SKEMA
BUSINESS SCHOOL

Introduction to
Artificial Intelligence

Dmitry A. Zaitsev
http://daze.ho.ua

EFMD EMBA | AACSB ACCREDITED | EFMD EQUIS ACCREDITED | skema BUSINESS SCHOOL

SKEMA
BUSINESS SCHOOL

Introduction to
Artificial Intelligence
Dmitry A. Zaitsev
http://daze.ho.ua

EFMD EMBA | AACSB ACCREDITED | EFMD EQUIS ACCREDITED | skema BUSINESS SCHOOL

Lesson 4
Basics of Data Science

Lesson 4
Basics of Data Science

Probability

Combinations and permutations . Conditional probability. Discrete and continuous random variables. Distribution.

Statistics

Popylation and samples. Data series. Mean, median, variance, correlation.

Data types

Classification of numeric and cathegorial data. Data types in programming languages and data bases.

Data visualization

Tables and graphs. Diagrams: pie, bar, line, scatter plot, heat map, quantile and quartile, histogram, box plot etc

Population and sample

Collecting data from a population

- A population is the entire group that you want to draw conclusions about.
- A sample is the specific group that you will collect data from.
- The size of the sample is always less than the total size of the population.
- An example:
    - Advertisements for IT jobs in the Netherlands
    - The top 50 search results for advertisements for IT jobs in the Netherlands on May 1, 2020

Data series

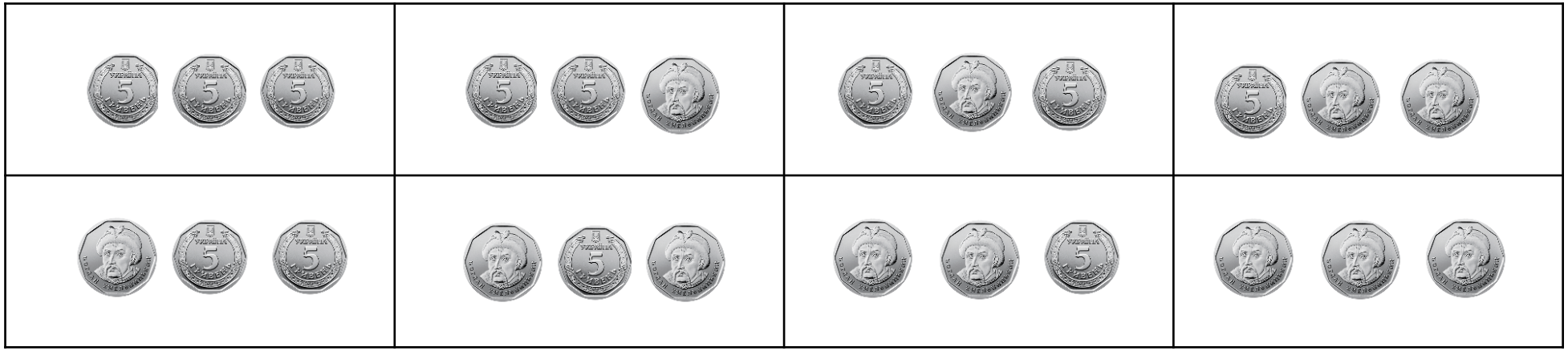
What is time series data?

- A data series is a row or column of numbers.
- Special case – time series indexed by instants (periods) of time
- Time series data, also referred to as time-stamped data, is a sequence of data points indexed in time order.
- These data points typically consist of successive measurements made from the same source over a fixed time interval and are used to track change over time.
- Time series data is a collection of observations obtained through repeated measurements over time.

Combinatorics

Enumerate cases and calculate their number

- Combinatorics is the branch of mathematics studying the enumeration, combination, and permutation of sets of elements and the mathematical relations that characterize their properties.



Probability

Ratio of specific case number to total number of cases

- Probability of an event – number of cases where the event takes place divided by the total number of cases
- One eagle: p=3/8=0.125
- Exactly two eagles: p=3/8=0.375
- At least two eagles: p=4/8=0.5

Basic combinatorial rules

Rules od sum, product, and conditional probability

Rule of sum: if there are a ways of doing something and b ways of doing another thing, and the two events cannot both occur, then there are a + b total possible outcomes for the events.

Rule of product: if there are a ways of doing something and b ways of doing another thing, then there are a · b ways of performing both actions

Conditional probability: a measure of the probability of an event occurring, given that another event has already occurred

