

Document Description

The following document contains a description of the commands and their context.

This formatting provides a brief description of the commands and the context of the entire exercise. It informs you about the next steps to be performed in the exercise.

This formatting provides instructions to be performed. These instructions are mandatory. Proper names are italicized in this section. Formula fragments, such as function names, partial cell addresses, and operators (e.g., **IF**, **A1:B3**), are always bold and uppercase.

*This formatting indicates the context for the current exercise element as well as more detailed information about additional options. Here you are given the description and effects of executing other available commands. Proper names in this section are not be italicized. Formula fragments, such as function names, partial cell addresses, and operators (e.g., **IF**, **A1:B3**), are always bold and uppercase.*

Formulas to be entered are written in the following format:

Cell where formula should be entered = Formula to be entered

This format always indicates that you should go to the cell with the given address (before the equal sign), and then enter the equal sign and the formula itself in the formula bar or inside the cell.

Introduction

Using MS Excel is much easier with its function keys. Keyboard shortcuts allow you to work faster and more efficiently without using a mouse. Below are some of the most commonly used keyboard shortcuts.

Keyboard shortcuts	Action
Ctrl + C	<i>copy</i>
Ctrl + V	<i>paste</i>
Ctrl + Z	<i>undo action</i>
Ctrl + Y	<i>retry action</i>
Ctrl + S	<i>quick save</i>
Ctrl + A	<i>select everything in the sheet</i>
Ctrl + P	<i>print</i>
Ctrl + B	<i>bold text</i>
Ctrl + F	<i>find</i>
Ctrl + N	<i>open new sheet</i>
Ctrl + U	<i>underline text</i>
Ctrl + W	<i>close sheet</i>
Ctrl + I	<i>write in italics</i>
Ctrl + X	<i>cut text from selected area</i>
Shift + Arrow	<i>select cells in selected column or row (arrow keys correspond to selection direction)</i>
Ctrl + Tab	<i>switch between workbooks</i>
Ctrl + Pgup/Pgdn	<i>switch between sheets</i>
Ctrl + Shift + Arrow	<i>quick select a given area</i>
Ctrl + 1	<i>go to cell formatting settings</i>
Ctrl + L/Ctrl + T	<i>create a table</i>
Ctrl + Space	<i>select current column</i>
Shift + Space	<i>select current row</i>
Ctrl + A/ Ctrl + Shift + Space	<i>select everything</i>
Ctrl + Arrow	<i>move to the end or beginning of a column</i>

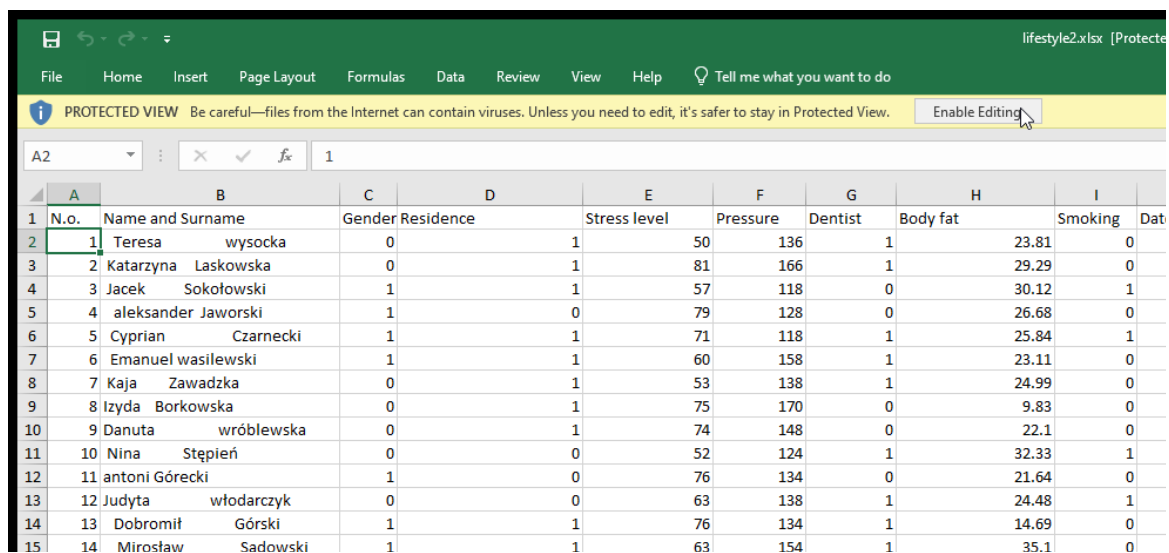
Exercise 1

In this exercise, we will prepare spreadsheets for further work. We will learn how to lock the first row of a spreadsheet.

Download the data file.

In a web browser, go to the Medical University of Białystok website. From the menu, select the "University" tab and navigate to "Faculty of Medicine". Select the "Units" field and click the "Department of Biostatistics and Medical Informatics" link. From the "Education" menu, select "Information Technology" under "MS Excel 2", click the "lifestyle 2.xlsx" link. Depending on the web browser used and its settings, the file may be saved in the default folder (e.g., Downloads) or elsewhere. We may also be asked to specify the destination folder. Save or move the downloaded file to your folder and open it in MS Excel.

The downloaded file may open in *Protected View*. Then click the *Enable Editing* button on the yellow bar to unlock the ability to edit the file [Fig. 1].



	A	B	C	D	E	F	G	H	I	J
1	N.o.	Name and Surname	Gender	Residence	Stress level	Pressure	Dentist	Body fat	Smoking	Date
2	1	Teresa wysocka	0	1	50	136	1	23.81	0	
3	2	Katarzyna Laskowska	0	1	81	166	1	29.29	0	
4	3	Jacek Sokołowski	1	1	57	118	0	30.12	1	
5	4	aleksander Jaworski	1	0	79	128	0	26.68	0	
6	5	Cyprian Czarnecki	1	1	71	118	1	25.84	1	
7	6	Emanuel wasilewski	1	1	60	158	1	23.11	0	
8	7	Kaja Zawadzka	0	1	53	138	1	24.99	0	
9	8	Izyda Borkowska	0	1	75	170	0	9.83	0	
10	9	Danuta wróblewska	0	1	74	148	0	22.1	0	
11	10	Nina Stępień	0	0	52	124	1	32.33	1	
12	11	antoni Górecki	1	0	76	134	0	21.64	0	
13	12	Judyta włodarczyk	0	0	63	138	1	24.48	1	
14	13	Dobromił Górski	1	1	76	134	1	14.69	0	
15	14	Miroslaw Sadowski	1	1	63	154	1	35.1	0	

Fig. 1 Disabling *Protected Mode*

The spreadsheet named "Data" of the "lifestyle2.xlsx" file contains information collected from a group of 200 employees at a certain company. The "Legend" spreadsheet provides explanations of the respondents' responses.

Freeze the first row of the sheet.

MS Excel allows you to freeze the first row or column to enhance data visibility and make it easier to navigate. Freezing the first row of a spreadsheet ensures that the headers are always visible, which is especially important when you work with a large number of rows and need to scroll down frequently.

To lock the first row, go to the *View* tab, then select the *Freeze Top Row* option from the *Freeze Panes* menu of the *Window* section [Fig. 2].

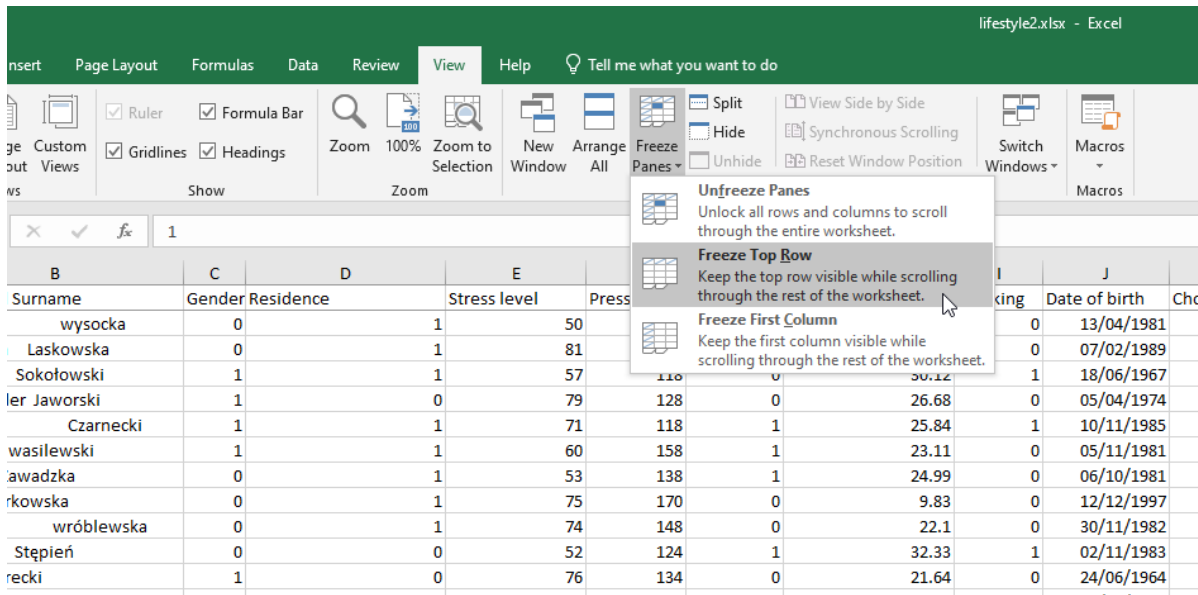


Fig. 2 Freezing the first row

Make a copy of the datasheet.

The following tasks will focus on applying functions from specific groups. For most of these, we will need a separate copy of the datasheet.

Right-click the "Data" sheet name and select the *Move or Copy...* command [Fig. 3]. In the *Move or Copy* window, select the following commands: (*move to end*) and *Create a copy* [Fig. 4]. Rename the copied sheet from "Data (2)" to "Lookup and count".

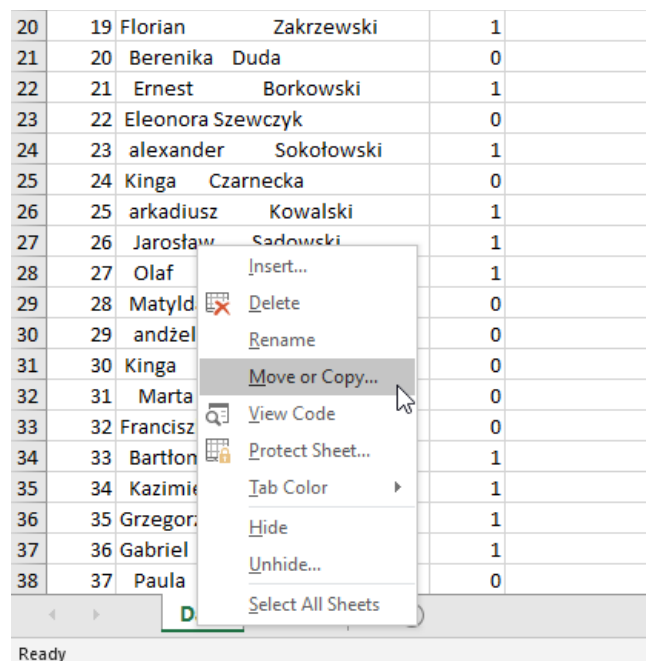


Fig. 3 *Move or Copy...* command

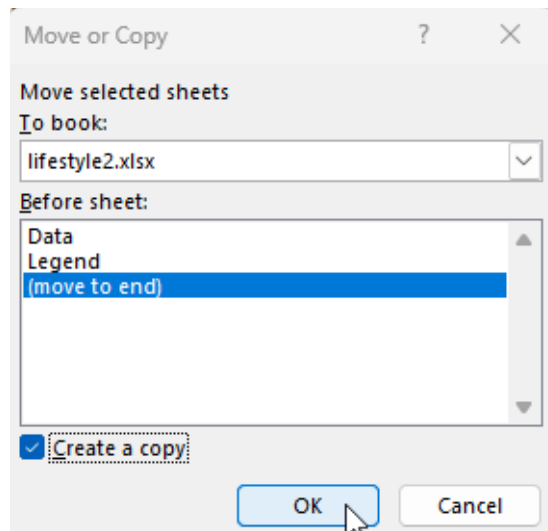


Fig. 4 *Move or Copy* window

Exercise 2

In this exercise, we will use search and counting functions.

Insert columns containing labels.

We will be performing this exercise in the "Lookup and count" spreadsheet. Right-click the name of column **D**, i.e., the one located next to the column labeled "gender". Once this is done, the entire column **D** should be selected. To insert a new column, select *Insert* from the menu [Fig. 5].

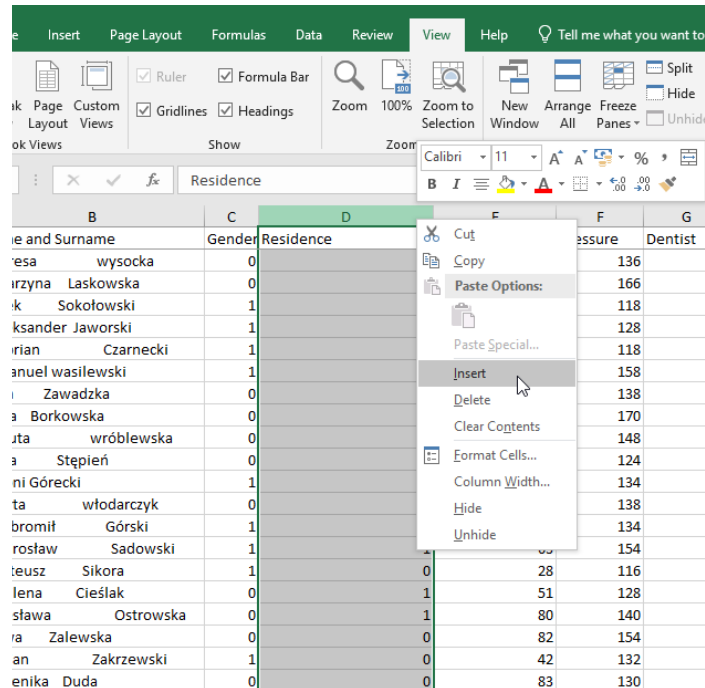


Fig. 5 *Insert* command

Create the name of the new column:

D1="Gender label"

Repeat the same steps to create the following column labels: "Residence label", "Dentist label" and "Smoking label" [Fig. 6].

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	N.o.	Name and Surname	Gender	Gender label	Residence	Residence label	Stress level	Pressure	Dentist	Dentist label	Body fat	Smoking	Smoking label	Date of birth	Cholesterol	Weight	Height
2	1	Teresa wysocka	0		1		50	136	1		23.81	0		13/04/1981	7.85	65	167
3	2	Katarzyna Laskowska	0		1		81	166	1		29.29	0		07/02/1989	4.03	81	175
4	3	Jacek Sokołowski	1		1		57	118	0		30.12	1		18/06/1967	4.34	93	181

Fig. 6 Adding a column

Fill the newly added columns with values from the dictionary.

*MS Excel contains two important functions for replacing values based on a dictionary: **VLOOKUP** and **HLOOKUP**. Both belong to the Lookup & Reference category [Fig. 7] (or Lookup & Reference [Fig. 8] on the Formulas tab in the Function Library section).*

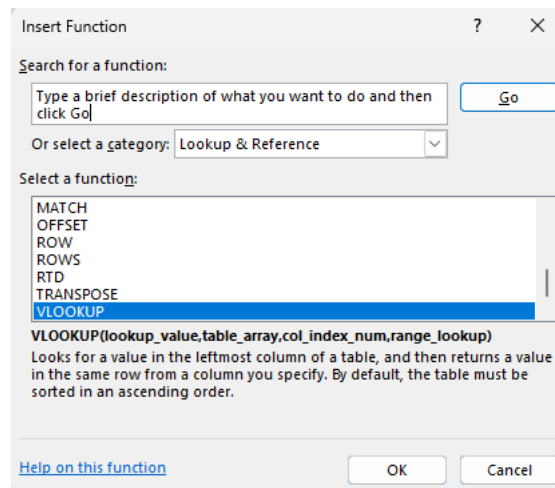


Fig. 7 *Insert* Function wizard

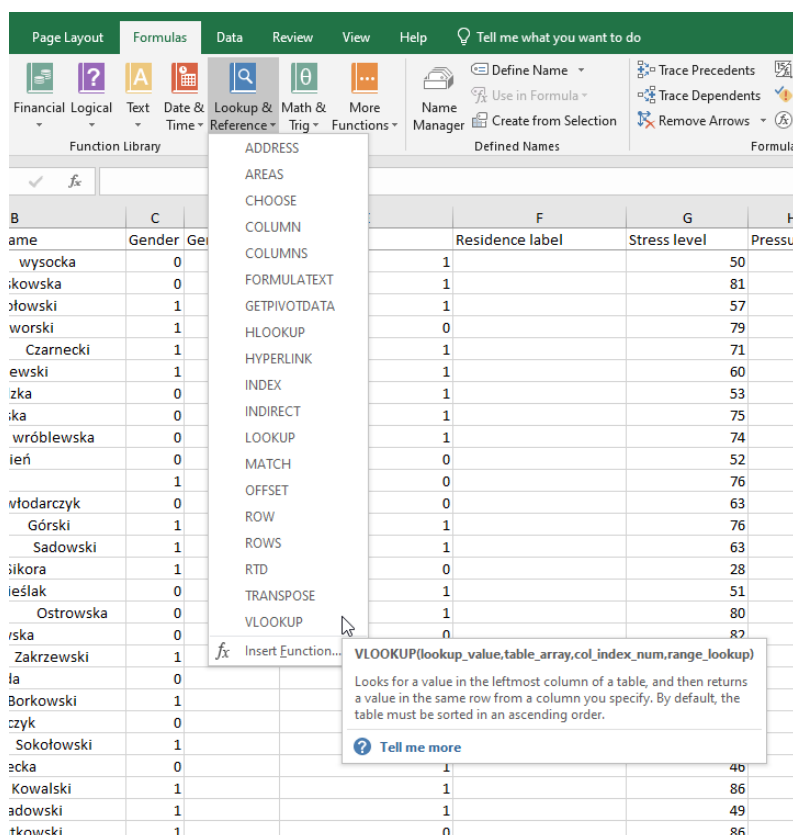


Fig. 8 Location of functions on the ribbon

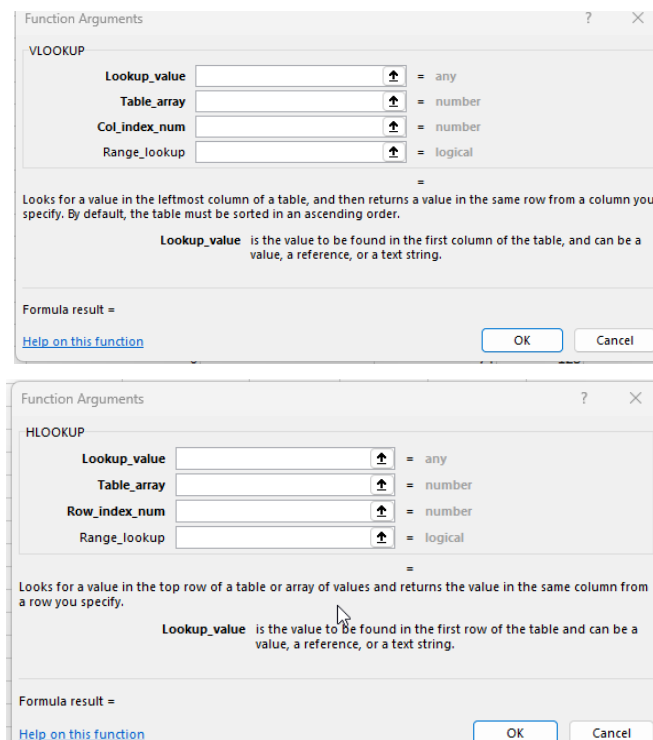


Fig. 9 Function Arguments wizard

[Fig. 9] shows the argument wizards for the **VLOOKUP** and **HLOOKUP** functions. The first argument (Lookup_Value or Reference) specifies the value or location of the value to be found in the dictionary. The next argument (Table_Array or Array) indicates the dictionary location and is a range of cells. Note that you are not specifying the dictionary header. The dictionary should contain the values being searched for in the first column or row. The target translation should be in one of the subsequent columns or rows. This is specified by the third argument (Col_Index_Num or Row_Number) of the function. The last argument (Lookup_range or Row) is a logical value (i.e., it takes one of only two values: **TRUE** (synonym – 1) or **FALSE** (synonym – 0), indicating whether an approximate or an exact match should be used during the search. For approximate searches, the search value is rounded down to its nearest value in the dictionary.

