

Business Research and Data Analytics

Lecture 3: Data Preprocessing and Transformation with MS Excel

Igor Vyshnevskyi, Ph.D.

Woosong University

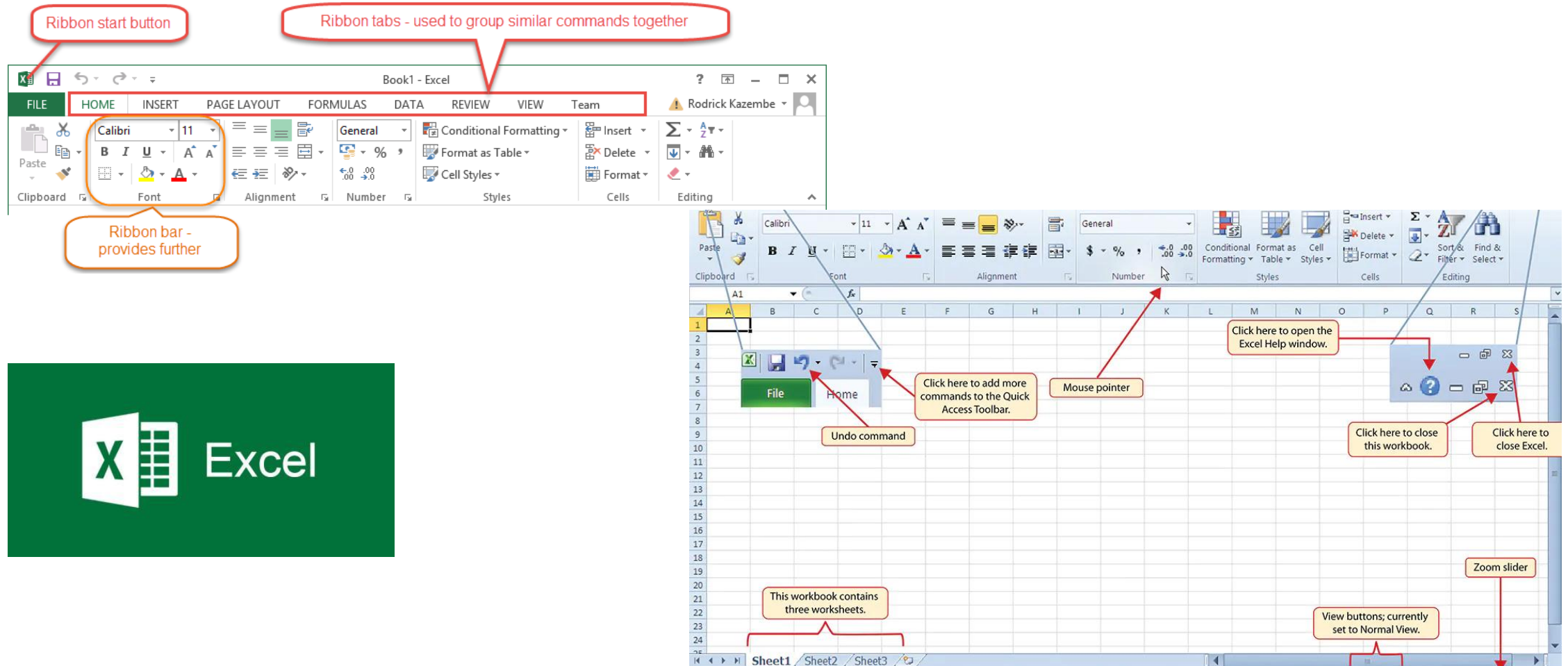
March 20, 2024

Agenda

1. Introduction to Data Preprocessing and Transformation
2. Processing and Transformation Techniques in MS Excel
3. In-class Assignment

MS Excel intro

No much details because you mostly know Excel (based on your background survey).



Source: Web pictures

1. Introduction to Data Preprocessing and Transformation

Data Processing is the overall process of manipulating, organizing, and structuring raw data into a more usable form.

It typically involves working with the ***dirty data*** such as cleaning data, removing duplicates, and formatting data for analysis.

Data Transformation is the process of converting data from one format or structure to another.

This may involve tasks such as splitting columns, merging data sets, or aggregating data.

Types of Dirty Data



Duplicate data



Outdated data



Incomplete data



Incorrect/inaccurate data



Inconsistent data

Types of Dirty Data

- To deal with dirty data it is better to develop your own check list which you can refer to and improve in the future.

