Data Organization in Spreadsheets: Best Practices

Or, how to save yourself and your collaborators headaches, frustration, and retractions

Data Organization in Spreadsheets

Karl W. Broman and Kara H. Woo

The American Statistician, 2017
https://doi.org/10.1080/00031305.2017.1375989

88%

of spreadsheets have errors

–*What We Know About Spreadsheet Errors*, Raymond Panko

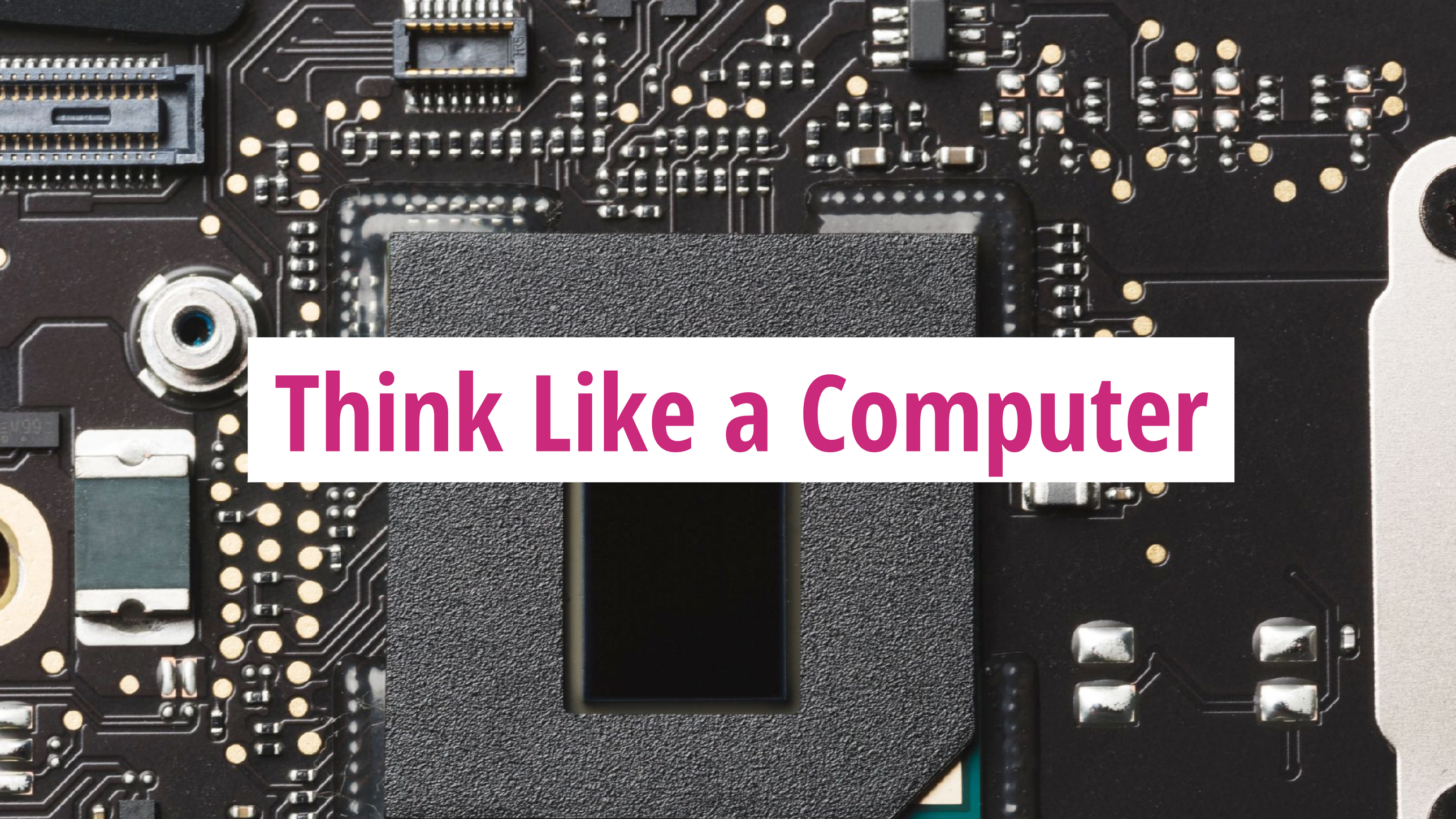
KANYE WEST
@kanyewest

burn that excel spread sheet 🔥😂

11:50 AM - Apr 25, 2018

34.2K 7,913 people are talking about this

1. Data is Correct
2. Data Results in Correct Analysis
3. Easy for Humans to Use
4. Easy for Computers to Read



Think Like a Computer

- **Naming and Consistency**
- Dates
- Cell Management
- Rectangles!
- Raw Data vs. Analysis
- Formatting, Styles, and Notes
- Data Validation
- Documentation

