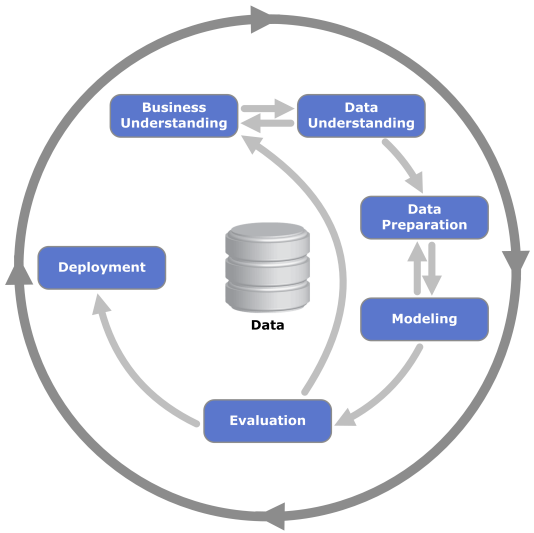
Machine learning

Data analysis process

Cross Industry Process for Data Mining CRISP-DM, 1996)



1. **Problem Understanding lub Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**

The quality of the data and the amount of useful information that it contains are key factors that determine how well a machine learning algorithm can learn.

Therefore, it is absolutely critical to ensure that we examine and preprocess a dataset before we feed it to a machine learning algorithm.

The word **data**, is the plural of the word **datum**.

datum (język łaciński) [edytuj]

znaczenia:

rzeczownik, rodzaj nijaki

(1.1) dar, podarek, datek[1]

(1.2) data[1]

odmiana:

(1.1) datum, datī (deklinacja II) [ukryj▲]

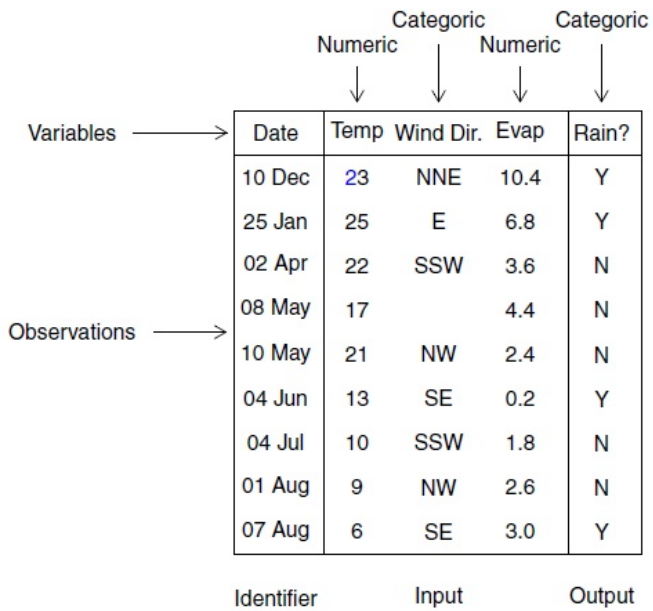
przypadek	liczba pojedyncza	liczba mnoga
mianownik	**datum**	**data**
dopełniacz	datī	datōrum
celownik	datō	datīs
biernik	datum	data
ablatyw	datō	datīs
wołacz	datum	data

Data

- The term was the first to be used by **Euclid**, in the work **Dedomena**.
- **data** is called, in Euclid's work, a quantity resulting directly from the terms of a given problem.
- **Euclid**

Data in the form of records

- each **record (object, sample, observation)** is described by a set of attributes (variables).
- each observation (record) has a fixed number of attributes (i.e. a fixed tuple length), so that it can be considered as a **vector in a multidimensional space** whose dimension is equal to the number of attributes.
- the data set can be represented as a matrix of type $m \times n$, where each of the m rows corresponds to an observation and each of the n columns corresponds to an attribute ($D = \{x_{ij}\}_{i=1}^{m}, j = 1, ..., n.$)

