DATA1002 Week 3 Tutorial

Monday 18/08/25

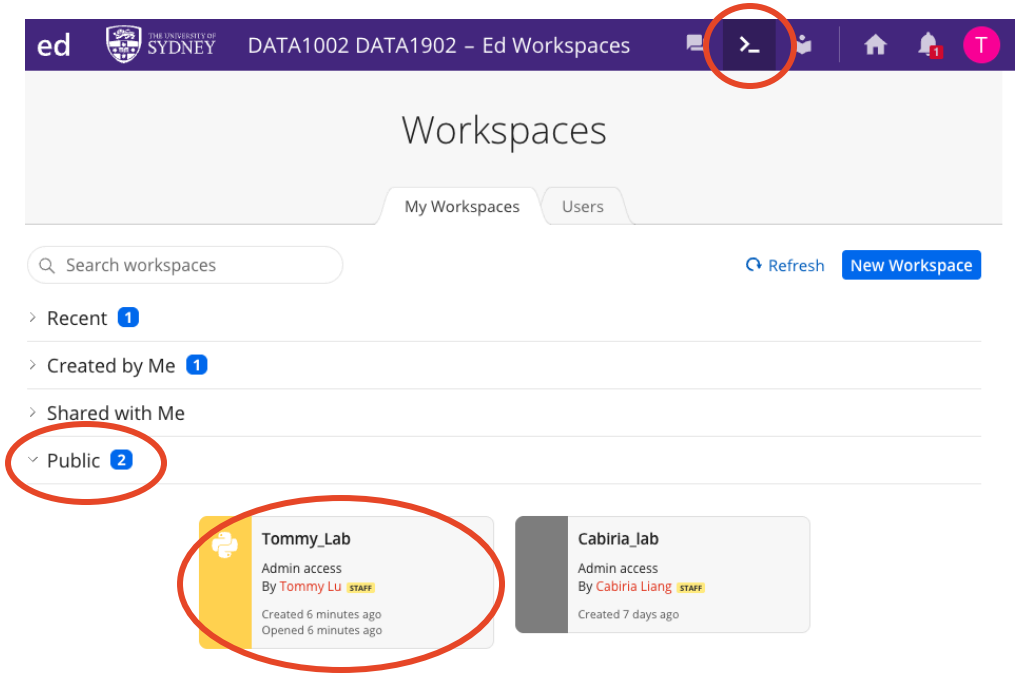
Tutorial Outline

- Content revision (Conditionals & Loops)
- Python demonstration
- Lab Pre-Work
- Lab Activities

THE UNIVERSITY OF SYDNEY

Tutor: *Tommy Lu*

Access the material for this tutorial through **Ed Workspaces**



Content Revision

Conditional Structure

if vs. if-else vs. if-elif

if

```
code-before-line-1
code-before-line-2
if condition:
    true-block-line1
    true-block-line2
code-after-line-1
code-after-line-2
```

if-else

```
code-before-line-1
code-before-line-2
if condition:
    true-block-line1
    true-block-line2
else:
    false-block-line-1
    false-block-line-2
code-after-line-1
code-after-line-2
```

if-elif

```
code-before-line-1
code-before-line-2
if condition1:
    block1-line1
    block1-line2
elif condition2:
    block2-line-1
    block2-line-2
elif condition3:
    block3-line-1
    block3-line-2
code-after-line-1
code-after-line-2
```

Tracing Exercise

```
x = int(input("Enter x: "))
y = int(input("Enter y: "))
if x < y:
    z = y - x
    if z > 10:
        print("y is much bigger")
    else:
        print("y is bigger by ",z)
    print("z is "+str(z))
else:
    print("x is bigger")
print("Goodbye")
```

For-Loop

```
code-before-line-1
code-before-line-2
for loop-control-variable in collection:
  body-line1
  body-line2
code-after-line-1
code-after-line-2
```

1	2	3	4	5	6	7	8	9	10

For-Loop

Example: reading lines from a text file

- Iterate through the lines of a text file
  - for line in open(filename):
    process-with-line
  - for line in open(filename).readlines():
    process-with-line
  - with open(filename) as f:
    for line in f:
    process-with-line

Python Exercises

Conditionals & Loops

Conditionals

```
# Exercise 1: Grades
# From an input, print out the associated grade from a given score

score = int(input("Enter your score: "))

# BONUS Exercise (if already familiar with Python): Bank Account v2
# Write a program that lets a user withdraw or deposit funds into their bank account.
# • The bank account should start with a default of $0.
# • The user should be able to indicate how many withdraw/deposit transactions they would
#   like to make.
# • Each transaction that is successful you should print the withdraw/deposit success and
#   display the new balance.
# • A transaction will fail if:
#   o The deposit is above the daily limit of $1000
#   o The withdrawal would leave the account with less than $0
# • If a transaction fails, inform the user.
# • If the user enters an invalid transaction type, inform the user.
# • Failed transactions and invalid transaction types will still count towards their total
#   transactions made.
# • After all transactions have finished, print the final balance.

# The following is last week's implementation of the bank account system:

print("Bank Account Simulator")
```

For code, go to Tommy_Lab Ed Workspace, Week 3

Loops

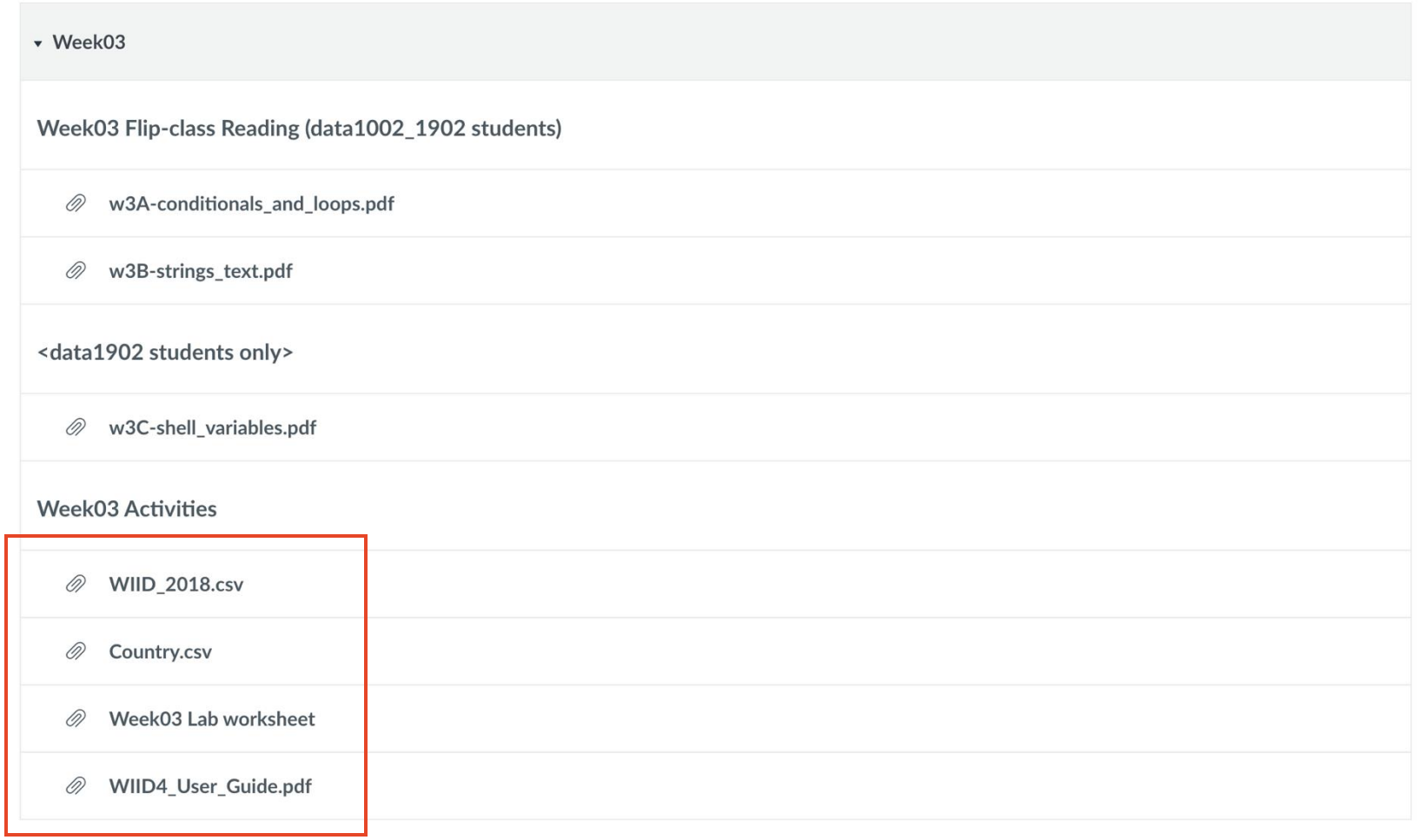
```
# Exercise 2: Pattern Printing
# The program below should print out the following pattern:
#   *
#   * *
#   * * *
#   * * * *
#   * * * * *
#   * * * *
#   * * *
#   * *
#   *
# However, there are some bugs in the code. You should be able to find and fix at
# least 3 bugs (and maybe find one that looks like a bug, but isn't!)

rows = 5
for i in range(0, rows-1):
    for j in range(0, i):
        print("*")
    print()

for j in range(rows, 0, -1):
    for i in range(0, j - 1):
        print("*")
    print()

# BONUS Exercise (if already familiar with Python): Fizzbuzz
# Print numbers 1-30
# If divisible by 3 → "Fizz"
# If divisible by 5 → "Buzz"
# If divisible by both → "FizzBuzz"
```

For code, go to Tommy_Lab Ed Workspace, Week 3

Lab Pre-Work

Analysing Data using Spreasheets

Download the material



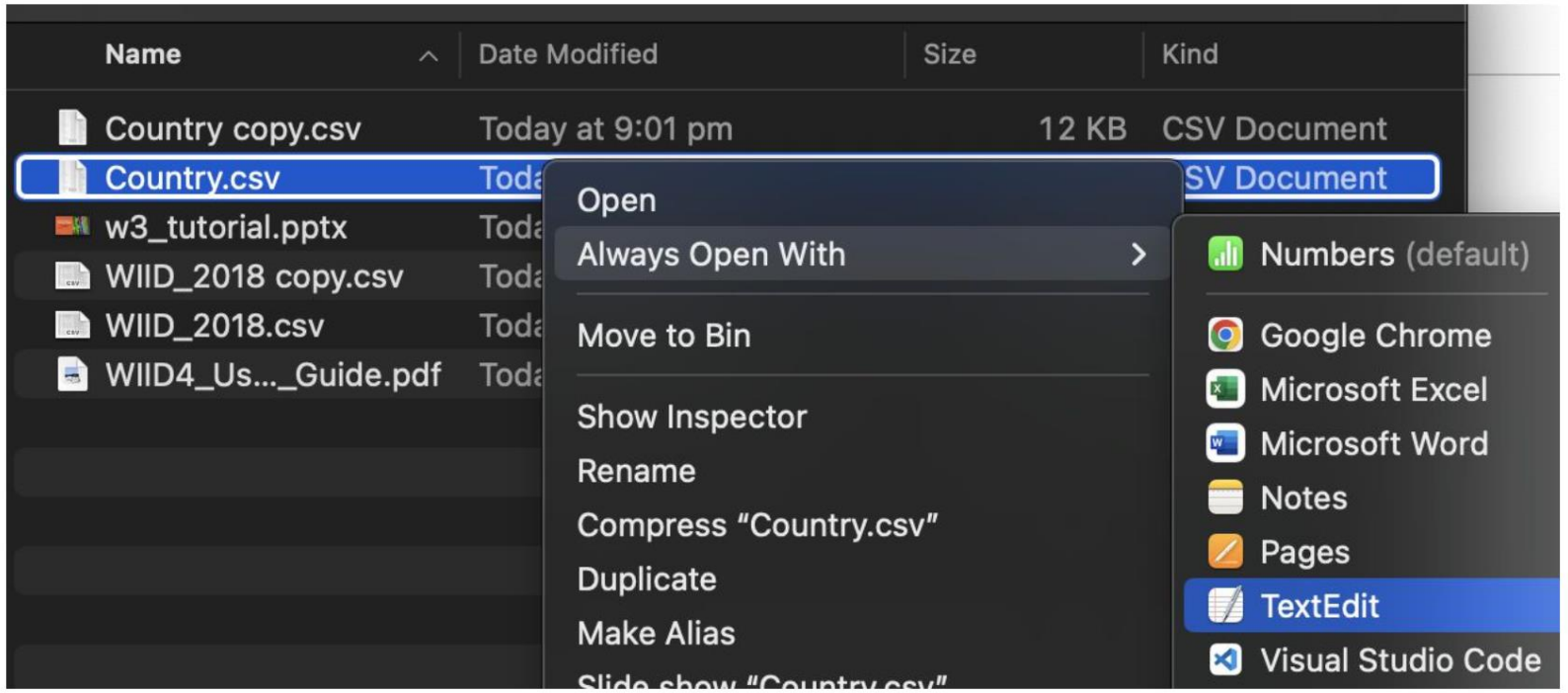
Initial Exploration and Setup

First, inspect the Country.csv file with a text editor. You will see in the row starting BHS a case where the third field (called country) has a value ("Bahamas, The") that has a comma within it, and the field is enclosed by double-quotes to prevent the comma being interpreted as a field-separator. There are other cases like this elsewhere in the file, and similar cases in WIID_2018.csv too.

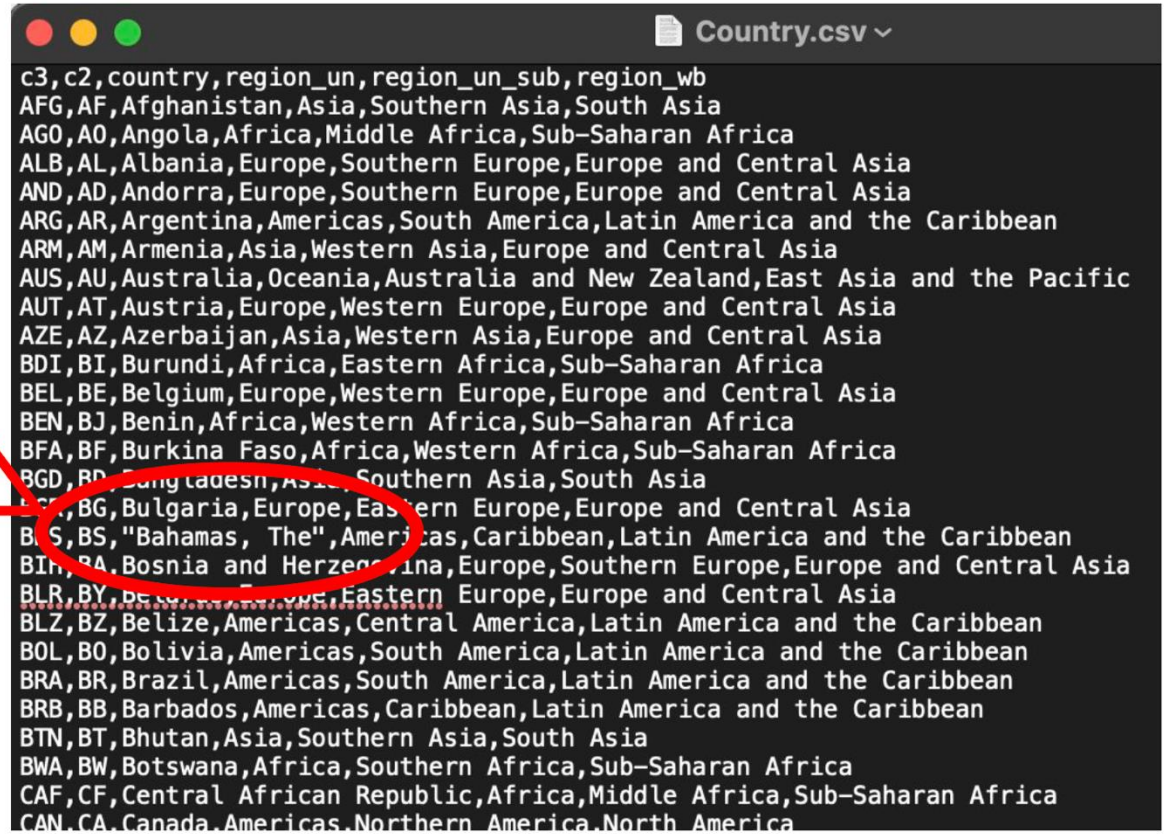
Now, import the Country.csv file in the spreadsheet tool you chose. Notice whether the tool deals correctly with the cases where a field has an embedded comma (Microsoft Excel does this!).

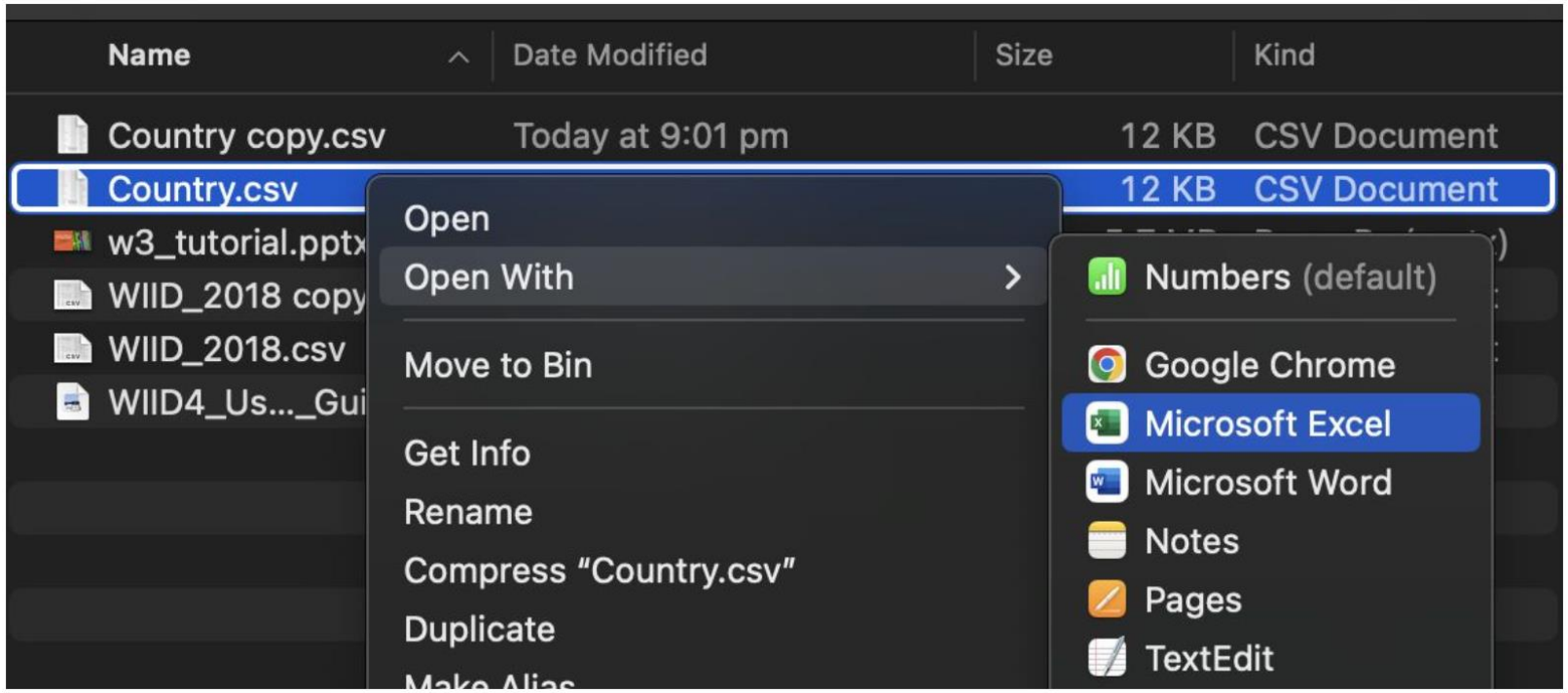
comma-separated values

A CSV (comma-separated values) file is a text file that has a specific format which allows data to be saved in a table structured format.

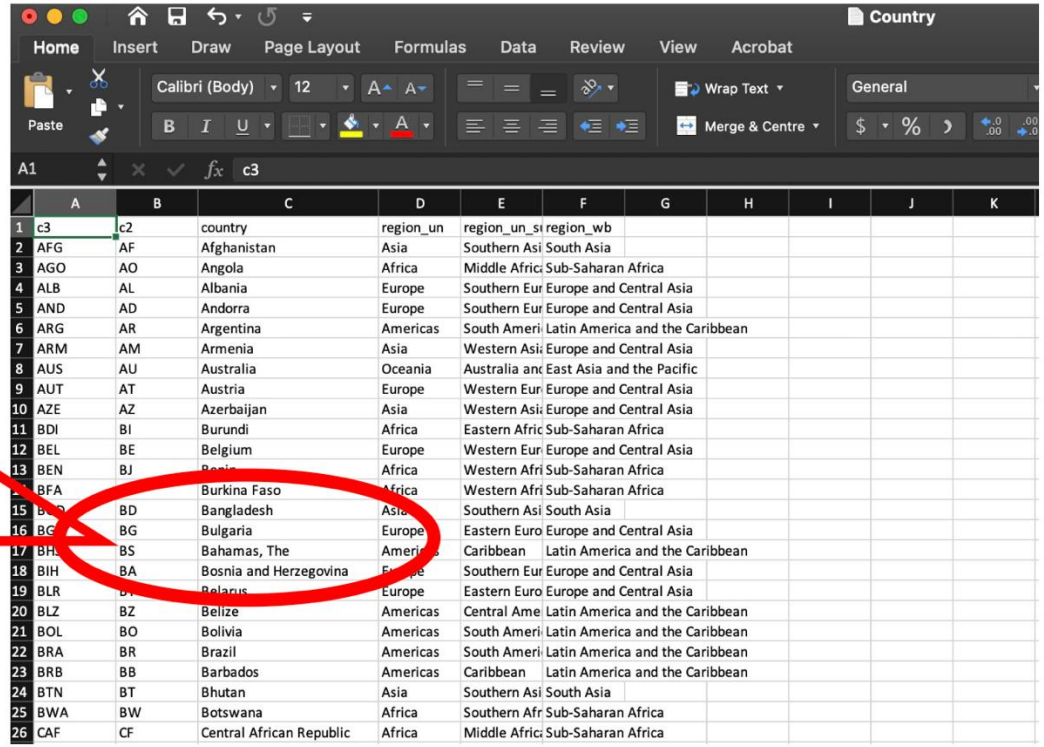

Name
Date Modified
Size
Kind
Country copy.csv
Today at 9:01 pm
12 KB
CSV Document
Country.csv
SV Document
w3_tutorial.pptx
WIID_2018 copy.csv
WIID_2018.csv
WIID4_Us..._Guide.pdf
Open
Always Open With
Move to Bin
Show Inspector
Rename
Compress “Country.csv”
Duplicate
Make Alias
Numbers (default)
Google Chrome
Microsoft Excel
Microsoft Word
Notes
Pages
TextEdit
Visual Studio Code

Bahamas, The is enclosed in " " so that its comma is not interpreted as a field-separator




Name
Date Modified
Size
Kind
Country copy.csv
Today at 9:01 pm
12 KB
CSV Document
Country.csv
12 KB
CSV Document
w3_tutorial.pptx
WIID_2018 copy
WIID_2018.csv
WIID4_Us..._Gui
Open
Open With
Move to Bin
Get Info
Rename
Compress “Country.csv”
Duplicate
Numbers (default)
Google Chrome
Microsoft Excel
Microsoft Word
Notes
Pages
TextEdit

Microsoft Excel handled this case well!



	A	B	C	D	E	F	G	H	I	J	K
1	c3	c2	country	region_un	region_un_s	region_wb					
2	AFG	AF	Afghanistan	Asia	Southern Asi	South Asia					
3	AGO	AO	Angola	Africa	Middle Afric	Sub-Saharan Africa					
4	ALB	AL	Albania	Europe	Southern Eur	Europe and Central Asia					
5	AND	AD	Andorra	Europe	Southern Eur	Europe and Central Asia					
6	ARG	AR	Argentina	Americas	South Ameri	Latin America and the Caribbean					
7	ARM	AM	Armenia	Asia	Western Asi	Europe and Central Asia					
8	AUS	AU	Australia	Oceania	Australia an	East Asia and the Pacific					
9	AUT	AT	Austria	Europe	Western Eur	Europe and Central Asia					
10	AZE	AZ	Azerbaijan	Asia	Western Asi	Europe and Central Asia					
11	BDI	BI	Burundi	Africa	Eastern Afric	Sub-Saharan Africa					
12	BEL	BE	Belgium	Europe	Western Eur	Europe and Central Asia					
13	BEN	BJ	Benin	Africa	Western Afri	Sub-Saharan Africa					
14	BFA		Burkina Faso	Africa	Western Afri	Sub-Saharan Africa					
15	BGD	BD	Bangladesh	Asia	Southern Asi	South Asia					
16	BGR	BG	Bulgaria	Europe	Eastern Euro	Europe and Central Asia					
17	BHS	BS	Bahamas, The	Americas	Caribbean	Latin America and the Caribbean					
18	BIH	BA	Bosnia and Herzegovina	Europe	Southern Eur	Europe and Central Asia					
19	BLR	BY	Belarus	Europe	Eastern Euro	Europe and Central Asia					
20	BLZ	BZ	Belize	Americas	Central Ame	Latin America and the Caribbean					
21	BOL	BO	Bolivia	Americas	South Ameri	Latin America and the Caribbean					
22	BRA	BR	Brazil	Americas	South Ameri	Latin America and the Caribbean					
23	BRB	BB	Barbados	Americas	Caribbean	Latin America and the Caribbean					
24	BTN	BT	Bhutan	Asia	Southern Asi	South Asia					
25	BWA	BW	Botswana	Africa	Southern Afr	Sub-Saharan Africa					
26	CAF	CF	Central African Republic	Africa	Middle Afric	Sub-Saharan Africa					

Following Pre-Work Activities

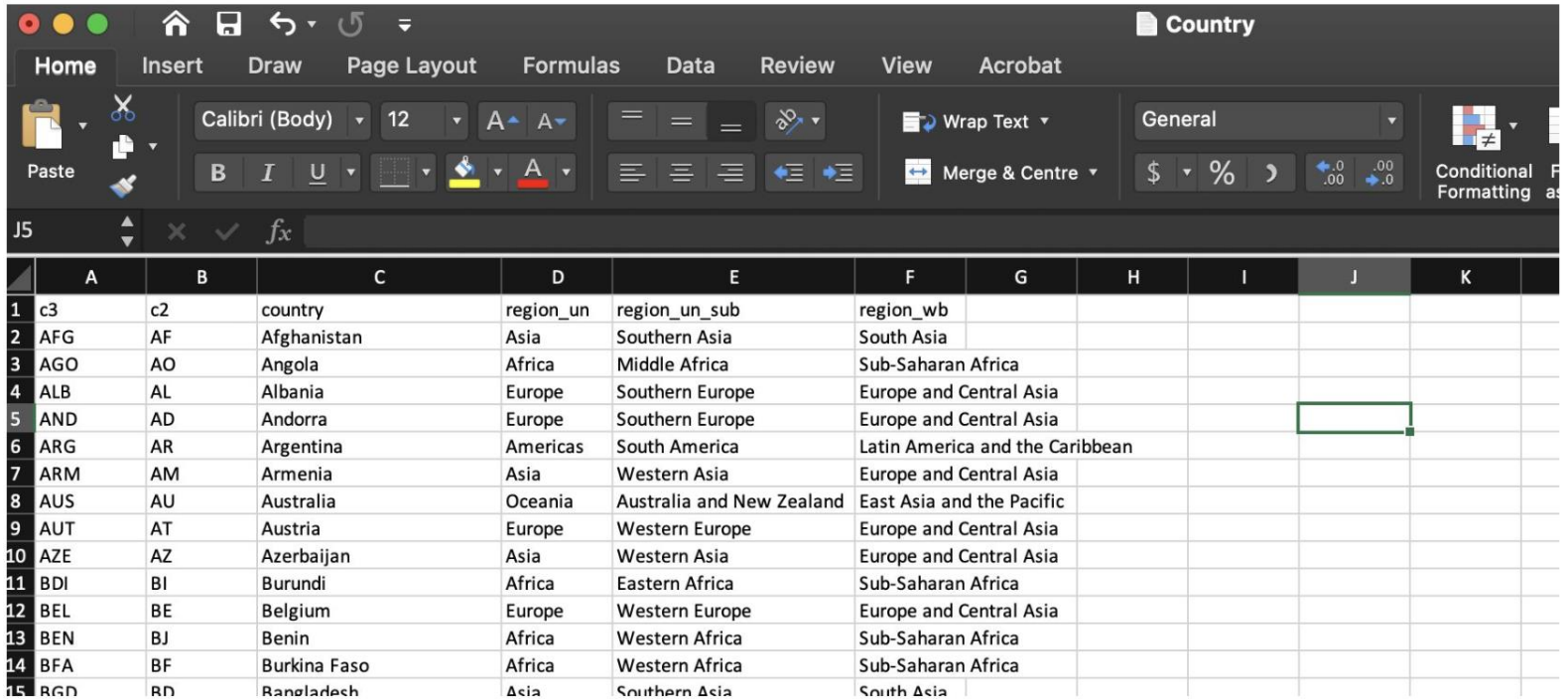
- [x] How many countries are part of the **World Bank region** (column region_wb) called **"Europe and Central Asia"?**
- [x] How many countries are in both the **World Bank region** (column region_wb) "Middle East and North Africa" and ALSO in the **United Regions** subregion (column region_un_sub) called **"Western Asia"**
- [x] **Index & Match**

Q1: How many countries are part of the World Bank region (column region_wb) called "Europe and Central Asia"?

Method:

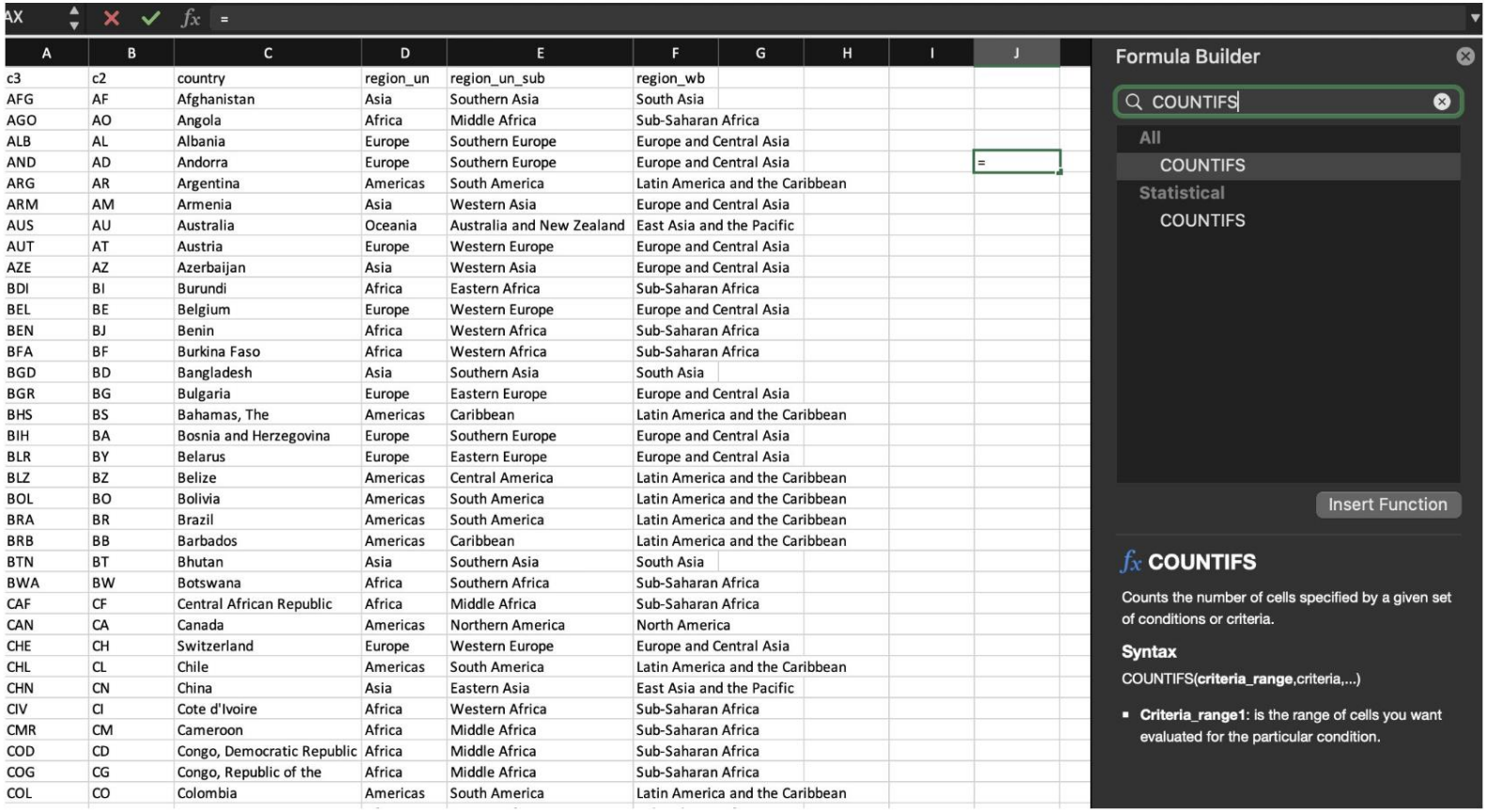
You can do this by sorting the data, with the region_wb column controlling the sort, and then the relevant rows will be placed together and can easily be counted. Or, you can leave the data in its original order, and choose an empty cell and put a formula there that uses COUNTIF. Still another way, is to create a new column, and put a formula there that gives the value 1 in rows where the region_wb is the one we are interested in, and instead gives value which is the empty string for the other rows; then we can then add up all the values in this new column, and that will count the appropriate rows. Try *each* of these approaches. [Hint: always, when working on a dataset, make a copy of the file and work on that, keeping the original dataset unchanged to return to if something goes wrong!]

Select an empty cell



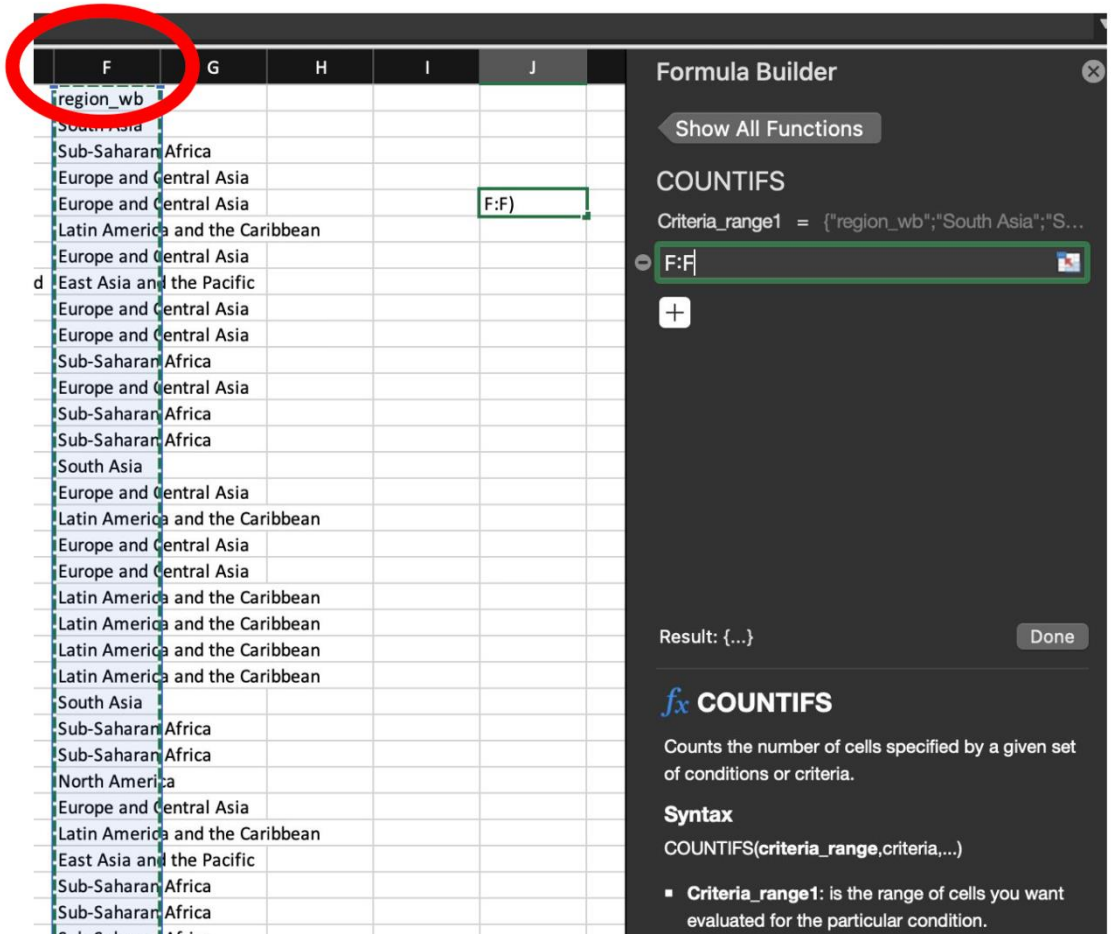
	A	B	C	D	E	F	G	H	I	J	K
1	c3	c2	country	region_un	region_un_sub	region_wb					
2	AFG	AF	Afghanistan	Asia	Southern Asia	South Asia					
3	AGO	AO	Angola	Africa	Middle Africa	Sub-Saharan Africa					
4	ALB	AL	Albania	Europe	Southern Europe	Europe and Central Asia					
5	AND	AD	Andorra	Europe	Southern Europe	Europe and Central Asia					
6	ARG	AR	Argentina	Americas	South America	Latin America and the Caribbean					
7	ARM	AM	Armenia	Asia	Western Asia	Europe and Central Asia					
8	AUS	AU	Australia	Oceania	Australia and New Zealand	East Asia and the Pacific					
9	AUT	AT	Austria	Europe	Western Europe	Europe and Central Asia					
10	AZE	AZ	Azerbaijan	Asia	Western Asia	Europe and Central Asia					
11	BDI	BI	Burundi	Africa	Eastern Africa	Sub-Saharan Africa					
12	BEL	BE	Belgium	Europe	Western Europe	Europe and Central Asia					
13	BEN	BJ	Benin	Africa	Western Africa	Sub-Saharan Africa					
14	BFA	BF	Burkina Faso	Africa	Western Africa	Sub-Saharan Africa					
15	BGD	BD	Bangladesh	Asia	Southern Asia	South Asia					

Insert the COUNTIFS function

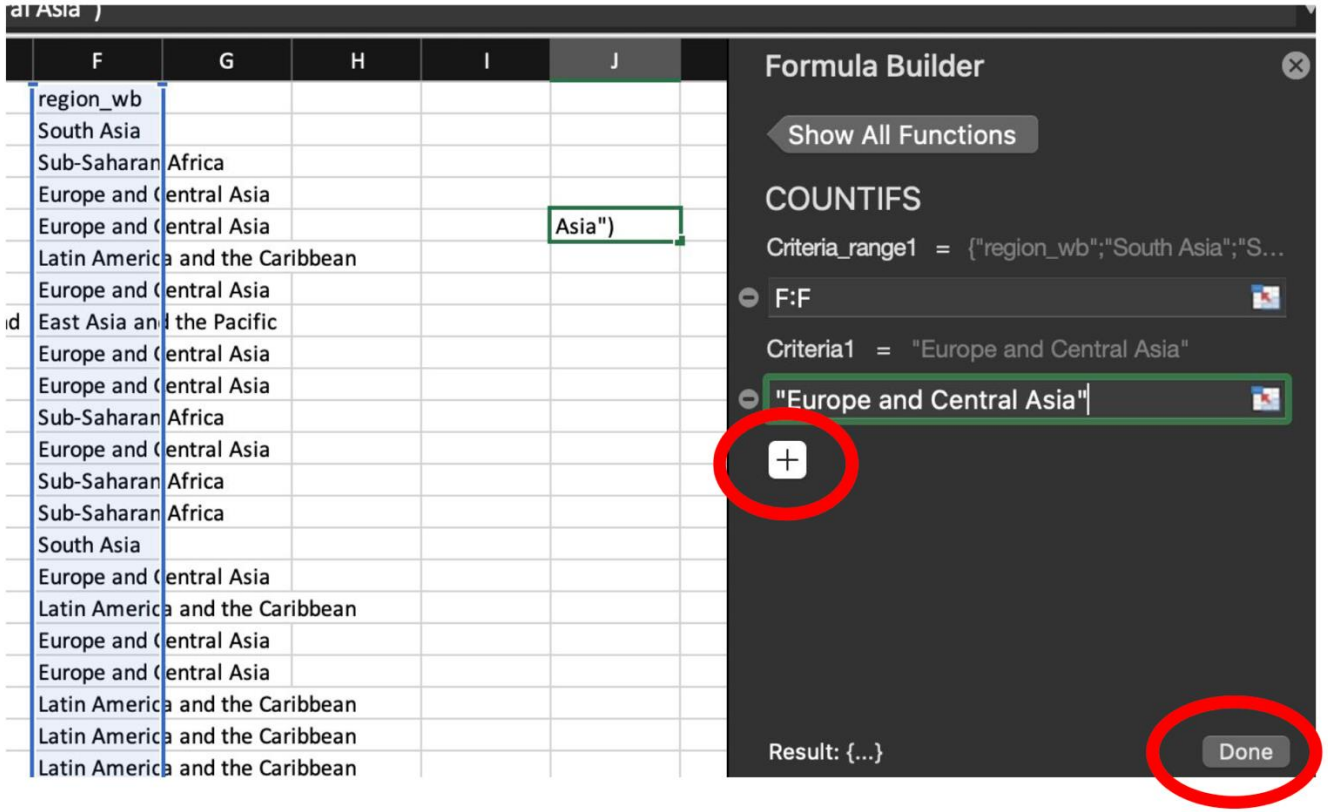


A	B	C	D	E	F
c3	c2	country	region_un	region_un_sub	region_wb
AFG	AF	Afghanistan	Asia	Southern Asia	South Asia
AGO	AO	Angola	Africa	Middle Africa	Sub-Saharan Africa
ALB	AL	Albania	Europe	Southern Europe	Europe and Central Asia
AND	AD	Andorra	Europe	Southern Europe	Europe and Central Asia
ARG	AR	Argentina	Americas	South America	Latin America and the Caribbean
ARM	AM	Armenia	Asia	Western Asia	Europe and Central Asia
AUS	AU	Australia	Oceania	Australia and New Zealand	East Asia and the Pacific
AUT	AT	Austria	Europe	Western Europe	Europe and Central Asia
AZE	AZ	Azerbaijan	Asia	Western Asia	Europe and Central Asia
BDI	BI	Burundi	Africa	Eastern Africa	Sub-Saharan Africa
BEL	BE	Belgium	Europe	Western Europe	Europe and Central Asia
BEN	BJ	Benin	Africa	Western Africa	Sub-Saharan Africa
BFA	BF	Burkina Faso	Africa	Western Africa	Sub-Saharan Africa
BGD	BD	Bangladesh	Asia	Southern Asia	South Asia
BGR	BG	Bulgaria	Europe	Eastern Europe	Europe and Central Asia
BHS	BS	Bahamas, The	Americas	Caribbean	Latin America and the Caribbean
BIH	BA	Bosnia and Herzegovina	Europe	Southern Europe	Europe and Central Asia
BLR	BY	Belarus	Europe	Eastern Europe	Europe and Central Asia
BLZ	BZ	Belize	Americas	Central America	Latin America and the Caribbean
BOL	BO	Bolivia	Americas	South America	Latin America and the Caribbean
BRA	BR	Brazil	Americas	South America	Latin America and the Caribbean
BRB	BB	Barbados	Americas	Caribbean	Latin America and the Caribbean
BTN	BT	Bhutan	Asia	Southern Asia	South Asia
BWA	BW	Botswana	Africa	Southern Africa	Sub-Saharan Africa
CAF	CF	Central African Republic	Africa	Middle Africa	Sub-Saharan Africa
CAN	CA	Canada	Americas	Northern America	North America
CHE	CH	Switzerland	Europe	Western Europe	Europe and Central Asia
CHL	CL	Chile	Americas	South America	Latin America and the Caribbean
CHN	CN	China	Asia	Eastern Asia	East Asia and the Pacific
CIV	CI	Cote d'Ivoire	Africa	Western Africa	Sub-Saharan Africa
CMR	CM	Cameroon	Africa	Middle Africa	Sub-Saharan Africa
COD	CD	Congo, Democratic Republic	Africa	Middle Africa	Sub-Saharan Africa
COG	CG	Congo, Republic of the	Africa	Middle Africa	Sub-Saharan Africa
COL	CO	Colombia	Americas	South America	Latin America and the Caribbean

Select the ”region_wb” column by clicking on F or type in F:F to the Criteria_range1 input box



Then click on the “+” button to add more arguments. Under Criteria 1, enter “Europe and Central Asia”
Finally, click Done to reveal the result
Result: I got 54, what did you get?



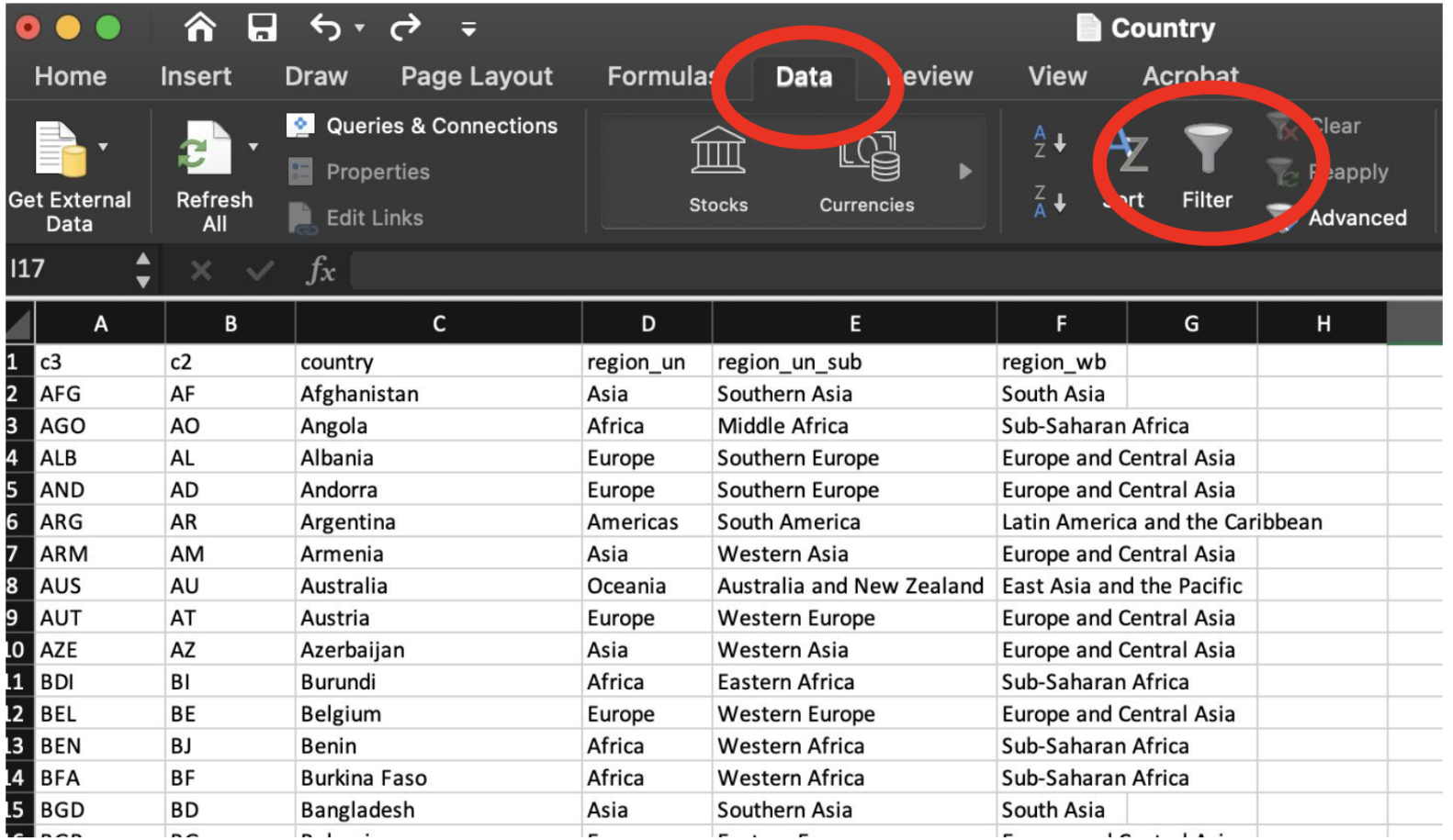
Delete the cell with the solution as we no longer need it

Q2: How many countries are in both the World Bank region (column region_wb) "Middle East and North Africa" and ALSO **in the United Regions subregion (column region_un_sub) called "Western Asia"**

Method:

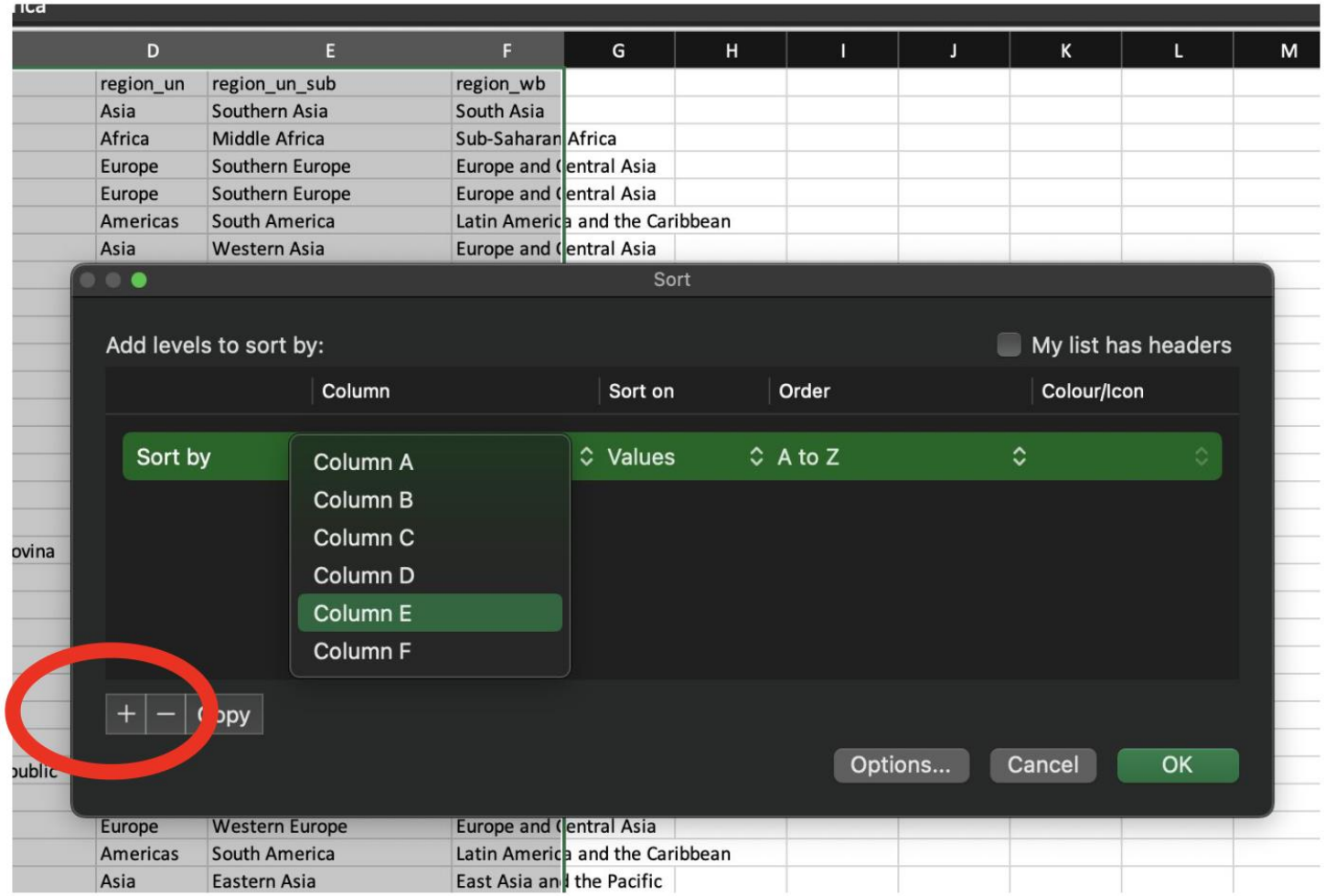
By the way, this is an interesting case where the geographic groupings from the two organizations are quite different: neither is a subset of the other. Do this in *at least one* of the following ways: by sorting using two columns (one as primary ordering, and then another to break ties); do this by a formula using COUNTIFS; do it by creating an auxiliary column which can be summed.

Select any cell with data (not empty)
Go to the ”Data” tab and click on Filter

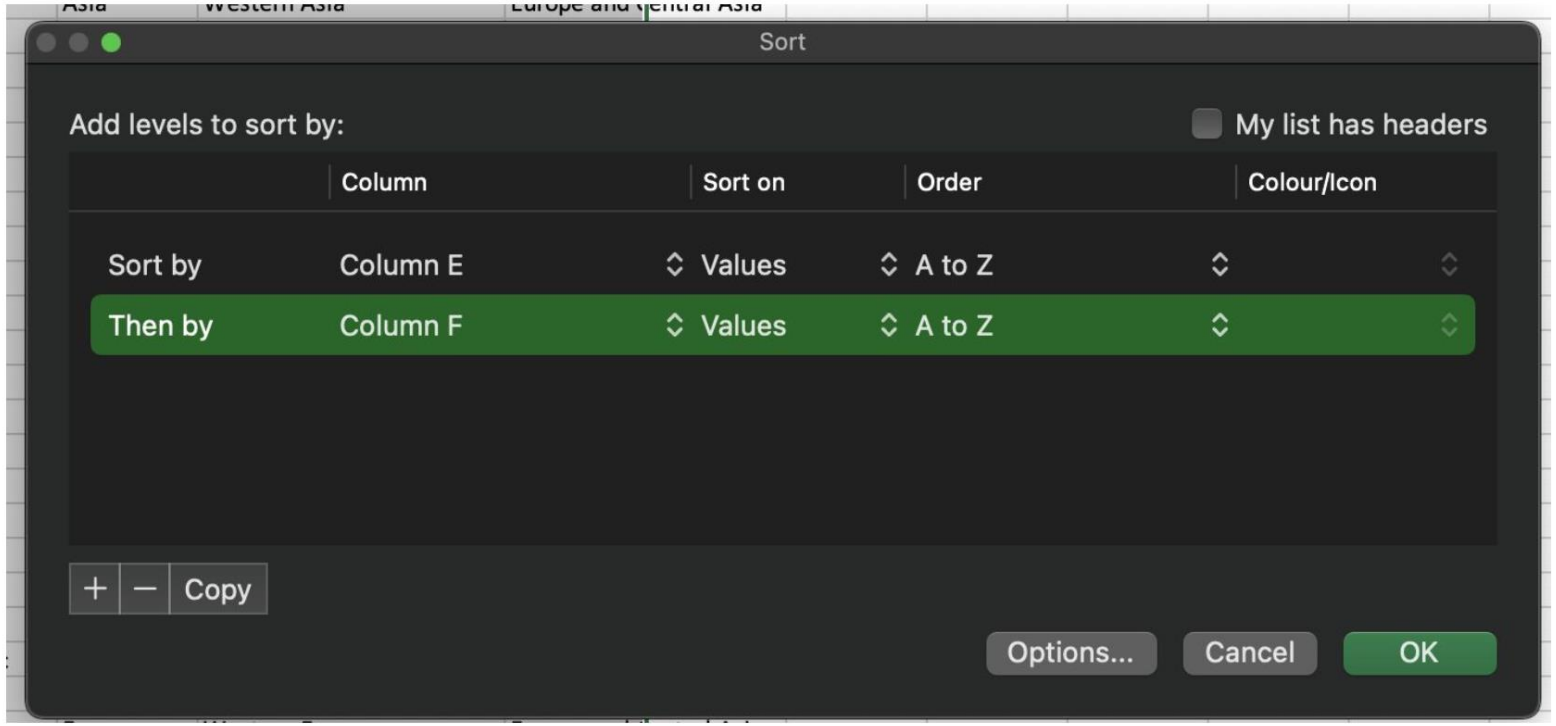


	A	B	C	D	E	F	G	H
1	c3	c2	country	region_un	region_un_sub	region_wb		
2	AFG	AF	Afghanistan	Asia	Southern Asia	South Asia		
3	AGO	AO	Angola	Africa	Middle Africa	Sub-Saharan Africa		
4	ALB	AL	Albania	Europe	Southern Europe	Europe and Central Asia		
5	AND	AD	Andorra	Europe	Southern Europe	Europe and Central Asia		
6	ARG	AR	Argentina	Americas	South America	Latin America and the Caribbean		
7	ARM	AM	Armenia	Asia	Western Asia	Europe and Central Asia		
8	AUS	AU	Australia	Oceania	Australia and New Zealand	East Asia and the Pacific		
9	AUT	AT	Austria	Europe	Western Europe	Europe and Central Asia		
10	AZE	AZ	Azerbaijan	Asia	Western Asia	Europe and Central Asia		
11	BDI	BI	Burundi	Africa	Eastern Africa	Sub-Saharan Africa		
12	BEL	BE	Belgium	Europe	Western Europe	Europe and Central Asia		
13	BEN	BJ	Benin	Africa	Western Africa	Sub-Saharan Africa		
14	BFA	BF	Burkina Faso	Africa	Western Africa	Sub-Saharan Africa		
15	BGD	BD	Bangladesh	Asia	Southern Asia	South Asia		

Select column E , “region_un_sub”
And then click the “+” button so that you can sort by a second column as well



Then select column F , “region_wb”. Finally, click OK



We sorted by column E first, and we want to only see countries which is classified as “Western Asia” in this category.
Then look at the subset of these countries which is also classified as “Middle East and North Africa” in the column F.
Result: I counted 10 countries matching this. How about you?

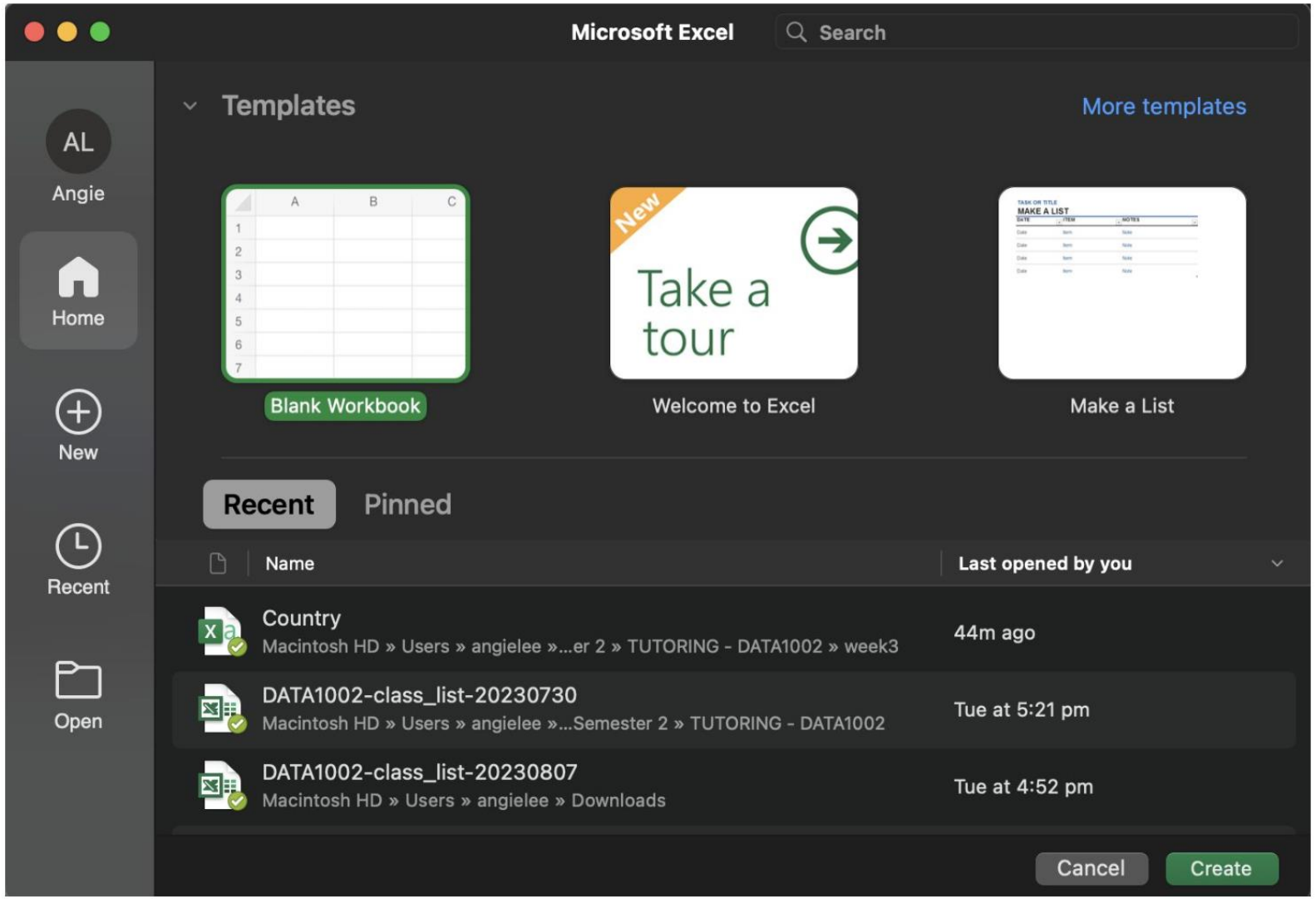
168	TGO	TG	Togo	Africa	Western Africa	Sub-Saharan Africa
169	ARM	AM	Armenia	Asia	Western Asia	Europe and Central Asia
170	AZE	AZ	Azerbaijan	Asia	Western Asia	Europe and Central Asia
171	CYP	CY	Cyprus	Asia	Western Asia	Europe and Central Asia
172	GEO	GE	Georgia	Asia	Western Asia	Europe and Central Asia
173	TUR	TR	Turkey	Asia	Western Asia	Europe and Central Asia
174	IRQ	IQ	Iraq	Asia	Western Asia	Middle East and North Africa
175	ISR	IL	Israel	Asia	Western Asia	Middle East and North Africa
176	JOR	JO	Jordan	Asia	Western Asia	Middle East and North Africa
177	LBN	LB	Lebanon	Asia	Western Asia	Middle East and North Africa
178	OMN	OM	Oman	Asia	Western Asia	Middle East and North Africa
179	QAT	QA	Qatar	Asia	Western Asia	Middle East and North Africa
180	SAU	SA	Saudi Arabia	Asia	Western Asia	Middle East and North Africa
181	SYR	SY	Syria	Asia	Western Asia	Middle East and North Africa
182	WBG	YU	West Bank and Gaza	Asia	Western Asia	Middle East and North Africa
183	YEM	YE	Yemen	Asia	Western Asia	Middle East and North Africa
184	AUT	AT	Austria	Europe	Western Europe	Europe and Central Asia
185	BEL	BE	Belgium	Europe	Western Europe	Europe and Central Asia
186	CHE	CH	Switzerland	Europe	Western Europe	Europe and Central Asia
187	DEU	DE	Germany	Europe	Western Europe	Europe and Central Asia
188	FRA	FR	France	Europe	Western Europe	Europe and Central Asia
189	LUX	LU	Luxembourg	Europe	Western Europe	Europe and Central Asia
190	NLD	NL	Netherlands	Europe	Western Europe	Europe and Central Asia
191						

Return the spreadsheet to how it was before sorting

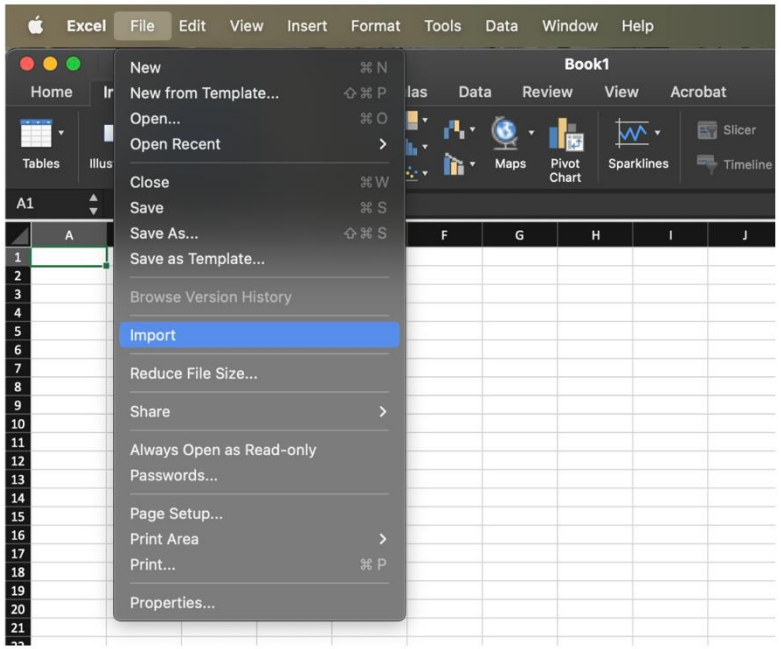
Further Setup

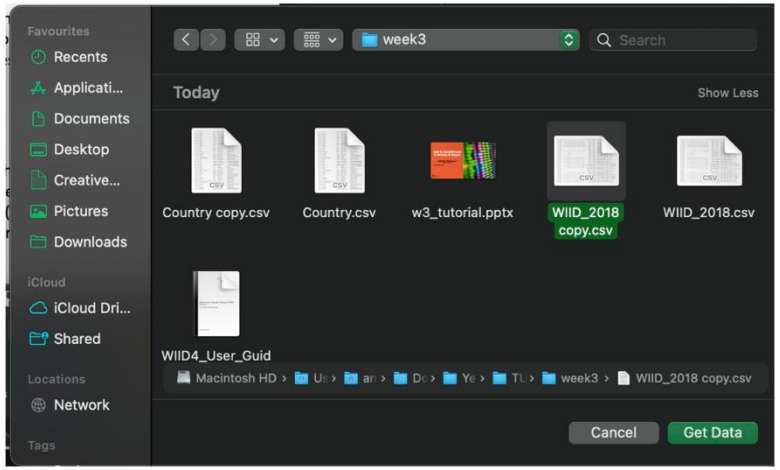
For the next stages of this analysis, create a new empty spreadsheet in whichever tool you use, then import the data from WIID_2018.csv into one worksheet, and then make a second worksheet within the spreadsheet, and put into that second sheet the data from Country.csv (for example, you might copy-paste the data from where you have it already open, into the correct place in the new spreadsheet). Make sure each worksheet has a sensible informative name.

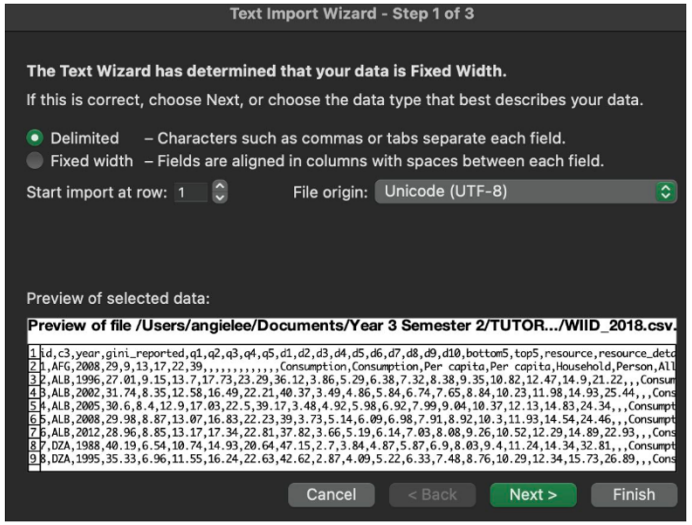
Create a new “workbook” using Microsoft Excel

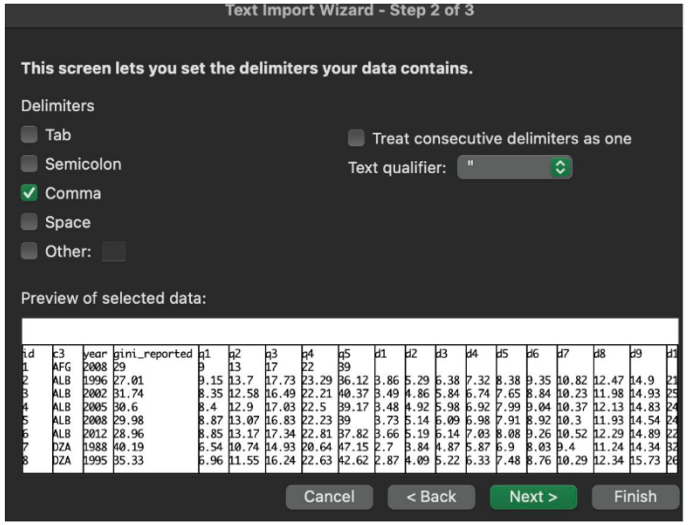


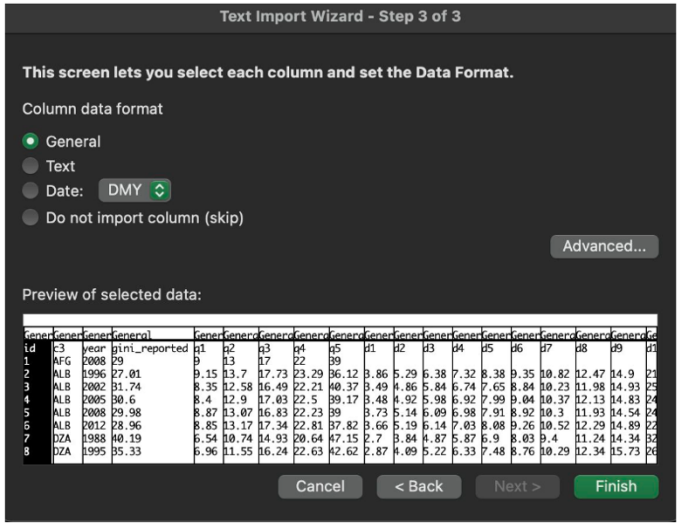
Import the data from WIID_2018.csv into one worksheet

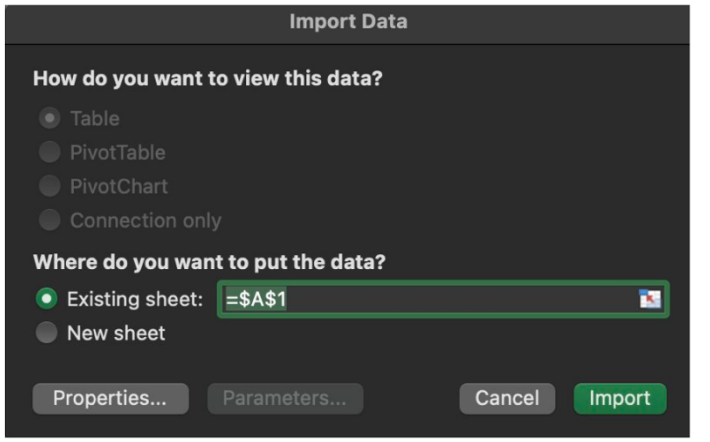


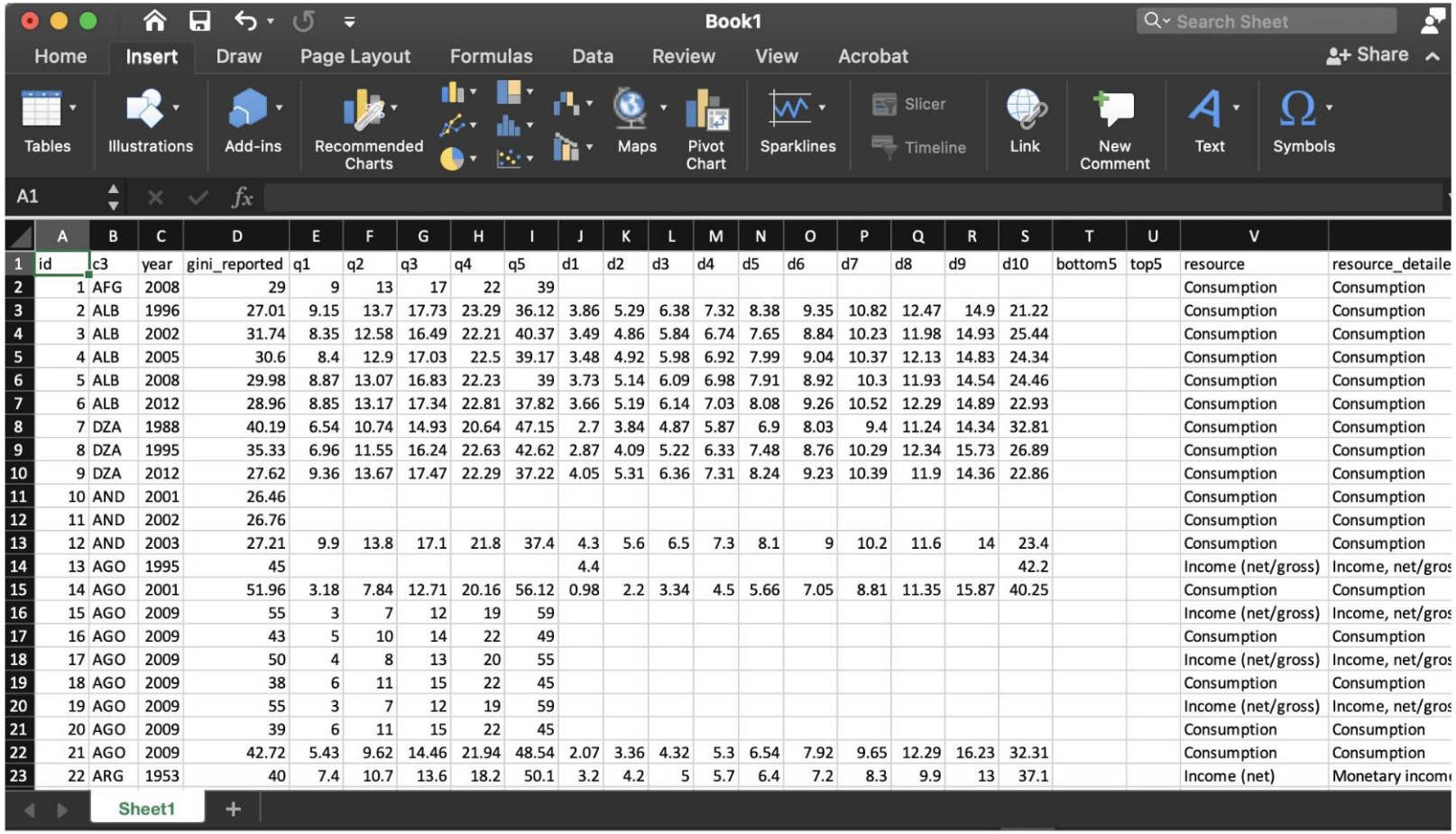



Text Import Wizard - Step 1 of 3
The Text Wizard has determined that your data is Fixed Width.
If this is correct, choose Next, or choose the data type that best describes your data.
Delimited – Characters such as commas or tabs separate each field.
Fixed width – Fields are aligned in columns with spaces between each field.
Start import at row: 1
File origin: Unicode (UTF-8)
Preview of selected data:
Preview of file /Users/angielee/Documents/Year 3 Semester 2/TUTOR.../WIID_2018.csv.
Cancel
< Back
Next >
Finish


Text Import Wizard - Step 2 of 3
This screen lets you set the delimiters your data contains.
Delimiters
Tab
Semicolon
Comma
Space
Other:
Treat consecutive delimiters as one
Text qualifier: "
Preview of selected data:
Cancel
< Back
Next >
Finish

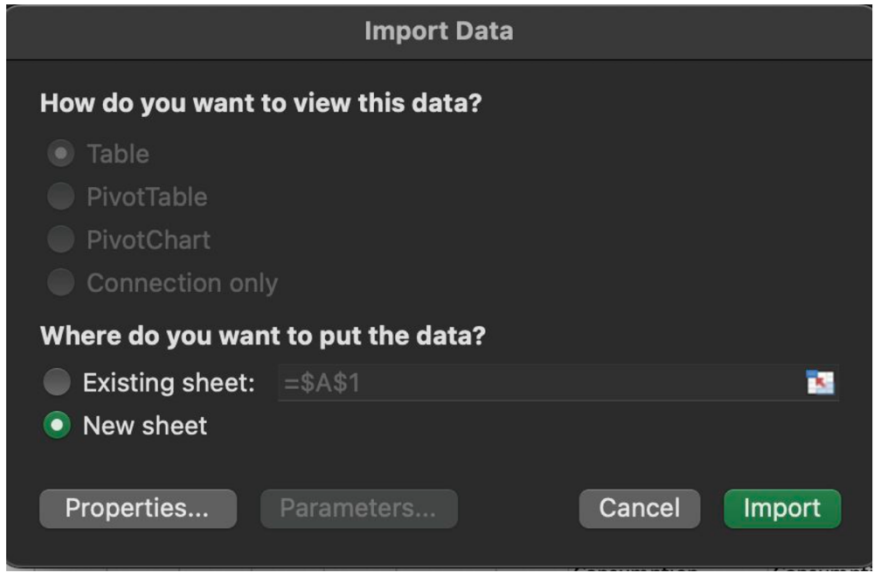

Text Import Wizard - Step 3 of 3
This screen lets you select each column and set the Data Format.
Column data format
General
Text
Date: DMY
Do not import column (skip)
Advanced...
Preview of selected data:
Cancel
< Back
Next >
Finish


Import Data
How do you want to view this data?
Table
PivotTable
PivotChart
Connection only
Where do you want to put the data?
Existing sheet: =$A$1
New sheet
Properties...
Parameters...
Cancel
Import

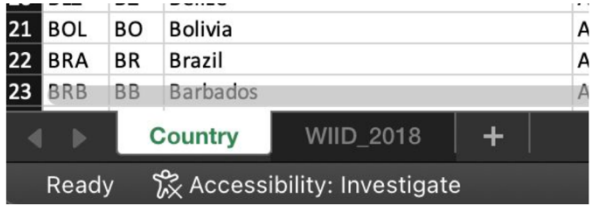


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	
1	id	c3	year	gini_reported	q1	q2	q3	q4	q5	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	bottom5	top5	resource	resource_detaile
2	1	AFG	2008	29	9	13	17	22	39													Consumption	Consumption
3	2	ALB	1996	27.01	9.15	13.7	17.73	23.29	36.12	3.86	5.29	6.38	7.32	8.38	9.35	10.82	12.47	14.9	21.22			Consumption	Consumption
4	3	ALB	2002	31.74	8.35	12.58	16.49	22.21	40.37	3.49	4.86	5.84	6.74	7.65	8.84	10.23	11.98	14.93	25.44			Consumption	Consumption
5	4	ALB	2005	30.6	8.4	12.9	17.03	22.5	39.17	3.48	4.92	5.98	6.92	7.99	9.04	10.37	12.13	14.83	24.34			Consumption	Consumption
6	5	ALB	2008	29.98	8.87	13.07	16.83	22.23	39	3.73	5.14	6.09	6.98	7.91	8.92	10.3	11.93	14.54	24.46			Consumption	Consumption
7	6	ALB	2012	28.96	8.85	13.17	17.34	22.81	37.82	3.66	5.19	6.14	7.03	8.08	9.26	10.52	12.29	14.89	22.93			Consumption	Consumption
8	7	DZA	1988	40.19	6.54	10.74	14.93	20.64	47.15	2.7	3.84	4.87	5.87	6.9	8.03	9.4	11.24	14.34	32.81			Consumption	Consumption
9	8	DZA	1995	35.33	6.96	11.55	16.24	22.63	42.62	2.87	4.09	5.22	6.33	7.48	8.76	10.29	12.34	15.73	26.89			Consumption	Consumption
10	9	DZA	2012	27.62	9.36	13.67	17.47	22.29	37.22	4.05	5.31	6.36	7.31	8.24	9.23	10.39	11.9	14.36	22.86			Consumption	Consumption
11	10	AND	2001	26.46																		Consumption	Consumption
12	11	AND	2002	26.76																		Consumption	Consumption
13	12	AND	2003	27.21	9.9	13.8	17.1	21.8	37.4	4.3	5.6	6.5	7.3	8.1	9	10.2	11.6	14	23.4			Consumption	Consumption
14	13	AGO	1995	45						4.4									42.2			Income (net/gross)	Income, net/gros
15	14	AGO	2001	51.96	3.18	7.84	12.71	20.16	56.12	0.98	2.2	3.34	4.5	5.66	7.05	8.81	11.35	15.87	40.25			Consumption	Consumption
16	15	AGO	2009	55	3	7	12	19	59													Income (net/gross)	Income, net/gros
17	16	AGO	2009	43	5	10	14	22	49													Consumption	Consumption
18	17	AGO	2009	50	4	8	13	20	55													Income (net/gross)	Income, net/gros
19	18	AGO	2009	38	6	11	15	22	45													Consumption	Consumption
20	19	AGO	2009	55	3	7	12	19	59													Income (net/gross)	Income, net/gros
21	20	AGO	2009	39	6	11	15	22	45													Consumption	Consumption
22	21	AGO	2009	42.72	5.43	9.62	14.46	21.94	48.54	2.07	3.36	4.32	5.3	6.54	7.92	9.65	12.29	16.23	32.31			Consumption	Consumption
23	22	ARG	1953	40	7.4	10.7	13.6	18.2	50.1	3.2	4.2	5	5.7	6.4	7.2	8.3	9.9	13	37.1			Income (net)	Monetary incom

Import data from Country.csv with the same settings, but this time in the final dialogue box choose “New sheet”



You can rename each sheet to something more helpful



Q3: Index & Match

In the original dataset (which you can download from https://www.wider.unu.edu/database/world-income-inequality-database-wiid4) there is actually a single file in which all the country information (the three character code for the country, the two-character code, the country name, the UN region, the UN subregion, and the World bank region) are present *in each row* along with the year and the various statistics such as gini coefficient. In the first worksheet (where the WIID data is), let's recreate this structure: **create new columns immediately after c3, one column for each of c2, country, region_un, region_un_sub, region_wb**. Produce a formula based around INDEX and MATCH functions as described in lecture 2B, to put into the region_wb column the correct region name that applies for each row, by looking up the c3 code in the Country worksheet.

Content-dependent retrieval

- A spreadsheet may contain more than one table- like region, perhaps on different worksheets
- We may want to write a formula in one cell, that looks-up a value from a cell whose exact location is not known, but varies with the data in a table
- It may come from whatever row of the referenced table, has a particular identifier in a given column
- Use the MATCH function, together with INDEX

Country and WIID_2018 each has a c3 column which contains a three-letter country code for each row.
We can use this to match rows in both tables together.

B1 c3

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	c3	year	gini_reported	q1	q2	q3	q4	q5	d1	d2	d3	d4
2	1	AFG	2008	29	9	13	17	22	39				
3	2	ALB	1996	27.01	9.15	13.7	17.73	23.29	36.12	3.86	5.29	6.38	7.32
4	3	ALB	2002	31.74	8.35	12.58	16.49	22.21	40.37	3.49	4.86	5.84	6.74
5	4	ALB	2005	30.6	8.4	12.9	17.03	22.5	39.17	3.48	4.92	5.98	6.92
6	5	ALB	2008	29.98	8.87	13.07	16.83	22.23	39	3.73	5.14	6.09	6.98
7	6	ALB	2012	28.96	8.85	13.17	17.34	22.81	37.82	3.66	5.19	6.14	7.03
8	7	DZA	1988	40.19	6.54	10.74	14.93	20.64	47.15	2.7	3.84	4.87	5.87
9	8	DZA	1995	35.33	6.96	11.55	16.24	22.63	42.62	2.87	4.09	5.22	6.33
10	9	DZA	2012	27.62	9.36	13.67	17.47	22.29	37.22	4.05	5.31	6.36	7.31
11	10	AND	2001	26.46									
12	11	AND	2002	26.76									
13	12	AND	2003	27.21	9.9	13.8	17.1	21.8	37.4	4.3	5.6	6.5	7.3
14	13	AGO	1995	45						4.4			
15	14	AGO	2001	51.96	3.18	7.84	12.71	20.16	56.12	0.98	2.2	3.34	4.5
16	15	AGO	2009	55	3	7	12	19	59				
17	16	AGO	2009	43	5	10	14	22	49				
18	17	AGO	2009	50	4	8	13	20	55				
19	18	AGO	2009	38	6	11	15	22	45				
20	19	AGO	2009	55	3	7	12	19	59				
21	20	AGO	2009	39	6	11	15	22	45				
22	21	AGO	2009	42.72	5.43	9.62	14.46	21.94	48.54	2.07	3.36	4.32	5.3
23	22	ARG	1953	40	7.4	10.7	13.6	18.2	50.1	3.2	4.2	5	5.7

Country | WIID_2018

A1 c3

	A	B	C	D	E	F
1	c3	c2	country	region_un	region_un_sub	region_wb
2	AFG	AF	Afghanistan	Asia	Southern Asia	South Asia
3	AGO	AO	Angola	Africa	Middle Africa	Sub-Saharan Africa
4	ALB	AL	Albania	Europe	Southern Europe	Europe and Central Asia
5	AND	AD	Andorra	Europe	Southern Europe	Europe and Central Asia
6	ARG	AR	Argentina	Americas	South America	Latin America and the Caribbean
7	ARM	AM	Armenia	Asia	Western Asia	Europe and Central Asia
8	AUS	AU	Australia	Oceania	Australia and New Zealand	East Asia and the Pacific
9	AUT	AT	Austria	Europe	Western Europe	Europe and Central Asia
10	AZE	AZ	Azerbaijan	Asia	Western Asia	Europe and Central Asia
11	BDI	BI	Burundi	Africa	Eastern Africa	Sub-Saharan Africa
12	BEL	BE	Belgium	Europe	Western Europe	Europe and Central Asia
13	BEN	BJ	Benin	Africa	Western Africa	Sub-Saharan Africa
14	BFA	BF	Burkina Faso	Africa	Western Africa	Sub-Saharan Africa
15	BGD	BD	Bangladesh	Asia	Southern Asia	South Asia
16	BGR	BG	Bulgaria	Europe	Eastern Europe	Europe and Central Asia
17	BHS	BS	Bahamas, The	Americas	Caribbean	Latin America and the Caribbean
18	BIH	BA	Bosnia and Herzegovina	Europe	Southern Europe	Europe and Central Asia

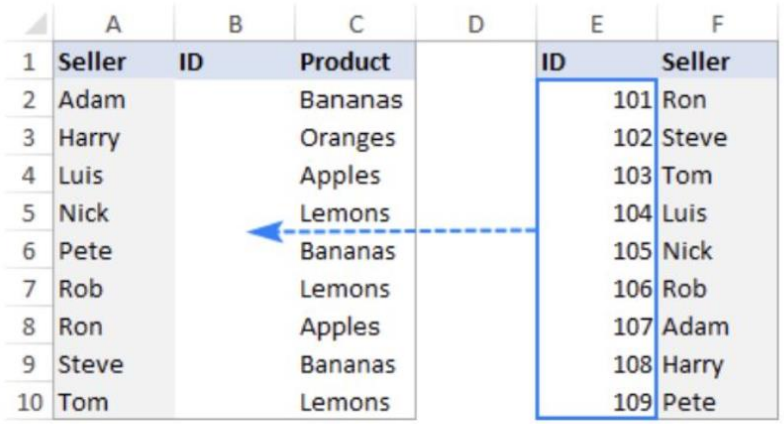
How to merge tables in Excel with INDEX MATCH

If you are looking for a more powerful and versatile alternative to the VLOOKUP function, embrace this INDEX MATCH combination:

```
INDEX (return_range, MATCH (lookup_value, lookup_range, 0))
```

The syntax is explained in detail in this tutorial: INDEX / MATCH in Excel. And here I will show you how to use this formula to **look up from right to left**, something that VLOOKUP is unable to do.

Let's say you have another lookup table with order IDs in the first column and you wish to copy those IDs to the main table by matching the seller names. For better visualization, both tables are put on the same sheet:



	A	B	C	D	E	F
1	**Seller**	**ID**	**Product**		**ID**	**Seller**
2	Adam		Bananas		101	Ron
3	Harry		Oranges		102	Steve
4	Luis		Apples		103	Tom
5	Nick		Lemons		104	Luis
6	Pete		Bananas		105	Nick
7	Rob		Lemons		106	Rob
8	Ron		Apples		107	Adam
9	Steve		Bananas		108	Harry
10	Tom		Lemons		109	Pete

To accomplish the task, you supply the following arguments to the Index Match formula:

- *Return_range* - $E$2:$E$10
- *Lookup_value* - $A2
- *Lookup_range* - $F$2:$F$10

Please notice the $ sign that locks the ranges to prevent them from changing as you copy the formula down the table:

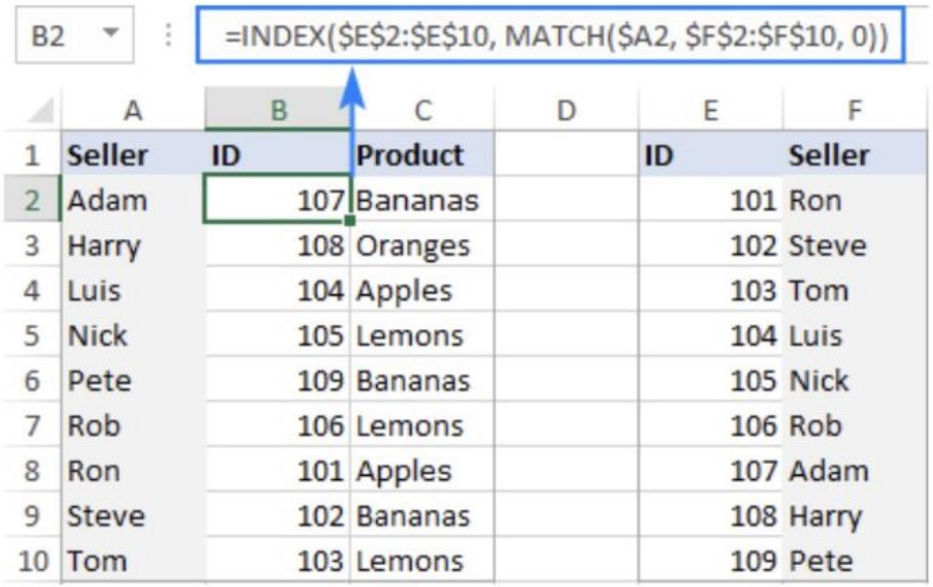
The completed formula looks as follows:

```
=INDEX($E$2:$E$10, MATCH($A2, $F$2:$F$10, 0))
```

The completed formula looks as follows:

```
=INDEX($E$2:$E$10, MATCH($A2, $F$2:$F$10, 0))
```

...and combines data from two tables perfectly:



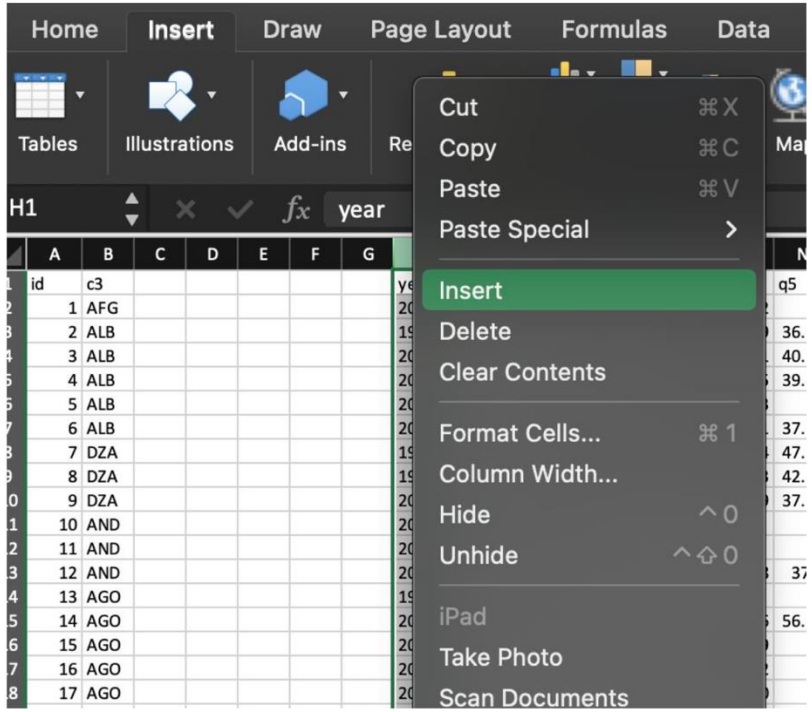
	A	B	C	D	E	F
1	**Seller**	**ID**	**Product**		**ID**	**Seller**
2	Adam	107	Bananas		101	Ron
3	Harry	108	Oranges		102	Steve
4	Luis	104	Apples		103	Tom
5	Nick	105	Lemons		104	Luis
6	Pete	109	Bananas		105	Nick
7	Rob	106	Lemons		106	Rob
8	Ron	101	Apples		107	Adam
9	Steve	102	Bananas		108	Harry
10	Tom	103	Lemons		109	Pete

In Excel 365, you can use the new XLOOKUP function for the same purpose:

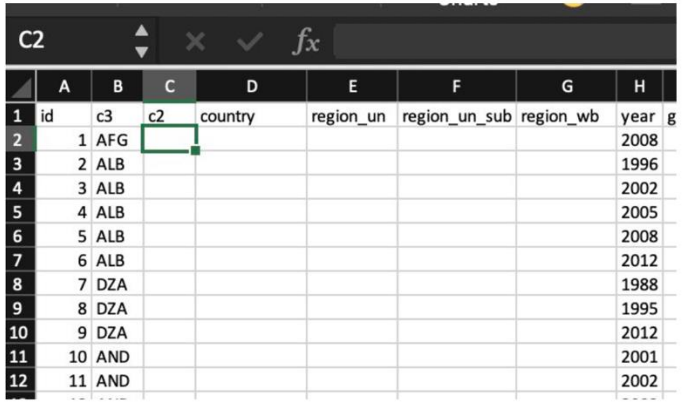
```
=XLOOKUP(A2, $F$2:$F$10, $E$2:$E$10, "Not found")
```

Helpful guide: https://www.ablebits.com/office-addins-blog/excel-merge-tables-matching-columns/

Insert five columns next to c3

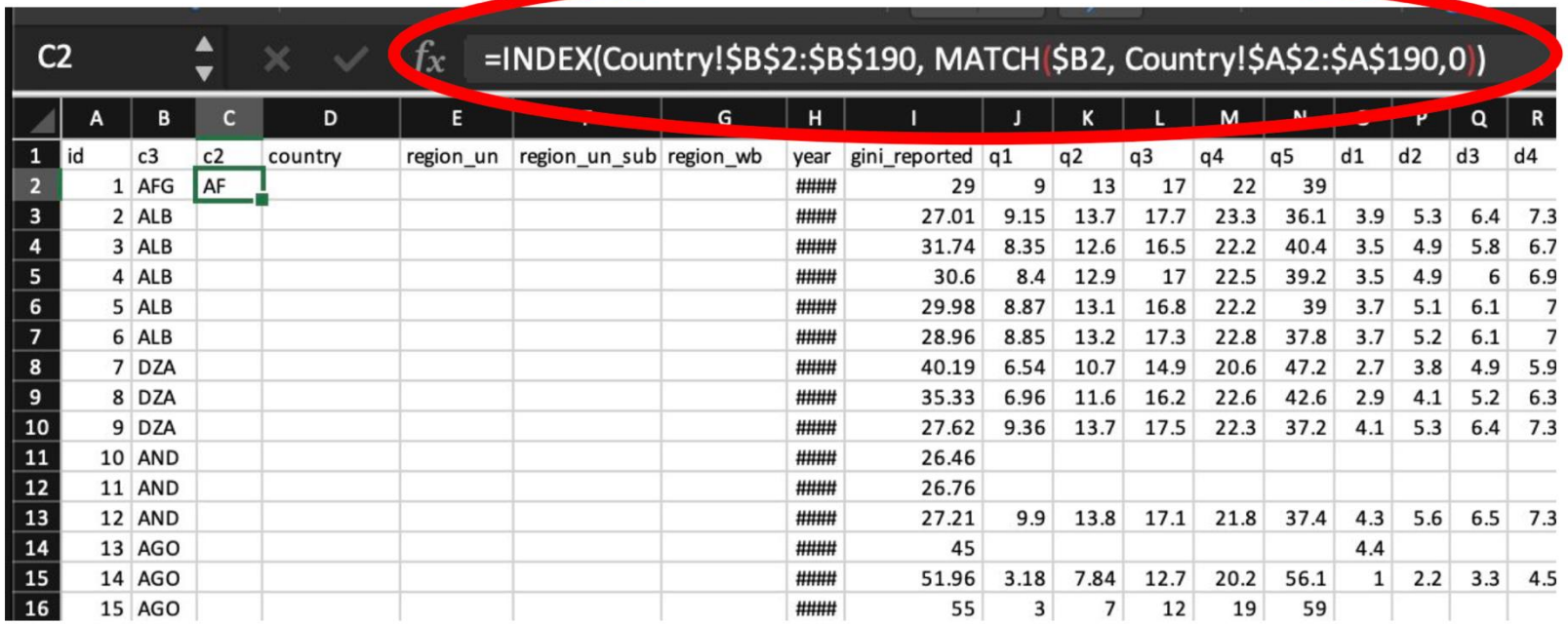


Give each column the appropriate name for our benefit
Then click on the first empty cell



	A	B	C	D	E	F	G	H
1	id	c3	c2	country	region_un	region_un_sub	region_wb	year
2	1	AFG						2008
3	2	ALB						1996
4	3	ALB						2002
5	4	ALB						2005
6	5	ALB						2008
7	6	ALB						2012
8	7	DZA						1988
9	8	DZA						1995
10	9	DZA						2012
11	10	AND						2001
12	11	AND						2002

Enter the following formula then press enter to see the result
If you check in Country.csv, c3=AFG does correspond to c2=AF



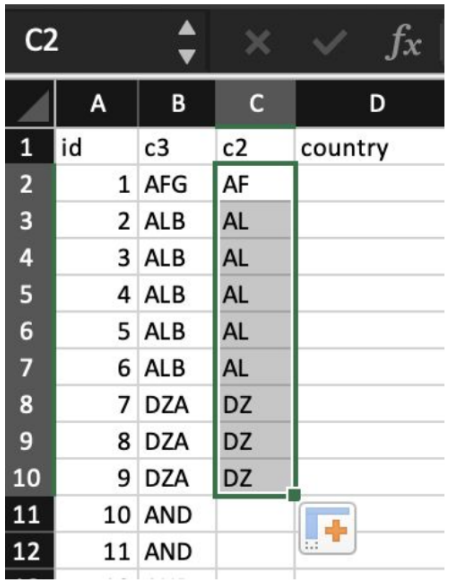
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	id	c3	c2	country	region_un	region_un_sub	region_wb	year	gini_reported	q1	q2	q3	q4	q5	d1	d2	d3	d4
2	1	AFG	AF					####	29	9	13	17	22	39				
3	2	ALB						####	27.01	9.15	13.7	17.7	23.3	36.1	3.9	5.3	6.4	7.3
4	3	ALB						####	31.74	8.35	12.6	16.5	22.2	40.4	3.5	4.9	5.8	6.7
5	4	ALB						####	30.6	8.4	12.9	17	22.5	39.2	3.5	4.9	6	6.9
6	5	ALB						####	29.98	8.87	13.1	16.8	22.2	39	3.7	5.1	6.1	7
7	6	ALB						####	28.96	8.85	13.2	17.3	22.8	37.8	3.7	5.2	6.1	7
8	7	DZA						####	40.19	6.54	10.7	14.9	20.6	47.2	2.7	3.8	4.9	5.9
9	8	DZA						####	35.33	6.96	11.6	16.2	22.6	42.6	2.9	4.1	5.2	6.3
10	9	DZA						####	27.62	9.36	13.7	17.5	22.3	37.2	4.1	5.3	6.4	7.3
11	10	AND						####	26.46									
12	11	AND						####	26.76									
13	12	AND						####	27.21	9.9	13.8	17.1	21.8	37.4	4.3	5.6	6.5	7.3
14	13	AGO						####	45						4.4			
15	14	AGO						####	51.96	3.18	7.84	12.7	20.2	56.1	1	2.2	3.3	4.5
16	15	AGO						####	55	3	7	12	19	59				

Return Range (C2)
i.e. results of Index Match is pulled from these cells

Look Up Range (C3)
i.e. look for C3 on WIID_2018 in this selected range

```
=INDEX (Country!B$2:B$190, MATCH(WIID_2018!$B2, Country!$A$2:$A$190,0))
```

Look Up Value (C3)
i.e. look for this value in the Country sheet



Drag the current formula down to the cells below to apply it

Try applying this same formula in order to fill in the columns: country, region_un and so on
Hint: all you really have to change is the cell selection highlighted below

=INDEX(Country!$B$2:$B$190, MATCH($B2, Country!$A$2:$A$190,0))

Lab Activities

Analyse Data & Discuss

Discussion 1: Tools and Approaches

How similar or different were your…

Approaches

Tools

Advantages

Disadvantages

Discussion 2: Index & Match

Advantages

Disadvantages

why it is not appropriate to put some properties of the country, such as population or income level, in the Country.csv file?

Exam-Style Questions

Question 1:

Discuss the role of string manipulation in data extraction and cleaning in the context of a data science project. Provide examples of common string operations used.

Question 2:

Evaluate the importance of tracing and debugging in the development of data science algorithms. How do these practices contribute to the reliability and accuracy of a data science project?

Exam-Style Questions (Q1)

String manipulation is crucial in data extraction and cleaning as it helps transform raw text data into a structured format suitable for analysis. Common string operations include:

- **Splitting Strings:** Using methods like `split()` to divide a string into a list based on a delimiter. For instance, splitting a CSV line into individual fields.
- **Stripping Strings**: Removing unwanted whitespace or characters using methods like `strip()`, `lstrip()`, or `rstrip()`. This is essential for standardising data entries.
- **Replacing Substrings:** Using `replace()` to substitute specific substrings with others, such as correcting misspellings or standardising text formats.
- **Concatenation:** Combining strings using the `+` operator or `join()` method to form new strings, which is useful for creating new fields from existing data. These operations help clean and prepare textual data, ensuring that it is consistent and ready for further analysis.

Exam-Style Questions (Q2)

Tracing and debugging are critical in developing reliable and accurate data science algorithms. Tracing involves following the flow of execution in the code to understand how data is processed and transformed. This practice helps identify logical errors and inefficiencies.

Debugging, on the other hand, involves finding and fixing errors in the code. Tools like print statements, logging, and debugging environments allow data scientists to monitor variable states and control flow. These practices ensure that algorithms perform as expected and produce accurate results. By systematically tracing and debugging code, data scientists can uncover hidden issues, such as incorrect data handling or flawed logic, ultimately improving the robustness and reliability of their models.

That’s it folks!

Remaining Ed Lessons, questions, etc.