understanding:
    - Hands on: Visualization

Contents

- Attributes and Datasets
- Data Summarization
- Data Visualization
- Proximity Measures
- Summary

Proximity Measures

Similarity: a numerical measure indicating how attribute/object/set are alike

Dissimilarity: a numerical measure indicating how attribute/object/set are different

- **Proximity** usually refers to similarity or dissimilarity

Applied to:

- Attributes/variables: Given the values for i^{th} attribute compare to values j^{th} attribute
- Objects: Given two data points x_i and x_j, the measure deals with the two points (i, j)
- Sets: Given two data groups $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots\}$ the measure addresses $\mathbf{X}$ and $\mathbf{Y}$ or characteristics of the sets

Proximity Measures

- **Similarity measure** or **similarity function**
  - Numerical measure that quantifies the similarity between attribute/object/set
  - Indicates how two attribute/object/set are alike
    - The higher value, the more alike
  - Often falls in the range [0,1]: 0 --> no similarity; 1: completely similar

Proximity Measures

- **Dissimilarity** (or **distance**) measure
  - Numerical measure that quantifies the difference between attribute/object/set
  - In some sense, the inverse of similarity:
    - The lower value, the more alike
  - Minimum dissimilarity is often 0: completely similar
  - Range [0, 1] or [0, ∞) , depending on the definition

Proximity Measures

Similarity and **dissimilarity** between two objects, ***x*** and ***y***

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

- <u>Nominal</u>: **Female** and **Male**
  - Dissimilarity = d = 1 , Similarity = s = 0
- <u>Ordinal</u>: **Good** and **Excellent** (Categories: Good, Very Good and Excellent)
  - Dissimilarity = d = |2 – 0|/2 = 1 , Similarity = s = 1 – d = 0

Distances (Dissimilarity): properties

A distance (dissimilarity) **d** has the following properties:

1. Non-negativity

   $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all ***x*** and ***y***

2. Identity

   $d(\mathbf{x}, \mathbf{y}) = 0$ iff $\boldsymbol{x = y}$

3. Symmetry

   $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all **x** and **y**

4. Triangle Inequality

   $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all **x**, **y**, and **z**

A distance that satisfies these properties is a **metric**

Similarities: properties

A similarity **s** has the following properties:

1. Maximum similarity

$$s(x, y) = 1 \text{ iff } x = y$$

2. Symmetry

$$d(x, y) = d(y, x) \text{ for all } x \text{ and } y$$

Data matrix and proximity matrix

- Data matrix
  - A data matrix of *n* data points with *D* dimensions
- Proximity matrix
  - *n x n* matrix
  - *n* data points, but registers only the proximity *p(i, j)* (typically metric)
  - Usually symmetric, thus a triangular **matrix**

$$Data = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{pmatrix} \qquad Prox = \begin{pmatrix} 0/1 & & & \\ p(2,1) & 0/1 & & \\ \vdots & \vdots & \ddots & \\ p(n,1) & p(n,2) & \ldots 0/1 & \end{pmatrix}$$

0 for a dissimilarity measure
1 for a similarity measure

Proximity measures: examples

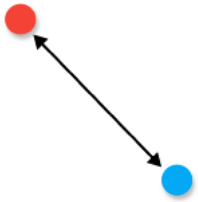
- Given the numerical vectors **x** and **y**

	Measure	Remarks
Euclidean	$d(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{D} (\boldsymbol{x_i} - \boldsymbol{y_i})^2}$	
Manhattan	$d(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{D} \lvert \boldsymbol{x_i} - \boldsymbol{y_i} \rvert$	Dissimilarity x = y → d = 0
Chebyshev	$d(\mathrm{x},\mathrm{y}) = \max_i \lvert \boldsymbol{x_i} - \boldsymbol{y_i} \rvert$	
Cosine	$s(\mathrm{x},\mathrm{y}) = \frac{\sum_{i=1}^{D} \boldsymbol{x_i}\boldsymbol{y_i}}{\sqrt{\left(\sum_{i=1}^{D} \boldsymbol{x_i}^2\right)\left(\sum_{i=1}^{D} \boldsymbol{y_i}^2\right)}}$	Similarity x = y → s = 1