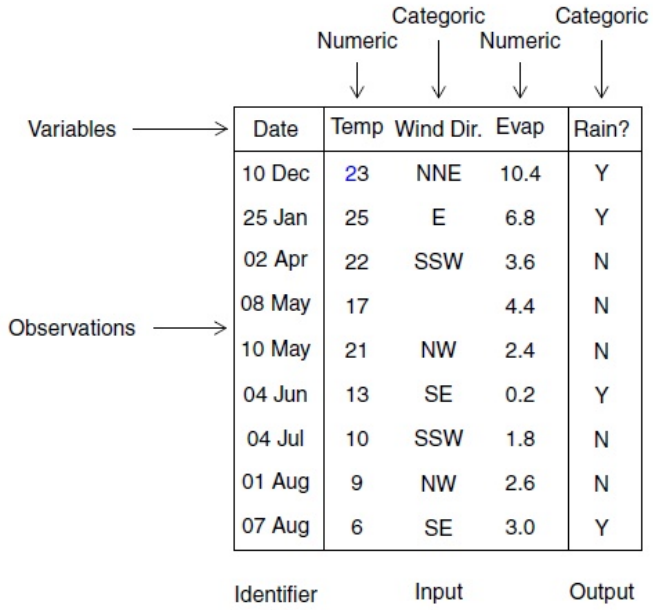


Date	Temp	Wind Dir.	Evap	Rain?
10 Dec	23	NNE	10.4	Y
25 Jan	25	E	6.8	Y
02 Apr	22	SSW	3.6	N
08 May	17		4.4	N
10 May	21	NW	2.4	N
04 Jun	13	SE	0.2	Y
04 Jul	10	SSW	1.8	N
01 Aug	9	NW	2.6	N
07 Aug	6	SE	3.0	Y

Data in the form of records

Diagnosis	GGT(u/l)	Cholesterol (mg/dL)	Albumin (g/dL)	Age (year)	Glycemia (mmol/L)	Sex
Cirrhosis	289	148	3.12	57	0.9	M
Hepatitis	255	258	3.5	65	1.1	M
Hepatitis	32	240	4.83	60	1.14	F
Hepatitis	104	230	4.06	36	0.9	F
Cirrhosis	585	220	2.7	55	1.5	F
Cirrhosis	100	161	3.8	57	1.02	M
Hepatitis	85	188	3.1	48	1.09	M
Cirrhosis	220	138	3.84	58	0.9	M
Cancer	1117	200	2.3	57	2.2	F
Cancer	421	309	3.9	44	1.1	M

Data in the form of records



Date	Temp	Wind Dir.	Evap	Rain?
10 Dec	23	NNE	10.4	Y
25 Jan	25	E	6.8	Y
02 Apr	22	SSW	3.6	N
08 May	17		4.4	N
10 May	21	NW	2.4	N
04 Jun	13	SE	0.2	Y
04 Jul	10	SSW	1.8	N
01 Aug	9	NW	2.6	N
07 Aug	6	SE	3.0	Y

- **Input variables** are also referred to as independent variables, observed variables or descriptive variables.
- **Output variables** are dependent on the input variables. They are referred to as target, response or dependent variables.