When using **VLOOKUP** (or **HLOOKUP**) for approximate searches, you need to ensure that the Table_array (or Array) argument is sorted in the ascending order and that Lookup_value is not smaller than the smallest value in the dictionary.

Go to cell **D2**. In the *Formulas* tab, in the *Function Library* section, from the *Lookup & Reference* menu, select the **VLOOKUP** function [Fig. 8]. The value we are looking for is the one in the adjacent cell, i.e., **C2**. The appropriate dictionary is contained in the "Legend" worksheet. Choose this worksheet in the *Table_array* edit and select the **A14:B15** area. Since the formula will be copied down, we need to lock the rows in the entered area. As the corresponding labels are contained in the second column, enter the value 2 in the *Colu_index_num* field. Set the last argument to **FALSE** [Fig. 10]. Copy the entered formula to the subsequent rows.

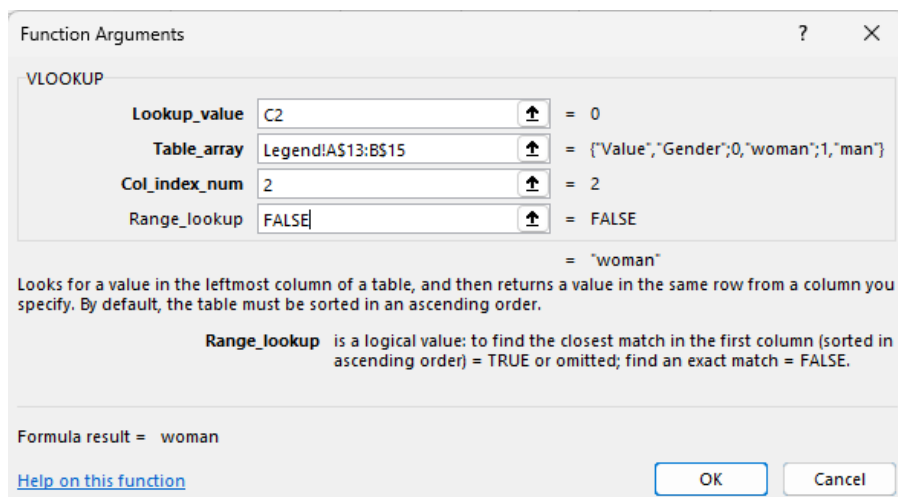


Fig. 10 VLOOKUP function arguments entered

Notice that the cell range entered from another sheet is preceded by the sheet name and an exclamation point. After selecting the appropriate area (or cell) in the wizard, MS Excel automatically referenced them correctly. This is a more complete syntax for referencing selected cells. It is worth noting that not only cells from another sheet can be referenced, but also cells from another workbook.

Proceed similarly in the following columns, entering the appropriate formulas:

F2=VLOOKUP(E2, Legend!D\$14:E\$15, 2, FALSE)

J2=VLOOKUP(I2, Legend!G\$14:H\$15, 2, FALSE)

M2=VLOOKUP(L2, Legend!J\$14:K\$15, 2, FALSE)

[Fig. 11] presents the final result.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	N.o.	Name and Surname	Gender	Gender label	Residence	Residence label	Stress level	Pressure	Dentist	Dentist label	Body fat	Smoking	Smoking label	Date of birth
2	1	Teresa wysocka	0	woman		1 urban	50	136	1	yes	23.81	0	no	13/04/15
3	2	Katarzyna Laskowska	0	woman		1 urban	81	166	1	yes	29.29	0	no	07/02/15
4	3	Jacek Sokotowski	1	man		1 urban	57	118	0	no	30.12	1	yes	18/06/15
5	4	aleksander Jaworski	1	man		0 rural	79	128	0	no	26.68	0	no	05/04/15

Fig. 11 The labeled data

Enter the labels to be counted.

Next to the data set, create a table containing the counts of employees of a given gender:

T4="woman"

T5="man"

Calculate the gender structure of the company.

MS Excel provides a function that counts the frequency of occurrence of a specific value, i.e., **COUNTIF**. This function belongs to the Statistical category (or Statistical in the More Formulas tab of the Function Library section). The first argument (Range) specifies the cell range in which non-blank cells are to be counted. The next argument (Criteria) specifies the criteria—expressed as a number, expression, or text—specifying which cells will be included in the count.

Enter the following formula [Fig. 12]:

U4=COUNTIF(D\$2:D\$201, T4)

and copy it to the cell below.

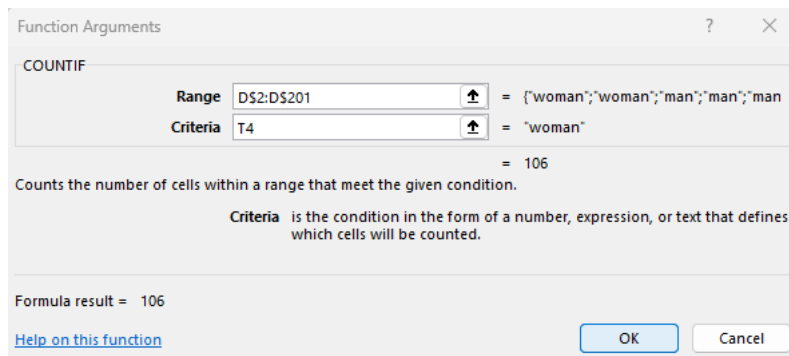


Fig. 12 COUNTIF function call

The following table is obtained: [Fig. 13]:

woman	106
man	94

Fig. 13 Table showing the number of respondents by gender.

Represent the structure on a pie chart.

Select the entire table area containing the counts [Fig. 14]. In the *Insert* tab, in the *Charts* section, select *Insert 3-D pie chart* [Fig. 15].

woman	106
man	94

Fig. 14 The selected table area

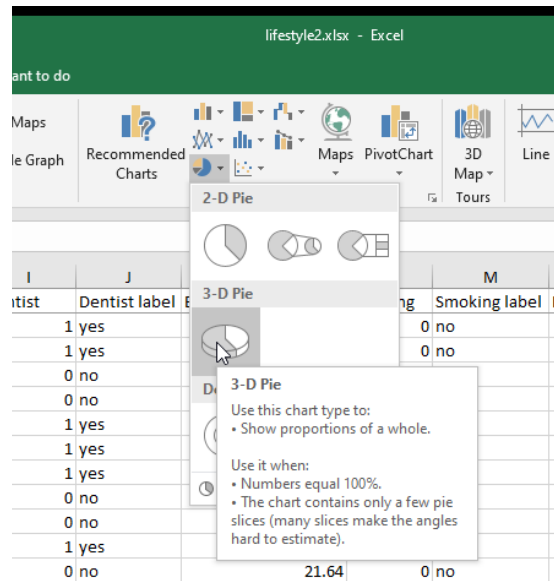


Fig. 15 Selecting the chart type

In the inserted chart, make the following changes:

- Change the chart title to "GENDER STRUCTURE" [Fig. 16].
- Add percentage values. To do this, click the "+" button located next to the chart. Then click the arrow located to the right of *Data Labels* and select *More Options...* from the drop-down menu [Fig. 17]. In the *Format Data Labels* dashboard, select the *Percentage* option and uncheck the *Value* box. Select *Center* as the label position [Fig. 18].
- Set the color and font size of the labels (color: *white*; size: *22 pts*). The appropriate options are located on the *Home* tab in the *Font* section. Label values must be selected when these changes are being made.
- Increase the font size of the title and the legend (*24 pts* and *16 pts*, respectively). The title or the legend must be selected when these changes are being made.

[Fig. 19] presents the final result of the introduced changes.

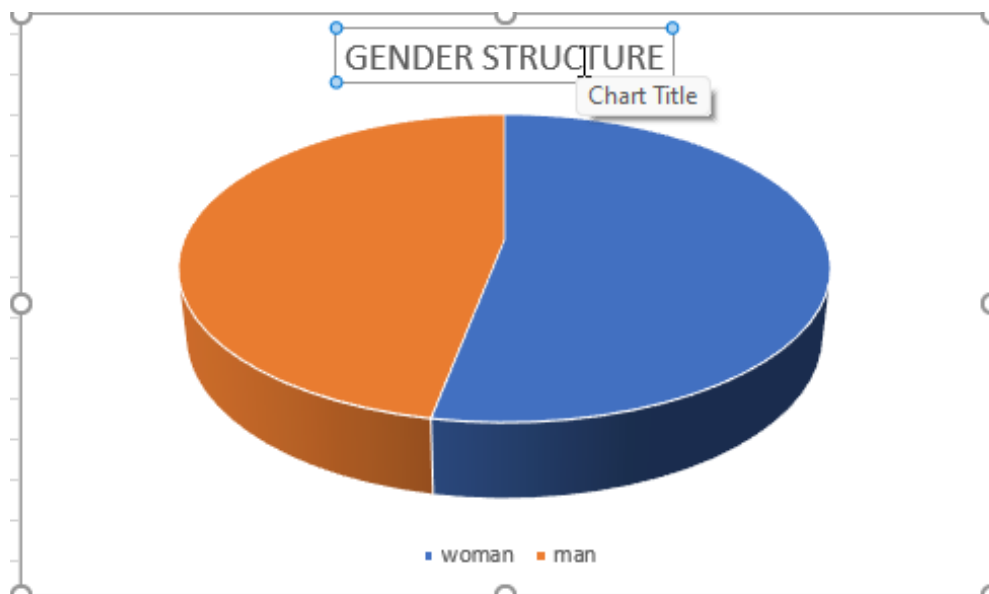


Fig. 16 Changing the title

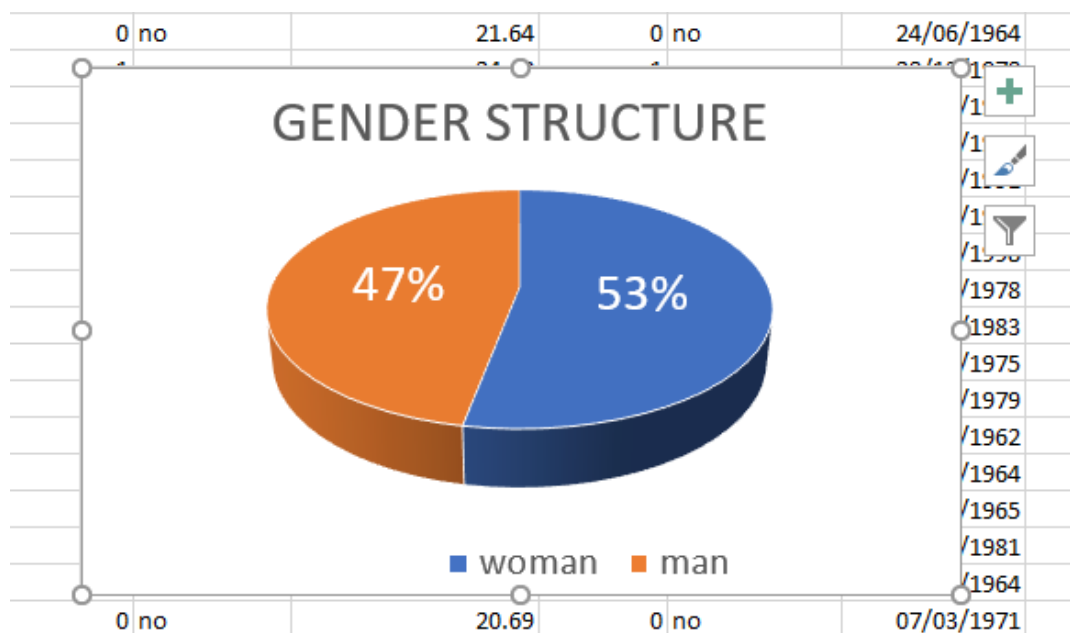


Fig. 19 The final result

Move the finished chart within the sheet to a position under the table containing the counts, so that it does not obscure the data.

Enter the labels to be counted.

Next to the data set, create a table containing the counts of employees of a given gender:

U8="smoker"
 U9="yes"
 V9="no"
 T10="woman"
 T11="man"

To increase the readability of a table, some of its cells are often merged.

To center the "smoking" label in the table, merge cells **U8** and **V8** and center the text. Select both cells and, from the *Home* tab, select the *Merge & Center* command from the *Alignment* section.

Calculate the gender structure of the company in relation to cigarette smoking.

*When working with data, you may need to count the occurrences of more than one condition. **COUNTIFS** is the appropriate function for this type of task. It belongs to the Statistical category (or Statistical on the Formulas tab of the Function Library section) and takes any number of arguments [Fig. 20], divided into pairs. Each pair of arguments has the same meaning as the first and second argument of the **COUNTIF** function.*

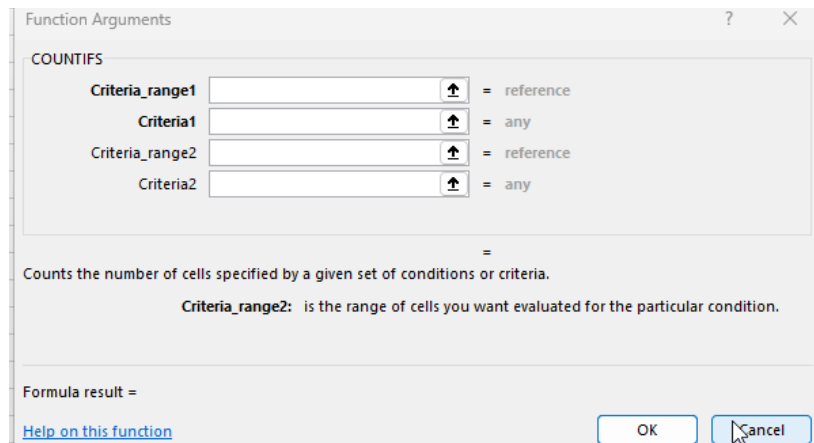


Fig. 20 COUNTIFS function wizard

The conditions for counting are located in the first row and the first column of the newly created table. When entering the formula into cell U10 (where female smokers will be counted), we must remember to block the addresses accordingly [Fig. 21], so that the completed formula can be copied to the remaining cells. Enter the following formula:

U10=COUNTIFS(\$D\$2:\$D\$201,\$T10,\$M\$2:\$M\$201,U\$9)

and copy it to the cells below and to the right. The final result is as follows [Fig. 22]

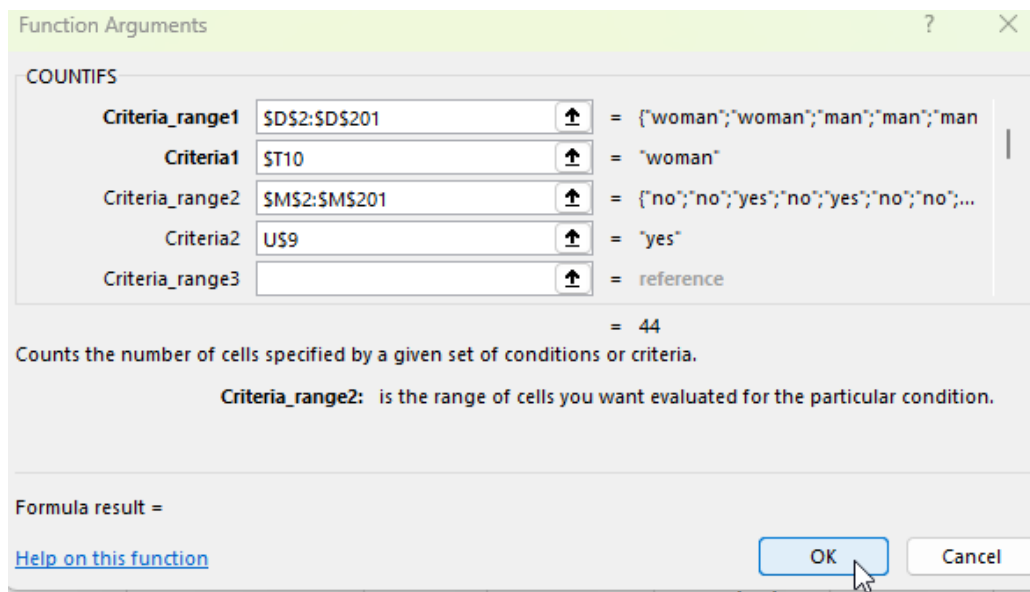


Fig. 21 COUNTIFS function call

	smoker		
	yes	no	
woman	44	62	
man	50	44	

Fig. 22 Gender structure of the company in relation to cigarette smoking

Exercise 3

In this exercise, we will use logical functions.

Make a copy of the datasheet.

Make another copy of the "Data" spreadsheet and rename it "Logical". We will continue the exercise in it. If necessary, the relevant instructions are included at the end of Exercise 1.

Enter the names of the next columns.

Create the names of the following columns:

N1="Height category"

O1="BMI"

P1="Correct BMI"

Q1="BMI category"

Calculate the height category.

*The **IF** function belongs to the Logical category (also found on the Formulas tab in the Function Library section). It allows you to perform a logical test, i.e., checking the validity of a Logical_test condition. It can be true or false (met or not met). [Fig. 23] shows the function's arguments. If the specified condition is true, the function returns the Value_if_true value; otherwise, it returns the Value_if_false value. In other words, the function displays one of the two values in the cell where it has been entered: Value_if_true or Value_if_false, depending on whether the logical test has been met or not.*

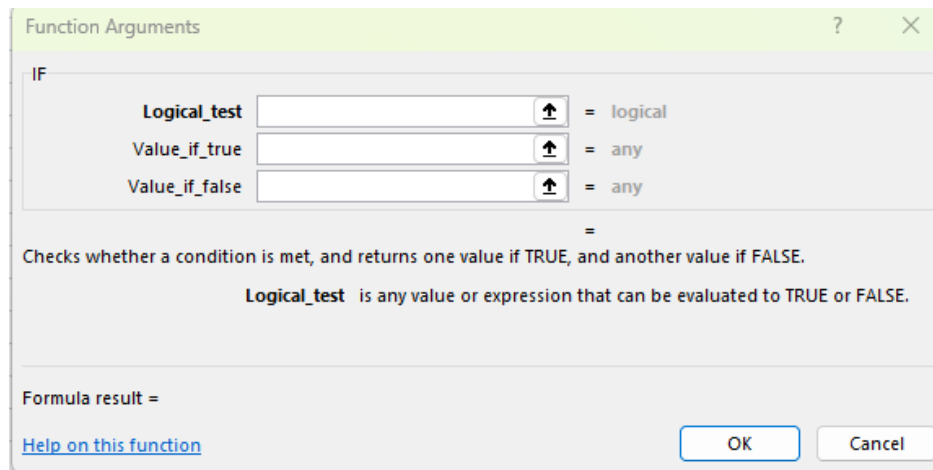


Fig. 23 **IF** function wizard

We assume 170 as the cutoff height. Taller individuals are labeled "tall", shorter individuals, "short". Enter the formula shown in [Fig. 24] and copy it to the cells below:

N2=IF(M2>170,"tall","short")

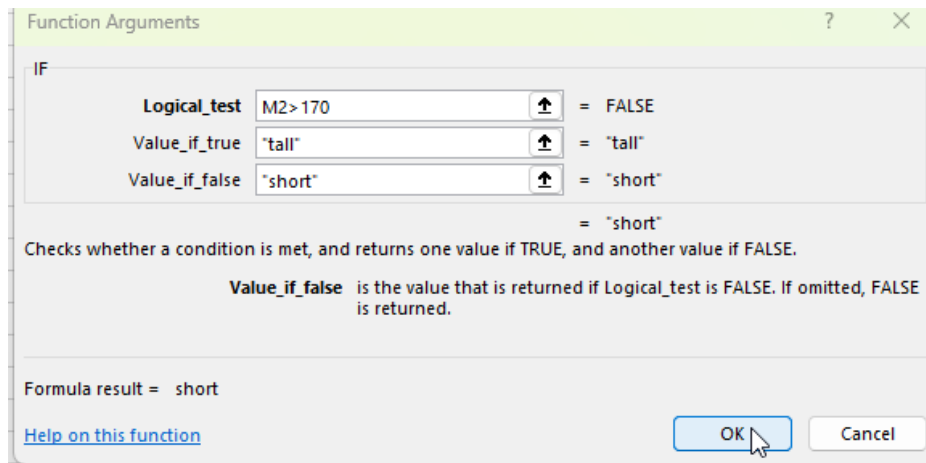


Fig. 24 IF function call

Calculate BMI.

For each respondent, calculate the body mass index (BMI) using the following formula:

$$BMI = \frac{weight(kg)}{height^2 (m)}$$

The correct formula is:

O2=L2/(M2/100)^2

The survey data provide height in centimeters, so when substituting the values into the BMI formula, remember to divide the height by 100 to convert the unit to meters. Copy the formula into the cells below.

