Data understanding

Lecturer: Assoc.Prof. Nguyễn Phương Thái

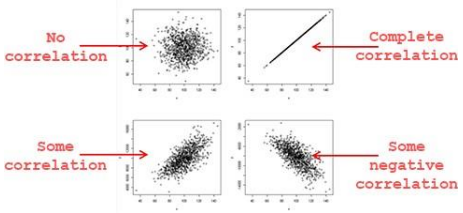
VNU University of Engineering and Technology
Slide: *from Assoc.Prof. Phan Xuân Hiếu, Updated: February 1, 2024*

Outline

Outline

Play with data first

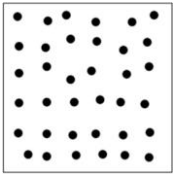
- Understand the nature and characteristics of data first.
  - Data types
  - Noisy, incomplete, missing, imbalanced . . . data
  - Data summarization (descriptive statistics)
  - Data shape and distribution
  - Correlation in data
  - Other explicit and implicit dependencies in data



- Machine learning and deep learning models later!

Do things right or do the right things?

- Is there a strong correlation between the inputs and the target?
  - Displaying ads based on preferences? why "yes" and why "no"?
  - Predicting users' gender based on reading behaviors?



- Doing data clustering for any given dataset?
- Doing topic analysis for very sparse data (lack of co-occurrences)?
- Mining frequent patterns from a database having several (very) popular items?

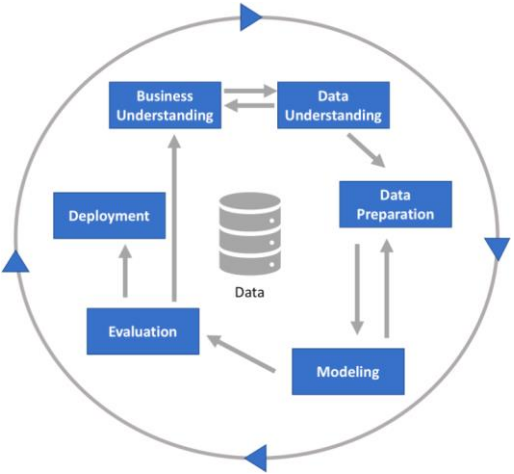
Cross–industry standard process for data mining (CRISP–DM)

- The **cr**oss–**i**ndustry **s**tandard **p**rocess for **d**ata **m**ining (CRISP–DM) is a process model that serves as the base for a data science process. It has six sequential phases:

  1. Business understanding
  2. **Data understanding**
  3. Data preparation
  4. Modeling
  5. Evaluation
  6. Deployment

- Published in 1999 to standardize data mining processes across industries, it has since become the most common methodology for data mining, analytics, and data science projects.

CRISP–DM process



CRISP–DM phase 1: Business understanding

This phase focuses on understanding the objectives and requirements of the project. It includes four tasks:

1. **Determine business objectives**:

   thoroughly understand, from a business perspective, what the customer/company really wants to accomplish, and then define business success criteria.

2. **Assess situation**:

   determine resources availability, project requirements, assess risks and contingencies, and conduct a cost–benefit analysis.

3. **Determine data mining goals**:

   in addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.

4. **Produce project plan**:

   select technologies and tools and define detailed plans for each project phase.

CRISP–DM phase 2: Data understanding

This phase drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase also has four tasks:

1. **Collect initial data**:

   acquire the necessary data and (if necessary) load it into your analysis tool.

2. **Describe data**:

   examine the data and document its surface properties like data format, number of records, or field identities.

3. **Explore data**:

   dig deeper into the data. Query it, visualize it, and identify relationships among the data.

4. **Verify data quality**:

   how clean/dirty is the data? Document any quality issues.

CRISP–DM phase 3: Data preparation

This phase prepares the final data for modeling. It has five tasks:

1. **Select data**:

   determine which data sets will be used and document reasons for inclusion/exclusion.

2. **Clean data**:

   often this is the lengthiest task. Without it, you will likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.

3. **Construct data**:

   derive new attributes that will be helpful. For example, derive someone's body mass index from height and weight fields.

4. **Integrate data**:

   create new data sets by combining data from multiple sources.

5. **Format data**:

   re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

Outline

Outline

Data attribute

- An *attribute* is a data field, representing a characteristic or feature of a data object.
- Often called:
  - *Field*, *column*, or *attribute* in database.
  - *Dimension* in data warehousing.
  - *Variable* in statistics.
  - *Feature*, *dimension* in machine learning.
- Example: attributes describing a student object can include:

  `student` – `id`, `name`, `gender`, `yob`, `phone` – `number`, `address`, etc.
- Observed values for a given attribute are known as *observations*.

Attribute data types

- Nominal or categorical attributes
- Binary attributes
- Ordinal attributes
- Numeric attributes
  - Interval–scaled attributes
  - Ratio–scaled attributes
  - Discrete vs. continuous attributes

Nominal or categorical attributes

- The values of a **nominal attribute** are symbols or names of things.
- Each value represents some kind of category, code, or state. Thus nominal attributes are also referred to as **categorical**.
- The values do not have any meaningful order. Math operations are not meaningful.
- Examples:
  - `Occupation`: *teacher*, *dentist*, *programmer*, *farmer*, etc.
  - `HairColor`: *black*, *brown*, *blond*, *red*, *gray*, *white*, etc.
  - `CustomerID`: values are numeric or alpha–numeric, but can be seen as strings.
- From the view of descriptive statistics:
  - No sense to find the *mean* (average) or *median* (middle) values.
  - The *mode* (the most commonly occurring value) may exist.

Binary attributes

- A **binary attribute** is a nominal attribute with only two categories/states: 0 or 1:
  - 0 typically means that the attribute is **absent**, and 1 means that it is **present**.
- It is also referred to as **boolean** if the two states correspond to *true* and *false*.
- Examples:
  - `IsStudent`: *no* or *yes*; *false* or *true*; 0 or 1.
  - `CovidTest`: *negative* or *positive*.
- A binary attribute is **symmetric** if both of its states are equally valuable and carry the same weight. One example is the `gender` with two states *male* and *female*.
- A binary attribute is **asymmetric** if the two states are **not equally important**.
  - Example: the *positive* and *negative* outcomes of a medical test for HIV.
  - By convention, we code the **most important** outcome, which is usually the **rarest one**, by 1 (e.g., *HIV positive*) and the other by 0 (e.g., *HIV negative*).

Ordinal attributes

- An **ordinal attribute**:
  - Has meaningful order or ranking among its values.
  - But the magnitude between successive values is not known.
- Examples:
  - Size of drinks: *small*, *medium*, *large*.
  - Grade: $A+$, A, $B+$, B, etc.
  - Rating: 1 (*very dissatisfied*), 2 (*dissatisfied*), 3 (*neutral*), 4 (*satisfied*), and 5 (*very satisfied*).
- Ordinal attributes may also be obtained from the **discretization** of numeric quantities by splitting the value range into a finite number of ordered categories.
- From the view of descriptive statistics:
  - The *mode* and *median* values are existing.
  - But the *mean* cannot be defined.

