DIGITAL CULTURE
ITMO UNIVERSITY

Data Processing Tools

High School of Digital Culture
ITMO University
dc@itmo.ru

Contents

1 Data Processing Tools

Researchers need simple and convenient tools to store and process data since it's quite difficult to keep the data in mind.

Data processing requires software that can manage huge volumes of data and solve data processing tasks. Files without formatting can keep small amounts of data. All you need is to create a document in the Notepad application, for example. Such files are called ASCII files. They save only unformatted characters.

Word processors, such as Word, are used to process unstructured texts. Tables will be the best choice for homogeneous data with an explicit structure. Comma-separated values (CSV) files are used to save tabular data as text. Each

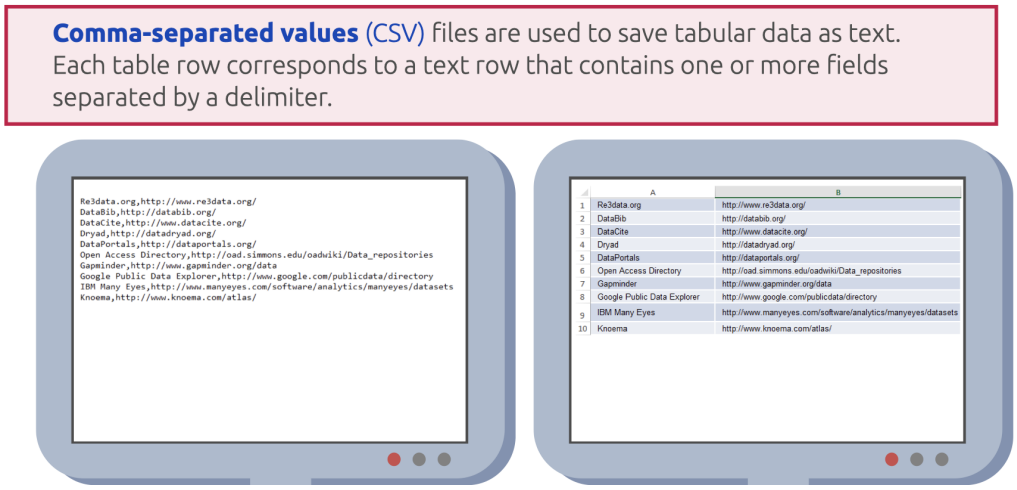


Figure 1: Comma-separated values

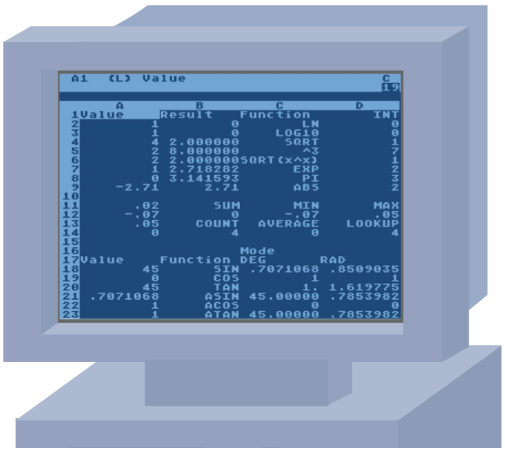
table row corresponds to a text row that contains one or more fields separated by commas or other separators, such as semicolons or tabs. Many apps that support CSV allow selecting the separator character.

Electronic tables were created to store such datasets. They help to calculate values, organize and filter data, as well as to transform, group, analyze, and visualize different types of data.

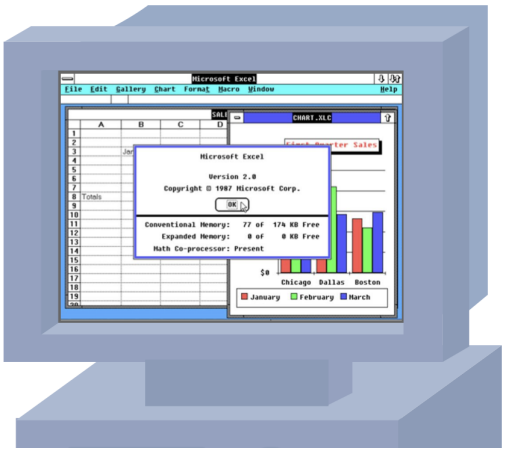
The first electronic-table program called VisiCalc was released in 1979. Over time, the use of electronic tables became one of the most common computer tasks. Microsoft's Excel has been leading the way on electronic-table software for more than 30 years. Microsoft released the first version of Excel in 1985. Excel is a part of the Microsoft Office suite that also includes Word for text, PowerPoint for presentations, Outlook for email, and other useful applications. Excel workbooks can consist of several worksheets. In 2016, one Excel worksheet could span more than 1 million rows and 16 thousand columns. It seems impressive until you realize that the amount of data can exceed this limit and that it's not that easy to work with a worksheet heavily loaded with data.

Electronic Tables

The first electronic-table program (VisiCalc) was released in 1979.

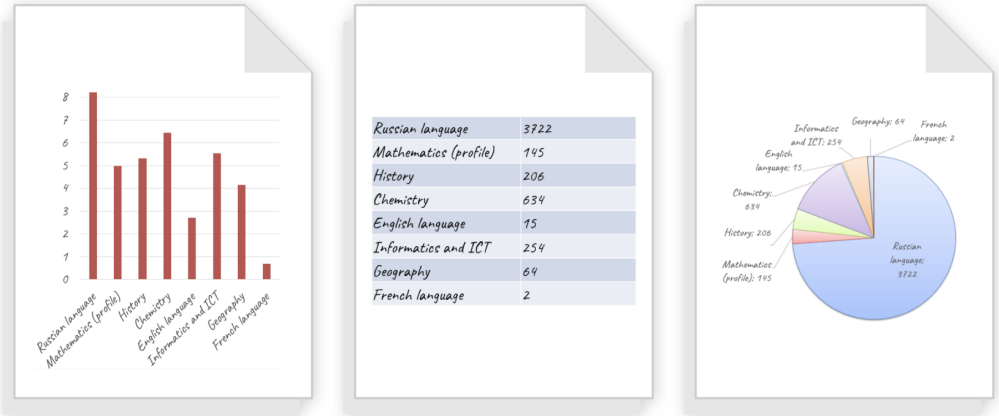


Microsoft released the first version of Excel in 1985.