Check if the BMI values are within the normal range.

*If you need to check whether more than one condition are met at the same time, the **AND** function can be used. If you need at least one of multiple conditions to be met, the **OR** function can be used.*

The BMI categories are presented in [Fig. 25]. The normal range of values is between 20 and 25:

Underweight	BMI < 20
Normal	20 < BMI < 25
Overweight	25 < BMI < 30
Obesity	30 < BMI

Fig. 25 BMI categories

We must not transfer the above inequalities directly into the formula (i.e., do not enter =20<=BMI<25), because the software would interpret it as an error. Instead, the **AND** function must be used [Fig. 26]:

P2=AND(20<=O2,O2<25)

Copy it to the cells below.

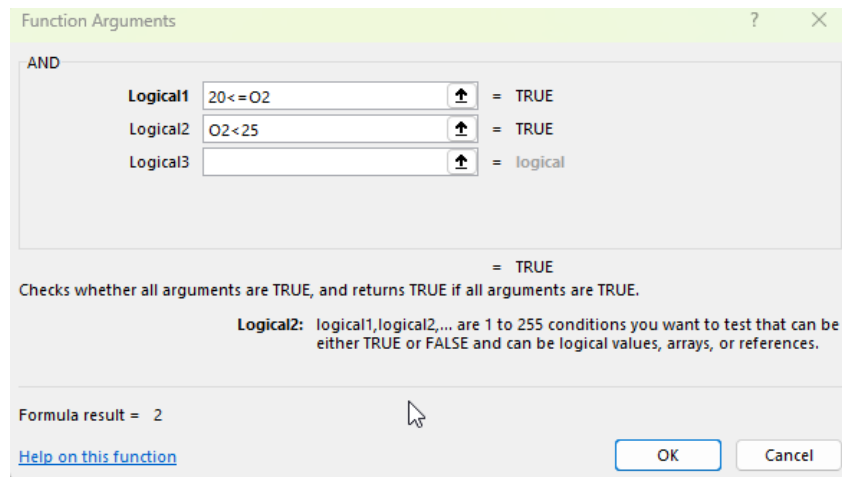


Fig. 26 **AND** function call

Determine the BMI category of individual employees.

To determine the BMI category, we will again use the **VLOOKUP** function. Four categories will be determined based on the calculated BMI value, as shown in [Fig. 25]:

Enter the BMI value dictionary.

Enter the values of the following cells:

S4=0

T4="underweight"

S5=20

T5="normal"

S6=25

T6="overweight"

S7=30

T7="obesity"

Enter the following formula [Fig. 27], remembering to block the appropriate addresses. Copy the following formula to the cells below:

Q2=VLOOKUP(O2,S\$4:T\$7,2,TRUE)

Function Arguments
?
X

VLOOKUP

Lookup_value	O2	=	23.30668005
Table_array	\$S\$4:\$T\$7	=	{0,"underweight";20,"correct";25,"over
Col_index_num	2	=	2
Range_lookup	TRUE	=	TRUE

= "correct"

Looks for a value in the leftmost column of a table, and then returns a value in the same row from a column you specify. By default, the table must be sorted in an ascending order.

Range_lookup is a logical value: to find the closest match in the first column (sorted in ascending order) = TRUE or omitted; find an exact match = FALSE.

Formula result = correct

[Help on this function](#)
OK
Cancel

Fig. 27 VLOOKUP function call

Exercise 4

In this exercise we will use conditional formatting.

Note: we will be performing this exercise in the same sheet ("Logical") as the previous one.

Conditional formatting contrasts data visually, making them much easier to work with. Only cells that meet specific conditions are formatted. These conditions can be applied multiple times within a given area. All conditional formatting options can be found on the Home tab, in the Styles section, in the Conditional Formatting group. To remove conditional formatting from an area, you simply need to select the area and, in the Conditional Formatting group, choose Clear Rules, then Clear Rules from Selected Cells. You can also remove conditional formatting from an entire worksheet as well as manage conditional formatting for any area within a workbook (Manage Rules).

Format the cells of the "BMI category" column depending on their values.

Select the whole "BMI category" column, apart from the header. On the *Home* tab, select *Conditional Formatting*, then the *Cell Highlight Rules* and *Equal to...* menus [Fig. 28].

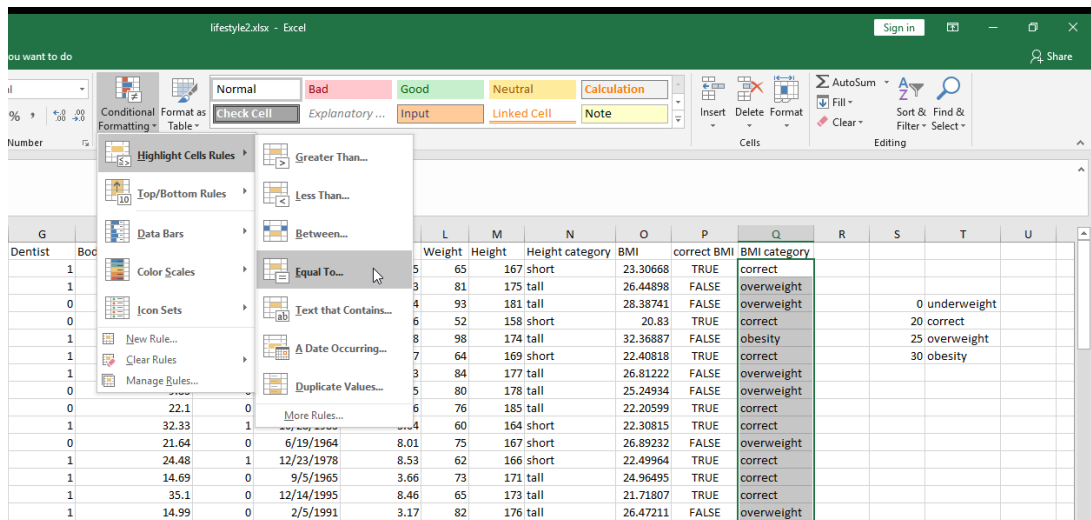


Fig. 28 Equals... option

In the *Equals* window, enter "obesity" and select *Light Red Fill with Dark Red Text* [Fig. 29].

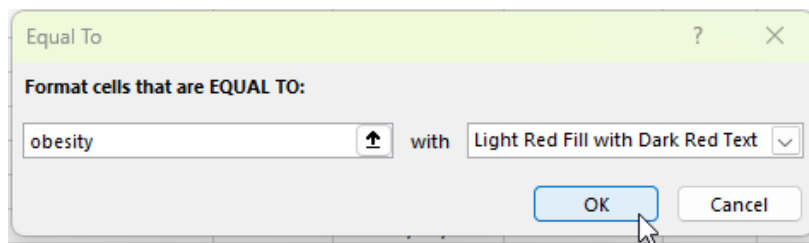


Fig. 29 Options for highlighting the "obesity" group

Proceed similarly for the groups:

- "overweight": Yellow Fill with Dark Yellow Text,
- "normal": Green Fill with Dark Green Text.

For the "underweight" group, select the *Custom Format...* option. In the *Format Cells* window, on the *Font* tab, select *Blue color* [Fig. 30]. In the *Fill* tab, select the *light blue* background color [Fig. 31]. Confirm by clicking the *OK* button.

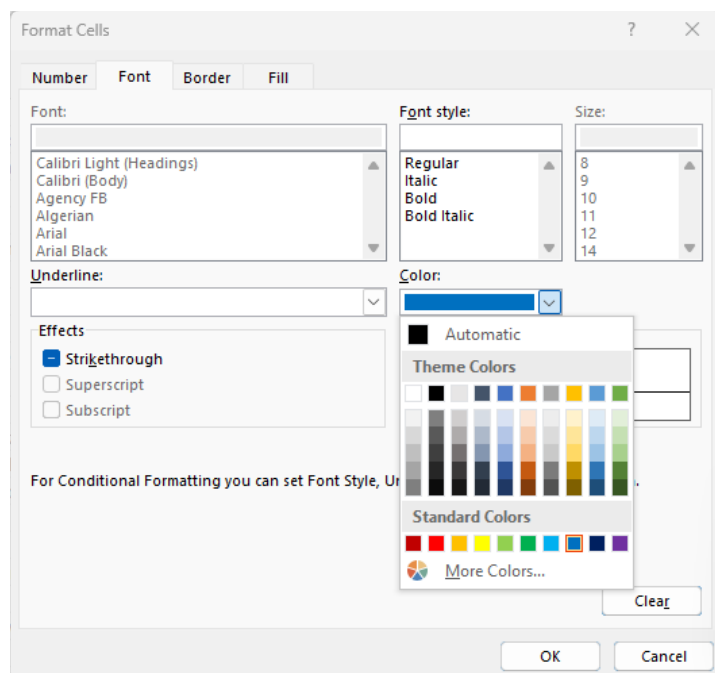


Fig. 30 Customizing the font color

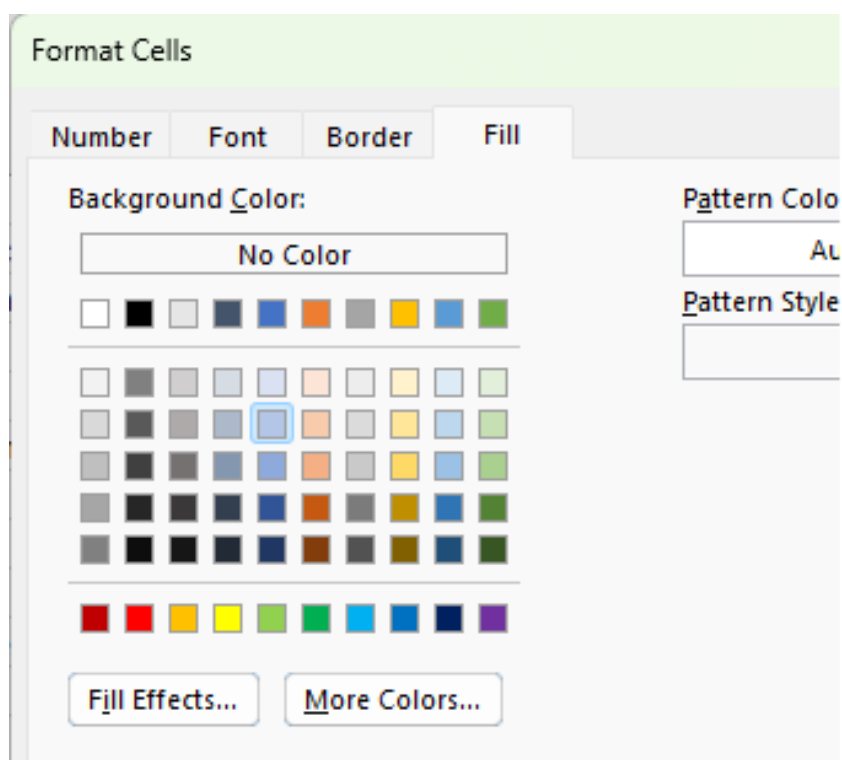


Fig. 31 Cell background color

Add data bars in the "Weight" column.

As before, select the whole "Weight" column, apart from the header. In the *Home* tab, select *Conditional Formatting*, then the *Data Bars* menu and the *Blue Data Bar* in the *Gradient Fill* group [Fig. 32].

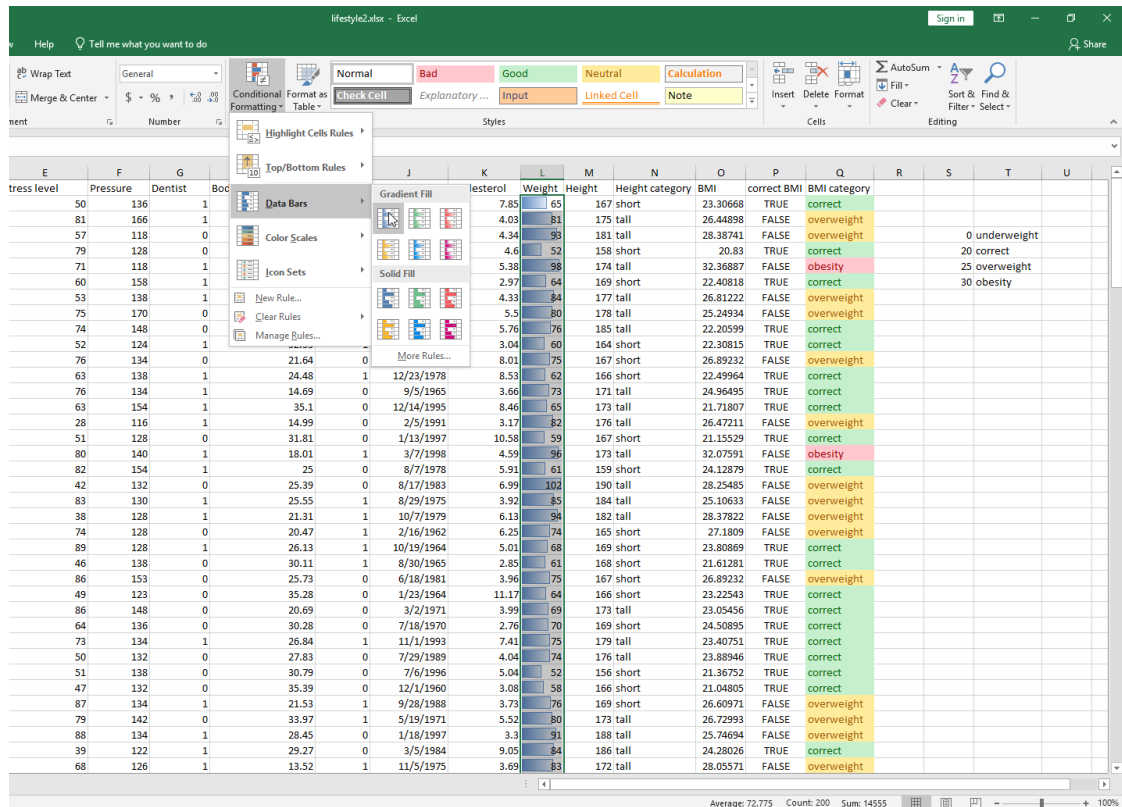


Fig. 32 The blue data bar

Exercise 5

In this exercise, we will use text functions.

Databases are unfortunately very susceptible to a number of random errors that occur during their creation. These can be the result of both human error (e.g., simple mistakes) and machine error (e.g., conversion errors).

Make a copy of the datasheet.

The database being analyzed contains the names and surnames of 200 employees, but the quality of the entered data is not satisfactory. First of all, there are extra spaces between some of the first and last names, as well as before a number of first names (possibly also after some last names, but it is impossible to tell by eye). Furthermore, some of the first and last names begin with a lowercase letter or contain capital letters in the middle. Therefore, our task in subsequent exercises is to make the appropriate corrections. Manually correcting these errors would be extremely time consuming, which is why we will use a number of text functions offered by MS Excel for this purpose.

Create another copy of the "Data" spreadsheet and rename it "Text". We will perform the rest of the exercise there. Add the names of the following columns:

N1="Removed spaces"
O1="Space position"
P1="Name"
Q1="Length"
R1="Surname"
S1="First letter of name"
T1="First letter of surname"
U1="Initials"
V1="Corrected initials"
W1="Name big letter"
X1="Surname big letter"
Y1="Name and surname"

Remove unnecessary spaces between employee names and surnames.

*A very common error when editing text is inserting excessive spaces between words – before, or after the text. Like any other error, excessive spaces can prove problematic during data analysis, because the same expressions, but one with extra spaces and the other without, can be misinterpreted as different data. MS Excel provides a function dedicated to removing extra spaces, i.e., **TRIM**. It belongs to the Text category (also found on the Formulas tab in the Function Library section). The function takes one argument (Text) that specifies the text or address from which the extra spaces will be removed.*

We will begin organizing employee names by removing unnecessary spaces. Enter the following formula [Fig. 33]:

N2=TRIM(B2)

Copy it to the cells below. Column N now contains employee names with unnecessary spaces removed.

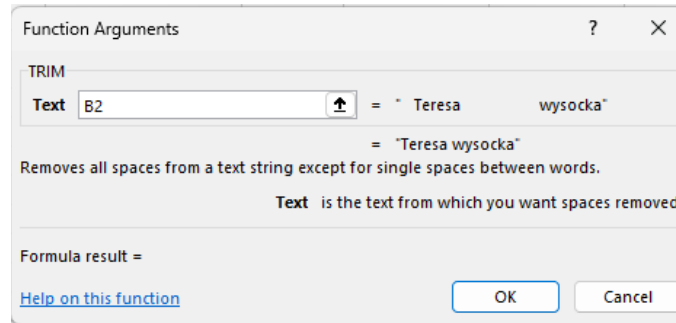


Fig. 33 **TRIM** option call

Find the position where there is a space between the employees' names and surnames.

*Conditional cell formatting of can depend on the presence of a specific text phrase, which is only part of the cell's full content. The **FIND** function is used to search for one text string within another and returns the starting position number of the searched text. This function belongs to the Text category (also found on the Formulas tab in the Function Library section) and takes three arguments. The first (Find_text) specifies the text (phrase) to be found. The second (Within_text) specifies the text containing the text to be found. The third (Start_num) specifies the character position at which the search should begin (if you want to start the search from the beginning of the text, this argument can be left blank). If the search phrase is not found at the specified position, an error is returned.*

The next task is to find the position where the space between the employee's name and surname is located and place it in the **O** column of the spreadsheet. This information will be used in the subsequent steps of the exercise. For this purpose, we will use the **FIND** function. Enter the following formula [Fig. 34]:

O2=FIND(" ",N2)

Copy it to the cells below. Column O now contains the sequence number of the position where there is a space between the employee name and surname.

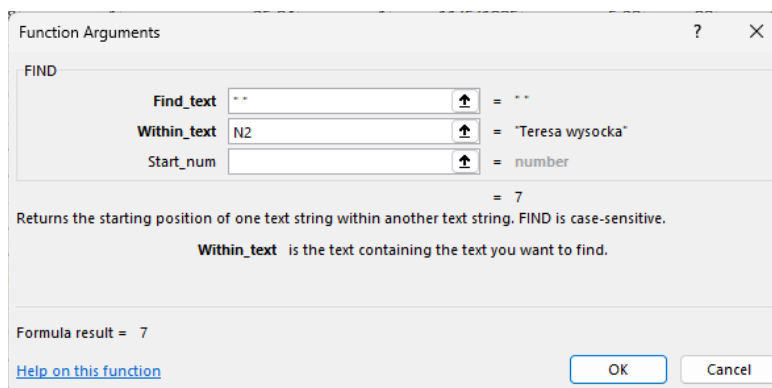


Fig. 34 **FIND** function call

Extract employee names.

If you want to extract a text fragment of a specified length from the beginning of an expression, you can use the **LEFT** function. It belongs to the Text category (also on the Formulas tab in the Function Library section) and returns at least one first character in a text string based on the specified number of characters. The function takes two arguments. The first (Text) specifies the text containing the string you want to extract. The second (Num_chars) specifies the number of characters to be extracted by the function. If omitted, it defaults to 1.

The next task is to extract employee names and place them in column **P** of the spreadsheet. We will use the **LEFT** function for this purpose. We already know the position of the space, so we can use this information to determine the number of consecutive characters in column **N** that make up the employee's name. Since we do not want the function to extract the name along with the space following it, we must remember to reduce the value of the *Num_chars* argument by 1 compared to the values specified in column **O**. Enter the following formula [Fig. 35]:

P2=LEFT(N2,O2-1)

and copy it to the cells below.

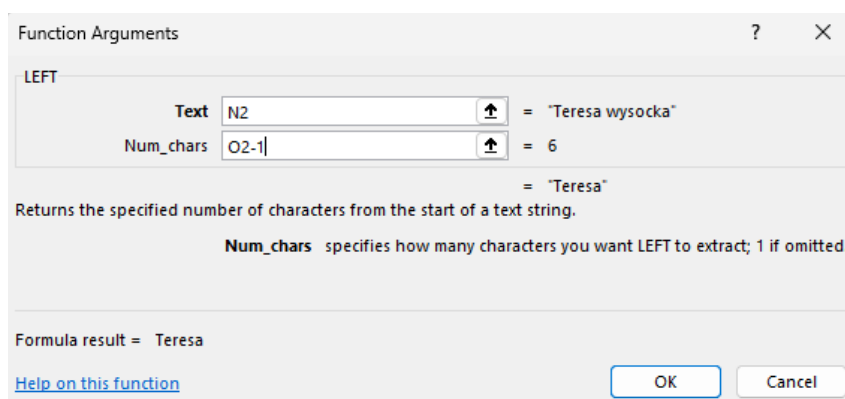


Fig. 35 **LEFT** function call

Specify the number of characters in employee names and surnames.

