Basic Spreadsheet Data Processing
S4 Summer 2021 GIS Institute

Frank Donnelly, GIS & Data Librarian

Introduction

This tutorial introduces a selection of basic data processing tasks with spreadsheets, geared towards GIS users. Data tables must follow strict formatting rules in order to be used in GIS, and a substantive amount of time is often needed for processing attribute data in preparation for mapping and analysis. Spreadsheets are powerful and accessible tools that are practical for processing data for small projects. These examples were written using MS Excel 2019, but can be replicated in other packages. Basic spreadsheet knowledge (selecting columns and rows, copying and pasting, saving files) is presumed. Before we begin, download the tutorial data file `data_processing_tutorial.zip` from the Canvas GIS Institute page, save it on your computer, and unzip the file.

Conventions used in this tutorial:

- Names of windows, tabs, and tools appear in *italic face*.
- Names of folders, files, and columns appear in `typewriter face`.
- Button-clicking steps are chained together as: *Tab on Ribbon - Button - Menu Option*

In this tutorial we will type formulas (also called functions) directly into cells. For help with formulas, you can insert them using the *fx* function button located above the cell block.

Notes on Sample Data

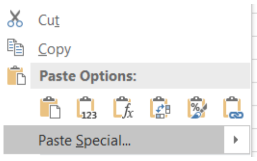
The sample data is a global spreadsheet of copper smelters created by USGS researchers in 2003 (metadata is included with the spreadsheet). This dataset was selected because it presents a number of common data processing challenges; like many spreadsheets it is formatted for display rather than data analysis. We will reformat this data with the intention of plotting the smelters in GIS using their coordinates, and summarizing it by country to join it to a country shapefile for thematic mapping. Additional data includes a country boundary file from *Natural Earth* `https://www.naturalearthdata.com/`, and a CSV file of ISO 3166 country codes.

1 Copy - Paste Special - Values

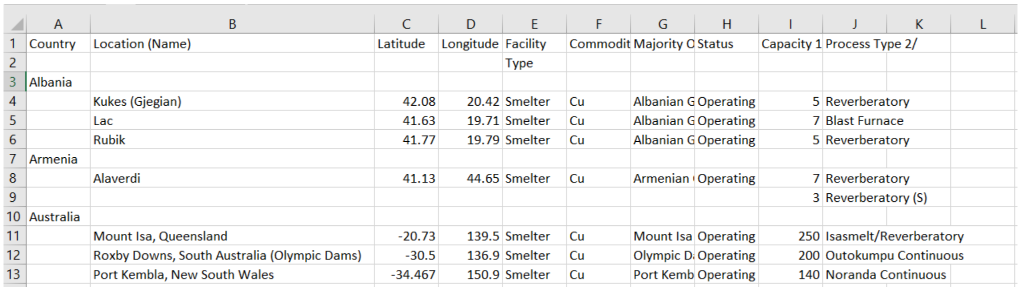
Never work directly in the original file, in case you need to refer back to it. Make a copy; you could copy the entire file, but in this case we will combine this with an additional step of removing formatting. The *Copy - Paste Special - Values* process is one you will use over and over again. Use it to eliminate formatting and to replace formulas with the actual values output from the formula.

1. Navigate into the `data_processing_tutorial - smelter` folders, and open the `CSTable.xls` file. Notice how the `Copper Smelters` worksheet is formatted for display purposes.

2. Select all the columns in this sheet, right click, and choose *Copy*. Then go to *File - New* and open a new, blank workbook. Click in cell `A1`, right click, and under *Paste* hit the *Values* button (little clipboard with 123). Alternatively, you can click on the *Paste Special* text, and choose the *values* radio button and hit *OK*.



3. Go to *File - Save As* and save the file in the `data_processing_tutorial - smelter` as `copper_smelters.xlsx` in the contemporary Excel format `.xlsx`.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Country	Location (Name)	Latitude	Longitude	Facility	Commodit	Majority O	Status	Capacity 1	Process Type 2/		
2					Type							
3	Albania											
4		Kukes (Gjegian)	42.08	20.42	Smelter	Cu	Albanian G	Operating	5	Reverberatory		
5		Lac	41.63	19.71	Smelter	Cu	Albanian G	Operating	7	Blast Furnace		
6		Rubik	41.77	19.79	Smelter	Cu	Albanian G	Operating	5	Reverberatory		
7	Armenia											
8		Alaverdi	41.13	44.65	Smelter	Cu	Armenian (	Operating	7	Reverberatory		
9									3	Reverberatory (S)		
10	Australia											
11		Mount Isa, Queensland	-20.73	139.5	Smelter	Cu	Mount Isa	Operating	250	Isasmelt/Reverberatory		
12		Roxby Downs, South Australia (Olympic Dams)	-30.5	136.9	Smelter	Cu	Olympic D	Operating	200	Outokumpu Continuous		
13		Port Kembla, New South Wales	-34.467	150.9	Smelter	Cu	Port Kemb	Operating	140	Noranda Continuous		

2 Remove Headers and Footers

Data tables can only have a single header row, and cannot have footnotes or subtotals.

1. Delete `Row 2` in-between the column header row and the first value row. Then scroll to the bottom of the worksheet and delete the two rows of footnotes. Save the workbook.

3 Making Values Explicit

The names of the countries in the sheet appear once on their own row as subheadings, and the rows that follow represent smelters in that country. In a data table, each row must represent a unit or record of data. We need to explicitly associate country names with each individual row, and eliminate the country subheader rows.

1. Select `Column A`. On *Home - Find and Select* choose *Go to Special*, and click the radio button for *Blanks*, and hit *OK*. `Cell A3` will be highlighted - do *not* click on the cell, just type: `=A2` without hitting enter. Then, hold down the CTRL key and hit enter. The blank cells will autofill with the country name above it.

2. Select Column A, right click and *Copy*, then right click again and choose *Paste Special - Values*. This overwrites the formulas in this column with the actual values output by the formulas.

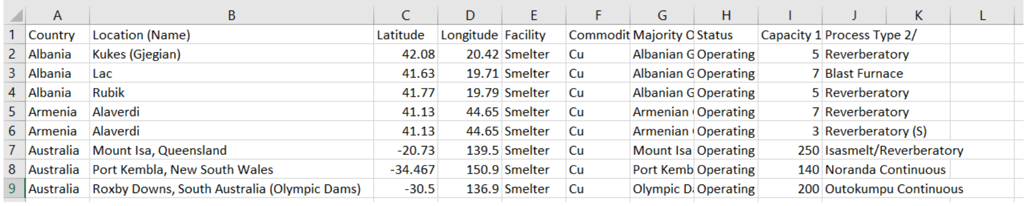
3. Scroll down and make manual corrections in this column. Row 83 wasn't filled because of a blank space, so type in the country name for Iran. Scroll to the bottom and delete the duplicate country names for Zimbabwe. Save the spreadsheet.

4. Return to the top of the list, and see that there are some gaps in the data. Row 8 for Armenia has a lot of missing information; this particular geographic location has two smelters, so the data for the second smelter was implied rather than hardcoded. Let's determine how widespread this issue is.

5. In Cell K3 type this formula: =IF(AND(ISNUMBER(I3),H3=""),"PROBLEM","") If the first condition in parentheses is true (there is a number in the capacity field but nothing "" in the status field (AND allows us to specify two criteria), then flag that we have a problem. Otherwise, print nothing. Copy and paste this formula down the length of the column.

6. Two problems appear; Row 8 for Armenia and Row 49 for China. Fill in the blank portions of these rows (from Location to Status with the data from the row above). Then select Column L and delete it to remove the formulas. Save the workbook.

7. Click in Cell A1, go to *Data - Sort*, check *My data has headers* box, and *Sort* by Location (Name). Scroll to the bottom of the sheet and delete all the rows that just have the country names (these were the country subheader rows). Go back and sort the data again, by Country, and *Add a level*, and add Location (Name). Save your work.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Country	Location (Name)	Latitude	Longitude	Facility	Commodit	Majority O	Status	Capacity 1	Process Type 2/		
2	Albania	Kukes (Gjegian)	42.08	20.42	Smelter	Cu	Albanian G	Operating	5	Reverberatory		
3	Albania	Lac	41.63	19.71	Smelter	Cu	Albanian G	Operating	7	Blast Furnace		
4	Albania	Rubik	41.77	19.79	Smelter	Cu	Albanian G	Operating	5	Reverberatory		
5	Armenia	Alaverdi	41.13	44.65	Smelter	Cu	Armenian	Operating	7	Reverberatory		
6	Armenia	Alaverdi	41.13	44.65	Smelter	Cu	Armenian	Operating	3	Reverberatory (S)		
7	Australia	Mount Isa, Queensland	-20.73	139.5	Smelter	Cu	Mount Isa	Operating	250	Isasmelt/Reverberatory		
8	Australia	Port Kembla, New South Wales	-34.467	150.9	Smelter	Cu	Port Kemb	Operating	140	Noranda Continuous		
9	Australia	Roxby Downs, South Australia (Olympic Dams)	-30.5	136.9	Smelter	Cu	Olympic D	Operating	200	Outokumpu Continuous		

4 Splitting Values with Text to Columns

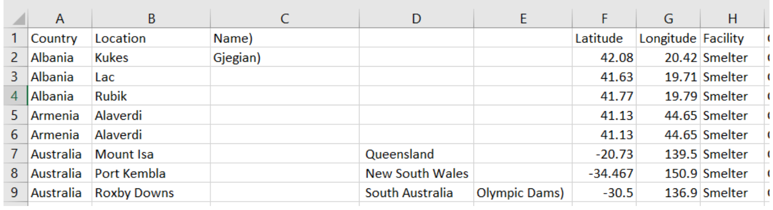
The Location (Name) column conflates several different variables, which is a bad practice. The geographic location appears first; if it is followed by a comma a state or province name is provided, and if a name appears in parentheses it represents a name of the facility. If we want to clearly identify each of these attributes, or wish to use the location as a label in GIS, we need to split these values apart. The *Text to Columns* tool will allow us to split text into different columns using a specific character found in the text.

1. Select Column C, right click, and *Insert* a blank column. NOTE - it's important to always insert a blank column beside the one you are splitting; if you don't do this data in the adjacent columns will be overwritten. Select Column B Locations, and go to *Data - Text to*

Columns. Choose *Delimited*, and in the next window check *Comma* and nothing else. Click *Finish*. This leaves Locations (with and without feature Names) in `Column B` and states (with and without feature names) in `Column C`, split apart based on the location of the comma.

2. Insert a new column to the right of `Column B`. Select `Column B - Locations`, *Data - Text to Columns*, *Delimited*, uncheck *Comma* and check *Other* and in the *Other* box type an open parentheses (. then click *Finish*. Now we have just the Locations in `Column B`, and if there was a feature name it's now in `Column C`.

3. Insert a new column to the right of `Column D`. Select `Column D - States`, *Data - Text to Columns*, *Delimited*, *Other* with an open parentheses (then *Finish*. Now we have just the States in `Column D`, and if there was a feature name it's now in `Column E`.

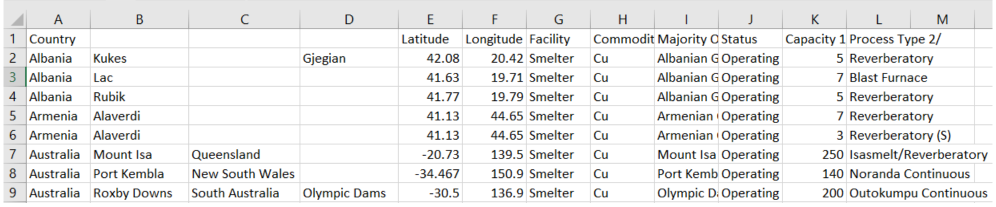


	A	B	C	D	E	F	G	H
1	Country	Location	Name)			Latitude	Longitude	Facility
2	Albania	Kukes	Gjegian)			42.08	20.42	Smelter
3	Albania	Lac				41.63	19.71	Smelter
4	Albania	Rubik				41.77	19.79	Smelter
5	Armenia	Alaverdi				41.13	44.65	Smelter
6	Armenia	Alaverdi				41.13	44.65	Smelter
7	Australia	Mount Isa		Queensland		-20.73	139.5	Smelter
8	Australia	Port Kembla		New South Wales		-34.467	150.9	Smelter
9	Australia	Roxby Downs		South Australia	Olympic Dams)	-30.5	136.9	Smelter

4. Insert a new column to the right of `Column E`. In this column we will combine the two separate feature name columns. In `Cell F2` type `=C2&E2`. Copy and paste this formula all the way down. Then do a *Copy - Paste Special - Values* on `Column F`. Delete `Column E`, then delete Column C. Save your work.

5. Let's clean up the feature column. Select `Column D`, go to *Home - Find and Select - Replace*. Type a closing parentheses) in the *Find* box and hit *Replace All*. This removes the parentheses and replaces it with nothing.

6. Many of our name values have trailing or leading spaces. Insert three empty columns to the right of `Column D`. In `Cell E2` type `=TRIM(B2)`. Copy this formula to the two adjacent cells in Columns E and F. Then copy the formula for these three columns all the way down the sheet. Do a *Copy - Paste Special - Values* on all three columns. Delete `Columns B,C,D`. Save your work.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Country				Latitude	Longitude	Facility	Commodit	Majority O	Status	Capacity 1	Process Type 2/	
2	Albania	Kukes		Gjegian	42.08	20.42	Smelter	Cu	Albanian G	Operating	5	Reverberatory	
3	Albania	Lac			41.63	19.71	Smelter	Cu	Albanian G	Operating	7	Blast Furnace	
4	Albania	Rubik			41.77	19.79	Smelter	Cu	Albanian G	Operating	5	Reverberatory	
5	Armenia	Alaverdi			41.13	44.65	Smelter	Cu	Armenian (	Operating	7	Reverberatory	
6	Armenia	Alaverdi			41.13	44.65	Smelter	Cu	Armenian (	Operating	3	Reverberatory (S)	
7	Australia	Mount Isa	Queensland		-20.73	139.5	Smelter	Cu	Mount Isa	Operating	250	Isasmelt/Reverberatory	
8	Australia	Port Kembla	New South Wales		-34.467	150.9	Smelter	Cu	Port Kemb	Operating	140	Noranda Continuous	
9	Australia	Roxby Downs	South Australia	Olympic Dams	-30.5	136.9	Smelter	Cu	Olympic D:	Operating	200	Outokumpu Continuous	

5 Naming Columns

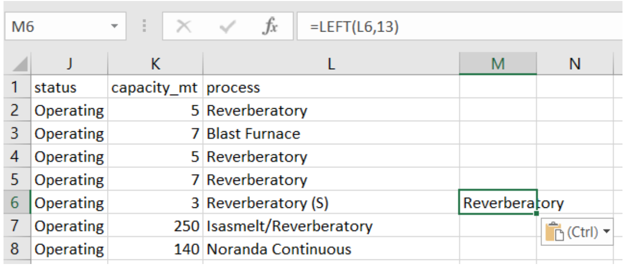
Tables that are being imported into GIS or a database cannot have column names that are: too long, contain spaces, contain punctuation (except underscores), or begin with numbers. Some like to use camel case (PctTotal) to distinguish different parts of the name, while others prefer snake case (pct_total). Use common sense abbreviations for values (pct, avg, tot) and units (km, vol, sqmi).

1. Name your columns from left to right as follows: `country`, `location`, `state`, `smelt_name`, `latitude`, `longitude`, `facility`, `commodity`, `company`, `status`, `capacity_mt`, `process`.

6 Splitting Text with Formulas

Formulas can also be used for splitting strings of text into distinct values. For example, the `process` column contains footnotes in parentheses. In GIS or a database, this would prevent us from summarizing data into categories using this field, as a single category would be split into two (one with and without a footnote). The following example illustrates how to use substring formulas for this type of work. We will strip the `process` name out of the cell without the footnote.

1. In `Cell M6` type `=LEFT(L6,13)`. This reads from the left of the cell and returns the first 13 characters, so the string `Reverberatory` is extracted to exclude the footnote `(S)`.



M6 =LEFT(L6,13)

	J	K	L	M	N
1	status	capacity_mt	process		
2	Operating	5	Reverberatory		
3	Operating	7	Blast Furnace		
4	Operating	5	Reverberatory		
5	Operating	7	Reverberatory		
6	Operating	3	Reverberatory (S)	Reverberatory	
7	Operating	250	Isasmelt/Reverberatory		
8	Operating	140	Noranda Continuous		

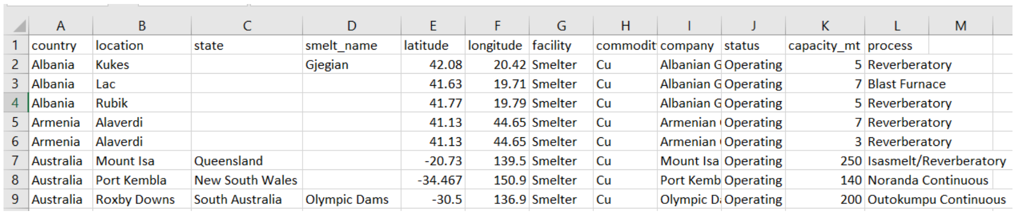
2. Replace the formula you just typed with this one: `=FIND("(",L6)`. This finds the first occurrence of the open parentheses and returns its position in the string.

3. Replace this formula with: `=LEFT(L6,FIND("(",L6)-2)`. This reads from the left of the cell, and for the character position we use the location of the first open parentheses, minus two positions (to avoid returning the parentheses and the leading space in the result).

4. Take this formula and copy / paste it into `Cell M5` directly above it. The formula returns a value error, because the string in `Cell L5` does not contain a parentheses.

5. Replace the formula in `Cell M5` with this one: `=IFERROR(LEFT(L5,FIND("(",L5)-2),L5)`. This says, if there is no error return the string before the parentheses. Otherwise, if an error is returned (because there is no parentheses), then simply return the full value that's in that cell.

6. Delete the existing formulas you created in the last steps. Enter this formula into cell `Cell M2`: `=IFERROR(LEFT(L2,FIND("(",L2)-2),L2)`. Copy and paste it down the column. Select the column and do a *Copy - Paste Special - Values*. Delete `Column L` which is the `process` column with the footnotes. Rename the new `process` column without the footnotes in `Column L`: `process`. Save your work.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	country	location	state	smelt_name	latitude	longitude	facility	commodit	company	status	capacity_mt	process	
2	Albania	Kukes		Gjegian	42.08	20.42	Smelter	Cu	Albanian G	Operating	5	Reverberatory	
3	Albania	Lac			41.63	19.71	Smelter	Cu	Albanian G	Operating	7	Blast Furnace	
4	Albania	Rubik			41.77	19.79	Smelter	Cu	Albanian G	Operating	5	Reverberatory	
5	Armenia	Alaverdi			41.13	44.65	Smelter	Cu	Armenian	Operating	7	Reverberatory	
6	Armenia	Alaverdi			41.13	44.65	Smelter	Cu	Armenian	Operating	3	Reverberatory	
7	Australia	Mount Isa	Queensland		-20.73	139.5	Smelter	Cu	Mount Isa	Operating	250	Isasmelt/Reverberatory	
8	Australia	Port Kembla	New South Wales		-34.467	150.9	Smelter	Cu	Port Kemb	Operating	140	Noranda Continuous	
9	Australia	Roxby Downs	South Australia	Olympic Dams	-30.5	136.9	Smelter	Cu	Olympic D	Operating	200	Outokumpu Continuous	

7 Unique Identifiers

Records in a data table should always have a unique identifier. If a table lacks an identifier, some GIS packages and database programs will create one, but others won't. Adding a unique integer is straightforward.

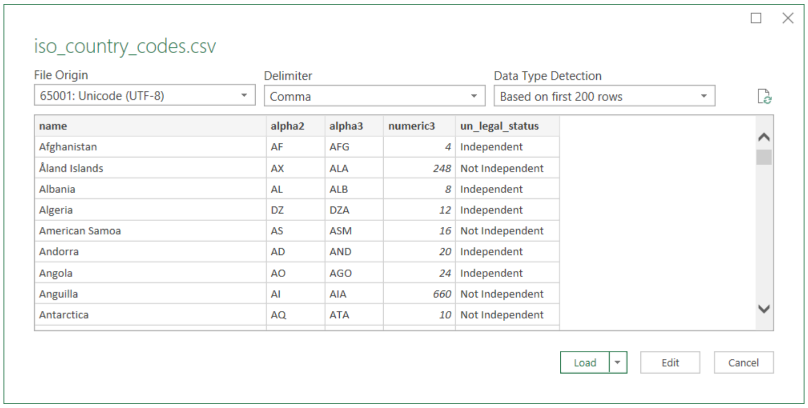
1. Insert a new column to the left of `Column A`. In `Cell A2` type 1. in `Cell A3` type `=A2+1`. Copy and paste the formula down the column. Select the column and do a *Copy - Paste Special Values*. Name the column `uid`. Save your work.

8 Importing CSVs and VLOOKUP

CSV files are plain text files where rows are stored on lines separated by the enter key, and row values are separated by a comma. Excel parses CSV files into rows and columns and makes judgements on whether values represent text or numbers; in some cases it judges poorly. ID codes often having leading zeros (think of USPS ZIP Codes), and if they are saved as numbers the zero is dropped and the value is stored incorrectly. MS Windows also uses an encoding standard that does not preserve all characters in non-English Latin and non-Latin alphabets.

In this example we will demonstrate how to import CSV files correctly to preserve their formatting. There are national and international standards for uniquely identifying certain features. Countries are uniquely identified using ISO 3166 country codes. These identifiers are better for identifying, relating, and aggregating data as opposed to using names, which can be highly variable. After importing the codes from CSV, we will see how to quickly relate values in two different worksheets using the VLOOKUP formula.

1. Go to *Data - Get Data - From File - From Text/CSV*. Browse into the `data_processing_tutorial - smelter_data` folders, select `iso_country_codes.csv`, and click *Import*. Under *File Origin* change the scheme from *1252: Western European (Windows)* to *65001: Unicode (UTF-8)*, and notice how the diacritics for country names are now properly rendered. Hit the *Edit* button at the bottom of the window. In the `numeric3` column hit the small *123* button and change this to *Text*. When prompted choose *Replace current*. Note the values are now saved as text with leading zeros preserved. Hit *Home - Close and Load*, and the file is loaded into `Sheet2`. Save your work.



iso_country_codes.csv

File Origin: 65001: Unicode (UTF-8)

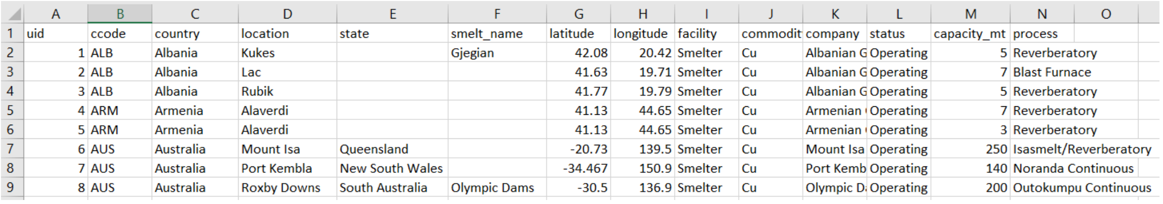
Delimiter: Comma

Data Type Detection: Based on first 200 rows

name	alpha2	alpha3	numeric3	un_legal_status
Afghanistan	AF	AFG	4	Independent
Åland Islands	AX	ALA	248	Not Independent
Albania	AL	ALB	8	Independent
Algeria	DZ	DZA	12	Independent
American Samoa	AS	ASM	16	Not Independent
Andorra	AD	AND	20	Independent
Angola	AO	AGO	24	Independent
Anguilla	AI	AIA	660	Not Independent
Antarctica	AQ	ATA	10	Not Independent

2. Tab back to `Sheet1`. Insert a new column to the left of `Column C Location`. In `Cell B2` type this formula: `=VLOOKUP(C2,Sheet2!A:C,3,FALSE)`. This takes the `Cell C2` `country` value and attempts to find that value in the *first* column (implied) in `Sheet2`. The 2nd parameter is the location of the full range of cells that the formula will look. If it finds the same value in the first column, it moves over to column 3 (specified in the 3rd parameter) and returns whatever is in that row. FALSE indicates that the match must be an exact match, not an approximate one. Copy and paste this formula down the column, then select the column and do a *Copy - Paste Special - Values*. Name the column `ccode`.

3. Some values did not return a match because of differences in the way country names were spelled in each table. Manually fix records that didn't match. `Iran - IRN, Korea, North - PRK, Russia - RUS, Serbia and Montenegro - SRB, United States - USA`. Change the name for `Serbia and Montenegro` to `Serbia` (since this dataset was created these countries have split into two separate ones, and the smelter is in Serbia). Save your work.



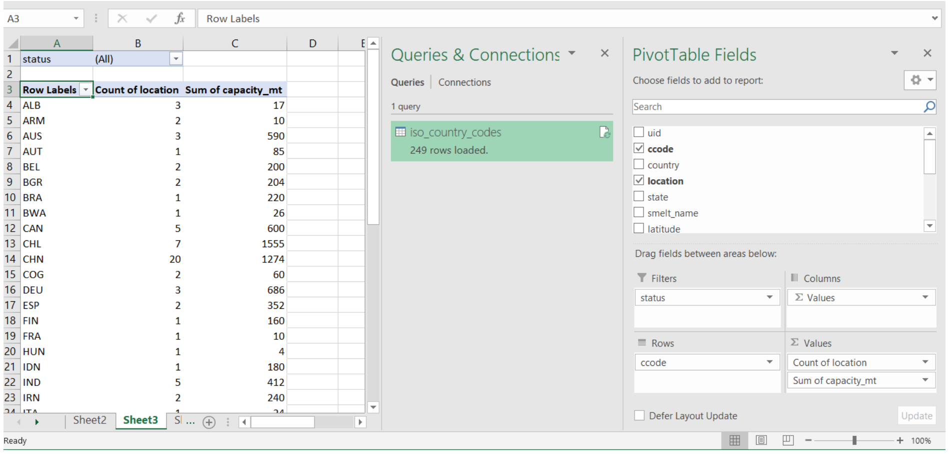
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	uid	ccode	country	location	state	smelt_name	latitude	longitude	facility	commodit	company	status	capacity_mt	process	
2	1	ALB	Albania	Kukes		Gjegian	42.08	20.42	Smelter	Cu	Albanian G	Operating	5	Reverberatory	
3	2	ALB	Albania	Lac			41.63	19.71	Smelter	Cu	Albanian G	Operating	7	Blast Furnace	
4	3	ALB	Albania	Rubik			41.77	19.79	Smelter	Cu	Albanian G	Operating	5	Reverberatory	
5	4	ARM	Armenia	Alaverdi			41.13	44.65	Smelter	Cu	Armenian	Operating	7	Reverberatory	
6	5	ARM	Armenia	Alaverdi			41.13	44.65	Smelter	Cu	Armenian	Operating	3	Reverberatory	
7	6	AUS	Australia	Mount Isa	Queensland		-20.73	139.5	Smelter	Cu	Mount Isa	Operating	250	Isasmelt/Reverberatory	
8	7	AUS	Australia	Port Kembla	New South Wales		-34.467	150.9	Smelter	Cu	Port Kemb	Operating	140	Noranda Continuous	
9	8	AUS	Australia	Roxby Downs	South Australia	Olympic Dams	-30.5	136.9	Smelter	Cu	Olympic D	Operating	200	Outokumpu Continuous	

9 Pivot Tables

The smelter data is now in a well formed state, and the coordinates in this file can be used to plot smelter locations. What if we wanted a summary table by country, so we could potentially join this table to a shapefile and map it? A pivot table makes this possible.

1. While in `Sheet1` go to *Insert - Pivot Table*. Click *OK* on the pivot window. In the *Pivot Table Fields* window on the right, choose `ccode` to add it as a *row*. Drag `location` and then `capacity_mt` into the *Values* section. It will assume to count the former and sum the latter (keep these defaults, but you could change them by hitting the drop down for each). Drag

`status` under the *Filters* section. In the table on the left, change the dropdown for `status` to `Operating` to remove smelters that are not open. Save your work.



A3 | Row Labels

	A	B	C
1	status	(All)	
2			
3	Row Labels	Count of location	Sum of capacity_mt
4	ALB	3	17
5	ARM	2	10
6	AUS	3	590
7	AUT	1	85
8	BEL	2	200
9	BGR	2	204
10	BRA	1	220
11	BWA	1	26
12	CAN	5	600
13	CHL	7	1555
14	CHN	20	1274
15	COG	2	60
16	DEU	3	686
17	ESP	2	352
18	FIN	1	160
19	FRA	1	10
20	HUN	1	4
21	IDN	1	180
22	IND	5	412
23	IRN	2	240

Queries & Connections

1 query

iso_country_codes
249 rows loaded.

PivotTable Fields

Choose fields to add to report:

□ uid
☑ ccode
□ country
☑ location
□ state
□ smelt_name
□ latitude

Drag fields between areas below:

Filters: status

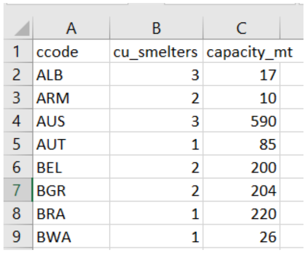
Columns: Σ Values

Rows: ccode

Values: Count of location, Sum of capacity_mt

□ Defer Layout Update

2. Copy *Columns A to C* in the pivot table. Hit the *plus* symbol at the bottom of the sheet to add a new sheet, *Sheet4*. Move into this sheet, click in `Cell A1` and do a *Paste Special - Values*. Delete the first two rows, then delete the last `Grand total` row. Rename the column headings: `ccode, cu_smelters,capacity_mt`. Save your work.



	A	B	C
1	ccode	cu_smelters	capacity_mt
2	ALB	3	17
3	ARM	2	10
4	AUS	3	590
5	AUT	1	85
6	BEL	2	200
7	BGR	2	204
8	BRA	1	220
9	BWA	1	26

10 Final Steps

1. Click on each sheet name and rename them: `Sheet1: smelters, Sheet2: ccodes, Sheet3: pivot, Sheet4: country_smelt`. Save your work.

2. Excel files can be added directly to GIS packages. When you add a file it will prompt you for an individual worksheet. The `country_smelt` sheet can be added as a table and joined to a shapefile of countries using the `ccode`. Try doing with this with the country layer from Natural Earth, saved in the `nat_earth_countries` folder.

3. The `smelters` data can be plotted as XY data using its longitude and latitude coordinates to create a point file. Most GIS packages require XY data to be in a CSV format. Go into the `smelters` sheet and choose *File - Save As*. Save it in the project folder as a CSV UTF-8 file (this will preserve diacritics in country names). It will save just the active sheet as a CSV. In a GIS package, when plotting specify longitude as the X coordinate and latitude as the Y coordinate. After plotting, transform it to a spatial data file by exporting it as a shapefile or feature class (otherwise your ability to manipulate the file will be limited).