Sometimes, the amount of data is large and its structure is complex, but the data still should be securely stored and easily accessed by users. Cases like this call for completely different tools, such as database management systems. The choice of tools mainly depends on the data structure. Relational DBMS, such as Oracle or PostgreSQL, process structured data very well. Opposed to them are NoSQL storages developed for unstructured or semi-structured data. An important step

Visual Presentation of Information



in understanding data is data visualization. Visual presentation of information is often more convenient than series of numbers. Excel, Google Sheets, and other worksheets offer a wide range of data visualization products.

However, you may need to use special-purpose suits to analyze data based on unknown parameters or visualize spatial or multidimensional data, for example. Tableau is an example of such data visualization software. It presents users with visualization of the analyzed data. Tableau, which can be called a system for interactive analysis, provides an in-depth and comprehensive analysis of large

amounts of information from a variety of sources. Data can be CSV or text files, PDFs, database files on internal storage devices or in the cloud. Tableau sells several products. To work with open data, one can use Tableau Online. It's a free cloud-based platform with a web interface, but keep in mind that all the solutions will be public and stored on a public server. The next level of data analytics is

Tableau: Interactive Data Visualization Software

Tableau provides an in-depth and comprehensive analysis of large amounts of information by means of its visualization

data mining. It's a systematic and consistent process of finding the underlying patterns in large datasets. Various machine learning algorithms are developed for it.

You don't have to be a programmer to put them in practice. You can simply use such software as Microsoft Azure, RapidMiner, or Weka. These tools allow you to clean and prepare the data, find patterns and anomalies, make predictions, and analyze texts. All these tools are user-friendly. You can simply upload your dataset, select the processing algorithm, apply it, and voilà, the problem is solved.

But what kind of challenges are posed in data analysis, what are the algorithms used to solve them, and how to interpret the results? To answer these questions, you need to learn more. And this course will provide you with this knowledge. But there is a long way to go.

Having studied algorithms for analysis and data processing, you may want to apply them through programming. The best choice then would be Python, a high-level general-purpose programming language that is fairly easy to learn. Python is widely used in education, scientific computing, big data, and machine learning, web and internet development, graphics, GUI, games, and etc.

Thanks to many packages developed for Python programming, it's easy to create an application by combining functions from different packages.

For data preprocessing, we will use electronic tables as our tool.

Ordinary electronic tables are tied to a single computer, which hampers the transfer of data. For example, it's almost impossible to restore the information if a file is accidentally deleted or lost in a computer crash. To avoid this, you can

use cloud storages. Clouds can be accessed from multiple devices. They safely keep all the files and also allow sharing the files with others.

In 2006, Google introduced electronic tables as a part of the Google Docs suite. With Google Sheets, users can create worksheets, co-edit them with others at the same time, and process data from any device connected to the internet.

Google Sheets looks and behaves just like any other worksheet tool. However, since it's an online application, it offers more features than many other tools. For example:

- You can use online tables anywhere; you cannot leave them at home, as opposed to a regular file saved locally.
- Online tables are available on any device in mobile apps for iOS and Android and in browsers.
- Google Docs is free, and it also offers Google Drive, Word, and PowerPoint to collaborate and share files, documents, and presentations online.
- It has all the common functions, and if you know how to use Excel, you'll feel at home in Google Sheets.
- You can download add-ons or create them and even write your code.
- It's online, which allows you to collect data using your table and perform different actions even if the worksheet is not open.

To be fair, other tools, such as Excel Online, also allow working remotely with worksheets. To use Excel Online, you will need to sign in with your Microsoft account. To access Microsoft Online, you have to register on outlook.live.com.

In a browser, you can work with Excel Online and create, view, edit, and save worksheets online.

Office Online combines the most popular Office features and real-time collaborative editing capabilities. It allows users to work together at school, home, or work and share documents, presentations, and worksheets.

Office Online also works seamlessly with the Office applications on the computer, so users can switch between the versions. Use Office Online to collaborate online and see each other's changes in real-time. With Office 365, you can also switch to the full-featured desktop apps, including Word, PowerPoint, and Excel on your macOS or Windows computer.

You may find it easier to complete the tasks in this course with Google Sheets because we will describe its functions in more detail. If you prefer Microsoft Excel or Excel Online, you are free to do so.

2 Goals of Data Visualization

Everything in this world can be counted or expressed in numbers. Our brain is extremely good at abstract thinking, but it cannot process too many numbers at once. That's why large amounts of heterogeneous information require visualization, which helps to transform arrays of data into pictures. Data visualization

Data Visualization

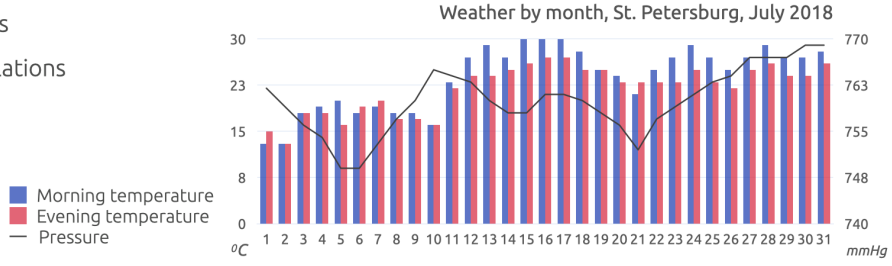
Date	Temperature morning	Temperature evening	Pressure
July 1	13	15	762
July 2	13	13	759
July 3	18	18	756
July 4	19	18	754
July 5	20	16	749
July 6	18	19	749
July 7	19	20	753
July 8	18	17	757
July 9	18	17	760
July 10	16	16	765
July 11	23	22	764
July 12	27	24	763
July 13	29	24	760
July 14	27	25	758
July 15	30	26	758
July 16	30	27	761
July 17	30	27	761
July 18	28	25	760
July 19	25	25	758
July 20	24	23	756
July 21	21	23	752
July 22	25	23	757

translates abstract information into graphics. It helps to find patterns, trends, and correlations that may go unnoticed in ordinary reports and tables (both on paper and in electronic form). Studies show that the human brain processes images 60-thousand times faster than text, and 90% of information transmitted to the brain is visual. Try to quickly find the minimum and maximum in the table