To calculate the length of a text, the **LEN** function is used. It belongs to the Text category (also found on the Formulas tab in the Function Library section) and is used to count the number of characters in a specified cell. The function takes only one argument (Text), which specifies the text whose length is to be calculated.

To extract employee names, we need data that will allow us to determine where the last name begins in the text. For this purpose, we will use the length of the first and surnames, including spaces, and the **LEN** function, and place the result in column **Q** of the spreadsheet. Enter the following formula [Fig. 36]:

Q2=LEN(N2)

Copy it to the cells below. Column **Q** now contains the character counts of the phrases that make up the employee's first and last names.

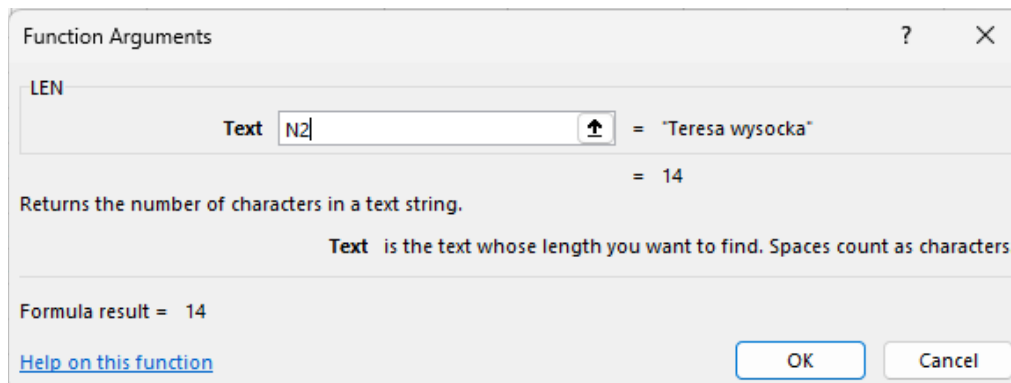


Fig. 36 **LEN** function call

Extract employee names.

Similarly to the **LEFT** function, the **RIGHT** function is used to extract a fragment of text from the end of a word. It belongs to the Text category (also in the Formulas tab, Function Library section) and takes two arguments. The first (Text) specifies the text containing the string of characters you want to extract. The second (Num) specifies the number of characters to be extracted by the function. If omitted, it defaults to 1.

MS Excel also has another text function, **MID.TEXT**, which allows you to extract a fragment of text, which is its first argument. The next two arguments specify the position within the original text and the length of the returned fragment. However, when you need to extract the last letter or a sequence of last letters, it is much easier to use the **RIGHT** function.

The next task is to extract the employee names and place them in column **R** of the spreadsheet. For this purpose, we will use the **RIGHT** function. We now know the position of the space and the total number of characters in the phrase "name and surname plus space", so we can use this information to extract the surname. The length of the surname will be the difference between the length of the full phrase "name and surname plus space" and the position of the space in the same phrase. Enter the following formula [Fig. 37]:

R2=RIGHT(N2,Q2-O2)

and copy it to the cells below.

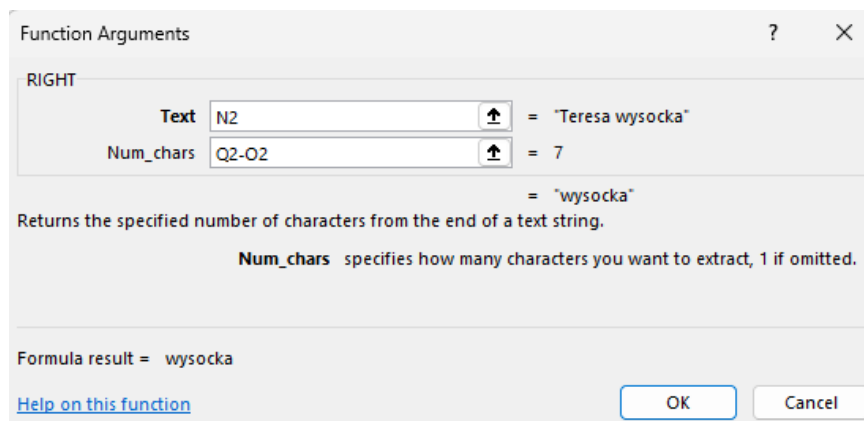


Fig. 37 **RIGHT** function call

Create employee initials.

With the employees' first and last names separated, their initials can be created easily. We will use the **LEFT** function again.

To create the initials, we will be using columns with separated first and last names. The function's *Num_chars* attribute will be left blank, as its default value is 1. Enter the following formulas:

S2=LEFT(P2)

T2=LEFT(R2)

and copy them to the cells below.

Combine the initials of the employees.

*The **CONCAT** function is the equivalent of the & (concatenation) character discussed earlier. It belongs to the Text category (also found on the Formulas tab in the Function Library section) and makes it possible to combine texts from different cells into a single text string. The function takes any number of arguments (Text), each of which is a text string or a range to be combined into a single text string.*

Since we want to ensure that the initials do not appear in separate columns, we will use the **CONCAT** function to concatenate them. We will also add periods after each letter. Enter the following formula [Fig. 38]:

U2=CONCAT(S2, ".", T2, ".")

and copy it to the cells below.

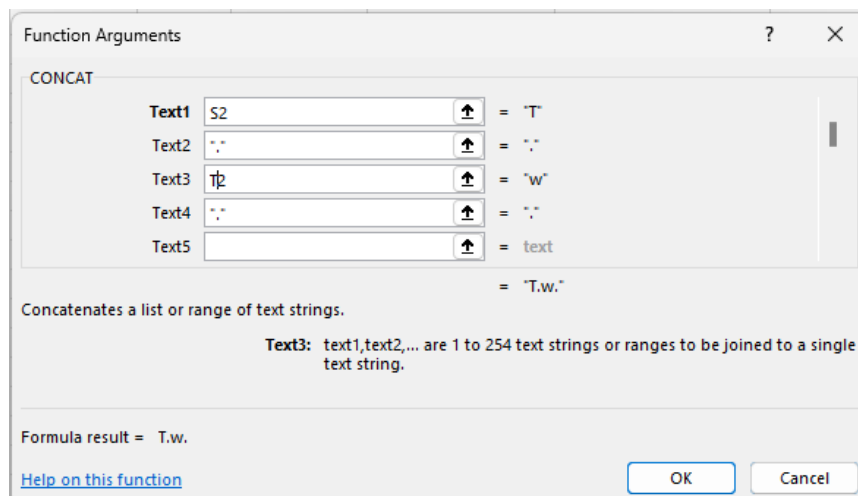


Fig. 38 **CONCAT** function call

Convert all initials to uppercase letters.

*The **UPPER** function allows you to convert letters in a text to uppercase. It is accompanied by a similar **LOWER** function. Both of these functions belong to the Text category (also found on the Formulas tab in the Function Library section) and take a single argument (Text) indicating a string or a range of characters to be converted in its entirety to uppercase or lowercase.*

We can see that column U contains initials, some of which are not capitalized. We can convert all lowercase letters to uppercase using the **UPPER** function. Enter the following formula [Fig. 39]:

V2=UPPER(U2)

and copy it to the cells below.

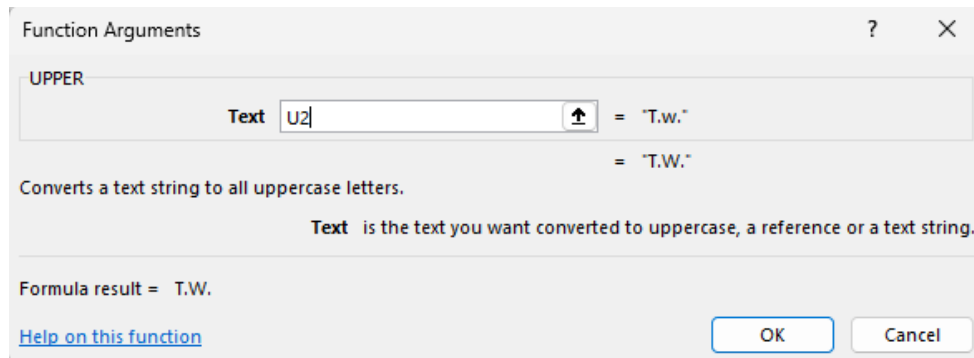


Fig. 39 **UPPER** function call

Capitalize employee names and surnames.

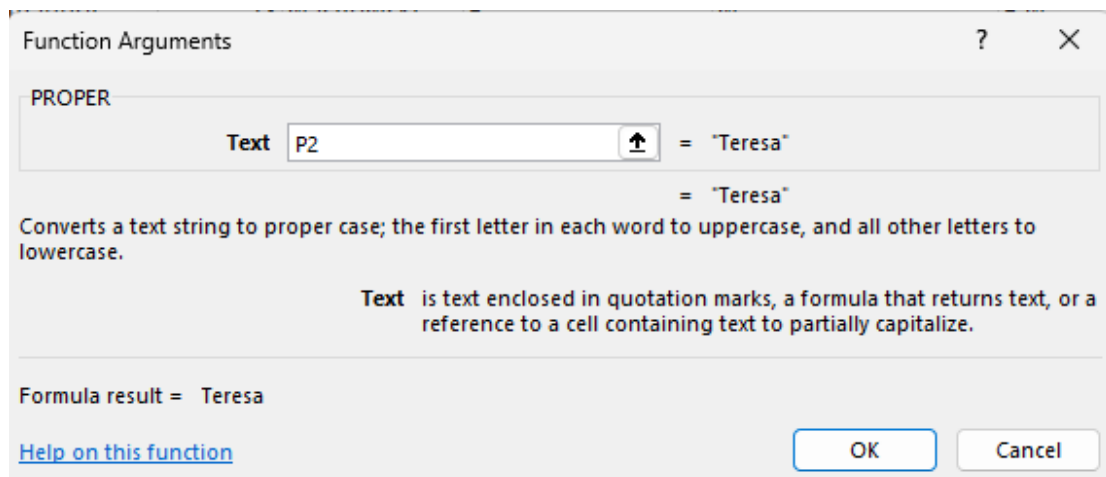
*In English and Polish, first names, last names, and proper names begin with a capital letter, followed by a lowercase letter. The **PROPER** function makes it possible to convert the first letter in a text to uppercase and the remaining letters to lowercase. It belongs to the Text category (also on the Formulas tab in the Function Library section) and takes a single argument (Text) indicating a string or a range for which the desired case conversion should occur.*

As noted earlier, some first and last names in the database are spelled incorrectly. Since we need words to begin with a capital letter, we will use the **PROPER** function. Enter the following formulas [Fig. 40]:

W2=PROPER(P2)

X2=PROPER(R2)

and copy them to the cells below.



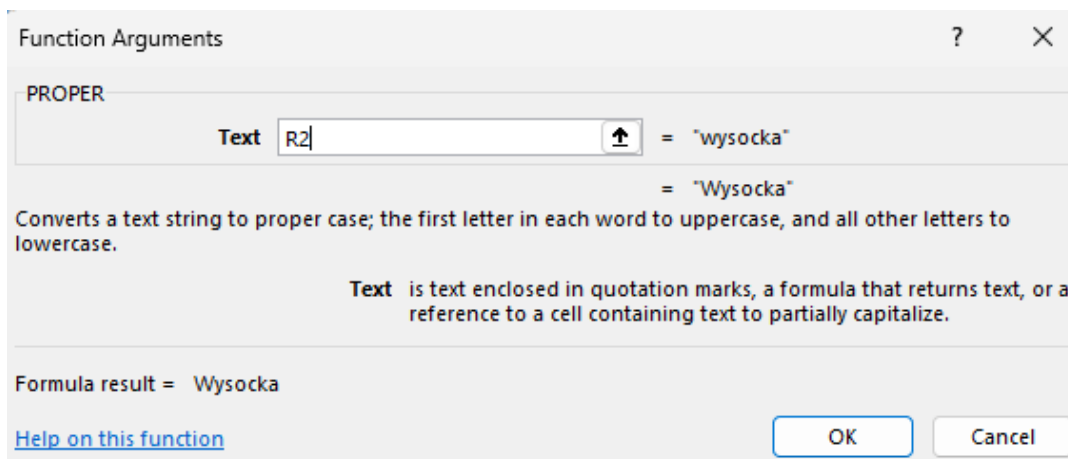


Fig. 40 **PROPER** function call

Create a list of employees in the following format: Name and Surname.

The **TEXTJOIN** function is used to combine text strings from multiple ranges. Its unique feature is the ability to use a delimiter (e.g., a space or a comma) between the individual texts being combined (i.e., there is no need to add a delimiter in the form of a separate text to be joined, as is the case with the **CONCAT** function). If no delimiter is specified, the function will concatenate the texts. The **TEXTJOIN** function belongs to the Text category (also found on the Formulas tab in the Function Library section). It takes at least four arguments. The first one (Delimiter) specifies a character or a string to be inserted between each text element. The second (Ignore_empty) enables default ignoring of empty cells (value TRUE). The remaining arguments (Text1 to Text252) specify the text strings or ranges to be joined.

Although we now have all the components needed to create a list of employees in the Name and Surname format, they are still located in separate cells. To combine all the elements into a single text, we will use the **TEXTJOIN** function. Note that the delimiter is a space. Enter the following formula [Fig. 41]:

Z2=TEXTJOIN(" ",,W2,X2)

and copy it to the cells below.

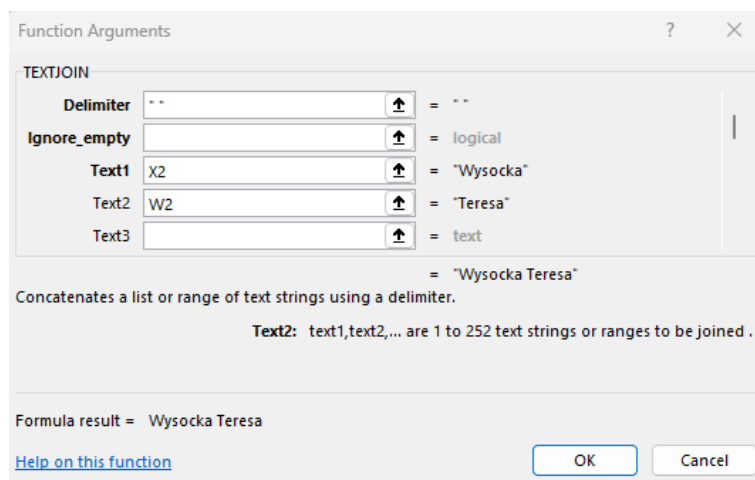


Fig. 41 **TEXTJOIN** function call

Exercise 6

In this exercise, we will use date and time functions.

The way dates and times are represented in MS Excel was discussed in detail in the previous tasks. Similarly to the number of days, a date makes it possible to perform arithmetic operations and comparisons. While using the calendar for calculations is not straightforward, various functions available in the spreadsheet can be helpful.

Create a copy of the datasheet.

Create another copy of the "Data" sheet and rename it "Date and time". We will perform the rest of the exercise in this file. If necessary, relevant instructions are provided at the end of Exercise 1.

Note: the dates in the "Date of Birth" column are different from those shown in the screenshots. However, the differences between subsequent dates in the rows are preserved. All individuals are between 24 and 66 years old.

Extract the components of the date of birth.

We can extract the components of a cell containing a date or time using the following functions:

- **YEAR**
- **MONTH**
- **DAY**
- **hour**
- **MINUTE**
- **SECOND**

*These functions belong to the Date & Time category (also found on the Formulas tab in the Function Library section). They take a single argument (a number, usually formatted as a date or time) and extract a specific component, consistent with its name. [Fig. 42] shows the **YEAR** function wizard.*

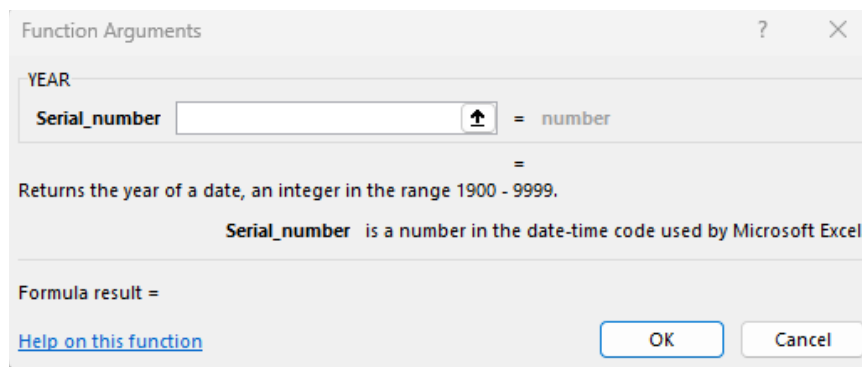


Fig. 42 **YEAR** function wizard

Enter the names of the following columns:

N1="Year"

O1="Month"

P1="Day"

Enter the appropriate formulas to calculate the selected date components [Fig. 43]:

N2=YEAR(J2)

O2=MONTH(J2)

P2=DAY(J2)

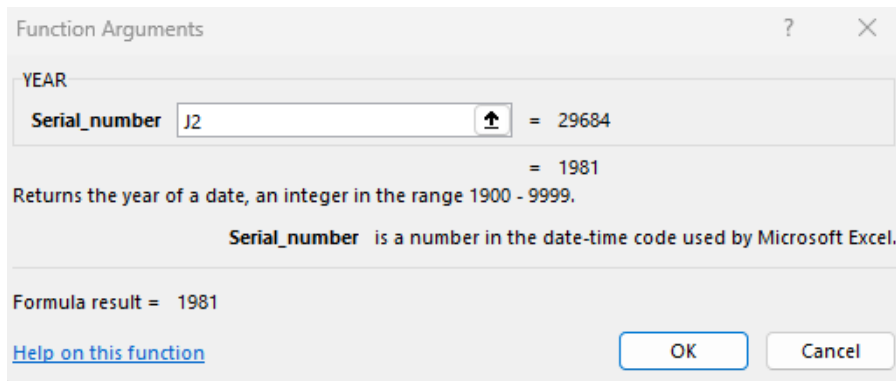


Fig. 43 **YEAR** function call

Calculate the current year.

*You can always calculate the current date using the **TODAY** function. If you are also interested in the exact time, you can use the **NOW** function. Both functions belong to the Date & Time category (also found on the Formulas tab in the Function Library section) and take no arguments. [Fig. 44] shows the **TODAY** function wizard.*

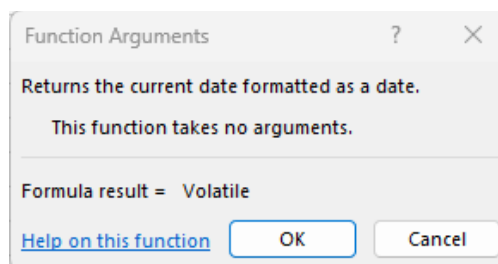


Fig. 44 **TODAY** function wizard

Complete the following cells:

W6="Current date"

X6="Current year"

We will calculate the current date using the **TODAY** function. Extract the current year using the **YEAR** function.

W7=TODAY()

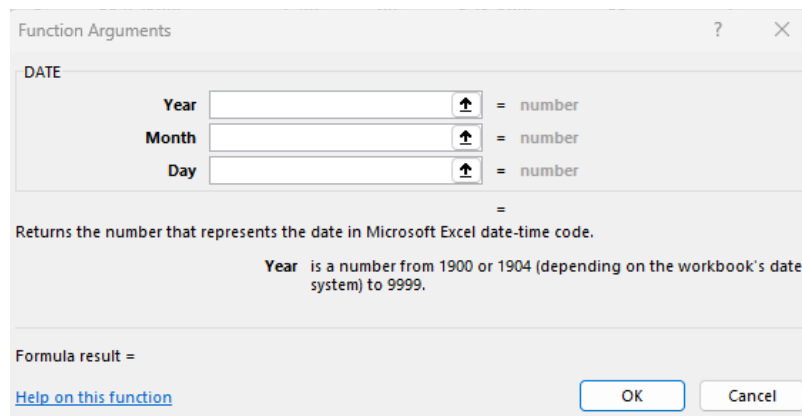
X7=YEAR(W7)

Calculate the date of birth for the current year.

*We have previously learned how to extract date and time components. The **DATE** & **TIME** functions are used to combine them. They belong to the Date & Time category (also found on the Formulas tab in the Function Library section) and take the following three arguments:*

- **DATE** function: Year, Month, Day [Fig. 45],
- **TIME** function: Hour, Minute, Second.

Note: entering numbers outside the specified ranges will result in an error.