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

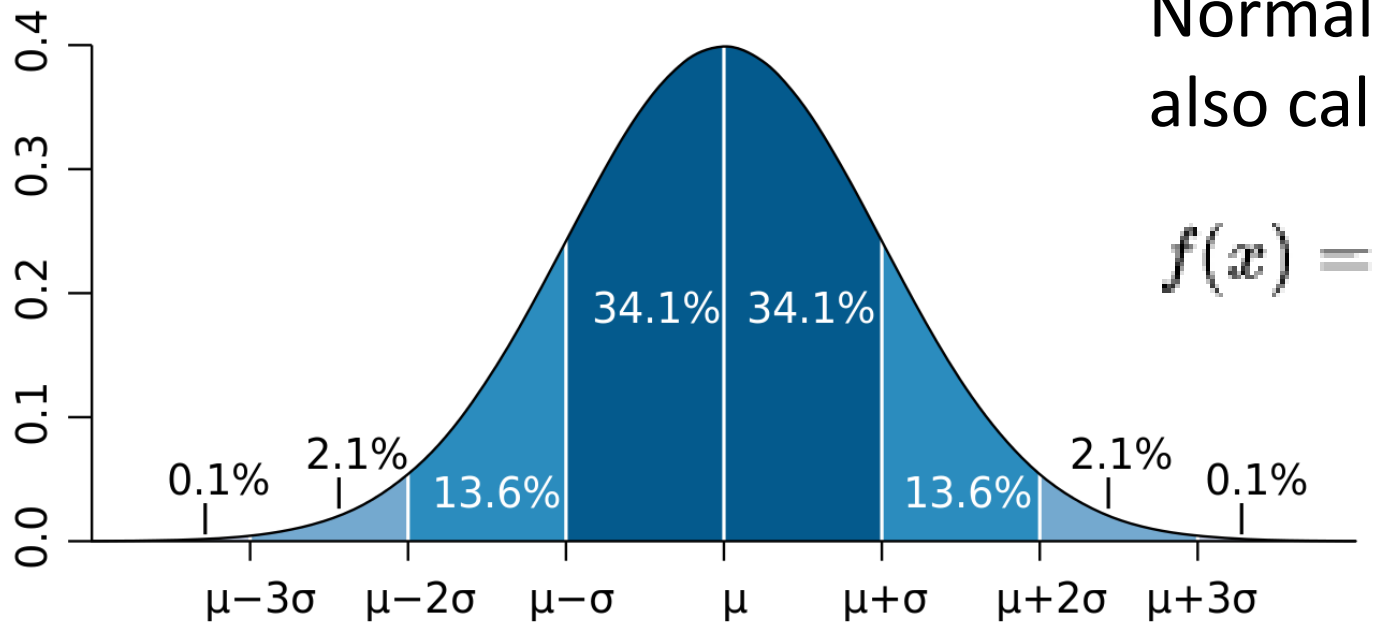
Random variables

Discrete and continuous random variables

- a variable whose possible values are numerical outcomes of a random phenomenon; a set of possible values from a random experiment
- a random variable is said to be discrete if the set of values it can take (its support) has either a finite or an infinite but countable number of elements
- a continuous random variable is a random variable that has only continuous values. Continuous values are uncountable and are related to real numbers

Distribution of random variable

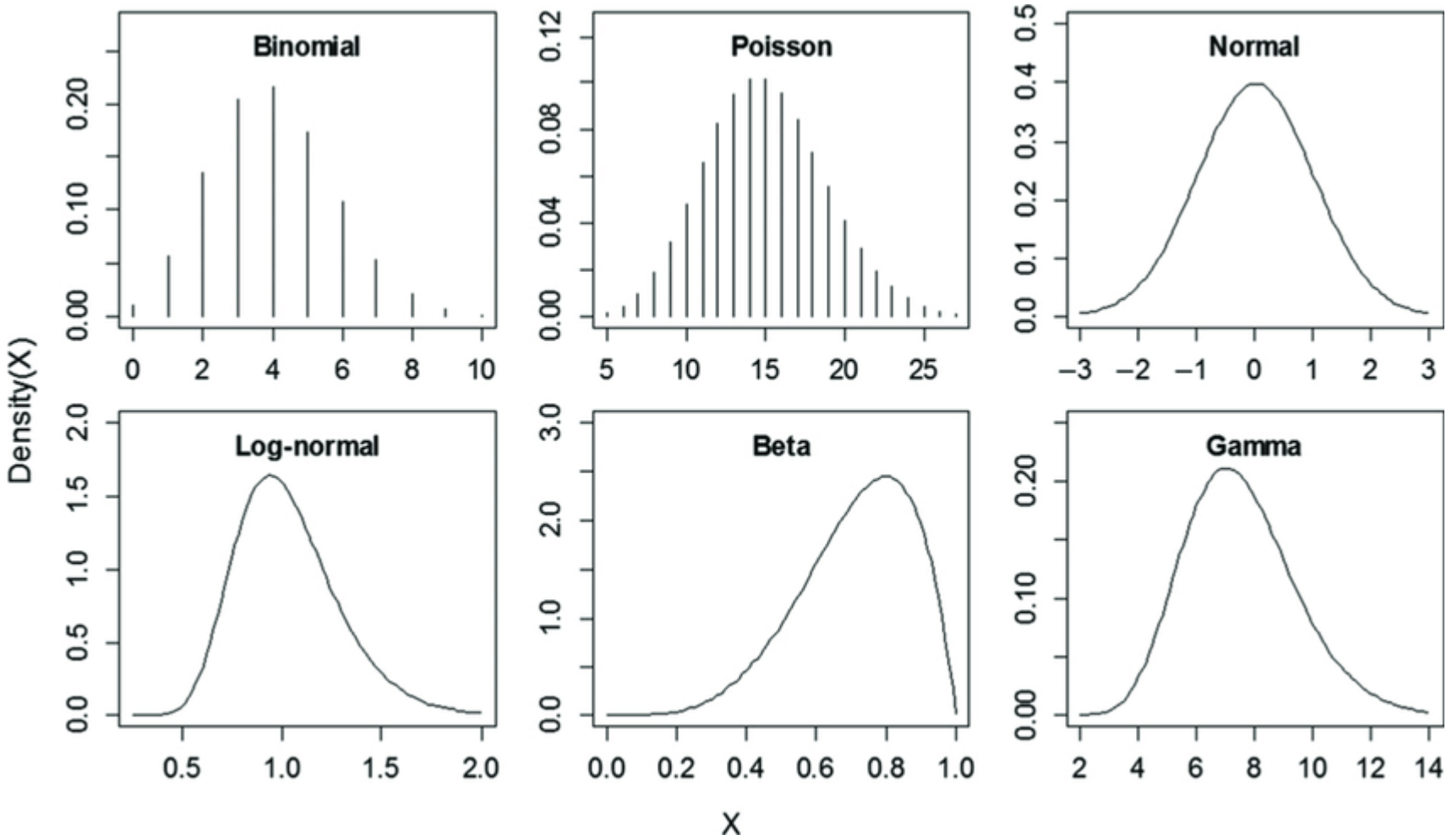
- The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable
- a probability density function is a function whose value at any point in the sample space can be interpreted as providing a relative likelihood that the value of the random variable would be equal to that sample



