



# Data Processing and Analysis. Working with Time Series.

High School of Digital Culture

ITMO University

[dc@itmo.ru](mailto:dc@itmo.ru)

---

# Contents

<b>1 Descriptive Statistics</b>	<b>2</b>
1.1 Range and Interquartile Range . . . . .	6
<b>2 Data Transformation</b>	<b>10</b>
2.1 Common Transformations . . . . .	11
2.2 Choosing the Right Transformation . . . . .	13
2.3 Units of Measurement and Inverse Transformations . . . . .	14
<b>3 Data Normalization</b>	<b>15</b>
3.1 Classes of Numerical Parameters . . . . .	16
3.2 Normalization of Unipolar Parameters . . . . .	16
3.3 Normalization of Bipolar Parameters . . . . .	17
3.4 Parameters of Students . . . . .	18
<b>4 Objective Function</b>	<b>22</b>
<b>5 Time Series</b>	<b>28</b>
5.1 Time Series Analysis . . . . .	28
5.2 Quality Metrics . . . . .	30
5.3 Time Series Smoothing . . . . .	32
5.4 Decomposing Time Series Data into Trends . . . . .	35
5.5 Drawing Trend Lines . . . . .	37

## 1 Descriptive Statistics

**Descriptive statistics** is used to systematize data from different sources. Descriptive statistics is the field of statistics that studies systematization methods, descriptions, and representation of the main characteristics of data. Descriptive statistics is used to analyze a particular sample and doesn't aim to characterize an entire population. That's why the concepts and definitions in descriptive statistics sometimes differ from those in inductive statistics. Descriptive statistics includes data collection, categorization, generalization, and representation. Descriptive statistics is actively used for exploratory data analysis and, in some cases, it turns out to be sufficient for a complete data analysis. Let's consider the main types of descriptive statistics and their application in practice.

A **measure of central tendency** in descriptive statistics is a number that describes all the values of a variable in a dataset. The measure of central tendency describes a whole set of data with a single value. The following characteristics of the central tendency are distinguished:

- Mean
- Mode
- Median

The **mean** of a variable is the sum of all the values of the variable divided by the number of values. It's designated by  $\bar{x}$  or **Mean**.

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

The mean is calculated only for numerical values. When it comes to dichotomous data, the mean takes 0 and 1. Each dataset has only one mean. Let's practice by calculating the mean of a student's grades. The university student got the following grades:

5, 4, 2, 5, 4, 3, 3, 4, 5, 3, 5, 5, 5, 2, 5

The mean is the sum of all the values divided by their number:

$$\text{Mean} = \frac{58}{14} \approx 4.14$$

We can also calculate the mean for dichotomous data. If two values of the variable are 0 and 1, the mean for such data indicates the share of ones in the sample. For example, for the following data:

1, 0, 0, 0, 1, 1, 1, 0, 0, 0,

40% of the sample values take the value of one.

$$\text{Mean} = \frac{4}{10} = 0.4$$

The **mode** is the value of the variable which occurs most frequently. It's designated by **Mo**. The mode can be found for the data of any measurement scale. There can be more than one mode. In this case, it's a bimodal or multimodal distribution of the variable values. If all the values in a dataset appear with the same frequency, there's no mode or all the values are the mode. The figure shows the example of finding the mode of the student's grades. The grade that appears most often is 5.

Example (calculating the mode of the student's grades)

grade	frequency
5	7
4	3
3	3
2	1

$$\text{Mo} = 5$$

The next characteristic of central tendency is called a median. The **median** is such a value of the feature that a half of the values are more than this value and the other half is less. To define the median, we will use the concept of a variational series.

Median is based on the concept a variational series

The **variational series** is the data arranged in ascending or descending order of variable values.

Example (variational series of the students' grades)

Input data:

5, 4, 5, 4, 3, 3, 4, 5, 3, 5, 5, 5, 2, 5

Variational series:

- in ascending order:

2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5

- in descending order:

5, 5, 5, 5, 5, 5, 4, 4, 4, 3, 3, 3, 2



The **variational series** is the data arranged in ascending or descending order of variable values. A series is called variational because it contains variations of the feature values. Let's look at the example of variational series of the students' grades (the input data is given in the figure above). Based on it, we create two variational series. The first row is sorted in ascending order, and the second in descending one.

At this point, we can define the concept of a median. The median (designated by  $Me$ ) is the value corresponding to the element in the middle of the variational series. The element in the middle is different for an even and odd number of variable values.

When the number of elements in a dataset is odd, the median is unique. However, for the even number of elements in a dataset, the median can be any number on the interval of middle numbers. More often, we find the median by calculating the half-sum of the middle elements.

For a dataset of  $n$  odd values, the middle element is  $\frac{n+1}{2}$ , and for the even number of  $n$ , the median is the arithmetic mean of two middle elements  $\frac{n}{2}$  and  $\frac{n}{2} + 1$ . The median is also found for numerical and ordinal data. Only one median exists for a dataset.

Take a look at the example of calculating the median of the student's grades. Here's the variational series of grades sorted in ascending order. It contains 14 elements. Therefore, the median is the average of the 7th and 8th elements, and

Example (calculating the median of the student's grades)

2, 3, 3, 3, 4, 4, **4, 5, 5, 5, 5, 5, 5**

14 elements

Median is the arithmetic mean of the 7-th and the 8-th elements:  $(4 + 5)/2$

$Me = 4.5$



it equals 4.5.

Another example is finding the median of the values of wind force at sea on the Beaufort scale. The following observations of the wind force on the Beaufort scale are available. We will use them to form a variational series of variable values sorted in ascending order. There are 13 elements. Therefore, the seventh element is the median.

Example (finding the median of the values of wind force at sea on the Beaufort scale)

0, 2, 2, 1, 1, 3, 3, 1, 1, 0, 0, 1, 2

Variational series (in ascending order)

0, 0, 0, 1, 1, 1, **1, 1, 2, 2, 2, 3, 3**

$Me = 1$  – light air

	0 Calm	4 Moderate breeze	8 Gale
	1 Light air	5 Fresh breeze	9 Strong gale
	2 Light breeze	6 Strong breeze	10 Storm
	3 Gentle breeze	7 High wind	11 Violent storm
			12 Hurricane

Let's summarize the issues we have discussed so far. We've discussed 3 characteristics of the central tendency. The next table shows the characteristics applicable to different measurement scales.

So, which characteristic is better? Which one do we apply if there is a choice? At first glance, it seems that the mean is the most capacious and widely

Characteristics of the Central Tendency	Nominal Data	Ordinal Data	Interval Data	Ratio Data
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓



used characteristic. It's true in terms of prominence but not in terms of use. Here's an often-used example.

Imagine a village that has 50 inhabitants. Among them, 49 people are residents with a monthly income of 1 thousand rubles, and one resident is a prosperous farmer with an income of 451 thousand rubles. So, what is the average income of the villagers? It equals 10 thousand rubles. This number doesn't fairly reflect the income of the villagers. In this case, it would be more appropriate to find the mode or median as a measure of central tendency (the mode or median are equal to 1 thousand rubles). However, one number cannot represent the income in this village.

#### Rural residents income

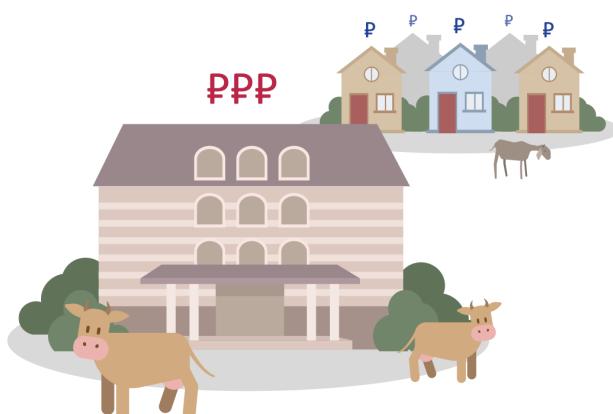
451000 rub. – 1 person

1000 rub. – 49 persons

**Mean** = 10000 rub.

**Mo** = 1000 rub.

**Me** = 1000 rub.



## 1.1 Range and Interquartile Range

The **measure of central tendency** is just a number used to describe a typical value in the studied sample. It doesn't show how diverse the data in the sample is. That's why a range and quartile (or interquartile) range were introduced.

The **range** is the difference between the highest and lowest values in the dataset. For the shown dataset of the student's grades, the range equals 3.

**The range** is the difference between the highest and lowest values in the dataset:

$$R = x_{\max} - x_{\min} .$$

Example (calculating the range of the student's grades)

5, 4, 2, 5, 4, 3, 3, 4, 5, 3, 5, 5, 5, 2, 5

$$R = 5 - 2 = 3$$



Another characteristic of the data range in the sample is called the **interquartile range**. It's based on the concept of quartiles. Quartiles are the values  $Q_1, Q_2, Q_3$  that divide the variational series into four equal parts.

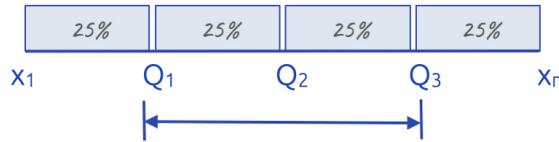
**Quartiles** are the values  $Q_1, Q_2, Q_3$ , that divide the variational series into four equal parts.