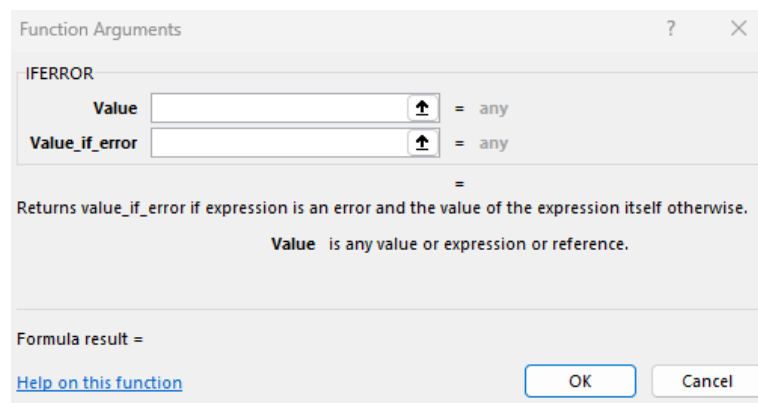


The image shows the 'Function Arguments' dialog box for the DATE function. The title bar says 'Function Arguments'. Inside, the function name 'DATE' is at the top. Below it, there are three input fields: 'Year', 'Month', and 'Day'. Each field has a small icon to its right. To the right of each field is an equals sign followed by the text '= number'. Below these fields, there is a description: 'Returns the number that represents the date in Microsoft Excel date-time code.' and a note: 'Year is a number from 1900 or 1904 (depending on the workbook's date system) to 9999.' At the bottom, there is a 'Formula result =' field, a 'Help on this function' link, and 'OK' and 'Cancel' buttons.

Fig. 45 **DATE** function wizard

The result of a formula may be an error. The term "error" in MS Excel is not the same as the colloquial meaning of "error", i.e., it is not, for example, a spelling error. In MS Excel, an error occurs due to incorrect formula syntax, an impossible mathematical operation, incorrect addressing, or an inappropriate data type.

*The **IFERROR** function, from the Logical category (also found on the Formulas tab in the Function Library section), can be helpful in such situations. It takes two arguments. If the first argument (Value) does not contain an error, the Value argument is returned as the result of the function call. Otherwise, the value specified in the Value_if_error argument is returned [Fig. 46].*



The image shows the 'Function Arguments' dialog box for the IFERROR function. The title bar says 'Function Arguments'. Inside, the function name 'IFERROR' is at the top. Below it, there are two input fields: 'Value' and 'Value_if_error'. Each field has a small icon to its right. To the right of each field is an equals sign followed by the text '= any'. Below these fields, there is a description: 'Returns value_if_error if expression is an error and the value of the expression itself otherwise.' and a note: 'Value is any value or expression or reference.' At the bottom, there is a 'Formula result =' field, a 'Help on this function' link, and 'OK' and 'Cancel' buttons.

Fig. 46 **IFERROR** function wizard

Enter the name of the next column:

Q1="Birthday"

According to the custom at the studied company, a person born on February 29th of a leap year celebrates their birthday on March 1st in a non-leap year. Note that if the current year is not a leap year, the February 29th date is invalid and will therefore cause an error when calling the **DATE** function. For this reason, we will use the **IFERROR** function with the appropriate arguments [Fig. 47].

Using the **DATE** function, we will combine the components of the date of birth (day and month) with the current year (cell **X7**). Using the **IFERROR** function allows us to check whether the current year is a leap year, thus finding people born on February 29th and replacing their day and month of birth with March 1st. Enter the following formula:

Q2=IFERROR(DATE(X\$7,O2,P2),DATE(X\$7,3,1))

Notice that cell **Q2** does not display the result as a date. To display the formula result correctly, we need to change the cell formatting. Select cell **Q2** and—in the *Home* tab, in the *Number* section to the right of the *Number Format* field—click the arrow and select the *Short Date* format from the drop-down menu [Fig. 48]. Then, copy the formula down.

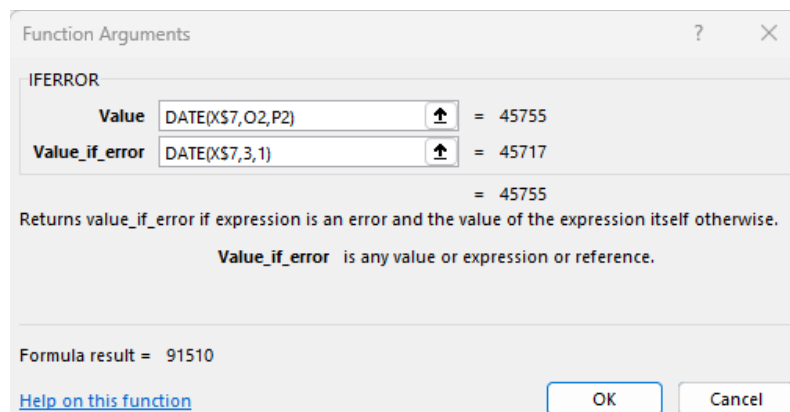


Fig. 47 Calculating the date of birth for the current year

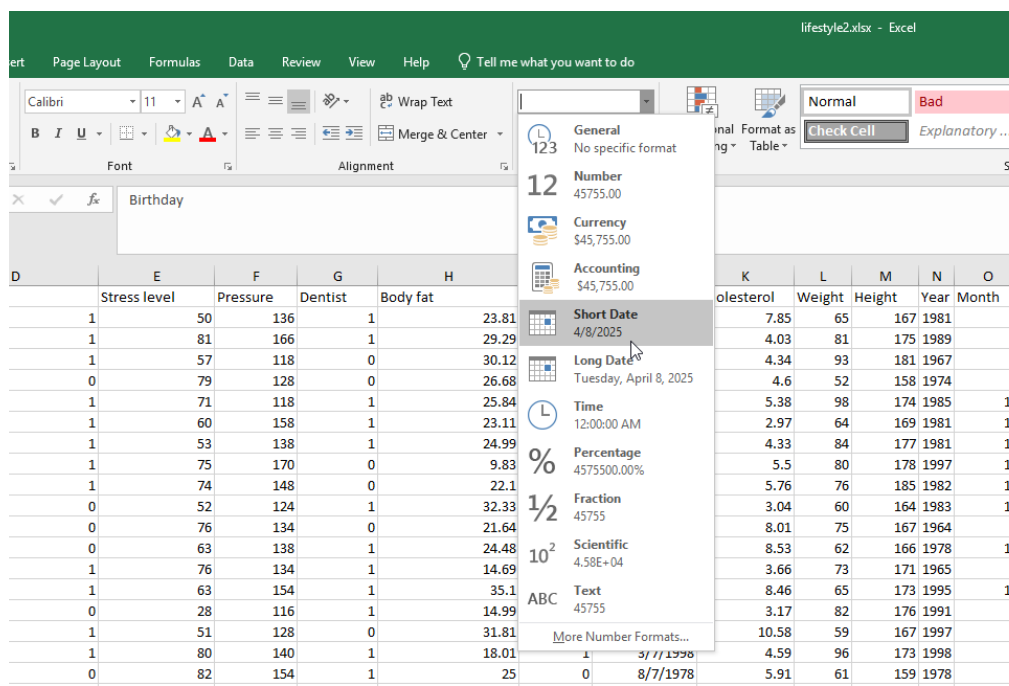
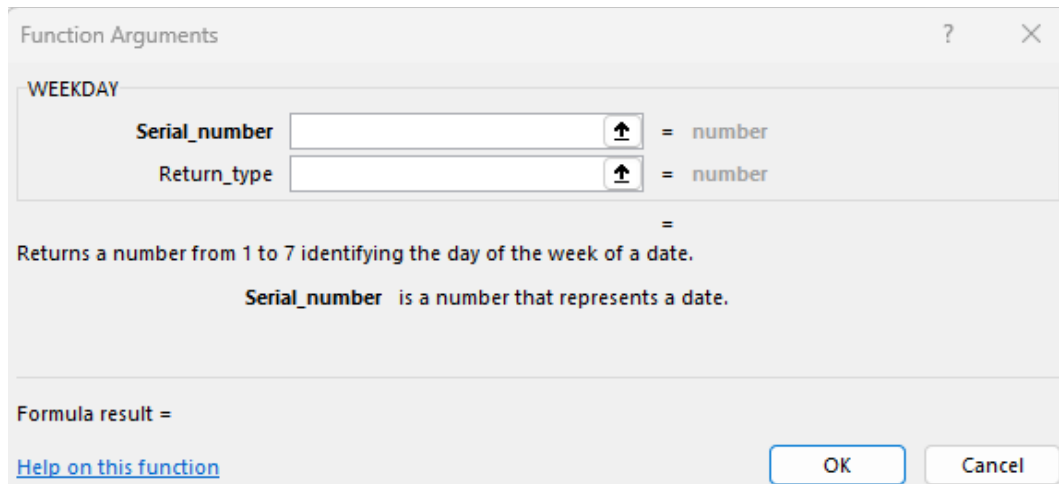


Fig. 48 *Short Date* format

Determine the day of the week of the birthday falls on in the current year.

*A week consists of seven days. The **WEEKDAY** function returns the ordinal number corresponding to the specified day of the week. This function belongs to the Date & Time category (also found on the Formulas tab in the Function Library section) and takes two arguments. The first one is the date, and the second (Return_Type) specifies the weekday numbering [Fig. 49]. If omitted, it defaults to 1.*



The screenshot shows the 'Function Arguments' dialog box for the **WEEKDAY** function. The 'Serial_number' field is empty, and the 'Return_type' field is also empty. Below the fields, it says 'Returns a number from 1 to 7 identifying the day of the week of a date.' and 'Serial_number is a number that represents a date.' At the bottom, there is a 'Formula result =' field, a 'Help on this function' link, and 'OK' and 'Cancel' buttons.

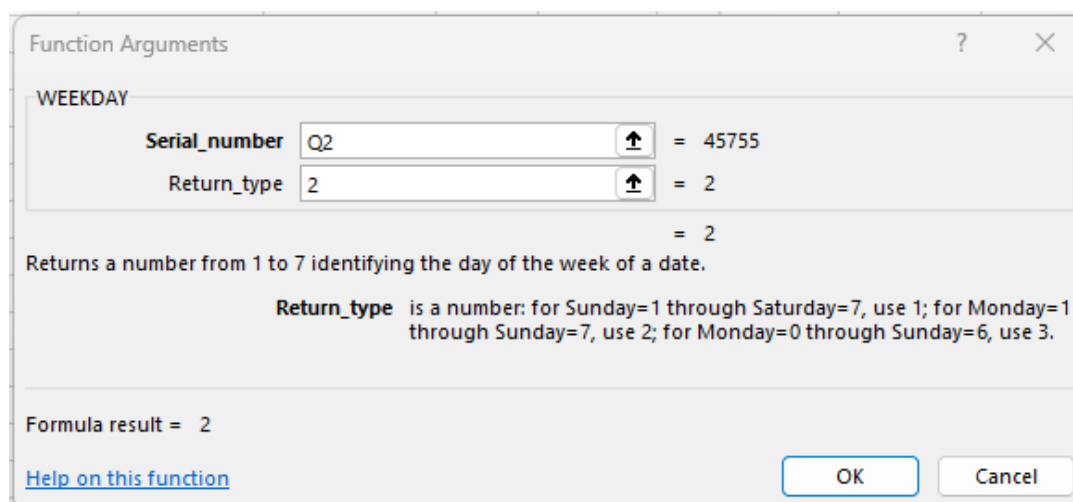
Fig. 49 **WEEKDAY** function wizard

Enter the name of the next column:

R1="Nr of the day of week"

We will calculate the ordinal number of the day of birth [Fig. 50] (we enter 2 as the second argument, which this means that the numbering of weekdays starts from Monday):

R2=WEEKDAY(Q2,2)



The screenshot shows the 'Function Arguments' dialog box for the **WEEKDAY** function. The 'Serial_number' field now contains 'Q2', and the 'Return_type' field contains '2'. Below the fields, it says 'Returns a number from 1 to 7 identifying the day of the week of a date.' and 'Return_type is a number: for Sunday=1 through Saturday=7, use 1; for Monday=1 through Sunday=7, use 2; for Monday=0 through Sunday=6, use 3.' At the bottom, the 'Formula result =' field shows '2', and there is a 'Help on this function' link, and 'OK' and 'Cancel' buttons.

Fig. 50 **WEEKDAY** function call

Based on the determined ordinal number of the day of the week, we will use the **VLOOKUP** function to determine the name of the day of the week. To do this, we will create a dictionary. Fill in the following cells:

V9=1

W9="Monday"
V10=2
W10="Tuesday"
V11=3
W11="Wednesday"
V12=4
W12="Thursday"
V13=5
W13="Friday"
V14=6
W14="Saturday"
V15=7
W15="Sunday"

Then enter the name of the next column and the appropriate formula:

S1="Birthday day"
S2=VLOOKUP(R2,V\$9:W\$15,2)

Calculate how many days are left until the birthday and format the results accordingly.

Complete the name of the next column:

T1="How many days left"

We will calculate how many days remain until the birthday in the current year. If the birthday has already passed, we will obtain a negative number. As a date is a number of days, simple subtraction can be used to find the corresponding difference in days:

T2=Q2-W\$7

Because cells **Q2** and **W7** are set to the *Date* format, the result in cell **T2** retains the same format, so we need to change it manually. Select cell **T2** and—in the *Home* tab, in the *Number* section, to the right of the *Number* Format field—click the arrow and select the *General* format from the drop-down menu [Fig. 51]. Then, copy the formula down.

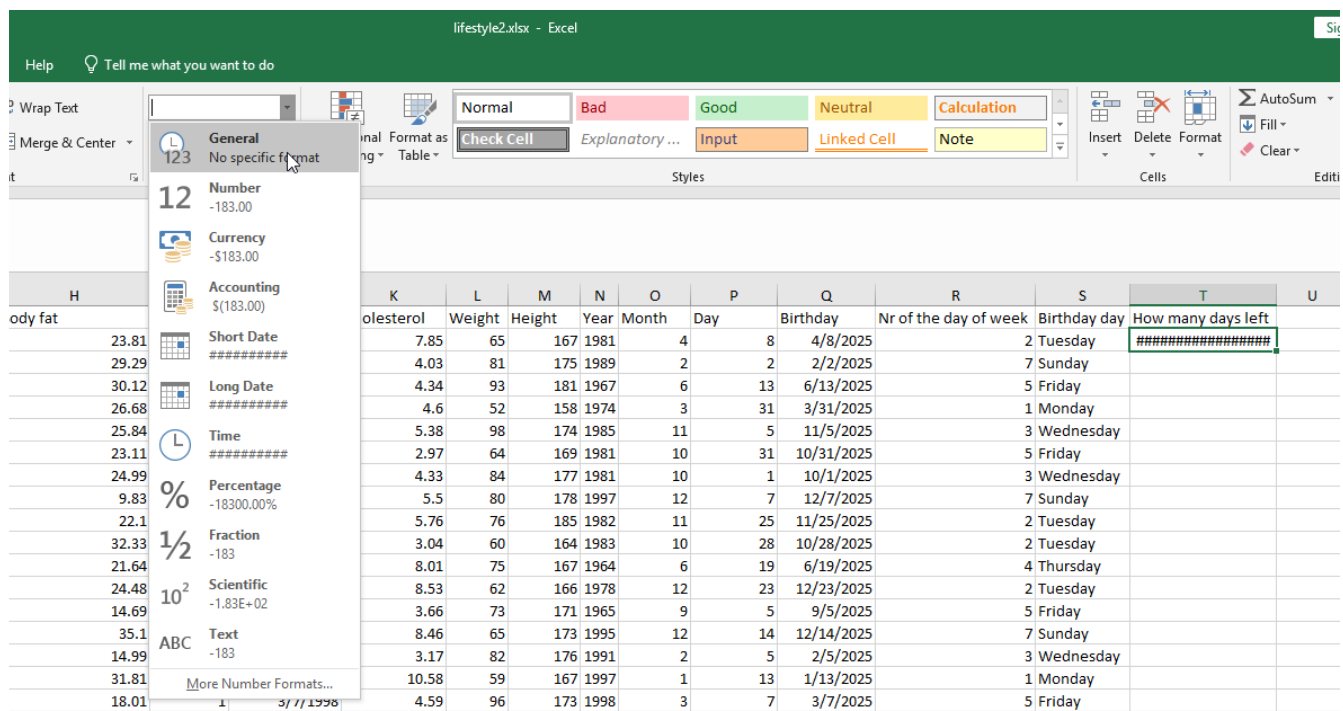


Fig. 51 General format

Finally, we will introduce a color-coded timeline for the birthday. The closer the birthday, the more noticeable the coloring will be:

- When the birthday is 7 or more days away, apply the *Green Fill with Dark Green Text* formatting;
- When the birthday is less than a week away, apply the *Yellow Fill with Dark Yellow Text* formatting;
- When the birthday falls on the current day, apply the *Light Red Fill with Dark Red Text* formatting.

Select the values in column T (without the header) and repeatedly use the *Conditional Formatting* command (by selecting the appropriate option from the *Home* tab, *Styles* section, *Conditional Formatting* command, *Highlight Cells Rules* menu) to set the following conditions:

- For values greater than 6, select the *Green Fill with Dark Green Text* formatting;
- For values between 1 and 6, select *Yellow Fill with Dark Yellow Text* formatting;
- For values equal to 0, select the *Light Red Fill with Dark Red Text* formatting.

Calculate the age at the time of the birthday.

Fill in the name of the next column:

U1="Which birthday"

Since we are looking for a difference in years, simple subtraction (between the current year and the year of birth) will yield an appropriate value. Enter the following formula and copy it down:

U2=X\$7-N2

Exercise 7

In this exercise, we will learn how to activate an MS Excel add-in.

MS Excel is equipped with two important add-ins: Solver and Data Analysis. Solver solves optimization problems (i.e., finding the minimum or maximum) of formulas by modifying selected cells. It can also be used to solve equations. The Data Analysis add-in contains a large number of statistical functionalities useful in data analysis. Once enabled, these add-ins typically remain active until disabled, meaning that the application remembers this setting. A major operating system update may disable them, however, in which case they have to be reactivated. If enabled, each add-in is visible on the Data tab [Fig. 52].

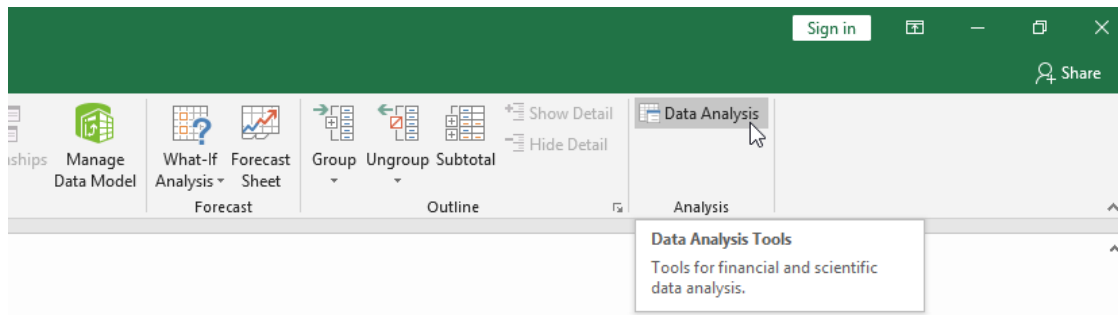


Fig. 52 *Data Analysis* command

Check the activation status of the *Data Analysis* add-in.

Go to the *Data* tab. If the *Data Analysis* command in the *Analysis* section is available [Fig. 52], proceed to the next exercise. Otherwise, continue with the current one.

Activate the *Data Analysis* add-in.

In the *File* tab, select *Options*. In the window that appears, open the *Add-ins* tab [Fig. 53], which displays the *Active Application Add-ins* and *Inactive Application Add-ins* lists.

Note: the *Data Analysis* add-in (shown as *Analysis ToolPak*) may be present on the *Active Application Add-ins* list, but the *Data Analysis* command may still be unavailable on the *Data* tab. In this case, disable and then re-enable the *Analysis Toolpak* add-in. To do this, select *Excel Add-ins* from the *Manage* list and click the *Go...* button [Fig. 53].