Data Cleaning Checklist	Preferred cleaning methods

2. Processing and Transformation Techniques in MS Excel

File to be used: *International-Logistics-Association-Memberships.csv*

Conditional formatting

- **Identify missing values** by conditional formatting

Let's apply conditional formatting to all columns in the table except for "Address 3", "Address 5" and "Certification" columns.

General instructions:

1. Select the cells you want to apply conditional formatting to.
2. Go to the Home tab on the ribbon.
3. Click on the Conditional Formatting option.
4. Choose the type of formatting you want to apply (e.g. highlight cells rules, top/bottom rules, data bars, color scales, icon sets, etc.).
5. Choose the formatting options you want to apply (e.g. select the colors, the minimum/maximum values, the criteria for highlighting, etc.).
6. Click OK to apply the formatting.

Excel ribbon: HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, VIEW, DEVELOPER, DESIGN

Font: Arial, 10, Bold, Italic, Underline, Color (A), Background Color (Yellow)

Alignment: General, Wrap Text, Merge & Center

Number: \$, %, .00, .00

Conditional Formatting: Highlight Cells Rules, Top/Bottom Rules, Data Bars, Color Scales, Icon Sets, New Rule..., Clear Rules, Manage Rules...

Conditional Formatting: Highlight Cells Rules

- Greater Than...
- Less Than...
- Between...
- Equal To...**
- Text that Contains...
- A Date Occurring...
- Duplicate Values...
- More Rules...

A	B	C	D	E
er ID	Last name	First name	Address 1	Address 2
1	Tsao	Danny	27 Wu Tzu St	Tamshui 251
2	Lei	Colleen	88 6th Avenue Teda	300457 TIANJIN
3	Roth	Nancy	Hoefenstrasse 31	Muehlethal
4	Meneses Contreras	Karl-Oscar	Poniente 134 Ste. 740	02300 México
5	Nunez	Helmut	Andador Pinos 345	45235 Zapopan
6	Fitzpatrick	Dmitry	22 Hemingford Pl	Whitby
7	Andreu	Leya	Nevada de Colima 104	20280 Aguascalientes
8	Ramsey	Stephen	Z-Block No 59	Chennai TN - 600040
9	Xiao-Hui	Michael	Unit B-E F19 XinMei Union Sq	200120 Shanghai
0	He	Jan	5055 Heather Leigh Avenue	Mississauga
1	Wisner	Ray	Chemin 15F	Vernier
2	Denturck	Bill	Hermeslaan 7	1831 DIEGEM
3	Arnout	Marco	Septestraat 27	2640 MORTSEL

Equal To

Format cells that are EQUAL TO:

with Light Red Fill with Dark Red Text

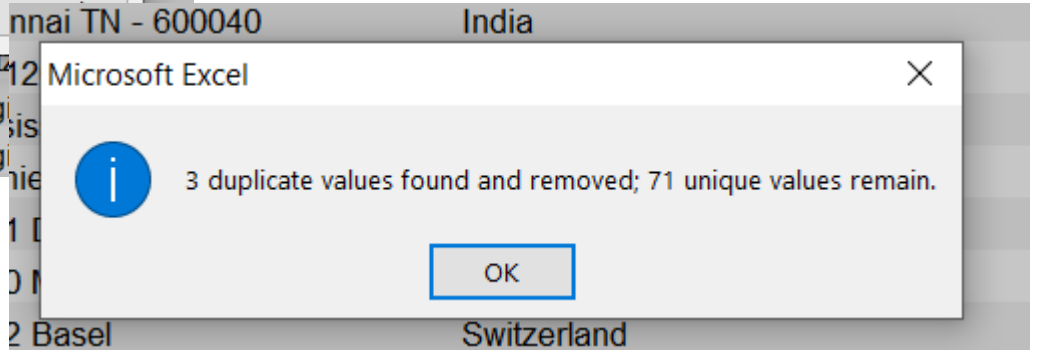
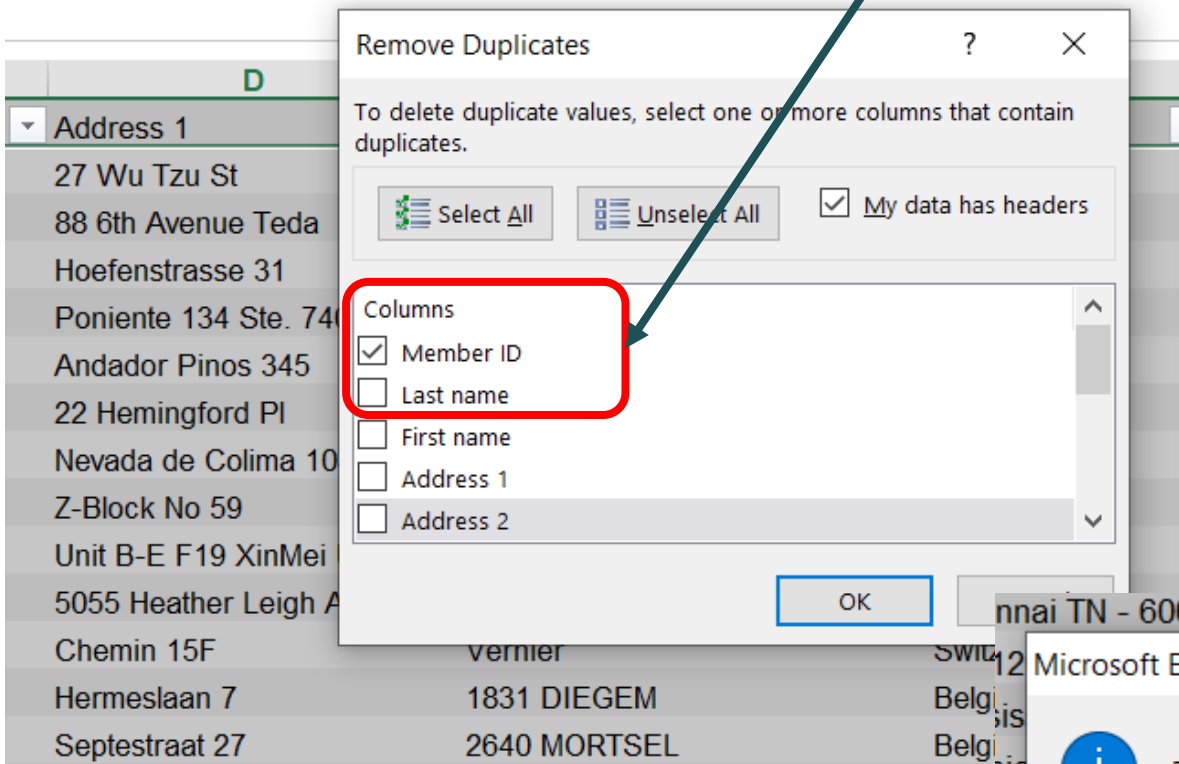
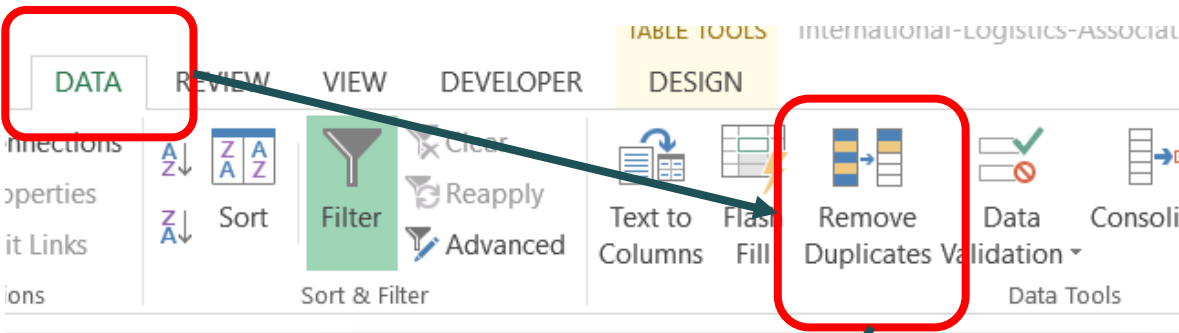
OK Cancel

Remove duplicates

Let's remove duplicates basing in Member ID

General instructions:

1. Click on the "Data" tab in the ribbon at the top of the screen.
2. Click on the "Remove Duplicates" button in the "Data Tools" section of the ribbon.
3. In the "Remove Duplicates" dialog box, select the columns that you want to check for duplicates.
4. Click the "OK" button.