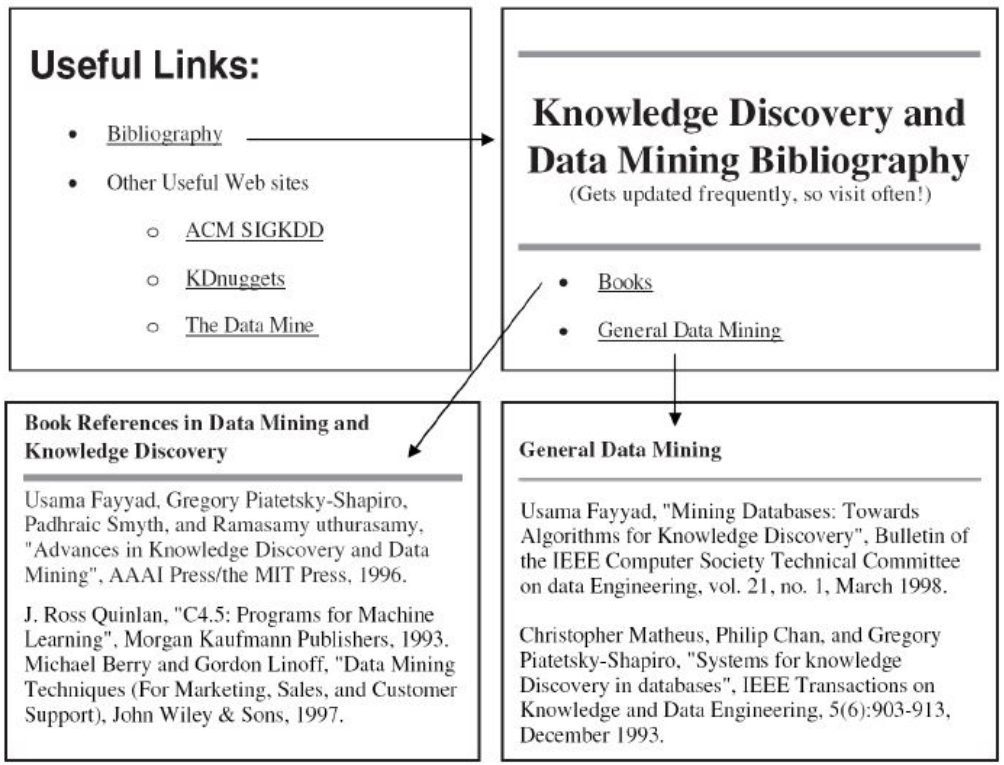
Transaction-based sets

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

- each transaction (purchase, observation) is assigned a vector.
- transaction components denote goods, objects, etc.

Data in graph form

The vertices of the graph are used to store the data, while the edges indicate the relationships between the data.



Data quality

- **Machine learning algorithms** are very sensitive to the quality of the source data;
- **GIGO (Garbage In, Garbage Out)** - results of processing incorrect data will be wrong regardless of the correctness of the processing procedure.

Data properties:

- completeness,
- correctness,
- actuality.

Noise in the data

- **label noise**
  - **inconsistent observations**
  - **classification errors** - observations that are labelled as a class other than the actual class.

Att 1	Att 2	Class
0.25	red	positive
0.25	red	negative
0.99	green	negative
1.02	green	positive
2.05	?	negative
=	green	positive

Often, the term **noised data** is used as a synonym for corrupted data.

Noise in the data

- **attribute noise** this refers to incorrect values of one or more attributes
  - **wrong attribute values** (1.02, green, class= positive) when we assume that an attribute has a bad value.
  - **missing or unknown attribute values**(2.05,?, class = negative) - we do not know the value of the second attribute.
  - **incomplete attributes or values** (=, green, class = positive) - the system cannot understand and correctly interpret values.
  - **outliers.**

Att 1	Att 2	Class
0.25	red	positive
0.25	red	negative
0.99	green	negative
1.02	green	positive
2.05	?	negative
=	green	positive

Types of Variable

The data type of attributes helps analysts select the correct method for data analysis and visualization plots.

We can divide attributes into just two types:

- **qualitative, (non-measurable)** (categorical) refers to names or labels of categorized variables; cannot be uniquely characterised by numbers; such as product name, brand name, zip code, state, gender, marital status or size of a T-shirt: small, medium, or large.
- **quantitative (numeric)** is presented as integer or real values .e.g. the height of a person, the number of items sold articles.