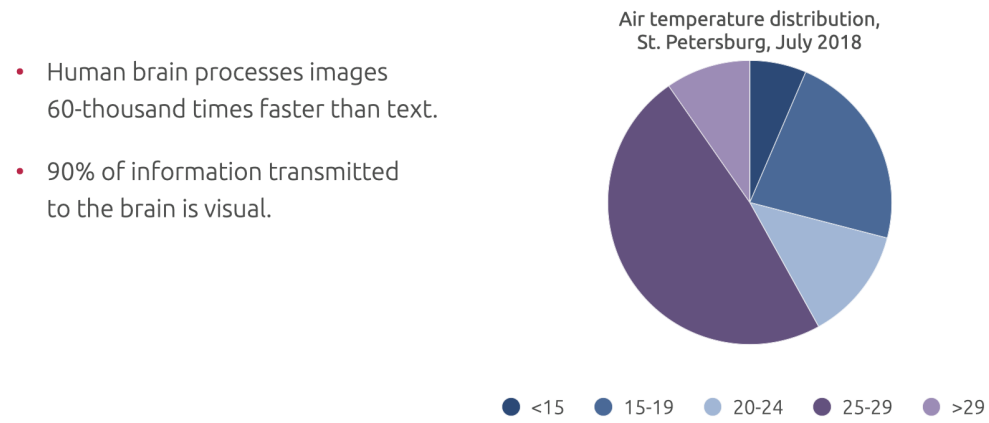
Data visualization translates abstract information into graphics.

Data visualization helps to find:

- patterns
- trends
- correlations

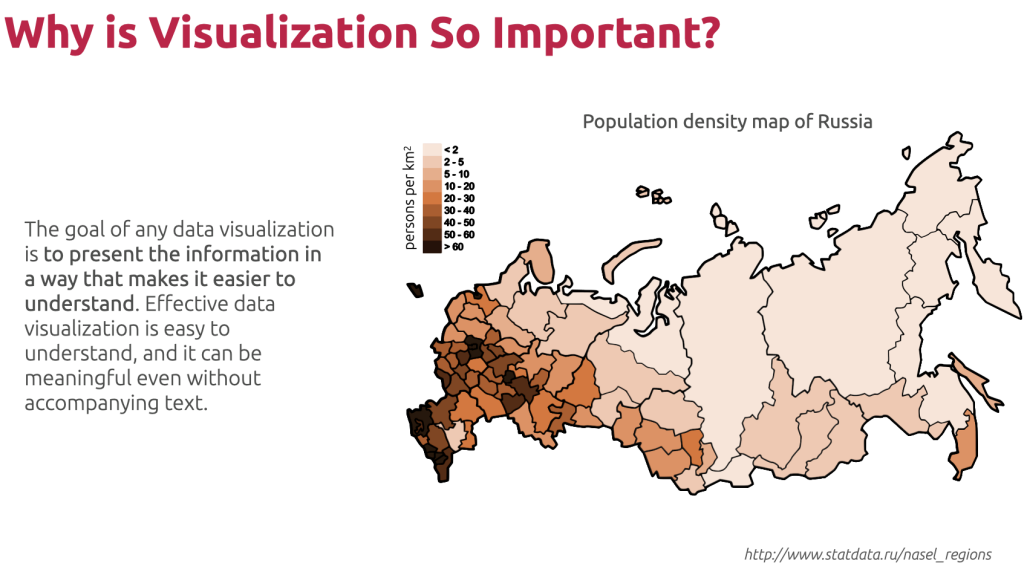


cells. Repeat the same with the graph. Pictorial instructions are clearer than a plain text. Visual or graphical representation of data is the use of graphs, charts, schemes, maps, and so on. Earlier, data visualization had a secondary role in data analysis. Nowadays, more and more studies prove its independence. Why is visualization so important?



- Human brain processes images 60-thousand times faster than text.
- 90% of information transmitted to the brain is visual.

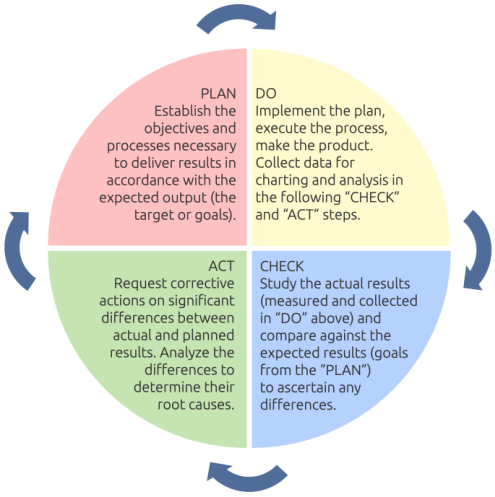
Data visualization presents information in a way that makes it easier to conclude instantly. Besides, effective data visualization is easy to understand, and it can be meaningful even without an accompanying text. Those who make



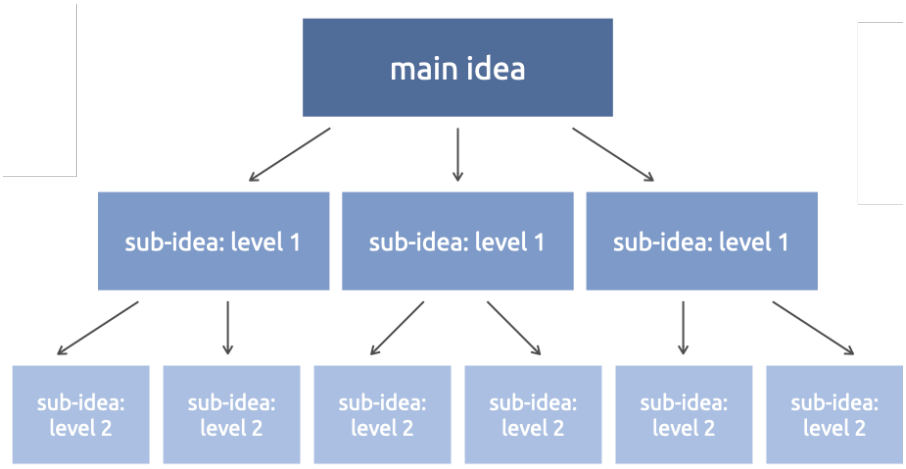
decisions usually don't bother to delve into the endless series of data. They need something to make a quick but good decision and assess the situation without going into details. That is why the quality of data visualization is crucial for decision-making.