Numeric attributes

- A **numeric attribute**
  - is a **quantitative** and **measurable** quantity, represented in integer or real values.
- Numeric attributes can be:
  - Interval-scaled
  - Ratio-scaled

Interval–scaled attributes

- **Interval–scaled attributes** are measured on a scale of **equal–size** units.
- The values have **order** and can be positive, 0, or negative.
- This attribute type allows to compare and quantify the difference between values.
- Examples:
  - `Temperature`: 15oC, 18oC, 20oC, 25oC, 30oC, 37oC, etc.
  - `CalendarYear`: 1010, 1945, 1954, 2010, 2020, 2022, etc.
- Temperature has no **true** zero–point (could not say 30oC is twice as warm as 15oC)
- From the view of descriptive statistics:

  Can compute the *mean*, *median*, and *mode* values.

Ratio–scaled attributes

- A **ratio–scaled attribute** is a numeric attribute with an **inherent zero–point**.
- If a measurement is ratio–scaled: a value is a multiple (or ratio) of another value.
- The is a meaningful order or ranking among values.
- Examples:
  - `Salary`: US$2000 is twice as much as US$1000.
  - Other examples: `Weight`, `Height`, `Longitude`, `Latitude`, etc.
- From the view of descriptive statistics:

  Can compute the *mean*, *median*, and *mode* values.

Discrete versus continuous attributes

- Mining algorithms often talk of attributes as being either *discrete* or *continuous*.
- Each type may be processed differently.
- A **discrete** attribute has a finite or countably infinite set of values:
  - May or may not be represented as integers.
  - `HairColor`, `Gender`, `DrinkSize` have a finite number of values, and so are discrete.
  - May have numeric values: 0 and 1 for binary attributes; 0 to 150 for the attribute `Age`.
- An attribute is *countably infinite* if the set of possible values is infinite but the values can be put in a one-to-one correspondence with natural numbers. For example, the attribute `CustomerID` is countably infinite.

Discrete versus continuous attributes (cont'd)

- If an attribute is not discrete, it is *continuous*.
- The terms *numeric attribute* and *continuous attribute* are often used interchangeably in the literature.
- This can be confusing because, in the classic sense, continuous values are real numbers, whereas numeric values can be either integers or real numbers.

Outline

Univariate data

- Univariate data is a type of data which consists observations from only one variable (X), i.e., a single characteristic or attribute:

$$\mathcal{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{1}$$

- Univariate data types: nominal/categorical (binary, finite, infinite) and numerical (discrete, continuous).

Univariate data (cont'd)

- Examples:
  - The Covid-19 test results of a sample of 12 persons:
    D = {*neg*, *pos*, *neg*, *neg*, *neg*, *pos*, *neg*, *neg*, *pos*, *neg*, *neg*, *neg*}
  - Marks of a student for the last 15 courses:
    D = {*A*, *A+*, *B*, *B*, *A*, *B+*, *A+*, *A*, *B*, *C*, *C*, *A*, *B+*, *B*, *A*}
  - Messi's goals for Barcelona from 2004–2005 to 2020–2021:
    D = {1, 8, 17, 16, 38, 47, 53, 73, 60, 41, 58, 41, 54, 45, 51, 31, 38}
  - The average high temperature in Hanoi for 12 months (oC):
    D = {19.7, 20.1, 22.9, 27.2, 31.4, 32.9, 33.1, 32.3, 31.2, 28.8, 25.3, 22.0}
- Univariate data analysis involves descriptive statistics like central tendency (mean, median, mode), dispersion (range, variance, quartiles, standard deviation), etc.
- Univariate data visualization can be frequency distribution tables, bar charts, histograms, pie charts, etc.

Bivariate data

- Bivariate data is a type of data which consists observations from two variables (X, Y), i.e., considering two characteristics or attributes at the same time:

$$\mathcal{D} = \begin{pmatrix} X & Y \\ \hline x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} \tag{2}$$

- Like univariate data, X and Y can be either categorical or numerical.

Bivariate data examples

Ad spend	Revenue
$14,500	$59,000
$19,000	$64,000
$22,400	$89,000
$28,900	$86,000
$30,000	$94,000
$32,000	$104,000
$29,000	$89,000
$28,000	$82,000
$32,000	$88,000
$35,000	$103,000
$29,000	$94,000
$38,000	$140,000

The advertising spend and the total revenue of 12 consecutive quarters

Season	Apps	Goals
2004-05	9	1
2005-06	25	8
2006-07	36	17
2007-08	40	16
2008-09	51	38
2009-10	53	47
2010-11	55	53
2011-12	60	73
2012-13	50	60
2013-14	46	41
2014-15	57	58
2015-16	49	41
2016-17	52	54
2017-18	54	45
2018-19	50	51
2019-20	44	31
2020-21	47	38

Messi's #apps, #goals for Barcelona

Hours	GPA
6	3.2
9	3.8
10	3.9
13	3.6
6	2.5
5	2.7
7	2.9
19	3.9
20	3.8
14	3.4
10	3.1
7	2.6
4	2.2
8	2.9
4	3.2

The number of hours studied per week and the GPA

Bivariate data analysis

- Determining the potential relationship (correlation, association, dependency, causality, etc.) between two variables X and Y.
- Regression (and classification):
  - Predicting or forecasting a value for one variable (e.g., Y –the dependent variable) if we know the value of the other (X –the independent variable); or
  - Inferring the causal relationships between the dependent and the independent variables.
- Bivariate analysis can also be descriptive (central tendency, variability) or inferential (e.g., inferring properties of an underlying distribution, properties of a population, etc.)
- Bivariate visualization with two–way frequency table, two–way proportion table, stacked bar chart, side–by–side bar chart, scatter plot, dot plot, grouped boxplots, grouped histograms, heatmap, etc.

Multivariate data

- Multivariate data is a type of data which consists observations from d variables/attributes $(X_1, X_2, \ldots, X_d)$, i.e., considering d characteristics or attributes at the same time. This is actually a $n \times d$ matrix as follows:

$$\mathcal{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \tag{3}$$

  In the row view, the data can be considered as a set of n points or vectors in the d-dimensional attribute space:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$$

- Like univariate and bivariate data, $X_1, X_2, \ldots, X_d$ can be either categorical or numerical.

Multivariate data analysis

- Multivariate mean vector:

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix} \tag{4}$$

- Covariance matrix:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix} \tag{5}$$

- Classification, regression, clustering, association mining, etc.
- Visualization: a lot of ways and tools.

IID data

- IID data mean the all the observations (data points) in the a data sample are independent and identically (distributed) drawn from variable(s).
- Univariate, bivariate, and multivariate data described above are assumed to be IID.
  - Univariate data: x_i $(i = 1..n)$ are i.i.d from variable/attribute X.
  - Bivariate data: (x_i, y_i) $(i = 1..n)$ are i.i.d from two variables/attributes (X, Y).
  - Multivariate data: $(x_{i1}, x_{i2}, \ldots, x_{id})$ $(i = 1..n)$ are i.i.d from d variables/attributes $(X_1, X_2, \ldots, X_d)$.
- However, i.i.d is a strong assumption. In reality, data points are not completely independent. There are some explicit or implicit dependencies among them.
- The dependencies among data points can be temporal (time), spatial (e.g., position, neighbors, etc.), structural (e.g., edges, links, etc.), referential, etc.
- Data with dependencies can be called *dependency–oriented data* or *complex and structured data*.