The screenshot shows a Microsoft Excel spreadsheet with a table of student grades in columns A and B. Cell C1 contains the formula `=QUARTILE.INC(A1:A14;1)` and cell B1 contains the formula `=QUARTILE.INC(A1:A14;3)`. The data in column A is: 11, 12, 2, 24, 34, 2, 33, 46, 32, 34, 22, 33, 44, 22. The data in column B is: 14,5 × ;

The second quartile  $Q_2$  is the same as the median.  $Q_1$  is the median for the values that are lower than  $Q_2$ .  $Q_3$  is the median for the values that are higher than  $Q_2$ . There are several slightly different ways to find the values of the quartiles. For example, when finding the first  $Q_1$  and the third quartile  $Q_3$ , we can include or exclude the median (in this case, "less" and "more" mean "strictly lower" and "strictly higher" or "not strictly lower" and "not strictly higher"). That's why many tools offer two variants of a quartile function to exclude or include the median. They are called QUARTILE.EXC() and QUARTILE.INC().

The **interquartile range** is the difference between the third and the first quartile, which is calculated by the formula:

$$IQR = Q_3 - Q_1$$

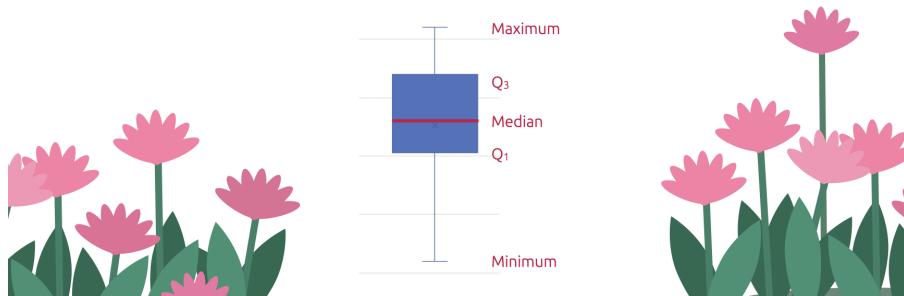


What is the fundamental difference between the range and interquartile range? The range is a very simple and rough measure of variability because only the smallest and greatest values of the variable are used to calculate the range. When calculating the interquartile range, the extreme values located outside the first and the third quartiles are ignored. 50% of the data falls between the third and first quartiles.

When running the exploratory data analysis, the so-called **box plot** comes in handy. An example of a box plot is shown in the figure. It can be drawn both horizontally and vertically. The box plot presents the minimum, maximum, and three quartiles. It displays the basic values of the data to the fullest.

## Boxplot

«box-and-whisker plot»  
«box-and-whisker diagram»



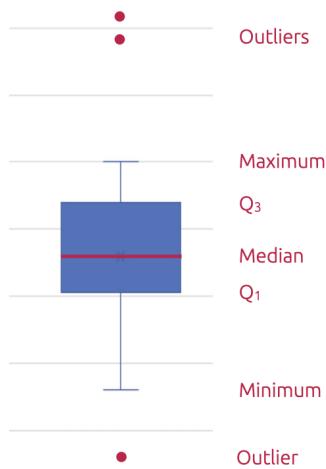
We can also use quartiles to identify so-called outliers, which are the values that are notably different from other data items.

Two types of outliers are called **mild** and **extreme**. **Mild outliers** lie  $1.5 \cdot IQR$  below the first quartile or above the third quartile but less than  $3 \cdot IQR$ . **Extreme outliers** lie below the first quartile or above the third quartile by more than  $3 \cdot IQR$ . The scheme that explains how to identify outliers is shown in the figure. Identification of outliers is very important for data preprocessing because it helps to remove the values of questionable validity.

The data is also used in a so-called box plot with an extension that illustrates outliers. It's calculated in two steps. The first step is to find the quartiles and



then outliers (shown on the plot as points). After that, the outliers are excluded from the data, and the minimum, maximum, and quartiles are recalculated and given as a box plot. The example of a box plot with outliers is illustrated in the figure.



Another useful statistical concept is **variance**. The variance of a dataset or a sample is a measure of the spread between numbers, or the deviation of values from the mean. To calculate the variance value, we can use the formula mentioned earlier or use a software tool that calculates this function.

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

In Google Sheets, this function is called `VAR()`. We specify the input sample for this function. In our case, it's the villagers' incomes. The result is shown in the figure.

Standard deviation, another commonly used statistic, is also tightly connected with the variance. **Standard deviation** is the square root of the sample variance. It's calculated by the formula:

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

It can be found in almost every tool. Google Sheets has the `STDEV()` function for standard deviation. The figure shows the calculated standard deviation of the villagers' income.

The **variance of a dataset or a sample** is the deviation of values from the mean. It is denoted as  $S^2$ .

A screenshot of a Microsoft Excel spreadsheet. The A column contains numerical values: 451000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, and 1000. The C column contains the formula =VAR(A1:A9), which has resulted in the value 4050000000. The status bar at the bottom shows the formula =VAR(A1:A9).

	A	B	C	D
1	451000		4050000000	x
2	1000		=VAR(A1:A9)	
3	1000			
4	1000			
5	1000			
6	1000			
7	1000			
8	1000			
9	1000			

**Standard deviation** is the square root of the sample variance.

A screenshot of a Microsoft Excel spreadsheet. The A column contains numerical values: 451000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, and 1000. The C column contains the formula =STDEV(A1:A9), which has resulted in the value 63639,61031. The status bar at the bottom shows the formula =STDEV(A1:A9).

	A	B	C	D
1	451000		variance	4050000000
2	1000			
3	1000			
4	1000		63639,61031	x
5	1000	st.deviation	=STDEV(A1:A9)	значение
6	1000			
7	1000			
8	1000			
9	1000			

This section has covered the main descriptive statistics that are beneficial to the exploratory data analysis.

## 2 Data Transformation

Data transformation is a common preprocessing technique used to identify the characteristics that are hidden in the data and cannot be seen in their original form.

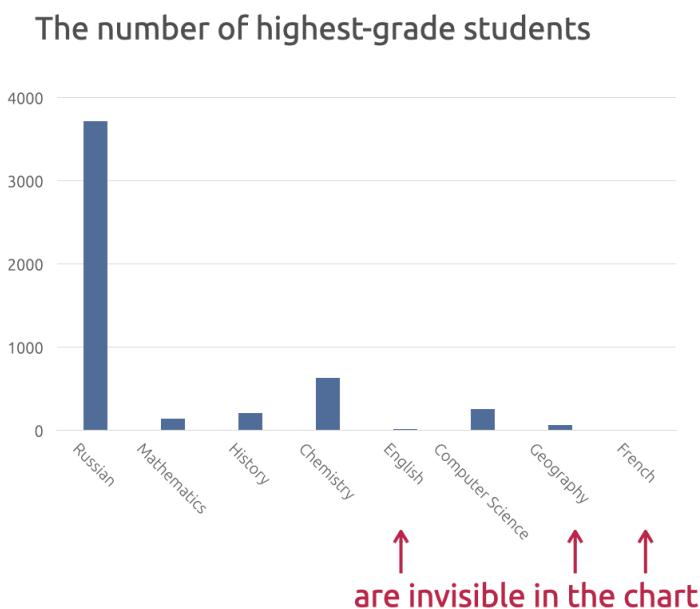


Data transformation helps to identify the characteristics that are hidden in the data and cannot be seen in their original form

The results of the exam

Subject	The number of highest-grade students
Russian	3722
Mathematics	145
History	206
Chemistry	634
English	15
Computer Science	254
Geography	64
French	2

To justify the need for transformations, we will use a specific example. The data on the 2018 exam results is available. The data is highly reliable. The exam results were published by the Federal Service for Education and Science Supervision of Russia.

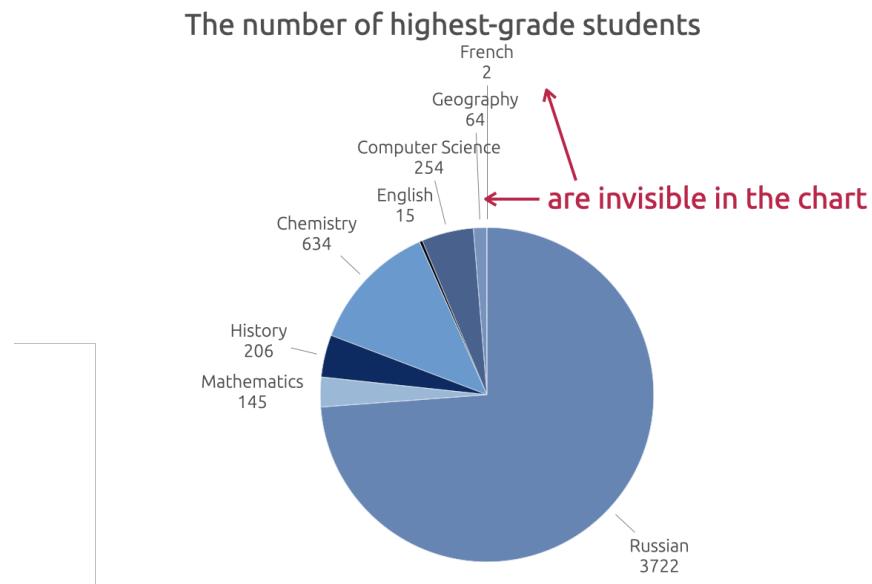


The table is overloaded with values, which makes it difficult to draw any conclusions. Let's try to display the number of highest-grade students in various subjects in simple graphs. The data corresponds to the aggregated values of different categories. Therefore, the data could be presented in a bar chart or pie chart. Here's what the data looks like in a bar chart.

Note that the data is reliable. This aggregated data is provided as is. However, the data cannot be visualized in this way because some values (for example,

corresponding to the categories of English, Geography, and French) are almost invisible in the chart.