Consistent Terms and Codes

`male` ≠ `Male` ≠ `m` ≠ `M` ≠ `male ` ≠ ` male`

Capitalization

Abbreviations

Extra Spaces

TRIM(text)

LOWER(text)

UPPER(text)

Variable Names: Baseline

- Same across files (including capitalization)
- Same conventions across variables
- No duplicates
- All in one row
- Every column has a name

	Life Expectancy		GDP	
	2010	2011	2010	2011
Albania				
Algeria				
Armenia				
Azerbaijan				

Country	Life Expectancy 2010	Life Expectancy 2011	GDP 2010	GDP 2011
Albania				
Algeria				
Armenia				
Azerbaijan				

Variable Names: Better

- No spaces
- Only: lowercase letters, numbers, underscore, hyphen
- Start with a letter
- Short, but meaningful (not too short)

country	lifeexp_2010	lifeexp_2011	gdp_2010	gdp_2011
Albania				
Algeria				
Armenia				
Azerbaijan				

Naming and Consistency

Dates

Cell Management

Rectangles!

Raw Data vs. Analysis

Formatting, Styles, and Notes

Data Validation

Documentation

Reference year for dates in Excel differs between Mac and PC

Programs assume order and fill in missing date components

4/3 >>> April 3, 2018

June 2017 >>> 6/1/2017

Dates are only numbers and hope

	Excel	
		January 1, 1850
0	1/0/00	January 0, 1900
1	1/1/00	January 1, 1900
432	3/7/01	March 7, 1901
4321	10/30/11	October 30, 1911
43256	6/5/18	June 5, 2018

Excel	Copied to Google Sheets	Full Google Sheets
1/0/00	1/0/00	1/0/00
1/1/00	1/1/00	January 1, 2000
3/7/01	3/7/01	March 7, 2001
10/30/11	10/30/11	October 30, 2011
6/5/18	6/5/18	June 5, 2018

Google Sheets

Original Number	Date from Number	Full Date from Number
-18,260	1/1/50	January 1, 1850
0	12/30/99	December 30, 1899
1	12/31/99	December 31, 1899
432	3/7/01	March 7, 1901
4321	10/30/11	October 30, 1911
43256	6/5/18	June 5, 2018

Does it look like a date (at all)? Spreadsheets want to be helpful...

Oct-4 >>> 10/4/2018

YYYY-MM-DD

YYYY-MM-DDTHH:MM:SS

All as one number: YYYYMMDD

20010101

18501225

19040210

Three separate columns

year	month	day
2001	1	1
1850	12	25
1904	2	10

BONUS

Spreadsheets will remove leading zeros on numbers

0183682943 >>> 183682943

ID variables won’t match back to other files!

Naming and Consistency

Dates

Cell Management

Rectangles!

Raw Data vs. Analysis

Formatting, Styles, and Notes

Data Validation

Documentation

What Goes In A Cell?

Something!

No Blank Cells

One thing

And only one thing!

Patient	Round	Concentration
1003	1	0.4
	2	3
	3	6
	4	10
1005	1	0.1
	2	0.2
	3	0.4
1006	1	3
	2	2
	3	4
	4	7

	1 min				5 min			
strain	normal		mutant		normal		mutant	
A	111	170	375	384	277	234	207	466
B	336	169	491	233	392	341	213	472

strain	genotype	minute	trial	response
A	normal	1	1	111
A	normal	1	2	170
A	mutant	1	1	375
A	mutant	1	2	384
A	normal	5	1	277
A	normal	5	2	234
A	mutant	5	1	207
A	mutant	5	2	466
B	normal	1	1	336
B	normal	1	2	169
B	mutant	1	1	491
B	mutant	1	2	233
B	normal	5	1	392
B	normal	5	2	341
B	mutant	5	1	213
B	mutant	5	2	472

Merged Cells = Empty Cells

Don’t do it!