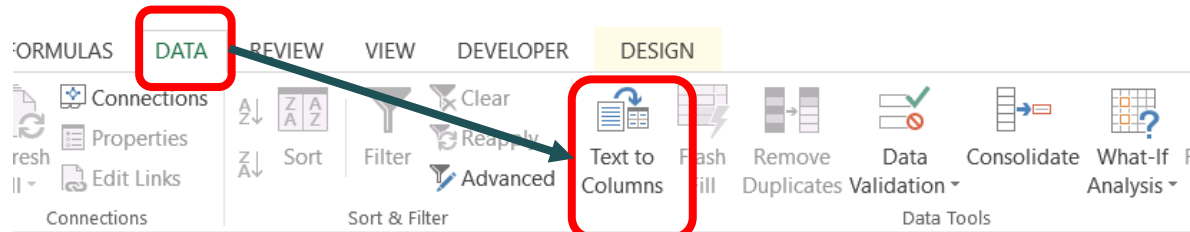


Splitting the data

Let's split certification info into different columns

General instructions:

- Select the cell(s) containing the data you want to split.
- Go to the Data tab on the ribbon.
- Click on the "Text to Columns" button.
- In the "Convert Text to Columns Wizard" dialog box that appears, choose the type of data you want to split (e.g. delimited, fixed width, etc.) and click "Next".
- Depending on the type of data you selected, you may need to choose additional options (e.g. specify the delimiter character, set the column widths, etc.). Follow the on-screen instructions and click "Next" to proceed.
- Choose the format for each of the columns you want to create (e.g. general, text, date, etc.) and click "Finish".
- Excel will split the data in the original cell(s) into different cells based on the criteria you specified.



ication

I	J	K	L
unt	Member type	Certification	
	Professional Member	CLTD	
	Concrete Member		

Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.
If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

☒ Delimited - Characters such as commas or tabs separate each field.

☐ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

1	Certification
2	CLTD
3	
4	CSCP, CLTD
5	
6	

Cancel < Back Next > Finish

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☒ Tab

☐ Semicolon

☒ Comma

☒ Space

☐ Other:

☒ Treat consecutive delimiters as one

Text qualifier: " " ' ' >

Data preview

Certification	
CLTD	
CSCP	CLTD

Cancel < Back Next > Finish

Filtering data

Let's check for inconsistent data

General instructions:

1. Select all range of cells that you want to filter.
2. Go to the "Data" tab in the ribbon menu at the top of the screen.
3. Click on the "Filter" button in the "Sort & Filter" group. This will add filter dropdowns to the header row of your data.
4. Click on the dropdown arrow in the header row of the column you want to filter. This will open the filter menu.

nt

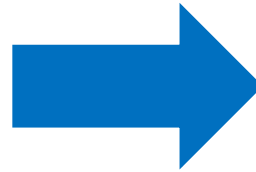
H	I	Mem
Address 5	Dues amount	Profe
4812		Corpc
		Stude
L1R 1G1		Corpc
		Profe
		Stude
		Profe
L5V 2R6		Corpc
1214		Profe
		Stude
		Stude
		Stude
V3J 1P1		Profe
31002		Profe
N3S 7P5	\$200	Corpc
		Profe

Let's say that we know that the dues range should be within \$100-500.

The filter menu shows us two inconsistent data. Let's choose them to examine in more details.

Fix inaccurate data.

H		
	<input type="text" value="Dues amount"/>	<input type="text" value="Member type"/>
	\$1,000	Student
	-\$200	Profess