Qualitative data:

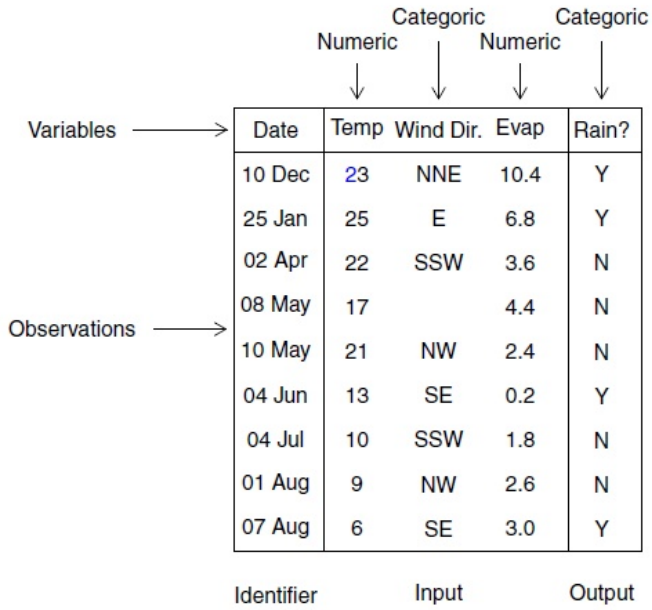
- the set of available values is always limited (e.g. in the case of months to 12); for this reason the values of categorical variables are called lables;
  - **nominal attributes** - the value of a nominal attribute can be the symbol or name of items. The values are categorical, qualitative, and **unordered in nature** such as product name, brand name, zip code, state, gender, and marital status.
  - **ordinal attributes** - refers to names or labels with a meaningful order or ranking.These types of attributes measure subjective qualities alone. That is why they are used in surveys for customer satisfaction ratings, product ratings, and movie rating reviews. Customer satisfaction ratings appear in the following order: **1: Very dissatisfied, 2: Somewhat dissatisfied, 3: Neutral, 4: Satisfied, 5: Very satisfied.**
- determining the distance between values is only possible within the framework of the adopted model;
- it is impossible to perform arithmetic operations on them.

Types of Variable

Quantitative data:

- the values can be compared with each other;
- it is possible to determine the distance between values;
- arithmetic operations can be performed on them;
- the values can be
  - discrete (integers) - a finite or countable set of values;
  - continuous (real numbers).

Types of Variable dataset



Date	Temp	Wind Dir.	Evap	Rain?
10 Dec	23	NNE	10.4	Y
25 Jan	25	E	6.8	Y
02 Apr	22	SSW	3.6	N
08 May	17		4.4	N
10 May	21	NW	2.4	N
04 Jun	13	SE	0.2	Y
04 Jul	10	SSW	1.8	N
01 Aug	9	NW	2.6	N
07 Aug	6	SE	3.0	Y

Types of Variable

- Variables can take one or multiple values.
- **Single-valued variables are otherwise known as constants or identifier**.

Single-valued variables should not be used in the data analysis process as they carry no information.

- remark - it should be checked whether a given variable is single-valued in the selected sample or in the entire source data set - if some values of a variable occur very rarely (e.g. once in 500 000 cases), we will probably not find them in a sample of 10 000 rows. By removing such a variable from the training dataset of a data mining model, we may not only degrade the accuracy of its results, but also **prevent it from recognising unusual cases, possibly the most interesting ones**.

Types of Variable

- **Some variables are used to uniquely identify the observation**, this could be, for example, some official identification number. The identifier can also be, for example, the date of the observation.
- **Identifiers** are not used in the model.

Types of Variable

Another type of variables not useful for predictive models are **monotonic variables.**

- The values of such variables constantly increase or decrease.
  - this type of variable is very common, for example: the values of all time-related variables (such as invoice date or date of birth) are increasing;

D. Larose

- Let us assume that we run a local shop and that we register all the details in the shop's database. We know our customers' details and what they buy each day.
- E.g. Alex, Jessica and Paul visit the shop every Sunday and buy candles. What we store in the database is just the **data**.
- Every time we want to know who the visitors are who buy the candles, we can search the database and get the answer. This is **information**. If we want to know how many candles were sold on each day of the week, then we can again direct a query to the database database and get the answer - this is also **information**.

