ISPA
Instituto de Investigación Sanitaria del Principado de Asturias

Los Diez Mandamientos de una buena base de datos

Patricio Suárez Gil
Coordinador Plataforma Bioestadística y Epidemiología

Fuente:

https://rtask.thinkr.fr/blog/the-ten-commandments-for-a-well-formatted-database/

Commandement 1 : Tu feras tenir toutes les données dans un seul tableau

Commandment 1: all your data shall fit into one single dataframe

Spreading your data over several spreadsheets or several files is **not** an option. All your data should fit in one single dataframe.

Commandement 2 : Tu respecteras une mise en page précise

Commandment 2: Thou shalt respect a precise formatting

For a perfect database, follow this rule: the

table must be filled cell by cell, **starting from the upper left corner** and, of course, **without skipping line(s)**.

	Session 1		Session 2	
ID	day	score	day	score
1	Monday	23	Thursday	56
2	Monday	54	Friday	43
3	Monday	12	Tuesday	56
4	Tuesday	23	Wednesday	89
5	Monday	56	Thursday	32
6	Monday	87	Thursday	34
7	Tuesday	45	Wednesday	5
8	Tuesday	3	Friday	17

ID	day Session 1	score Session 1	day Session 2	score Session 2
1	Monday	23	Thursday	56
2	Monday	54	Friday	43
3	Monday	12	Tuesday	56
4	Tuesday	23	Wednesday	89
5	Monday	56	Thursday	32
6	Monday	87	Thursday	34
7	Tuesday	45	Wednesday	5
8	Tuesday	3	Friday	17

Commandement 3 : Une ligne = un individu statistique

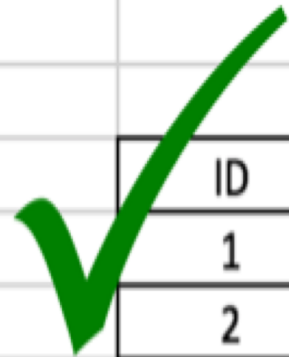
Commandment 3: A line = a statistical individual

It is not always easy to define what a statistical individual is. In a nutshell, **an answer to a survey corresponds to one individual** in your experimental protocol. If a physical person answers two times, you can enter these on two lines.

Commandement 4 : Une colonne = une variable

Commandment 4: A column = a variable

ID	Man	Woman
1	yes	no
2	no	yes
3	no	yes
4	yes	no
5	yes	no
6	yes	no
7	no	yes



ID	Gender
1	Man
2	Woman
3	Woman
4	Man
5	Man
6	Man
7	Woman

ID	Source of pain
1	foot and head
2	head and elbow
3	elbow
4	
5	foot and head
6	foot
7	elbow
8	no pain

ID	pain	source_foot	source_head	source_elbow
1	yes	yes	yes	no
2	yes	no	yes	yes
3	yes	no	no	yes
4	yes			
5	yes	yes	yes	no
6	yes	yes	yes	no
7	yes	no	no	yes
8	no			

ID	pain	source_foot	source_head	source_elbow
1	yes	x	x	
2	yes		x	x
3	yes			x
4	yes			
5	yes	x	x	
6	yes	x	x	x
7	yes			x
8	no			

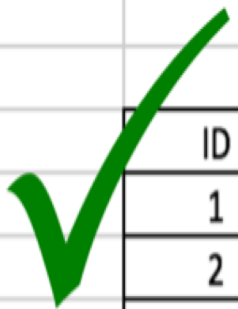
ID	pain	source_foot	source_head	source_elbow
1	yes	yes	yes	no
2	yes	no	yes	yes
3	yes	no	no	yes
4	yes			
5	yes	yes	yes	no
6	yes	yes	yes	no
7	yes	no	no	yes
8	no			

Commandement 5 : Tu ne coderas pas tes variables qualitatives

Commandment 5: Thou shalt not encode thy qualitative variables

Avoid as much as possible encoding qualitative variables: a 1/2 coding for the gender is useless, prefer the use of “woman” and “man” for a better readability. If you really can’t prevent yourself from using a code, then stick to 1/0 to mean presence/absence or yes/no. Not more !