Let's try to create a pie chart with the same dataset. Well, it's a little bit



better now. Geography now looks more convincing. However, English and French are still barely visible in the chart. The reason is the large spread of values in the aggregated data. What to do with this data and how to visualize it? A possible solution is to transform the data in a way that reduces the spread of values and makes the data commensurable. There are many different transformation methods. Let's consider the most common transformations and discuss how they affect visualization.

## 2.1 Common Transformations

The table shows the most common transformations and peculiarities of their use.



Transformation	Properties	Not applicable to
$\ln(x)$	preserves the order of the values	zero values
$\log_{10}(x)$		
$\sqrt{x}$	<ul style="list-style-type: none"> <li>preserves the order of the values</li> </ul>	negative values
$x^2$	<ul style="list-style-type: none"> <li>preserves the order of the values</li> <li>suitable for zero values</li> </ul>	for the data containing both positive, and negative values
$1/x$	changes the order of the values	zero values

For example, natural and common logarithms are well applicable to data transformation. They also preserve the order of the values but cannot be used if the input data contains zero values. The square root transformation also preserves the order of the values, and it can be applied to zero values but not negative ones.

The transformation by the reciprocal formula changes the order of the values (which can be useful sometimes), but it does not apply to zero values.

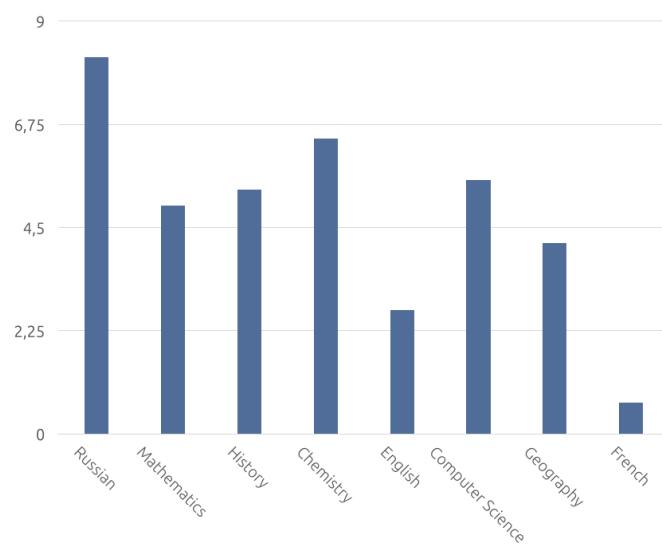
**Log transformation.** Let's consider the example of a log transformation to see how the data transformations are applied. A log transformation requires calculating the logarithm of each value in a dataset and use that transformed data instead of the input. Log transformations significantly affect the shape of

**The results of the exam**

Subject	The number of highest-grade students	ln(The number of highest-grade students)
Russian	3722	8.22
Mathematics	145	4.98
History	206	5.33
Chemistry	634	6.45
English	15	2.71
Computer Science	254	5.54
Geography	64	4.16
French	2	0.69

the distribution. The figure shows the bar chart of the highest-grade students after applying a natural log transformation.

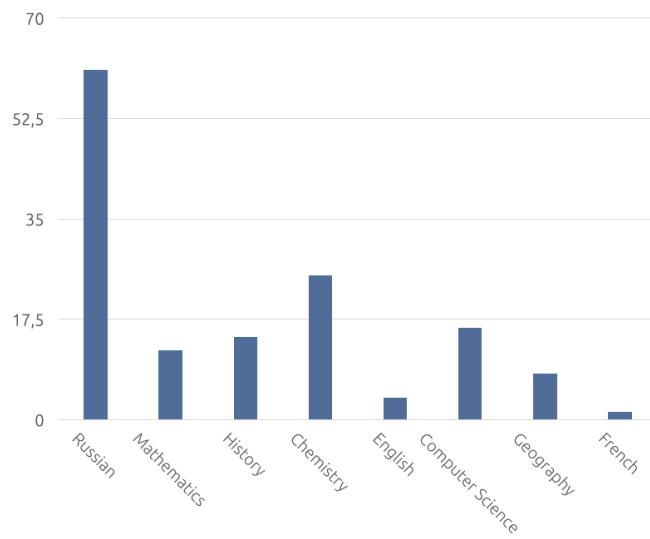
**The results of the exam:  
ln(The number of highest-grade students)**



**Square root transformation.** The square root transformation moderately affects the shape of the distribution.

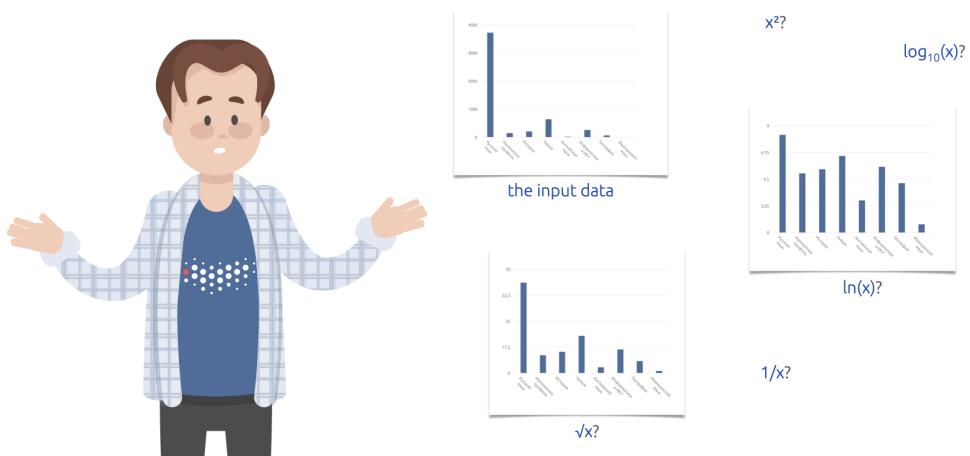
The next one is the bar chart of highest-grade students after applying the square root transformation.

**The results of the exam:**  
 $\sqrt{(\text{The number of highest-grade students})}$



## 2.2 Choosing the Right Transformation

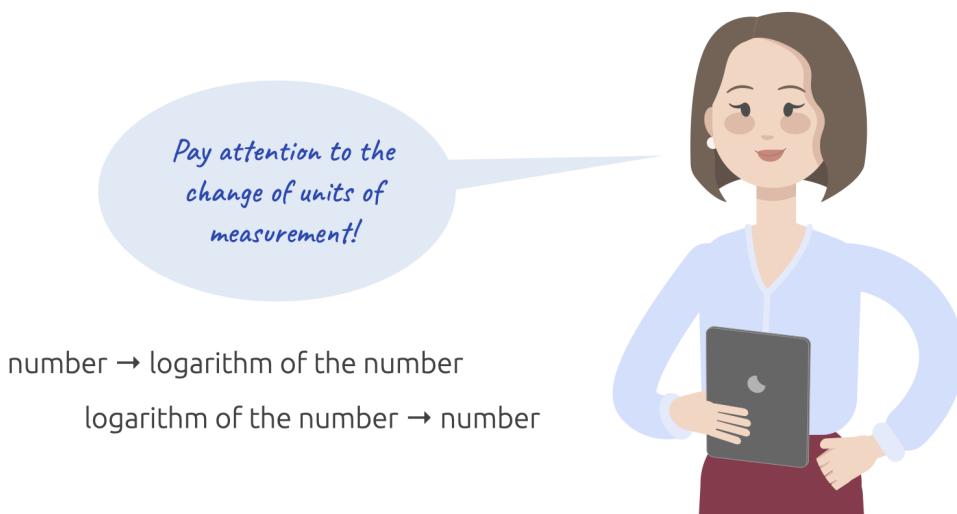
There are many possible transformations. So, how do we choose the right one? The answer to this question is not obvious even though we can use the formal statistical methods to choose the transformation. If you don't delve into these theories, the overall strategy for choosing a transformation is to define the purpose of the transformation (for example, visualization of a certain type), apply the most common transformations such as logarithms, a square root, square, reciprocal, and then choose the best method based on the purpose and obtained results.



## 2.3 Units of Measurement and Inverse Transformations

Since the transformation methods apply mathematical functions to the input data, we should pay attention to the change of units of measurement. For example, if you apply a logarithmic function to the variable corresponding to the number of highest-grade students, the logarithm of that number becomes the unit of measurement.

Therefore, when presenting data in charts, it's important to show the performed transformations and units of measurement. If the transformed data was used to calculate statistics, the inverse transformation is performed to represent the result in the initial units. For example, if a square root transformation was applied, we should perform the inverse transformation to square the result.



### 3 Data Normalization

Data normalization is a data preprocessing technique. It is used to compare, aggregate, and visualize the values of variables measured on different scales. Some algorithms (and in particular, machine learning) regularly require variables normalization.

The next example will justify the need for normalization. Suppose that a school has the following information about students and wants to decide who will be given a free trip. Let's assume that all the certificates of achievements are of

**Task:**

*find a criterion to choose  
the best student*



Information about students

	Average grade	The number of certificates in arts	The number of certificates of achievements in intellectual competitions	The number of certificates in sports
Ian	4.5	5	2	8
Alex	4.9	0	4	4
Simon	4	2	3	10
Kate	4.5	5	3	0
Mathew	4.2	2	6	9
Ann	5	2	5	0

the same order, and none of them is more impressive than others. The school should find a formal criterion for assessing the students, select one student who meets the selected criterion, and offer a free trip. What is the best way to find this criterion? The task is to match each student to one number representing the student's achievements, use the obtained numbers to rank the students, and then find a winner. If all the variables and grades were measured on the same scales and in the same units, we would summarize all the values. This approach is very rough because sports certificates are awarded more often than other types of certificates, and sports achievements can surpass, for example, scientific or academic achievements. The solution is to normalize the variable values to calculate the final criterion based on the normalized values.