Normal distribution, also called Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Examples of distributions



Statistical values

Basic statisticalvalues to calculate and analyze

- minimum, maximum, range
- mean
- median, mode
- dispersion, variance
- higher rank statistical moments
- covarience and correlation

Mean and median

Central tendency

- The mean is the average of the numbers. It is easy to calculate: add up all the numbers, then divide by how many numbers there are.
- The mean is the integral of a continuous function of one or more variables over a given range divided by the measure of the range.
- The median is the middle value of a sorted array having an odd length; for an even length, two middle numbers are summed and divided by two.
- The mode is the value that appears most often in a set of data values.

Mean and median – how to compute

Central tendency

mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{n}$

$$\text{median}(x) = x_{(n+1)/2}$$

odd n

$$\text{median}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

even n

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	*Sum of values of a data set divided by number of values*	(1+2+2+3+4+7+9) / 7	**4**
Median	*Middle value separating the greater and lesser halves of a data set*	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	*Most frequent value in a data set*	1, **2**, **2**, 3, 4, 7, 9	**2**

Dispersion

Spread out of data

Range – difference between maximum and minimum: $r = x^{max} - x^{min}$

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

Standard deviation: σ

Dispersion - example

Spread out of data

$$X = \{1,2,2,3,4,7,9\}$$

Range: $r = 9 - 1 = 8$

Variance:

$$\sigma^2 = \frac{(1-4)^2+(2-4)^2+(2-4)^2+(3-4)^2+(4-4)^2+(7-4)^2+(9-4)^2}{7-1} = \frac{52}{6} \approx 8.7$$