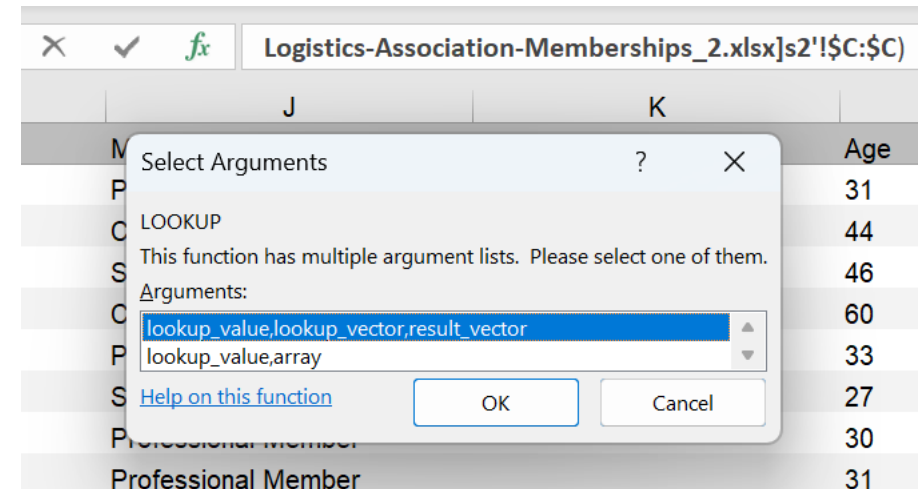
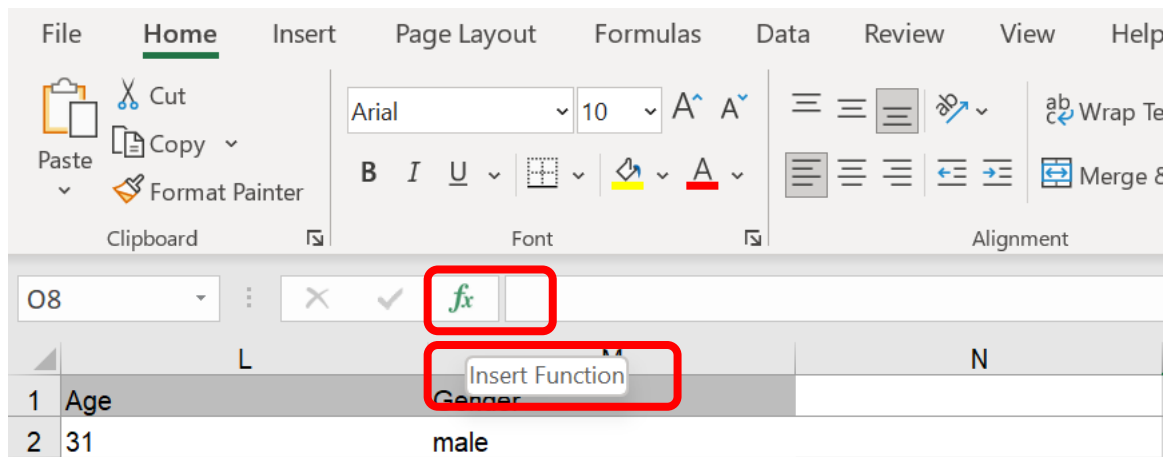
	<input type="text" value="Dues amount"/>	<input type="text" value="Member type"/>
	\$100	Student Ass
	\$200	Professiona

Functions in Excel

A **function** in Excel is a preset formula, that helps perform mathematical, statistical and logical operations.

Once you are familiar with the function you want to use, all you have to do is enter an equal sign (=) in the cell, followed by the name of the function and the cell range it applies to.

To check the function's description and arguments:

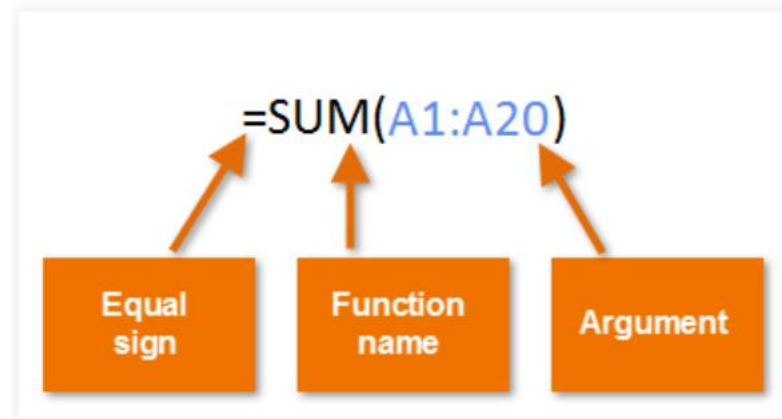


Functions in Excel

In order to work correctly, a function must be written a specific way, which is called the **syntax**.

The basic syntax for a function is an **equals** sign (=), the function **name** (SUM, for example), and one or more **arguments**. Arguments contain the information you want to calculate.

The function in the example below would add the values of the cell range A1:A20.



Basic Functions in Excel

The image displays the Microsoft Excel ribbon with the **Formulas** tab selected. The ribbon includes tabs for File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Developer. The **Formulas** tab contains the following groups:

- Function Library**: A group of icons for different function categories: Financial, Logical, Text, Date & Time, Lookup & Reference, Math & Trig, and More Functions. This group is highlighted with a red border.
- Defined Names**: Includes options like Define Name, Use in Formula, and Create from Selection.