There are several data visualization tasks. The first one is idea illustration. Idea illustration is aimed at training and clarification. It replaces a detailed description. Its typical examples are organizational charts and business process diagrams.

Data visualization often helps to generate ideas when the goal is to solve a problem or find out the truth. It is also used in brainstorming. A typical example is a mind map.



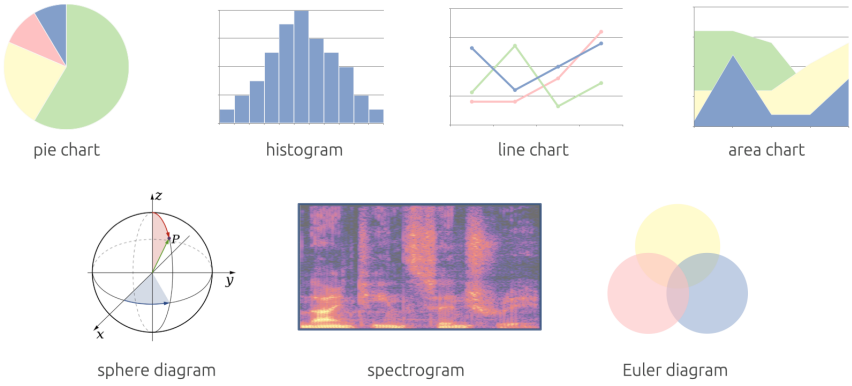
As has been noted, data visualization is an independent type of analysis. This type of analysis can be called a visual study. It's used to find patterns



and better understand them. This includes complex multivariate representations. Data visualization is called routine when a message is put in context. It's used for reports and presentations for management and business partners. This is the way to inform others about things.

Data visualization varies a lot. It represents quantitative information as schemes. This group includes varieties of pie charts, line charts, histograms, spectrograms, tables, and different scatter plots. In visualization, data can be transformed in a way that enhances the visual perception and helps in analysis, for example, maps and polar charts, timelines and graphs with parallel axes, and the Euler diagrams.

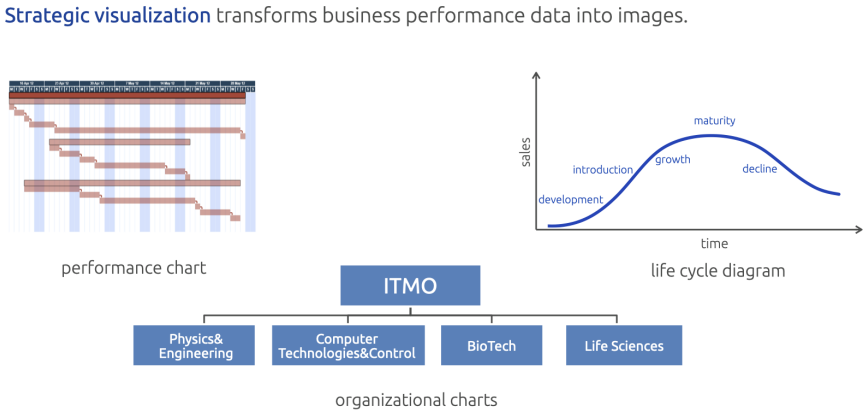
Visualization Types



Concept visualization helps to develop complex ideas and plans using concept maps, Gantt charts, graphs with shortest paths graphs, and other similar types of charts.

Strategic visualization transforms business performance data into images. It uses all kinds of performance charts, lifecycle diagrams, and organizational charts.

Strategic Visualization



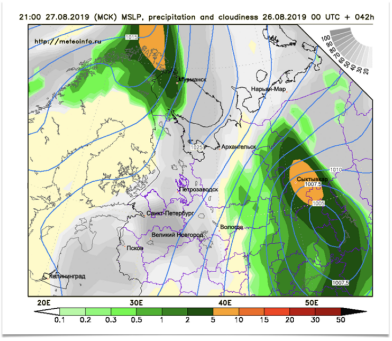
Metaphorical visualization graphically arranges structured information in pyramids, trees, and maps. A subway map is an example of metaphorical visualization. Combined visualization puts several complex charts into one just like on a weather map.

Visualization techniques can do the following:

- Provide visuals that are easy to understand
- Describe dataset regularities without excessive details
- Compress information
- Identify data gaps

Combined Visualization

Combined visualization puts several complex charts into one.



- Detect noise and outliers in data

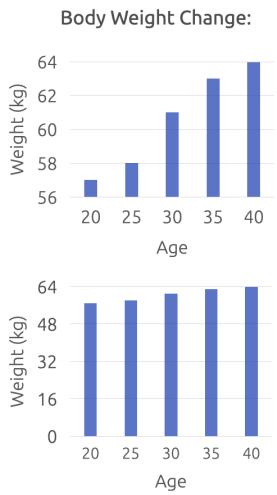
Benjamin Disraeli said, 'Figures often beguile me, particularly when I have the arranging of them myself.' There's also a saying popularized by Mark Twain, 'There are three kinds of lies: lies, damned lies, and statistics.'

Well, just like the correct visualization helps to analyze data, the incorrect visualization can be very misleading and create a distorted view of the data.

For example, consider the weight change data in the table. Let's create a bar chart using the data in the table. It looks like a rapid weight gain.

Correctness of Visualization

Age	Weight (kg)
20	57
25	58
30	61
35	63
40	64



Now let's create another chart for the same dataset. The second chart shows gradual weight changes.