Standard deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{8.7} = 2.9$

Covariance and correlation

Mutual influence

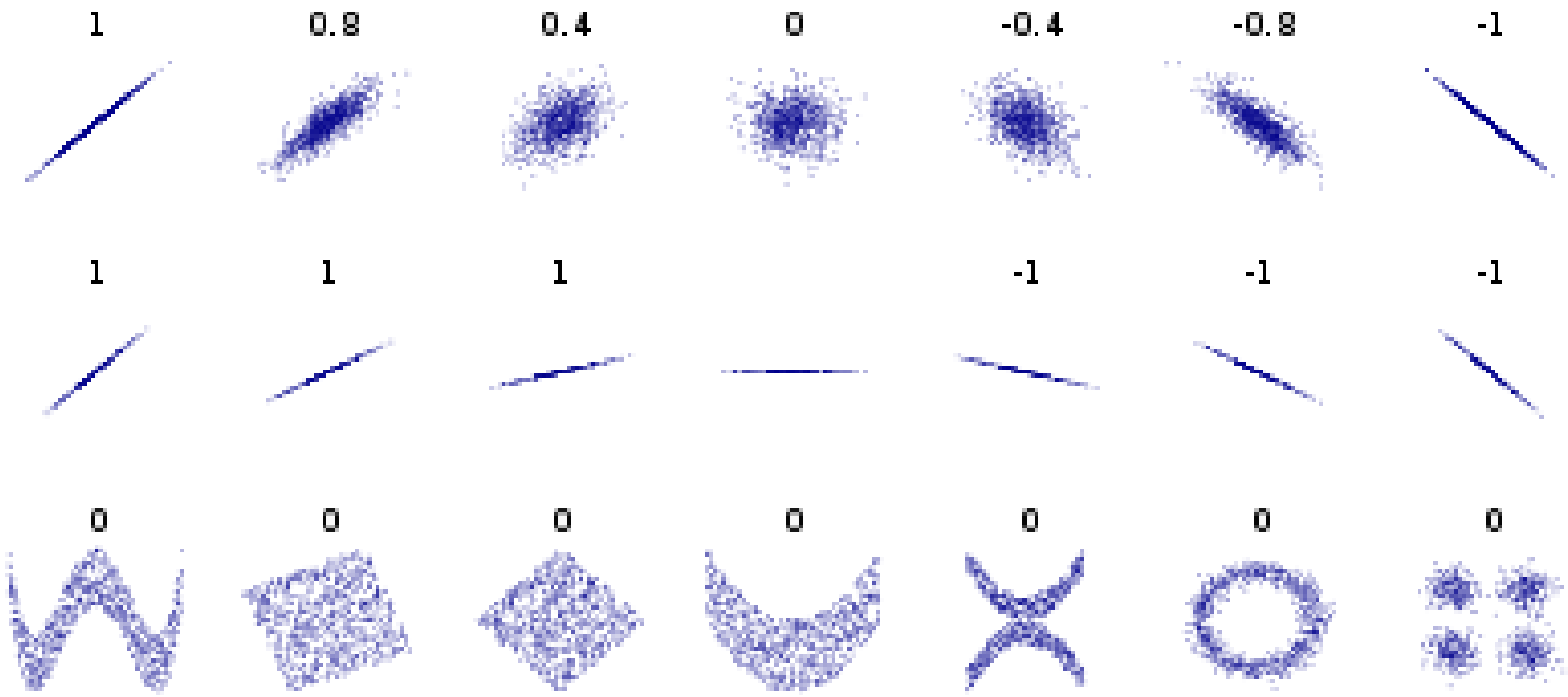
Covariance is a measure of the joint variability of two random variables:

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

Correlation or dependence is any statistical relationship, whether causal or not, between two random variables:

$$corr(X,Y) = \rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

Correlation shape and value



Task 3

Compute manually basic statistical values

- for your company, compute basic statistical values on time series for one year and monthly frequency (12 values):

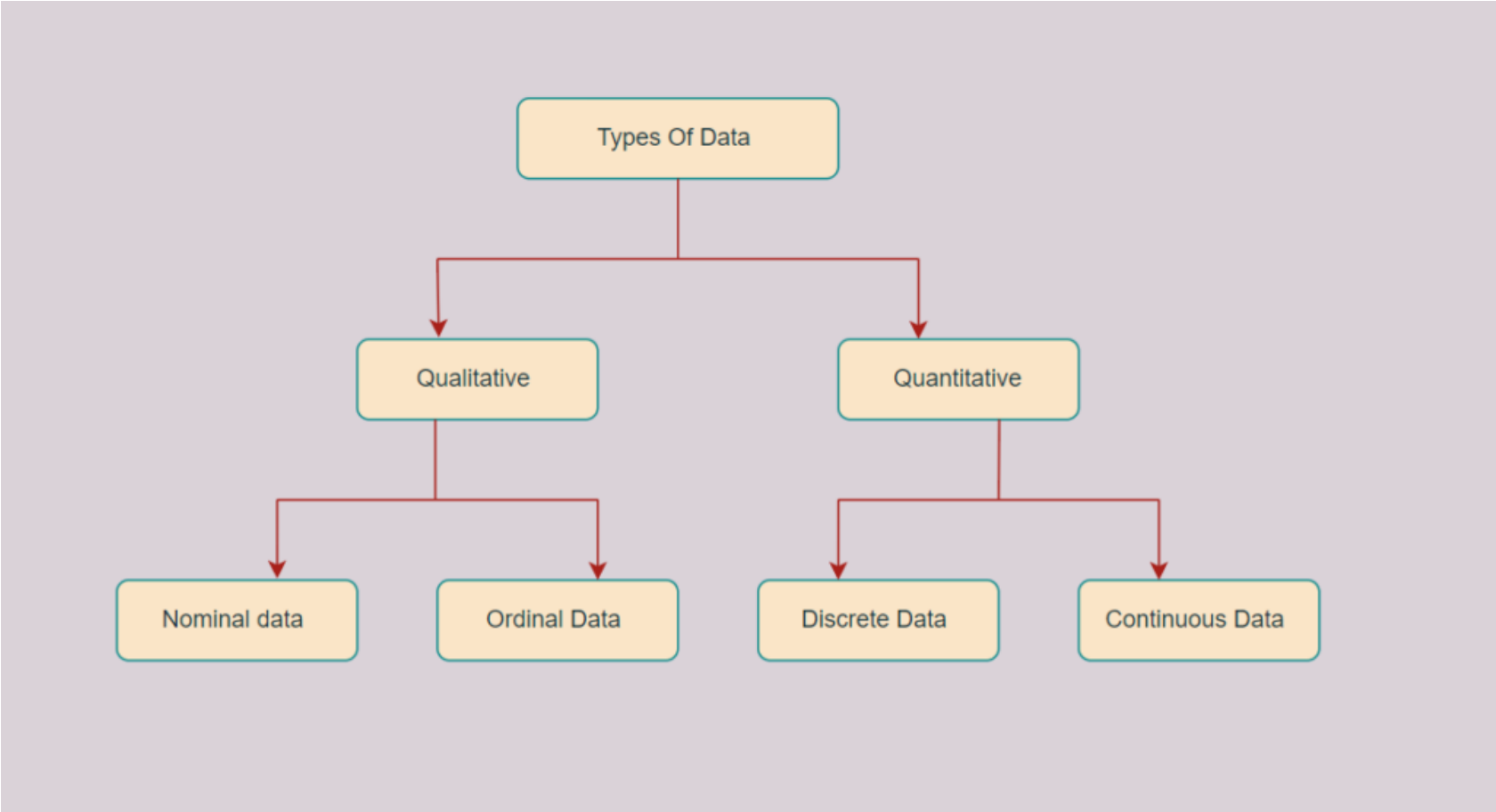
minimum, maximum, range; mean, median, mode; variance, deviation

