DATA
PROCESSING
GUIDE

AUTHORS: AVRAM CALIN, GLIGOR ADRIAN, NYULAS VICTORIA and DUMITRU ROMAN

This guide was created within the project entitled: **"Innovative methods needed by teaching and research staff for AI data analysis and processing"**, which took place between February 2022 and February 2023, with the University of Medicine, Pharmacy, Science and Technology as partners "G.E.Palade" Targu Mures and SINTEF - Norway.

CONTENT

INTRODUCTION

In order to conduct valid and valid research, an essential critical element is the accurate and efficient collection and processing of data for analysis. Most medical researchers have a low level of training in how to manage and process collected medical data, which can lead to an increase in the time spent cleaning data, as well as the risk of collection or recording errors.

Students, PhD students, young researchers, all are or can be involved in data collection, planning, execution, and research analysis. Often, due to lack of funding, experience, or both, this data is entered into the computer using an ad hoc process that results in incorrectly coded data, incorrectly formatted values, incomplete information, and typographical errors. It is often a common situation for researchers to find that the data collected must be extensively "cleaned" before it can be properly analyzed.

Consequently, we have produced this guide which provides practical, concise and illustrative guidance on how to process data to create a well-designed, homogeneous and verified data set. Since Microsoft Excel is the software most used by students, researchers for collecting and organizing data, the guide presents specific methods that can be implemented in the Microsoft Excel spreadsheet program.

This guide is a comprehensive tutorial that provides a good insight into the latest and advanced features available in Microsoft Excel. Within the guide we have included a lot of screenshots that explain how to use a particular feature, step by step.

1. DESCRIPTION OF DATA ELEMENTS. TYPES OF DATA

Accurate data collection starts with planning. Before collecting any data, an investigator should define the research questions and determine what measurements are needed to answer them. Typically, a research data set includes at least one outcome variable (dependent variable) and one or more predictor variables (independent). Other demographic, covariate or control measures may also be recorded. It is important that each subject investigated in a data set is assigned a unique identification number. For each study, the researcher should use this information to check that all variables needed to perform an analysis are included. It is discouraging to realize too late that a variable needed to complete an analysis was not collected.

Defining Big Data

In the field-specific literature, Big Data is characterized as data that includes:

- High volume - both in terms of data elements and dimensionality.
- High speed - reaches a very high rate, usually with an assumption of low latency between the arrival of the data and the value obtained.
- High variety - includes the ability to evolve over time for the form and meaning of data.

And which requires:

- Cost-effective processing
- Innovative types of analysis
- New business value

Table 1. Data types – data categorization

Data	Description	Example
Nominal or categorical data	Data that can be placed into mutually exclusive categories	Name of a country, Gender, Residence, Presence of an illness, Marital status
Ordinal data	Data is ordered or classified according to certain characteristics	Likert scale, Educational level, Blood type, Patient's level of statistics, Socio-economic status, Disease stage
Range data	Data characterized and ordered by some distance between each observation, but without a natural zero	IQ Test, Temperature in degrees Fahrenheit, Celsius,
Report type data	Data that has a natural zero	Age, height, weight, basal blood glucose value, ventricular rate

2. DATA ANALYSIS OVERVIEW

Data analysis is a process of inspecting, cleaning, transforming, and modeling data with the goal of uncovering needed information, suggesting conclusions, and supporting decision-making.

Statistician John Tukey in 1961 defines the process of data analysis as follows "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the collection of data to make their analysis easier, more precise, or more accurate , as well as all cars. and the results of the (mathematical) statistics that are applied to the analysis of the data." Accordingly, data analysis is a process for obtaining large, unstructured data from various sources and transforming it into useful information for:

- Answering research questions
- Hypothesis testing
- Making decisions
- Denial of theories

Types of data analysis

There are several data analysis techniques spanning various fields such as business, science, social science, etc. with a variety of names. The main approaches to data analysis and exploitation are:

- Data Mining
- Business intelligence
- Statistical analysis of data
- Predictive data analysis
- Text analysis

Data Mining is the analysis of large amounts of data to extract interesting, previously unknown data patterns, unusual data, and dependencies. Note that the goal is to extract patterns and knowledge from large amounts of data, not to mine the data itself.

Data Mining analysis involves computer methods at the intersection of artificial intelligence, machine learning, statistics and database systems.

Business Intelligence

Business Intelligence techniques and tools are used to acquire and transform large amounts of unstructured business data for use in identifying, developing and creating new strategic business opportunities. The goal of Business Intelligence is to enable easy interpretation of large volumes of data to identify new opportunities.

Statistical data analysis

Statistics deals with the study of collecting, analyzing, interpreting, presenting and organizing data. In the statistical analysis of the data, two main statistical methodologies are used:

1. Descriptive statistics: data from the entire population or a sample are summarized with numerical descriptive indicators.
2. Inferential statistics: uses patterns from sample data to draw inferences about the represented population. These inferences can be: hypothesis testing, estimation, correlation and regression analyses.

Predictive data analysis

Predictive data analysis uses statistical models to analyze current and past data for forecasting (predictions) about future or unknown events, which can help in decision making.

Text analysis

Text analysis is the process of extracting high-quality information from text. Text mining typically involves the process of structuring the input text, deriving patterns within the structured data using means such as statistical model learning, and finally evaluating and interpreting the results.

3. DATA ANALYSIS PROCESS

The data analysis process consists of the following stages which are iterative in nature:

- Data requirements specifications
- Data collection
- Data processing
- Data cleaning
- Data analysis
- Data communication

Data Requirements Specifications

The data required for analysis is based on a research question or experiment. Based on the requirements that drive the analysis, the data required as inputs to the analysis are identified (eg, Hypertension Patient Population). Specific characteristics relating to a population (eg age, gender and income) can be specified and obtained. Data can be numeric or categorical.

Data collection

Data collection is the process of gathering information about target characteristics marked as data requirements. The focus is on ensuring accurate and correct data collection. Data collection ensures that the data collected is accurate so that related decisions are valid.

Data processing

The source of collected data is diverse, from organizational databases to information from web pages. Data obtained in this way may not have a coherent structure and may contain irrelevant information. The collected data must therefore be subject to data processing and cleaning.

Data cleaning

The processed and organized data may be incomplete, contain duplicate values, or contain errors. Data cleansing is the process of preventing and correcting these errors. Depending on the type of data, we can use several types of data cleaning.

Data analysis

Data that is processed, organized and cleaned is ready and suitable for the next stage, that of analysis. Based on the requirements, we can use different methods of data analysis to understand, interpret and draw conclusions.

Data communication

The results of data analysis must be reported in a format required by users to support their decisions and further actions. Feedback from users could lead to further analysis. Data visualization and data illustration techniques such as tables and charts can be used to communicate data, which help communicate the message clearly, concisely and effectively to users. Analysis tools provide ways to highlight needed information with color coding and formatting in tables and charts.

4. DATA ANALYSIS WITH EXCEL – OVERVIEW

Data analysis with Excel

Microsoft Excel provides many different ways to analyze and interpret data. The provenance of the data can be from various sources, and the ways of converting and formatting the data are also various. For data analysis, Excel provides the following techniques, function commands, and tools: conditional formatting, ranges, tables, text functions, date functions, time functions, financial functions, subtotals, quick analysis, audit formulas, investigation tools, IF analysis , solver, data models, PowerPivot, PowerView, PowerMap, etc.

As mentioned and listed, Excel contains and provides commands, functions, and tools that make data analysis tasks easier.

Ranges and Tables

Data can be presented in a range or in a table. Certain commands on the data can be performed regardless of how the data is presented, range or table. However, certain commands are more efficient if the data is presented as a table rather than a range, and Excel also provides functions exclusively for tables.

Data cleaning – Text, date and time functions

Before the data analysis process, the stage of data cleaning and structuring must be carried out, below are presented how many functions can be used to structure data:

- With text functions
- With functions for given values
- With functions for time values

Conditional formatting

To color specific cells, text font, or insert, based on predefined criteria, different symbols next to the values contained in the cells, Microsoft Excel provides commands for conditional formatting.

Sorting and filtering

In the preparation stage of the data analysis, or to expose various specific data, you can use the Data Filtering and Sorting functions provided by Excel.

Subtotals with ranges and PivotTables

To summarize the data in Excel, you can use the PivotTable tool, but for grouping/ungrouping the data or summarizing it, there is another tool provided by Excel, you can use Subtotals with ranges.

Data visualization

Excel provides numerous data visualization techniques (band charts, gantt, cascade chart, boxplot, scatterplot, etc.), all of which can be used from the Excel charts section of the Insert menu.

Data validation

Data validation can be used to restrict the type of data or values that users enter into cells. With data validation commands, one can easily set data validation values for a cell, an input message prompting the user what is expected to be entered in the cell, validating input values against defined criteria, and displaying an error message in case of incorrect entries.

Working with multiple worksheets

If it is necessary to perform identical calculations in several worksheets, with the help of Excel functions, this can be done in a few simple steps.

5. IMPORTING DATA INTO EXCEL

A large number of users use a connection to import external data into Excel as an updatable snapshot. The advantage is that it creates an independent document that can be used to work offline, but updated with new data when online. Because the data is contained in Excel, users can also transform it to account for their own context or analysis needs.

A few key challenges must be considered when importing Big Data into Excel:

- Querying Big Data — Data sources designed for Big Data, such as SaaS, HDFS, and large relational sources, can sometimes require specialized tools. In Excel we can turn to the Power Query tool, which provides several modern sets of connectors for Excel users, including connectors for relational, HDFS, SaaS (Dynamics CRM, SalesForce), etc.
- Data transformation— similar to other types of data, Big Data is not very often absolutely clean. Power Query provides the ability to create an orderly and reproducible set of steps for processing data.

In Excel, data can be imported from various sources such as: Microsoft Access databases, web pages, text files, SQL Server tables, SQL Server analysis cube, XML files, etc.

Importing external data

Data in compatible external files can be imported into Excel without the need to completely re-enter the data. Depending on the format of the data to be imported, different techniques can be used, including opening and saving in Excel, connecting to databases, importing data, copying and pasting it into Excel.

5.1 Importing data from a database

We will exemplify this action in a spreadsheet with no content. The purpose of this action is to import data into Microsoft Excel from a Microsoft Access database.

In a folder on the computer, we have saved a file containing an Access database, named Patient_Database.

1. In Excel 2016, we open a blank worksheet.

2. We click on DATA > Get External Data > From Access.
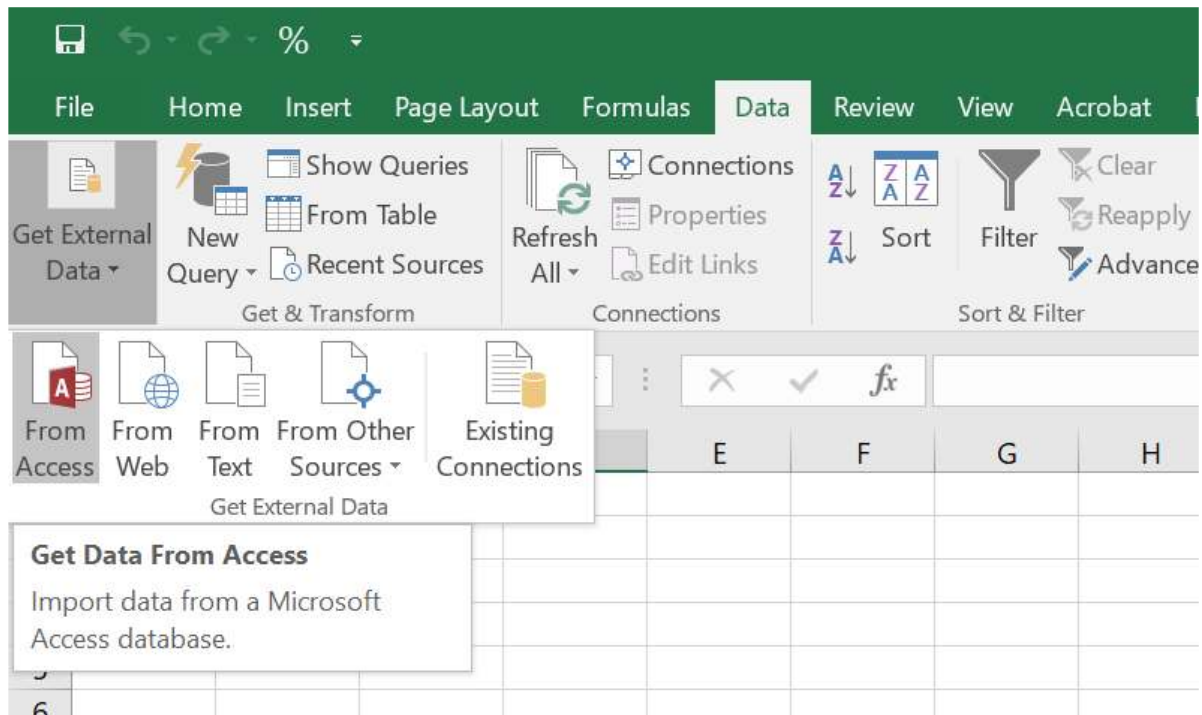
3. We will select the file Patient_Database.accdb from the folder and click Open. Another window appears showing the tables found in the Access database. Tables in a database are similar to worksheets in Excel. Next, check the checkbox Enable selection of multiple tables and select all the tables. Then click OK.
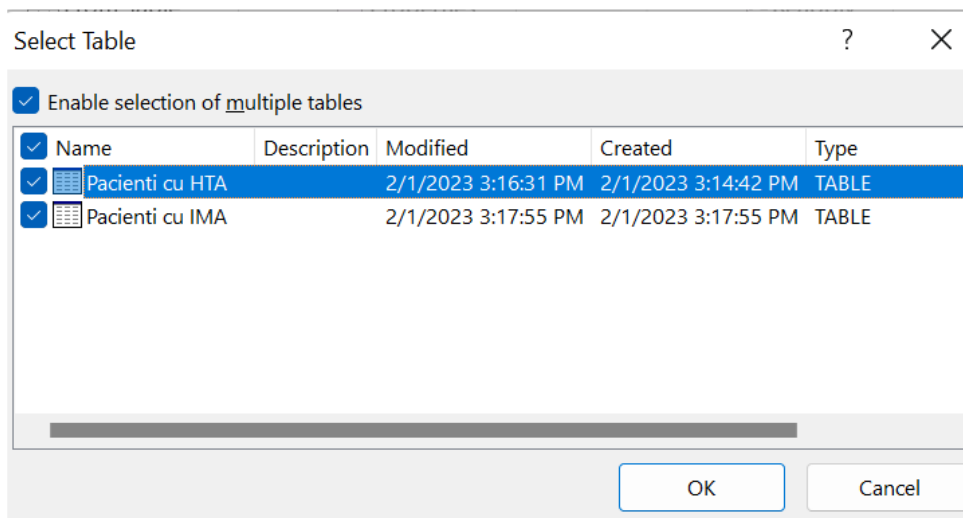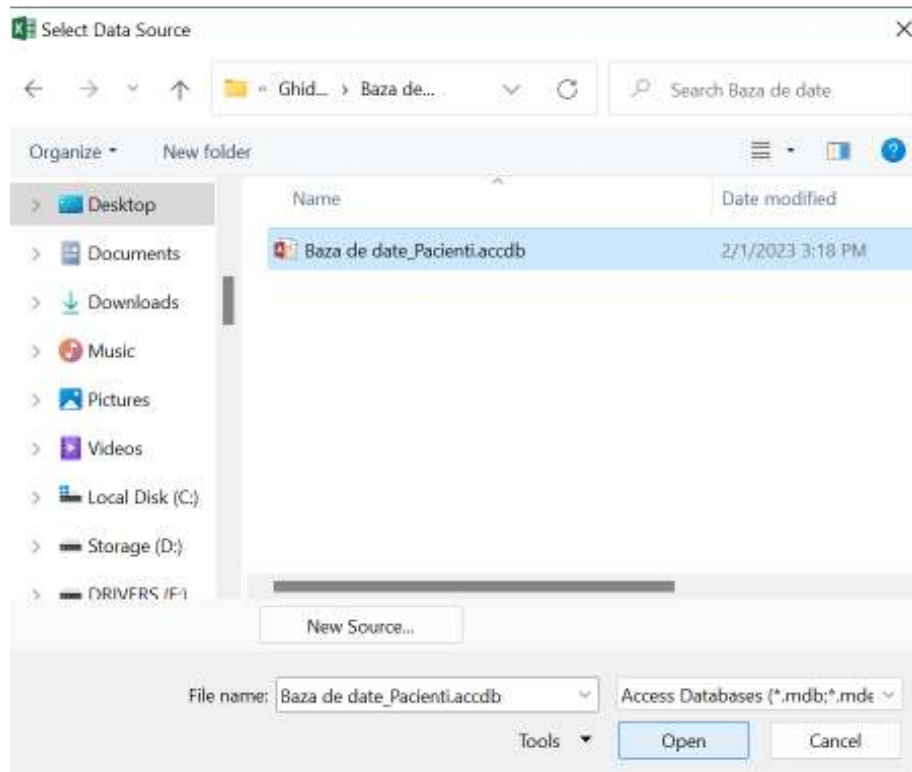
4. The Import Data window appears. We select the PivotTable Report box, which imports the tables into Excel and prepares a PivotTable for analyzing the imported tables, and we check the Existing worksheet box, then click OK.
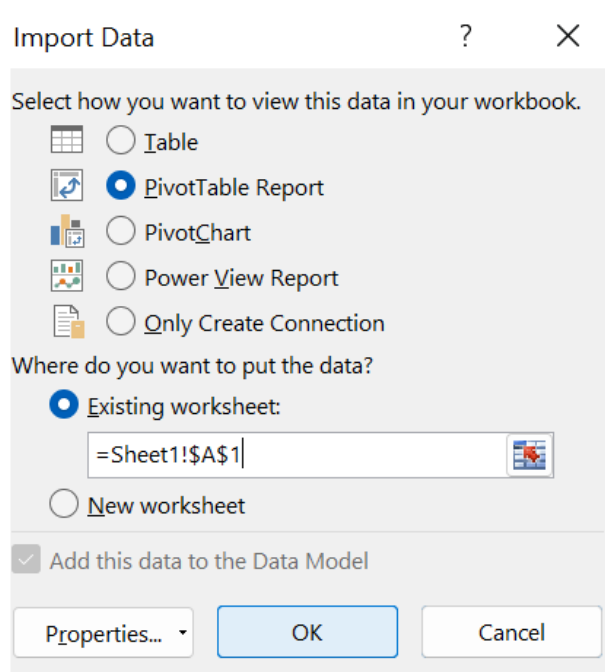
5. After the data is imported, a PivotTable is created using the imported tables. Exploring imported data is easy using a PivotTable. In a PivotTable, you drag fields (similar to columns in Excel) from tables (such as tables just imported from an Access database) to different areas of the PivotTable report to adjust how it presents the data. A PivotTable has four fields: FILTERS, COLUMNS, ROWS, and VALUES.
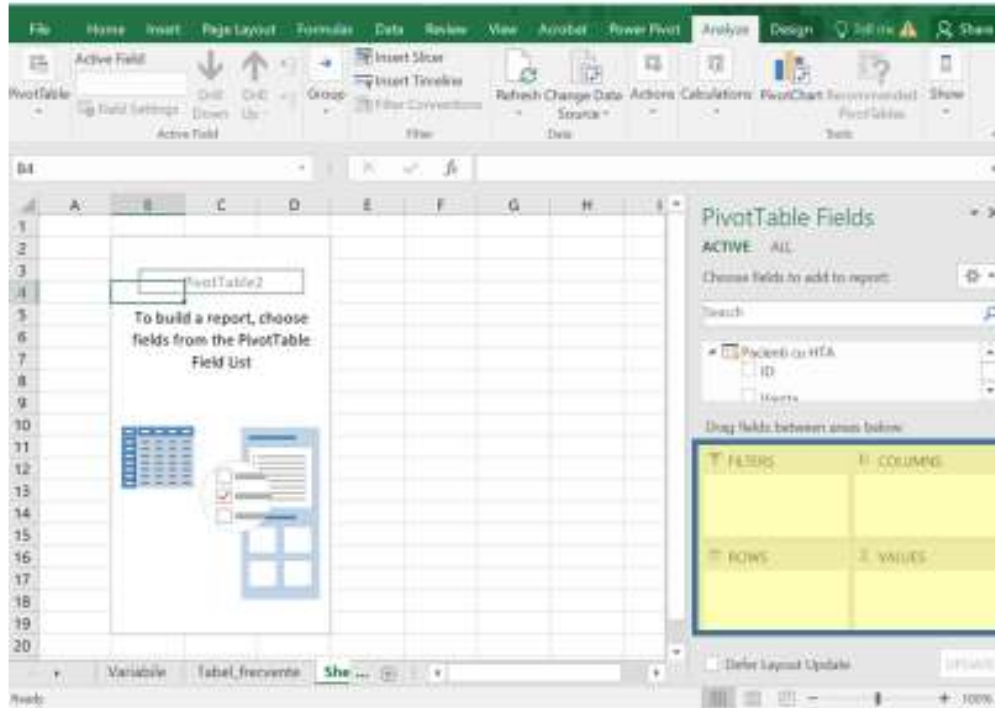
It may take several tries to identify in which area a field should be selected. Depending on the user's wishes, as many or as few fields from the tables can be selected, until the PivotTable displays the data as desired. An example of a PivotTable report is shown in the figure below.