The impression differs because of the change in the initial point of the vertical axis. The counting starts from 50 in the first case and from 0 in the second.

This example clearly illustrates how different visualization of the same dataset can change the impression.

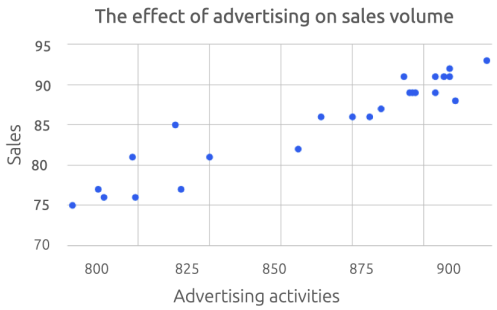
3 Visualization Techniques

Given the number of dimensions, visualization techniques are divided into two groups:

- Visualization techniques for one, two, and three dimensions
- Visualization techniques for more than three dimensions

Data visualization is aimed at transmitting information for achieving the purpose of communication. For example, researchers may want to find dependencies

Relationships in data are the relations and dependencies between data items.

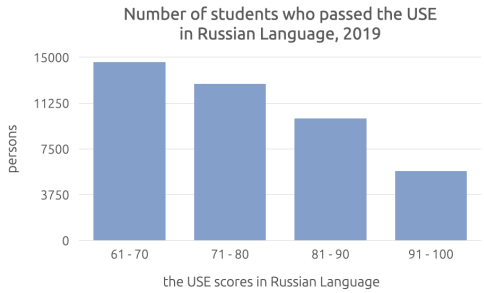


in data, visualize data distribution and composition, or compare data.

Relationships in data are the relations and dependencies between data items. Relationships help to identify the dependencies between variables. If the main idea in data can be described with such words as `refers to` or `decreases/increases when`, it makes sense to find the relationships in data.

Distribution displays the number of observations in which the feature takes values within the given intervals. In this case, the main idea is expressed

Distribution displays the number of observations in which the feature takes values within the given intervals.



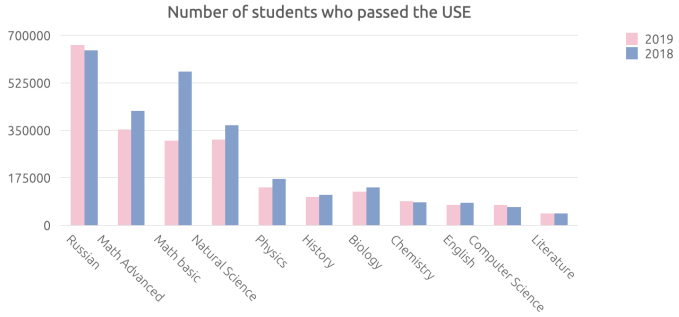
with such phrases as `ranges from x to y`, `concentration`, `frequency`, and `distribution`.

Data composition is combining data to analyze the overall picture or compare the elements that represent the percent of the whole. The key phrases for the composition are `amounted to x%`, `rate`, `percent of the whole`.

Data composition is combining data to analyze the overall picture or compare the elements that represent the percent of the whole.

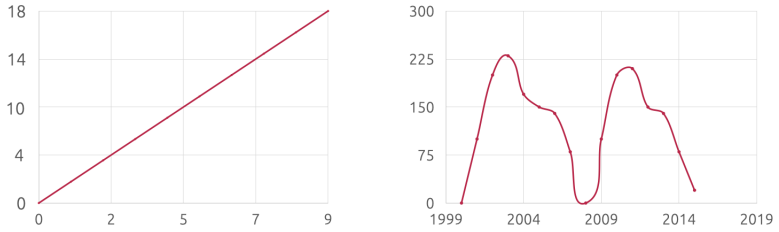
Data comparison combines data to compare the features and understand how objects relate to each other. It's also used to compare the components that change over time. When it comes to data comparison, the key phrases are `more than/less than`, `equals`, `changes`, `increases/decreases`.

Data comparison combines data to compare the features and understand how objects relate to each other.

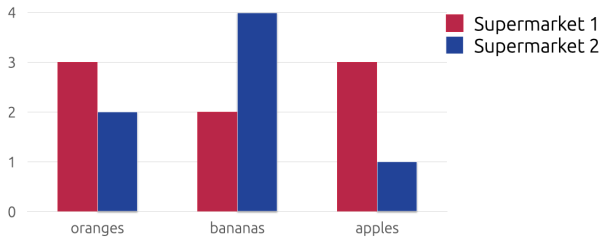


Having defined the visualization goal, it's time to define the data type. Data types and structures can be very heterogeneous. The most common are continuous numerical and time-series data, discrete data, geographic, and logical data. Continuous numerical data contains information about the dependence of one numeric value on another, for example, graphs of functions such as $y = 2x$. Continuous time-series data contains information about events within a time interval, such as a graph of the daily temperature. Discrete data can contain dependencies of categorical variables, for example, a graph of sales by stores. Geospatial data typically combines information about location, geology, and other geographic features; an ordinary map is a good example of geospatial data.

Continuous data contains information about the dependence of one numeric value on another.

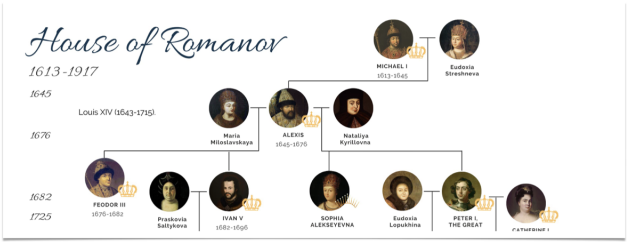


Discrete data is quantitative data expressed by a limited set of values (usually integers).



Logical data shows the logical position of items relative to each other, for example, on a family tree.