- But suppose we have many other customers who also buy candles from us every Sunday (mostly with some level of freedom), and all of them are Christians (going to church). So so we can conclude that Alex, Jessica and Paul must also be Christians. The religion of Alex, Jessica and Paul was not recorded in the database, so we could not retrieve it as information from it. We learned this information indirectly. It is a **knowledge** that we discovered.
- Of course, it is likely that our findings as to Alex, Jessica and Paul may be wrong. Therefore it is important that our knowledge and findings are evaluated correctly.

Data pre-processing

Data pre-processing involves cleaning and transformation of data to prepare it for mining. It is estimated that data preprocessing is 70-80% of the process of of knowledge discovery.

For example, a database may contain

- fields that are out of date or redundant,
- records with missing values,
- outliers,
- data in a format unsuitable for machine learning models,
- values incompatible with principles or common sense.

To illustrate the necessity of data cleaning, we will examine the following example (**D.Larose**):

We will analyse the personal data from the following table

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	78,000	C	M	5000
1002	J2S7K7	F	−40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

Example

Data Mining Map. http://www.saedsayad.com/data_mining_map.htm

The field of **statistics** helps us to gain an understanding of our data, and to quantify what our data and results look like.
It also provides us with mechanisms to measure how well our application is performing and prevent certain machine learning pitfalls (such as under/overfitting).

Distribution

A **distribution** is a representation of how often values appear within a dataset.

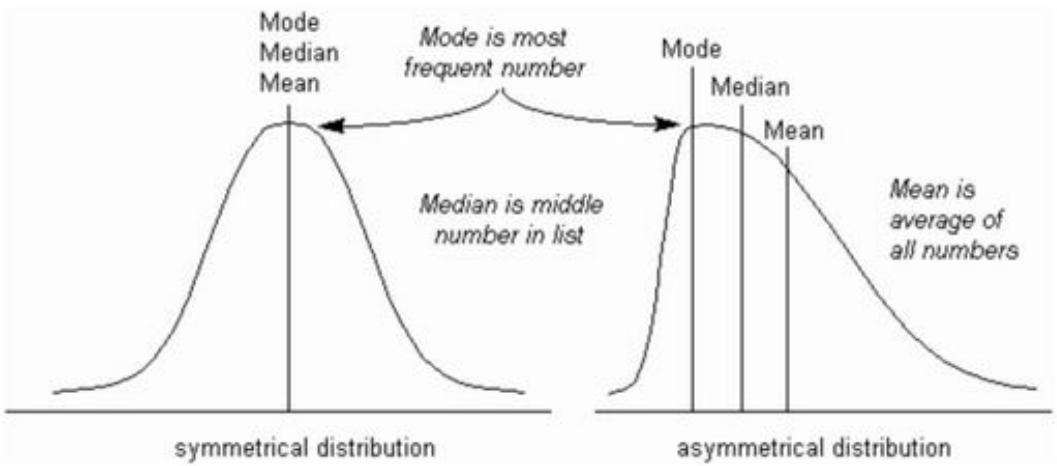
There are two types of these measures:

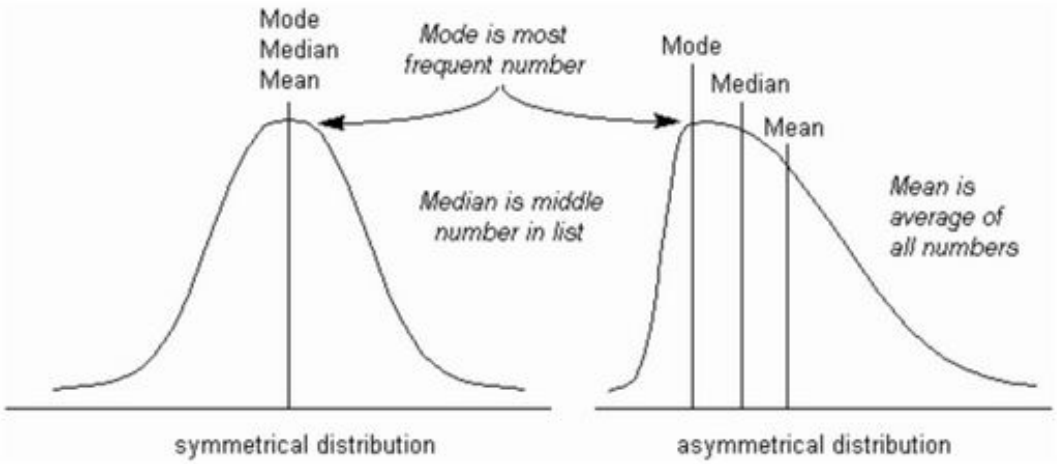
- **central tendency measures** - these measure where most of the values are located, or where the center of the distribution is located;
- **spread or dispersion measures** - these measure how the values of the distribution are spread across the distribution's range (from the lowest value to the highest value).

Measures of central tendency include the following:

- **Mean** - this is what you might commonly refer to as an average. We calculate this by summing all of the numbers in the distribution and then dividing by the count of the numbers. The mean of a sample $x_1, x_2, \ldots, x_n$, usually denoted by $\bar{x}$

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

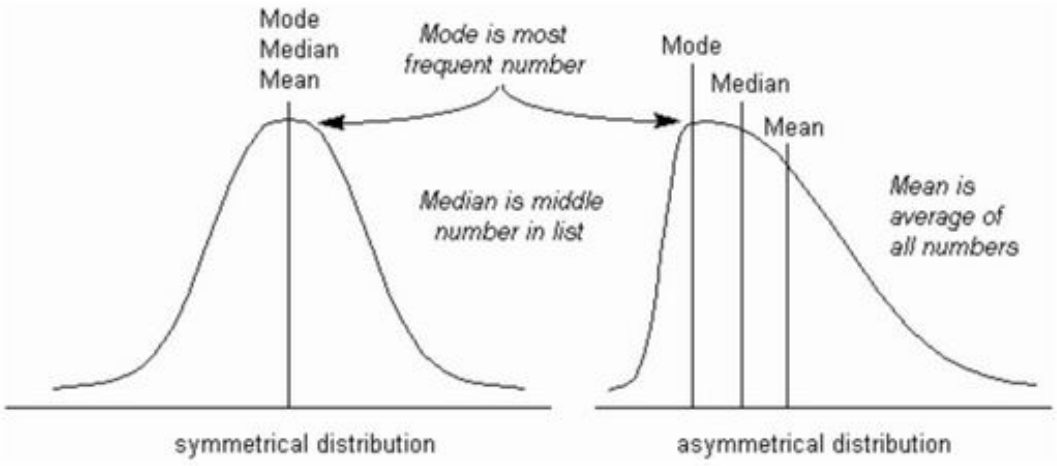




- **Median** - if we sort all of the numbers in our distribution from the lowest to highest, this is the number that separates the lowest half of the numbers from the highest half of the numbers. The median of a sample $x_1, x_2, \ldots, x_n$, where $x_1 \leq x_2 \leq \ldots \leq x_n$