Missing Values

1. Do not leave blank
2. Use text, not numerical values
3. Be consistent

NA - . N/A missing unknown

trial	temperature	height
1	77 F	12m
2	87 F	15m
3	90 F	19m
4		18m
5	68 F	16m
6	69 F	13m

trial	temperature	height
1	77 F	12m
2	87 F	15m
3	90 F	19m
4	NA	18m
5	68 F	16m
6	69 F	13m

trial	temperature_f	height_m
1	77	12
2	87	15
3	90	19
4	NA	18
5	68	16
6	69	13

document	on topic?
483294	yes
578420	no
238302	yes (subtle)
123934	no
893832	no - unsure
723939	yes

sample ID	plate well
34	13 1A
35	14 2C
36	13 1B
38	14 2A

sample_id	plate	well_row	well_column
34	13	1	A
35	14	2	C
36	13	1	B
38	14	2	A

Naming and Consistency

Dates

Cell Management

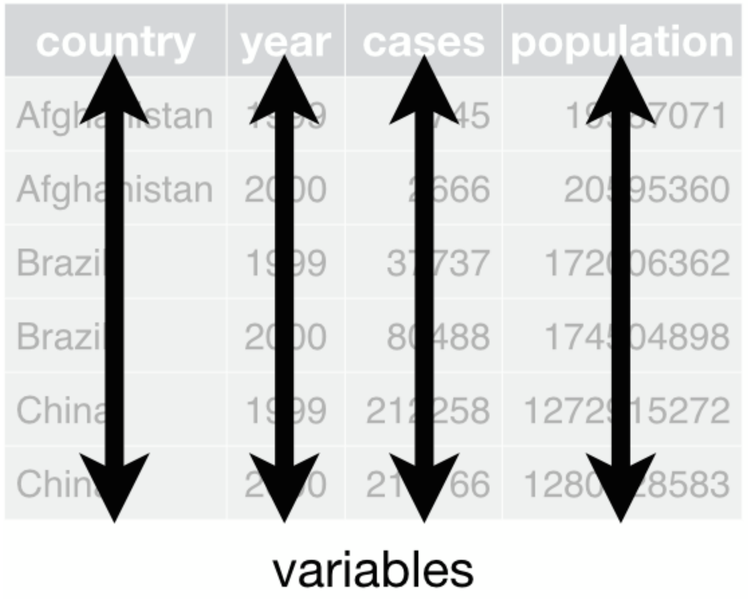
Rectangles!