**Why do we need normalization?** The qualities are usually expressed in numbers. The variable  $x$  varies from its minimum  $x_{min}$  (the lack of that quality) to its maximum  $x_{max}$  (the highest degree of that quality). A quality criterion allows comparing two objects based on one parameter only. However, the parameter may vary, and it's important to know the extent. It's also important to consider very diverse ranges of values and units of measurement of variables. Moreover, sometimes it's also necessary to estimate how close the value is to the ends of the range or the middle of the range. And comparing or aggregating data based on

different parameters is not an easy task. Yet the parameter of quality is interpreted as the **degree of intensity** of the quality. The good thing is that degrees of intensity can be and should be compared and aggregated! But the parameters should be adjusted to one scale so that the minimum and maximum values of different variables match. This transformation is called **normalization**. After this transformation, we can compare and aggregate various parameters obtained with different techniques.

### 3.1 Classes of Numerical Parameters

With all the variety of numerical characteristics, two broad classes can be distinguished:

- **Unipolar** expressing only the degree of the presence of quality or quantity (for example, intense color or very good grade)
- **Bipolar** reflecting not only the degree of the presence of quality but also its direction

The normalization methods differ for these classes. Let's consider some of these methods.

### 3.2 Normalization of Unipolar Parameters

Unipolar values are often normalized in the range from 0 to 1. Any continuous increasing function  $y = f(x)$  with the minimum of 0 and maximum of 1 can be a normalization function:



Let's consider possible variants of this function that have the discussed characteristics. There are two types of normalization **exponential**

$$y(x) = 1 - \exp \left( 1 - \frac{x}{x_{\min}} \right)$$

or **linear**.

$$y(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

- Option I

### Exponential Normalization

$$y(x) = 1 - \exp(1 - \frac{x}{x_{\min}})$$

- distributes the initial values in the range from 0 to 1 more evenly

- Option II

### Linear Normalization

$$y(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

- simple

The linear transformation is simple, so it's used more often than other transformations. The advantage of the exponential function is that it distributes the initial values in the range from 0 to 1 more evenly. Moreover, slight modifications to this formula make it easy to increase the evenness of distribution in certain cases.

### 3.3 Normalization of Bipolar Parameters

Bipolar parameters are usually normalized in the range from -1 to 1. Any continuous increasing function  $y = f(x)$  with the minimum of negative one and maximum of positive one can become a normalization function. Here's an example of such a linear function:

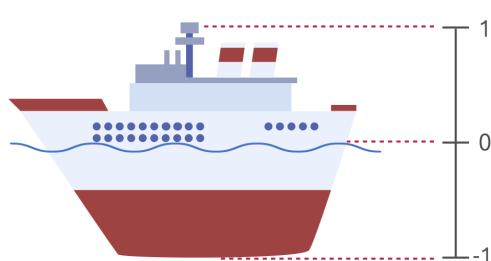
$$y(x) = \frac{2x - (x_{\max} + x_{\min})}{x_{\max} - x_{\min}}.$$

Sure, other normalization options are possible, and some of them are not linear and sometimes tied to the specifics of the domain where the parameters are normalized. However, this transformation is sufficient for subsequent analysis in most cases. Now armed with the knowledge, we can go back to the example of the students' grades.

Normalization function of bipolar parameters

$$y(x_{\min}) = -1; \quad \frac{dy}{dx} > 0$$

$$y(x_{\max}) = 1;$$



### 3.4 Parameters of Students

Here are all the parameters in the example. All of them are unipolar. The average grade is a unipolar parameter that shows the unidirectional quality of educational success, and it is often measured from 1 to 5. The number of certificates

	Average grade	The number of certificates in arts	The number of certificates of achievements in intellectual competitions	The number of certificates in sports
Ian	4.5	5	2	8
Alex	4.9	0	4	4
Simon	4	2	3	10
Kate	4.5	5	3	0
Mathew	4.2	2	6	9
Ann	5	2	5	0

in arts is a unipolar parameter (a positive integer). The number of certificates in sports is a unipolar parameter (a positive integer). The number of certificates of achievements in intellectual competitions is also a unipolar parameter (a positive integer). Thus, we can normalize the data by applying any normalization to the unipolar parameters. For ease of explanation, we'll take linear normalization. For

$$y(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

$$x_{\min} = 0;$$

$$x_{\max} = 5;$$

$$y = \frac{x}{5}$$

	The number of certificates in arts	
	Before normalization	After normalization
Ian	5	1.00
Alex	0	0.00
Simon	2	0.40
Kate	5	1.00
Mathew	2	0.40
Ann	2	0.40

example, we can normalize parameter “The number of certificates in arts”. To begin with, we find the minimum and maximum for this parameter. The minimum is 0, and the maximum is 5. Next, we plug these values in the formula for linear normalization of the parameter value. Take a look at the results.

We normalize the input data by the linear unipolar transformation formula for all the parameters (while keeping in mind that each parameter should be normalized separately). That's how we get the normalized values for each parameter. What's next?

### Normalized parameters of students

	Average grade	The number of certificates in arts	The number of certificates of achievements in intellectual competitions	The number of certificates in sports
Ian	0.50	1.00	0.00	0.80
Alex	0.90	0.00	0.50	0.40
Simon	0.00	0.40	0.25	1.00
Kate	0.50	1.00	0.25	0.00
Mathew	0.20	0.40	1.00	0.90
Ann	1.00	0.40	0.75	0.00

Now, we should set an objective function based on the normalized values (corresponding to students). What is the objective function? It's a mathematically expressed criterion for an object quality (process, solution). The objective function is set to obtain one parameter instead of a large number of qualitative parameters for each studied object. After that, it is used to find the object, at which an extremum is attained, based on the maximum or minimum of the function. So, what value of the function to use? Maximum? Or minimum? It depends on the specifics of the task and the form of the function.

$$\begin{aligned}
 \text{Objective function} & \quad \text{normalized average grade } (x) \\
 (\text{option A}) & \quad + \\
 F(x) = & \quad \text{normalized number of certificates in arts } (x) \\
 & \quad + \\
 & \quad \text{normalized number of certificates of} \\
 & \quad \text{achievements in intellectual competitions } (x) \\
 & \quad + \\
 & \quad \text{normalized number of certificates in sports } (x)
 \end{aligned}$$

*The student corresponding to the maximum of the function will win.*

For example, if this function reflects the cumulative positive qualities of a student, then it is probably the maximum. And when we talk about the total cost, it makes more sense to use the minimum. In our case, we can use the sum of the normalized values as an objective function because each of the values reflects

some positive qualitative characteristics of the student, and the maximum of the function is the best value. So, the student corresponding to the maximum of the function will win. There are at least two ways in which we can obtain the result with this function. We can add one more column to the table to calculate the sum of normalized parameters and find a student with the maximum of the objective function. Another option is to use a wonderful visualization tool called a stacked column chart. It is available in many visualization tools. This chart sums up the parameters, so our job is just to set them correctly and find the column with the maximum stacked value. We will show both ways.

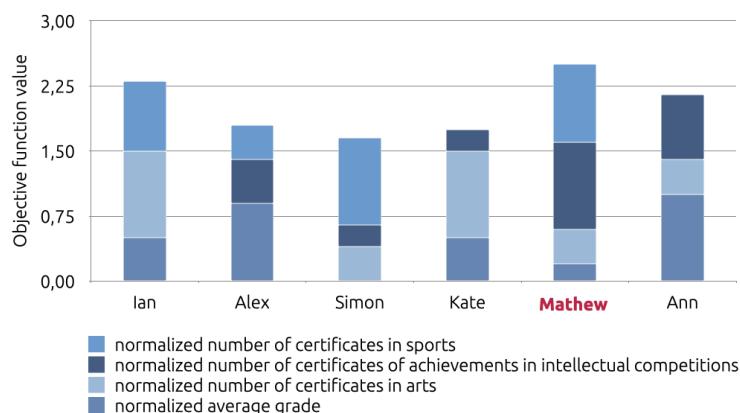
**The table with the objective function**

	Average grade	The number of certificates in arts	The number of certificates of achievements in intellectual competitions	The number of certificates in sports	Objective function value (option A)
Ian	0.50	1.00	0.00	0.80	2.30
Alex	0.90	0.00	0.50	0.40	1.80
Simon	0.00	0.40	0.25	1.00	1.65
Kate	0.50	1.00	0.25	0.00	1.75
<b>Mathew</b>	<b>0.20</b>	<b>0.40</b>	<b>1.00</b>	<b>0.90</b>	<b>2.50</b>
Ann	1.00	0.40	0.75	0.00	2.15

The last column in the table contains the value of the objective function (option A is the sum of all normalized parameters). The maximum of the objective function corresponds to Mathew. So, Mathew is the winner! The second way to find the winner is a stacked column chart.

We create the stacked column chart based on the normalized values and observe the same result. So, the column with the accumulated value of the maximum height corresponds to Mathew. In this case, all you need to take care of is to set the type of the chart and the input values correctly.