- using other company compute manually:

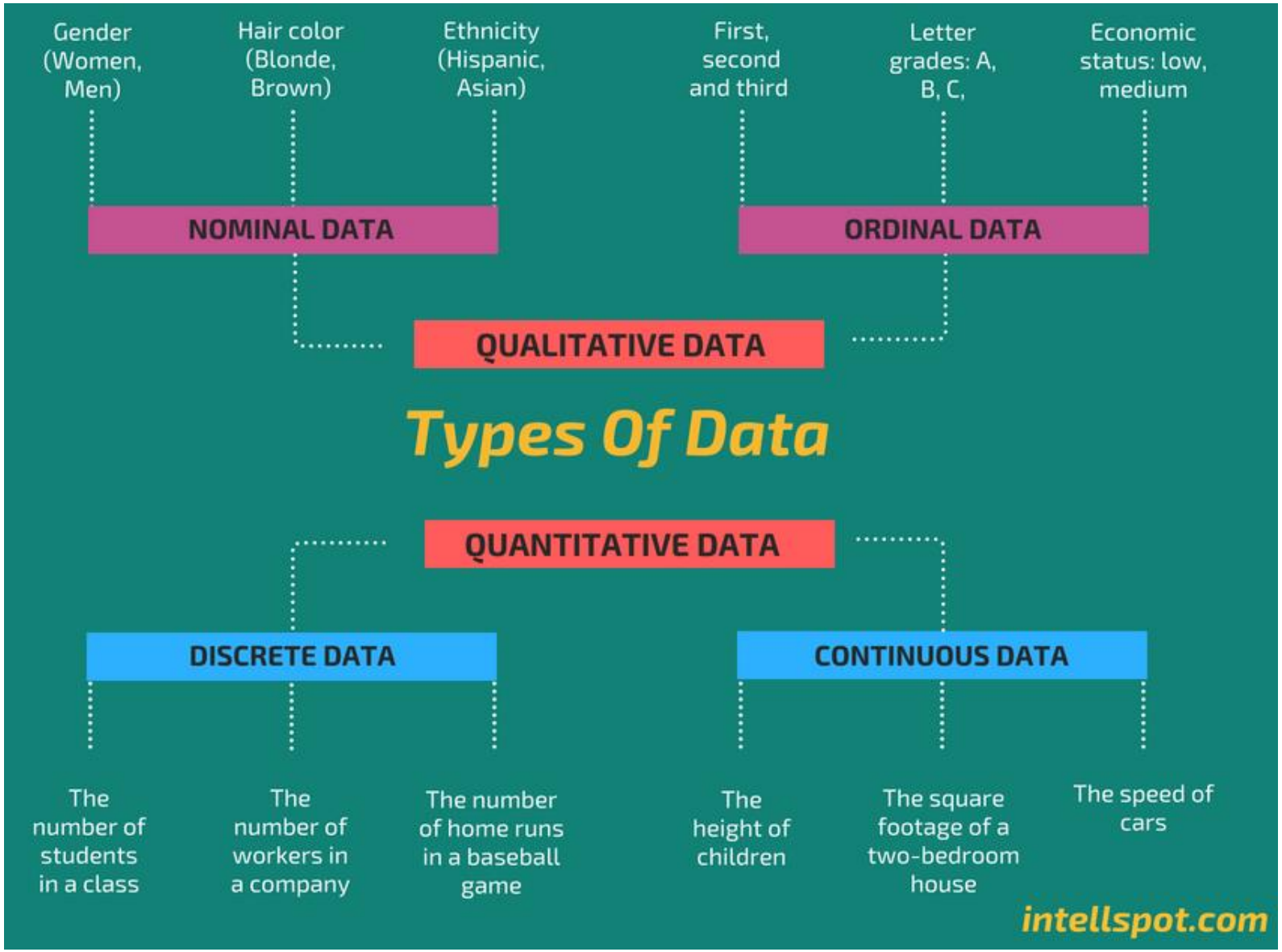
covariance and correlation

- represent the process of computations with intermediate results
- compare results of computations with values obtained in Orange