Below the ribbon, several basic Excel functions are listed in a grid format, each with its formula syntax and a brief description:

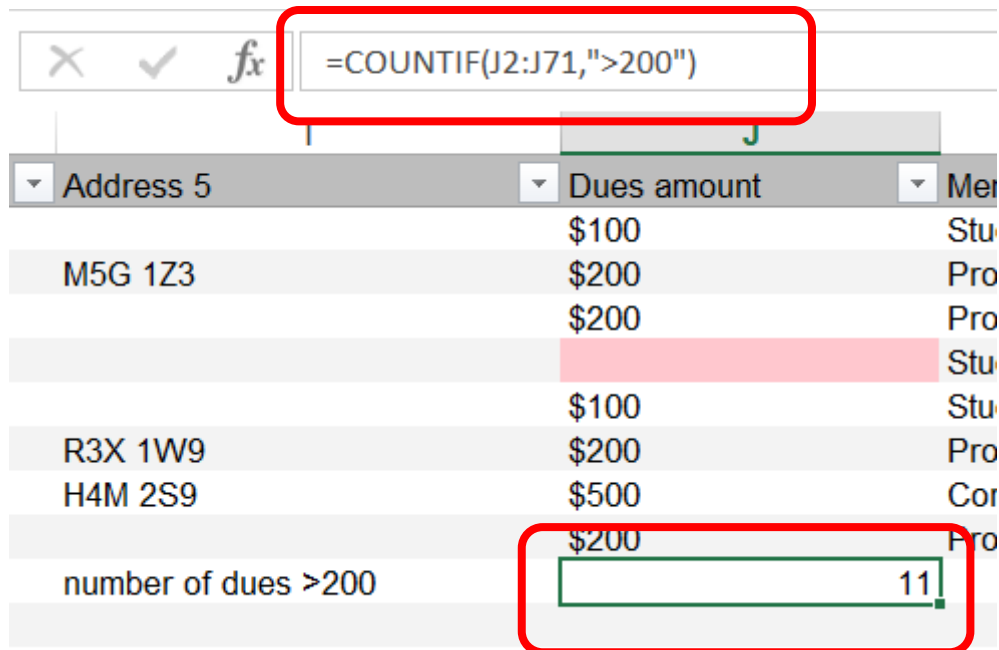
=SUM() SUM(number1, [number2], ...)	=MAX() MAX(number1, [number2], ...)	=MIN() MIN(number1, [number2], ...)	=NOW() NOW()
=AVERAGE() AVERAGE(number1, [number2], ...)	=COUNT() COUNT(value1, [value2], ...)	=COUNTA() COUNTA(value1, [value2], ...)	=LEN() LEN(text)
=ABS() ABS(number)	=RAND() RAND()	=RANDBETWEEN() RANDBETWEEN(bottom, top)	=UPPER() UPPER(text)
		=LOWER() LOWER(text)	
		=PROPER() PROPER(text)	

At the bottom left, a small grid shows the numbers 7, 8, and 9.

COUNTIF()

- count the number of cells in a range that contain numeric values with certain condition

Let's count the number of dues of the amount more than \$200.



The screenshot shows a spreadsheet with a data table and a summary row. A formula bar at the top displays the formula `=COUNTIF(J2:J71,">200")`, which is highlighted with a red box. The data table has three columns: 'Address 5', 'Dues amount', and 'Member'. The 'Dues amount' column contains values: \$100, \$200, \$200, \$100, \$200, \$500, and \$200. The 'Member' column contains values: Stu, Pro, Pro, Stu, Pro, Cor, and Pro. A summary row at the bottom is labeled 'number of dues >200' and shows the result '11' in a cell, which is also highlighted with a red box.