Outline

Complex and structured data

- Time–series data
- Discrete sequences and strings
- Spatial data
- Spatiotemporal data
- Network and graph data
- Other forms of data

Time–series data

- Time–series data contain values that are typically generated by continuous measurement over time.
- For example, an environmental sensor will measure the temperature continuously, whereas an electrocardiogram (ECG) will measure the parameters of a subject's heart rhythm.
- Such data typically have implicit dependencies built into the values received over time. For example, the adjacent values recorded by a temperature sensor will usually vary smoothly over time, and this factor needs to be explicitly used in the data mining process.
- The nature of the temporal dependency may vary significantly with the application. For example, some forms of sensor readings may show periodic patterns of the measured attribute over time.

Time–series data (cont'd)

- Attributes in time–series data are classified into two types:
  - *Contextual attributes*: These are the attributes that define the context on the basis of which the implicit dependencies occur in the data. For example, in the case of sensor data, the time stamp at which the reading is measured may be considered the contextual attribute. Sometimes, the time stamp is not explicitly used, but a position index is used.
  - *Behavioral attributes*: These represent the values that are measured in a particular context. In the sensor example, the temperature is the behavioral attribute value. It is possible to have more than one behavioral attribute. For example, if multiple sensors record readings at synchronized time stamps, then it results in a multidimensional time–series data set.
- The contextual attributes typically have a strong impact on the dependencies between the behavioral attribute values in the data.

Time–series data definition

Multivariate time–series data [2]:

A time–series of length n and dimensionality d contains d numeric features at each of n time stamps $t_1, t_2, \ldots, t_n$. Each timestamp contains a component for each of the d series. Therefore, the set of values received at timestamp t_i is $\bar{Y}_i = (y_i^1, y_i^2, \ldots, y_i^d)$. The value of the j^{th} series at timestamp t_i is y_i^j. The time–series data are $\{\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n\}$.

- For example, consider the case where two sensors at a particular location monitor the temperature and pressure every second for a minute. This corresponds to a multidimensional series with $d = 2$ and $n = 60$. In some cases, the timestamps $t_1, t_2, \ldots, t_n$ may be replaced by index values from 1 through n, especially when the timestamp values are equally spaced apart.
- Time-series data are relatively common in many sensor applications, health, weather forecasting, and financial market analysis.

Discrete sequences and strings

- Discrete sequences can be considered the categorical analog of time–series data. As in the case of time–series data, the contextual attribute is a time stamp or a position index in the ordering. The behavioral attribute is a categorical value;
- Therefore, discrete sequence data are defined in a similar way to time–series data.

Multivariate discrete sequence data [2]:

A discrete sequence of length n and dimensionality d contains d discrete feature values at each of n different timestamps $t_1, t_2, \ldots, t_n$. Each of the n components $\bar{Y}_i$ contains d discrete behavioral attributes $(y_i^1, y_i^2, \ldots, y_i^d)$, collected at the i^{th} timestamp.

Discrete sequences and strings (cont'd)

- For example, consider a sequence of Web accesses, in which the Web page address and the originating IP address of the request are collected for 100 different accesses. This represents a discrete sequence of length $n = 100$ and dimensionality $d = 2$.
- A particularly common case in sequence data is the *univariate* scenario, in which the value of d is 1. Such sequence data are also referred to as *strings*.
- In theory, it is possible to have series that are mixed between categorical and numerical data.
- Another important variation is the case where a sequence does not contain categorical attributes, but a set of any number of unordered categorical values. For example, supermarket transactions may contain a sequence of sets of items. Each set may contain any number of items. Such setwise sequences are not really multivariate sequences, but are univariate sequences.
- Thus, discrete sequences can be defined in a wider variety of ways, as compared to time–series data because of the ability to define sets on discrete elements.

Discrete sequences and strings (cont'd)

- In some cases, the contextual attribute may be a position based on physical placement. In such cases, the timestamp may be replaced by an index representing the position of the value in the string, starting at `0` or `1`. Examples:
  - *Event logs*: A wide variety of computer systems, Web servers create event logs on the basis of user activity. An example of an event log is a sequence of user actions at a financial Web site:

    `Login Password Login Password Login Password ...`

    This particular sequence may represent a scenario where a user is attempting to break into a password-protected system, and it may be interesting in anomaly detection.
  - *Biological data*: In this case, the sequences may correspond to strings of nucleotides or amino acids. The ordering of such units provides information about the characteristics of protein function. Therefore, the data mining can be used to determine interesting patterns reflecting different biological properties.
- Discrete sequences are often more challenging for mining algorithms because they do not have the smooth value continuity of time–series data.

Spatial data

- In spatial data, many nonspatial attributes (e.g., temperature, pressure, image pixel color intensity) are measured at spatial locations.
- For example, sea-surface temperatures are often collected by meteorologists to forecast the occurrence of hurricanes. In such cases, the spatial coordinates correspond to contextual attributes, whereas attributes such as the temperature correspond to the behavioral attributes.
- Typically, there are two spatial attributes. As in the case of time–series data, it is also possible to have multiple behavioral attributes. For example, in the sea-surface temperature application, one might also measure other behavioral attributes such as the pressure.

Spatial data (cont'd)

Spatial data [2]:

A d-dimensional spatial data record contains d behavioral attributes and one or more contextual attributes containing the spatial location. Therefore, a d-dimensional spatial data set is a set of d dimensional records $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n$, together with a set of n locations $L_1, L_2, \ldots, L_n$, such that the record $\bar{X}_i$ is associated with the location L_i.

- The aforementioned definition provides broad flexibility in terms of how record $\bar{X}_i$ and location L_i may be defined.
- For example, the behavioral attributes in record $\bar{X}_i$ may be numeric or categorical, or a mixture of the two. In the meteorological application, $\bar{X}_i$ may contain the temperature and pressure attributes at location L_i.
- Furthermore, L_i may be specified in terms of precise spatial coordinates, such as latitude and longitude, or in terms of a logical location, such as the city or state.

Spatial data (cont'd)

- Spatial data mining is closely related to time–series data mining, in that the behavioral attributes in most commonly studied spatial applications are continuous, although some applications may use categorical attributes as well.
- Therefore, value continuity is observed across contiguous spatial locations, just as value continuity is observed across contiguous time stamps in time–series data.

Spatiotemporal data

- A particular form of spatial data is spatiotemporal data, which contains both spatial and temporal attributes.
- The precise nature of the data also depends on which of the attributes are contextual and which are behavioral. Two kinds of spatiotemporal data are most common:
  - *Both spatial and temporal attributes are contextual*: This kind of data can be viewed as a direct generalization of both spatial data and temporal data. This kind of data is particularly useful when the spatial and temporal dynamics of particular behavioral attributes are measured simultaneously. For example, consider the case where the variations in the sea-surface temperature need to be measured over time. In such cases, the temperature is the behavioral attribute, whereas the spatial and temporal attributes are contextual.
  - *The temporal attribute is contextual, whereas the spatial attributes are behavioral*: Strictly speaking, this kind of data can also be considered time–series data. However, the spatial nature of the behavioral attributes also provides better interpretability and more focused analysis in many scenarios. The most common form of this data arises in the context of trajectory analysis.

Network and graph data

- In network and graph data, the data values may correspond to nodes in the network, whereas the relationships among the data values may correspond to the edges in the network.
- In some cases, attributes may be associated with nodes in the network.
- Although it is also possible to associate attributes with edges in the network, it is much less common to do so.

Network data [2]:

A network $G = (N, E)$ contains a set of nodes N and a set of edges E, where the edges in E represent the relationships between the nodes. In some cases, an attribute set $\bar{X}_i$ may be associated with node i, or an attribute set $\bar{Y}_{ij}$ may be associated with edge (i, j).