Proximity measures: examples

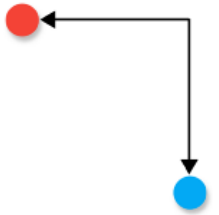
Euclidean distance (L_2 norm)

$$d(x,y) = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2}$$



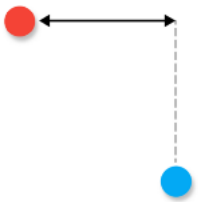
Manhattan distance (L_1 norm)

$$d(x,y) = \sum_{i=1}^{D} |x_i - y_i|$$



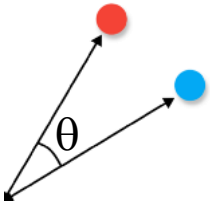
Chebyshev distance

$$d(x,y) = max_i\ |x_i - y_i|$$



Cosine similarity

$$s(x,y) = \frac{x^T y}{\|x\|\|y\|} = cos(\theta)$$



$\|\cdot\|$ is the length of the vector

Euclidean distance

Euclidean distance

$$d(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{D} (\boldsymbol{x_i} - \boldsymbol{y_i})^2}$$

where

- $\boldsymbol{D}$ is the number of dimensions (attributes)
- $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ are, respectively, the entries of the $\boldsymbol{i^{\text{th}}}$ attribute of observations **x** and **y**

Remark:

- Standardization is necessary, if scales differ

Minkowski distance

Minkowski distance is a generalization of **Euclidean distance**

$$d(\mathrm{x},\mathrm{y}) = \left(\sum_{i=1}^{D} |\boldsymbol{x_i} - \boldsymbol{y_i}|^{\boldsymbol{r}} \right)^{1/\boldsymbol{r}}$$

where

- $\boldsymbol{r}$ is a parameter
- $\boldsymbol{D}$ is the number of dimensions (attributes)
- $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ are, respectively, the entries of the $\boldsymbol{i^{th}}$ attribute of observations **x** and **y**

Minkowski Distance: special cases

- $r = 1$: Manhattan distance (City block , taxicab, L_1 norm)

$$d(i,j) = |x_{i1} - y_{i1}| + |x_{i2} - y_{i2}| + \cdots + |x_{iD} - y_{iD}|$$

  - A common example of this for binary vectors is the Hamming distance, which is just the number of bits that are different between two binary vectors

- $r = 2$: Euclidean distance

$$d(i,j) = \sqrt{|x_{i1} - y_{i1}|^2 + |x_{i2} - y_{i2}|^2 + \cdots + |x_{iD} - y_{iD}|^2}$$

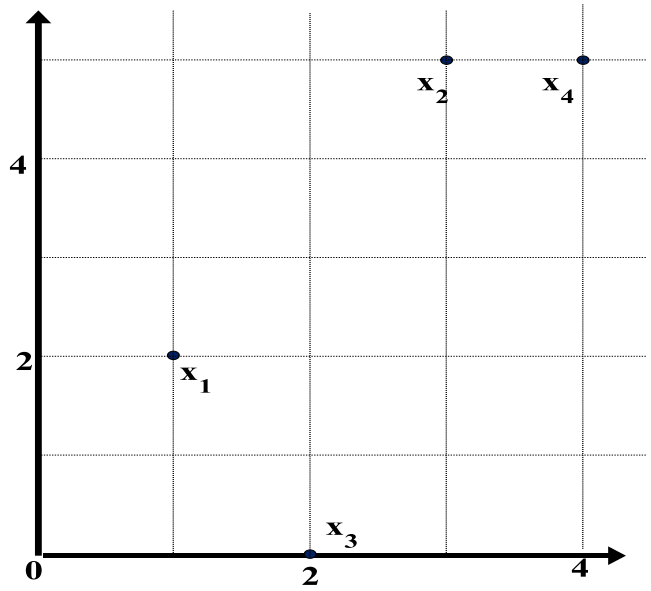
- $r \to \infty$: "supremum" distance (L_{max} norm, L_∞ norm)

$$d(i,j) = \lim_{\boldsymbol{r}\to\infty} \sqrt[r]{|x_{i1} - y_{i1}|^{\boldsymbol{r}} + |x_{i2} - y_{i2}|^{\boldsymbol{r}} + \cdots + |x_{iD} - y_{iD}|^{\boldsymbol{r}}}$$

  - This is the maximum difference between any component of the vectors

Data matrix and distance matrix: examples

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L_1	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L_2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Proximity measures: binary data

Given **two objects** (vectors) with Binary entries

- Compute the contingency table
  - f_{11} = number of attributes where x was 1 and y was 1 (matching)
  - f_{00} = number of attributes where x was 0 and y was 0 (matching)
  - f_{10} = number of attributes where x was 1 and y was 0 (not matching)
  - f_{01} = number of attributes where x was 0 and y was 1 (not matching)

		Object y		
		1	0	sum
Object x	1	f_{11}	f_{10}	$f_{11}+f_{10}$
	0	f_{01}	f_{00}	$f_{01}+f_{00}$
	sum	$f_{11}+f_{01}$	$f_{10}+f_{00}$	