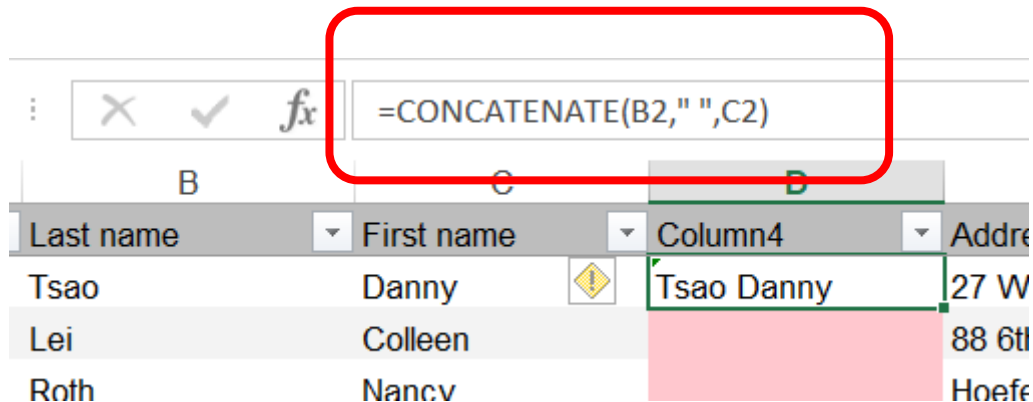
Address 5	Dues amount	Member
	\$100	Stu
M5G 1Z3	\$200	Pro
	\$200	Pro
		Stu
	\$100	Stu
R3X 1W9	\$200	Pro
H4M 2S9	\$500	Cor
	\$200	Pro
number of dues >200	11	

CONCATENATE()

- combine text from different cells into a single cell

Let's combine last and first name of the members into one cell.

For this purpose create new column and write corresponding function.



The screenshot shows an Excel spreadsheet with columns B, C, and D. Column B is labeled 'Last name', Column C is labeled 'First name', and Column D is labeled 'Column4'. The first row of data shows 'Tsao' in B2 and 'Danny' in C2. The formula bar at the top shows the formula '=CONCATENATE(B2," ",C2)' which is highlighted with a red rounded rectangle. The result of the formula, 'Tsao Danny', is displayed in cell D2. The second row shows 'Lei' in B3 and 'Colleen' in C3. The third row shows 'Roth' in B4 and 'Nancy' in C4. The 'Column4' header is highlighted in green, and the cell D2 is also highlighted in green.

B	C	D	Address
Last name	First name	Column4	
Tsao	Danny	Tsao Danny	27 W
Lei	Colleen		88 6th
Roth	Nancy		Hofe

Files to be used: *International-Logistics-Association-Memberships.csv*
International-Logistics-Association-Memberships_2.csv

INDEX() & MATCH()

- Manual appending of data from different tables

Let's move columns 'Age' from Tab 2 to Tab 1.

Important: have a unique identifier (similar column in both files).

L2			
			<div>=INDEX('[International-Logistics-Association-Memberships_2.xlsx]s2'!\$B:\$B,MATCH([@[Member ID]], '[International-Logistics-Association-Memberships_2.xlsx]s2'!\$A:\$A,0),)</div>
	J	K	L
1	Member type	Certification	Age
2	Professional Member	CLTD	23

LOOKUP()

- Manual appending of data from different tables

File to be used: Tab 1 - International-Logistics-Association-Memberships

Tab 2 - International-Logistics-Association-Memberships_2

Let's move columns 'Gender' from Tab 2 to Tab 1.

Important: have a unique identifier (similar column in both files).

The screenshot shows an Excel interface with a formula bar and a data table. The formula bar contains the following formula:

```
=LOOKUP([@[Member ID]], '[International-Logistics-Association-Memberships_2.xlsx]s2'!$A:$A, '[International-Logistics-Association-Memberships_2.xlsx]s2'!$C:$C)
```

The data table below has the following structure:

	L	M	N
1	Age	Gender	
2	31	male	

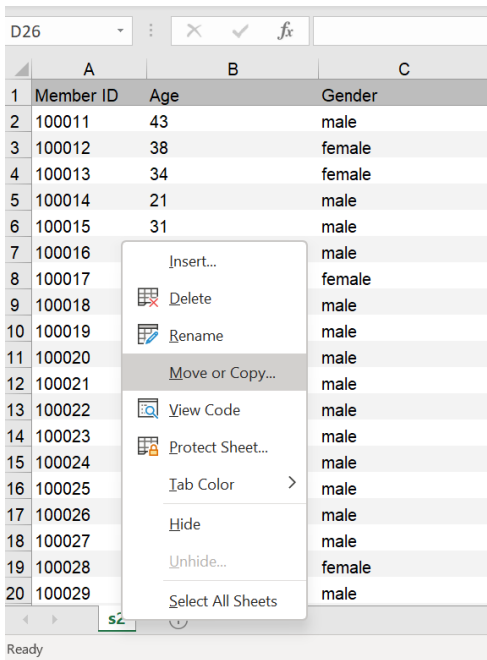
Copying Sheets

- Merge two Excel files into one by copying sheets

Let's move Sheet 's2' from Tab 2 to Tab 1.