5.2 Import data from a worksheet by copying and pasting

Now let's import data from another source, this time from an existing workbook, and then specify the relationships between our existing data and the new data. Relationships allow you to analyze collections of data in Excel and create interesting and engaging visualizations from imported data.

Let's start by creating a blank worksheet, then import data from an Excel workbook. To import data from an existing Excel worksheet in the computer, the following steps can be taken into account:

1. Insert a new Excel workbook and name it Example.
2. We open the Excel file called Patient_Database.xlsx from the folder on the computer.
3. We select and copy the data from the Variables spreadsheet. To select the entire content of the workbook, use the CTRL+A key combination.
4. On the Example worksheet, we position the cursor in cell A1 and paste the data (Right Click-Paste or CTRL+V) selected from the previous spreadsheet.
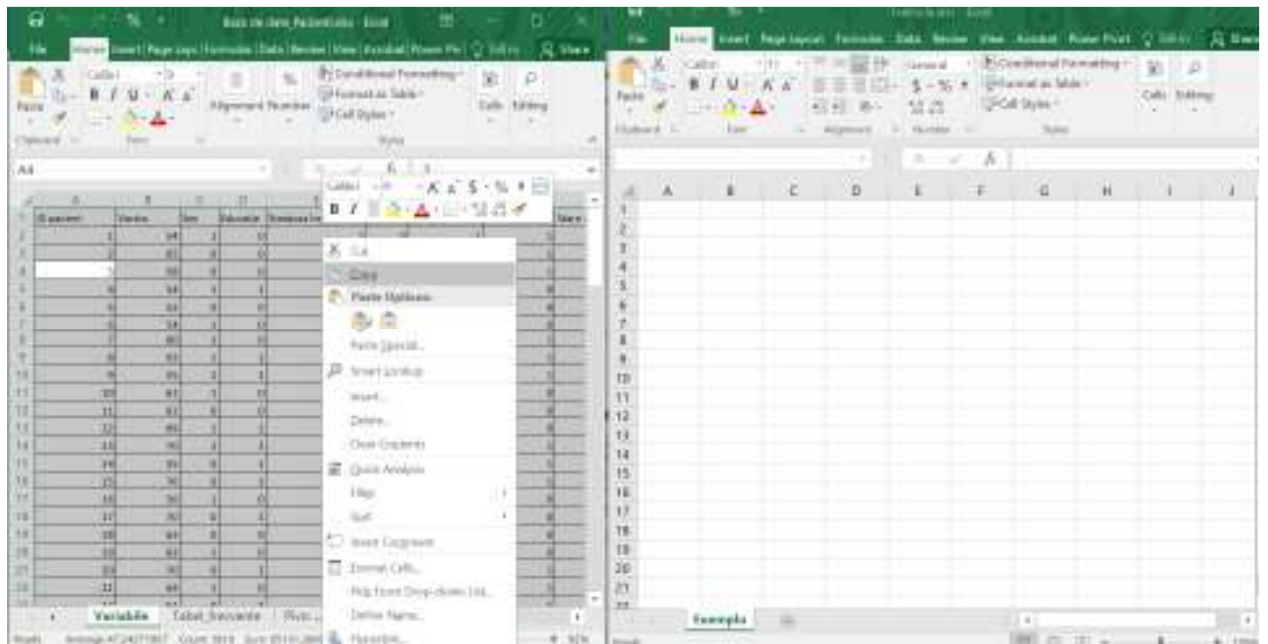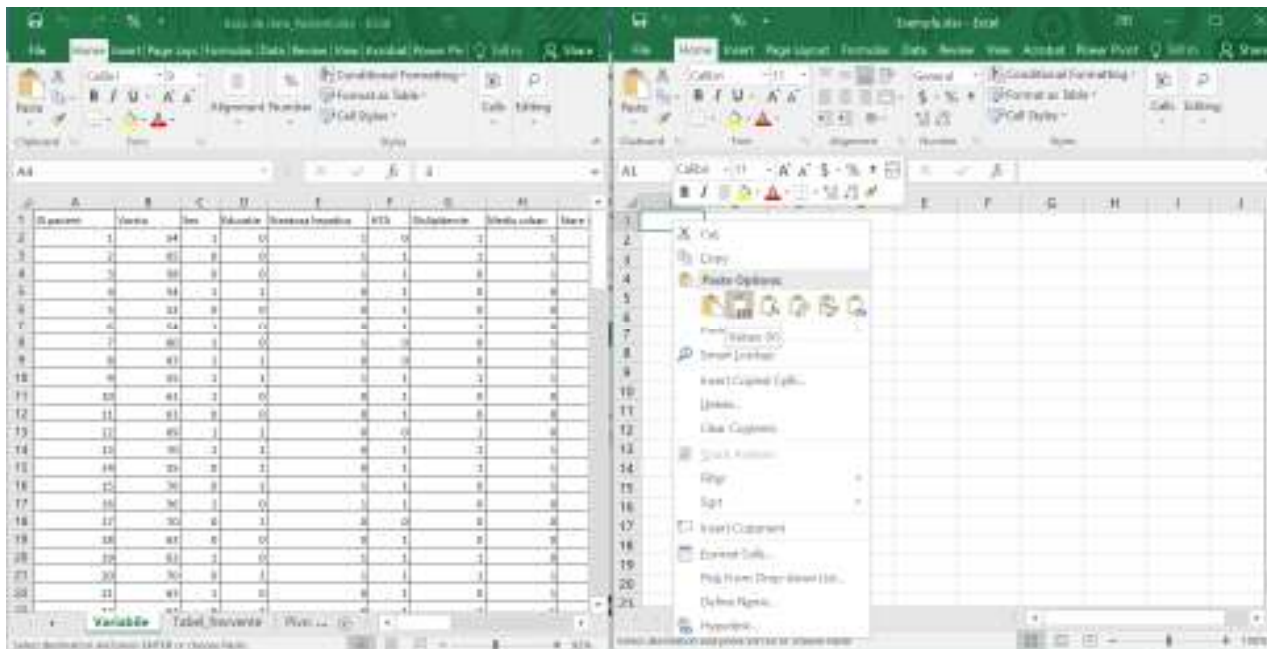
5. Next, with the data selected and highlighted in the register called Example, we use the CTRL+T key combination to format the inserted data as a table. We can also format the data as a table from the ribbon by selecting INSERT – TABLES – CREATE A TABLE. Since each column has a header, check My table has headers in the Create a table window that appears later and click OK at the end.
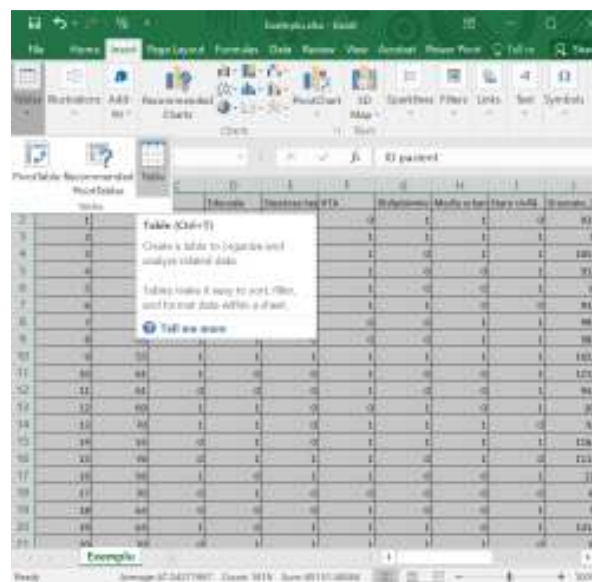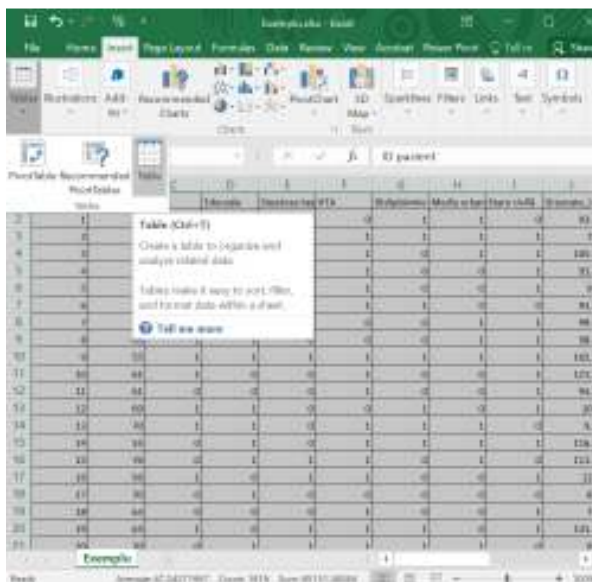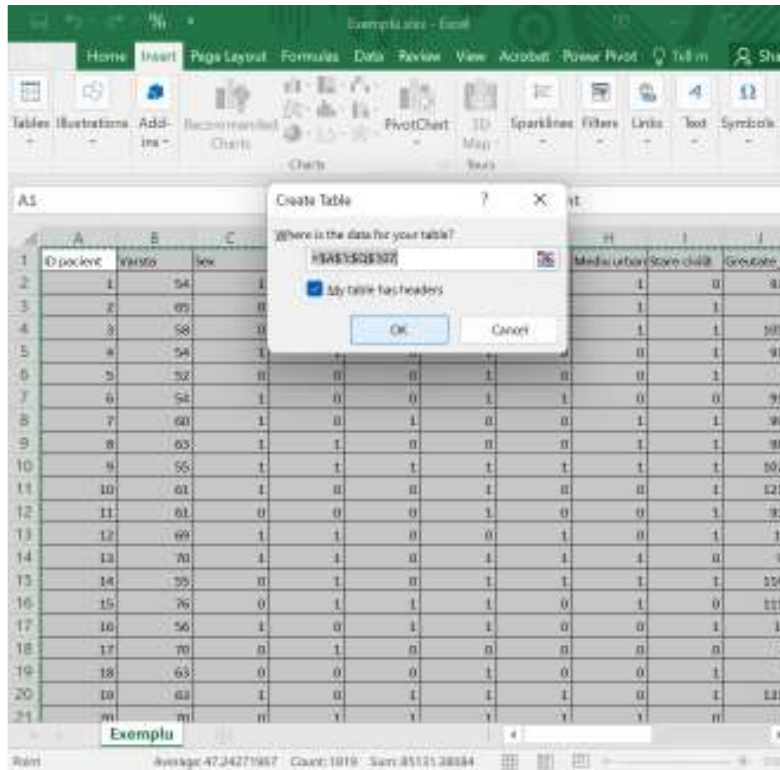
6. To name the table, in the TABLE TOOLS - DESIGN - PROPERTIES menu, find the TABLE
NAME field and type Tabel_Example. And finally we save the Excel file.
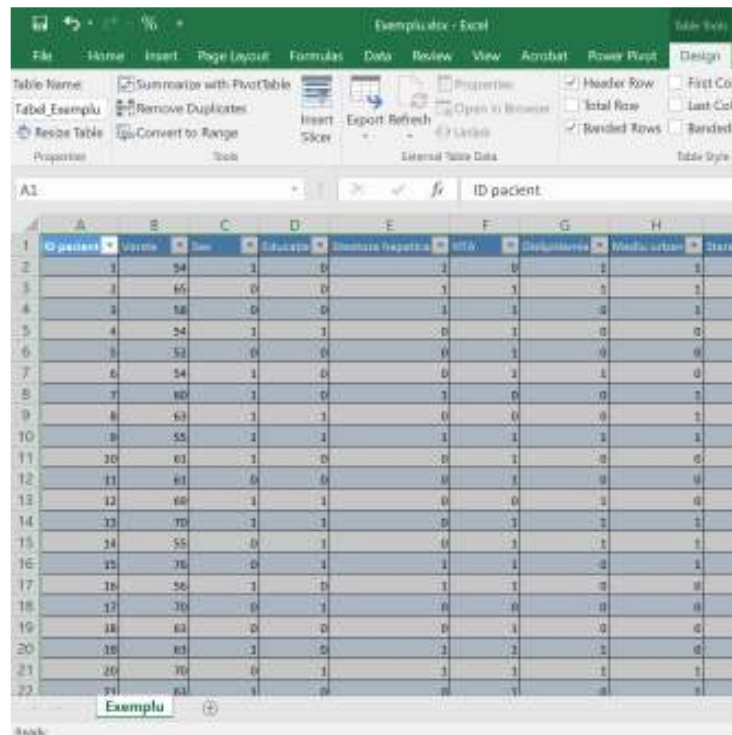
5.3 Importing data from websites

Data and information from websites and other sources can be imported into Excel if they are written in an appropriate format that Microsoft Excel also recognizes.

To import information from websites we will follow the steps explained below:

1. We enter the website www.worldometers.info, which contains information and data about the number of cases of COVID-19 worldwide, copy the URL of the web page: https://www.worldometers.info/coronavirus/# countries.

2. We open an Excel register that we name Coronavirus.

3. From the Data menu, select Get External Data – From Web. After these steps the dialog box opens to enter the URL of the web page with the data we want to import.

4. Paste the URL into the From Web dialog box and click OK. The navigation pane will open with a list of data that can be imported into Excel.
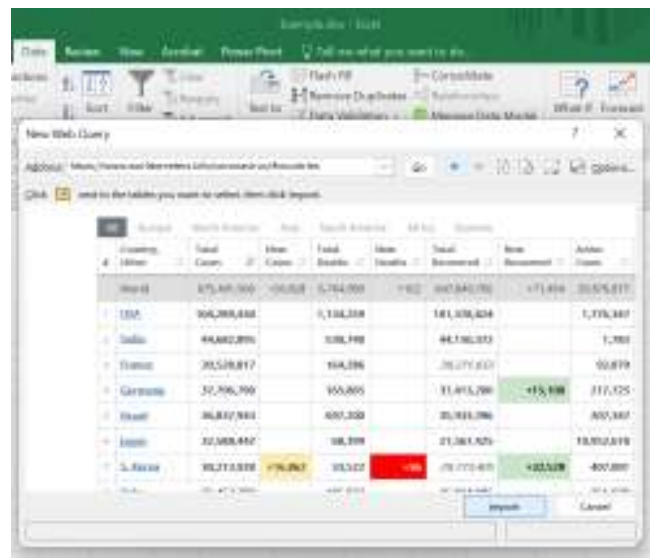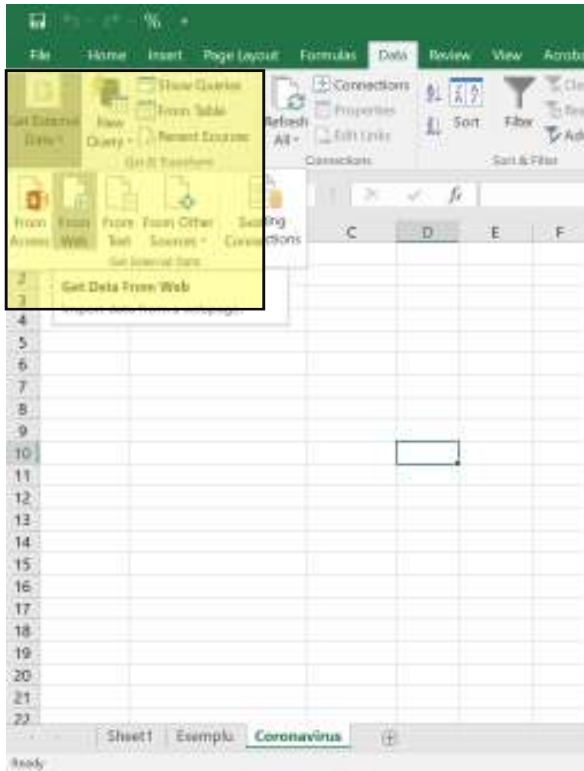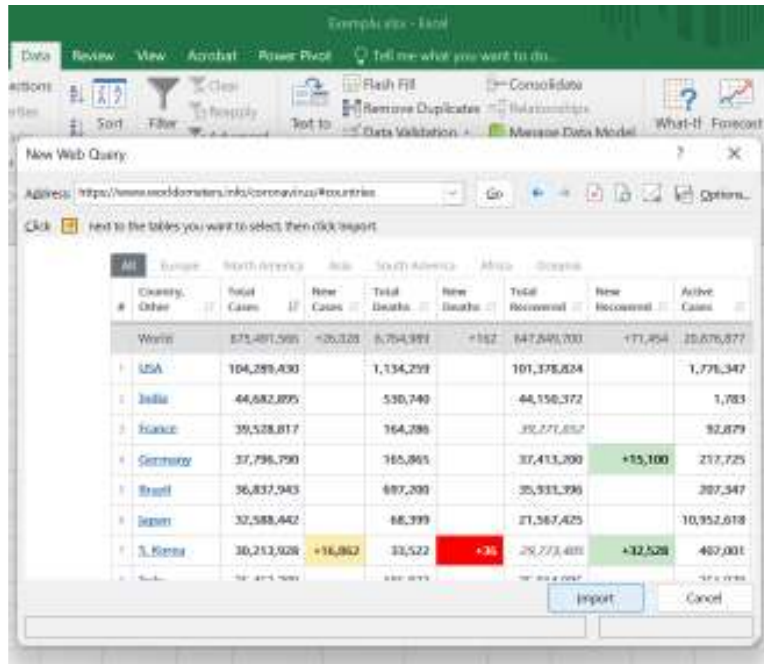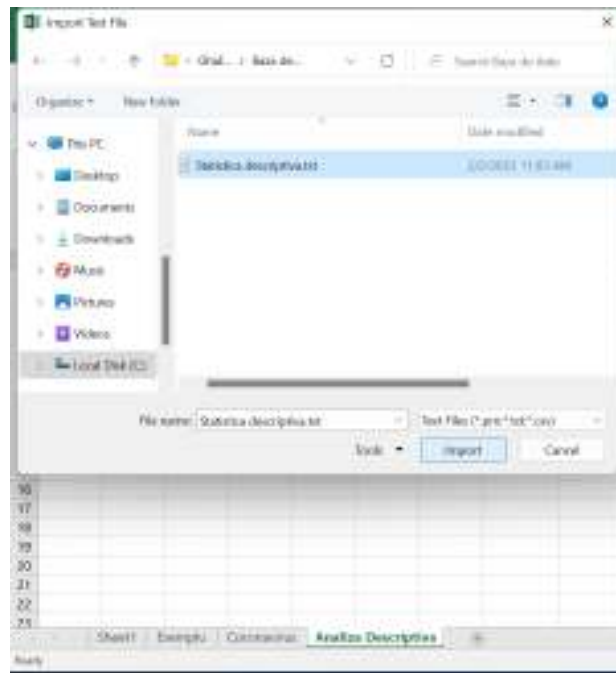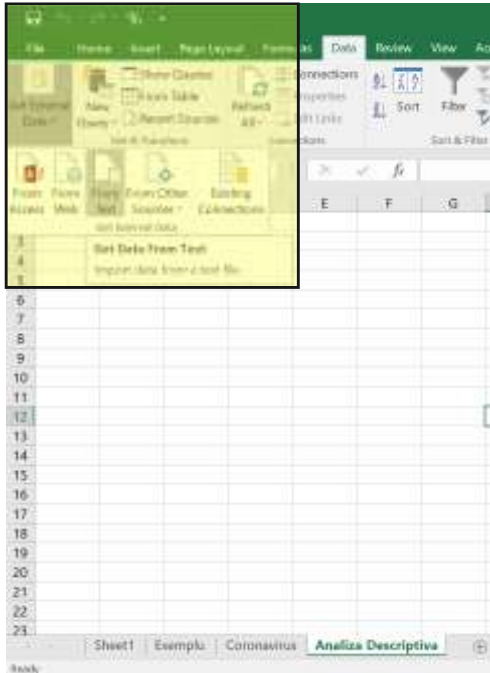
5. In the next window, select the spreadsheet where we want the data to be imported and click OK.
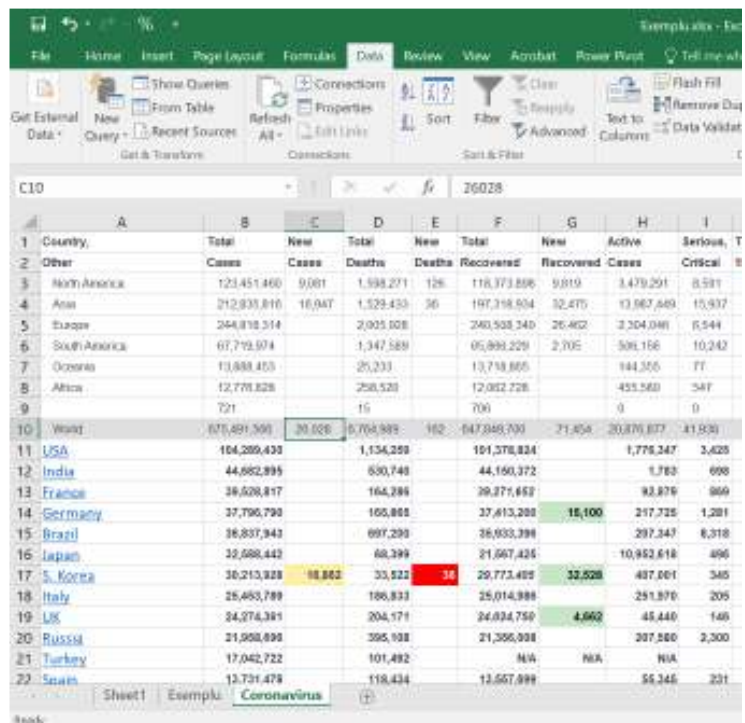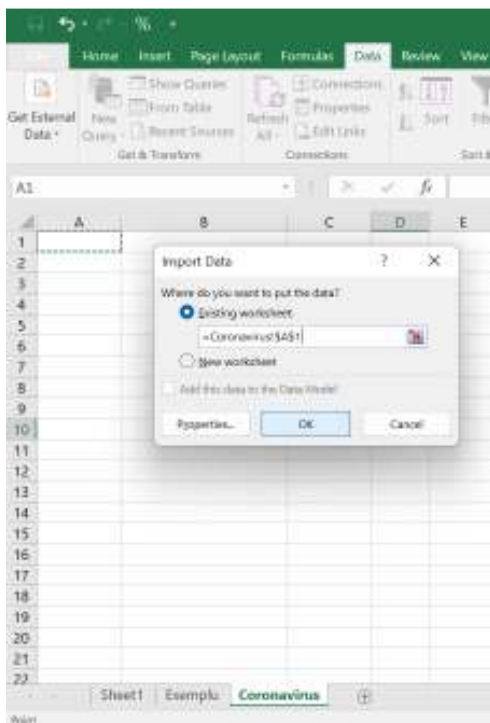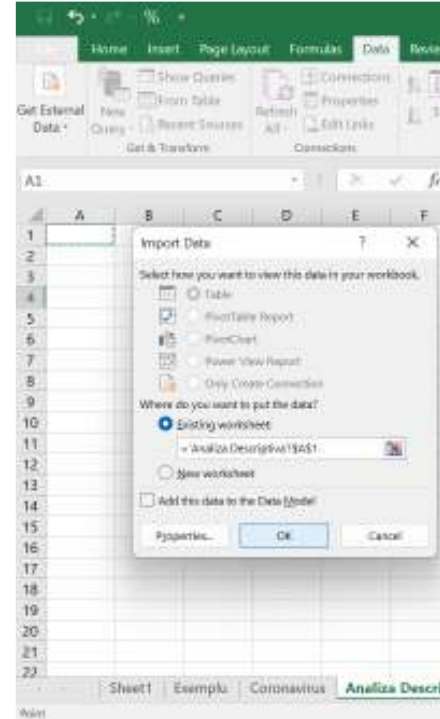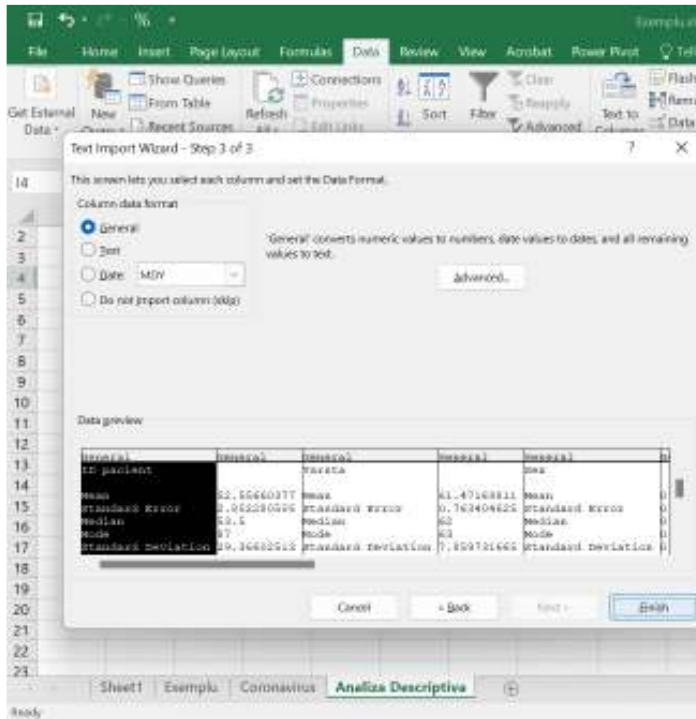
5.4 Importing data from a file