Proximity measures: binary data

Object x \ Object y	1	0	sum
1	f_{11}	f_{10}	$f_{11}+f_{10}$
0	f_{01}	f_{00}	$f_{01}+f_{00}$
sum	$f_{11}+f_{01}$	$f_{10}+f_{00}$	

Simple matching coefficient

- $SMC = \frac{number\ of\ matches}{number\ of\ matches + number\ of\ no{-}matches} = \frac{f_{11}+f_{00}}{f_{11}+f_{00}+f_{10}+f_{01}}$

Jaccard coefficient

(useful for asymmetric binary attributes)

- $J = \frac{number\ of\ 1{-}1\ matches}{number\ of\ non\ 0{-}0\ matches} = \frac{f_{11}}{f_{11}+f_{10}+f_{01}}$

Hamming distance

(useful for asymmetric binary attributes)

- $SMC = \frac{number\ of no{-}matches}{number\ of\ matches + number\ of\ no{-}matches} = \frac{f_{10}+f_{01}}{f_{11}+f_{00}+f_{10}+f_{01}}$

Proximity measures: binary data

Example:

x = [1 0 0 0 0 0 0 0 0 0]

y = [0 0 0 0 0 0 1 0 0 1]

- $f_{11} = 0$
- $f_{00} = 7$
- $f_{10} = 1$
- $f_{01} = 2$

Object x \ Object y	1	0	sum
1	0	1	1
0	2	7	9
sum	2	8	

- Simple matching coefficient: $SMC = \frac{f_{11}+f_{00}}{f_{11}+f_{00}+f_{10}+f_{01}} = \frac{7}{10} = 0.7$
- Jaccard coefficient: $J = \frac{f_{11}}{f_{11}+f_{10}+f_{01}} = \frac{0}{3} = 0$
- Hamming distance: $SMC = \frac{f_{10}+f_{01}}{f_{11}+f_{00}+f_{10}+f_{01}} = \frac{1+2}{10} = 0.3$

Proximity measures: example
Dissimilarity between asymmetric binary variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Mary

Jack	1	0	Σ_{row}
1	2	0	2
0	1	3	4
Σ_{col}	3	3	6

- Gender is a symmetric attribute (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0

Jim

Jack	1	0	Σ_{row}
1	1	1	2
0	1	3	4
Σ_{col}	2	4	6

- Distance: $DABV = \frac{f_{10}+f_{01}}{f_{11}+f_{01}+f_{10}}$
- $d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$
- $d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$
- $d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$

Mary

Jim	1	0	Σ_{row}
1	1	1	2
0	2	2	4
Σ_{col}	3	3	6

Proximity measures and attributes types

Nominal

- Hamming distance (normalized)
  - Example: color, profession

Proximity measures and attributes types

Ordinal: Can be treated as interval-scaled

- Replace an *ordinal variable value* by its rank: $r_{if} \in \{1, \ldots, M_f\}$
- Map the range of each variable into [0, 1] by replacing $\boldsymbol{i^{th}}$ object in the $\boldsymbol{f^{th}}$ variable by $z_{if} = \frac{r_{if}-1}{M_f-1}$
- Compute the dissimilarity using methods for interval-scaled variables
  - **Euclidian** distance, **Manhattan** (city block), **cosine** similarity, etc.

Example

- Acceptable -> 0; Good -> 1/3; Very good -> 2/3; Excellent -> 1

Then distance: **d(Acceptable, Excellent) = 1**, **d(Very good , Excellent) = 1/3**

Proximity measures and attributes types

Mixed type

- A data set may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- Global proximity measure $\boldsymbol{d(i,j)}$
  - a combination of measures applied individually to each attribute

$$d(i,j) = \frac{\sum_{f=1}^{D} w_f \, d_f(i,j)}{\sum_{f=1}^{D} w_f}$$

where

- $\boldsymbol{w_f}$ is a weight for the $\boldsymbol{f^{\text{th}}}$ attribute
- to control missing values on the data ($\boldsymbol{w_f = 0}$, if the $\boldsymbol{f^{\text{th}}}$ attribute is missing either in $\boldsymbol{i}$ or $\boldsymbol{j}$ objects)