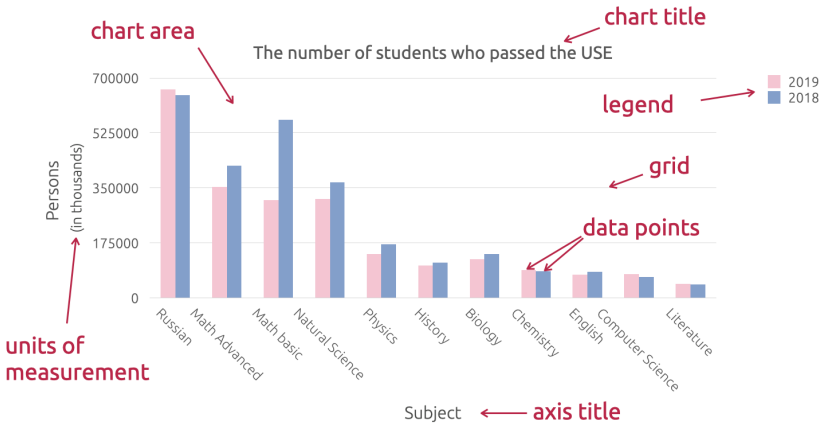
Logical data shows the logical position of items relative to each other.



https://https://www.palaces-of-europe.com

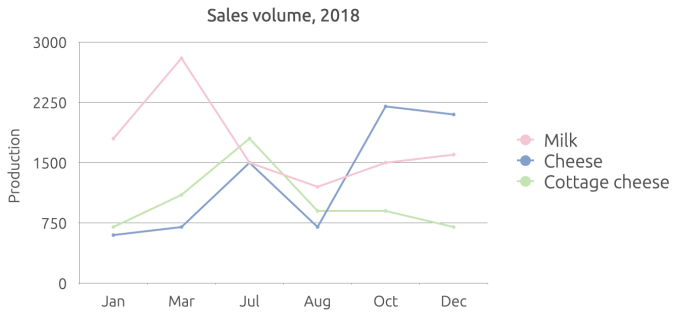
A chart consists of different elements, including names of axes, units of measurement, title, legend, and other things. Such elements make charts easier to understand.

Chart Elements



Line charts are the most common charts. Line charts connect a set of data points with a continuous line. Line charts are used to display a quantitative

Line Chart



- is used to display **trends over a continuous interval** and to show **relationships between categories**

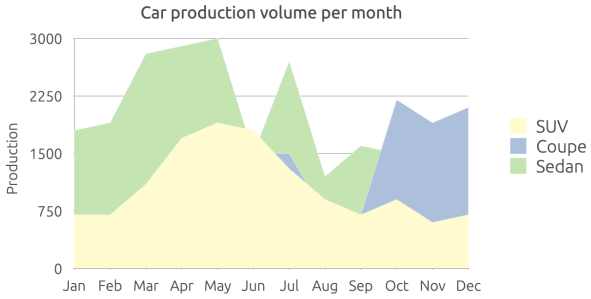
value over a continuous interval. They are frequently used to show trends and relationships between categories (when grouped with other lines). Line charts also help to see a bigger picture for the period of time and corresponding changes.

Use different colors when multiple lines are plotted in a line chart and explain in the legend what the colors mean.

Don't overload the chart with too much information. 4 or 5 is the optimal number of different data types or categories, otherwise, it's better to create more charts.

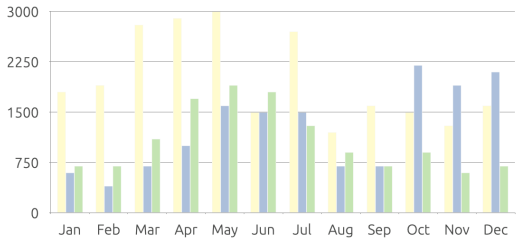
Area charts are based on line charts. The area between the axis and the line is usually emphasized with colors, textures, and strokes. Area charts are often used to compare two or more data series. Use area charts and stacked bar charts to display the relative value that each value contributes to the total by time or category.

Area Chart



Bar charts emphasize different categories with colors and show the amounts for each category. Two ways to display categories are vertical and horizontal.

Bar Chart

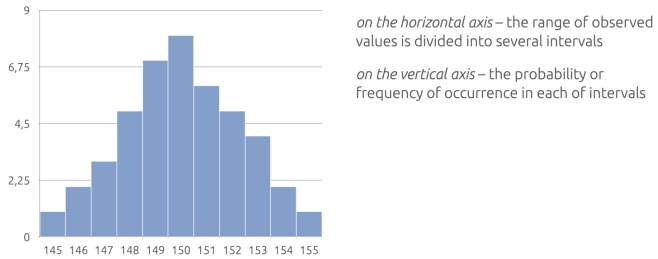


- emphasizes different categories with colors and shows **the amounts for each category**

Categories are shown in colors explained in the legend. Histograms show the quantitative ratios of a measured object as rectangles. The width of rectangles usually remains the same for ease of interpretation, but their height reflects the ratio of the displayed parameter.

Histograms are used in statistics to visualize the probability distribution of values of a random variable. The range of observed values is divided into several

Histogram



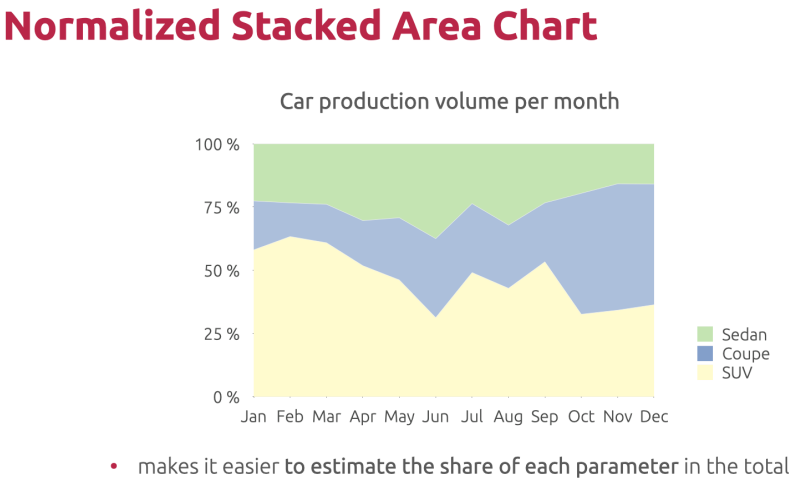
- is used in statistics to visualize **the probability distribution of values of a random variable**

intervals and plotted on the horizontal axis of the histogram, and the probability or frequency of occurrence in each of them is plotted on the vertical axis. A

rectangle reflects the values of the measured objects for the interval on which it is based.