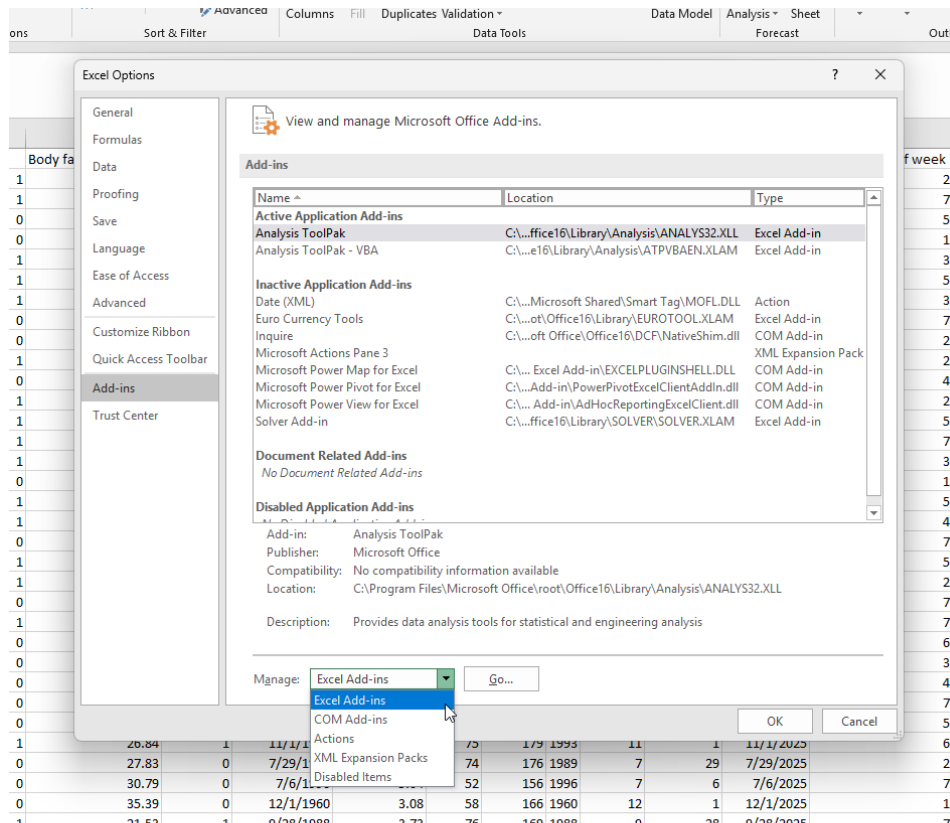


Fig. 53 Active and inactive application add-ins

In the *Add-ins* window [Fig. 54], check that the *Analysis ToolPak* option is selected. If it is, deselect it, click *OK*, and again click the *Go...* button in the previous window. Now, in the *Add-ins* window, select the checkbox next to the *Analysis ToolPak* option and click *OK*.

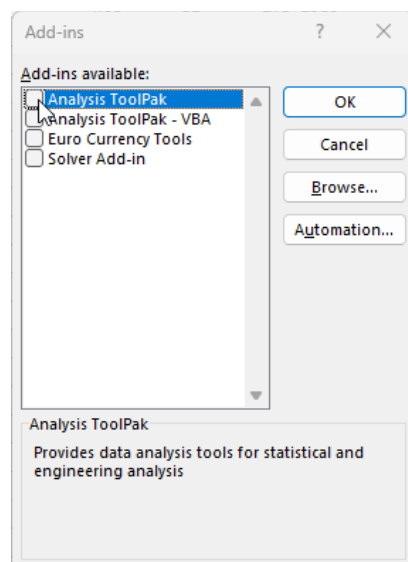


Fig. 54 *Add-ins* window

Exercise 8

In this exercise, we will calculate selected descriptive statistics.

Descriptive statistics are used to describe the most important information about the variables analyzed in a study. They are used to determine the number of observations, mean scores, variation among observations, among others. Descriptive statistics include:

- *measures of occurrence, e.g., number of observations, cumulative percentage,*
- *measures of location, e.g., mean, median, modal,*
- *measures of variability, e.g., standard deviation, variance, kurtosis,*
- *measures of asymmetry, e.g., skewness.*

Most parameters included in descriptive statistics can be calculated using appropriate functions from the Statistical category (or the Statistical function on the Formulas tab in the Function Library section). However, when using the data analysis module, they are all stored under a single command, so you do not need to execute each one individually.

Prepare the data.

Make another copy of the "Data" spreadsheet and rename it "Data Analysis". If necessary, relevant instructions are provided at the end of Exercise 1.

Since calculating descriptive statistics only makes sense for numerical data, we need to delete other types of data. In order to do it, right-click each of the column labels, i.e., **J, I, G, D, C, B,** and **A** (maintain the specified deletion order to avoid problems with changing the column labels), and select *Delete* [Fig. 55]. After the task is completed, the following columns should be seen: "Stress level", "Blood pressure", "Body fat", "Cholesterol", "Weight", and "Height".

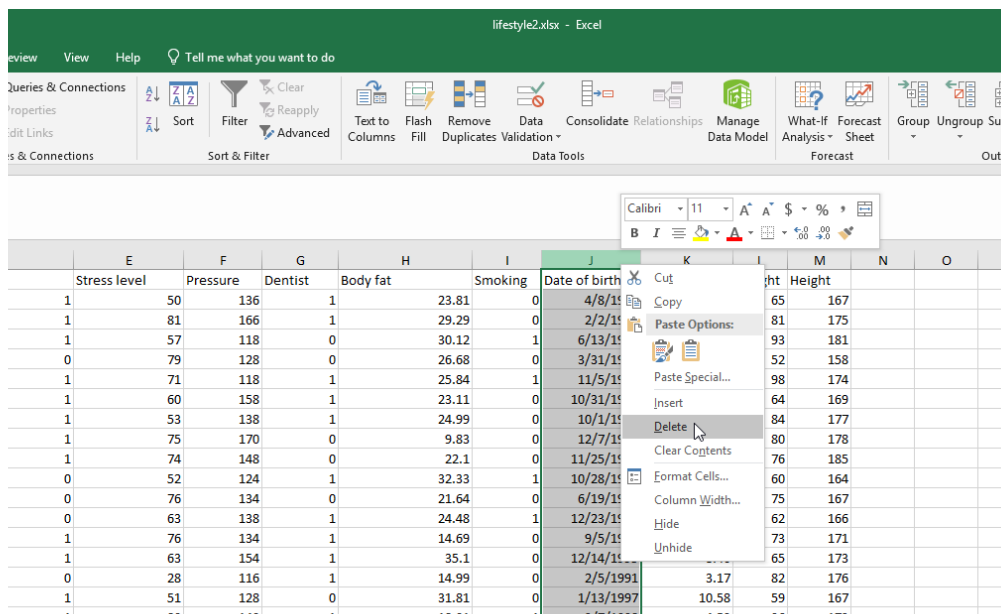


Fig. 55 Deleting columns

Calculate descriptive statistics for the following variables: "Stress level", "Blood pressure", "Body fat", "Cholesterol", "Weight", and "Height".

In the *Data* tab, in the *Analysis* section, click *Data Analysis* [Fig. 52]. In the window that opens [Fig. 56], select *Descriptive Statistics* and click *OK*.

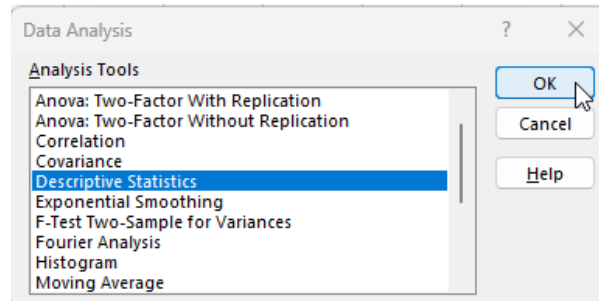


Fig. 56 *Data analysis* window

The Descriptive Statistics window [Fig. 57] contains a number of options for the analyzed data (Input) and for output results (Output Options).

The Input Range edit field should contain the range of cells for which calculations will be performed. Data arrangement is selected using one of the Grouping by Columns or Rows options. Data are most often collected in columns, so usually the Columns option is selected. The first row or column may contain a header, as opposed to the first value. This should be selected in the Labels in First Row checkbox (when the Rows option is selected, this field is called Labels in First Column).

The first three Output Options indicate the location of the results. This may be a highlighted location in the current workbook (Output Range), a New Sheet, or a different file (New Workbook). Note: if the Output Range overlaps with existing data, it will be overwritten. This is problematic because this operation cannot be undone, so it is safest to use the New Sheet option. To calculate descriptive statistics, you should select the Summary Statistics option. You may also be interested in determining the highest or lowest values present in the analyzed data, or determining the confidence level for the mean.

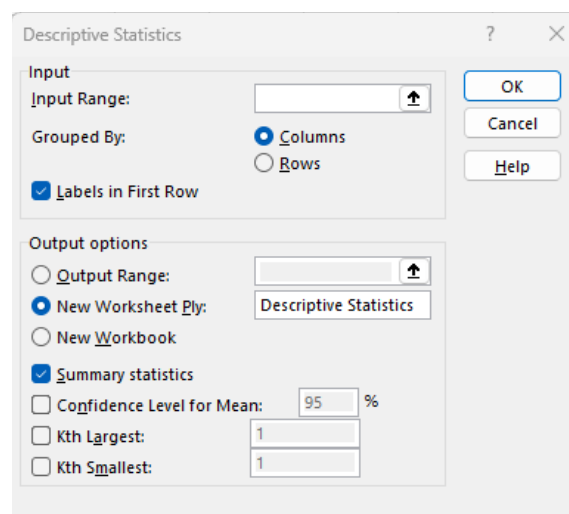


Fig. 57 *Descriptive Statistics* window

We will calculate descriptive statistics for all columns. Go to the *Input Range* edit field. Select all columns, i.e., the **A1:F201** range. We have selected the data, including the headers, so the application will automatically load the names of the analyzed features. This is done by

selecting the *Labels in First Row* option, which should be enabled. Due to the size of the data table, it is best to display the results in a new spreadsheet. To do this, select the *New Worksheet Ply* checkbox and name it "Descriptive Statistics" in the adjacent window. We expect summary statistics as the result, so select the *Summary Statistics* option. All these options are presented in [Fig. 58]. Confirm by clicking *OK*.

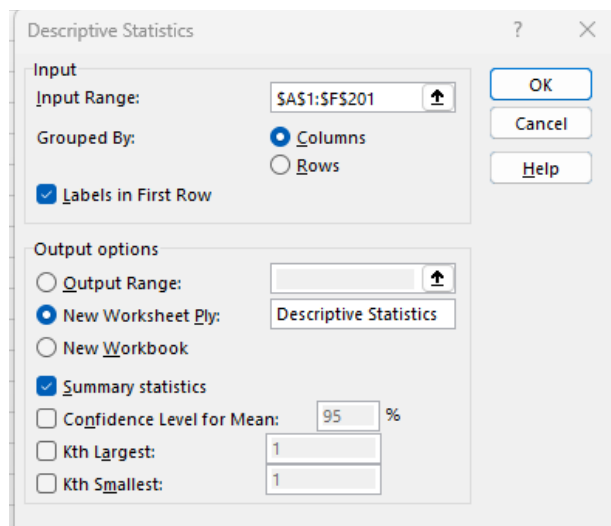


Fig. 58 *Statistics* settings

Format the resulting sheet.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Stress level		Pressure		Body fat		Cholesterol		Weight		Height	
2												
3	Mean	65.855	Mean	136.065	Mean	24.2744	Mean	5.05005	Mean	72.775	Mean	171.445
4	Standard Error	1.120893	Standard Error	1.243315	Standard Error	0.502746	Standard Error	0.13862	Standard Error	0.842509	Standard Error	0.548351
5	Median	69	Median	133	Median	24.48	Median	4.795	Median	72	Median	170
6	Mode	79	Mode	134	Mode	14.69	Mode	8.07	Mode	75	Mode	169
7	Standard Deviation	15.85183	Standard Deviation	17.58313	Standard Deviation	7.109905	Standard Deviation	1.960388	Standard Deviation	11.91487	Standard Deviation	7.754848
8	Sample Variance	251.2804	Sample Variance	309.1666	Sample Variance	50.55075	Sample Variance	3.843122	Sample Variance	141.9642	Sample Variance	60.13766
9	Kurtosis	-0.00484	Kurtosis	2.478788	Kurtosis	-0.81963	Kurtosis	0.145128	Kurtosis	-0.26745	Kurtosis	-0.05979
10	Skewness	-0.54215	Skewness	1.368412	Skewness	-0.00316	Skewness	0.654206	Skewness	0.293719	Skewness	0.492994
11	Range	80	Range	100	Range	33.08	Range	10.19	Range	58	Range	41
12	Minimum	16	Minimum	108	Minimum	7.97	Minimum	0.98	Minimum	49	Minimum	153
13	Maximum	96	Maximum	208	Maximum	41.05	Maximum	11.17	Maximum	107	Maximum	194
14	Sum	13171	Sum	27213	Sum	4854.88	Sum	1010.01	Sum	14555	Sum	34289
15	Count	200	Count	200	Count	200	Count	200	Count	200	Count	200

Fig. 59 The calculated statistics

[Fig. 59] shows the result of the performed steps. To make it more readable:

- delete the second (empty) row (deleting a row is performed similarly to deleting a column, i.e., right-click on its label (2) and select *Delete*),
- move the contents of the cells containing the variable names to the adjacent columns containing the values of the individual variables (e.g., "Stress level" from A1 to B1),
- remove duplicate names in every other column (i.e., remove columns K, I, G, E, and C, but not A),
- expand the columns.

We have obtained the table shown in [Fig. 60]. Note the range of "Height" values.

	A	B	C	D	E	F	G
1		<i>Stress level</i>	<i>Pressure</i>	<i>Body fat</i>	<i>Cholesterol</i>	<i>Weight</i>	<i>Height</i>
2							
3	Mean	65.855	136.065	24.2744	5.05005	72.775	171.445
4	Standard Error	1.120893342	1.243315342	0.502746215	0.138620383	0.842508742	0.548350542
5	Median	69	133	24.48	4.795	72	170
6	Mode	79	134	14.69	8.07	75	169
7	Standard Deviation	15.85182566	17.58313419	7.109905161	1.960388254	11.91487289	7.75484773
8	Sample Variance	251.2803769	309.166608	50.5507514	3.843122108	141.964196	60.13766332
9	Kurtosis	-0.004836922	2.478788018	-0.81962862	0.145127775	-0.267445998	-0.059794018
10	Skewness	-0.542146403	1.368412079	-0.003156784	0.654205579	0.293718513	0.492993557
11	Range	80	100	33.08	10.19	58	41
12	Minimum	16	108	7.97	0.98	49	153
13	Maximum	96	208	41.05	11.17	107	194
14	Sum	13171	27213	4854.88	1010.01	14555	34289
15	Count	200	200	200	200	200	200

Fig. 60 The formatted results

Exercise 9

In this exercise, we will present the distribution of values in a selected column graphically.

A histogram is an approximate representation of data distribution. To construct a histogram, you need to define intervals that divide the data across their entire range. Within each interval, you count the values contained within it. Intervals are usually defined so that they cover the full range of a given variable consecutively and without overlapping. For this reason, the next interval should begin at the end of the previous one. They are also often required to be of equal width. A corresponding bar chart displays the calculated frequencies of occurrence in each interval.

You can plot the so-called cumulative density function on a histogram. It shows the proportion of observations with values lower than the current position on the ordinate axis.

Select the appropriate intervals for the "Height" column.

Move to the "Data Analysis" spreadsheet. In the previous exercise, we noted that the height of the study group varied between 153 and 194 cm. Let us assume that we need 10 cm-wide intervals. For simplicity, the range ends should be multiples of 10 cm. We will define the following height ranges:

- up to and including 160 cm,
- from 160 cm to and including 170 cm,
- from 170 cm to and including 180 cm,
- from 180 cm to and including 190 cm,
- above 190 cm.

This gives a total of 5 ranges separated by four values constituting the upper limits of each of the ranges, i.e., 160, 170, 180, and 190 cm. Enter these values, along with the appropriate heading:

H1="Ranges"

H2=160

H3=170

H4=180

H5=190

Create a histogram for the "Height" column.

In the *Data* tab, in the *Analysis* section, click *Data Analysis*. In the window that opens, select *Histogram* and click *OK*.

The Histogram window [Fig. 61] contains a number of options for the analyzed data (Input) and the output results (Output Options).

The Input Range: edit field should contain the range of cells for which the histogram will be generated. The next edit field, Bin Range: should contain the range of cells defining the appropriate intervals. The first rows of these ranges should contain headers. This should be selected in the Labels checkbox.

Note: it is not possible to specify columns where only one column has a header. In this case, either the header would be treated as a value (often resulting in the "Input data contains

non-numeric values" error), or the first value would be the header. The first three Output Options indicate the location of the results and function similarly to the Descriptive Statistics window. Selecting the Output Chart checkbox results in a graphical representation of the histogram as a bar chart. An additional option, Cumulative Percentage, plots the cumulative density function on the chart, along with an additional vertical axis.

Fig. 61 The *Histogram* window

Go to the *Cell Range* edit field. Select the "Height" data column, i.e., the **F1:F201** range. In the *Set Range:* checkbox, enter the address of the area that contains the entered limit values for the subsequent ranges (from 160 to 190), also with a header, i.e., the **H1:H5** range.

As before, select the *Title* option to automatically load the "Height" characteristics name and the column name with the specified ranges. Again, select the location of results as *New Wroksheet Ply:*, which we name "Histogram" in the adjacent checkbox. Select the *Chart Output* and *Cumulative Percentage* options to display a graphical interpretation of the histogram. All these options are presented in [Fig. 62]. Confirm by clicking the *OK* button.

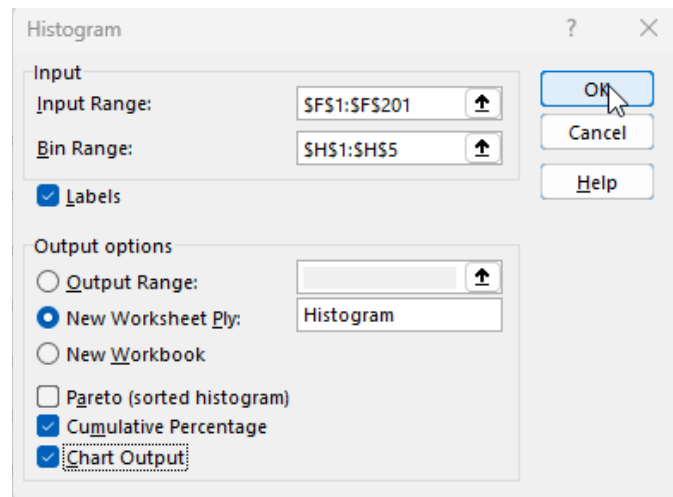


Fig. 62 The histogram options

We have obtained a spreadsheet that contains a table and a chart. To improve their readability, the appropriate columns can be widened, as well as the chart itself [Fig. 63].

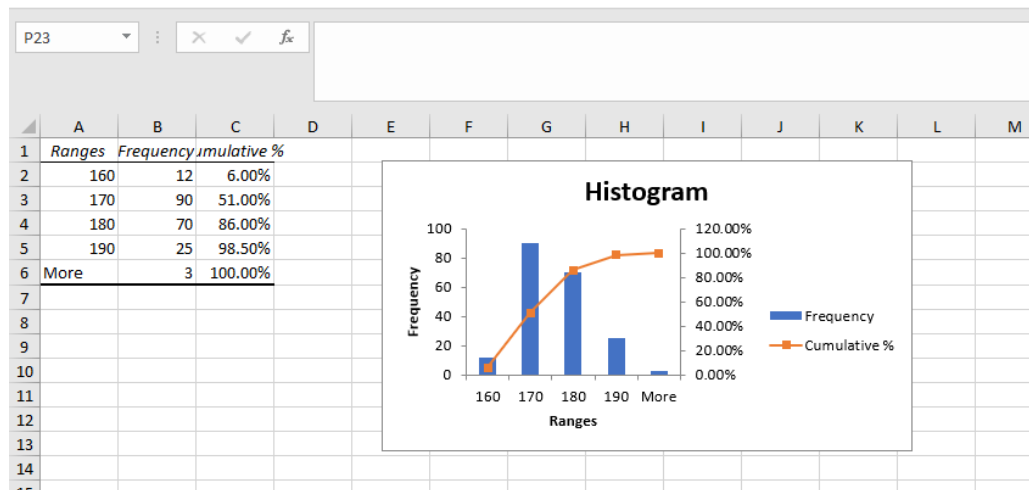


Fig. 63 *Histogram*: the final result

We can see (in the table or by hovering the cursor over the appropriate area [Fig. 64]), for example, that 70 people from the analyzed group have a height ranging from 170 cm to 180 cm inclusive. People up to 180 cm tall constitute 86% of the entire group.