1. To import some data from a file, we created an empty spreadsheet called Descriptive Analysis, and from the Data menu we select Get External Data – From Text. In the next step, the Import Text File window will be displayed where we will have to locate and select the file called Descriptive Statistics.txt from which we want to import the data.

2. Click on Import, and then in the Text Import Wizard window that appears, check Delimited, click Next, in the next window check Tab and General and click Finish.
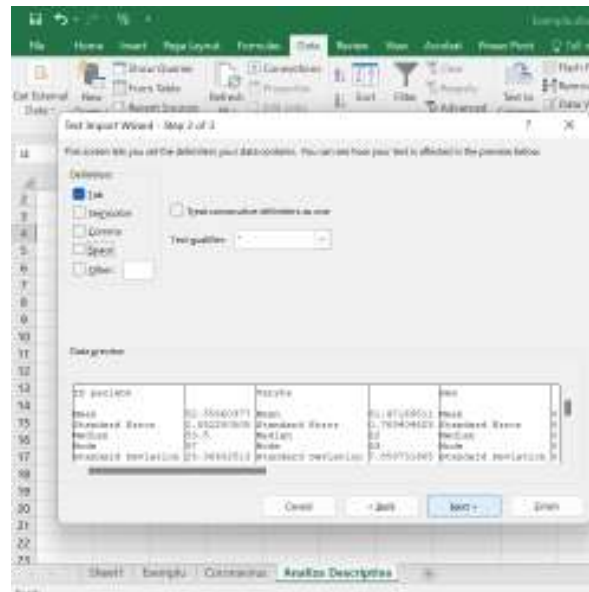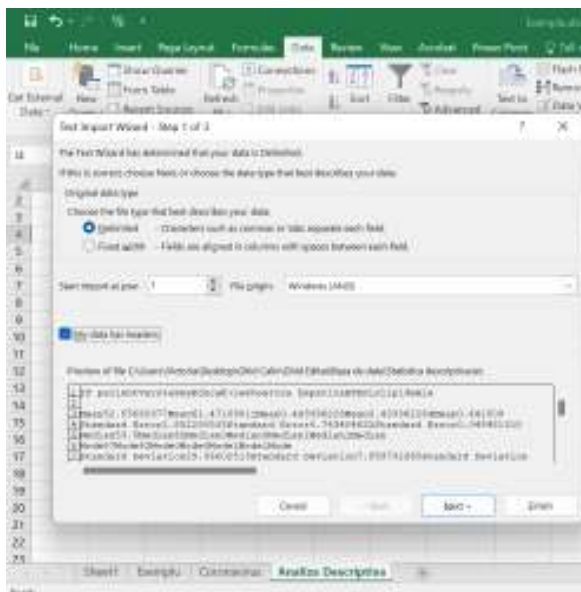
3. In the Import Data window, select the spreadsheet in which we want to import the data. Click OK.

4. The data imported from the .txt file is displayed as follows:

6. DATA CONSOLIDATION

Data consolidation refers to the collection and integration of data from multiple sources into a single destination. During this process, different data sources are brought together or consolidated into a single data warehouse. Because data comes from a wide range of sources, consolidation allows users to more easily present data while facilitating effective data analysis.

Data consolidation techniques reduce inefficiencies such as data duplication, costs related to reliance on multiple databases, and multiple data management points. Therefore it can be of real help to explain the steps to follow for setting up a summary worksheet that consolidates data from existing configured sources. Excel provides this consolidation technique in a few simple and quick steps.

6.1 Preparing data for consolidation

The first thing we must ensure before resorting to the consolidation technique is the existence of consistency between the data sources owned. Data must be arranged in a certain way and have a logic, and this is true for every data source.

Characteristics of the data to be consolidated:

- Each data area should be on a separate worksheet, or separate register.
- Each data area should have list format, with labels in the first row.
- All data areas have the same appearance.
- All data areas contain similar actions
- There should be no empty content (rows and columns) in each area.

6.2 Consolidate data in the same workbook

To consolidate data that is in the same Excel file, we follow the following steps:

1. We have created a separate worksheet for hospitalized patients in the months of July, August, September from several counties of the country, namely July Patients, August Patients, September Patients, we also created a blank worksheet called Report.

2. Next we will work in the Report spreadsheet, where we will click on the cell in which we want the consolidation results to be displayed, then click on the DATA menu, follow the Data Tools section - click on CONSOLIDATE.
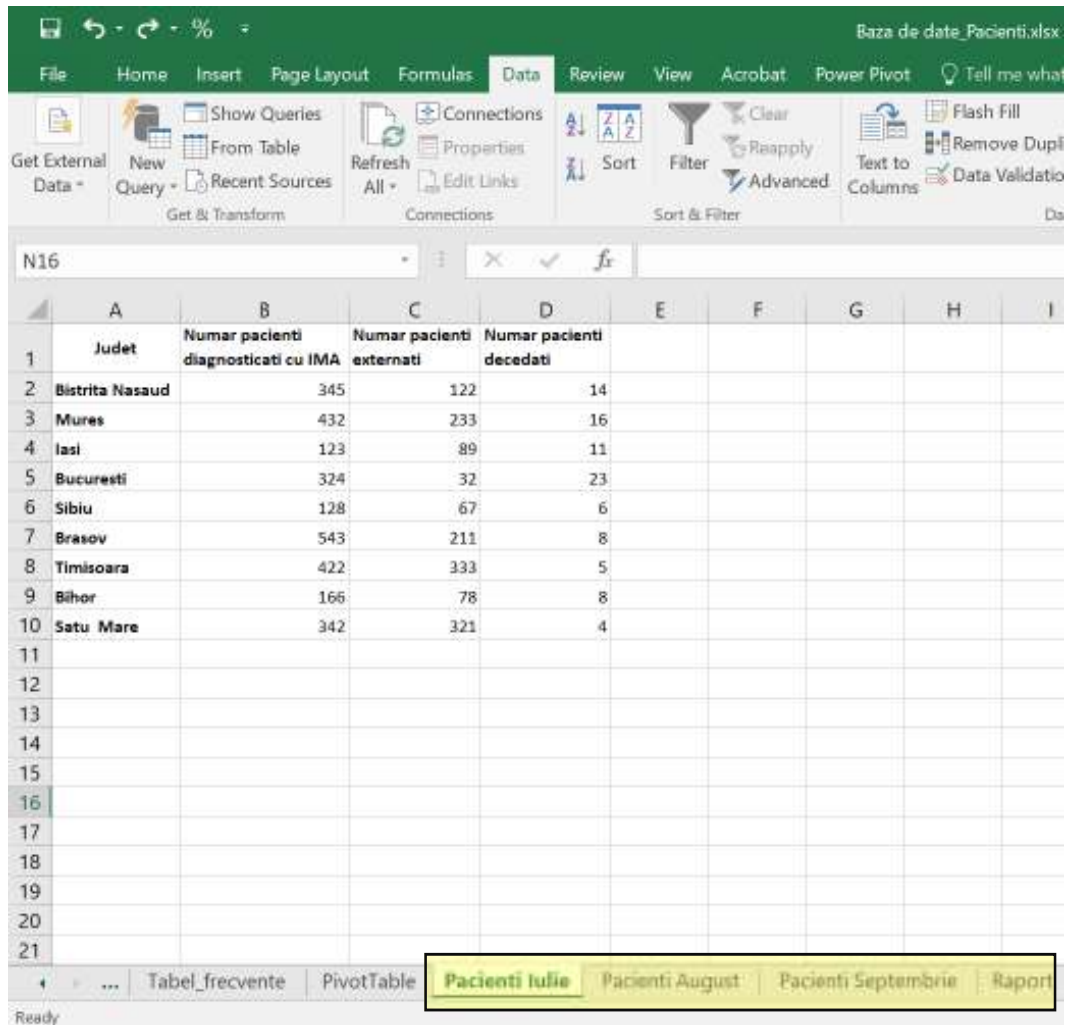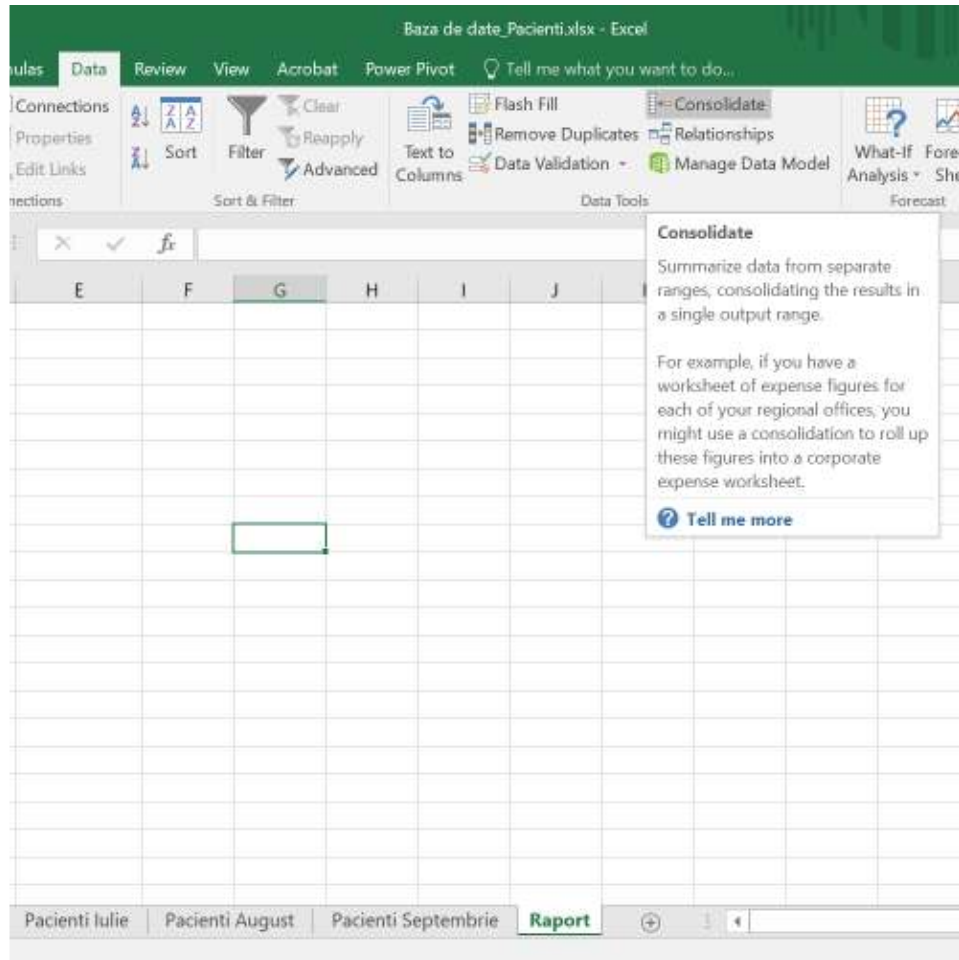
3. The CONSOLIDATED dialog window is displayed where we can select and add the reference areas from the 3 spreadsheets mentioned above. In the Function field we select the option that interests us, namely SUM, COUNT, etc., we check Use labels in First row and Left column, for a good structure of the report we want to create, after which we click OK.

4. We can view the results of the CONSOLIDATED command in the images below.

Screenshot 1 (Data ribbon, Consolidate dialog):

Consolidate

Function:
Sum

Reference:
Pacienti Septembrie'!$A$1:$D$10

All references:
'Pacienti August'!$A$1:$D$10
'Pacienti Iulie'!$A$1:$D$10
'Pacienti Septembrie'!$A$1:$D$10

Use labels in
☑ Top row
☑ Left column   ☐ Create links to source data

OK   Close

	Numar pacie diagnosticati
Judet	
Bistrita Nasa	
Mures	
Iasi	
Bucuresti	
Sibiu	
Brasov	
Timisoara	
Bihor	
Satu Mare	

Sheets: Tabel_frecvente | PivotTable | Pacienti Iulie | Pacienti August | **Pacienti Septembrie** | Raport

---

Screenshot 2 (Raport sheet):

		Numar pacienti diagnosticati cu IMA	Numar pacienti externati	Numar pacienti decedati	
	Bistrita Nasaud	1035	366	42	
	Mures	1296	699	48	
	Iasi	369	267	33	
	Bucuresti	972	96	69	
	Sibiu	384	201	18	
	Brasov	1629	633	24	
	Timisoara	1266	999	15	
	Bihor	498	234	24	
	Satu Mare	1026	963	12	

Sheets: Tabel_frecvente | PivotTable | Pacienti Iulie | Pacienti August | Pacienti Septembrie | **Raport**

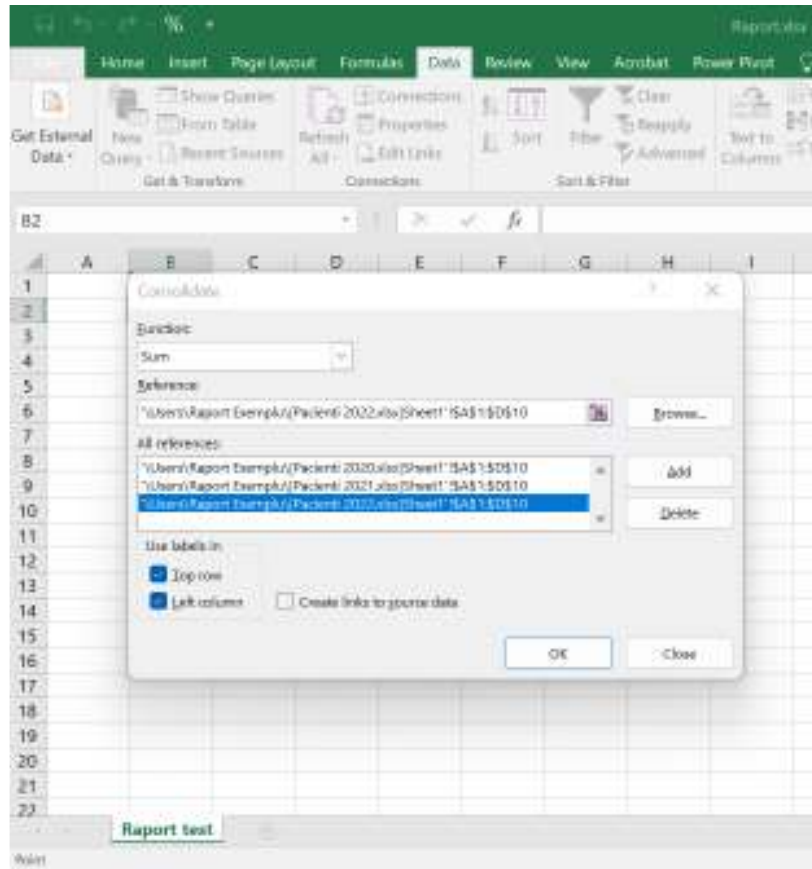**6.3 Consolidation of data from different work registers**

In the situation where we need to consolidate certain data and variables that are stored in different registers, we can refer to the following steps to make the consolidation report.

1. We open the workbooks containing the data and information, named Patients 2020, Patients 2021, Patients 2022.

2. We open a blank Excel workbook called Report, create a worksheet called Test Report, click in the cell where we want to insert the report, then click on the DATA menu, follow the Data Tools section - click on CONSOLIDATE.

3.The CONSOLIDATED dialog box is displayed where we can select and add reference areas from the 3 Excel workbooks mentioned above. In the Function field we select the option that interests us, namely SUM, COUNT, etc., we check Use labels in First row and Left column, for a good structure of the report we want to create, after which we click OK, and we can view the Summary Report .

Consolidate

Function:
Sum

Reference:
"\Users\Raport Exemplu\[Pacienti 2022.xls]Sheet1'!$A$1:$D$10

All references:
"\Users\Raport Exemplu\[Pacienti 2020.xls]Sheet1'!$A$1:$D$10
"\Users\Raport Exemplu\[Pacienti 2021.xls]Sheet1'!$A$1:$D$10
"\Users\Raport Exemplu\[Pacienti 2022.xls]Sheet1'!$A$1:$D$10

Use labels in
☑ Top row
☑ Left column   ☐ Create links to source data

OK   Close

Raport test

	Numar pacienti diagnosticati cu IMA	Numar pacienti externati	Numar pacienti decedati
Bistrita Nasaud	1035	366	42
Mures	1296	699	48
Iasi	369	267	33
Bucuresti	972	96	69
Sibiu	384	201	18
Brasov	1629	633	24
Timisoara	1266	999	15
Bihor	498	234	24
Satu Mare	1026	963	12

Raport test

7. TABLES IN EXCEL

When we refer to a table, we can mention that it represents a rectangular area of structured data. Some of the key features of a table are:

- Each row in the table corresponds to a single data record. An example: patient information.
- Each column contains a specific piece of information. An example: Columns can contain data such as patient identification code, gender, educational level, residence and age, etc.
- The first row from the top exposes the information contained in each column and is called the header row.
- Each entry in the top row is called a column header.