ID	disease
1	1
2	1
3	2
4	2
5	1
6	2
7	2

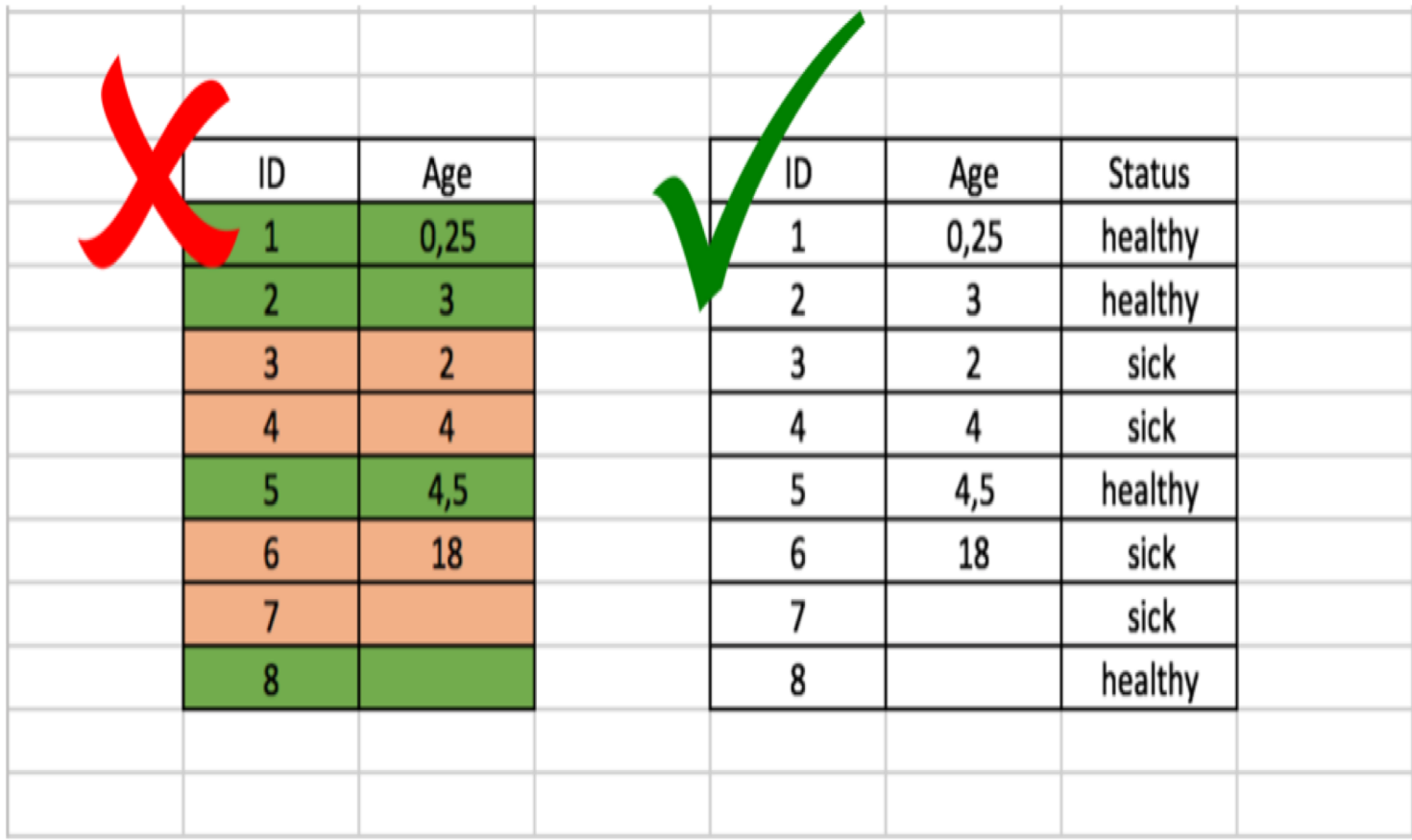


ID	disease
1	cold
2	cold
3	flu
4	flu
5	cold
6	flu
7	flu

Commandement 6 : Ta base ne contiendra que les données

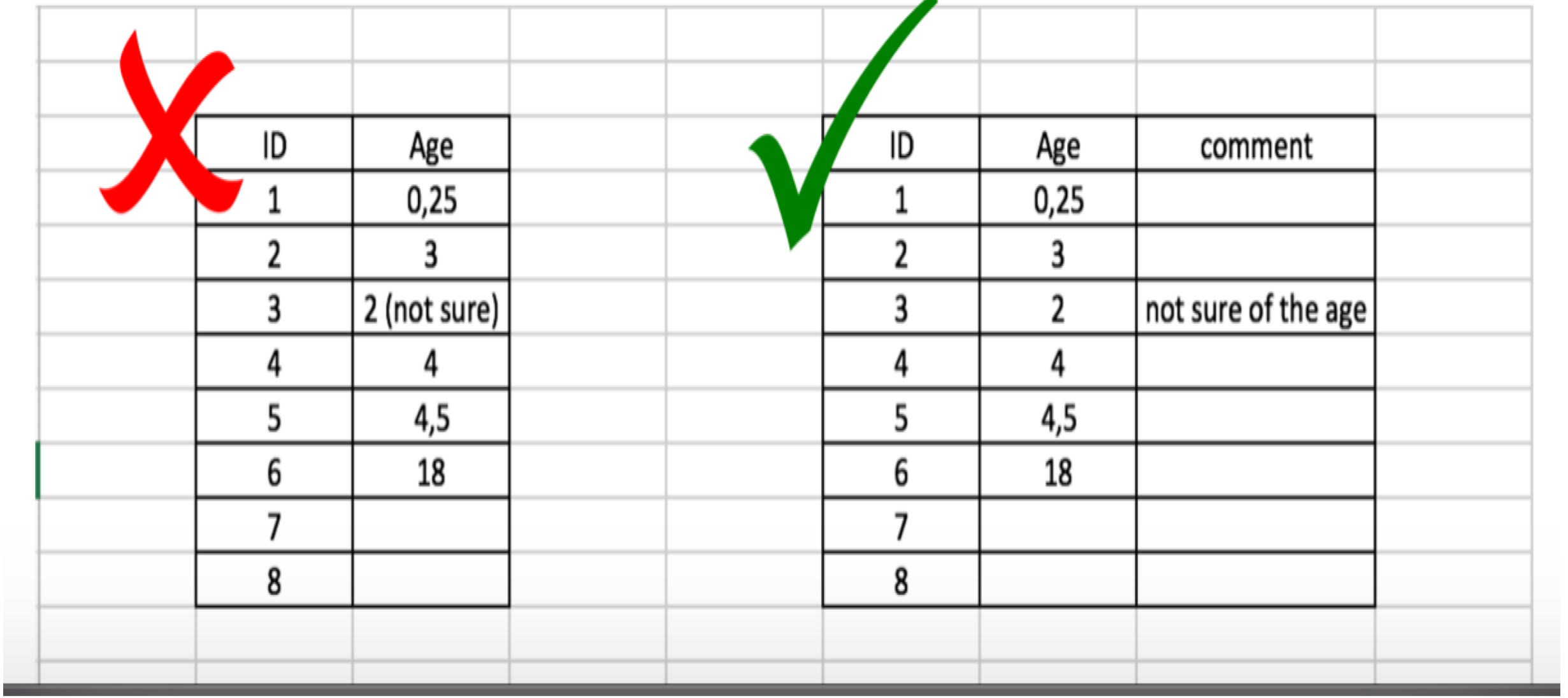
Commandment 6: Thy database shall only contain data

For Excel lovers: don't code your variables using colors or bold/italic/underlined. No "control in blue, test case in red": nothing worse than such a format to loose information when exporting those data.



ID	Age
1	0,25
2	3
3	2
4	4
5	4,5
6	18
7	
8	

ID	Age	Status
1	0,25	healthy
2	3	healthy
3	2	sick
4	4	sick
5	4,5	healthy
6	18	sick
7		sick
8		healthy



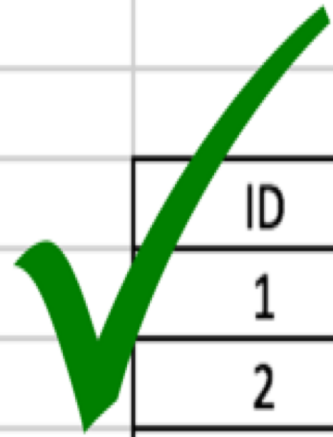
ID	Age
1	0,25
2	3
3	2 (not sure)
4	4
5	4,5
6	18
7	
8	