Basic data types



Data types examples



Compound data

Certain (spatial) structure of elementary data

- **array** – an indexed sequence of data of same type
- **matrix** – a multidimensional lattice of data of same type
- **list** – a sequence of data of any type (including list)
- **set** – no more than one copy
- **bag** – nonnegative integer number of copies
- **record** – a cortege of data of any type – Cartesian product
- **table** – an array of records – a relation

Data visualization

A picture is worth a thousand words

graph of function

scatter plot

line chart

bar chart

pie chart

histogram

bubble chart

box plot

heatmaps

doughnut chart

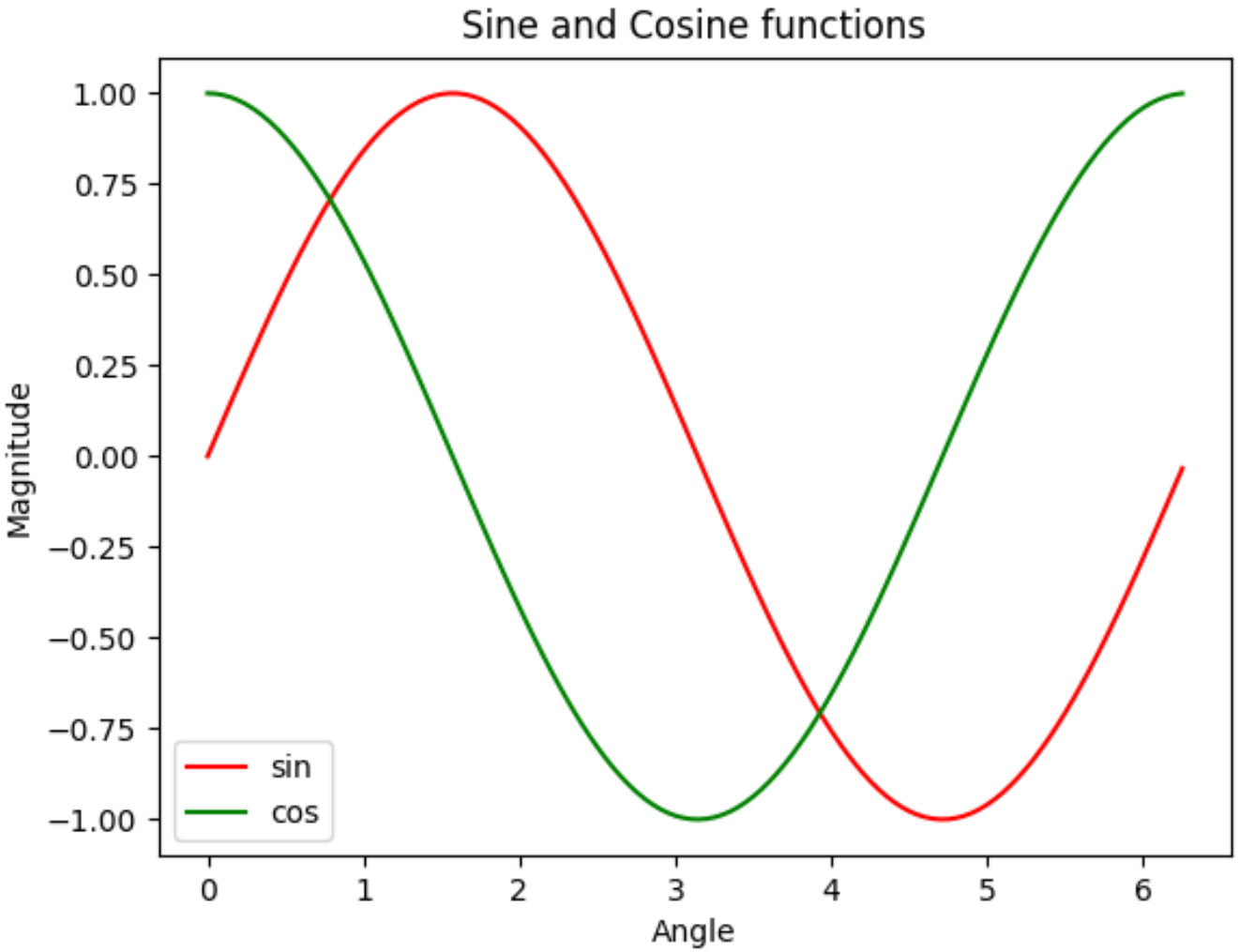
area chart

treemap chart

mosaic display

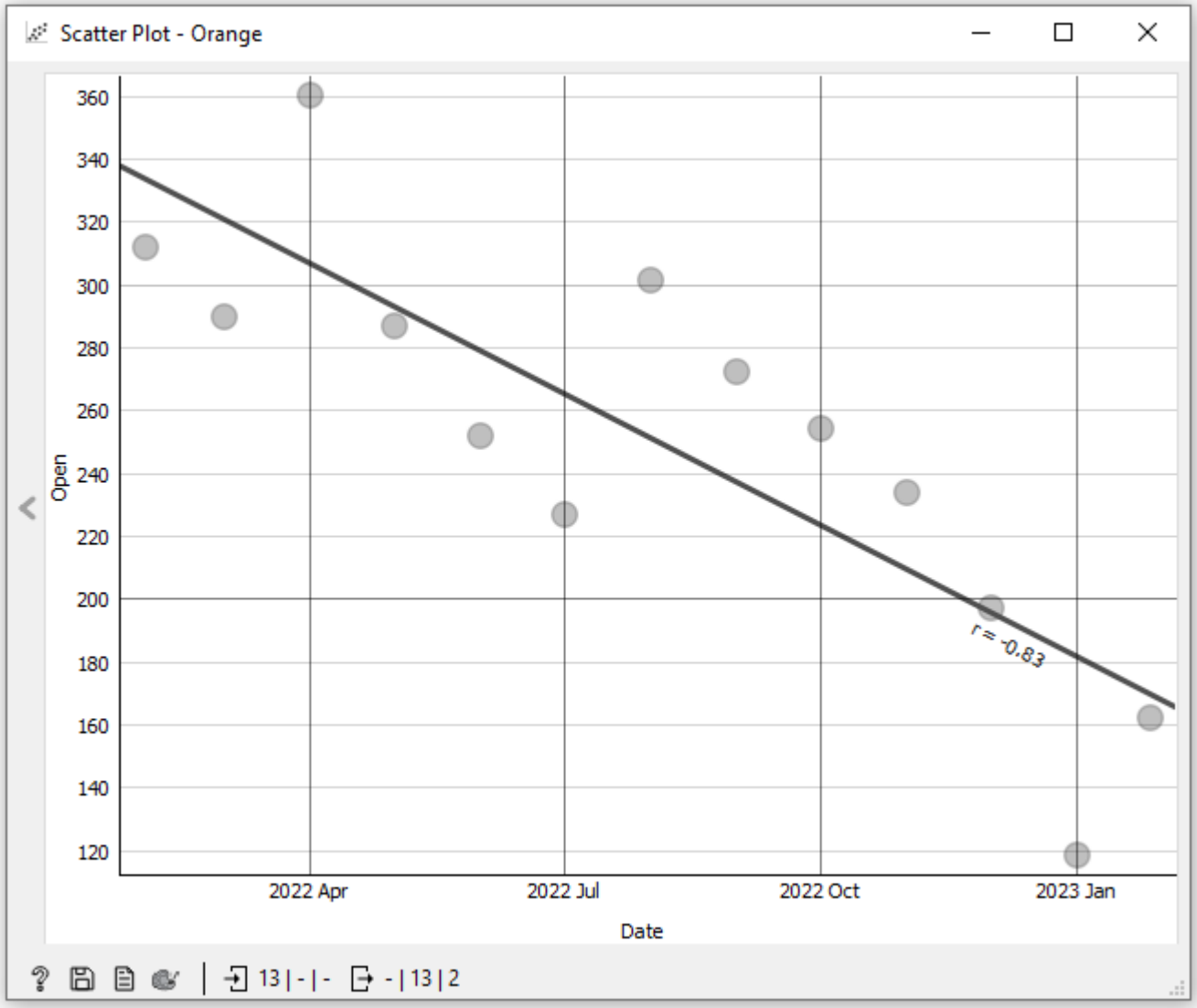
violin plot

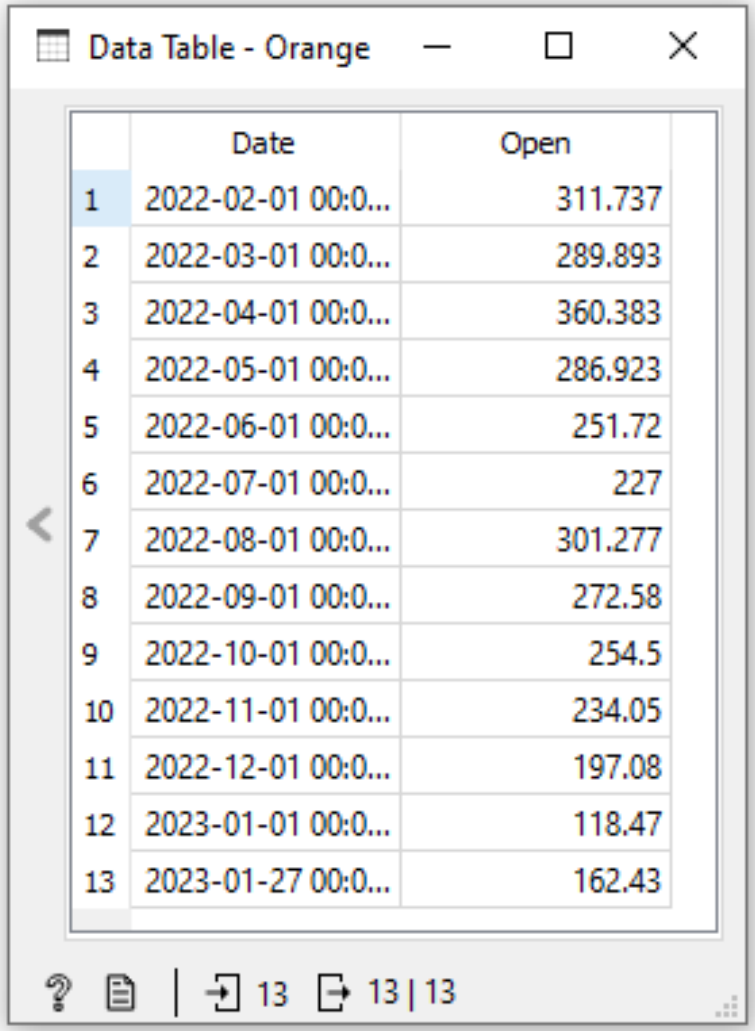
Graphs of functions




```
import matplotlib.pyplot as plt
import numpy as np
import math
X = np.arange(0, math.pi*2, 0.05)
y = np.sin(X)
z = np.cos(X)
plt.plot(X, y, color='r', label='sin')
plt.plot(X, z, color='g', label='cos')
plt.xlabel("Angle")
plt.ylabel("Magnitude")
plt.title("Sine and Cosine
functions")
plt.legend()
plt.show()
```