Making tables in Excel is a very fast and easy way to process, manage and analyze the data you have. Filtering, Sorting and Shading techniques are time-saving and greatly facilitate data synthesis and reporting activities.

Another key feature of using tables in Excel is extending formulas and functions from one cell to another.

7.1 Distinction between table and range (area)

Below we will present some aspects related to the differences between a table and a range or area in Excel:

- The use of tables makes the way of working more structured and ordered than intervals.
- In a few steps we can convert a range into a table, and therefore automatically the data will be more structured, the Excel table containing:
  - Table name
  - Column header names
  - Data formatting (font and cell color) for viewing and highlighting
- The existence of additional features for tables that are not available for ranges, such as:
  - Table editing tools in the ribbon, from properties to styles.
  - The existence of the Filter button in each column header to sort the data or filter the table so that only the rows that meet the desired criteria are displayed.
  - The table header is preserved when scrolling the rows or the spreadsheet.
  - Extending formulas and functions from one cell to another.
  - Expand the size of the table by adding more rows or more columns by clicking and dragging the small triangular control in the lower right corner of the lower right cell.

7.2 Creating a table in Excel

Below we will present the steps to follow to create a table with existing data in the form of a range:

1. We will select the range of cells that we want to convert into a table. Cells can contain information or be empty. In the following image, we will show a spreadsheet that contains a range with 107 rows of patient data. The first row contains the column label (header).

2. We will select the entire range, after which from the Insert menu we will click on Table, in the next window called Create Table, we check if the data range selected in the Where is the data for your table? dialog box. Check My table has headers, and click OK.
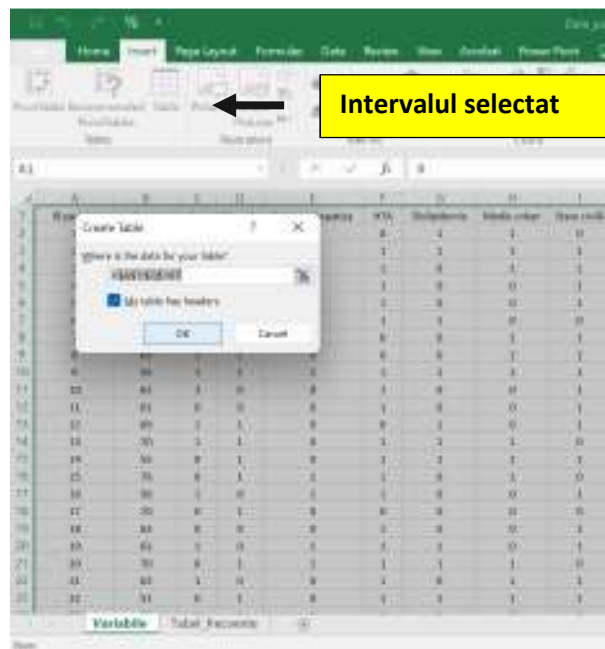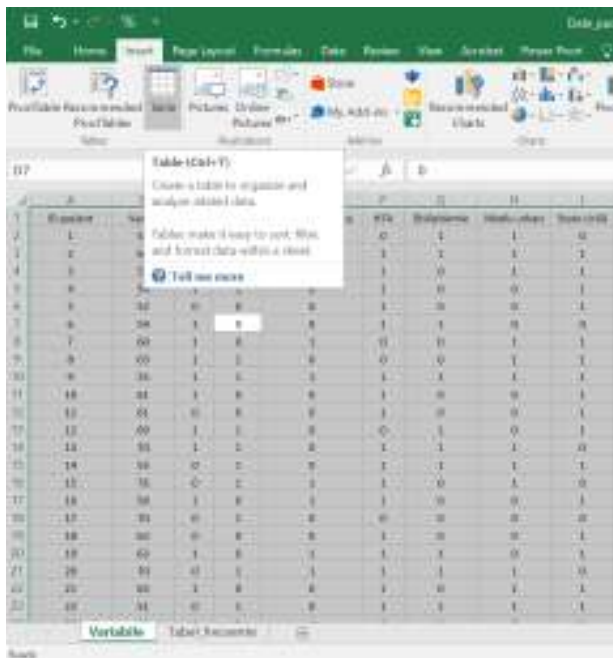
3. The selected range is converted to Table with the default style.
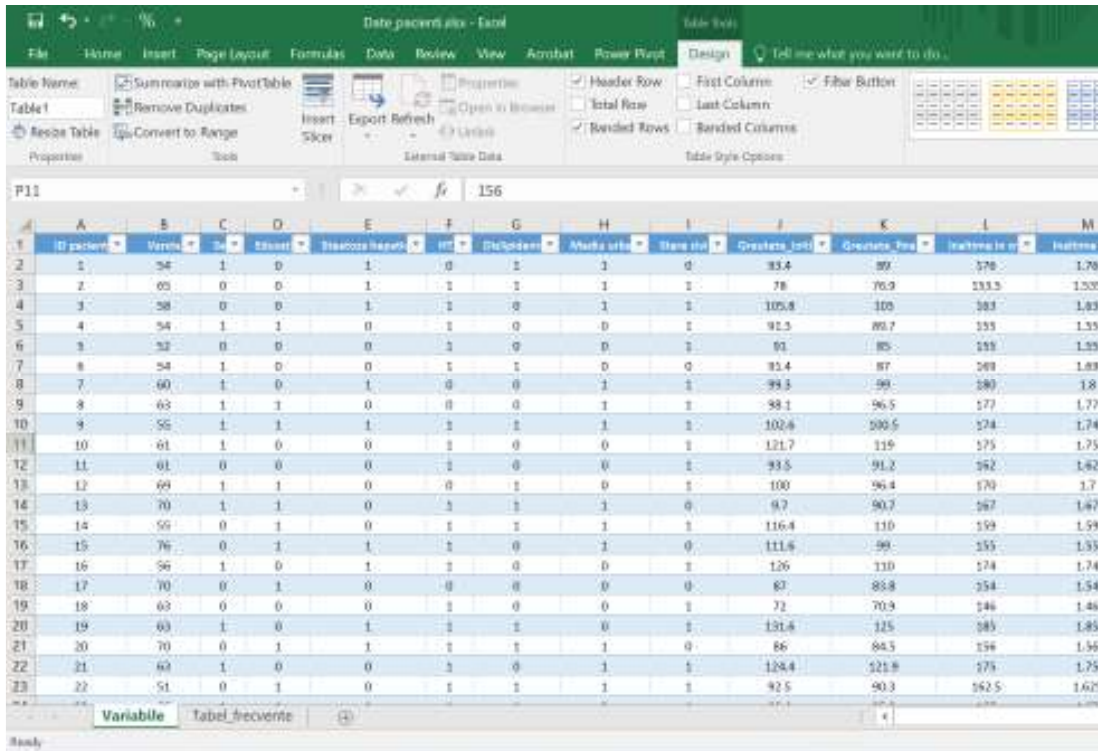
4. In the Table Tools - Design menu, we can assign names to the created table. In the Properties section, the Table Name box will be displayed and can be changed from there.
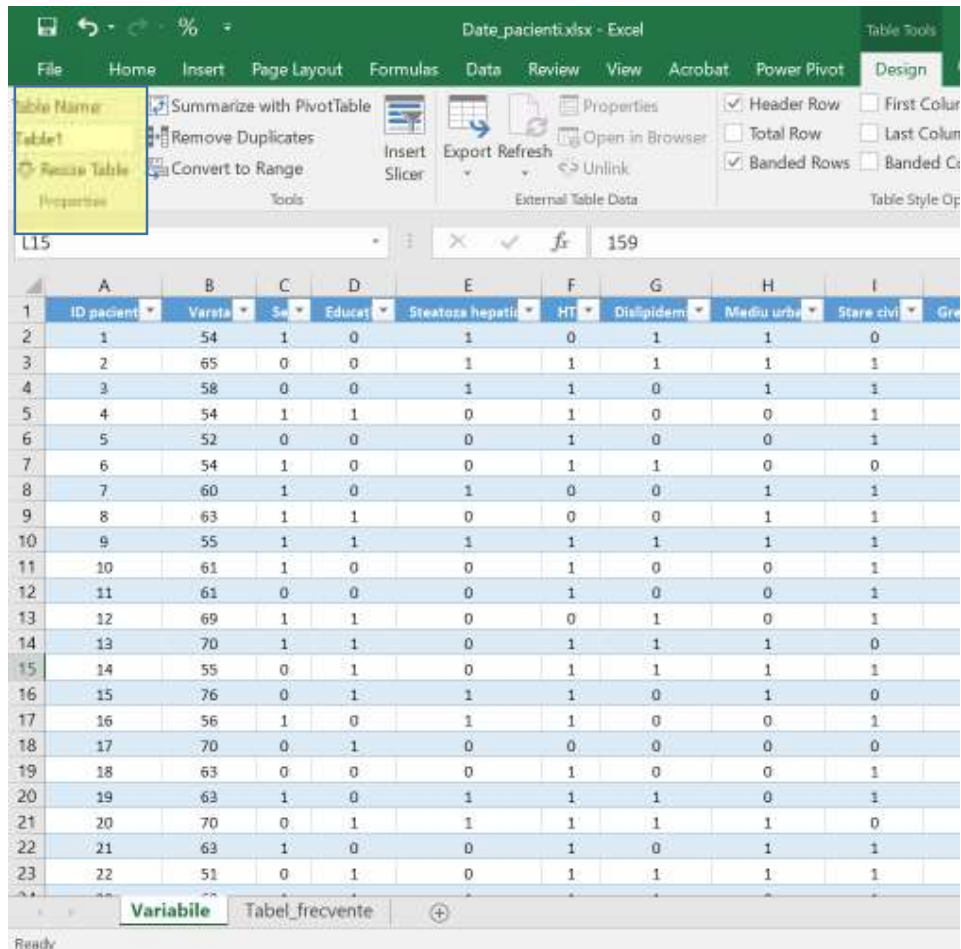
7.3 Managing names in a table

With the help of Name Manager we can manage the names of the created tables in the same way as you manage the names of the ranges.

1. Click on the inserted table, then click on Formulas – Defined Names - Name Manager. The Name Manager dialog box appears and we can find the names of the tables in the spreadsheet named Variables.
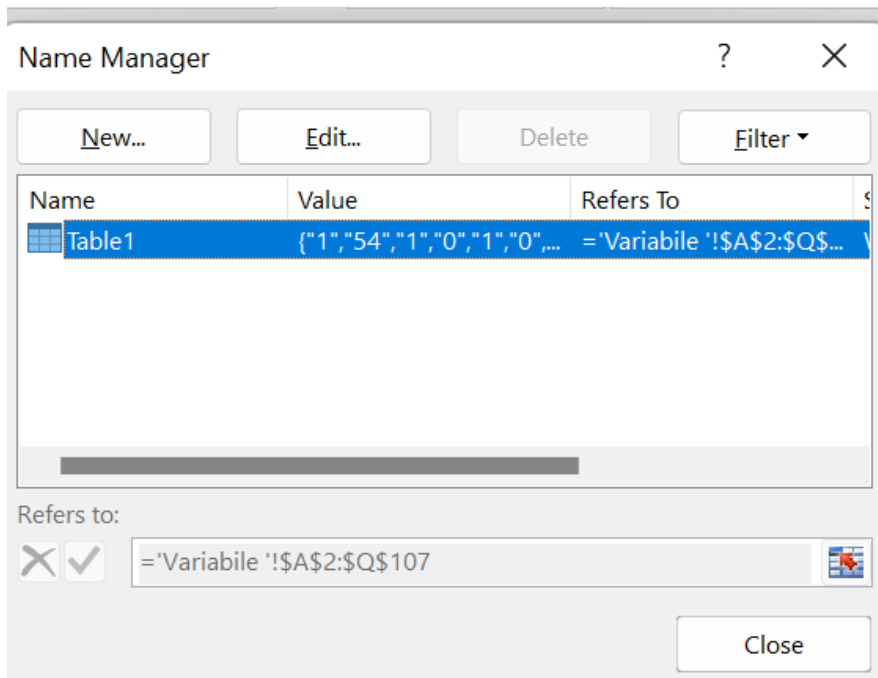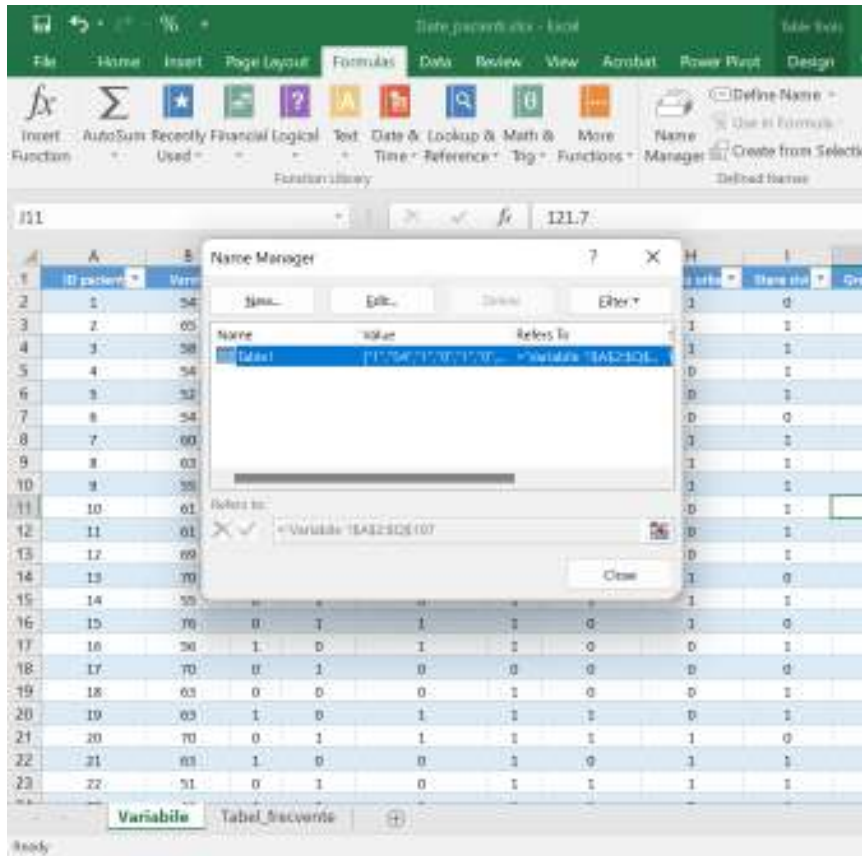
2. The table we created is named Table1, as can be seen in the images above, but we want to change the name from Table1 to Patients, to do this in the Name Manager window we click on Edit.
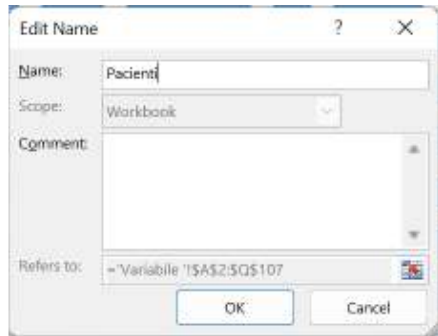
Also, in addition to the name of the table, we can also modify the names of the column headers. For example, we would like to change the column header called Education to Education_Level, and to do that we must: click on the Education column header in the table, then click on Name Manager from the Formulas menu (as in the previous steps). After the dialog box appears, click New. In the Name box, find the Header of the column we want to edit.
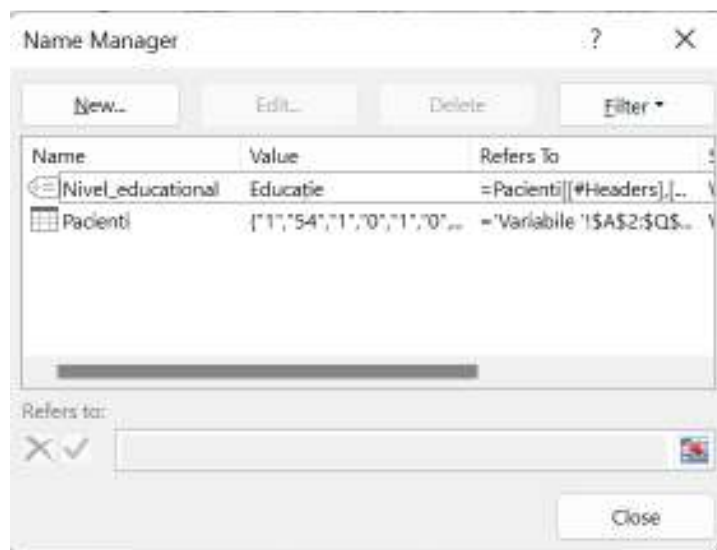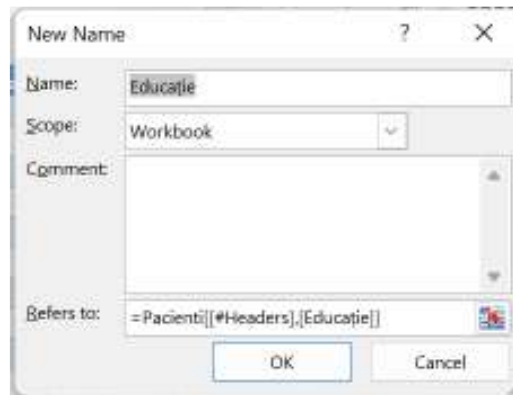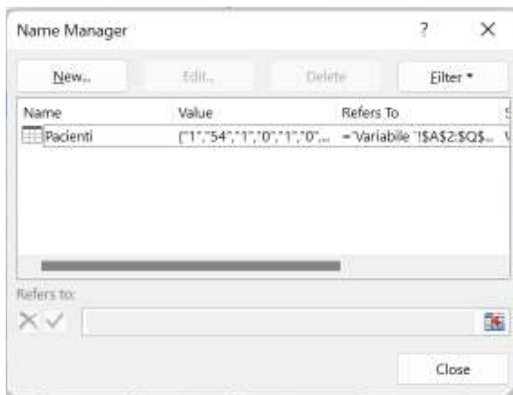
7.4 Propagation of a formula in a table

In the table called Patients, we also have the height and weight anthropometric indicators recorded, and in the following example we will demonstrate how we calculate the Body Mass Index (BMI) by propagating the formula to all patients in the table.

1. We will create 2 new (blank) columns in the table, named IMC_initial and IMC_final, and in any of the cells of the IMC_initial column we type the calculation formula for BMI, =[@[Initial_Weight]]/([@[Height in m]]^2) and type Enter.

Stare civil	Greutate_initial	Greutate_final	Inaltime in cm	Inaltime in m	IMC_initial	IMC_final	Glicemie 1	Glicemie 2
0	93.4	89	176	1.76			149	109
1	78	76.9	153.5	1.535			73	100
1	105.8	103	163	1.63			121	94
1	91.3	89.7	155	1.55			105	93
1	91	85	155	1.55			105	93
0	91.4	87	169	1.69			123	90
1	99.3	99	180	1.8			131	100
1	98.1	96.5	177	1.77			70	70
1	102.6	100.5	174	1.74			103	85
1	121.7	119	175	1.75			156	70
1	93.5	91.2	162	1.62			117	92
1	100	96.4	170	1.7			91	91
0	9.7	90.7	167	1.67			88	80
1	116.4	110	159	1.59			115	107
0	111.6	99	155	1.55			103	102
1	126	110	174	1.74			206	145
0	87	83.8	154	1.54			157	132
1	72	70.9	146	1.46			144	125
1	131.6	125	185	1.85			103	90
0	86	84.5	156	1.56			139	139
1	124.4	121.9	175	1.75			104	104
1	92.5	90.3	162.5	1.625			118	120

Stare civi	Greutate_initial	Greutate_final	Inaltime in cm	Inaltime in	IMC_initial	IMC_final
0	93.4	89	176	1.76	30.15	
1	78	76.9	153.5	1.535	33.10	
1	105.8	103	163	1.63	39.82	
1	91.3	89.7	155	1.55	38.00	
1	91	85	155	1.55	37.98	
0	91.4	87	169	1.69	=[@[Greutate_initi	
1	99.3	99	180	1.8	30.65	
1	98.1	96.5	177	1.77	31.31	
1	102.6	100.5	174	1.74	33.89	
1	121.7	119	175	1.75	39.74	
1	93.5	91.2	162	1.62	35.63	
1	100	96.4	170	1.7	34.60	
0	9.7	90.7	167	1.67	3.48	
1	116.4	110	159	1.59	46.04	
0	111.6	99	155	1.55	46.45	
1	126	110	174	1.74	41.62	
0	87	83.8	154	1.54	36.68	
1	72	70.9	146	1.46	33.78	
1	131.6	125	185	1.85	38.45	
0	86	84.5	156	1.56	35.34	
1	124.4	121.9	175	1.75	40.62	
1	92.5	90.3	162.5	1.625	35.03	

The calculation formula for BMI is automatically propagated to the other cells in that column of the table.

