Business Research and Data Analytics

Lecture 3: Data Preprocessing and Transformation with MS Excel

Iegor Vyshnevskyi, Ph.D.

Woosong University

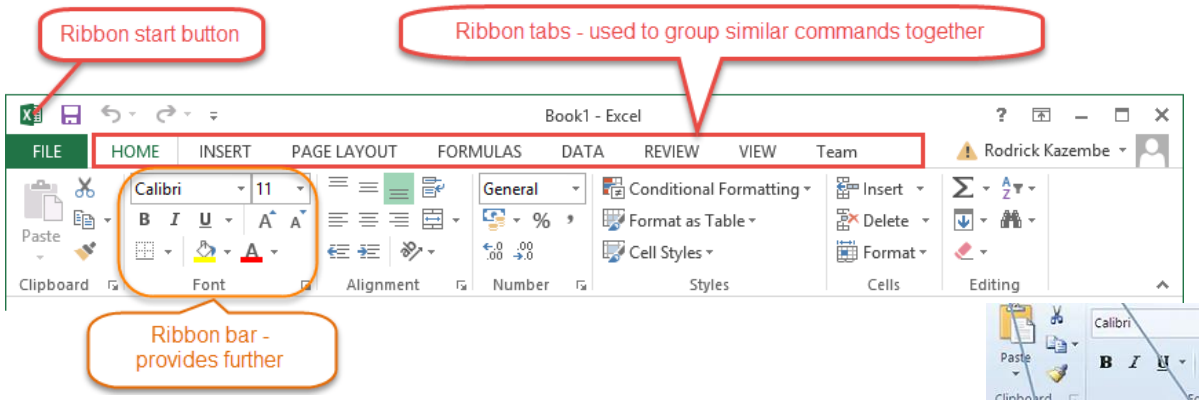
March 20, 2024

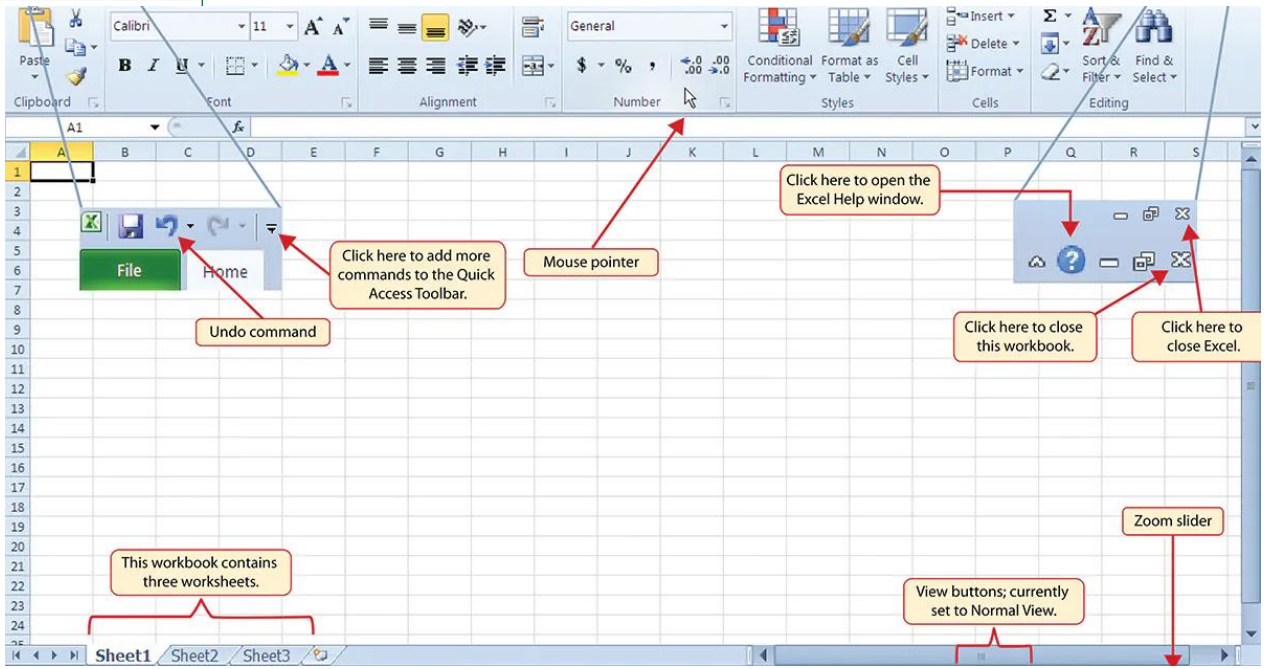
Agenda

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Outdated data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

Source: Coursera course "Process Data From Dirty to Clean"