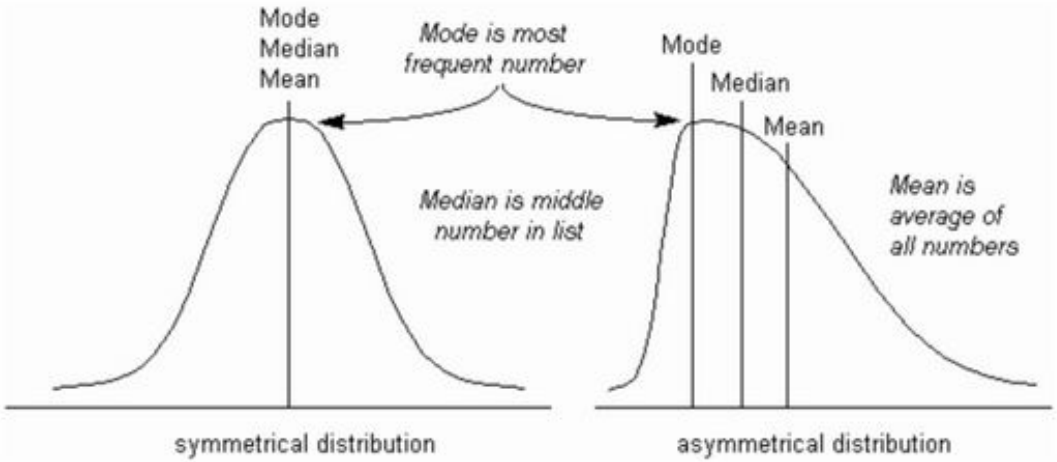
$$m_e = \begin{cases} x_{(n+1)/2} & n \text{ is odd,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \end{cases}$$

Measures of central tendency



- **Mode** - this is the most frequently occurring value in the distribution.

Measures of central tendency



- If the **mean** and **median** are significantly different, then we can expect that some observations for a given distribution are far from the mean. These are **outlier points**.

Measures of spread or dispersion

Measures of dispersion include the following:

- **Maximum** - the highest value of the distribution
- **Minimum** - the lowest value of the distribution
- **Range** - the difference between the maximum and minimum
- **Variance** - this measure is calculated by taking each of the values in the distribution, calculating each one's difference from the distribution's mean, squaring this difference, adding it to the other squared differences, and dividing by the number of values in the distribution
- **Standard deviation** - the square root of the variance
- **Quantiles/quartiles** - similar to the median, these measures define cut-off points in the distribution where a certain number of lower values are below the measure and a certain number of higher values are above the measure

Variance

The variance of a sample $x_1, x_2, \ldots, x_n$ is of the form

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

where $\bar{x}$ is the mean value.

The standard deviation is calculating a sort of average distance of how far the data values are from the arithmetic mean.

- by taking $x - \bar{x}$, you are finding the literal difference between the value and the mean of the sample;
- by squaring the result, $(x_i - \bar{x})^2$, we are putting a **greater penalty on outliers** because squaring a large error only makes it much larger.
- by dividing by the number of items in the sample, we are taking (literally) the average squared distance between each point and the mean.

Standard deviation

The standard deviation of a sample $x_1, x_2, \ldots, x_n$ is of the form

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

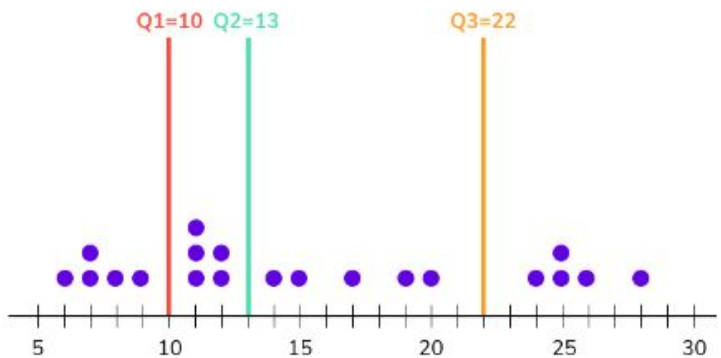
where $\bar{x}$ is the mean value.

A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Quartile

A **quartile** divides the number of sorted data points into four parts, or quarters, of more-or-less equal size. The three main quartiles are as follows:

- **the first quartile** Q_1 is defined as the middle number between the smallest number (minimum) and the median of the data set. It is also known as the lower or 25th empirical quartile, as 25% of the data is below this point.
- **the second quartile** Q_2 is the median of a data set; thus 50% of the data lies below this point.
- **the third quartile** Q_3 is the middle value between the median and the highest value (maximum) of the data set. It is known as the upper or 75th empirical quartile, as 75% of the data lies below this point.