Proximity measures and attributes types

Mixed type

- A data set may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- Global proximity measure $\boldsymbol{d(i,j)}$
  - a combination of measures applied individually to each attribute

$$d(i,j) = \frac{\sum_{f=1}^{D} w_f \, d_f(i,j)}{\sum_{f=1}^{D} w_f}$$

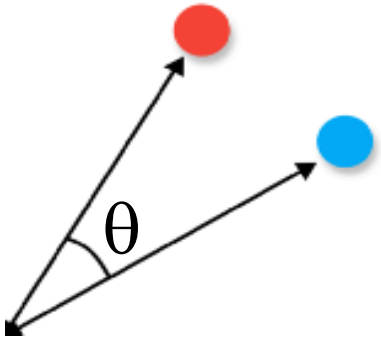
  - $\boldsymbol{w_f}$ is a weight for the $\boldsymbol{f^{\text{th}}}$ attribute
  - to control missing values on the data ($\boldsymbol{w_f = 0}$, if the $\boldsymbol{f^{\text{th}}}$ attribute is missing either in $\boldsymbol{i}$ or $\boldsymbol{j}$ objects)

Cosine similarity of two vectors

Cosine measure

- If d_1 and d_2 are two vectors, then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{||d_1|| \times ||d_2||} = \cos(\theta)$$



where

- $\bullet$ indicates vector dot product

$||d||$ the length of vector d

Cosine similarity of two vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other example: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.

Cosine similarity of two vectors: example

Find the **similarity** between documents 1 and 2.

d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

1. calculate vector dot product

$d_1 \bullet d_2 = 5x3 + 0x0 + 3x2 + 0x0 + 2x1 + 0x1 + 0x1 + 2x1 + 0x0 + 0x1 = 25$

2. calculate $||d_1||$ and $||d_2||$

$$||d_1|| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$||d_2|| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

3. Calculate cosine similarity

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{||d_1|| \times ||d_2||} = 25/(6.481 \times 4.12) = 0.94$$

Proximity measures: issues in calculation

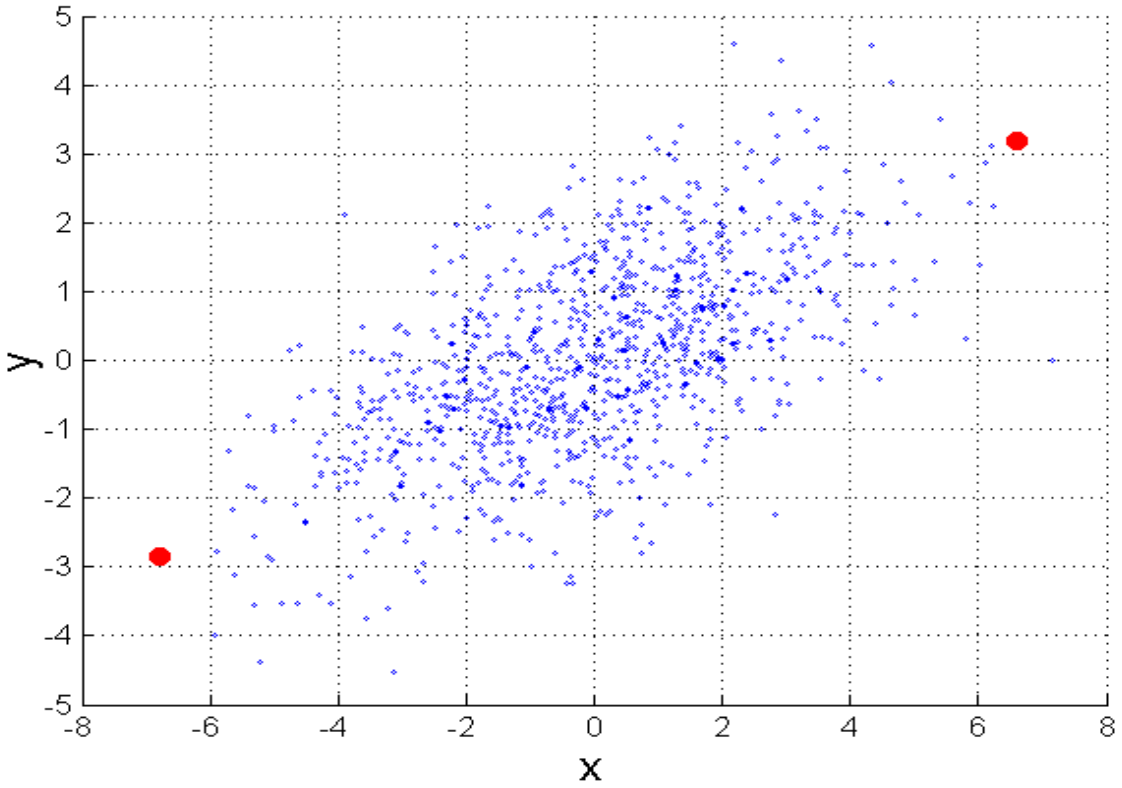
How to deal with

- features having different ranges
  - normalization before calculation (advisable Euclidean distance)
- correlated features
  - Mahalanobis distance is better than Euclidean
- features of different types (quantitative and qualitative)