Network and graph data (cont'd)

- The edge (i, j) may be directed or undirected. For example, the Web graph may contain directed edges corresponding to directions of hyper–links between pages, whereas friendships in Facebook are undirected.
- A second class of graph mining problems is that of a database containing many small graphs such as chemical compounds. The challenges in these two classes of problems are very different. Some examples of data that are represented as graphs:
  - *Web graph*: The nodes correspond to the Web pages, and the edges correspond to hyperlinks. The nodes have text attributes corresponding to the content in the page.
  - *Social networks*: The nodes correspond to social network actors; the edges correspond to friendship links. The nodes may have attributes corresponding to social page content. Some specialized forms of social networks are email or chat-messenger networks, the edges may have content associated with them.
  - *Chemical compound databases*: The nodes correspond to the elements and the edges correspond to the chemical bonds between the elements. The structures in these chemical compounds are very useful for identifying important reactive and pharmacological properties of these compounds.

Network and graph data (cont'd)

- Network data are a very general representation and can be used for solving many similarity–based applications on other data types;
- For example, multidimensional data may be converted to network data by creating a node for each record in the database, and representing similarities between nodes by edges. Such a representation is used quite often for many similarity–based data mining applications, such as clustering.
- It is possible to use community detection algorithms to determine clusters in the network data and then map them back to multidimensional data.

Other forms of data

- *Text data*: can be seen as discrete sequences; each element is a word or token.
- *Natural language data*: can be seen as discrete sequences; each element can be character, token, word, phrase, sentence, paragraph.
- *Speech data*: can be seen as discrete sequences or time–series data.
- *Image data*: can be seen as spatial data.
- *Video data*: can be seen as discrete sequences or time–series data where each element is a frame.

Outline

Outline

Measuring the central tendency of data

- Suppose $\mathsf{D} = \{x_1, x_2, \ldots, x_n\}$ is a data sample (i.e., univariate data) consisting of n observations of a variable or attribute X.
- If we were to plot the observations in D, where would most of the values fall? This gives us an idea of the central tendency of the data.
- Measures of central tendency include:
  - Mean
  - Median
  - Mode
  - Midrange

Mean

- The most common and effective numeric measure of the "center" of a set of data is the (arithmetic) *mean*, i.e., the average.
- The **mean** of $D = \{x_1, x_2, \ldots, x_n\}$, denoted $\bar{x}$, is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{6}$$

- Examples:
  - Messi's goals for Barcelona from 2004–2005 to 2020–2021:
    $D = \{1, 8, 17, 16, 38, 47, 53, 73, 60, 41, 58, 41, 54, 45, 51, 31, 38\}$

$$\bar{x} = \frac{1 + 8 + 17 + \cdots + 51 + 31 + 38}{17} = \frac{672}{17} = 39.53$$

  - Salary (in \$k) of 12 employees in a company (shown in increasing order):
    $D = \{30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215\}$

$$\bar{x} = \frac{30 + 36 + 47 + \cdots + 70 + 110 + 215}{12} = \frac{840}{12} = 70$$

Weighted mean or weighted average

Sometimes, each value x_i in D may be associated with a weight w_i (for $i = 1..n$). The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, the weighted mean (weighted average) is:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \tag{7}$$

Mean and its issues

- Although the mean is the most well–known quantity for describing a data sample, it is not always the best way of measuring the center of the data.
- A major problem with the mean is its **sensitivity to extreme values** (e.g., **outliers**). Even a small number of extreme values can corrupt the mean.
- For example, the mean salary above may be substantially pushed up by that of a few highly paid managers ($110k, $215k). Similarly, the mean score of a class in an exam could be pulled down quite a bit by a few very low scores.
- To offset the effect caused by a small number of extreme values, we can instead use the **trimmed mean**, which is the mean obtained after chopping off values at the high and low extremes. For example, we can sort the values observed for salary and remove the top and bottom 2% before computing the mean.

Median

- For **skewed** (asymmetric) data, a better measure is the **median**.
- Median is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half.
- In probability and statistics, the median generally applies to numeric data. However, it may apply to ordinal data.
- Given $D = \{x_1, x_2, \ldots, x_n\}$ is a sample from a variable/attribute X, let $D' = \{x'_1, x'_2, \ldots, x'_n\}$ is the list of values after sorting D in increasing order.
- If n is odd, then the median is the middle value of D'. If n is even, then the median is not unique; it is the two middlemost values and any value in between. If X is a numeric attribute in this case, by convention, the median is taken as the average of the two middlemost values.

Median (cont'd)

- In other words, if X is numeric, the median is:

$$median = \frac{x'_a + x'_b}{2} \tag{8}$$

where $a = \lceil \frac{n}{2} \rceil$ (is the smallest integral value greater than or equal to $\frac{n}{2}$) and $b = \lfloor \frac{n}{2} + 1 \rfloor$ (is the largest integral value less than or equal to $\frac{n}{2} + 1$).

- Examples:
  - Messi's goals for Barcelona from 2004–2005 to 2020–2021:
    $D = \{1, 8, 17, 16, 38, 47, 53, 73, 60, 41, 58, 41, 54, 45, 51, 31, 38\}$.
    $D' = \{1, 8, 16, 17, 31, 38, 38, 41, 41, 45, 47, 51, 53, 54, 58, 60, 73\}$

$$median = x'_9 = 41$$

  - Salary (in \$k) of 12 employees in a company (shown in increasing order):
    $D = \{30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215\}$

$$median = \frac{52 + 56}{2} = \frac{108}{2} = 54 \text{ (more reasonable than } \bar{x} = 70)$$

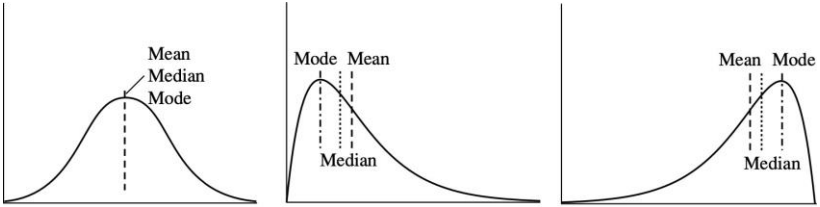
Mode

- The **mode** is another measure of central tendency.
- The mode for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for both qualitative and quantitative attributes.
- It is possible for the greatest frequency to correspond to several different values, which results in more than one mode. Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**. In general, a data set with two or more modes is **multimodal**.
- If each data value occurs only once, then there is no mode.
- Examples:
  - Messi's goals for Barcelona from 2004–2005 to 2020–2021:
    {1, 8, 17, 16, 38, 47, 53, 73, 60, 41, 58, 41, 54, 45, 51, 31, 38}. **mode** = 38 or 41 (bimodal).
  - Salary (in $k) of 12 employees in a company:
    {30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215}. **mode** = 52.
  - Marks of a student for the last 15 courses:
    {A, A+, B, B, A, B+, A+, A, B, C, C, A, B+, B, A}. **mode** = A.

Midrange

- The **midrange** can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set.
- Examples: salary (in $k) of 12 employees in a company:
  D = {30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215}. **midrange** = $\frac{30+215}{2}$ = 122.5.
- The midrange value is even more affected by outlier values than the mean.