ID	Age	comment
1	0,25	
2	3	
3	2	not sure of the age
4	4	
5	4,5	
6	18	
7		
8		

Commandement 7 : Homogène, tu resteras

Commandment 7: Homogeneous thou shalt be

ID	Gender
1	Masculine
2	Woman
3	woman
4	M
5	Man
6	Man
7	Woman



ID	Gender
1	Man
2	Woman
3	Woman
4	Man
5	Man
6	Man
7	Woman

Commandement 8 : Tu respecteras tes variables numériques

Commandment 8: Thy numerical variables with respect thou shalt treat

ID	Age
1	3 months
2	3 years
3	between 1 and 3 years
4	4
5	4,5
6	18
7	??
8	-

ID	Age
1	0,25
2	3
3	2
4	4
5	4,5
6	18
7	
8	

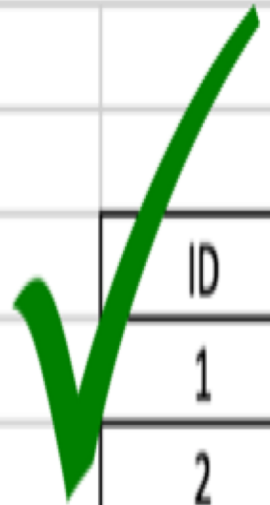
ID	Date of birth
1	01-Nov-86
2	Thirteenth of April 1985
3	01/01/85
4	14/08/04
5	01/01/08

ID	Date of birth
1	1986/11/01
2	1985/04/13
3	1985/01/01
4	2004/08/14
5	2008/01/01