Line charts, area charts, and histograms can contain multiple values in a single argument for a single category, and those values also add up to the total. To visualize and compare totals, a stacked area chart is often used. It compares not only the data series but also the totals.

A normalized stacked area chart is similar to the previous one, but the values in a stacked area chart are normalized (given in percentages). The total is always 100%. This chart makes it easier to estimate the share of each parameter in the total.



The tag cloud is a visualization of word frequency in a text. Colors can be used to divide words into categories (by the frequency of use). The tag cloud doesn't provide accurate values, but it is simple to understand.

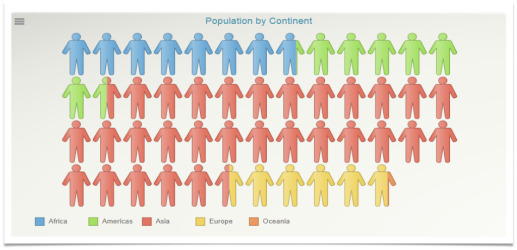
Pictogram charts use icons to give a more engaging overall view of small sets of discrete data. Icons usually represent the subject or category of data, for example, data on population would use icons of people. All icons should have the same size. Fractions are usually represented as a part of the icon. Each icon can be a unit or any number of units (for example, each icon represents 10).

As you can see, the data visualization options are numerous. However, certain data transformations are sometimes required sometimes before visualization.

Pie charts help to show proportions and percentages between categories by dividing a circle into slices. The arc length of each slice is proportional to the category it represents, and the entire circle is the sum of the entire dataset, which is equal to 100%. Pie charts are ideal for showing percentage or proportional data. The main disadvantage of pie charts is that they don't go well with more than 3-5 values because as the number of displayed values increases, the size of each segment or slice becomes smaller. This explains why they are unsuitable for large amounts of data.

For ease of comparison, the slices are arranged in descending order of the arc

Pictogram

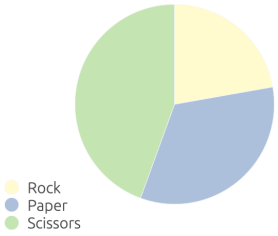


Icons:

- represent the subject or category of data
- have the same size
- fraction is represented as a part of the icon

- gives an overall view of small sets of discrete data

Pie Chart

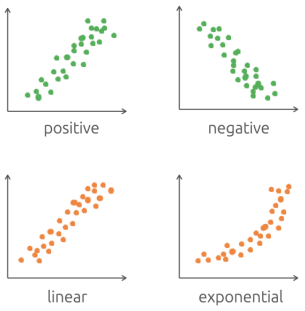


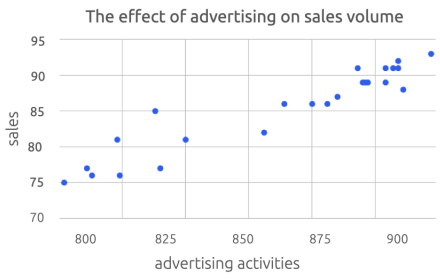
Raw Data			
Rock	Paper	Scissors	*Total*
2	3	4	*9*
Percentage			
(2/9)*100% = 22%	(3/9)*100% = 33%	(4/9)*100% = 44%	*100%*
Degrees for each slide of the pie			
(2/9)*360° = 80°	(3/9)*360° = 120°	(4/9)*360° = 160°	*360°*

- shows **proportions and percentages** between categories by dividing a circle into slices

lengths. A scatter plot (scatter chart, scatter graph, scatter diagram) uses Cartesian coordinates to display values for two variables as points in the plane. The display of variables on each axis helps to make assumptions about the relationship or correlation between two variables.

Scatter Plot

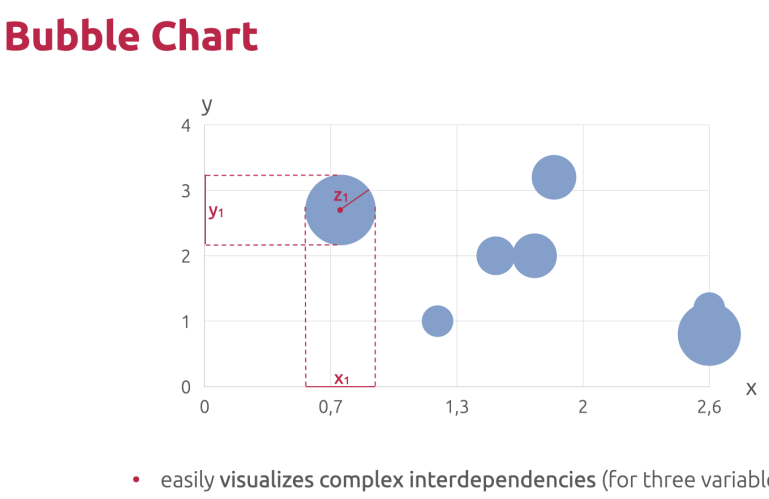




- helps **to make assumptions** about the relationship or correlation between two variables