Mean, median, and mode for skewed data



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Mean, median, and mode of symmetric versus positively and negatively skewed data [1]

- In a unimodal frequency curve with perfect **symmetric** data distribution, the mean, median, and mode are all at the same center value (figure **a**).
- Data in most real applications are not symmetric. They may instead be either **positively skewed**, where the mode occurs at a value that is smaller than the median (figure **b**), or **negatively skewed**, where the mode occurs at a value greater than the median (figure **c**).

Outline

Measuring the dispersion of data

1. Range
2. Quantiles
3. Quartiles
4. Interquartile range
5. Five–number summary
6. Boxplots
7. Outliers
8. Variance and standard deviation

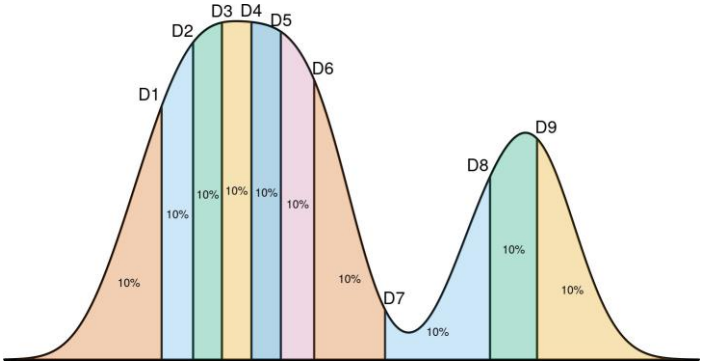
Range

- Let $D = \{x_1, x_2, \ldots, x_n\}$ be univariate data consisting of n observations sampling from a numeric variable/attribute X.
- The **range** of D is the difference between the largest (max()) and smallest (min()) values in S.
- Examples:
  - Messi's goals for Barcelona from 2004–2005 to 2020–2021:
    $D = \{1, 8, 17, 16, 38, 47, 53, 73, 60, 41, 58, 41, 54, 45, 51, 31, 38\}$
    Range = 73 - 1 = 72.
  - Salary (in $k) of 12 employees in a company:
    $D = \{30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215\}$
    Range = 215 - 30 = 185.

Quantiles

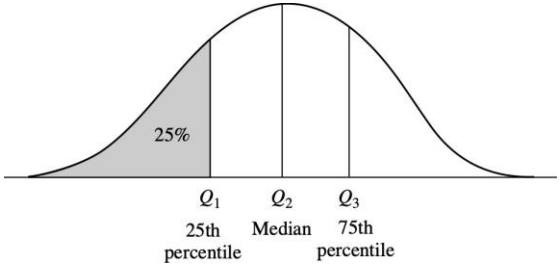
- Suppose that the data points/values in the data sample D are sorted in increasing numeric order.
- Imagine that we can pick certain data points so as to split D into equal-size consecutive sets/parts.
- These data points are called *quantiles*. **Quantiles** are points taken at regular intervals of a data sample, dividing it into essentially equal-size consecutive sets/parts.
- The k^{th} *q-quantile* for a given data sample is the value x such that at most k/q of the data values are less than x and at most $(q-k)/q$ of the data values are more than x, where k is an integer such that $0 < k < q$. There are $q-1$ q-quantiles.
- There are some well-known q-quantiles:
  - **4-quantiles** (Quartiles): 3 quartiles split the data into four parts
  - **10-quantiles** (Deciles): 9 deciles split the data into 10 parts
  - **100-quantiles** (Percentiles): 99 percentiles split the data into 100 parts

10–quantiles or deciles



10–quantiles (also called *deciles*) are 9 values (D_1, D_2, ..., D_9) dividing the data distribution into 10 equal–size consecutive parts [source: Internet]

Quartiles



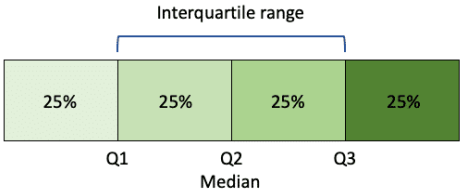
Quartiles divide the distribution into four equal–size consecutive subsets [1]

- Quartiles are three values (Q_1, Q_2, Q_3) dividing the (sample) distribution into four equal–size consecutive parts.
- The first quartile (Q_1) corresponds to the 25^{th} percentile.
- The second quartile (Q_2) corresponds to the 50^{th} percentile, i.e., the median.
- The third quartile (Q_3) corresponds to the 75^{th} percentile.

Interquartile range

- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **interquartile range** (**IQR**) and is defined as

$$IQR = Q_3 - Q_1 \tag{9}$$



- Example:
  - Salary (in $k) of 12 employees in a company (shown in increasing order): D = {30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215}.
  - Q_1 = $47,000 and Q_3 = $62,000.
  - IQR = $62,000 - $47,000 = $15,000.

Five–number summary

- No single numeric measure of spread (e.g., IQR) is very useful for describing skewed distributions.
- In the symmetric distribution, the median (and other measures of central tendency) splits the data into equal–size halves. This does not occur for skewed distributions.
- Therefore, it is more informative to also provide the two quartiles Q_1 and Q_3, along with the median.
- Because Q_1, the median, and Q_3 together contain no information about the end–points (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well.
- This is known as the *five–number summary*. The **five–number summary** of a distribution consists of the median (Q_2), the quartiles Q_1 and Q_3, and the smallest and largest individual observations, written in the order of

$$\{Minimum, Q_1, Median, Q_3, Maximum\}$$

Boxplots (a.k.a box-and-whisker plots)

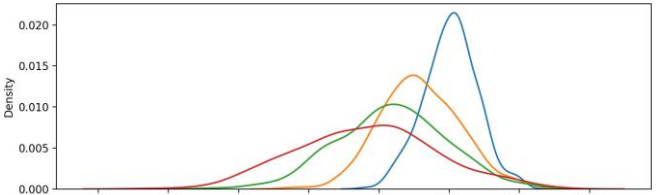
Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five–number summary as follows:

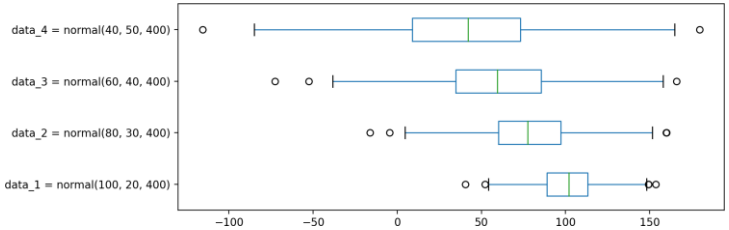
- Typically, the ends of the box are at the quartiles (Q_1 and Q_3) so that the box length is the interquartile range.
- The median is marked by a line within the box.
- Two lines (called *whiskers*) outside the box extend to the smallest ($Minimum$) and largest ($Maximum$) observations/values.

Boxplots with **outliers**:

- When dealing with a moderate number of observations, it is worthwhile to plot potential outliers individually.
- The whiskers are extended to the extreme low and high observations *only if* these values are less than $1.5 \times IQR$ beyond the quartiles.
- Otherwise, the whiskers terminate at the most extreme observations occurring within $1.5 \times IQR$ of the quartiles.