7.5 Resizing the table

Excel also provides the option to resize an inserted table, we can remove or add columns/rows within the table. We have the table called Patients which contains 107 rows, and we want to add 3 more patients, that is we will insert 3 more rows.

1. We click on the table and we will scroll the page to the end of the table, and in the lower right corner of the table there is a dot called Control, we will click on Control and drag down to insert 3 rows and to the right to insert a new one columns.

O	P	Q	R
.06	106	92	
.09	264	164	
.06	172	102	
.08	114	120	
.07	124	123	
.10	148	104	
.06	153	113	
.06	292	203	**Control** ←

Inaltime i ▼	IMC_initial ▼	IMC_final ▼	Glicemie 1 ▼	Glicemie 2 ▼
1.73	44.44	0.06	106	92
1.73	33.81	0.09	264	164
1.6	43.83	0.06	172	102
1.66	30.66	0.08	114	120
1.485	33.28	0.07	124	123
1.8	29.94	0.10	148	104
1.7	41.90	0.06	153	113
1.65	41.10	0.06	292	203

Control ←

After inserting the blank rows and column, the table will look like the image below.

	Inaltime i ▾	IMC_initial ▾	IMC_final ▾	Glicemie 1 ▾	Glicemie 2 ▾	Column1 ▾	S
100	1.73	44.44	0.06	106	92		
101	1.73	33.81	0.09	264	164		
102	1.6	43.83	0.06	172	102		
103	1.66	30.66	0.08	114	120		
104	1.485	33.28	0.07	124	123		
105	1.8	29.94	0.10	148	104		
106	1.7	41.90	0.06	153	113		
107	1.65	41.10	0.06	292	203		
108							
109							
110							
111							
112							

7.6 Remove duplicate data in Excel

Depending on the source of the collected data and how it is collected, it is very likely that the databases contain duplicate data and information. For correct and valid data analysis, removing duplicate values and cleaning the database is absolutely necessary.
Look at the following data where you have information about various products of different brands. Let's say you want to remove duplicates from this data.

1. We have a table containing active ingredient and OTC drug name, and we want to check and remove duplicate values. We click on the table that contains the data of interest, and from the Table Tools - DESIGN menu, click on Remove Duplicates in the Tools section.

2. The Remove Duplicates window is displayed, which contains the names of the existing columns in the table, check the columns and click OK.

After scanning for duplicates, a message will be displayed about how many rows with duplicate values will be removed and how many unique values remain.
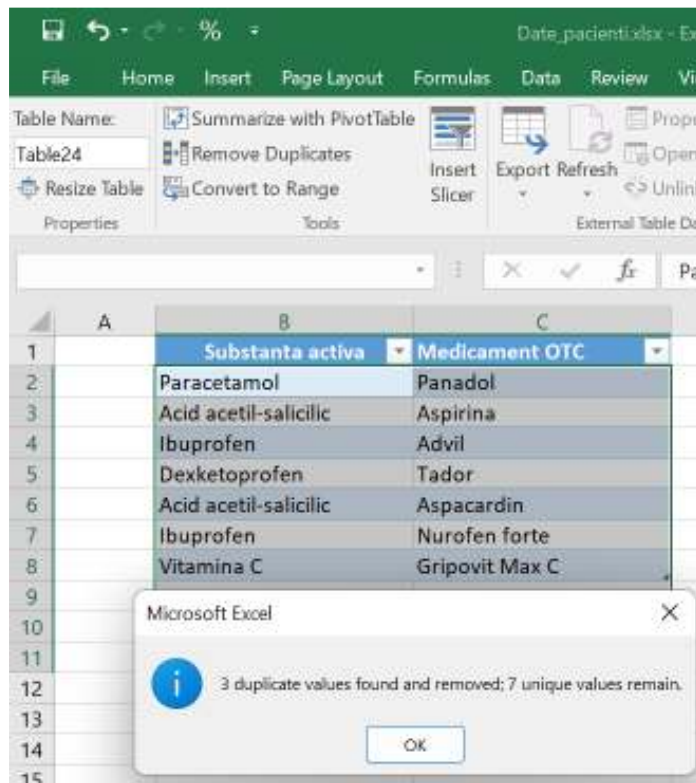

7.7 Convert table to range (area)

If Excel makes available the conversion of a range to a table, we of course also have the opposite option, the conversion of a table to a range. As an example, we have a table that contains the active ingredient and OTC drug name, and we want to convert it to range. We click on the table that contains the data of interest, and from the Table Tools – DESIGN menu, click on Convert to Range in the Tools section.
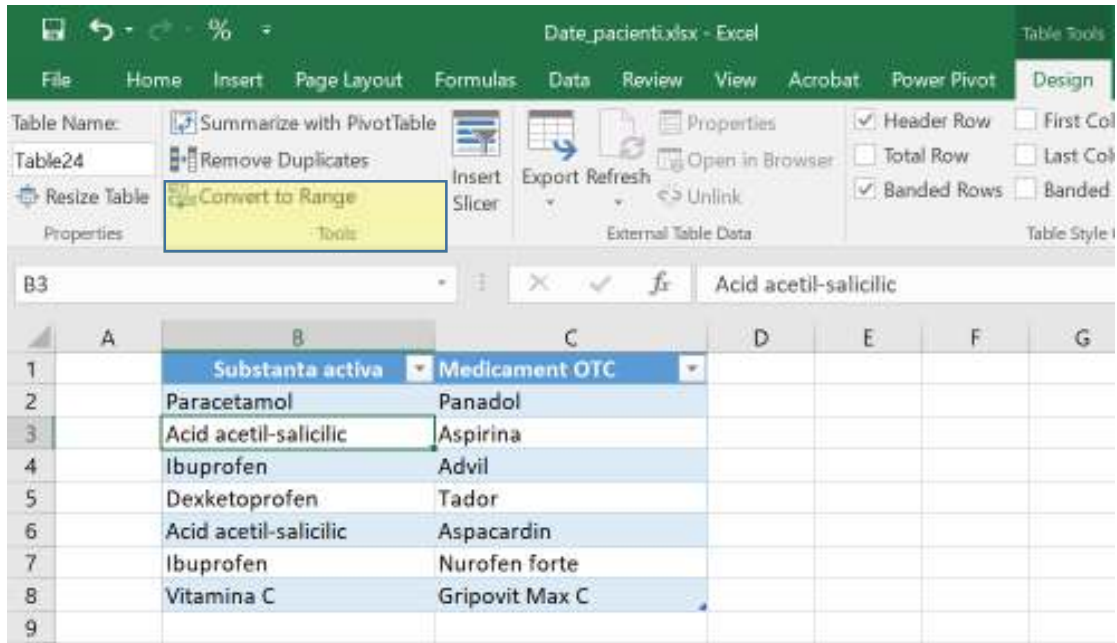
The result can be seen in the figures below.
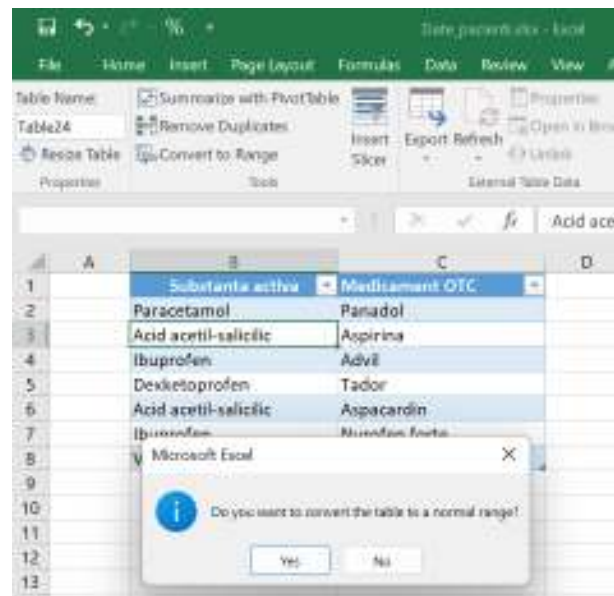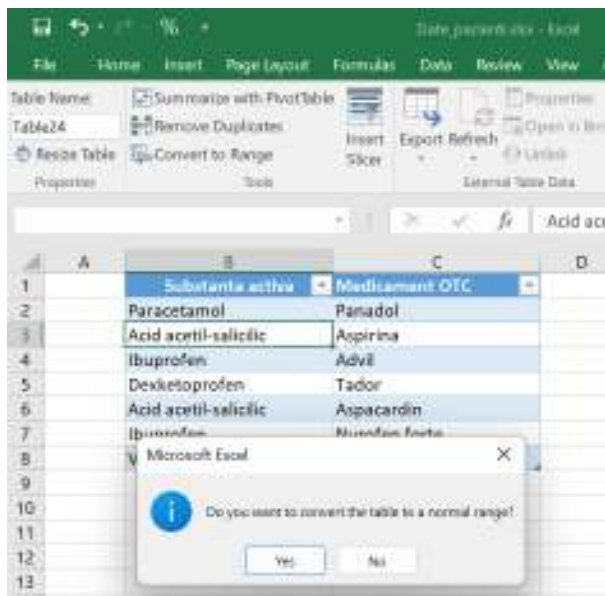
8. DATA CLEANING WITH TEXT FUNCTION

As we have mentioned several times throughout this guide, it is very likely that the data we obtain from various data collection sources will not present a format ready for proper analysis.

Once collected in its raw form, the data must be cleaned, and the process of data cleaning is the removal of unwanted characters from the text. Data structuring procedures for analysis are:

- Identifying required text patterns by using text functions.
- Extracting data values from text.
- Formatting data using text functions.
- Performing data operations with text functions.

8.1 Removing unwanted characters from text

In some cases it happens that the imported data has non-printable characters and/or excess spaces. Excess spaces can be leading spaces and/or extra spaces between words.

We present an example of such data in the image below, where you can view the raw information of a report related to the active substance, the name of the medicine, the brand and the price of the medicine, and in the following steps we will present techniques to clean the data of unwanted characters. The data is imported in a single column and the character"|" separates the field in each row, and some rows contain extra spaces that need to be removed.

Raport
Substanta activaMedicament OTCBrandPret
Paracetamol PanadolSandoz34.5
Acid acetil-salicilicAspirinaGee|on55.6
IbuprofenAdv|ilBiopharma123
DexketoprofenTadorFarmexim22
Acid acetil-salicilicAspacardinFarmexpert134
IbuprofenNurofen fort|eSandoz322
Vitamina CGripovit Max CBioe|el123

To remove excess spaces we can use CLEAN and the TRIM function.

- CLEAN - Removes all non-printable characters from text
- TRIM - Removes spaces from text

fx	=TRIM(CLEAN(F5))	
	F	**G**
	Raport	**Date curatate**
	Substanta activaMedicament OTCBrandPret	=TRIM(CLEAN(F5))
	Paracetamol PanadolSandoz34.5	
	Acid acetil-salicilicAspirinaGee|on55.6	
	IbuprofenAdv|ilBiopharma123	
	DexketoprofenTadorFarmexim22	
	Acid acetil-salicilicAspacardinFarmexpert134	
	IbuprofenNurofen fort|eSandoz322	
	Vitamina CGripovit Max CBioe|el123	

We select the area in the adjacent column corresponding to the area containing the collected data, type =TRIM (CLEAN (F5)) and then press CTRL+Enter, and the formula propagates throughout the area, cleaning the data of excess spaces.

Raport	**Date curatate**
Substanta activaMedicament OTCBrandPret	Substanta activaMedicament OTCBrandPret
Paracetamol PanadolSandoz34.5	Paracetamol PanadolSandoz34.5
Acid acetil-salicilicAspirinaGee|on55.6	Acid acetil-salicilicAspirinaGee|on55.6
IbuprofenAdv|ilBiopharma123	IbuprofenAdv|ilBiopharma123
DexketoprofenTadorFarmexim22	DexketoprofenTadorFarmexim22
Acid acetil-salicilicAspacardinFarmexpert134	Acid acetil-salicilicAspacardinFarmexpert134
IbuprofenNurofen fort|eSandoz322	IbuprofenNurofen fort|eSandoz322
Vitamina CGripovit Max CBioe|el123	Vitamina CGripovit Max CBioe|el123

8.2 Identifying text patterns needed with text functions

In order to perform the data structuring process, it is sometimes necessary to perform some text pattern matching based on which we can extract the data values. For this purpose we can use the text functions:

• EXACT - Checks if two text values are identical

• FIND - Finds a text value within another value (case sensitive)

• SEARCH - Finds a text value within another value (case insensitive)

8.3 Extracting data values from text

In the example of point 8.1, we have imported the data in a single column, but in order to analyze it, we need to process it so that it is displayed in 4 columns: Active substance, OTC drug, Brand and Price. To extract the data, we can use the following methods:

a. Extracting data values with the Text to Columns Wizard into columns

b. Extracting data values with text functions

c. Extract data values with Flash Fill

8.3.1 Extracting data values with the Text to Columns Wizard

We use the Text to Columns Wizard to extract data values into Excel columns if the fields are character delimited or column aligned with spaces between each field. In our example, the fields are separated by the "|" character.

1. We select the data, copy it and paste it in the same place in the form of Paste as values so that the formula previously applied to clean the data disappears.
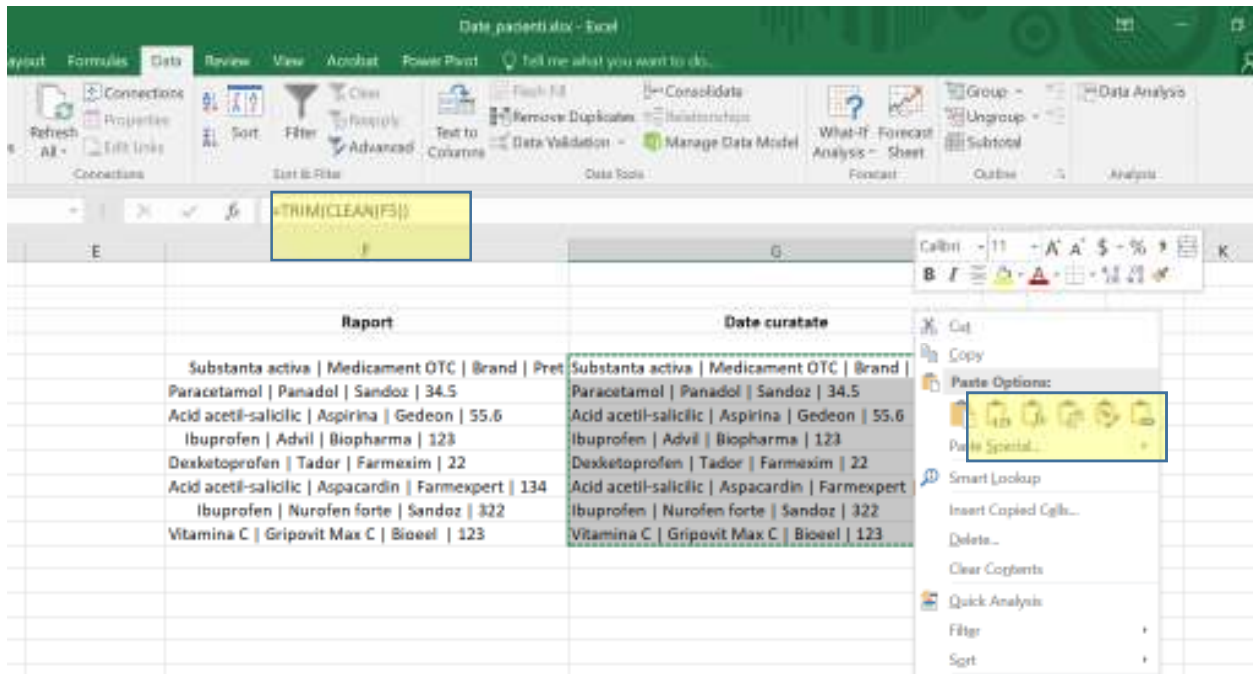
2. Then select the data again, and from the Data menu - Data Tools section, select Text to Columns Wizard. And in the first dialog box that appears, check Delimited and click Next.
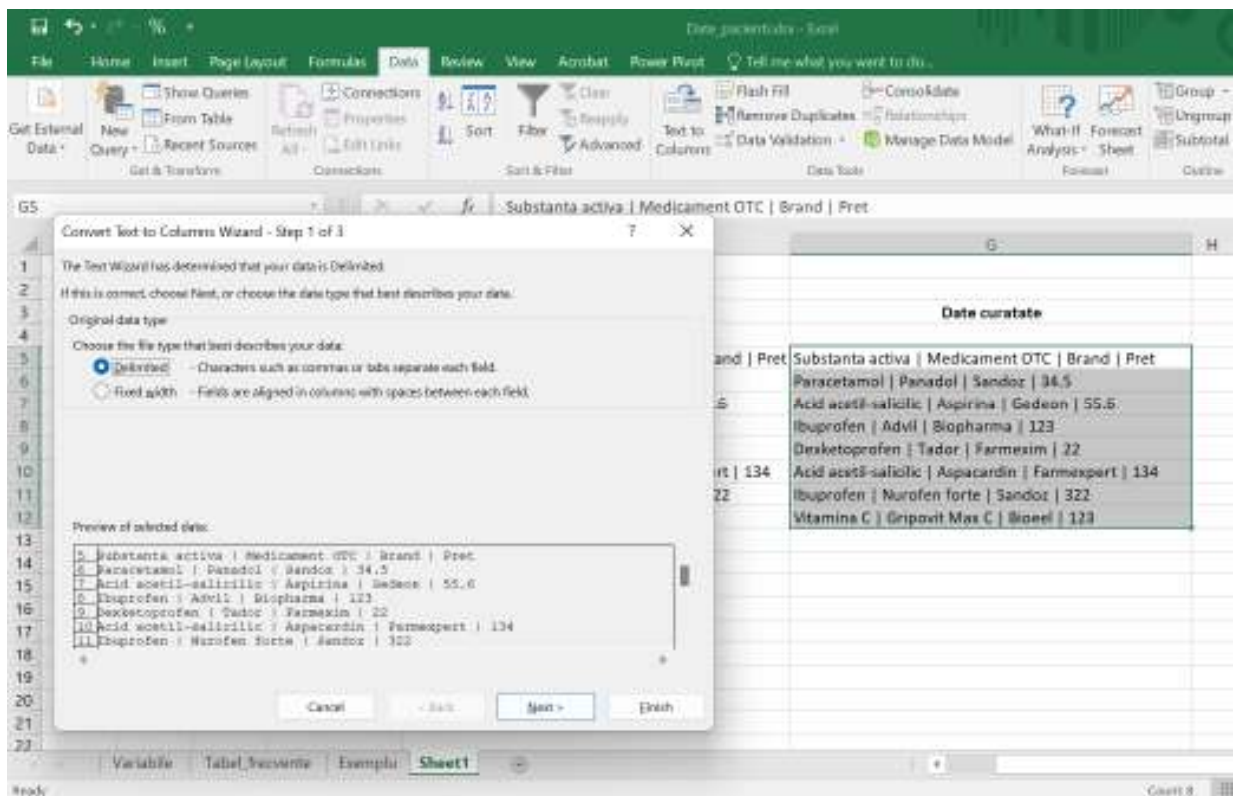
3. In the next step we select the delimiter contained in our data, in this case we will check Other, and we will manually enter the "|" character, after which we click on Next.
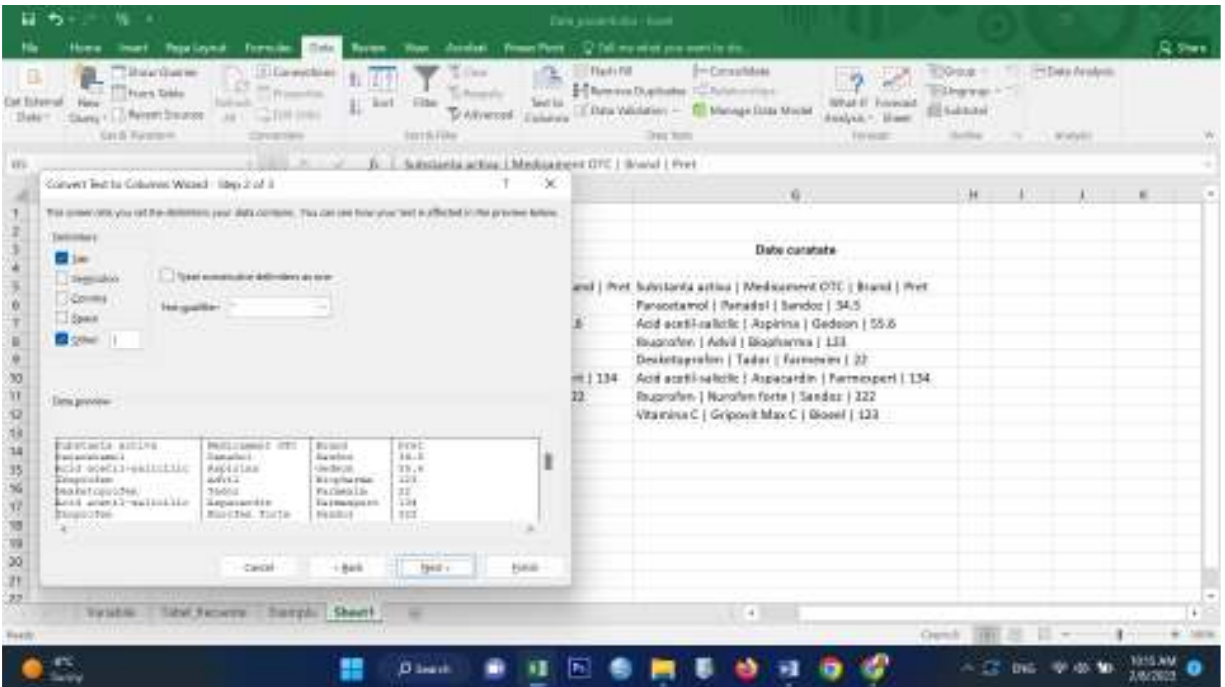
4. In the last step, we select the destination where we want to insert the processed data and we can view a preview of the data to be displayed after clicking on Finish.
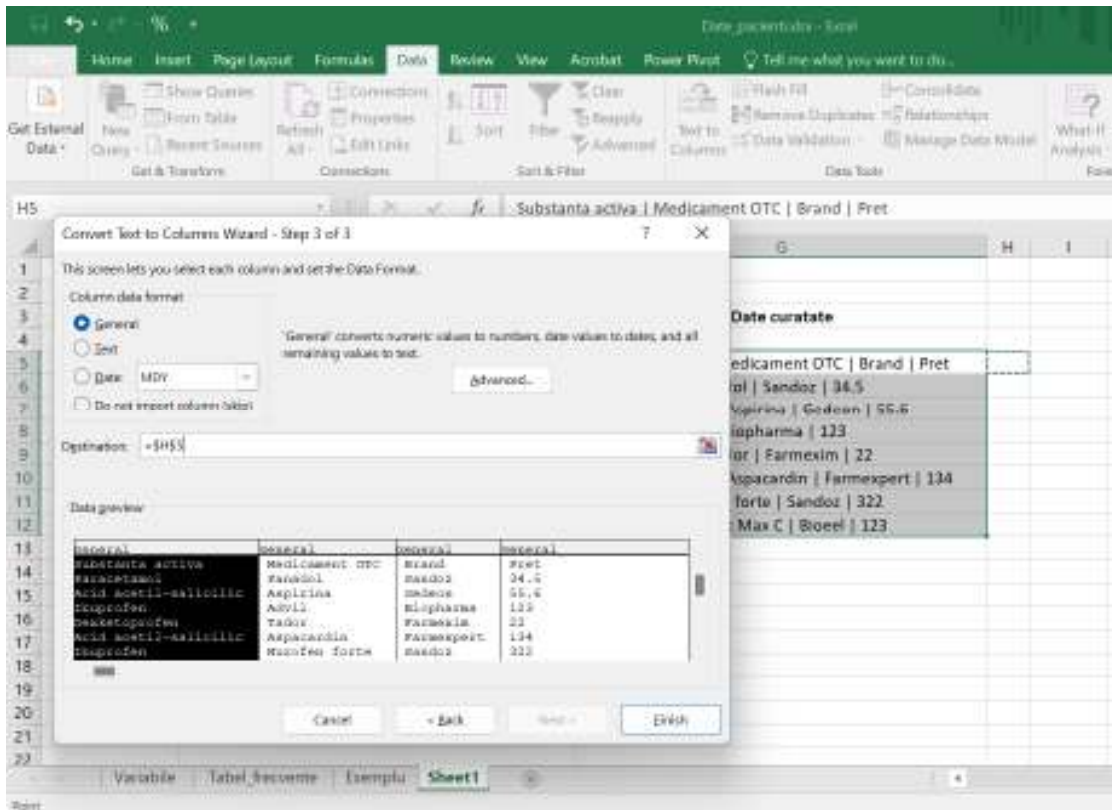
5. In the figure below we can see the result, the text from one column transposed into the 4 columns, and each of the 4 columns presents a header.