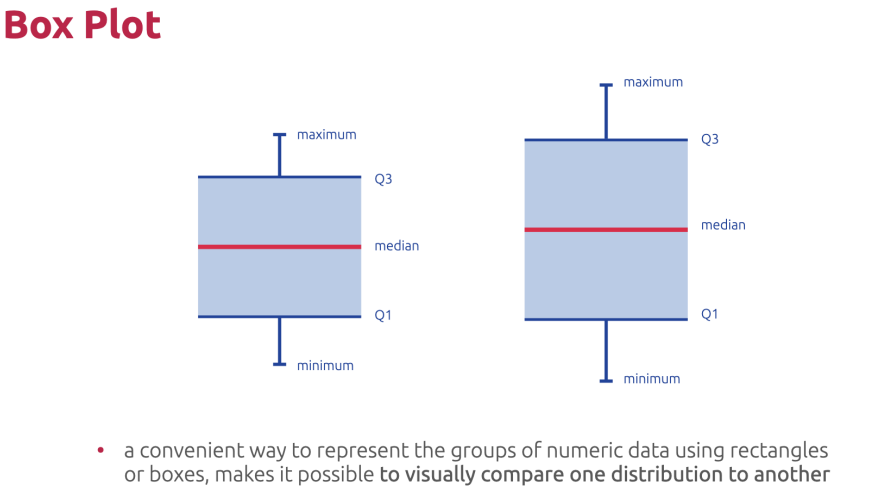
Bubble charts are very similar to scatter plots because the position of each bubble is defined by two coordinates. In addition, the size of the circle at each point represents an additional dimension. Because of this, bubble charts make it possible to compare three variables. They easily visualize complex interdependen-

cies that are invisible in charts for two variables. It's also possible to use colors



to distinguish categories or represent an additional variable.

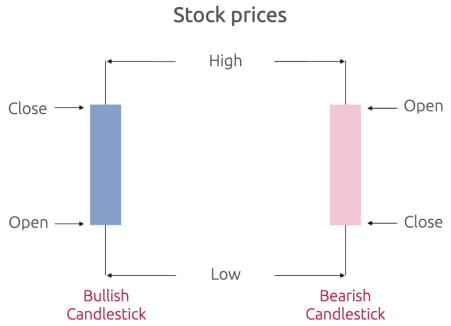
But let's continue discussing the types of data visualization. One of such types is called a box plot. A box plot is a convenient way to represent the groups of numeric data using rectangles or boxes. The box borders are the first and third quartiles, and the line in the middle is a median. The lines extending from the boxes are called whiskers. The ends of the whiskers are the minimum and the maximum, exclusive of outliers that show the variability outside the upper and lower quartiles. Whisker boxes can be drawn vertically or horizontally.



Some of these boxes are drawn side by side to compare one distribution to another. They can be placed horizontally or vertically. The interquartile range is a measure of statistical dispersion and data asymmetry.

A candlestick chart is a tool for visualization and analysis of price movements of a security, derivative, currency, and so on. Each candlestick typically shows important information over the fixed interval, such as open and close, high and low. The color represents the candles with a higher open and a lower close and vice versa.

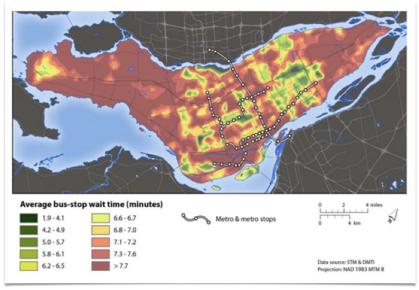
Candlestick Chart



- a tool for visualization and analysis of **price movements of a security, currency**, etc.

Heat tables visualize data by measuring its variations in color. They make the important variables appear in color as a function of two other variables.

Heat Map



- **makes the important variables appear in color** as a function of two other variables

Population density. The simplest example of a color map is a map of a region showing the population density in color. You can order the regions by the corresponding population density, or you can visualize the data on a heat map that illustrates the necessary information.

A heat map for taxi services. Here's an example of heat map use. Taxi services use heat maps to show the drivers the areas of high demand. A heat map of the city indicates in red the areas with the highest numbers of taxi orders within the last hour.

Heat maps in tables. Heat maps make it easy to analyze large amounts of data, and they are not necessarily applied to geographic maps. Look how an ordinary table changes when the colors are mapped onto it. With a heat map, it's easier to analyze data.

Researchers use special visualization techniques when working with more than three dimensions.

The most widely used techniques for multivariate data representations are parallel coordinates, radar charts, and Chernoff faces. In parallel coordinates, a

chart is a union of two-dimensional projections of a multivariate dataset. Parallel projections may be horizontal or vertical.

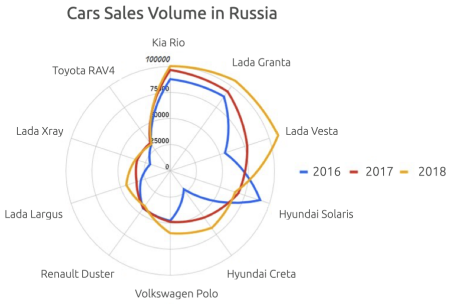
Parallel Coordinates

- is a union of two-dimensional projections of a multivariate dataset

A common way to represent stock market data is a parallel coordinates chart. One projection shows the time and price of a deal, and the second time displays time and amount. It's possible to extend the plot with two projections. The time and number of purchase orders and the time and number of sales orders.

Radar charts are used to compare the values of multiple qualitative variables (if they are measurable). Each variable has an axis that begins at the center. All axes are radial and equally distanced from each other. The lines of the web are

Radar Chart

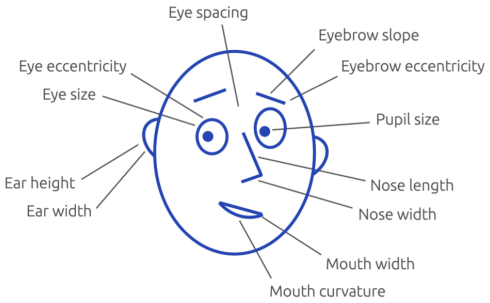


- compares the values of multiple (3 and more) qualitative variables **regarding the central point**

used as guiding lines, the lines of one axis are connected to the lines of another axis. Each value of the variable is drawn along its separate axis. All the plotted values are connected to form a polygon. Each observation has its polygon.

The main idea of visualization with the Chernoff faces method is to display the values of variables in the shape of a human face. Each observation has its face. Each face has such parts as eyes, ears, mouth, and nose. Those parts represent relative values of the variables by their shape, size, placement, and orientation. This method is based on the fact that humans easily recognize faces and notice the changes intuitively.

The Chernoff Faces



- displays the values of variables **in the shape of a human face**

The Chernoff faces method was applied to urban indicators in Los Angeles. Chernoff faces showed the unemployment rates, income levels, white population proportion, and other indicators.