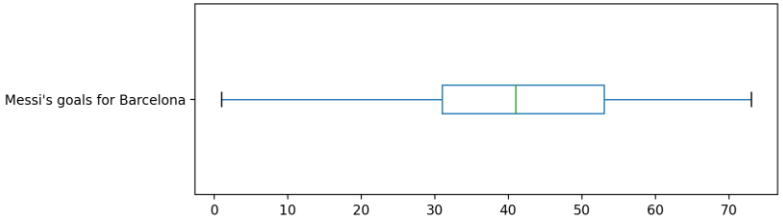
Boxplot examples





Boxplot examples (cont'd)

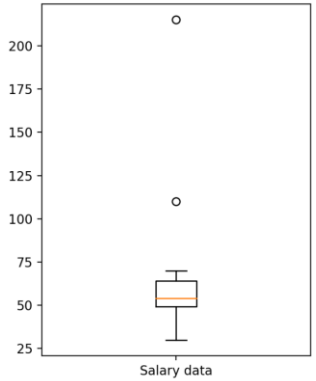
Messi's goals for Barcelona from 2004–2005 to 2020–2021:
D = {1, 8, 17, 16, 38, 47, 53, 73, 60, 41, 58, 41, 54, 45, 51, 31, 38}
D′ = {1, 8, 16, 17, 31, 38, 38, 41, 41, 45, 47, 51, 53, 54, 58, 60, 73}



- $Minimum = 1$
- $Q_1 = 31$
- $Median\ (Q_2) = 41$
- $Q_3 = 53$
- $Maximum = 73$

Boxplot examples (cont'd)

Salary (in \$k) of 12 employees in a company:
$\mathcal{D} = \{30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215\}$



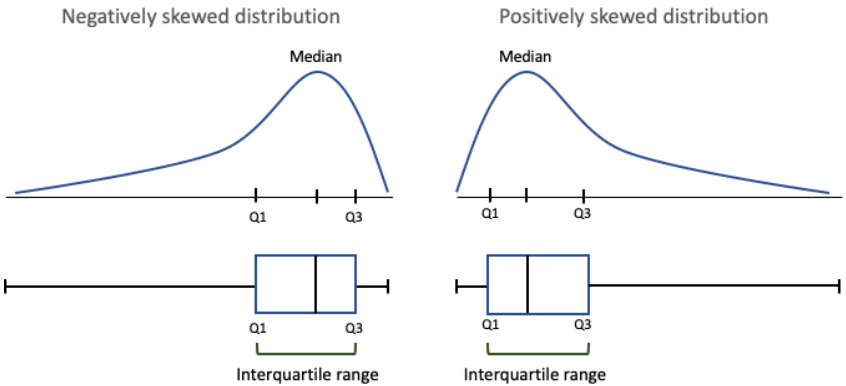
- $Minimum = \$30,000$
- $Q_1 = \$48,500$
- $Median\ (Q_2) = \$54,000$
- $Q_3 = \$66,000$
- $Maximum = \$215,000$
- $IQR = Q_3 - Q_1 = \$17{,}500$
- $1.5 \times IQR = \$26{,}250$
- $Outliers$ are 110 and 215

What is boxplot useful for?

Boxplots are especially useful for showing the central tendency, the data dispersion, the skewness of data distributions, and the outliers:

- The position of the median → the central tendency.
- The box length → the data dispersion.
- The position of the box relative to the both ends (i.e., $Minimum$ and $Maximum$) → the skewness of the distribution.
- The data values too far from Q_1 (i.e., smaller than $Q_1 - 1.5 \times IQR$) and Q_3 (i.e., larger than $Q_3 + 1.5 \times IQR$) → outliers.

How boxplots depict skewed data



Boxplots for skewed distributions [source: Internet]

Variance and standard deviation

- Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is.
- A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.
- Let $\mathsf{D} = \{x_1, x_2, \ldots, x_n\}$ be a sample of n data observations/values from a numeric variable/attribute X. Let $\bar{x}$ be the mean of D. The **variance** (σ^2) of the observations in S is

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 \qquad (10)$$

- The **standard deviation**, σ, of the observations is the square root of the variance.

Examples of variance and standard deviation

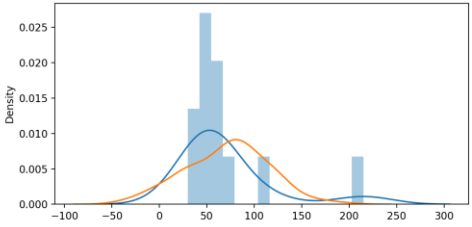
- Salary (in \$k) of 12 employees in a company (shown in increasing order):
  D = {30, 36, 47, 50, 52, 52, 56, 60, 62, 70, 110, 215} with **mean** $\bar{x} = 70$.

$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + \cdots + 110^2 + 215^2) - 70^2 \approx 2276.5$$
$$\sigma \approx \sqrt{2276.5} \approx 47.71$$

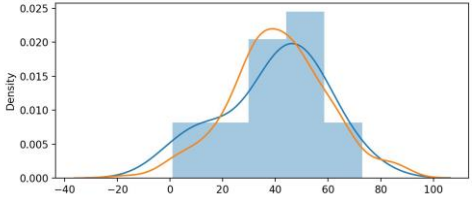
- Messi's goals for Barcelona from 2004–2005 to 2020–2021:
  D = {1, 8, 17, 16, 38, 47, 53, 73, 60, 41, 58, 41, 54, 45, 51, 31, 38} with **mean** $\bar{x} = 39.53$.

$$\sigma^2 = \frac{1}{17}(1^2 + 8^2 + \cdots + 31^2 + 38^2) - 39.53^2 \approx 359.38$$
$$\sigma \approx \sqrt{359.38} \approx 18.96$$

Sample vs. inferred distributions: salary and Messi cases



Salary case: blue line is the sample distribution; yellow line is the inferred (normal) distribution – numpy.random.normal(70, 47.71, 200)



Messi case: blue line is the real sample distribution; yellow line is the inferred (normal) distribution – numpy.random.normal(39.53, 18.96, 200)

Outline

Outline

Correlation

Correlation is a type of statistical relationship that reflects the strength and direction of association between two random variables or two data attributes (i.e., bivariate data).

- **Positive correlation** means that both variables change in the same direction.
- **Negative correlation** means that the variables change in opposite directions.
- **Zero correlation** means there is no dependency or relationship between the variables.

One of the most familiar measure of dependence between two variables is **Pearson correlation**. It is a measure of linear correlation between two sets of data. Pearson correlation works well if the relationship between variables is linear and if the variables are roughly normal. However, it is not robust for handling outliers, non–linear as well as skewed distributions. In this case, **Spearman's rank correlation** should be used.

Pearson correlation coefficient for a population

- Pearson correlation coefficient of two variables X and Y is the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{11}$$

where:

  - $cov(X, Y)$ is the covariance between X and Y.
  - σ_X is the standard deviation of X.
  - σ_Y is the standard deviation of Y.

- Because $cov(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$:

$$\rho_{XY} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{12}$$

where:

  - μ_X and μ_Y are the means of X and Y.
  - E is the expectation.

Pearson correlation coefficient for a population (cont'd)

- Because:
  - $\mu_X = \mathbb{E}[X]$
  - $\mu_Y = \mathbb{E}[Y]$
  - $\sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
  - $\sigma_Y^2 = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$
  - $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
    $= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- Then:

$$\rho_{XY} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}} \tag{13}$$

Pearson correlation coefficient for a sample

- Let D_{xy} be a bivariate data sample consisting of n data points being sampled from two variables/attributes X and Y as follows:

$$\mathcal{D}_{xy} = \begin{pmatrix} X & Y \\ \hline x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} \tag{14}$$

- The sample Pearson correlation coefficient between the two variables in D_{xy}, denoted as r_{xy}, is defined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{15}$$

where:

- $\bar{x}$ and $\bar{y}$ are the sample means of $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$, respectively.