**stacked column chart:  
objective function value (option A)**



It might be the end but... it turned out that the school wants to double the significance of the normalized score for academic performance. So, we should plug all the normalized values in the objective function with the coefficient equal to 1 except the normalized average score that we multiply by the significance coefficient equal to 2; such coefficients are called weights. The final formula for the objective function is as follows.

$$\begin{array}{l}
 \text{Objective Function} \\
 \text{(option B)} \\
 F(x) = 
 \end{array}
 \begin{array}{l}
 \text{normalized average grade (x)} *2 \\
 + \\
 \text{normalized number of certificates in arts (x)} \\
 + \\
 \text{normalized number of certificates of} \\
 \text{achievements in intellectual competitions (x)} \\
 + \\
 \text{normalized number of certificates in sports (x)}
 \end{array}$$

↑  
significance  
coefficient

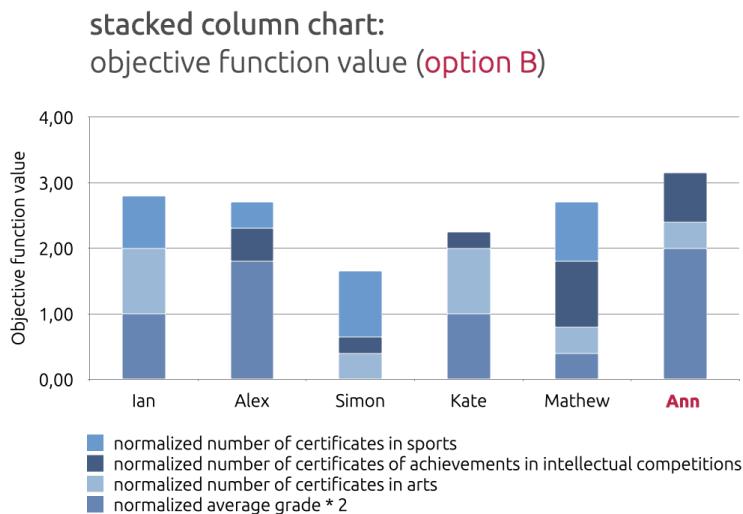
Let's try to find the winner with the new objective function. Option B in the table is the objective function. So, the maximum of the objective function (2.833) corresponds to Anna.

The table with the objective function

	Average grade	The number of certificates in arts	The number of certificates of achievements in intellectual competitions	The number of certificates in sports	Objective function value (option B)
Ian	0.50 *2	1.00	0.00	0.80	2.8
Alex	0.90 *2	0.00	0.50	0.40	2.7
Simon	0.00 *2	0.40	0.25	1.00	1.65
Kate	0.50 *2	1.00	0.25	0.00	2.25
Mathew	0.20 *2	0.40	1.00	0.90	2.7
Ann	1.00 *2	0.40	0.75	0.00	<b>3.15</b>

So, in this case, the winner is Anna! Another option is shown in the stacked column chart.

We cannot change the manner of stacking in the chart, but we can double the values of the performance parameter in the input data for the chart and create an ordinary stacked bar chart. As expected, the winner is Anna.



## 4 Objective Function

Sometimes, we contrast real-world objects to find a better solution to the problem at hand. When the object has one feature, the task seems simple. For example, when we are trying to find the cheapest apartment, the car with the biggest trunk, or the highest point in the world above sea level. However, when objects have multiple relevant features, the question is how to find the best of them. When it comes to numerical data, we can normalize it to align the order of the possible values and adjust the ranges of feature values to fixed boundaries. The normalized data is used to construct the objective functions used to solve applied problems. How to define the objective function in the general case?

The **objective function** is a numerical function of several variables (parameters) that is optimized to solve an optimization problem.

What does **optimization** mean in this context? It's searching for possible values of the parameters that maximize or minimize the function depending on the problem.

The process of solving applied problems often reduces to solving optimization problems. And the plan to solve the latter often includes the following steps:

1. Normalize the values of input variables (features).
2. Construct the objective function (the function being optimized) based on the task.
3. Find the largest (or the smallest) value of the objective function on the set.

Rather than finding the value of the objective function, the last step often finds the value of the argument at which the function attains this value. This point is often called the optimum of the objective function or the global maximum or

minimum.

$$F(X) = a_1 \cdot x_1 + a_2 \cdot x_2 + \dots + a_n \cdot x_n,$$

where  $x_i$  are normalized parameters,  $a_i$  are coefficients (weights).

In the simplest cases, the objective function is a sum of the normalized numerical parameters (with weights at times). Normalization aligns the values of different features, while the coefficients (weights) of the objective function allow us to make a feature more or less significant. However, the objective function is not always so easy to set. Moreover, the construction of the function in most cases reflects the subjective opinion of the researcher solving the problem. The subjectivity is prominent when the function is constructed and the features are not numerical (for example, nominal or ordinal). So, the researcher decides what data needs transformation into the numerical form, and then constructs the objective function. Nowadays, many tools offer built-in techniques and libraries that allow data transformation.

Nominal data		Numerical data			
Object's ID	Color	Object's ID	blue?	red?	white?
1	blue	1	1	0	0
2	red	2	0	1	0
3	white	3	0	0	1

Here's one of the most versatile methods for the transformation of nominal data into the numerical form. Each nominal value is matched to the additional feature with two possible values. For example, the Color variable takes 3 possible values, blue, red, and white, so, we can introduce 3 additional features for the Color feature blue?, red?, and white?, so that each takes 0 or 1 when describing a particular object.

Ordinal data is often handled differently. Ordinal data is easy to order (that's how it got its name). Therefore, we can match each data item to a numerical value to reflect the order. For example, if we consider the answers For, Against,

bipolar case		unipolar case	
Absolutely against	2	Absolutely against	0
Against	1	Against	1
Don't care	0	Don't care	2
For	-1	For	3
Absolutely for	-2	Absolutely for	4

Absolutely against, Absolutely for, Don't care, we can order these answers on a numerical scale reflecting the measure of agreement and match the numerical values to the answers while preserving the order. These values in the bipolar and unipolar cases would look as follows:

Let's consider the example of constructing an objective function to choose the best rental car. The input data is the information on the cars of a car-sharing company. We assume that we don't need to choose the brand of a car. The cars are manufactured by the same company, but the options available in those cars differ. It's known that:

- Some of the cars run on diesel, others on gasoline.
- The gearboxes can be manual, automatic, or robotic.
- Some of the cars are with a heated steering wheel while others are not.
- The cars have trunks of different sizes.
- The year of manufacture and daily rental price also differ.

How do we choose the best car? First of all, we will transform the data into numbers, then normalize this data, construct the objective function, and use its extremum (minimum or maximum) to choose the best car.

Let's have a look at the table with the list of available cars: The next step

Option No.	Fuel	Gearbox	Heated steering wheel	Trunk size (l)	Year of manufacture	Daily rental price (rub)
1	diesel	automatic	yes	270	2015	2000
2	gasoline	automatic	yes	700	2014	2100
3	diesel	manual	no	370	2018	2150
4	gasoline	automatic	yes	750	2018	2200
5	diesel	automatic	yes	370	2016	1900
6	diesel	automatic	no	700	2017	2500
7	gasoline	manual	no	350	2019	2300
8	diesel	automatic	yes	700	2017	2400
9	gasoline	automatic	yes	750	2017	2600
10	gasoline	automatic	no	800	2017	2700
11	gasoline	robotic	no	850	2019	2800

is data normalization. Let's begin with the option called a heated steering wheel. It's dichotomous data with two possible values of yes or no. We can translate this data to 0 and 1 to construct the objective function later on. A heated steering wheel is a great option. Therefore, the value yes will be 1 and no will be 0. The data is immediately normalized and requires no further transformation.

Heated steering wheel (input data)	Heated steering wheel (numerical data)
yes	1
no	0

How to transform the gearbox data? From a driver's point of view (which is subjective), this data can be represented as ordinal and, therefore, numerical data. Given that many drivers prefer an automatic gearbox, we can arrange the numbers as follows:

Gearbox (input data)	Gearbox (numerical values)
automatic	3
robotic	2
manual	1

Note that the maximum value matches the best value. However, this transformation is insufficient to construct the objective function. So, we should normalize the obtained values. We apply the classic linear normalization to obtain the following data:

Gearbox (input data)	Gearbox (numerical values)	Gearbox (normalized numerical values)
automatic	3	1
robotic	2	0,5
manual	1	0

The next option is the type of fuel. It's nominal data with two possible values of diesel or gasoline. Thus, it's also dichotomous data that we can easily transform into the numbers of 0 and 1. Many drivers choose diesel (which is also subjective), so, it makes sense to assign 1 to diesel and 0 to gasoline.

As a result, we obtain normalized data that requires no further transformation.

The next thing to consider is the volume of a car trunk. The data is numerical, so we should only normalize it.

The next option is the year of manufacture. The data is also numerical, so we will simply normalize it.

The last option is the rental price. It is also numerical data to be normalized.

We've normalized the car data, and now we can construct the objective function  $F$ , find its maximum (or minimum), and see for which car the function attains that value. As has been said, the construction of an objective function is a subjective process that reflects the bias of the analyst who constructs the objective function. That's why it's critical to rely on expert opinions when constructing the objective function. Experts with knowledge in the field can explain why some features are more important than others. In the car example, we will be experts.

Imagine that we've developed the following requirements for constructing the objective function:

1. All the parameters that may affect the choice of a car are equally important to us (thus, no coefficients in the objective function are required).
2. When choosing between two cars with different types of fuel, we would prefer the one running on diesel (it corresponds to the maximum normalized value).
3. When choosing a gearbox, we would prefer the car with the gearbox that has the largest normalized value.
4. When the only difference is the absence or the presence of the heated steering wheel option, we would prefer the car with that option (it corresponds to the maximum normalized value).
5. When choosing a trunk size, we would prefer the car with the largest trunk (which corresponds to the maximum normalized value).
6. When choosing between two cars of different years of manufacture, we would prefer the car with the latest year of manufacture (which corresponds to the maximum normalized value).
7. When the only difference is a rental price, we would prefer the car with the lowest rental price (which corresponds to the minimum normalized value).