Raport	Date curatate	Substanta activa	Medicament OTC	Brand	Pret
Substanta activa | Medicament OTC | Brand | Pret	Substanta activa | Medicament OTC | Brand | Pret				
Paracetamol | Panadol | Sandoz | 34.5	Paracetamol | Panadol | Sandoz | 34.5	Paracetamol	Panadol	Sandoz	34.5
Acid acetil-salicilic | Aspirina | Gedeon | 55.6	Acid acetil-salicilic | Aspirina | Gedeon | 55.6	Acid acetil-salicilic	Aspirina	Gedeon	55.6
Ibuprofen | Advil | Biopharma | 123	Ibuprofen | Advil | Biopharma | 123	Ibuprofen	Advil	Biopharma	123
Dexketoprofen | Tador | Farmexim | 22	Dexketoprofen | Tador | Farmexim | 22	Dexketoprofen	Tador	Farmexim	22
Acid acetil-salicilic | Aspacardin | Farmexpert | 134	Acid acetil-salicilic | Aspacardin | Farmexpert | 134	Acid acetil-salicilic	Aspacardin	Farmexpert	134
Ibuprofen | Nurofen forte | Sandoz | 322	Ibuprofen | Nurofen forte | Sandoz | 322	Ibuprofen	Nurofen forte	Sandoz	322
Vitamina C | Gripovit Max C | Bioeel | 123	Vitamina C | Gripovit Max C | Bioeel | 123	Vitamina C	Gripovit Max C	Bioeel	123

8.3.2 Extracting data values with text functions

Let's take an example where the collected data is not character delimited nor aligned in columns with spaces between each field, in this case we can call text functions that extract the data values. These functions are:

• LEFT - Returns the leftmost characters in a text value

• RIGHT - Returns the rightmost characters in a text value

• MID - Returns a specified number of characters in a text string starting at the position you specify

• LEN - Returns the number of characters in a text string

Depending on the data held, the above functions can be combined to extract the required values.

### 8.3.3 Extracting data values with Flash Fill

Another technique for extracting data values from text is Flash Fill, but this technique is useful when Excel can identify a pattern in the data.
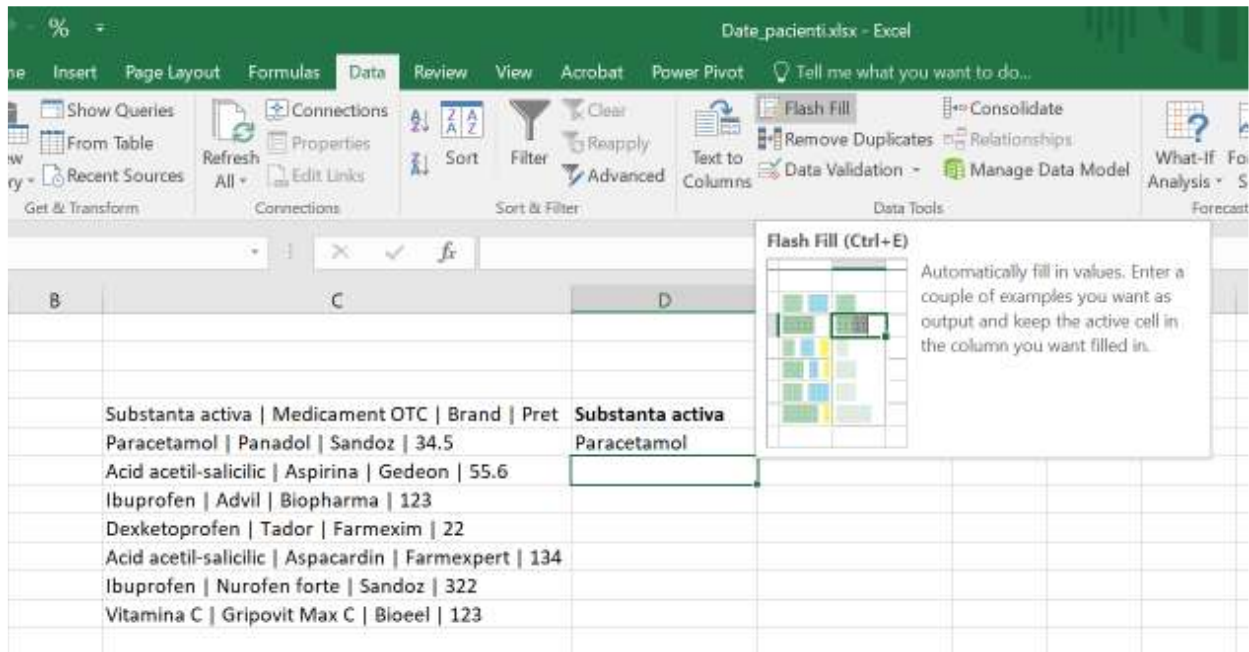
In the following we will create an example.

1. We will create 2 columns named Active Substance and OTC Drug.
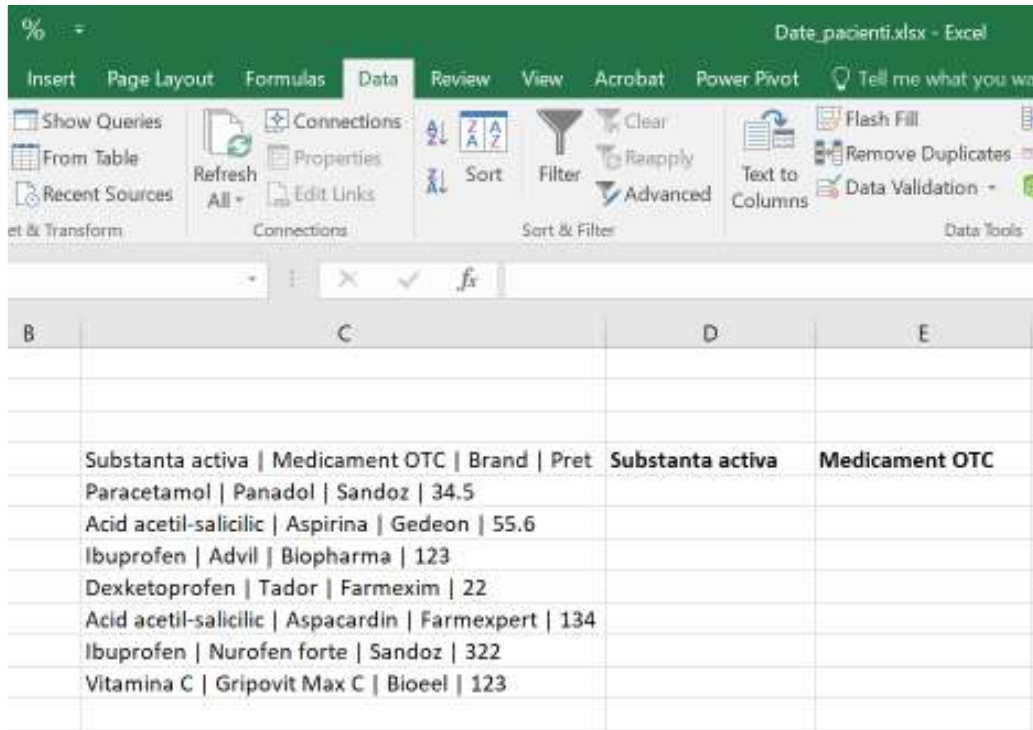
2. After that we will fill the first row from the created columns with the information from the raw data.
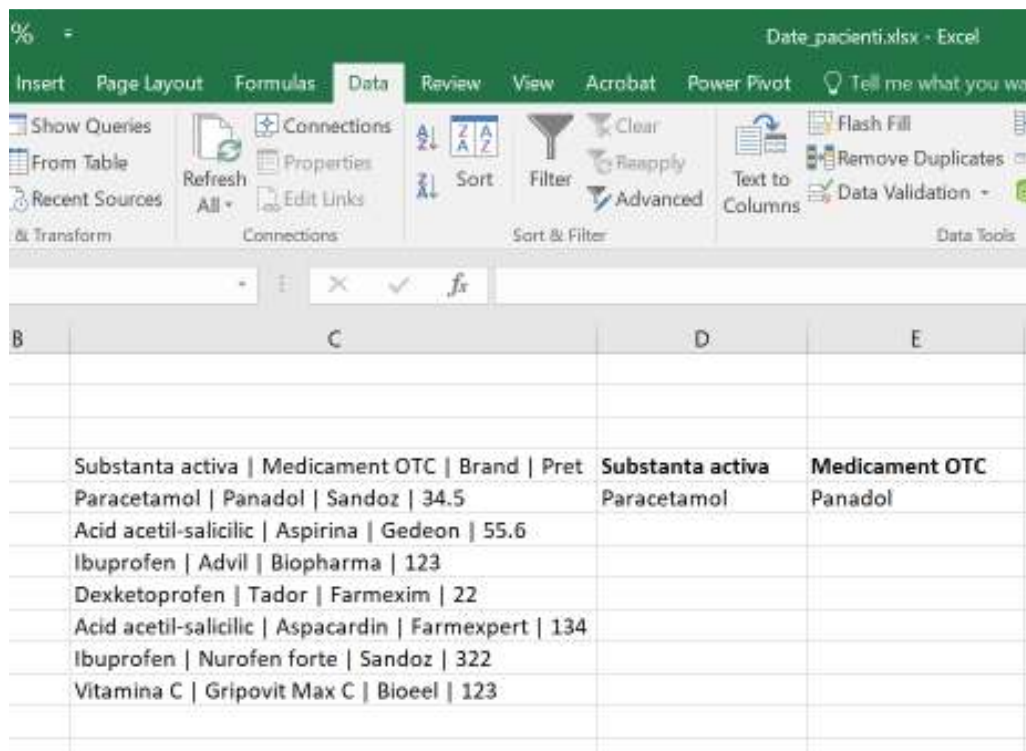
3. We will click in the second row, and from the Data – Data Tools menu, we will click on Flash Fill.

4. Using Flash Fill, the Active Substance column was automatically filled with the information from the Raw Data column. We repeated these steps for the OTC Drug column as well.
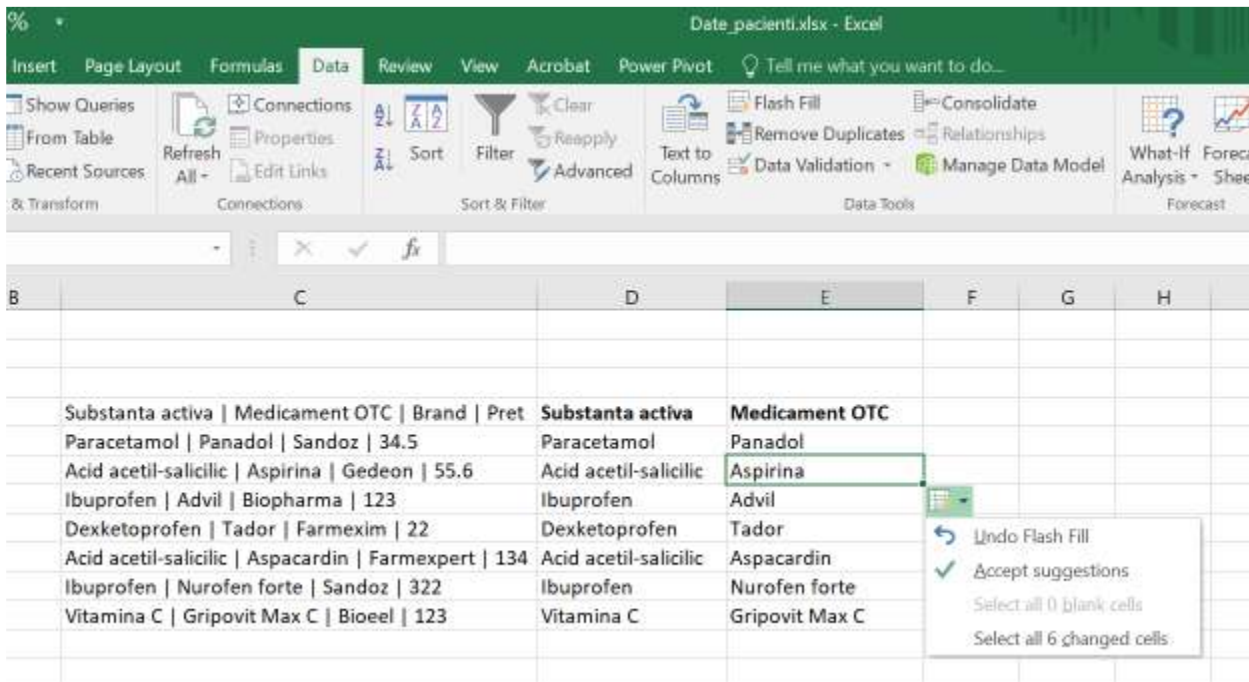
8.4 Formatting data with text functions

To format data containing text as needed, Microsoft Excel provides the following text functions:

- LOWER - Converts text to lowercase
- UPPER - Converts text to uppercase
- PROPER - Capitalizes the first letter of each word of a text value
- DOLLAR - Converts a number to text, using the $ (dollar) currency format.
- FIXED - Formats a number as text with a fixed number of decimal places
- TEXT - Formats a number and converts it to text
- VALUE - Converts a text argument to a number

Text functions are useful for text operations on data containing text:

- REPLACE - Replaces characters in the text
- SUBSTITUTE - Replaces new text with old text in a text string
- CONCATENATE - Joins multiple text elements into a single text element
- CONCAT - Combines text from multiple ranges and/or strings, but does not supply the delimiter or IgnoreEmpty arguments.
- TEXTJOIN - Combines text from multiple ranges and/or strings and includes a delimiter that you specify between each text value to be combined. If the delimiter is an empty text string, this function will effectively concatenate the ranges.
- REPT - Repeats the text a certain number of times

9. CLEANING DATA CONTAINING DATA VALUES

The diversity of raw data acquisition sources may result in different formats for data and values. In the following, we will present methods for processing data that contain date values, such as data conversion in serial format - month-day-year format and vice versa, extracting the year, month, day from the date, calculating the number of years, months and days between two dates and so on.

9.1 Data Formats
Microsoft Excel accepts data values in two types:

- Serial format
- In different Year-Month-Day formats
  We can transform:
- One date in serial format to one date in year-month-day format
- One date in year-month-day format to one date in serial format

Date in serial format

Date in serial format is represented by a positive integer representing the number of days between the Current Date and January 1, 2019. Both the Current Date and January 1, 2019 are included in the count. For example, 43341 is a date that represents 08/29/2018. Date in month-day-year formats

Excel supports different date formats depending on the locale (location) you choose. Therefore, we must first determine the compatibility of the data formats we have and the data analysis at hand. Certain date formats are prefixed with *(asterisk), this means that:

- Date formats starting with *(asterisk) respond to changes in regional date and time settings that are specified for the operating system.
- Data formats without *(asterisk) are not affected by operating system settings.