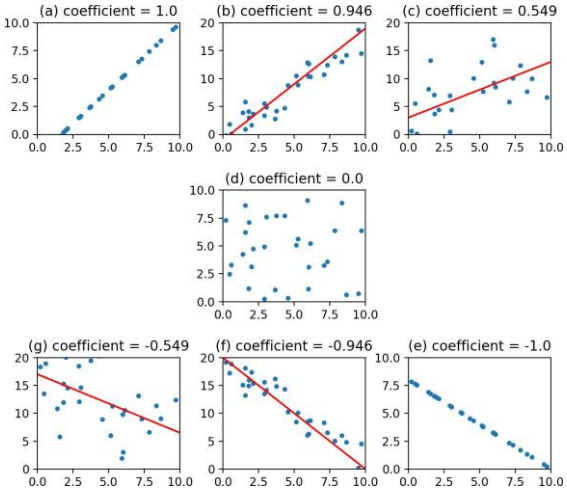
Pearson correlation coefficient for a sample (cont'd)

- Let $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$ for $1 \leq i \leq n$. The correlation coefficient r_{xy} is exactly the cosine of θ – the angle between the vector $\mathbf{a} = \{a_1, a_2, \ldots, a_n\}$ and the vector $\mathbf{b} = \{b_1, b_2, \ldots, b_n\}$:

$$r_{xy} = \cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{16}$$

- The value of r_{xy} varies in $[-1, 1]$.
  - $r_{xy} = -1$: completely negative correlation;
  - $-1 < r_{xy} < 0$: negative correlation;
  - $r_{xy} = 0$: no correlation, i.e., X and Y are independent;
  - $0 < r_{xy} < 1$: positive correlation;
  - $r_{xy} = 1$: completely positive correlation.

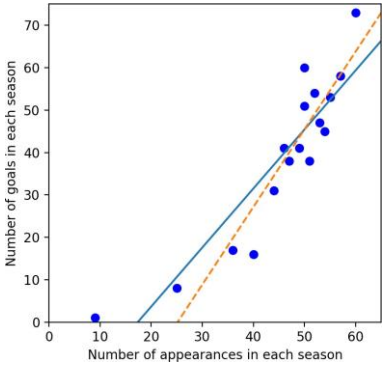
Levels of correlation (Pearson correlation coefficients)



Pearson correlation between the numbers of Messi's appearances & goals

Season	**Apps**	**Goals**
2004-05	9	1
2005-06	25	8
2006-07	36	17
2007-08	40	16
2008-09	51	38
2009-10	53	47
2010-11	55	53
2011-12	60	73
2012-13	50	60
2013-14	46	41
2014-15	57	58
2015-16	49	41
2016-17	52	54
2017-18	54	45
2018-19	50	51
2019-20	44	31
2020-21	47	38

Messi's appss and goals for Barcelona

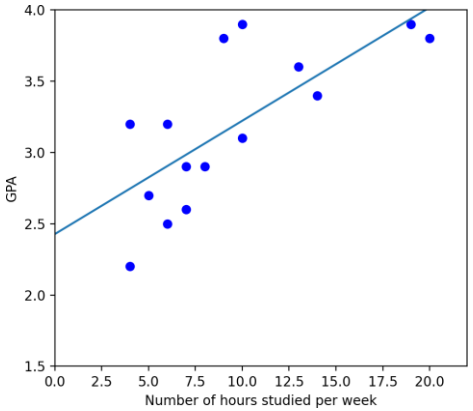


- Pearson correlation coefficient = 0.9021
- Blue line: regression with all data
- Yellow line: regression after removing the data of the first season

Pearson correlation between hours studied and GPA

Hours	GPA
6	3.2
9	3.8
10	3.9
13	3.6
6	2.5
5	2.7
7	2.9
19	3.9
20	3.8
14	3.4
10	3.1
7	2.6
4	2.2
8	2.9
4	3.2

The number of hours studied per week and the GPA



- Pearson correlation coefficient = 0.7315
- Blue line: linear regression

Outline

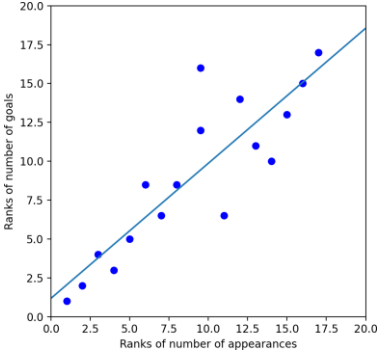
Spearman's rank correlation

- Spearman's rank correlation is an alternative that mitigates the effect of outliers and skewed distributions;
- To compute Spearman's correlation, we have to compute the rank of each value, which is its index in the sorted sample;
- For example, in the sample $[1, 2, 5, 7]$ the rank of the value 5 is 3, because it appears third in the sorted list. Then we compute Pearson's correlation for the ranks.

Spearman correlation between Messi's apps and goals

Season	**Apps (Ranks)**	**Goals (Ranks)**
04-05	9 (**1.0**)	1 (**1.0**)
05-06	25 (**2.0**)	8 (**2.0**)
06-07	36 (**3.0**)	17 (**4.0**)
07-08	40 (**4.0**)	16 (**3.0**)
08-09	51 (**11.0**)	38 (**6.5**)
09-10	53 (**13.0**)	47 (**11.0**)
10-11	55 (**15.0**)	53 (**13.0**)
11-12	60 (**17.0**)	73 (**17.0**)
12-13	50 (**9.5**)	60 (**16.0**)
13-14	46 (**6.0**)	41 (**8.5**)
14-15	57 (**16.0**))	58 (**15.0**)
15-16	49 (**8.0**)	41 (**8.5**)
16-17	52 (**12.0**)	54 (**14.0**)
17-18	54 (**14.0**)	45 (**10.0**)
18-19	50 (**9.5**)	51 (**12.0**)
19-20	44 (**5.0**)	31 (**5.0**)
20-21	47 (**7.0**)	38 (**6.5**)

Messi's apps and goals for Barcelona



- Spearman correlation coefficient = 0.8692

Outline

Pearson's chi–square test for categorical data

- For nominal/categorical data, a correlation relationship between two attributes/variables X and Y can be discovered by a χ^2 (chi-square) statistical test.
- Suppose X has r distinct values, namely $a_1, a_2, \ldots, a_r$; and Y has c distinct values, namely $b_1, b_2, \ldots, b_c$.
- Suppose we have a bivariate data sample consisting of n data tuples drawn from both X and Y as follows:

$$\mathcal{D} = \begin{pmatrix} X & Y \\ \hline x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} \tag{17}$$

- Let (X_i, Y_j) denote the joint event that X takes on value a_i and Y takes on value b_j, or $X = a_i$ and $Y = b_j$ for short.