We can combine these expert opinions to get the following objective function that will attain its maximum when the best car is chosen:

$$F(X) = NF(X) + NB(X) + NH(X) + NV(X) + NY(X) + (1 - NP(X)),$$

where

- $NF(X)$  is a normalized value corresponding to the type of fuel
- $NB(X)$  is a normalized value corresponding to the gearbox
- $NH(X)$  is a normalized value corresponding to the heated steering wheel option
- $NV(X)$  is a normalized value corresponding to the car trunk
- $NY(X)$  is a normalized value corresponding to the year of manufacture
- $NP(X)$  is a normalized value corresponding to the rental price

Note that when considering the rental price parameter, the term in the objective function was not the normalized value but the expression  $(1 - NP(X))$ . We are trying to consider that, other parameters being equal, we would prefer the car with the lowest cost. The next table shows the values of the objective function for all the cars in the input data. Looks like we got lucky, and the options are

Option No.	Value of the objective function $F(X)$
1	4,09
2	3,52
3	2,69
4	4,29
5	4,57
6	3,67
7	1,69
<b>8</b>	<b>4,79</b>
9	3,65
10	2,62
11	2,50



limited. We have been able to calculate the value of the objective function for each possible variant. The table shows that the maximum of the objective function corresponds to variant 8. It is the best choice. In more complex cases, it may be difficult to calculate all possible values of the objective function. To find the best values, we can use special mathematical methods. However, we will not discuss them this time. To learn more, we invite you to the Machine Learning course.

## 5 Time Series

### 5.1 Time Series Analysis

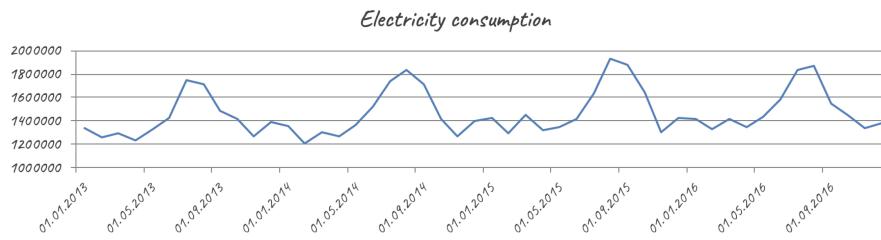
A time series is a series of observations in time order:

$$y_1, y_2, \dots, y_n,$$

where  $y_i$  are the values of a variable at  $n$  equally spaced time points:

$$t_1, t_2, \dots, t_n.$$

Examples of time series are consistently measured data (collected every day, hour, minute, and so on) on power generation, production volume, sales, consuming, transport services, etc. The figure shows an example of the process given by a time series.



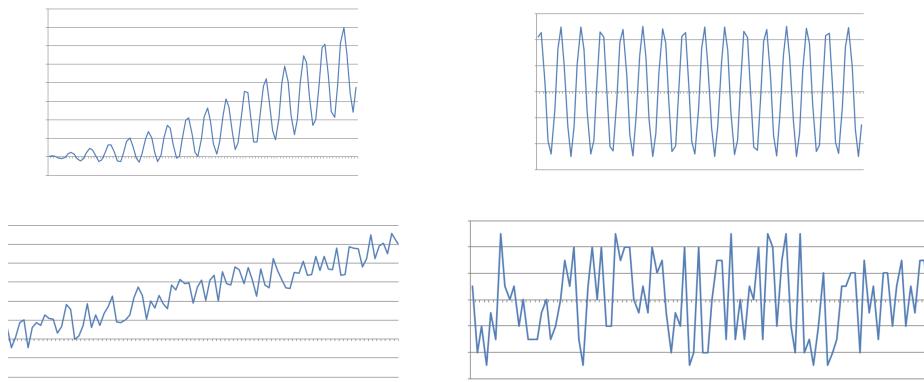
Time series analysis assumes that the past values of time-series data could be used to predict future values.

The problems of time series analysis are generally divided into two classes:

- Analyzing time-series **components** to extract meaningful statistics and other characteristics of the data.
- **Forecasting** the time-series behavior in the future.

Let's consider these problems in turn. The time-series components are:

- A **trend** reflects smooth, long-term changes in the level of the series.
- **Seasonality** describes periodic changes in the level of the series in a fixed period of time (for example, monthly power consumption corresponds to a period of 12 months).
- A **cycle** reflects the repeating period in the series (for example, economic cycles, solar cycles, and so on).
- **Noise** is an unpredictable random variation in the series.



The figure shows several examples of the series and the respective components.

Time-series forecasting assumes the construction of such a function  $f$  that gives the series predicted value for the points  $t + h$  based on the values of the time series  $y_1, y_2, \dots, y_t$  and the additional parameter  $h$ :

$$f(y_1, y_2, \dots, y_t, h) = \hat{y}_{t+h},$$

where  $h$  is a parameter that lies within the interval from 1 to  $H$ , and  $H$  is the **forecasting horizon**. Based on the horizon value, the forecasting problem is generally divided into **short-term** and **long-term** forecasting. However, the time scale is conditional, as it often depends on the characteristics of the time series.

There are some things to consider while solving these problems. Let's elaborate on them.

- It is not always easy to reveal underlying patterns in the history of a series (for example, to measure the duration of the periods of time or choose the appropriate analytical function for a trend).
- Patterns (if there are any) can be distorted by noise in the data. These distortions are especially characteristic of data collected by sensors. That is why the analysis includes data preprocessing to remove noise by way of different smoothing techniques.
- The dynamics of the series in the past do not guarantee similar behavior of the series in the future because the behavior can significantly change due to external factors (for example, oil price dynamics tends to change dramatically in response to oil production quotas, political events in the oil-producing countries, and so on).

Despite these problems, we can and should develop time-series models because their benefits to business development outweigh the known risks. However,

some mathematical methods in statistics more accurately estimate the so-called **prediction interval**. The prediction interval gives an interval within which we expect the predicted value to fall with the probability not lower than a specified one.

Time-series models are usually classified as **additive** and **multiplicative**. An additive model predicts a series by recursively adding or subtracting some

### Additive model

$$y_t = u_t + s_t + e_t$$

### Multiplicative model

$$y_t = u_t \cdot s_t \cdot e_t$$

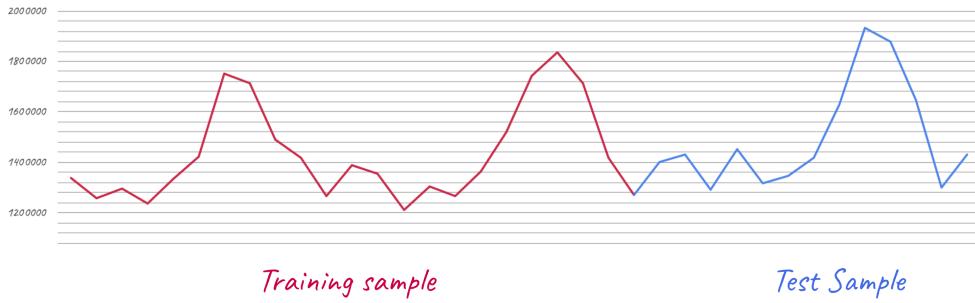
*ut is the trend component,  
st is the seasonal component,  
et is the random component.*

increments with respect to the known values of the time series. A multiplicative model suggests that the known elements are multiplied by some coefficients to forecast a series. For example, we build an additive model and know that the average monthly increase in demand for a certain product is 100 pieces. Then, the predicted value for the next month's demand is defined as the previous series value plus 100 pieces. The multiplicative model shows this as the demand increase by 10 percent. Hence, the predicted value for the demand in the next month can be calculated as the previous series value multiplied by 1.1. Seasonality, which periodically repeats in the series, can also be analyzed by using the additive or multiplicative model.

This lecture will cover simple examples of time-series modeling. But first, let's discuss how to assess the quality of obtained time-series models and how to compare them. Such comparisons require so-called quality metrics. The **metric** is a function that defines the distance between any two members of the set. There are many different forms of such metrics, but for now, we will look at some of them.

## 5.2 Quality Metrics

Before deciding which metric to use, we need to find out more about the data. To assess the forecast quality, we need predicted values and actual time-series data. Moreover, we need to build a model based on one part of the data and verify it against another part. The data used to construct a model is known as a training sample, while the data for the model assessment is called a test sample. How to split a time series into a training and a test sample? This can be solved in many different ways. However, in the case of a time series, the simplest and most logical approach is to split the series into three parts and use the first two



as a training sample while building the model. The third part of data can be a test sample for assessing the quality of the developed model.

Now we can move on to the metrics. Most of the quality metrics are based on the concept of a **forecast error** (that we will designate by  $e_t$ ). A forecast error at time  $t$  is the difference between the predicted and observed values of the variable at time  $t$ . Therefore,

$$e_t = \hat{y}_t - y_t,$$

where  $\hat{y}_t$  is the predicted value, and  $y_t$  is the actual value of the variable.

All the metrics that we use assume that the model with minor errors in the past will have minor errors in the future.