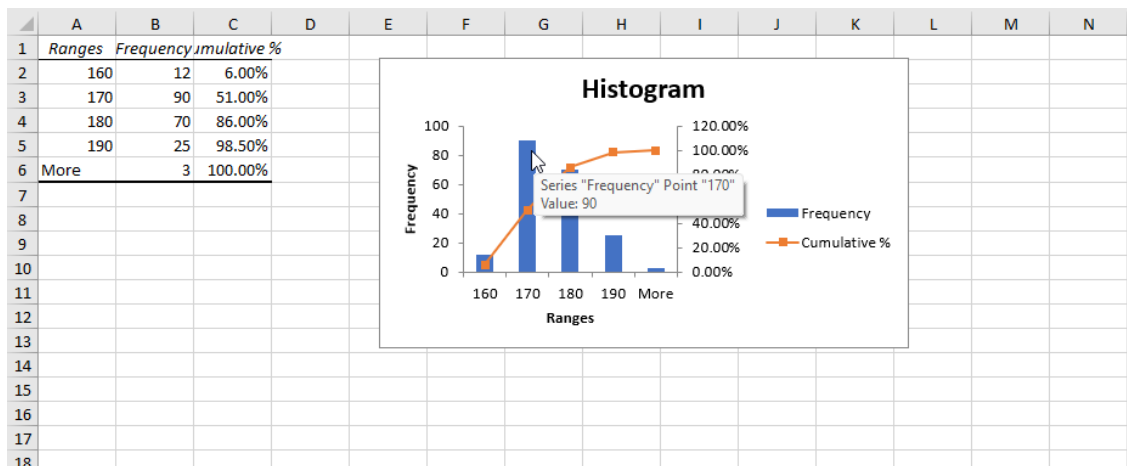


Fig. 64 Checking the value of a selected histogram area

Exercise 10

In this exercise, we will calculate the correlation coefficient between selected variables.

The term "correlate" is often used colloquially and can be translated as "to be related" or "to be interdependent with something". Statistics has various measures of correlation (or coefficients), one of which is the so-called correlation coefficient, which ranges from -1 to +1. Correlation coefficient values close to the extreme values indicate a strong relationship between the pair of variables for which the coefficient is calculated. A positive value indicates that as one variable increases or decreases, the other follows a similar trend. A negative value indicates the opposite.

Calculate the correlation coefficient between the selected variables.

Go to the "Data Analysis" spreadsheet. In the *Data* tab, in the *Analysis* section, click the *Data Analysis* command. In the window that opens, select the *Correlation* command and confirm with the *OK* button.

The Correlation window [Fig. 65] contains a number of options for the analyzed data (Input) and the output results (Output Options). Their use is the same as in the Descriptive Statistics window.

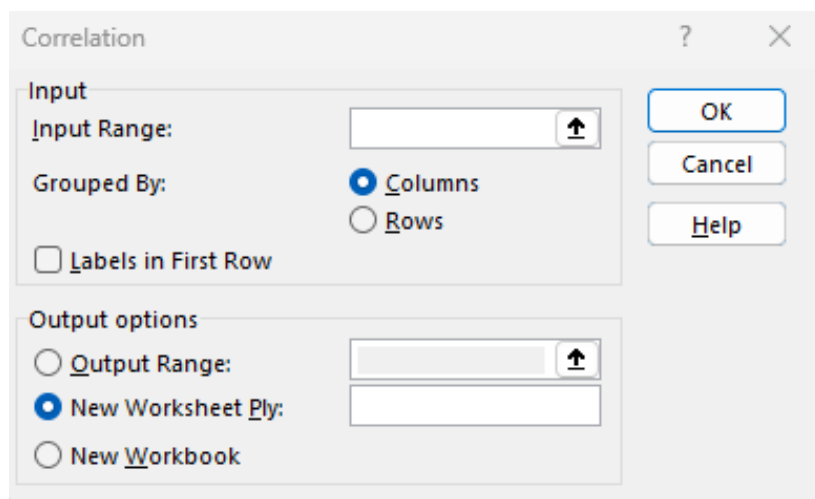


Fig. 65 Correlation window

We will calculate the correlation coefficients for all the six variables from the "Data Analysis" spreadsheet. Go to the *Input Range* edit field and select all columns, i.e., the **A1:F201** range. We have selected the data, including the headers, so the application will automatically load the names of the analyzed variables. This is done by selecting the *Labels in First Row* option, which should be enabled. Due to the size of the data table, it is best to display the results in a new spreadsheet. To do it, select the *New Worksheet Ply* checkbox. Name it "Correlations" in the adjacent window. All these options are presented in [Fig. 66]. Confirm by clicking the *OK* button.

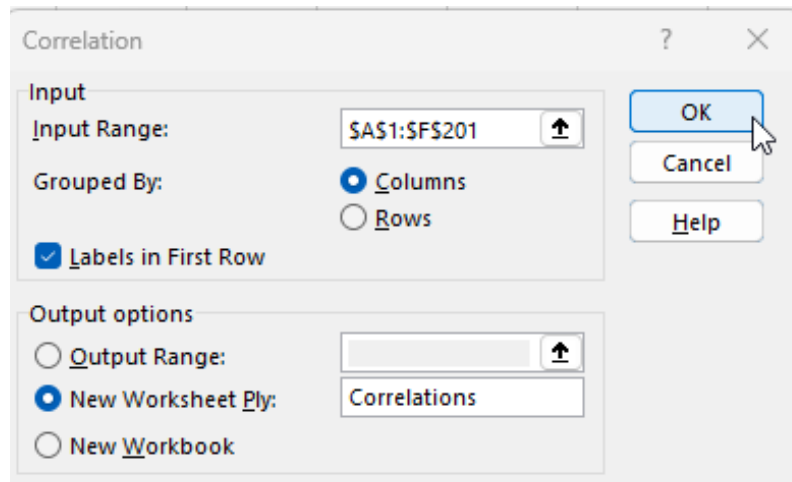


Fig. 66 Correlation options

Format the resulting sheet.

	A	B	C	D	E	F	G
1		<i>Stress level</i>	<i>Pressure</i>	<i>Body fat</i>	<i>Cholestero</i>	<i>Weight</i>	<i>Height</i>
2	Stress level	1					
3	Pressure	0.206502	1				
4	Body fat	-0.22462	-0.04996	1			
5	Cholesterol	-0.09658	0.034043	0.248884	1		
6	Weight	0.070784	-0.00907	0.022233	0.090849	1	
7	Height	0.077175	-0.04311	0.089297	0.116929	0.798763	1

Fig. 67 The calculated correlation coefficients

[Fig. 67] shows the result of the executed commands. To make it more readable, move the columns further apart. To better visualize the data, we will apply conditional formatting to the **B2:G7** area. Values greater than 0.5 are conditionally formatted (*Home* tab, *Styles* section, *Conditional Formatting* command, *Highlight Cell Rules* group, *Greater than...* option) using the *Light Red Fill with Dark Red Text* option [Fig. 68]. Of course, each variable is perfectly correlated with itself, but the correlation between the "Height" and "Weight" pair of variables is much more interesting.

	A	B	C	D	E	F	G
1		<i>Stress level</i>	<i>Pressure</i>	<i>Body fat</i>	<i>Cholesterol</i>	<i>Weight</i>	<i>Height</i>
2	Stress level	1					
3	Pressure	0.206501953	1				
4	Body fat	-0.224619991	-0.049959553	1			
5	Cholesterol	-0.09658016	0.034043281	0.248884006	1		
6	Weight	0.070784376	-0.009068578	0.022232636	0.090848601	1	
7	Height	0.07717462	-0.043110534	0.089297171	0.116929083	0.798763	1

Fig. 68 The formatted correlation matrix

Exercise 11

In this exercise, we will create a scatterplot and add a regression line to it.

A scatterplot displays points whose coordinates correspond to the values of two variables (located in two columns).

Create a scatterplot for the variables "Height" and "Weight".

Go to the "Data Analysis" worksheet. In the previous exercise, we identified a correlation between height and weight, so now will present this relationship graphically.

When selecting the data to be charted, MS Excel defaults to the first selected column as the one containing values to be placed on the X-axis, and the second one as containing the values to be placed on the Y-axis. If you need to plot the data in reverse, you have to start by inserting a blank chart and then add the appropriate data.

We will place weight on the Y-axis and height on the X-axis. Click any empty cell in the worksheet. On the *Insert* tab, in the *Charts* section, in the *Scatterplots* group, select the first chart [Fig. 69].

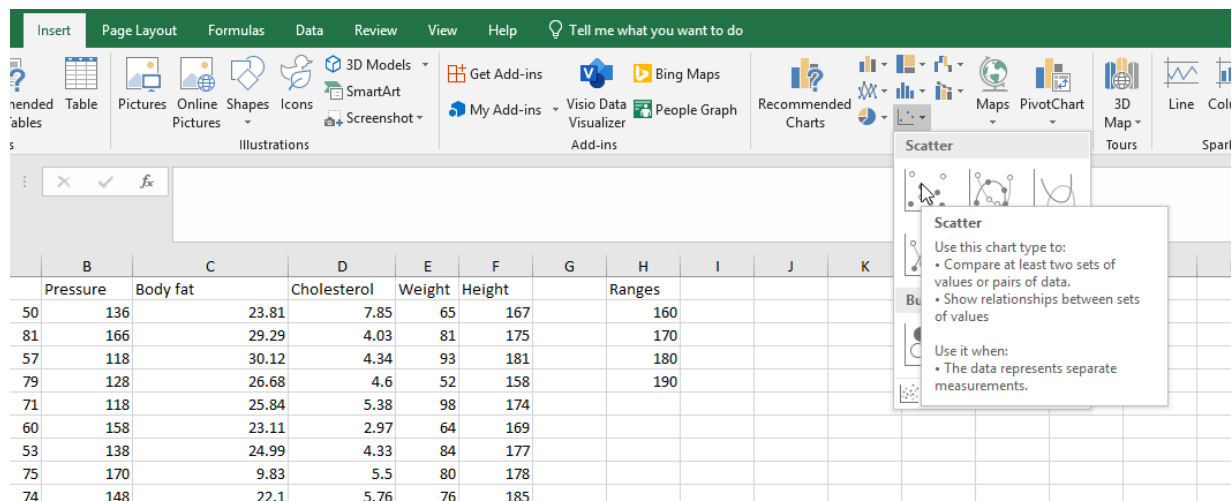


Fig. 69 The scatter plot

This action creates a blank chart area in the spreadsheet. Right-click on it and choose *Select Data...* from the drop-down menu [Fig. 70].

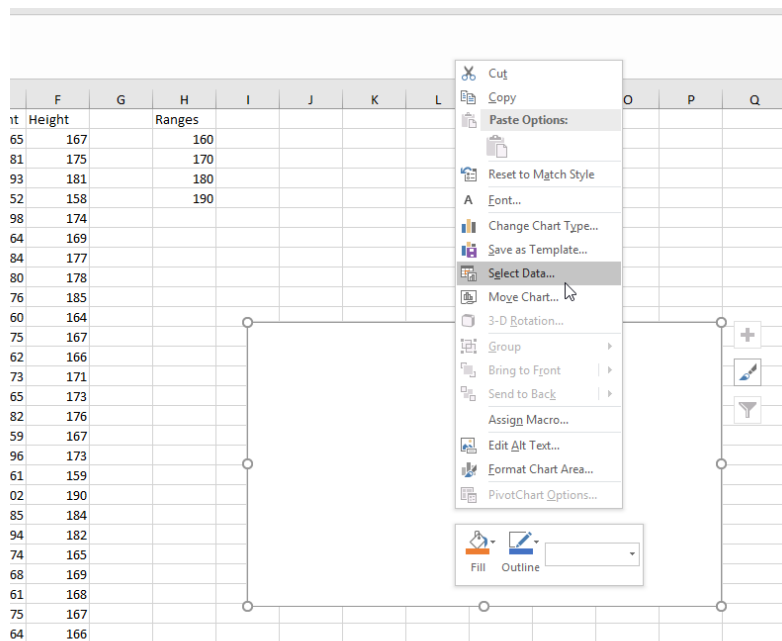


Fig. 70 *Select Data...* command

In the *Select Data Source* window, in the *Legend Entries (Series)* section, click the *Add* button [Fig. 71].

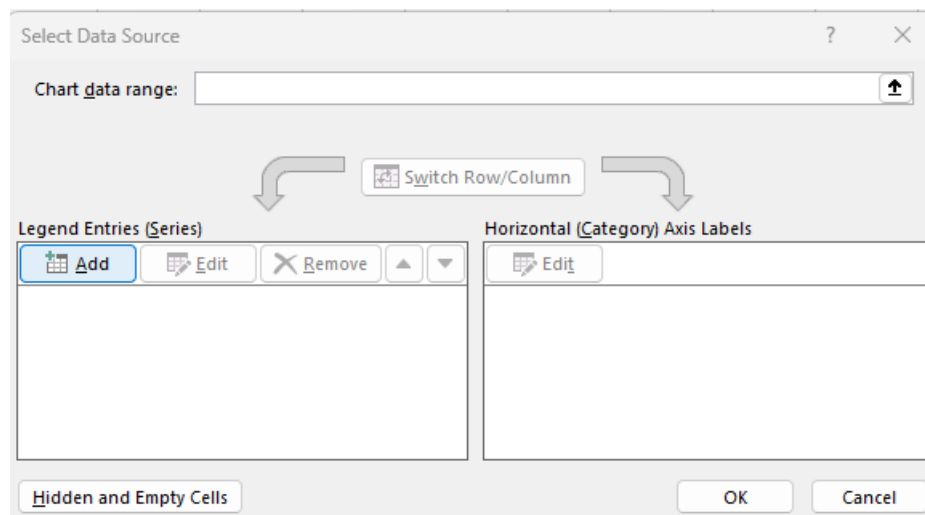


Fig. 71 Selecting a data source

In the *Edit Series* window, in the *Series X values* field, select the values for the "Height" column (without the header). In the *Series Y Values* field, the default value $=\{1\}$ is inserted, which we must remove manually. After having removed it, select the values for the "Weight" column (without the header) in its place [Fig. 72]. Click the *OK* button and confirm the data selection in the *Select Data Source* window [Fig. 73].

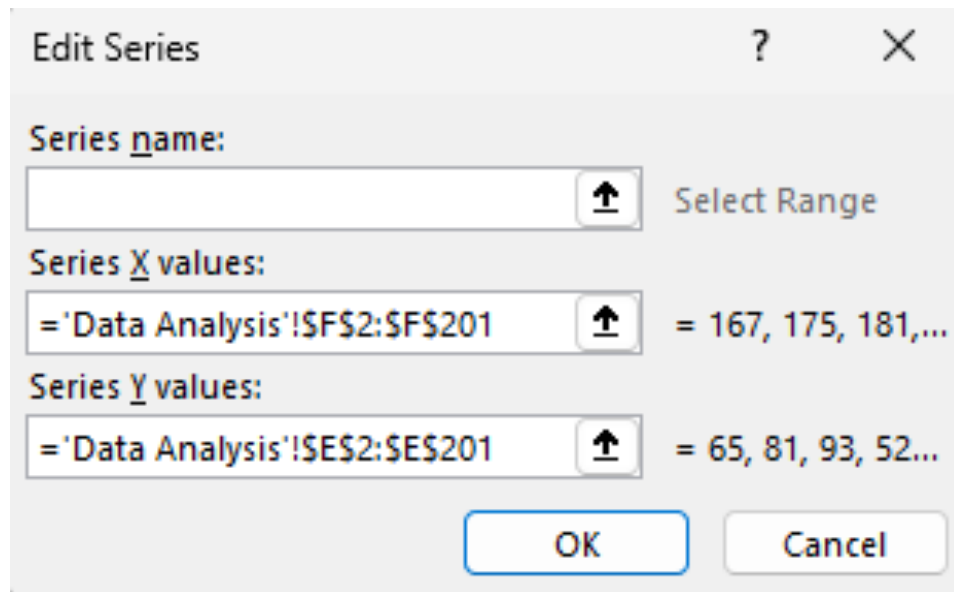


Fig. 72 Editing a series

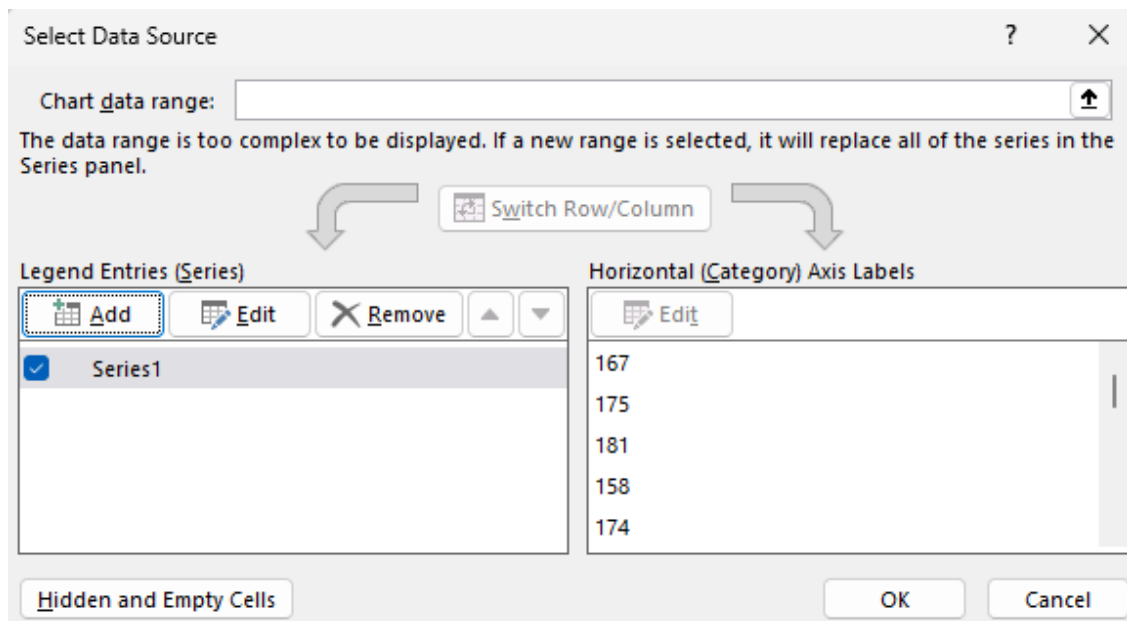


Fig. 73 Data confirmation

As a result, we obtain a scatter plot showing the correlation between weight and height [Fig. 74].

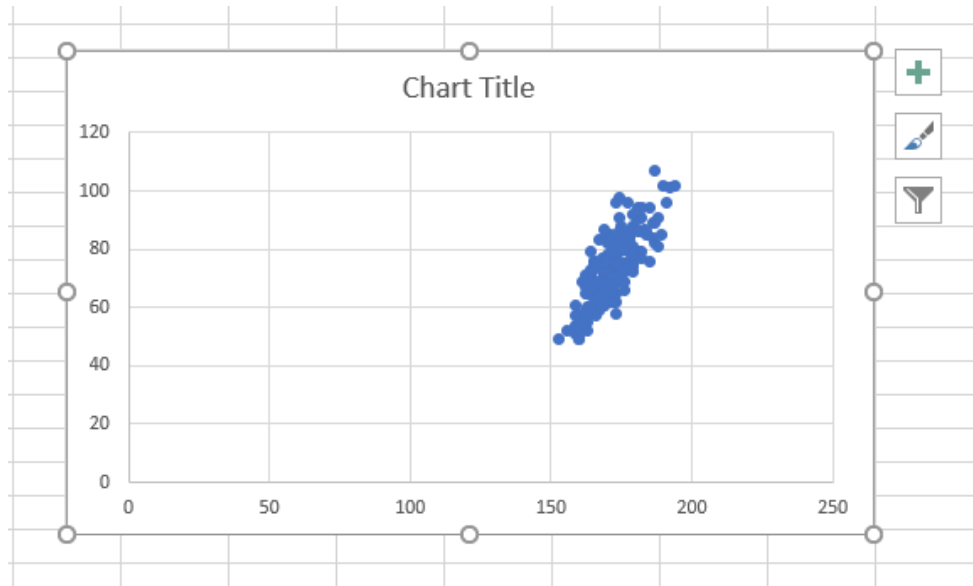


Fig. 74 The obtained scatter plot

The resulting graph is unreadable. To improve its readability, we need to change the starting and ending values on both axes. First, we will do this for height. Double-click on the X-axis values. The axis formatting options will open on the right. Change the minimum value from 0 to 150 and the maximum value to 200 (all the height values for the analyzed data are between 150 cm and 200 cm). Change the value in the *Major* field to 5 (the values on the axis are labeled every 5 units) [Fig. 75].