Contingency table

- The n data tuples of $\mathcal{D}$ form a **contingency table**, with the r distinct values of X making up the rows and the c distinct values of Y making up the columns:

	b_1	b_2	$\cdots$	b_c	Total
a_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_1
a_2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_2
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_r
Total	m_1	m_2	$\cdots$	m_c	n

where:

- n_{ij} $(i = 1..r,\ j = 1..c)$ is the *observed frequency*, i.e., the actual count of the joint event (X_i, Y_j) in $\mathcal{D}$. That is, n_{ij} is the number of data tuples in $\mathcal{D}$ where $X = a_i$ and $Y = b_j$.
- $n_i = \sum_{j=1}^{c} n_{ij}$ is the number of data tuples in $\mathcal{D}$ where $X = a_i$.
- $m_j = \sum_{i=1}^{r} n_{ij}$ is the number of data tuples in $\mathcal{D}$ where $Y = b_j$.
- $n = \sum_{i=1}^{r} n_i = \sum_{j=1}^{c} m_j = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$

Pearson's chi–square test statistic calculation

- The χ^2 value (also known as the Pearson χ^2 statistic) is computed as follows:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{18}$$

where e_{ij} is the *expected frequency* of the joint event (X_i, Y_j). e_{ij} is computed as:

$$e_{ij} = \frac{count(X = a_i)\ count(Y = b_j)}{n} = \frac{n_i m_j}{n} \tag{19}$$

- Actually, e_{ij} is computed under the assumption that X and Y are independent. $\hat{P}(X = a_i) = \frac{n_i}{n}$ and $\hat{P}(Y = b_j) = \frac{m_j}{n}$, then:

$$e_{ij} = n\hat{P}(X_i, Y_j) = n\hat{P}(X = a_i)\hat{P}(Y = b_j) = n\frac{n_i}{n}\frac{m_j}{n} = \frac{n_i m_j}{n}$$

Pearson's chi–square test hypotheses

- The sum in equation (18) is computed over all of the $r \times c$ cells of the contingency table.
- Note that the cells that contribute the most to the χ^2 value are those for which the actual count (n_{ij}) is very different from the expected frequency (e_{ij}).
- For the Pearson's chi–square test, the **null hypothesis** H_o is that X and Y are independent, i.e., there is no correlation between them. The **alternative hypothesis** H_a is that there is a correlation relationship between X and Y.
- The test is based on a significance level, with $(r-1) \times (c-1)$ degrees of freedom.
- If the null hypothesis can be rejected, then we say that X and Y are statistically correlated.

Pearson's chi–square test example

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

2 × 2 contingency table between `gender` and `preferred_reading` [1]

- A group of 1500 people was surveyed. The gender of each person was noted. Each person was polled as to whether his or her preferred type of reading material was *fiction* or *nonfiction*.
- Thus, we have two attributes, `gender` and `preferred_reading`. The observed frequency (or count) of each possible joint event is summarized in the contingency table above. In which, the numbers in parentheses are the expected frequencies. For example:

$$e_{12} = \frac{count(female) \times count(fiction)}{n} = \frac{1200 \times 450}{1500} = 360$$

Pearson's chi–square test example (cont'd)

- Using equation (18) for χ^2 computation, we get:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93$$

- For this 2×2 table, the degrees of freedom are $(2-1)(2-1) = 1$.
- For 1 degree of freedom, the χ^2 value needed to reject the null hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the χ^2 distribution, typically available from any textbook on statistics).
- Since our computed value is above this, we can reject the null hypothesis that `gender` and `preferred_reading` are independent and conclude that the two attributes are (strongly) correlated for the given group of people.

Upper–tail critical values of chi–square distribution

Upper–tail critical values with ν degrees of freedom

	Probability less than the critical value				
ν	0.90	0.95	0.975	0.99	0.999
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252

Outline

Correlation versus causation

- **Correlation** describes an association between variables: when one variable changes, so does the other. A correlation is a statistical indicator of the relationship between variables.
- **Causation** means that changes in one variable brings about changes in the other. There is a cause–and–effect relationship between variables. The two variables are correlated with each other, and there is also a causal link between them.
- If variables X and Y are correlated, there are several possible explanations: X causes Y, or Y causes X, or X and Y cause each other, or some other set of factors causes both X and Y. The last case is the **third variable(s)** problem.
- Causation means correlation, but correlation does not necessarily imply causation.

Outline

Excercises

1. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) {13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}.
   a. What is the *mean* of the data? What is the *median*?
   b. What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
   c. What is the *midrange* of the data?
   d. Can you find (roughly) the first quartile (Q_1) and the third quartile (Q_3) of the data?
   e. Give the *five–number summary* of the data.
   f. Show a *boxplot* of the data.
   g. What is the *variance* and *standard deviation* of the data?

Excercises (cont'd)

2. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

   a. Calculate the mean, median, variance, and standard deviation of *age* and *%fat*.
   b. Draw the boxplots for *age* and *%fat*.
   c. Draw a *scatter plot* for these two variables.
   d. Calculate the Pearson and Spearman correlation coefficients between the two attributes.

Excercises (cont'd)

$\mathcal{D} =$

Club	*Season*	*Appearances*	*Goals*	*Assists*
Man United	03/04	5	0	1
Man United	04/05	7	0	2
Man United	05/06	6	0	0
Man United	06/07	11	3	5
Man United	07/08	11	8	1
Man United	08/09	12	4	3
Real Madrid	09/10	6	7	1
Real Madrid	10/11	12	6	4
Real Madrid	11/12	10	10	3
Real Madrid	12/13	12	12	1
Real Madrid	13/14	11	17	5
Real Madrid	14/15	12	10	4
Real Madrid	15/16	12	16	4
Real Madrid	16/17	13	12	6
Real Madrid	17/18	13	15	3
Juventus	18/19	9	6	2
Juventus	19/20	8	4	1
Juventus	20/21	6	4	2
Man United	21/22	7	6	0

Excercises (cont'd)

3. Let D (in the previous slide) be the number of appearances, goals, and assists of Cristiano Ronaldo in Champions League from season 2003/2004 to 2021/2022.
   a. Calculate the *mean*, *median*, *mode*, *variance*, *standard deviation* of three attributes appearances, goals, and assists.
   b. Find the *first quartile* (Q_1) and the *third quartile* (Q_3) of three three attributes appearances, goals, and assists.
   c. Show the *boxplots* of the three attributes appearances, goals, and assists.
   d. Show the *scatter plots* of appearances and goals; of appearances and assists; of goals and assists.
   e. Compute the *covariance matrix* of the three attributes appearances, goals, and assists.

Excercises (cont'd)

4. 800 students of three majors (math, history, and computer science) were asked whether they like playing computer games or not. The survey data are summarized in the following contingency table:

	Like games	Do not like games	Total
Math	130	100	230
History	35	165	200
Computer science	280	90	370
Total	445	355	800

a. Calculate the expected frequency of each cell in the table above.
b. Using Pearson's chi–square test to confirm that there is a correlation between study major and playing computer games or not?

References

1 J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Elsevier, 2012 [Book1].

2 C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015 [Book2].

3 J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2014 [Book3].

4 M. J. Zaki and W. M. Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2013 [Book4].

5 D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010 [Book5].

6 J. VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly, 2017 [Book6].

7 J. Grus. *Data Science from Scratch: First Principles with Python*. O'Reilly, 2015 [Book7].

Summary

- Explaining why data understanding is critical before doing any data analysis and mining tasks.
- Knowing different types of data attributes such as categorical, binary, numeric, etc.
- Introducing various data forms like univariate, bivariate, multivariate as well as more complex and structured data like time–series, sequences, spatial, network/graph, and stream data.
- Understanding and knowing how to measure the central tendency as well as the variability of data; knowing some statistics and visualization tools like boxplot to understand the sample distribution.
- Studying the potential relationships between data variables/attributes via important concepts like covariance, correlation, Pearson, Spearman correlation, Pearson's chi–square test, as well as explaining the difference between correlation and causation.