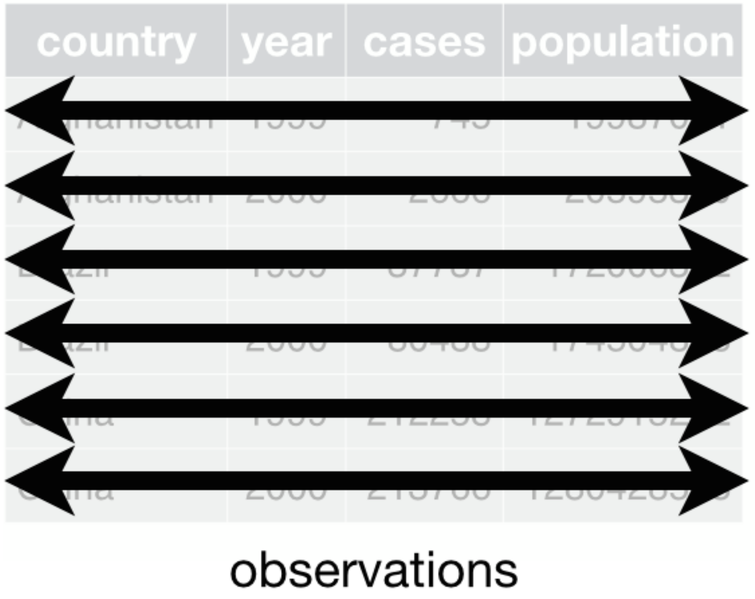
Raw Data vs. Analysis

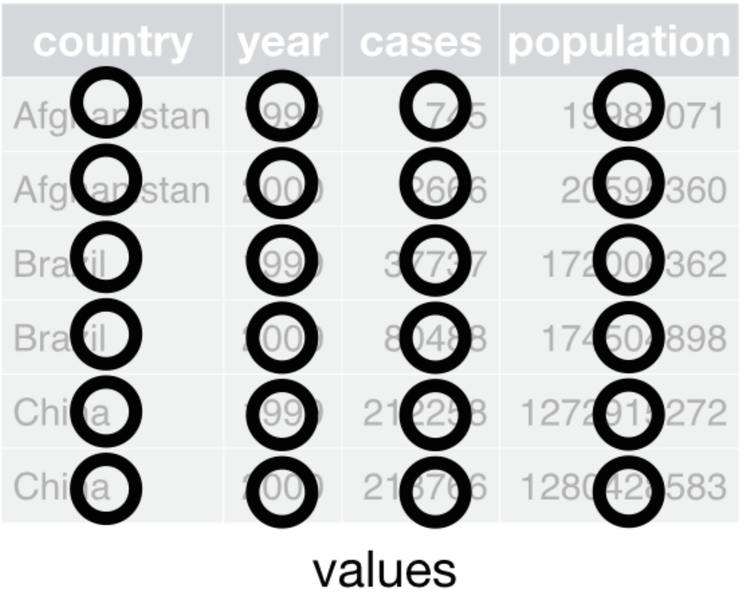
Formatting, Styles, and Notes

Data Validation

Documentation


country
year
cases
population
variables


country
year
cases
population
observations


country
year
cases
population
values

Date	3/4/2015					
Days on diet	143					
Mouse #	14					
sex	f					
experiment		values			mean	SD
control		0.482	0.121	0.386	0.33	0.187
treatment A		0.574	0.226	0.65	0.484	0.226
treatment B		0.944	0.404	0.849	0.732	0.288
fold change		values			mean	SD
treatment A		0.381	0.163	0.029	0.191	0.178
treatment B		0.715	0.148	0.517	0.46	0.288

trial_id	date	diet_days	mouse	sex
13	20150304	143	14	f

group	trial	round	exp_measure
control	13	1	0.482
control	13	2	0.121
control	13	3	0.386
treatment A	13	1	0.574
treatment A	13	2	0.226
treatment A	13	3	0.65
treatment B	13	1	0.944
treatment B	13	2	0.404
treatment B	13	3	0.849

group	trial	round	fold_change
treatment A	13	1	0.381
treatment A	13	2	0.163
treatment A	13	3	0.029
treatment B	13	1	0.715
treatment B	13	2	0.148
treatment B	13	3	0.517

Naming and Consistency

Dates

Cell Management

Rectangles!

Raw Data vs. Analysis

Formatting, Styles, and Notes

Data Validation

Documentation

RAW DATA

Collected or Generated Values Only
Backup Thoroughly, Write Protect
Do NOT Touch
Keep All Versions
Plain Text Format Preferred

ANALYSIS

Copy Raw Data
Summary Statistics, Calculations
Charts and Tables
Pretty Formatting and Layouts

Naming and Consistency

Dates

Cell Management

Rectangles!

Raw Data vs. Analysis

Formatting, Styles, and Notes

Data Validation

Documentation

PLAIN TEXT

CSV: Comma Separated Values

Department of State

Bureau of Population, Refugees, and Migration

Office of Admissions - Refugee Processing Center

Demographical Profile of Refugee Arrivals

Calendar Year

Afghanistan

as of 21-February-2017

Arrivals by Demographic

Report Start Date: 1-January-2002
Report End Date: 21-February-2017

Characteristic	CY 2002			CY 2003			CY 2004			CY 2005			CY 20	
	F	M	Total	F	M	Total	F	M	Total	F	M	Total	F	M
Under 14	413	440	853	187	236	423	116	147	263	141	155	296	93	101
Age 14 to 20	286	185	471	156	204	360	97	114	211	82	90	172	69	88
Age 21 to 30	161	37	198	106	90	196	58	49	107	71	49	120	32	56
Age 31 to 40	205	21	226	88	63	151	56	38	94	63	59	122	47	31
Age 41 to 50	128	20	148	92	58	150	55	51	106	36	82	118	33	45
Age 51 to 64	57	39	96	32	35	67	12	26	38	14	24	38	13	18
Age 65 and Over	15	19	34	9	14	23	9	8	17	5	5	10	1	6
Total	**1,265**	**761**	**2,026**	**670**	**700**	**1,370**	**403**	**433**	**836**	**412**	**464**	**876**	**288**	**345**

Data prior to 2002 was migrated into WRAPS from a legacy system therefore we have more confidence in post-2002 data.