The first metric, **MAE (mean absolute error)**, is obtained by dividing the sum of the absolute values of the forecast errors by the number of points in the test sample:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

The second metric, **MSE (mean squared error)**, is calculated as the sum of squared forecast errors divided by the number of points in the test sample:

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2$$

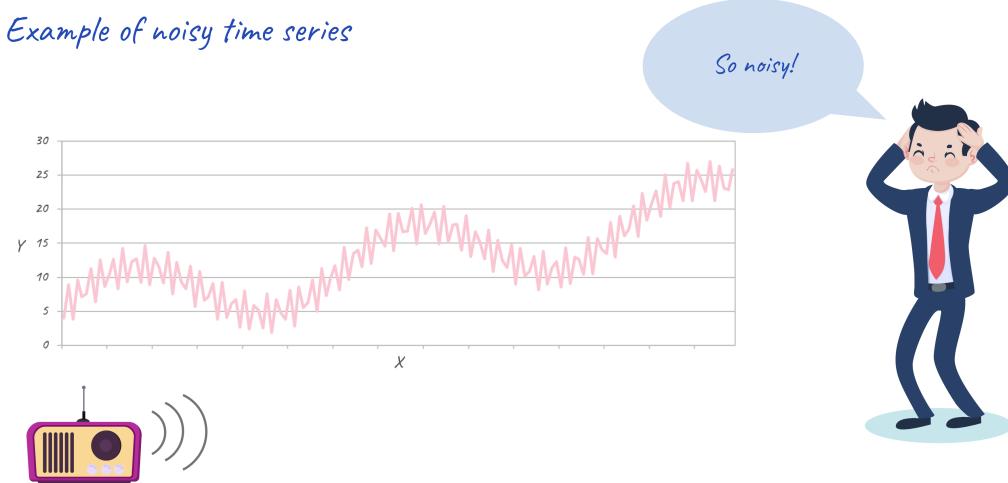
The third metric, **MAPE (mean absolute percentage error)**, is defined as a percentage of the sum of forecast errors divided by actual values of the time series over the number of points in the test sample. In the formulas,  $e_t$  is the forecast error,  $y_t$  is the actual value of the variable, and  $n$  is the number of points in the test sample:

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right| \cdot 100\%$$

We will use these metrics to evaluate time-series models. When choosing the model, we will use metrics with minimum values.

### 5.3 Time Series Smoothing

As we have already noted, some time-series values are noisy (in the sense that they reflect a random error). Noisiness is often found in series that are based on the data obtained from different sensors. An example of a noisy time series is shown in the figure. When analyzing time series, you need to identify their structure and estimate all major components, which is complicated because of noise. The good news is that there are many ways to remove it. Let's consider some of the well-known methods widely used for denoising.

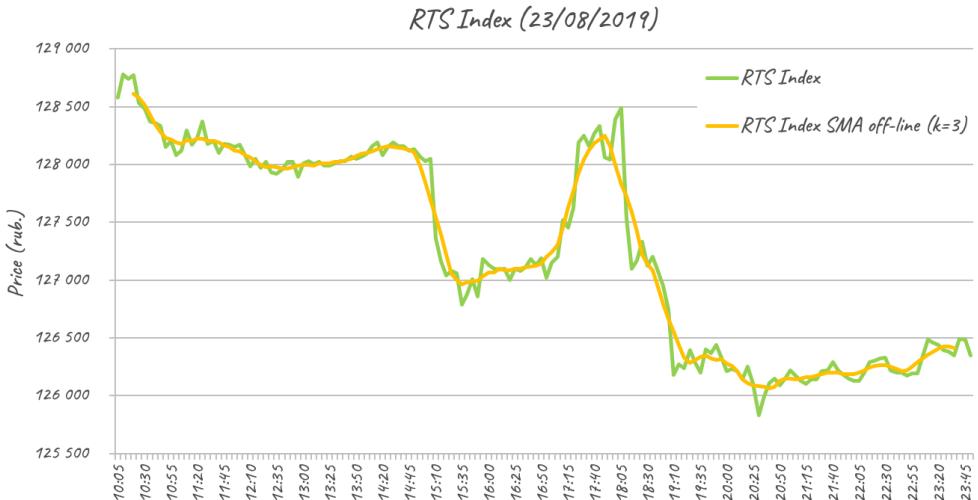


The first method is a **moving average**. The moving average requires setting a window width for each value of the series variable. The window is placed over neighboring values of the series (ideally,  $k$  values before and  $k$  values after the smoothed value). Next, these neighbors and the initial value of the series are used to calculate the arithmetic mean. Here's the smoothing formula:

$$s_i = \frac{1}{2k+1} \sum_{j=-k}^k y_{i+j}$$

In this formula,  $y_i$  is the value of the initial series,  $s_i$  is the value of the smoothed series, and  $2k + 1$  is the window width. The degree of smoothing depends on the window width. A large window size results in rough smoothing and a possible loss of information about the series dynamics. A small window size of 5-7 points can result in insufficient denoising. There is no universal window size. A window size heavily depends on the domain and the goals of averaging in each particular case.

Smoothing is widely used in the technical analysis of market quotes, and it is built in all the tools for stock analysis. Let's look at the example of the initial and smoothed RTS index chart after applying the moving average when  $k$  equals 3. Please note that the calculations of the moving average do not allow finding



the smoothed value for the first  $k$  and last  $k$  points of the series. The moving average formula always requires access to future measurements. Therefore, it can be calculated only offline (when the previous and next values for each smoothed series point are known). In practice, the moving average is often calculated online (which means we know only the past values and the point on the smoothed time series). This smoothing is often used in the technical analysis of financial data, like stock prices. Can we apply the moving average to online calculations?

Yes, we can. However, the calculation formula will include only the neighbors preceding the smoothed value:

$$s_i = \frac{1}{n+1} \sum_{j=-n}^0 y_{i+j},$$

where  $y_i$  is the value of the initial series,  $s_i$  is the value of the smoothed series, and  $n+1$  is the series width. This smoothing results in a delay of the smoothed data stream, but still, it can be successfully used to perform specific tasks. There's one more technique often used in online smoothing. It is **exponential smoothing**. Exponential smoothing is very simple because it is given by the recursive formula.

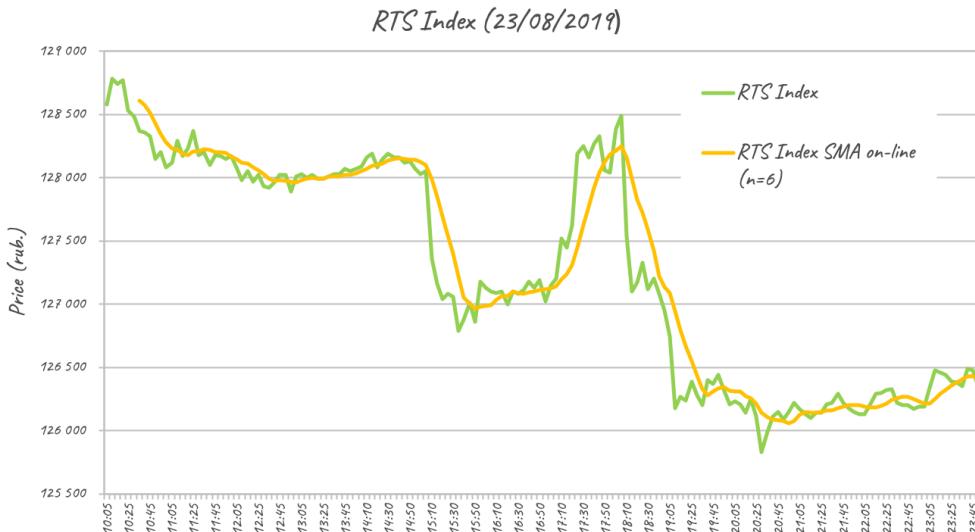
Assume  $Y = \{y_1, \dots, y_T\}$  is a time series. **Exponential** smoothing is performed according to the following recursive formula:

$$s_t = \alpha y_t + (1 - \alpha)s_{t-1},$$

where  $y_t$  is the value of the initial series at point  $t$ ,  $s_{t-1}$  is the value of the smoothed series at point  $t-1$ ,  $\alpha \in (0, 1)$  is the smoothing parameter. The initial value  $s_1$  is defined as the first point on the time series:

$$s_1 = y_1.$$

Choosing an appropriate value of smoothing parameter  $\alpha$  is crucial. The current smoothed value is obtained from the previous one and some error rate



of the previous smoothing. The value of the error used for adjustment is defined by smoothing parameter  $\alpha$ . The closer the value  $\alpha$  is to 1, the larger part of the discrepancy between the smoothed and real values is considered acceptable for calculating the next value. The closer value  $\alpha$  is to zero, the larger part of the discrepancy between the smoothed and real values is considered random, therefore, the smaller part of discrepancy is used to calculate the next value.

We can rewrite the exponential smoothing formula in the non-recursive form:

$$s_t = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \alpha(1 - \alpha)^2y_{t-2} + \alpha(1 - \alpha)^3y_{t-3} + \dots$$

The formula shows that the smoothed value is the weighted sum of all the previous values of the time series, and the coefficients decrease when the distance between the series value and the present moment of time increases. For example, if  $\alpha = 0.1$ , the formula takes the form:

$$s = 0.1y_t + 0.09y_{t-1} + 0.081y_{t-2} + 0.0729y_{t-3} + \dots$$

How do we choose the smoothing parameter? A formally correct procedure for choosing  $\alpha$  doesn't exist. In practice (at least in exchange activities),  $\alpha$  lies between 0.1 and 0.3. We can say that the value of the smoothing parameter is based on the statistician's judgment on the sustainability of change in the parameter at hand.

The figure shows examples of exponential smoothing of the RTS index with various values of the smoothing parameter. The smoothing parameter  $\alpha = 0.1$  is on the first chart. On the second chart, the smoothing parameter  $\alpha$  is 10 times smaller and equals 0.01. Note how the choice of the smoothing parameter affects the result.



## 5.4 Decomposing Time Series Data into Trends

Earlier, we discussed the major components of a typical time series such as trend, seasonality, and noise. We also noted that it is possible to denoise data using special techniques at the stage of preliminary analysis of time series.

And denoising makes the pattern more salient. How to identify a trend and seasonality? Can they be found analytically (by applying a time-dependent mathematical function)? How will it contribute? If we learn to characterize the behavior of a time series, we will be able to predict its behavior. For business, it means that we can predict car sales or the number of passengers, consumers, guests, and many more. So, this is actually very interesting.