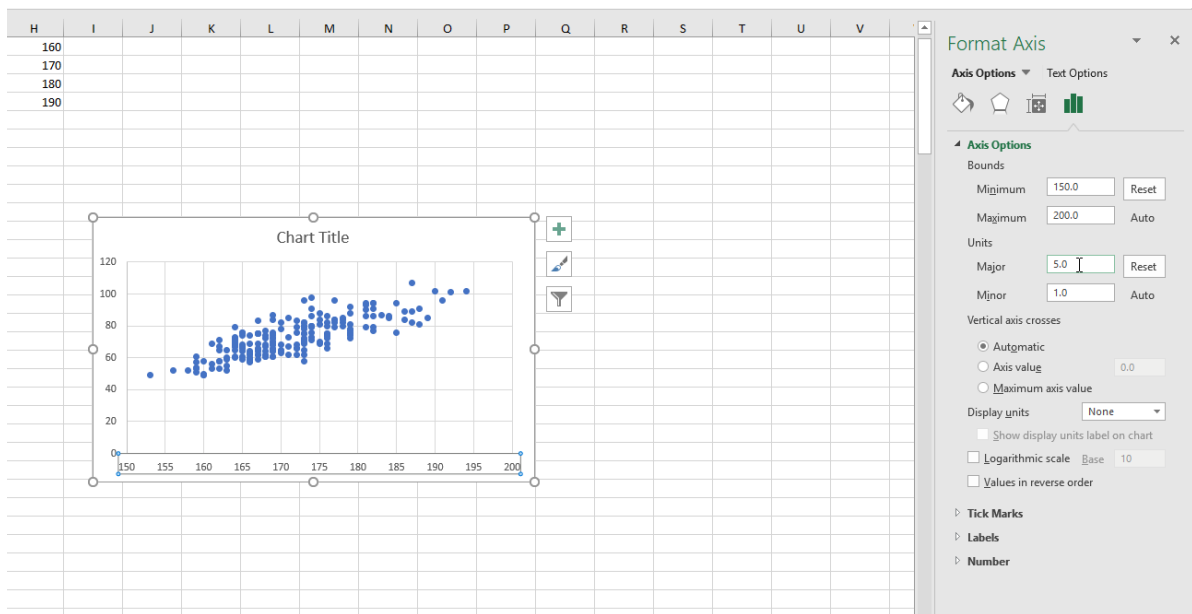


Fig. 75 X-axis formatting

Change the values on the Y axis in a similar manner. Change the minimum value from 0 to 40, the maximum value to 110 (weight varies between 40 kg and 110 kg for all persons), and the value in the *Major* field to 5. The final result is shown in [Fig. 76]. Close the axis formatting options.

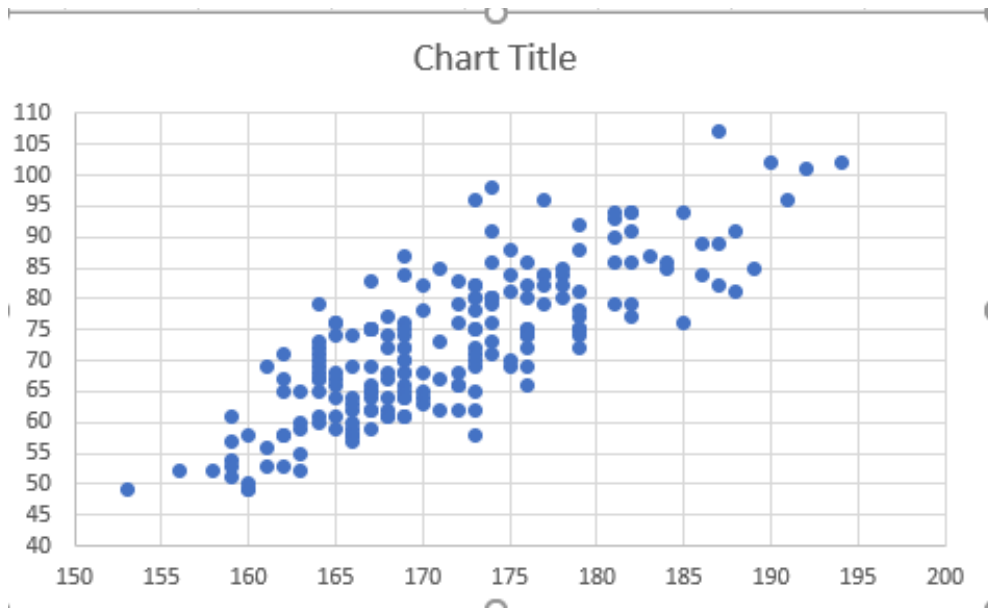


Fig. 76 The formatted axes

Change the chart title to "Scatterplot". Click the green plus symbol in the upper right corner of the chart and select the *Axis Titles* checkbox [Fig. 77].

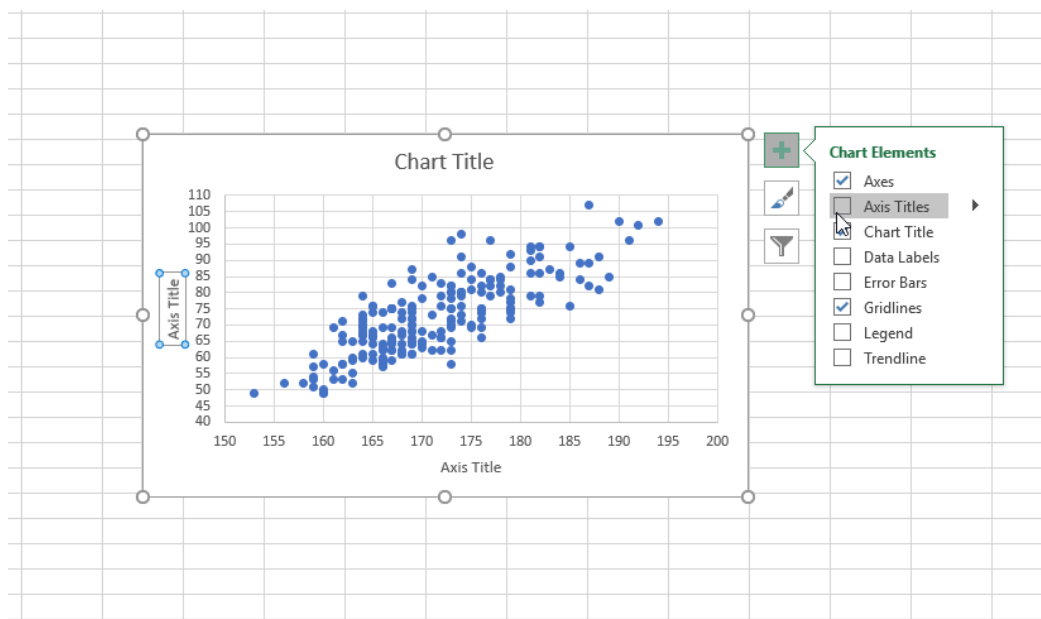


Fig. 77 Adding axis titles

In a similar manner to the chart title, change the title of the vertical axis to "Weight" and the horizontal axis to "Height". The final result is shown in [Fig. 78].

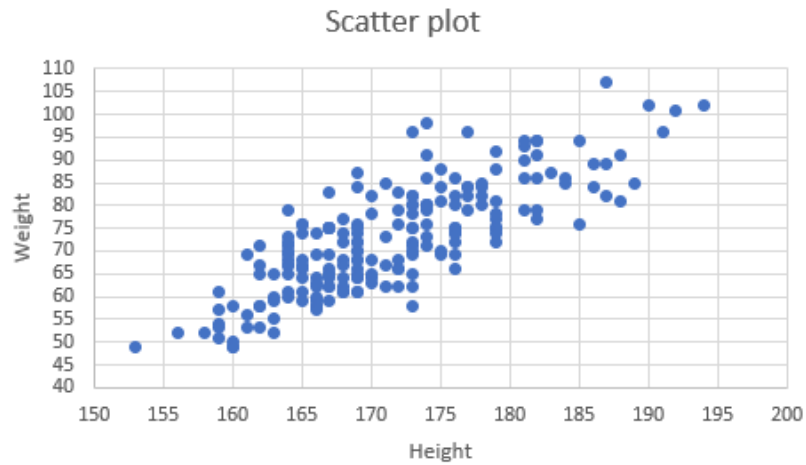


Fig. 78 Changed axis titles

Add a trend line.

Click the green plus symbol in the upper right corner of the chart and select the *Trendline* checkbox [Fig. 79].

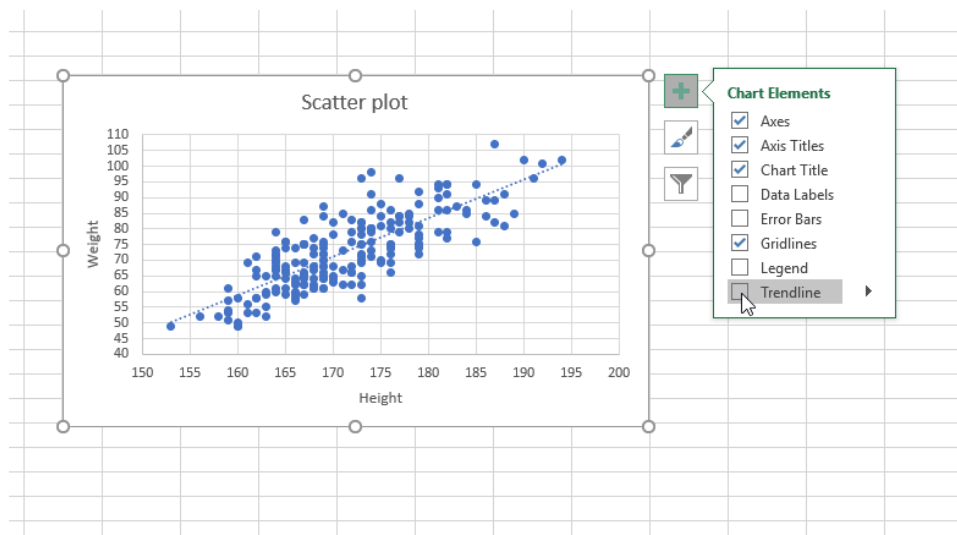


Fig. 79 Chart with *trendline* added

Add its equation and the R-squared value to the graph with the trend line.

Double-click the trend line added to the chart. This opens the *Trendline Formatting* dashboard. Select the following checkboxes: *Display Equation on chart* and *Display R-squared values on chart* [Fig. 80].

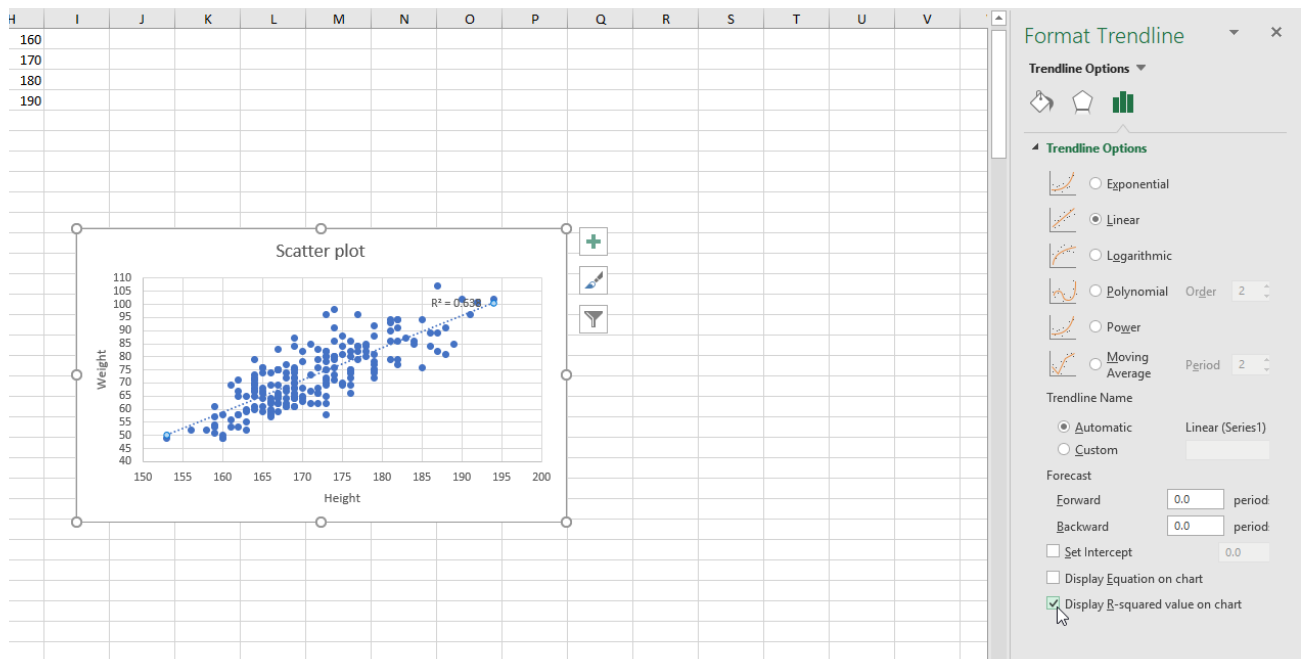


Fig. 80 Graph with the equation and R-squared value added on the trend line

Change the trend line color to red.

Finally, to improve the visibility of the trend line, we will change its color to red. After completing the previous exercise, we are now in the *Trendline Formatting* dashboard. Go to the first tab, i.e., *Fill & Line* [Fig. 81]. Change the *Color* to *Red*, the *Width* to "2 pts", and the *Dash Type* to *Solid* [Fig. 82].

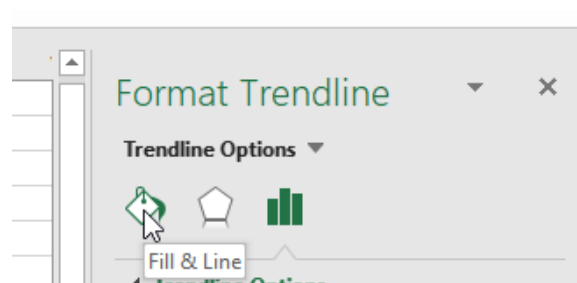


Fig. 81 *Fill and Line* tab

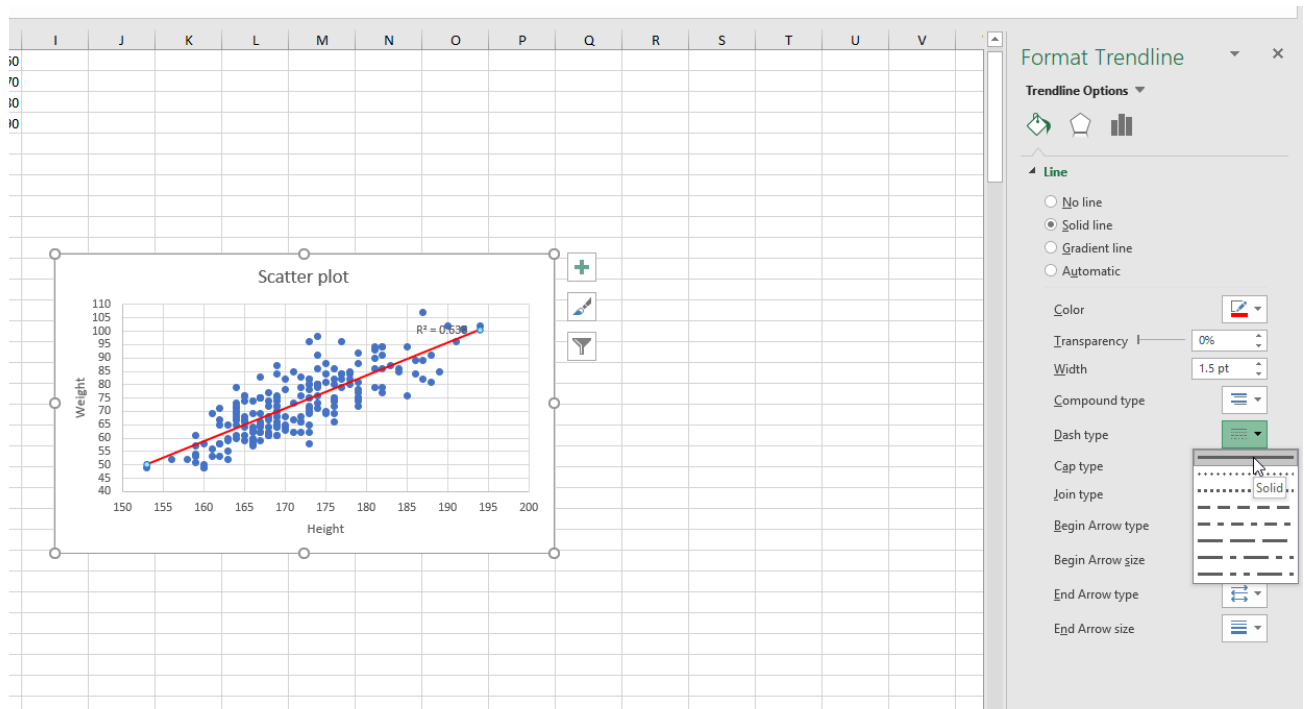


Fig. 82 The weight vs. height scatterplot after formatting

Exercise 12

In this exercise, we will perform a basic regression analysis.

In the previous exercise, we created a linear trend graph adapted to a selected pair of variables. The trend line was calculated using the appropriate statistical methodology. We can see its equation on the graph, but we would also like to see the corresponding coefficients as values in the cells. We will use a dedicated command in the Data Analysis add-in for this purpose.

Perform a regression analysis.

Remain in the "Data Analysis" spreadsheet. In the *Data* tab, in the *Analysis* section, click the *Data Analysis* command. In the window that opens, select the *Regression* command and click *OK*.

The Regression window [Fig. 83] contains a number of options for the analyzed data (Input) and the output results (Output Options). Regression is an advanced statistical method that describes the relationship between one variable and another. Due to the complexity of this topic, only selected options are described below.

The regression equation is expressed as:

$$Y = aX + b$$

The Input Range Y: field contains the dependent variable range, while the Input X Range X: field contains the independent variable range. Select the Labels checkbox if you have selected variable names in the variable ranges, you also need to check the Titles checkbox. The three output options have the same use as those in the Correlation or Descriptive Statistics windows. The remaining options are not of interest to us.

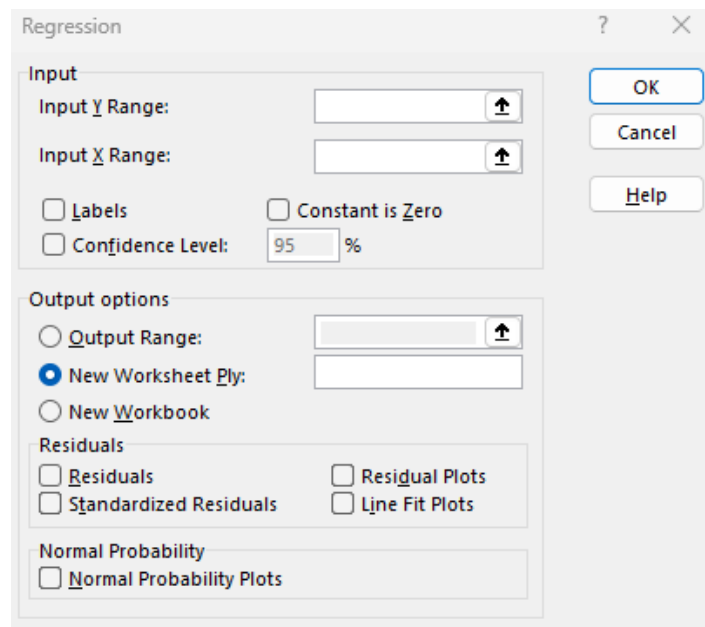


Fig. 83 Regression window

We want to explain a person's weight using their height. Go to the *Input Range Y:* edit field and select the "Weight" column, i.e., the **E1:E201** range. In the *Input Range X:* edit field, select the "Height" column, i.e., the **F1:F201** range. Select the *Labels* option. From the output

options, select *New Worksheet Ply*, which we name "Regression" in the adjacent window [Fig. 84]. Confirm by clicking *OK*. To increase the readability of the results, expand the columns.

Fig. 84 *Regression* options

We can now compare the calculated regression coefficients [Fig. 85] with the values on the scatterplot. A keen eye will also notice the R-squared value (cell **B5**).

1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.798763239							
5	R Square	0.638022712							
6	Adjusted R Square	0.636194544							
7	Standard Error	7.186609008							
8	Observations	200							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	1	18024.69989	18024.69989	348.995645	1.44168E-45			
13	Residual	198	10226.17511	51.64734904					
14	Total	199	28250.875						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	-137.6314419	11.27432713	-12.20750829	6.42875E-26	-159.8646121	-115.3982718	-159.8646121	-115.3982718
18	Height	1.227253299	0.065693773	18.68142513	1.44168E-45	1.097704031	1.356802568	1.097704031	1.356802568
19									

Fig. 85 The regression analysis result