,,,
Department of State,,,
"Bureau of Population, Refugees, and Migration",,,
Office of Admissions - Refugee Processing Center,,,
Demographical Profile of Refugee Arrivals,,,
Calendar Year,,,
Afghanistan,,,
as of 21-February-2017,,,

,,,
Arrivals by Demographic,,,

,,,
Report Start Date:,1-January-2002,,
Report End Date:,21-February-2017,,

,,,
Characteristic,CY 2002,,,,CY 2003,,,CY 2004,,,,CY 2005,,,CY 2006,,,,CY 2007,,,CY 2008,,,CY 2009,,,CY 2010,,,CY 2011,,,CY 2012,,,CY
,F ,M ,Total,,F ,M ,Total,F ,,M ,Total,F ,M ,Total,F ,,M ,Total,F ,M ,Total,F ,M ,Total,F ,M ,Total,F ,M ,Total,F ,M ,Total,F ,M ,Total,F ,M ,To
Under 14,413,440,853,,187,236,423,116,,147,263,141,155,296,93,,101,194,65,55,120,78,85,163,55,51,106,75,78,153,44,61,105,83,96,17
Age 14 to 20,286,185,471,,156,204,360,97,,114,211,82,90,172,69,,88,157,34,55,89,63,66,129,29,43,72,48,68,116,47,38,85,56,79,135,63,
Age 21 to 30,161,37,198,,106,90,196,58,,49,107,71,49,120,32,,56,88,39,24,63,50,42,92,24,32,56,45,67,112,46,37,83,41,48,89,66,64,130,
Age 31 to 40,205,21,226,,88,63,151,56,,38,94,63,59,122,47,,31,78,34,26,60,47,42,89,31,26,57,35,38,73,30,32,62,50,43,93,57,37,94,51,27
Age 41 to 50,128,20,148,,92,58,150,55,,51,106,36,82,118,33,,45,78,20,23,43,38,53,91,15,22,37,24,17,41,23,12,35,31,21,52,35,28,63,35,
Age 51 to 64,57,39,96,,32,35,67,12,,26,38,14,24,38,13,,18,31,12,18,30,15,27,42,3,5,8,13,15,28,11,6,17,12,13,25,16,19,35,10,15,25,28,18,
Age 65 and Over,15,19,34,,9,14,23,9,,8,17,5,5,10,1,,6,7,3,5,8,2,3,5,2,0,2,2,2,4,3,2,5,2,3,5,7,4,11,3,5,8,5,7,12,26,23,49,5,2,7,207,1.43%
Total,"1,265",761,"2,026",,670,700,"1,370",
403,,433,836,412,464,876,288,,345,633,207,206,413,293,318,611,159,179,338,242,285,527,204,188,392,275,303,578,328,350,678,380,3

,,,
Data prior to 2002 was migrated into WRAPS from a legacy system therefore we have more confidence in post-2002 data.,,,,,,,,,,,,,

Plain Text

No formatting
No highlighting or colors
No column/row widths
No column types >> All text
No embedded notes/comments
No hidden columns
Merged cells leave blanks
Single sheet
Single rectangle of data
Rows should be able to be reordered
Single header row

document	date	topic
19594	20161001	econ
16766	20170804	politics
51540	20151230	econ
15667	20150316	econ
14914	20160705	politics
73681	20170326	social
23931	20161010	politics
49176	20160819	politics

document	date	topic	check
19594	20161001	econ	1
16766	20170804	politics	0
51540	20151230	econ	0
15667	20150316	econ	0
14914	20160705	politics	1
73681	20170326	social	0
23931	20161010	politics	0
49176	20160819	politics	0

Naming and Consistency

Dates

Cell Management

Rectangles!

Raw Data vs. Analysis

Formatting, Styles, and Notes

Data Validation

Documentation

Whole Number in a Range

Decimal Number in a Range

Date in a Range

List of Specific Values

Text Length

Custom

Naming and Consistency

Dates

Cell Management

Rectangles!

Raw Data vs. Analysis

Formatting, Styles, and Notes

Data Validation

Documentation

Variable Name

Data Type

Description

Format

Example

Units

Allowed Values

Source

Missing Allowed?

Notes

File/Table

Relations

Additional Resources

- http://dataabinitio.com/
- Good Enough Practices in Scientific Computing
- Research Data Alliance
- European Spreadsheet Risks Interest Group
- Data Carpentry: Data Organization in Spreadsheets