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*

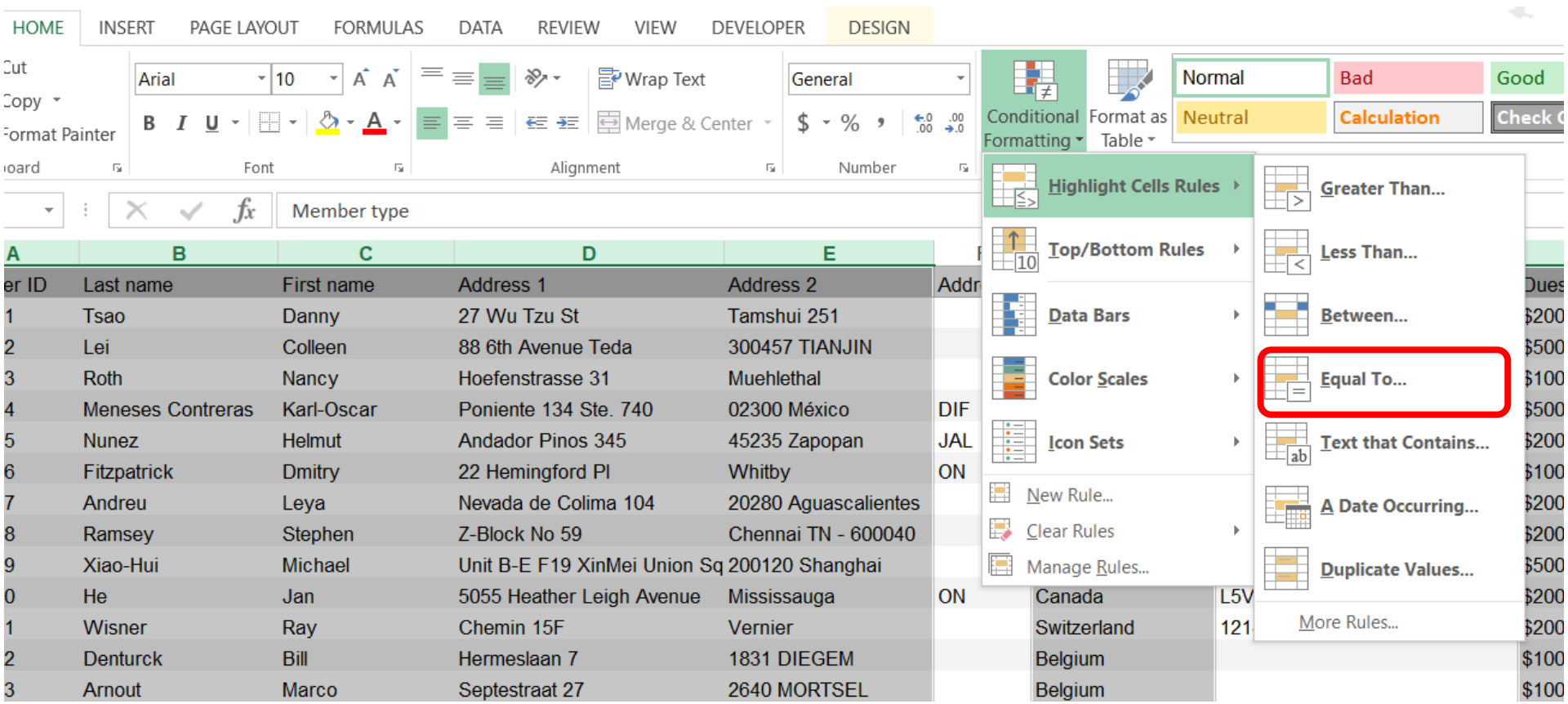
Conditional formatting

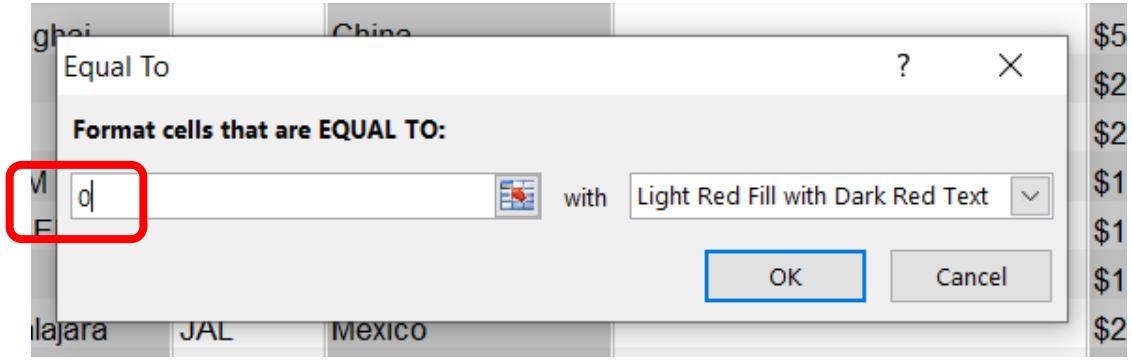
- **Identify missing values** by conditional formatting

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.


HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
REVIEW
VIEW
DEVELOPER
DESIGN
Cut
Copy
Format Painter
Arial
10
Wrap Text
Merge & Center
General
Font
Alignment
Number
Conditional Formatting
Format as Table
Normal
Bad
Good
Neutral
Calculation
Member type
Highlight Cells Rules
Top/Bottom Rules
Data Bars
Color Scales
Icon Sets
New Rule...
Clear Rules
Manage Rules...
Greater Than...
Less Than...
Between...
Equal To...
Text that Contains...
A Date Occurring...
Duplicate Values...
More Rules...
Last name
First name
Address 1
Address 2
Tsao
Danny
27 Wu Tzu St
Tamshui 251
Lei
Colleen
88 6th Avenue Teda
300457 TIANJIN
Roth
Nancy
Hoefenstrasse 31
Muehlethal
Meneses Contreras
Karl-Oscar
Poniente 134 Ste. 740
02300 México
Nunez
Helmut
Andador Pinos 345
45235 Zapopan
Fitzpatrick
Dmitry
22 Hemingford Pl
Whitby
Andreu
Leya
Nevada de Colima 104
20280 Aguascalientes
Ramsey
Stephen
Z-Block No 59
Chennai TN - 600040
Xiao-Hui
Michael
Unit B-E F19 XinMei Union Sq
200120 Shanghai
He
Jan
5055 Heather Leigh Avenue
Mississauga
Canada
Wisner
Ray
Chemin 15F
Vernier
Switzerland
Denturck
Bill
Hermeslaan 7
1831 DIEGEM
Belgium
Arnout
Marco
Septestraat 27
2640 MORTSEL
Belgium

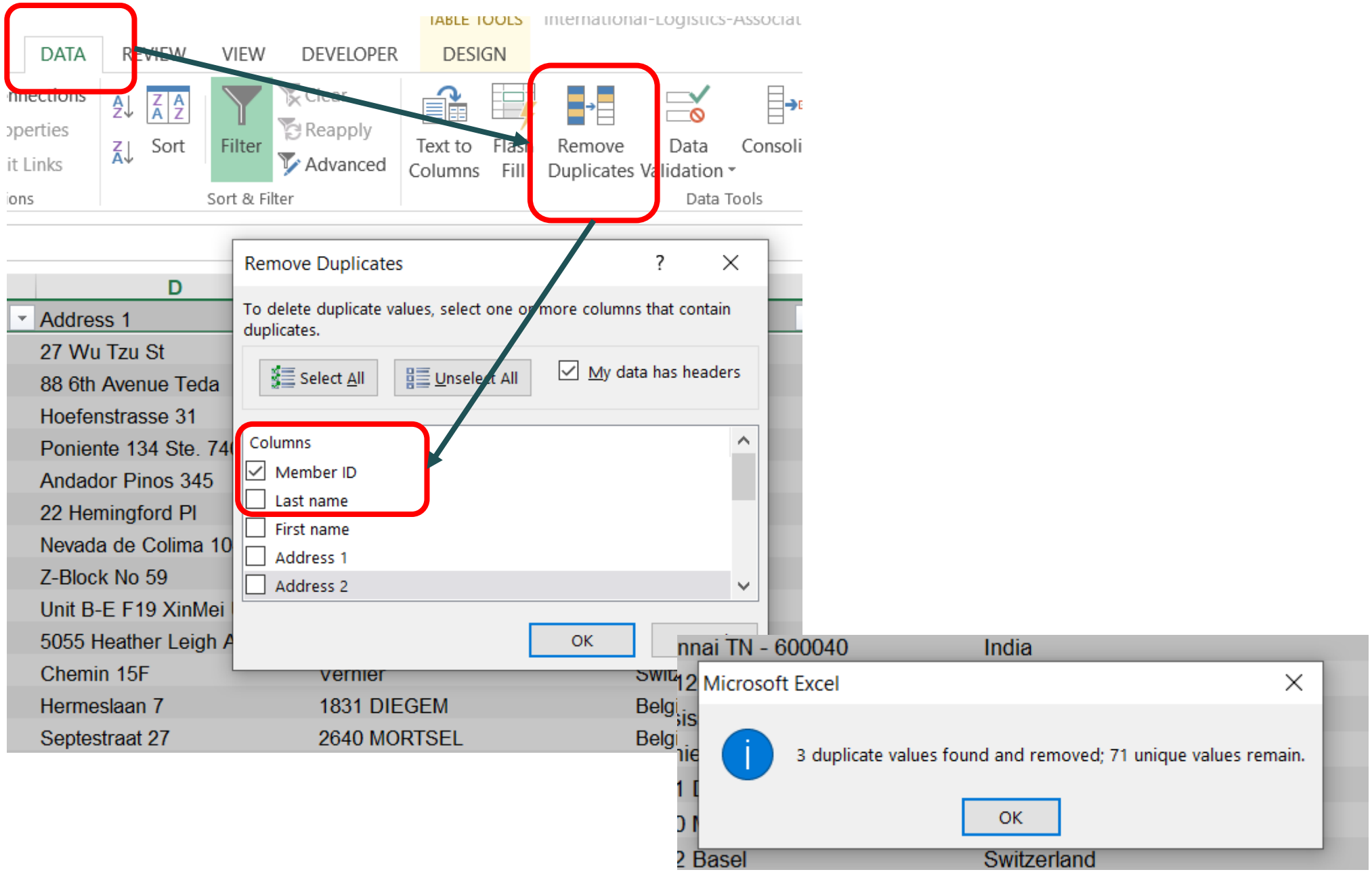

Equal To
Format cells that are EQUAL TO:
0
with
Light Red Fill with Dark Red Text
OK
Cancel

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.

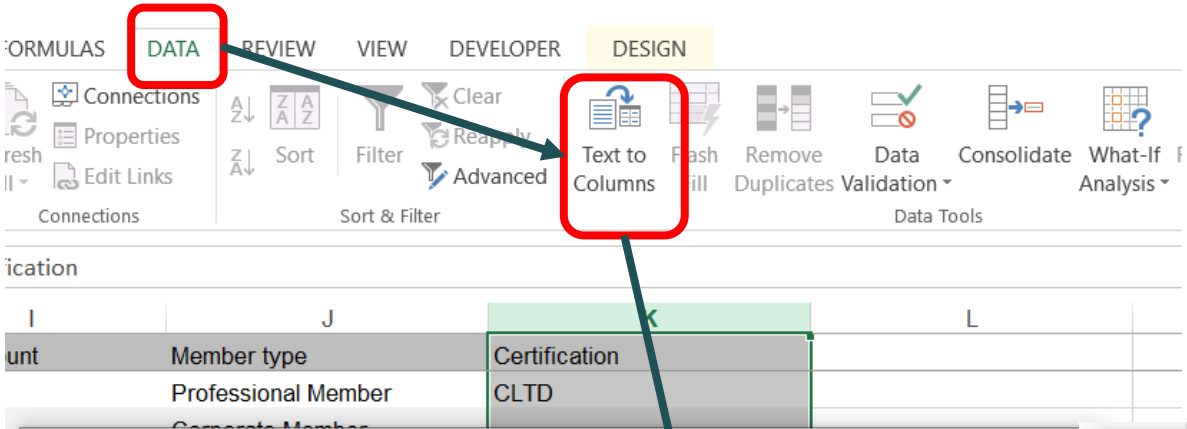

DATA
VIEW
DEVELOPER
DESIGN
Sort
Filter
Reapply
Advanced
Sort & Filter
Text to Columns
Remove Duplicates
Data Validation
Data Tools
Remove Duplicates
To delete duplicate values, select one or more columns that contain duplicates.
Select All
Unselect All
My data has headers
Columns
Member ID
Last name
First name
Address 1
Address 2
OK
Address 1
27 Wu Tzu St
88 6th Avenue Teda
Hoefenstrasse 31
Andador Pinos 345
22 Hemingford Pl
Z-Block No 59
Chemin 15F
Hermeslaan 7
Septestraat 27
1831 DIEGEM
2640 MORTSEL
India
Microsoft Excel
3 duplicate values found and removed; 71 unique values remain.
OK
Switzerland

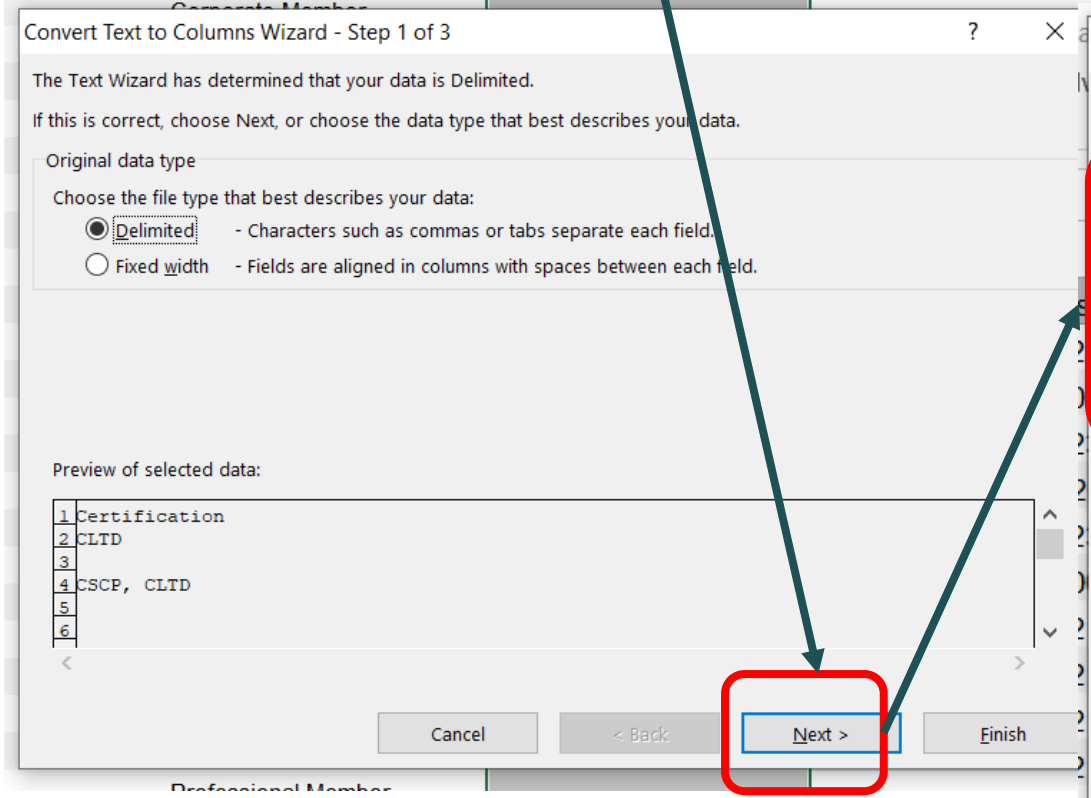
Splitting the data

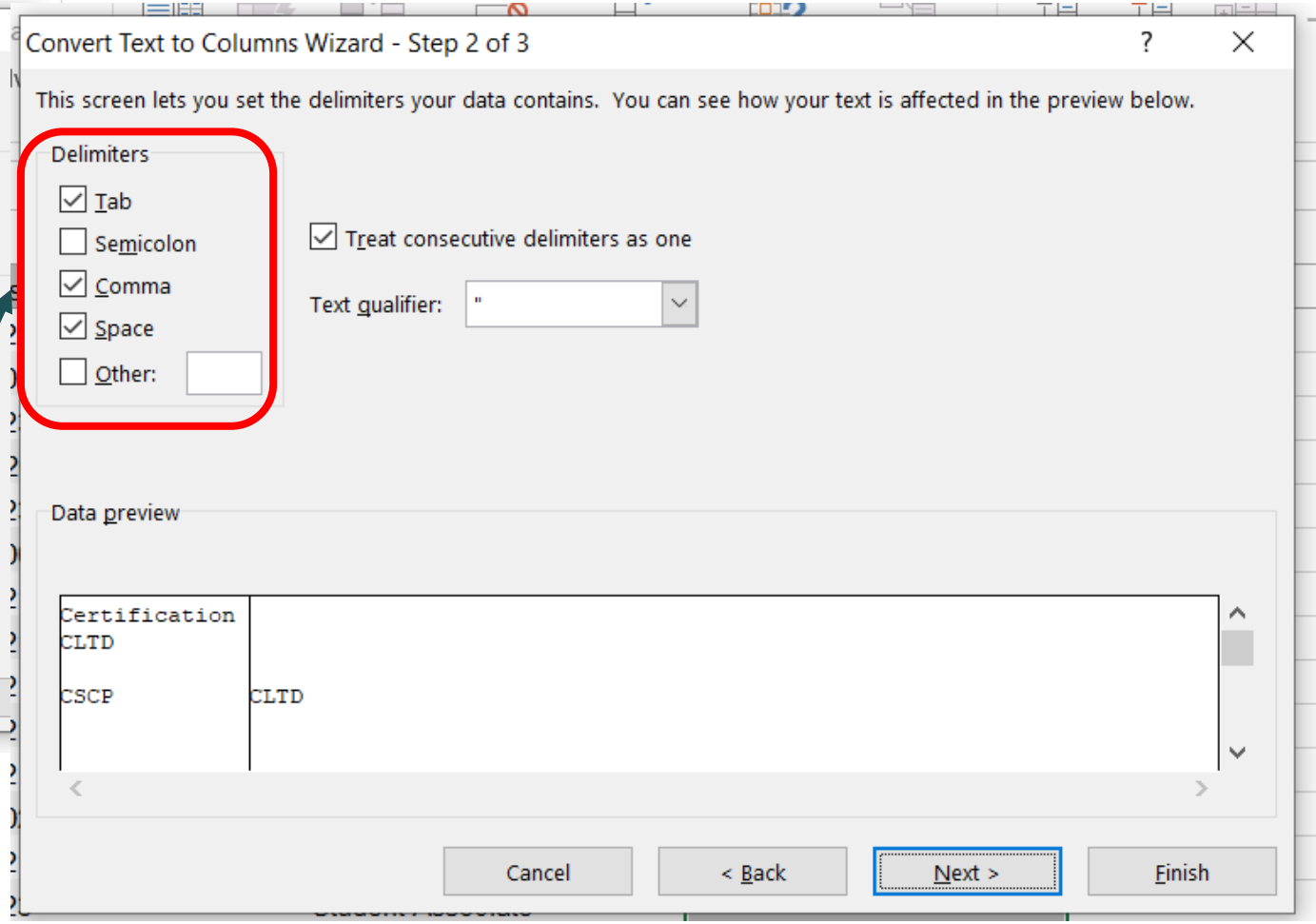
Let’s split certification info into different columns

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.


FORMULAS
DATA
REVIEW
VIEW
DEVELOPER
DESIGN
Connections
Properties
Edit Links
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Flash Fill
Remove Duplicates
Data Validation
Consolidate
What-If Analysis
Data Tools
I
J
K
L
Member type
Certification
Professional Member
CLTD


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

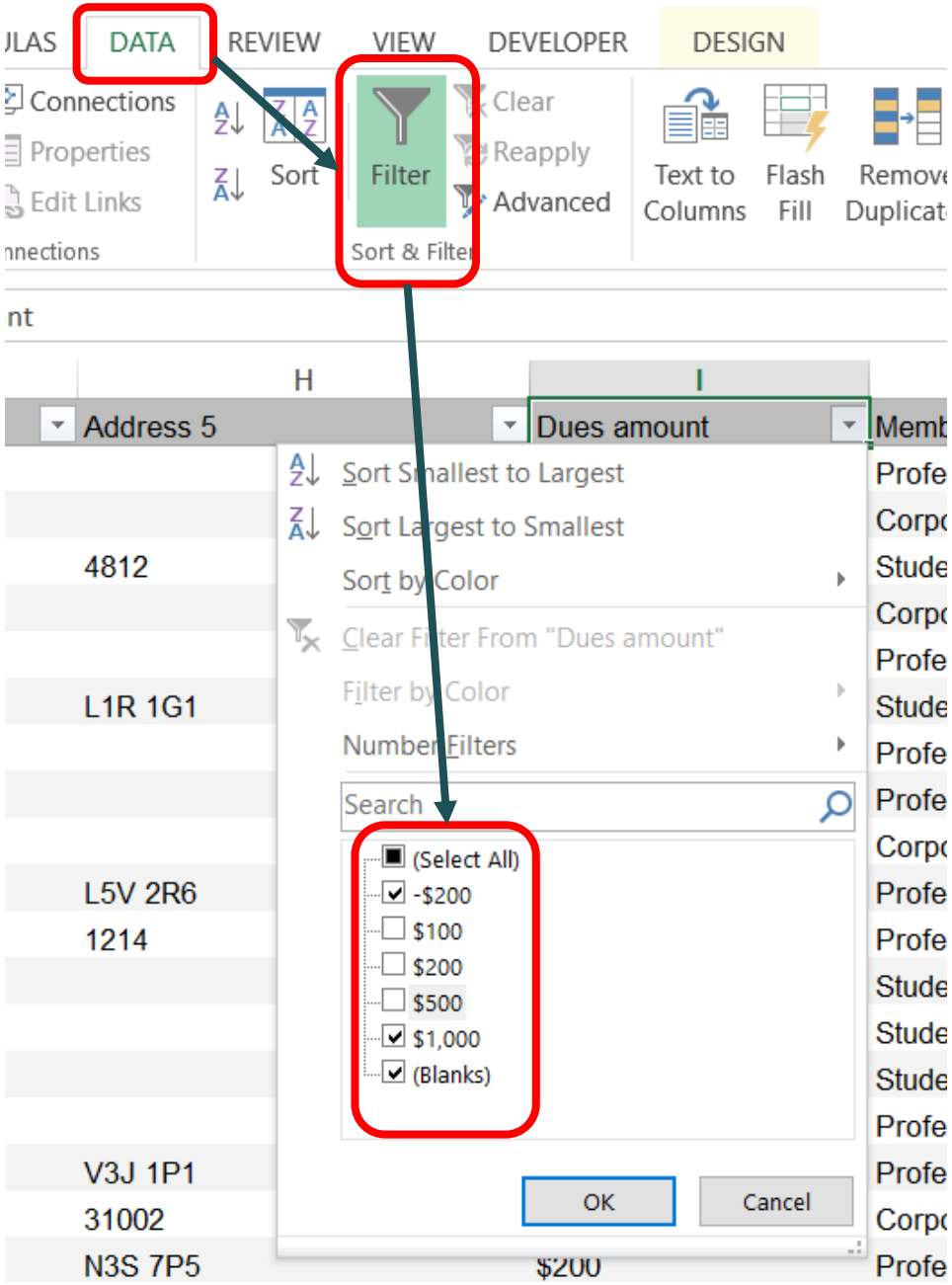

Convert Text to Columns Wizard - Step 2 of 3
This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.
Delimiters
Tab
Semicolon
Comma
Space
Other:
Treat consecutive delimiters as one
Text qualifier:
"
Data preview
Certification
CLTD
CSCP
CLTD
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.



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	I	
	Dues amount	Membe
	$1,000	Student
	-$200	Profess

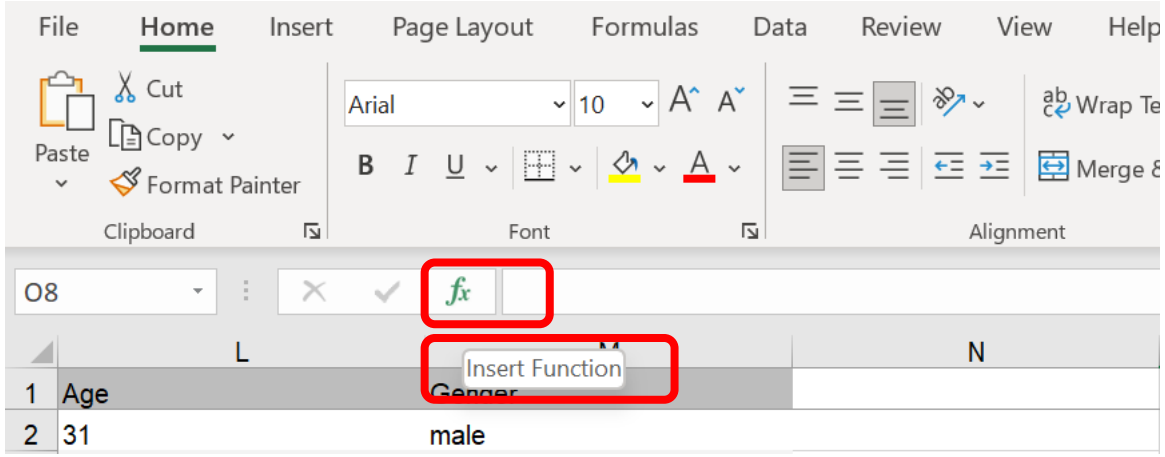
	I	
	Dues amount	Member typ
	$100	Student Ass
	$200	Professiona

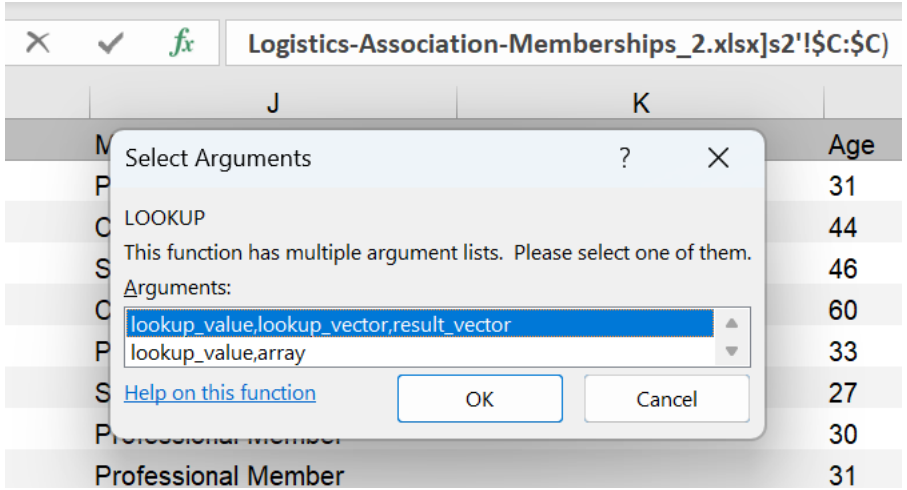
Functions in Excel

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:



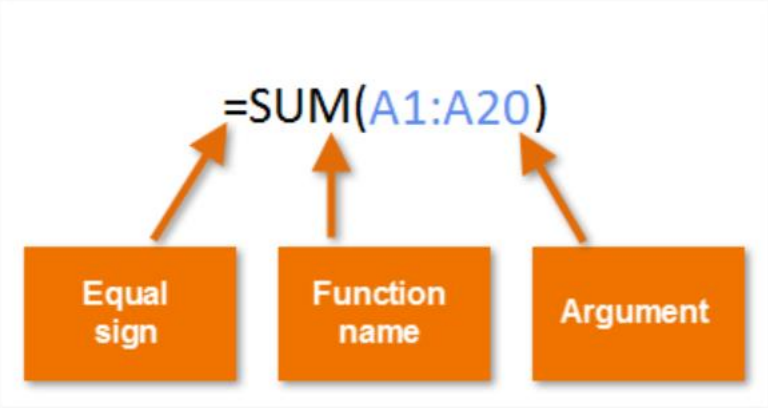


Functions in Excel

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

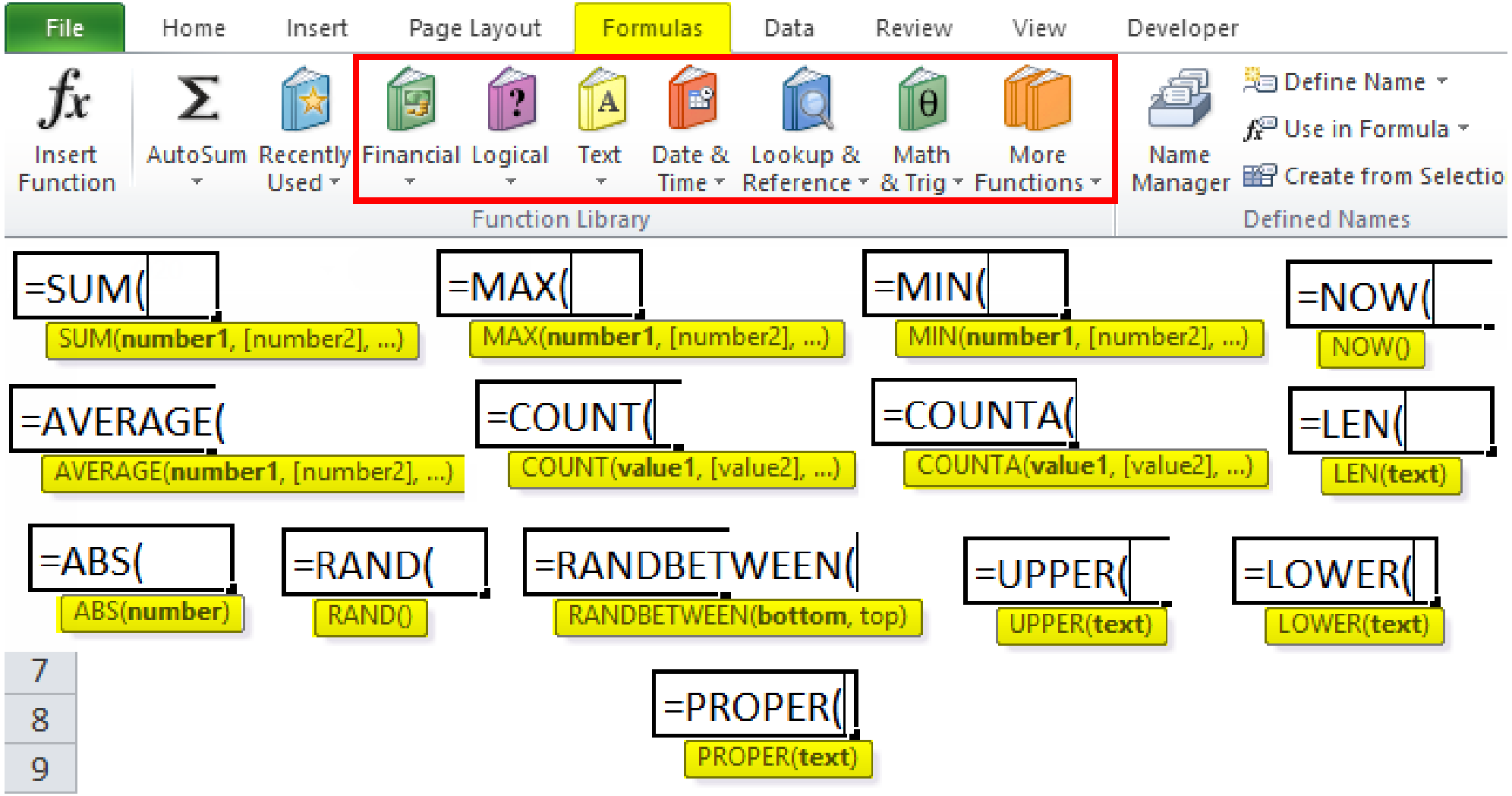
The basic syntax for a function is an **equals** sign (=), the function **name** (SUM, for example), and one or more **arguments**. Arguments contain the information you want to calculate.

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



[Top 30 Excel Formulas and Functions]

Basic Functions in Excel

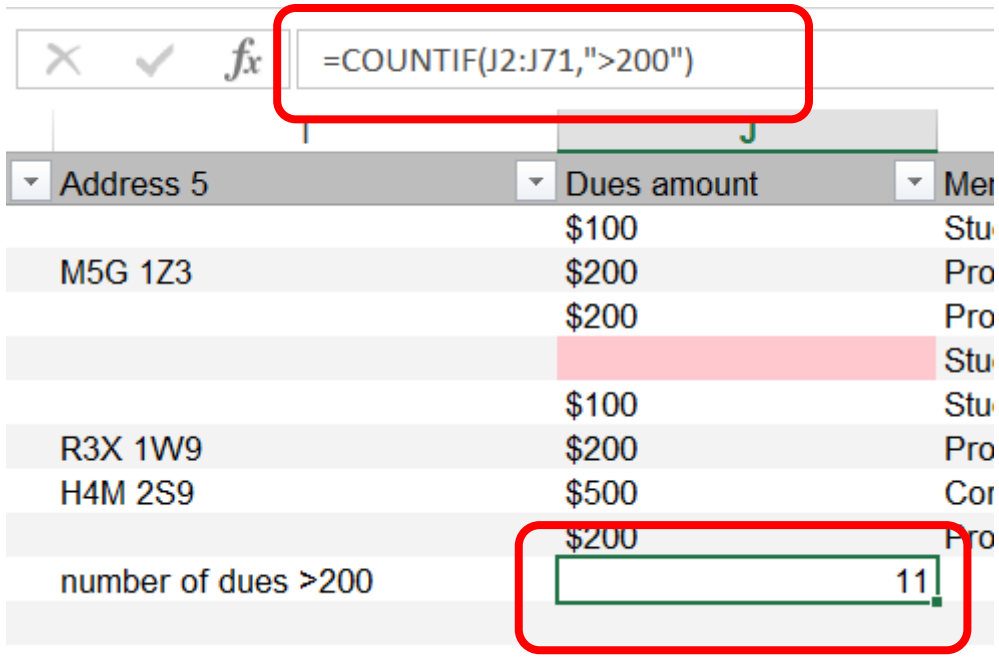


Source: ProjectCubicle

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.

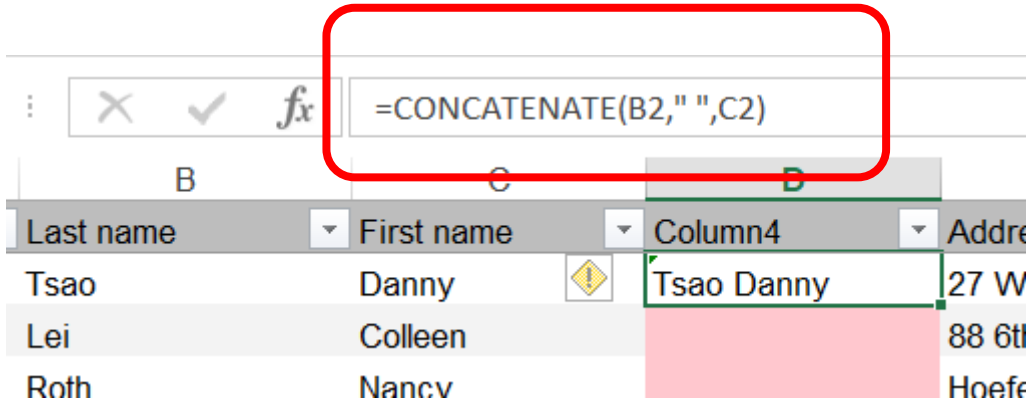


CONCATENATE()

- 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.



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

=INDEX('[International-Logistics-Association-Memberships_2.xlsx]s2'!$B:$B,MATCH([@[Member ID]], '[International-Logistics-Association-Memberships_2.xlsx]s2'!$A:$A,0),)

	J	K	L
1	Member type	Certification	Age
2	Professional Member	CLTD	23

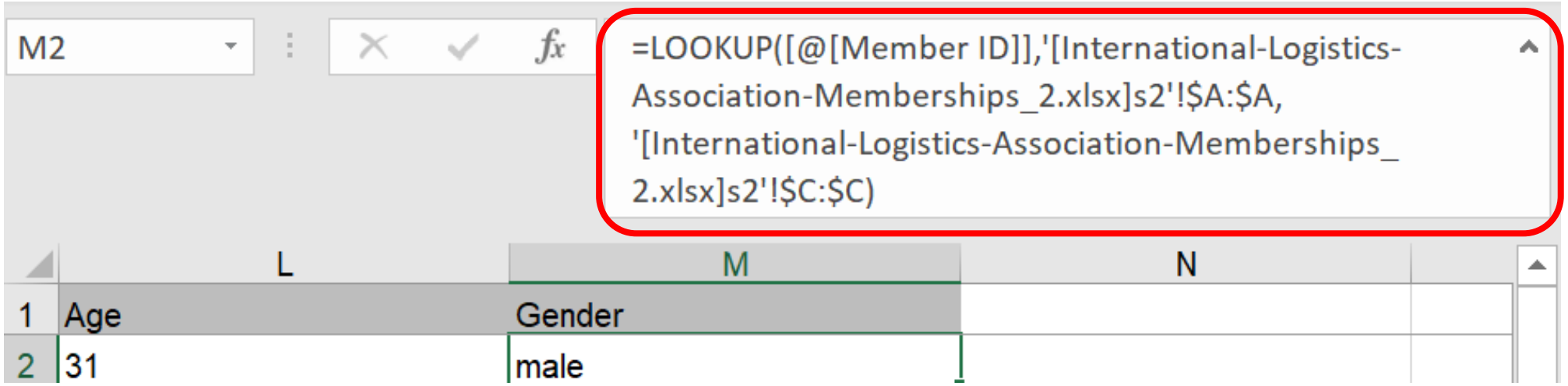
LOOKUP()

- 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).



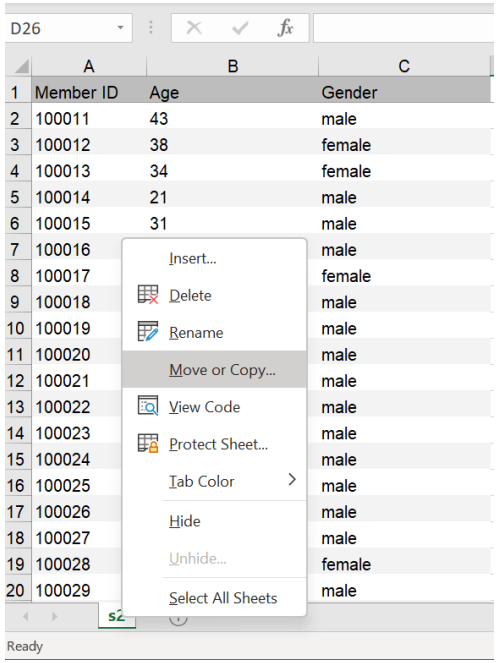
Copying Sheets

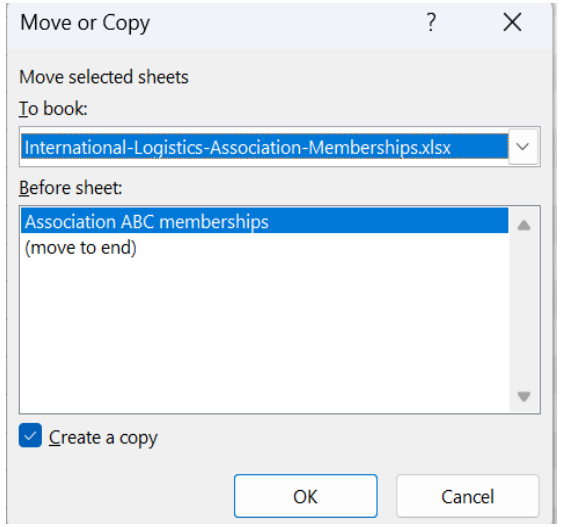
- Merge two Excel files into one by copying sheets

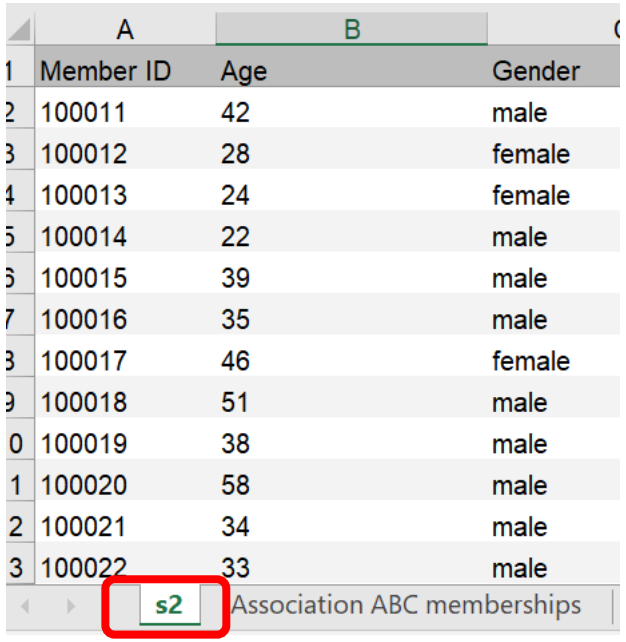
Let’s move Sheet ‘s2’ from Tab 2 to Tab 1.

-> Right click on ‘s2’.

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







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!