9.2 Conversion of data in serial format to Month-Day-Year format
The conversion of data from the Serial format to the Moon-Day-Year format is carried out in the following way:

We select the date (the number 43212), right-click on the number, click on Format Cells and then click on Number from the Format cells dialog box. In the Category we select Date, we choose a date type format and the location after which OK. In the Sample field we can see the result of the conversion of the serial date 43212 in the format Month-Day-Year, 04/22/2018.
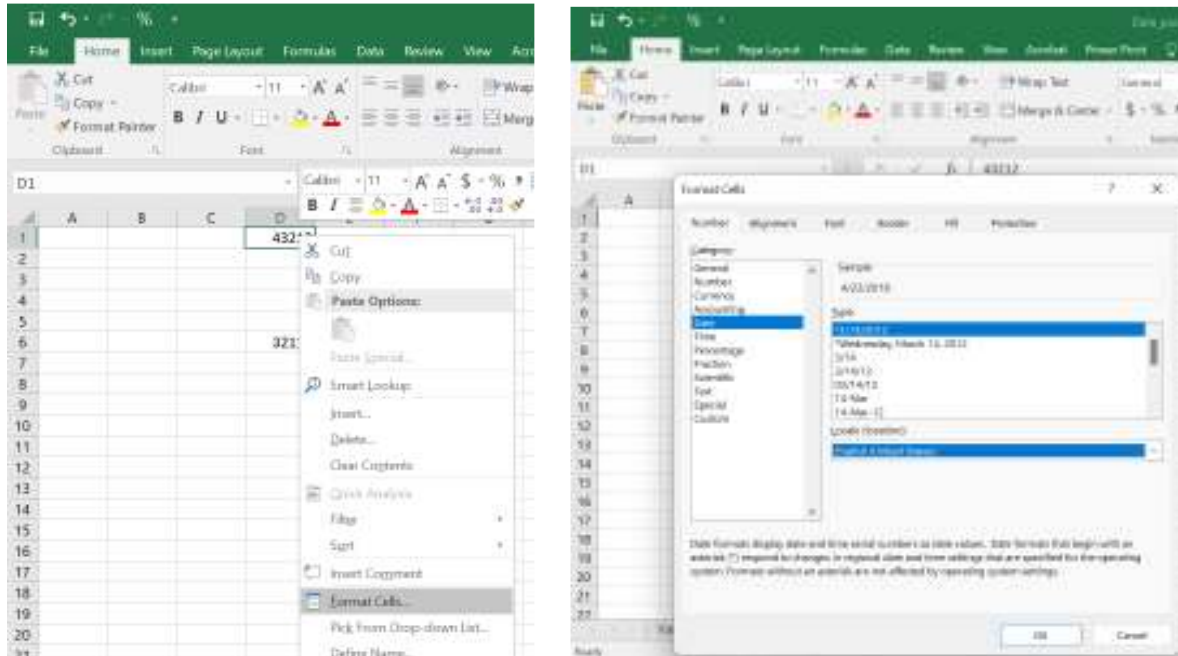
9.3 Convert month-day-year data to serial format

Data conversion from Month-Day-Year format to Serial format can be done in two ways:

- By using the Format Cells dialog box
- By using the Excel DATEVALUE function

➢ *Format Cells dialog box*

We will select the date 8/2/2023, right-click on the date, click on Format Cells, under Category we select General. In the Sample field we can see the result of the conversion of the date Day-Month-Year, 8/2/2023 in Serial 45140 format.
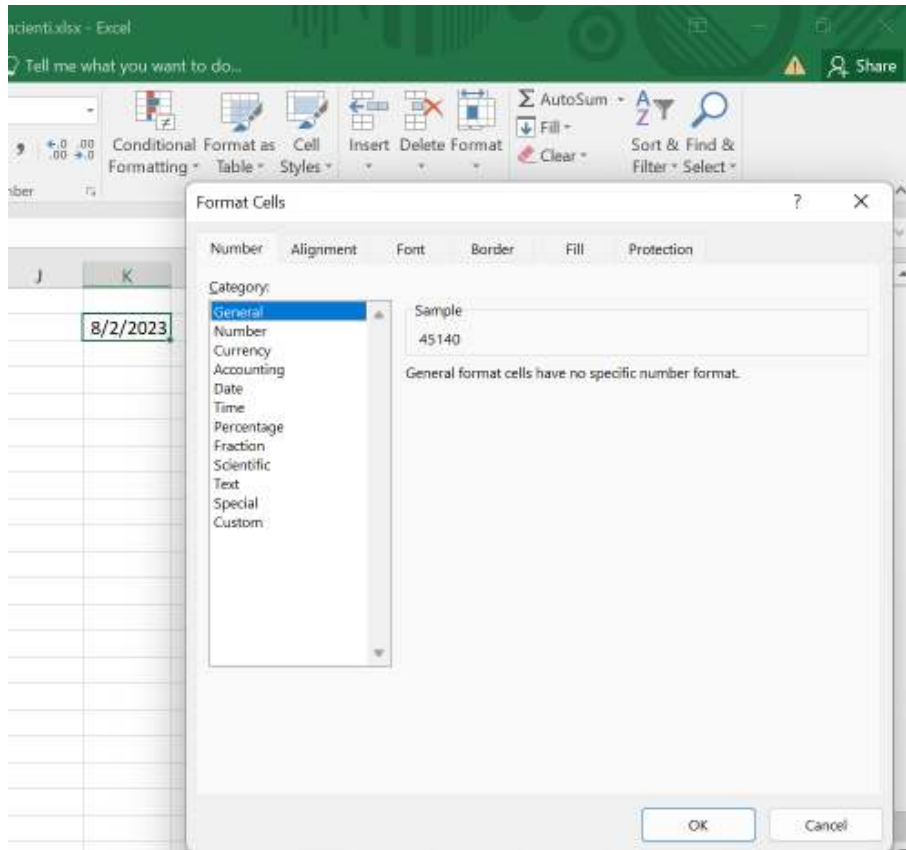
> *Using the Excel DATEVALUE function*

We use the Excel DATEVALUE function to convert a date format into a serial number. To remember! The date must be entered between the characters " " (quotation marks), =DATEVALUE ("8/2/2023") returns 45140.
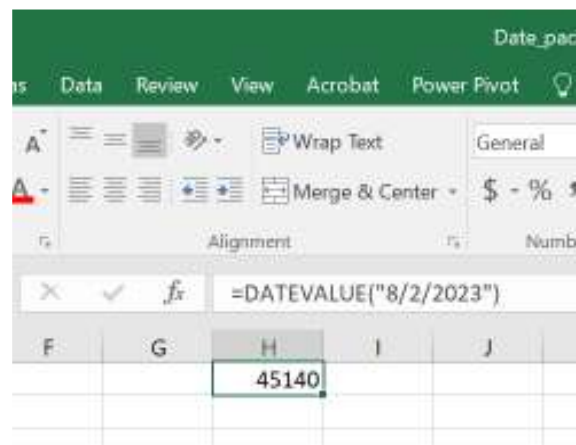
9.4 Calculation of the current date

We can calculate the current date using the TODAY function.
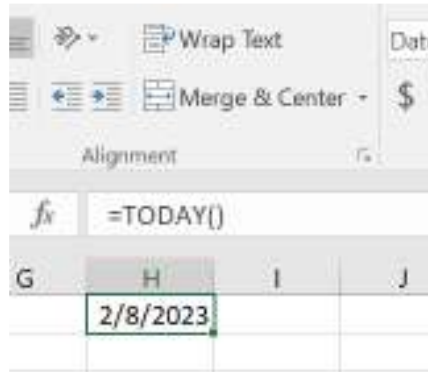
9.5 Identifying a business day by specified days

We can use Excel functions WORKDAY and WORKDAY.INTL to identify a workday after specified days.

- WORKDAY - Returns the serial number of the working days of the date before or after a certain one

- WORKDAY.INTL - Returns the serial number of the date before or after a specified number of workdays using parameters to indicate which and how many days are weekend days.

An example would be to calculate the next business day, 10 days from today.
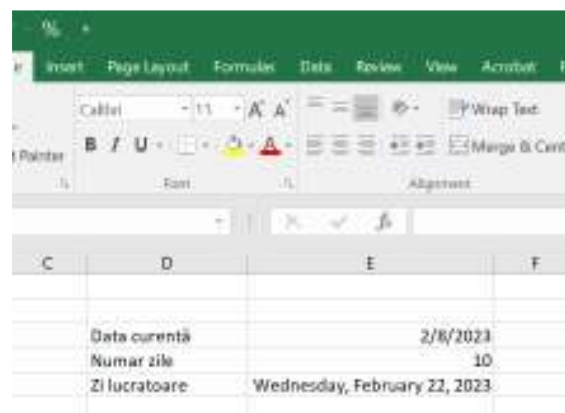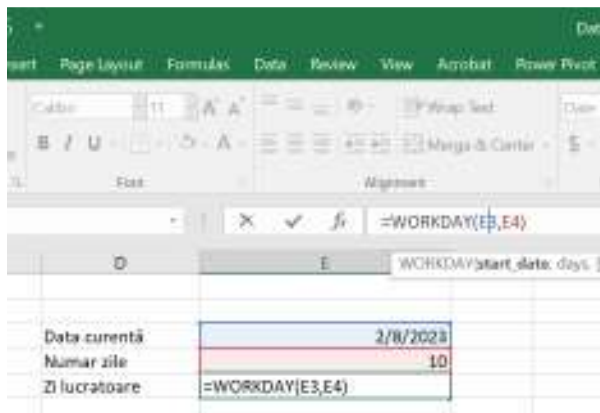
9.6 Extract Year, Month, Day from Date

We can extract from each date in a list of dates the corresponding day, month and year using the excel functions DAY, MONTH and YEAR.

We have the following data from the figure below, from each position we can extract the day, month and year.
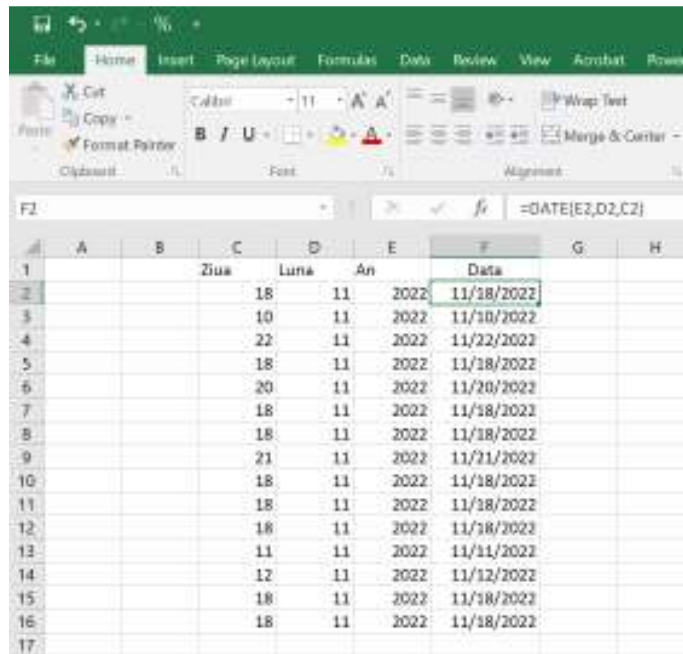
9.7 Extracting the day of the week from data

With the help of the WEEKDAY function, the corresponding day of the week can be extracted from each date in a list of dates.

9.8 Obtaining date from year, month and day

With the help of the DATE function we can get the date values, if the day, month and year are written in separate columns.

9.9 Calculating years, months and days between two dates

It may be necessary to calculate the time spent from one date to another. We may need this information in the form of years, months and days. As an example, we will calculate the current age of a person by the difference operation between the date of birth and today. We call the Excel functions DATEDIF, TODAY and CONCATENATE.

10. WORKING WITH TIME VALUES

Excel supports time values in two ways:

- Serial format
- In various hour-minute-second formats

We can convert:

- Time in serial format to time in hour-minute-second format
- Time in hour-minute-second format to time in serial format

10.1 Time in serial format

Time in serial format is a positive number that represents the Hour as a fraction of a 24-hour day, starting at midnight.

In Excel, we can combine Date and Time in the same cell. The serial number is the number of days after January 1, 1900 and the fraction of time associated with the given time.

10.2 Time in hour-minute-second format

Excel allows us to specify the time in hour-minute-second format with a colon (:) after the hour and another colon before the seconds. For example, 7:40 AM, 7:40 PM, or just 7:40 using 12-hour format, or as 7:40, 19:40 in 24-hour format. 7:40:40 AM is 7 hours, 40 minutes and 40 seconds.

We can also specify the date and time together. For example, if we type February 8, 2023 9:25 in a cell, it will be displayed as 2/8/2023 9:25 and represents 2/8/2023 9:25:00 AM.

Microsoft Excel supports different time formats depending on the locality (location) we choose. Therefore, you must first determine the compatibility of the time formats and data analysis at hand.

10.3 Convert time in serial format to hour-minute-second format

To convert serial time format to hour-min-sec format, we will make an example.

In a cell I typed the time 14:00, then right-click on the cell, choose Format Cells – Number – Category – Time, and from the drop-down list choose a format and a location for the time, which can be viewed in the Sample field, after which click OK.
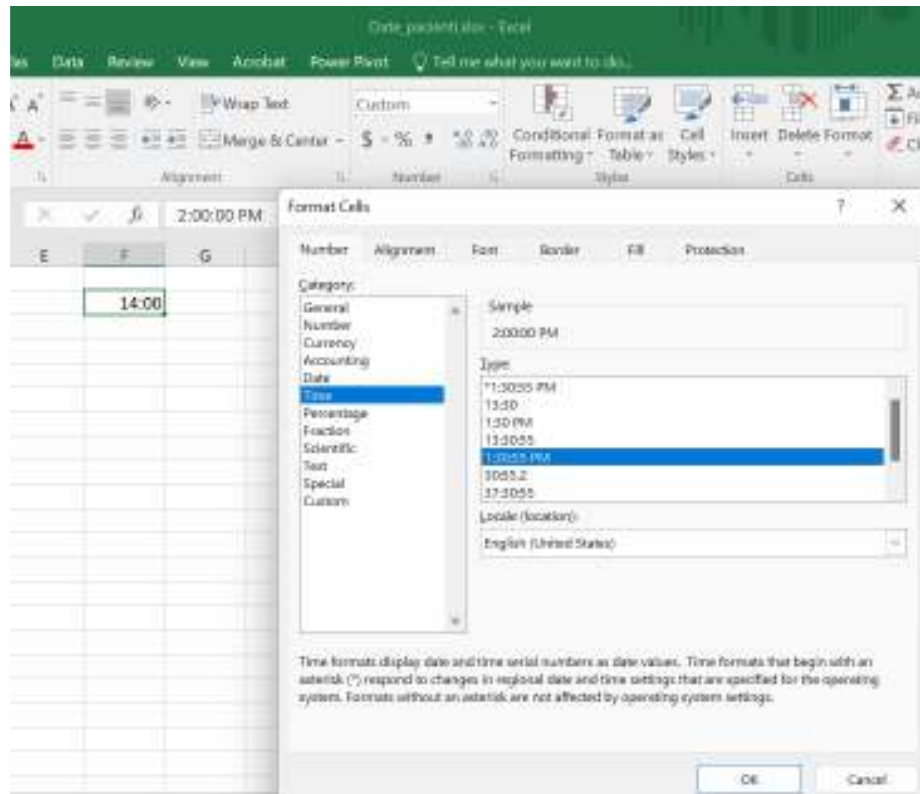

10.4 Convert times in hour-minute-second format to serial format

We can convert the time from hour-minute-second format to serial format in two ways:

- Using the Format Cells dialog box
- Using the Excel TIMEVALUE function

Conversion using the Format Cells dialog box

In a cell I typed the time 2:00:00 PM, then right-click on the cell, choose Format Cells – Number – Category – General, and the serial time that can be viewed in the Sample field, then click OK.
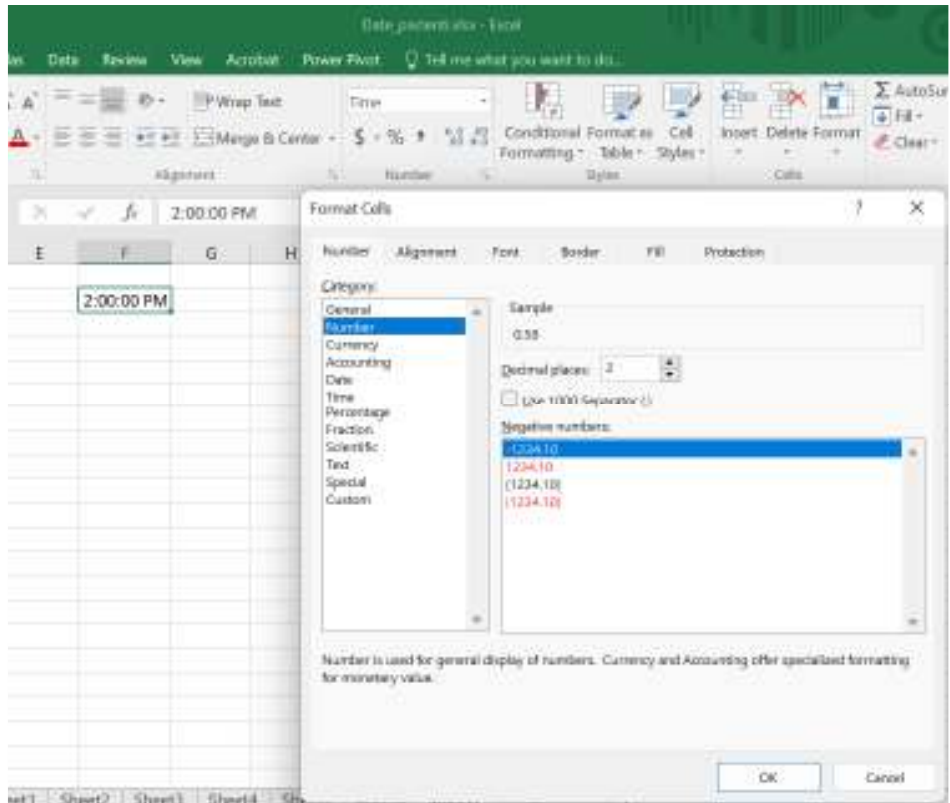
Conversion using the Excel TIMEVALUE function

An example is the following: TIMEVALUE ("14:35:15") results in 0.6078125, remember to insert the characters " " (quotes).
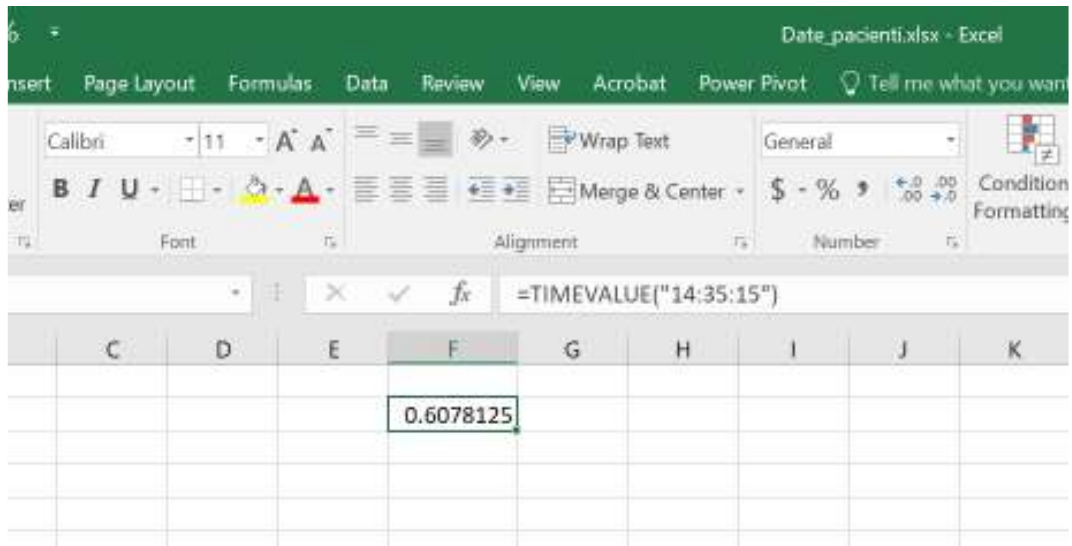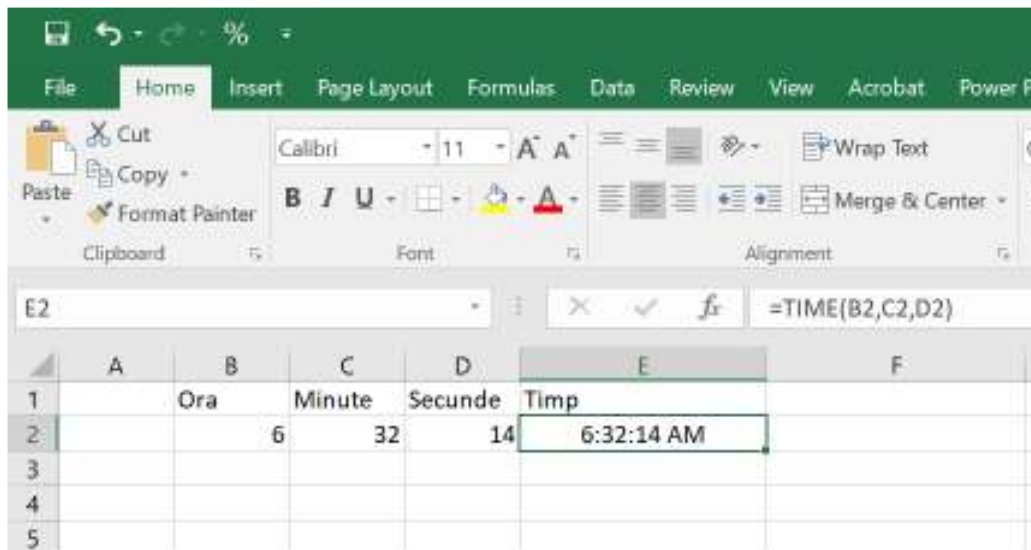
Obtaining the current time

To calculate the current time we use the Excel function NOW(). The result reflects the date and time the function is used.
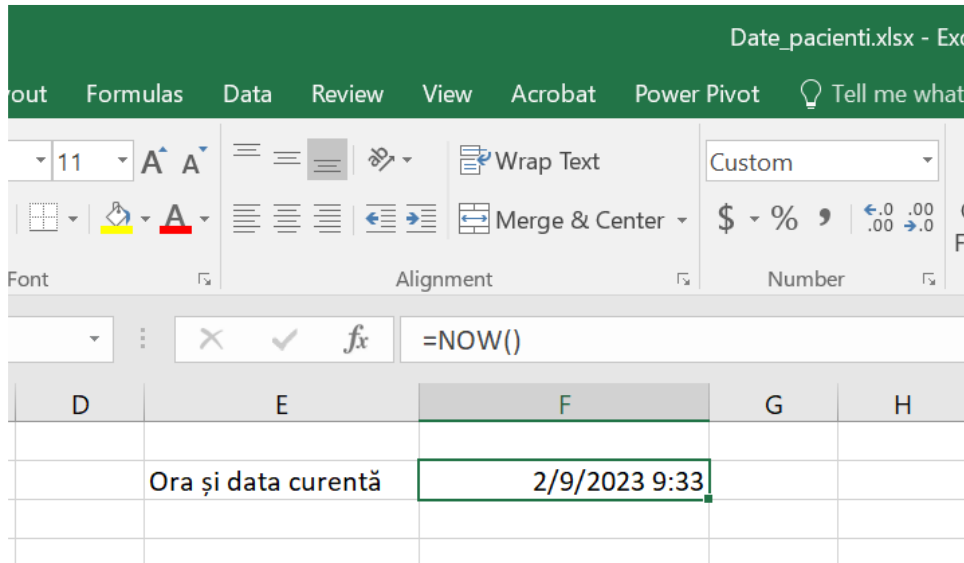
10.5 Obtaining the time from hour, minute and second

If our time data is imported in separate columns, we can join this information to find the time using the TIME function.