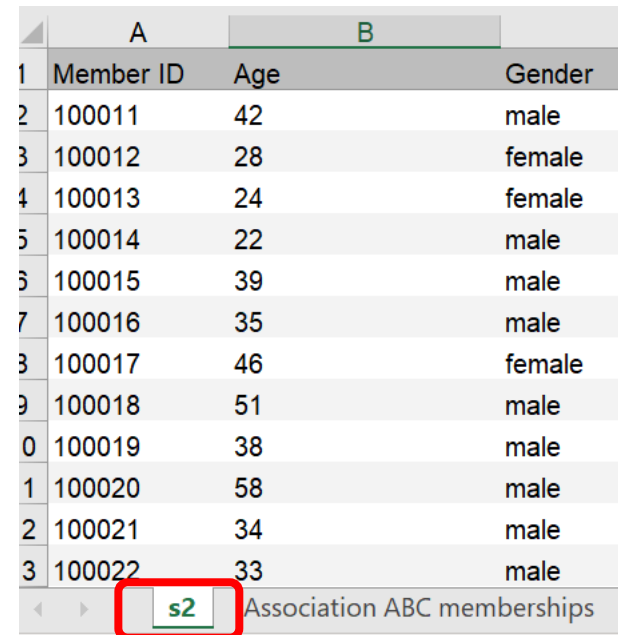
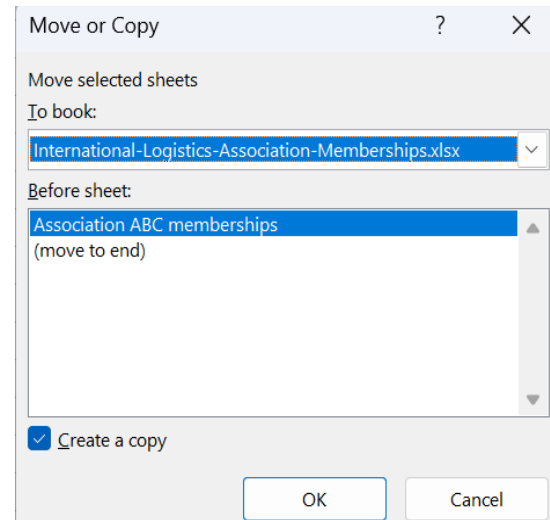
-> Right click on 's2'.

-> Remember to choose the destination file and tick 'Create a copy' option.



This screenshot shows an Excel spreadsheet with columns A, B, and C. Column A contains 'Member ID', B contains 'Age', and C contains 'Gender'. Rows 1 through 20 contain data. A right-click context menu is open over the sheet tab 's2' at the bottom. The menu options are: Insert..., Delete, Rename, Move or Copy..., View Code, Protect Sheet..., Tab Color, Hide, Unhide..., and Select All Sheets. The 'Move or Copy...' option is highlighted.

	A	B	C
1	Member ID	Age	Gender
2	100011	43	male
3	100012	38	female
4	100013	34	female
5	100014	21	male
6	100015	31	male
7	100016		male
8	100017		female
9	100018		male
10	100019		male
11	100020		male
12	100021		male
13	100022		male
14	100023		male
15	100024		male
16	100025		male
17	100026		male
18	100027		male
19	100028		female
20	100029		male



This screenshot shows the Excel spreadsheet after moving sheet 's2' to Tab 1. The sheet tab 's2' is highlighted with a red box. The spreadsheet data is the same as in the first screenshot.

	A	B	C
1	Member ID	Age	Gender
2	100011	42	male
3	100012	28	female
4	100013	24	female
5	100014	22	male
6	100015	39	male
7	100016	35	male
8	100017	46	female
9	100018	51	male
0	100019	38	male
1	100020	58	male
2	100021	34	male
3	100022	33	male

3. In-class Assignment

Instructions

Please open the DataCamp Group and do the following:

- Complete at least Chapters 1 & 2 of the Data Preparation in Excel course.
- Please don't use the DataCamp in-build AI helper.
- Submit the screenshot showing the completion of these chapters.

It's an individual assignment.

Max score: 10 points

Thank you!