Well, how to identify a trend in the time series? In practice, we use such analytical functions as linear, polynomial, exponential, and logarithmic.

### Analytical Functions for Trend Detection

$$\text{Linear } f(x) = a + bx$$

$$\text{Polynomial } f(x) = a + b_1x + b_2x^2 + \dots + b_kx^k$$

$$\text{Exponential } f(x) = ce^{a+bx}$$

$$\text{Logarithmic } f(x) = a \log_b x + c$$

Of course, we can also use other functions, but these four are more common,

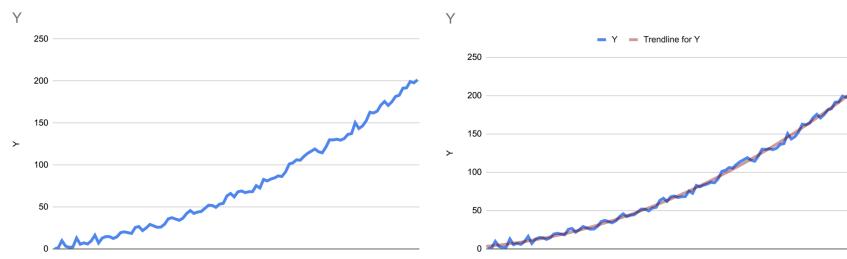
and they are embedded in many tools. How do we choose the right function for each case?

There are formal methods, but the simplest way is to recall how the graphs of the respective functions look like and use the graph of a time series to choose the best function. Let's consider the examples.

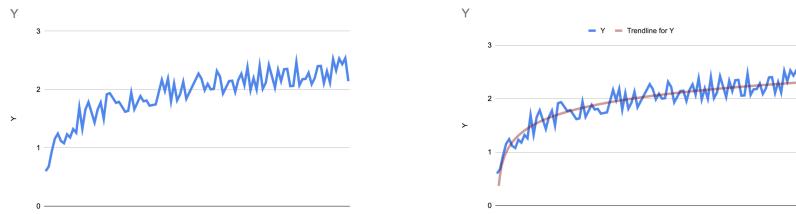
*Example 1 (Linear Trend)*



*Example 2 (Polynomial Trend)*



*Example 3 (Logarithmic Trend)*



However, to model the behavior of the series, it is not enough to define the type of the function, which corresponds to the trend line. We need to find the exact function parameters. So, how to do this? The corresponding mathematical methods exist for each mentioned type, and we can use these methods to find the parameters.

For example, to find the parameters of a linear trend, we can use a least-squares fit. It helps to find the parameters of the function by using the formulas:

$$y = ax + b,$$

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2},$$

$$b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n},$$

where  $n$  is the number of measurements,  $y_i$  is the element of the time series,  $x_i$  is the time.

To find the parameters of the other trends, we can apply specific mathematical techniques. This time, we will not describe them in detail. However, note that the trends and the corresponding analytical functions are well identified by the simplest tools such as **Google Sheets** or **Microsoft Excel**.

Before we move on to identifying trends, we would like to make one more note. Sometimes, it is hard to choose the trend line among several options. Are there formal criteria for choosing the type of the trend? In fact, such criteria exist! And it's called the **coefficient of determination** (designated by  $R^2$ ):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - y_{avg})^2},$$

where  $y_i$  is the value of the time series at the time  $i$ ,  $f_i$  is the value of the trend at the time  $i$ , and  $y_{avg}$  is the average value of the time series.

A coefficient of determination can be used to assess the quality of the selected trend equation. It takes the values from 0 to 1. For the acceptable trend models, it is assumed that a coefficient of determination should not be less than 0.5. Models with a coefficient of determination greater than 0.8 are considered good enough. When the coefficient of determination  $R^2 = 1$ , there's a functional relationship between the variables (between the initial time series and the trend). The figure shows the accurate formula for a coefficient of determination.

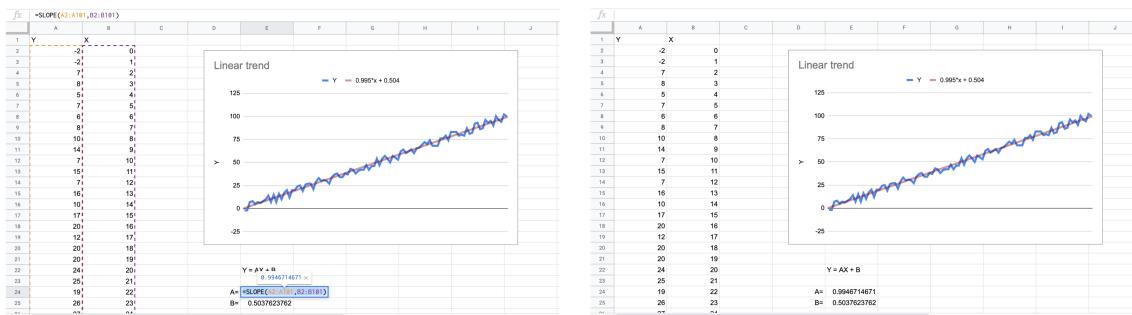
We will use a coefficient of determination that is implemented in many tools (for example, **Google Sheets** or **Microsoft Excel**) to assess the quality of the trend lines.

## 5.5 Drawing Trend Lines

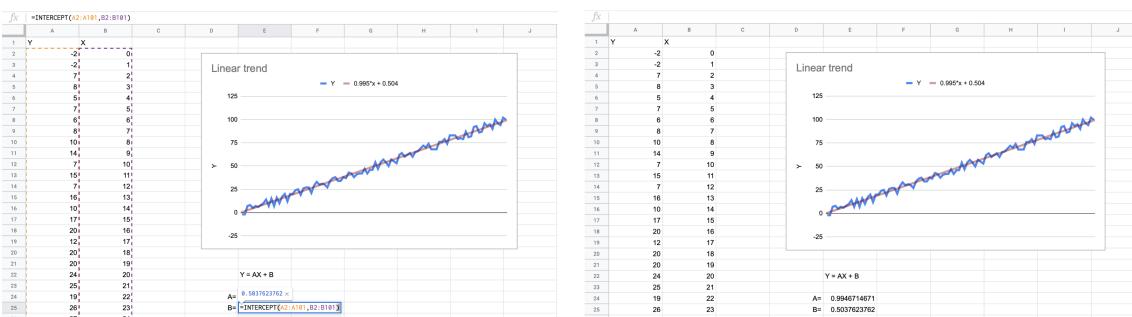
Data processing tools can not only display the trend equation on a chart but also offer options that allow calculating the parameters of the linear trend without creating a graph. These functions are **SLOPE** and **INTERSECT**. For the trend equation  $y(x) = ax + b$ , **SLOPE** calculates the parameter  $a$ , while **INTERSECT** calculates the parameter  $b$ . In some cases, these functions are preferable because they give a more accurate result compared to the equation with the rounded parameters (it is shown on the trend chart).

Let's look at the example of using the **SLOPE** and **INTERSECT** in **Google Sheets** and the results:

## SLOPE Example



## INTERCEPT Example



Compare the accuracy of the representation for the parameter  $a$  in the equation on the chart and after applying the function.

Well, now you know how to identify trends and the corresponding analytical functions. How to apply this knowledge in practice? For example, you can forecast the behavior of simple time series with noise and trends but without seasonality. To put it differently, you can model a time series based on the analytical function that corresponds to the trend. And we have such a time series to practice.

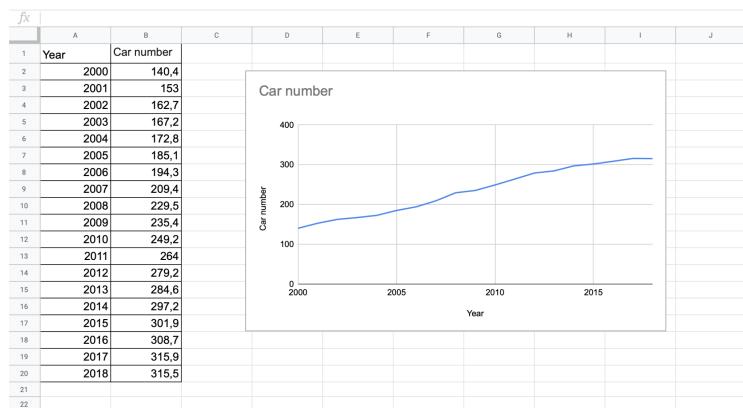
We've discussed this time series before. It refers to the cars for 1.000 persons in the Central Federal District. Data reflects the period from 2000 to 2018. Let's create a linear graph corresponding to this time series and make sure that there's no seasonality. Therefore, we can model the series based on the analytical function  $ax + b$ .

Let's see how effectively the points are modeled by this function. To do this, we will split the series into training and test samples, find the parameters of the trend on the training sample, forecast on the test sample, and then assess the forecast quality with the MAPE metric.

The training sample is highlighted in yellow, and the test sample in blue. To find the parameters of the linear trend, we will use **SLOPE** and **INTERSEPT**.

We can calculate the predicted values on the test sample (for the years from 2013 to 2018) by the formula  $ax + b$ . The result is shown in the **Forecast**

### Cars per 1.000 Persons in the Central Federal District



column. To obtain this result in **Google Sheets**, you can also use the built-in **FORECAST.LINEAR** function that outputs the predicted value based on the model of the linear trend for the specified time series. Next, we can assess the forecast quality by examining the **MAPE** metric. To do so, we add a new column named **Error** and calculate the terms in the numerator for the **MAPE** metric. All we've got left is to sum up all the values in the **Error** column and divide it by the number of test sample elements. The result of 2.9% shows an average deviation from the actual values and seems reasonable.