10.6 Extracting the hour, minutes and seconds from time

Depending on the analysis we are performing, we may sometimes need information about hours, minutes and seconds separately. We can use the Excel HOUR, MINUTE, SECOND functions to get the time values in separate columns.

11. DATA SORTING

An important step in the process of data analysis and processing is the sorting operation. By sorting we can view a list in alphabetical order, or from the largest to the smallest and vice versa, in ascending or descending order, or we can perform a personalized sorting according to needs. Sorting can be done by columns or by rows.

Criteria for sorting data in one or more columns according to:

- text (from A to Z or from Z to A)
- numbers (from smallest to largest, ascending or from largest to smallest, descending)
- dates and times (oldest to newest and newest to oldest)
- custom list (eg large, medium and small)
- format, including cell color, font color, or icon set.

11.1 Sorting by numbers

We can sort a table using a column that contains text. In the table in the image below, information about patients admitted to a private clinic is collected.

As an example, we want to sort the patient_ID column, to display the data in descending order (low to high). To perform the sorting, we click in the column of interest,

and from the HOME menu, Editing section, we will select Sort&Filter, and check sort from Z to A (from high to low).

5	52	0	0	0	1	0	0	1	91	85	155	1.55	37.88	0.06
6	54	1	0	0	1	1	0	0	91.4	87	169	1.69	32.00	0.08
7	60	1	0	1	0	0	1	1	99.3	99	180	1.8	30.65	0.11
8	63	1	1	0	0	0	1	1	98.1	96.5	177	1.77	31.31	0.10
9	55	1	1	1	1	1	1	1	102.6	100.5	174	1.74	33.89	0.09
10	61	1	0	0	1	0	0	1	121.7	119	175	1.75	39.74	0.08
11	61	0	0	0	1	0	0	1	93.5	91.2	162	1.62	35.63	0.07
12	69	1	1	0	0	1	0	1	100	96.4	170	1.7	34.60	0.08
13	70	1	1	0	1	1	1	0	9.7	90.7	167	1.67	3.48	7.50
14	55	0	1	0	1	1	1	1	116.4	110	159	1.59	46.04	0.05
15	76	0	1	1	1	0	1	0	111.6	99	155	1.55	46.45	0.05
16	56	1	0	1	1	0	0	1	126	110	174	1.74	41.62	0.06
17	70	0	1	0	0	0	0	0	87	83.8	154	1.54	36.68	0.06
18	63	0	0	0	1	0	0	1	72	70.9	146	1.46	33.78	0.06
19	63	1	0	1	1	1	0	1	131.6	125	185	1.85	38.45	0.08
20	70	0	1	1	1	1	1	0	86	84.5	156	1.56	35.34	0.07

The result can be viewed in the figure below, where we can see the Patient ID column has the values sorted in descending order. The same can be done for a text list.

11.2 Sorting by text

We want to sort the gender column so that gender M appears first in the list. To perform the sorting, we click in the column of interest, and from the HOME menu, Editing section, we will select Sort&Filter, and check sort from Z to A.
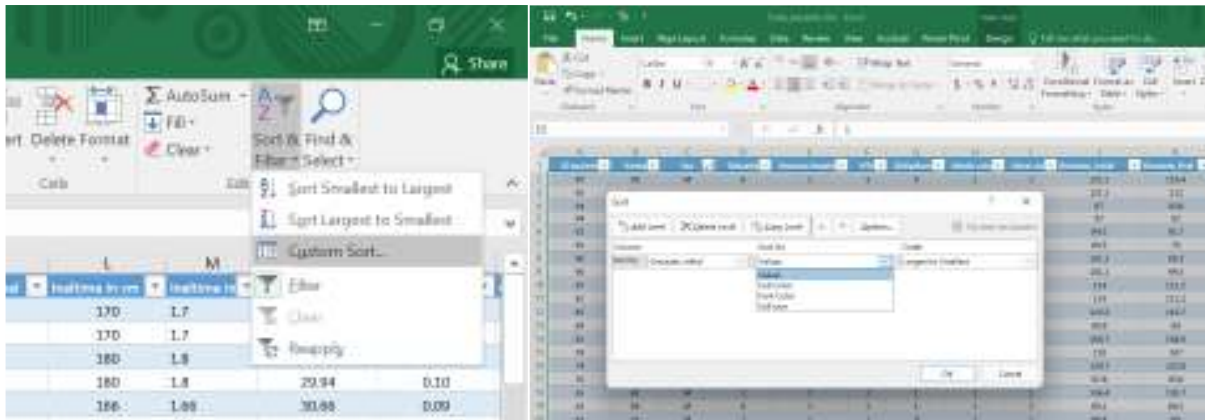

The result can be viewed in the figure below, where we can see the Sex column has the values ordered alphabetically descending, from M to F.
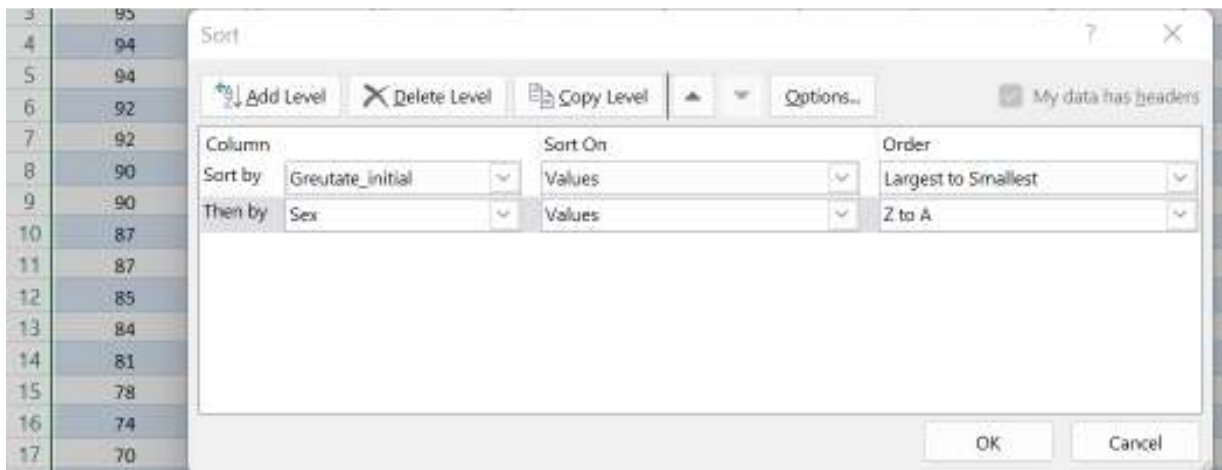
11.3 Custom sort

To create a custom sort, access the HOME menu, Editing section, select Sort&Filter, and check Custom Sort.

We want to do a double sort, in the sense that we want the male patients with the highest weight to appear first, and then the female patients to appear after that.

After completing the above steps, the Sort dialog box opens, and at Sort By, we select from the drop-down list the columns we want to sort, if we have more columns we check Add level, as many times as we want to add a new column when sorting, at Sort on, we tick depending on what exactly to sort and at Order, we select the order in which the data appear.

And thus we have achieved the custom sorting, in the figure below.

ID pacient	Varsta	Sex	Educatie	Boactopa hepatu	HTA	Dislipidem	Mediu urban	Stare civil	Greutate_initial	Greutate_final	Inaltime in cm
31	52	F	0	0	1	0	0	1	145.5	140.7	160
89	45	F	1	1	1	0	1	1	135	126.4	171
89	45	F	1	1	1	0	1	1	135	126.4	171
19	65	M	0	1	1	1	0	1	131.6	125	185
79	61	F	0	1	1	1	1	1	130	125	157
41	48	M	0	1	0	0	1	1	127	123.4	180
16	56	M	0	1	1	0	0	1	126	110	174
45	61	F	0	0	1	1	1	0	124.8	123	163
74	70	M	0	1	1	0	1	1	124.7	120.8	178
21	63	M	0	0	1	0	1	1	124.4	121.9	175
85	63	M	0	1	0	1	0	1	124.3	119.7	175
10	61	M	0	0	1	0	0	1	121.7	119	175
95	59	M	0	1	1	0	1	1	121.1	116.4	170
95	59	M	0	1	1	0	1	1	121.1	112	170
34	40	F	0	1	0	1	1	0	120.6	117.2	164
69	52	F	0	1	1	1	1	1	120	114.9	194
32	71	M	0	1	1	1	0	1	119.6	116.3	165
14	55	F	1	0	1	1	1	1	116.4	110	159
64	67	F	1	0	1	1	1	0	116.1	114.9	173
87	57	M	1	1	1	1	1	1	114	111.2	180
87	57	M	1	1	1	1	1	1	114	111.2	180

12. DATA FILTERING

To extract data that meets certain defined criteria from a range or table, we use the Filtering operation. Filtering is a quick and easy way to display only the information we need from a large database.

The filtering method can be done by:

• Selected values

• Text filters if the selected column contains text

• Data filters if the selected column contains data

• Number filters if the selected column contains numbers

• Cell color if the selected column contains cells with color

• Font color if the selected column contains colored font

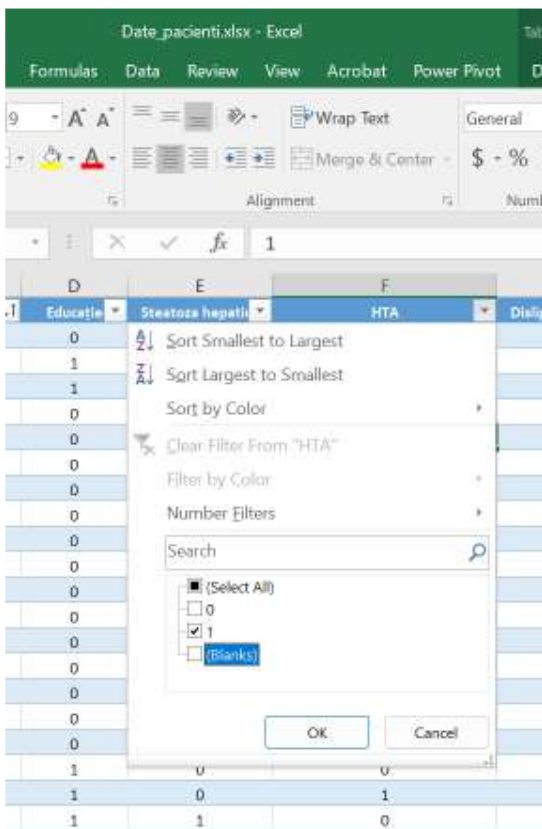• Cell icon if the selected column contains cell icons

• Advanced filter

If our data is arranged as a table in Excel, the column headers of the table automatically show the filter tag, called AutoFilter. The auto filter is represented by the arrow next to the column header. Each automatic filter has filtering options based on the type of data we have in that column.

If our data is organized as a range, we can label each column with a filter by going to the menu Home – Editing – Sort&Filter – check Filter, this will enable Auto Filter for range. Also in this mode the Filter option can be disabled.

12.1 Filtering by selected values

In the following example, we want to make a filter in which only patients presenting HTA=1 are displayed. We click on the Filter arrow next to the header of the HTA column, the dialog box that presents the filtering options will open, and we uncheck 0, Blanks, after which we click on OK. After the filtering has been done, you can see that a new symbol appears next to the Filter arrow, a funnel, this symbol shows us that there is an active filter on that column.
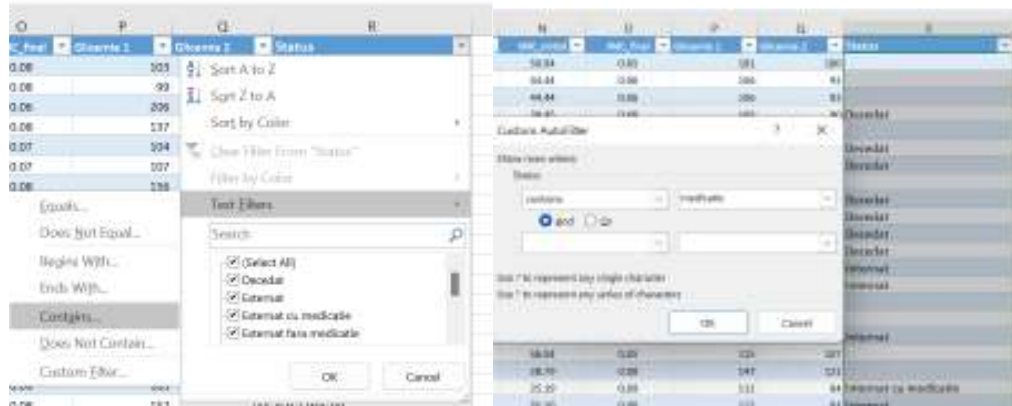
12.2 Filtering by text

We want to filter the database so that the Gender column shows only male patients and non-responders, so we exclude female patients.

We click on the Filter arrow next to the header of the Sex column, the dialog box that presents the filter options will open, and we uncheck Select All, after which we check only the options of interest, Male and No answer.
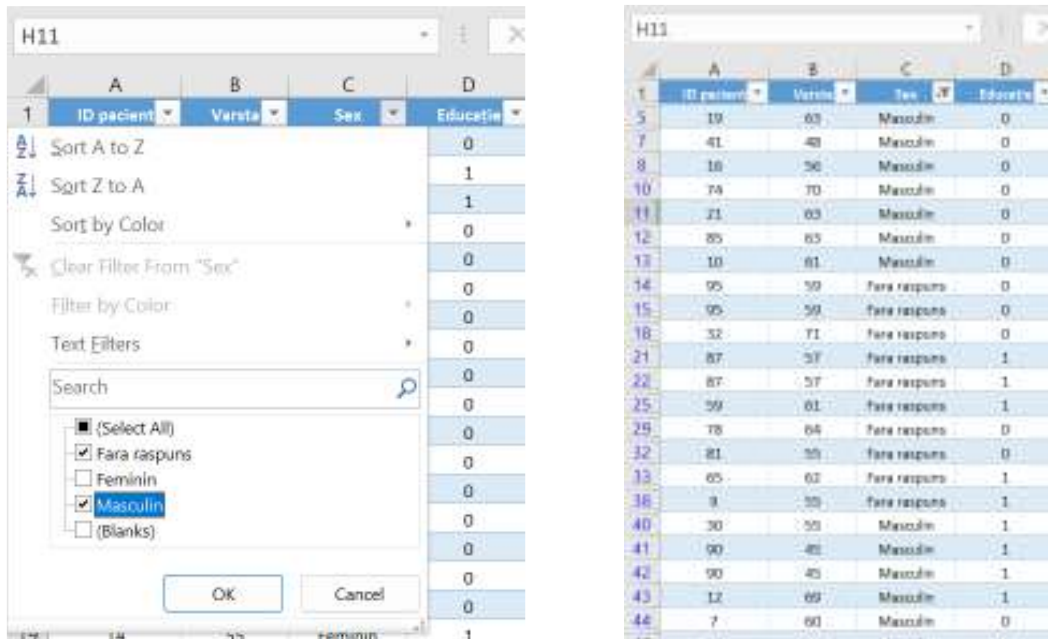

Also Filter by text contains many other options, another example is as follows. In the Status column, we want to display only patients who have received or are receiving medication, regardless of their status. For this, we will select Text Filters – Contains, and in the Custom AutoFilter dialog box, in the Status column, we will type medication, click on OK. Only rows where the Title contains Medication will be displayed.
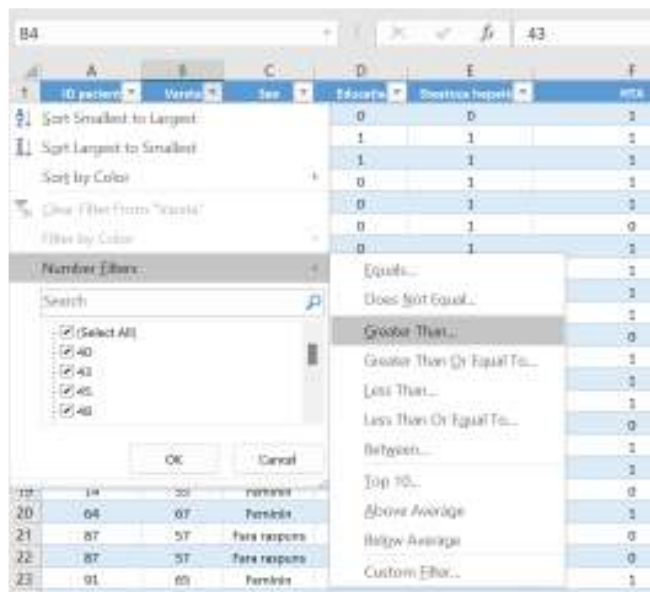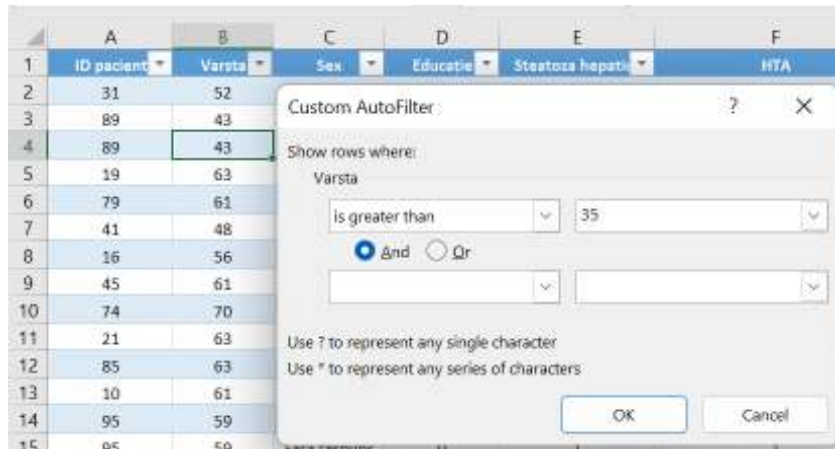
12.3 Filtering by date

We want to perform a filtering by the date of admission of the patients, namely to display all patients admitted before the year 11/18/2022. The steps and the result can be seen in the images below.
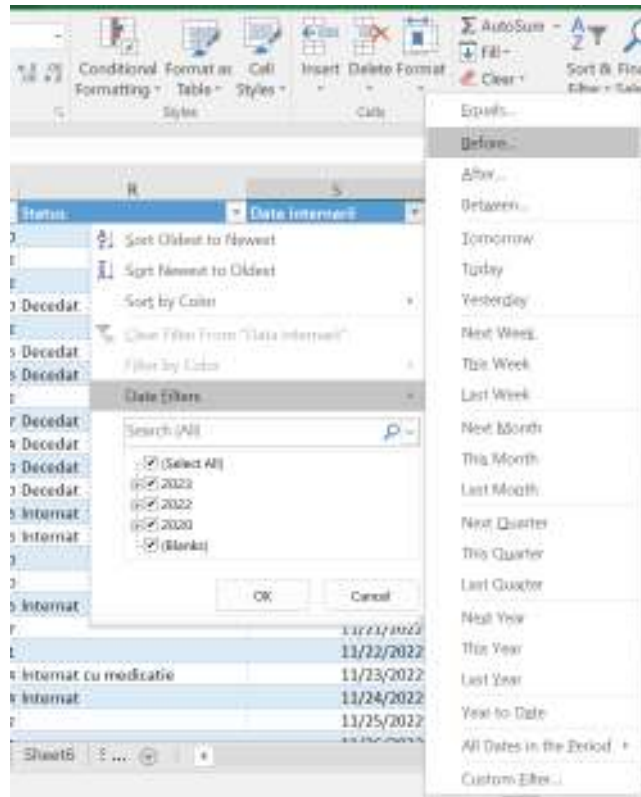
10.3 Filtering by numbers

We only want to view patients over the age of 35. To perform this filtering we click on the arrow next to the header of the Age column. From the drop-down list, click Number Filters. The Number Filters options appear. Click Greater than. The Auto Custom Filter dialog box opens. We type 35 in the box next to Greater than.

The result can be seen in the figure above.

12.4 Clearing the filter

Removing filters is called Clear Filter in Excel. We can delete filter applied from a single column or all filters applied in the spreadsheet simultaneously.

To remove a filter from a particular column, we click on the arrow in the table header of that column. From the drop-down menu, click Clear Filter from "Age".

The filter in the Age column is removed. And to remove filtering from the entire worksheet, select Clear in the Sort&Filter editing group on the Home menu. All filters in the worksheet are removed at once.