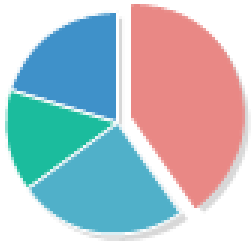
Data series and scatter plots



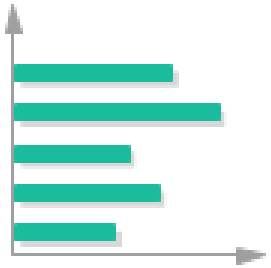


	Date	Open
1	2022-02-01 00:0...	311.737
2	2022-03-01 00:0...	289.893
3	2022-04-01 00:0...	360.383
4	2022-05-01 00:0...	286.923
5	2022-06-01 00:0...	251.72
6	2022-07-01 00:0...	227
7	2022-08-01 00:0...	301.277
8	2022-09-01 00:0...	272.58
9	2022-10-01 00:0...	254.5
10	2022-11-01 00:0...	234.05
11	2022-12-01 00:0...	197.08
12	2023-01-01 00:0...	118.47
13	2023-01-27 00:0...	162.43

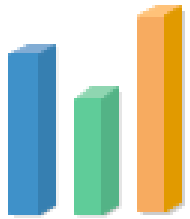
Basic diagrams (charts)



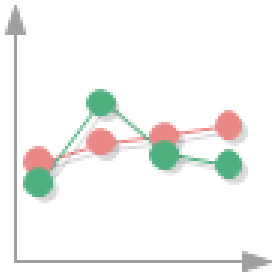
Pie



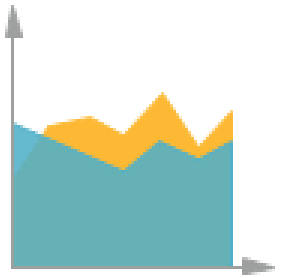
Bar



Column

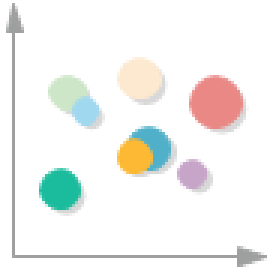


Line

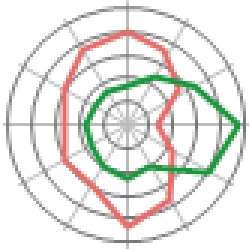


Area

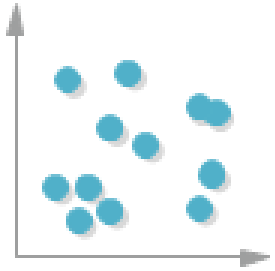
Doughnut



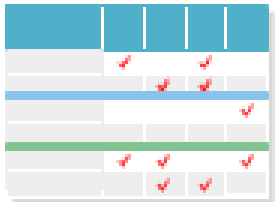
Bubble Chart



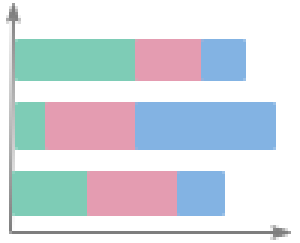
Spider and Radar



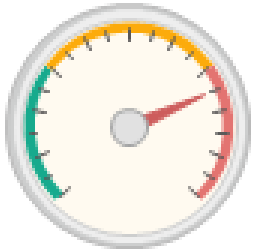
Scatter



Comparison Chart



Stacked bar chart



Gauges

Battle of the Data Science Venn Diagrams


SKEMA
BUSINESS SCHOOL
Introduction to
Artificial Intelligence
Dmitry A. Zaitsev
http://daze.ho.ua
EFMD EMBA
AACSB
EFMD EQUIS
skema
BUSINESS SCHOOL

Lesson 4
Basics of Data Science