Quantiles

q-quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes.

- In statistics and probability, quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.
- Common quantiles have special names, such as
  - quartiles (four groups),
  - quintiles (five groups),
  - deciles (ten groups),
  - percentiles (100 groups).

Skewness

Skewness measures the symmetry of a distribution.

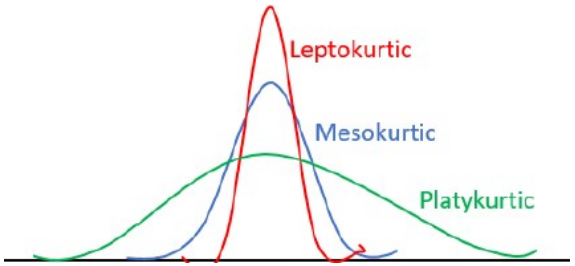
- It shows how much the distribution deviates from a normal distribution.
- Its values can be **zero, positive, and negative**.
  - A zero value represents a perfectly normal shape of a distribution.
  - Positive skewness is shown by the tails pointing toward the right—that is, outliers are skewed to the right and data stacked up on the left.
  - Negative skewness is shown by the tails pointing toward the left—that is, outliers are skewed to the left and data stacked up on the right.
- Positive skewness occurs when the mean is greater than the median and the mode.
- Negative skewness occurs when the mean is less than the median and mode.

Kurtosis

Kurtosis measures the tailedness (thickness of tail) compared to a normal distribution.

- **High kurtosis is heavy-tailed, which means more outliers are present in the observations, and low values of kurtosis are light-tailed, which means fewer outliers are present in the observations.**
- There are three types of kurtosis shapes:
  - mesokurtic
  - platykurtic
  - leptokurtic

Kurtosis



A normal distribution having zero kurtosis is known as a mesokurtic distribution.

A platykurtic distribution has a negative kurtosis value and is thin-tailed compared to a normal distribution.

A leptokurtic distribution has a kurtosis value greater than 3 and is fat-tailed compared to a normal distribution.

Covariance

Covariance measures the relationship between a pair of variables.

It shows the degree of change in the variables—that is, how the change in one variable affects the other variable. Its value ranges from -infinity to + infinity.

The problem with covariance is that it does not provide effective conclusions because it is not normalized.

Correlation

Correlation shows how variables are correlated with each other.

Correlation ranges from -1 to 1.

- A negative value represents the increase in one variable, causing a decrease in other variables or variables to move in the same direction.
- A positive value represents the increase in one variable, causing an increase in another variable, or a decrease in one variable causes decreases in another variable.
- A zero value means that there is no relationship between the variable or that variables are independent of each other.

The method parameter can take one of the following three parameters:

- pearson: Standard correlation coefficient
- kendall: Kendall's tau correlation coefficient
- spearman: Spearman's rank correlation coefficient

Spearman's rank correlation coefficient is Pearson's correlation coefficient on the ranks of the observations.

- It is a non-parametric measure for rank correlation.
- It assesses the strength of the association between two ranked variables.
- Ranked variables are ordinal numbers, arranged in order.
- First, we rank the observations and then compute the correlation of ranks.
- It can apply to both continuous and discrete ordinal variables.
- When the distribution of data is skewed or an outlier is affected, then Spearman's rank correlation is used instead of Pearson's correlation because it doesn't have any assumptions for data distribution.

Kendall's rank correlation coefficient or Kendall's tau coefficient is a non-parametric statistic used to measure the association between two ordinal variables. It is a type of rank correlation. It measures the similarity or dissimilarity between two variables. If both the variables are binary, then Pearson's = Spearman's = Kendall's tau.

Thank you for your attention!!!