Proximity measures: correlated features

Mahalanobis distance

$$d(\mathrm{x},\mathrm{y}) = \left((\mathrm{x}-\mathrm{y})^T \Sigma^{-1} (\mathrm{x}-\mathrm{y})\right)^{-1/2}$$, where Σ is the covariance matrix

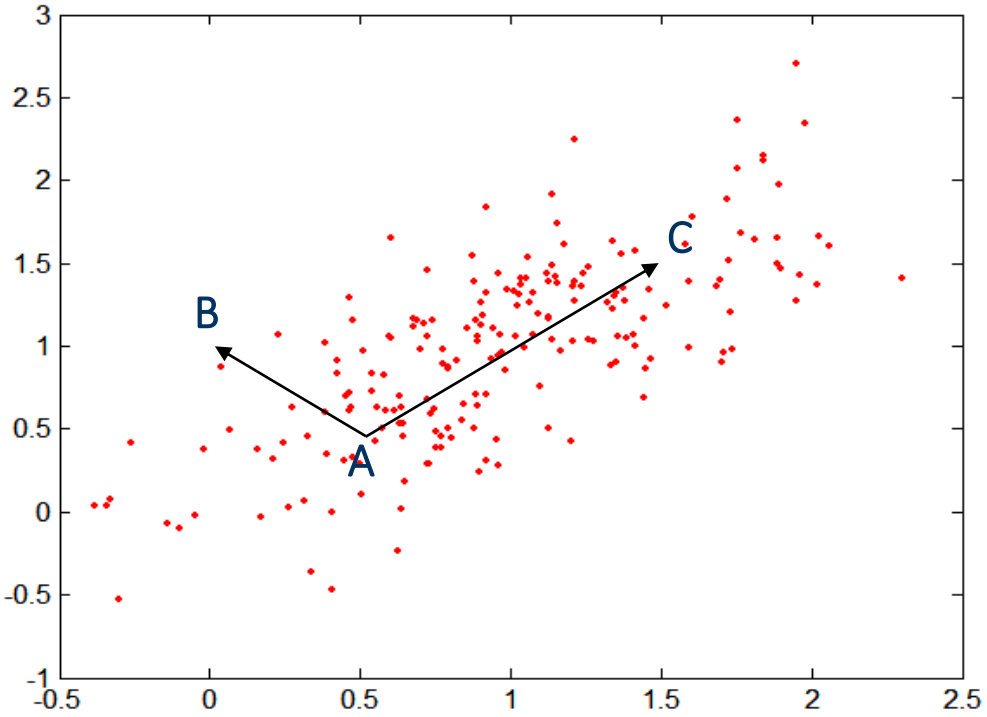


- For red points:
  - Euclidean distance is 14.7,
  - Mahalanobis distance is 6.

Proximity measures: correlated features

Dispersion (spread)

- Covariance matrix $\Sigma = \begin{bmatrix} 0.58 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$



A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A, B) = 5

Mahal(A, C) = 4

Correlation vs Cosine vs Euclidean Distance

Compare the three proximity measures according to their behavior under variable transformation

- scaling: multiplication by a value
- translation: adding a constant

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

- Consider the example
  - $\mathbf{x} = (1, 2, 4, 3, 0, 0, 0)$, $\mathbf{y} = (1, 2, 3, 4, 0, 0, 0)$
  - $y_s = y * 2$ (scaled version of y), $y_t = y + 5$ (translated version)

Measure	(x , y)	(x , y_s)	(x , y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

Homework…

- Assignment I (see eLearning)
  - Notes with exercises: Ex. 1.3 and 1.4

Contents

- Attributes and Datasets
- Data Summarization
- Data Visualization: Amounts, Distributions, Associations, Trends
- Proximity Measures
- Summary

Summary

- Types and scales of attributes
  - nominal, ordinal, interval-scaled, ratio-scaled
- Many types of data sets
- Gain insight into the data by:
  - Basic statistical data description: frequency, central tendency, dispersion / spread
  - Data visualization: map data onto graphical primitives
  - Measure data similarity

Data Understanding - the beginning of Data Preparation/Preprocessing

Bibliography

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, *Pearson*, 2019 (chap 2.1 & 2.4)

Data Mining, the Textbook, Charu C. Aggarwal, *Springer*, 2015 (chap 1.3 & chap 2)

Fundamentals of Data Visualization, Claus O. Wilke, *O'Reilly*, 2022

https://www.python-graph-gallery.com/parallel-coordinate-plot-plotly

https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa