esri Australia

GIS For Schools SKILLS

Data Cleaning

What is data cleaning?

Data cleaning, not to get confused with actual cleaning, is the process of identifying and removing inaccurate, unfinished, unreliable or non-relevant parts from your data set.

What cleaning needs to take place?

- Delete extra spaces
- Remove duplicates
- Remove source(s)
- Fix column headers (no spaces)

Did you know?

Data scientists spend 80% of their time cleaning and manipulating their data and only 20% of their time actually analysing it.

Let's take a look at what is wrong with this data set.

	A	B	C	D
1	Data Sourced From Esri Australia			
2	Date: 2018			
3				
4	Country	World Happiness Score	Life Expectancy	
5	Australia	81	81	
6	Dem. Rep. Congo	63	71	
7	NZ	80	81	
8				
9	Downloaded from: madeupstatistics.com.au			
10				
11				
12				
13				

The headers in column B and C need underscores instead of spaces. It should be World_Happiness_Score And Life_Expectancy

Row 6 contains a shortened version of the full country name. It needs to say Democratic Republic of the Congo

Row 7 contains the acronym for NZ. It needs to say the full name, New Zealand

Rows 1, 2 and 3 contain blank spaces and unnecessary information. They need to be deleted

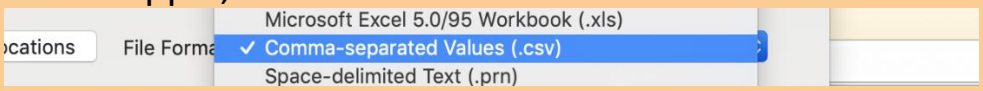
Row 9 contains the website link. This needs to be deleted (after you have written it down in your reference list!)

	A	B	C	D
1	Country	World_Happiness_Score	Life_Expectancy	
2	Australia	81	81	
3	Democratic Republic of Congo	63	71	
4	New Zealand	80	81	
5				
6				

This is what the data set looks like when it is cleaned. Once you have cleaned your dataset, save it as a **CSV**.

If you are on Windows, it will look like this:

If you are on Apple, it will look like this:



The below is a real data set from the World Bank. We have highlighted everything we would delete.

	A	B	C	D	E	F	G	H	I	J
1	Data Source	World Development Indicators								
2										
3	Last Updated Date	3/21/2019								
4										
5	Country Name	Country Code	Indicator Name	Indicator Code	2013	2014	2015	2016	2017	2018
6	Aruba	ABW	Forest area (sq. km)	AG.LND.FRST.K2	4.2	4.2	4.2	4.2		
7	Afghanistan	AFG	Forest area (sq. km)	AG.LND.FRST.K2	13500	13500	13500	13500		
8	Angola	AGO	Forest area (sq. km)	AG.LND.FRST.K2	581056	579808	578560	577312		
9	Albania	ALB	Forest area (sq. km)	AG.LND.FRST.K2	7734.2	7724.6	7715	7705.4		
10	Andorra	AND	Forest area (sq. km)	AG.LND.FRST.K2	160	160	160	160		
11	Arab World	ARB	Forest area (sq. km)	AG.LND.FRST.K2	386635.8	384200.2	381764.6	379329		
12	United Arab Emirates	ARE	Forest area (sq. km)	AG.LND.FRST.K2	3204.8	3215.4	3226	3236.6		
13	Argentina	ARG	Forest area (sq. km)	AG.LND.FRST.K2	277056	274088	271120	268152		
14	Armenia	ARM	Forest area (sq. km)	AG.LND.FRST.K2	3316	3318	3320	3322		
15	American Samoa	ASM	Forest area (sq. km)	AG.LND.FRST.K2	176.1	175.8	175.4	175		
16	Antigua and Barbuda	ATG	Forest area (sq. km)	AG.LND.FRST.K2	98	98	98	98		
17	Australia	AUS	Forest area (sq. km)	AG.LND.FRST.K2	1241350	1244430	1247510	1250590		
18	Austria	AUT	Forest area (sq. km)	AG.LND.FRST.K2	38654	38672	38690	38708		
19	Azerbaijan	AZE	Forest area (sq. km)	AG.LND.FRST.K2	10869.6	11131.8	11394	11656.2		
20	Burundi	BDI	Forest area (sq. km)	AG.LND.FRST.K2	2668	2714	2760	2806		
21	Belgium	BEL	Forest area (sq. km)	AG.LND.FRST.K2	6825.2	6829.6	6834	6838.4		
22	Benin	BEN	Forest area (sq. km)	AG.LND.FRST.K2	44110	43610	43110	42610		
23	Burkina Faso	BFA	Forest area (sq. km)	AG.LND.FRST.K2	54696	54098	53500	52902		
24	Bangladesh	BGD	Forest area (sq. km)	AG.LND.FRST.K2	14342	14316	14290	14264		
25	Bulgaria	BGR	Forest area (sq. km)	AG.LND.FRST.K2	37886	38058	38230	38402		