Commandement 9 : Tu anonymiseras ta base

Commandment 9: Anonymous thy database thou shalt keep

First name	Surname	Score	test
Alan	Smith	45	1
Kevin	Hasseloff	23	1
Helen	Marple	38	1
Alan	John	42	2
Judie	Roger	60	2

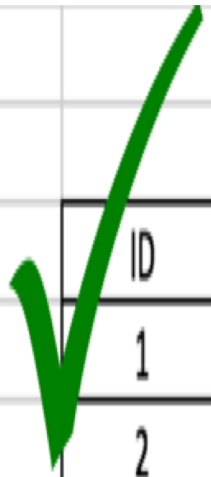


ID	Score	test
1	45	1
2	23	1
3	38	1
4	42	2
5	60	2

Commandement 10 : Tu garderas ta base compréhensible par l'humain

Commandment 10: Human readable thy database shall be

ID	day session	score session	day session	score session
1	Monday	23	Thursday	56
2	Monday	54	Friday	43
3	Monday	12	Tuesday	56
4	Tuesday	23	Wednesday	89
5	Monday	56	Thursday	32
6	Monday	87	Thursday	34
7	Tuesday	45	Wednesday	5
8	Tuesday	3	Friday	17



ID	day Session 1	score Session 1	day Session 2	score Session 2
1	Monday	23	Thursday	56
2	Monday	54	Friday	43
3	Monday	12	Tuesday	56
4	Tuesday	23	Wednesday	89
5	Monday	56	Thursday	32
6	Monday	87	Thursday	34
7	Tuesday	45	Wednesday	5
8	Tuesday	3	Friday	17

Algo más…

- Hacer una copia en formato texto (.txt, .csv …)

- Cuidado con los nombres de las variables: evitar símbolos excepto punto (.) y guión bajo (_)


HOJA FINAL.csv
UNREGISTERED
HOJA FINAL.csv
Historia;EDAD;Sexo;Tabaco;HTA;DL;Diabetes;Talla;Peso;IMC;FEVI;Diagn—stico;Intervenci—n;V. AORTICA;V. MITRAL;V. MIXTA;CORONARIA;MIXTA;CORONARIA+Ao;CORONARIA+MITRAL;OTROS PROCEDIMIENTOS;BENTALL;PROGRAMADA;URGENTE;EMERGENTE;Tiempo CEC;Tiempo Clampaje;Hemorragia;Reintervenci—n;Fecha de Aparici—n;TiempoDesdeIqx;FA postQx;Tratamiento;Farmacol—gico;Cardioversi—n;Ambos;Tratamiento al alta;EXITUS
3086873;59;1;0;1;1;1;1,82;102;30,79338244;60;I330;1/9/16;1;0;0;0;0;0;0;0;1;0;0;41;37;1;0;;-42001;0;;0;0;0;0;
268802;62;0;0;1;1;1;1,59;78;30,85320992;60;I350;2/9/16;1;0;0;0;0;0;0;0;1;0;0;88;75;0;0;4/9/16;2;1;;1;0;0;0;
3086518;66;1;1;0;1;1;1,62;65,3;24,88187776;28;I25100;2/9/16;0;0;0;1;0;0;0;0;0;1;0;0;0;0;0;0;;-42002;0;;0;0;0;0;
402071;69;0;1;1;0;0;1,63;68;25,59373706;60;I712;2/9/16;0;0;0;0;0;0;0;1;0;0;1;0;95;50;0;0;4/9/16;2;1;;1;0;0;0;EXITUS
1235714;78;0;0;1;;1;1,6;67;26,171875;;Z952;2/9/16;0;0;0;0;0;0;0;1;0;1;0;0;0;0;0;0;;-42002;0;;0;0;0;0;
1114052;73;1;0;1;1;0;1,63;70;26,34649403;40;I25100;5/9/16;0;0;0;1;0;0;0;0;0;1;0;0;78;71;0;0;;-42005;0;;0;0;0;0;
3087267;54;1;1;1;1;0;;111;;60;I21;5/9/16;0;0;0;1;0;0;0;0;0;1;0;0;68;62;0;0;;-42005;0;;0;0;0;0;
639159;64;1;2;1;1;0;1,72;62,5;21,12628448;40;I340;5/9/16;0;0;0;0;1;0;1;0;0;1;0;0;144;93;0;0;9/9/16;4;1;;1;0;0;1;
1427522;78;0;;0;0;0;1,51;59;25,87605807;60;I350;6/9/16;1;0;0;0;0;0;0;0;0;1;0;0;68;54;1;0;11/9/16;5;1;;1;0;0;0;
918063;47;1;1;1;0;0;1,76;78;25,18078512;60;Q231;6/9/16;1;0;0;0;0;0;0;0;0;1;0;0;81;67;0;0;;-42006;0;;0;0;0;0;
1167947;65;0;0;0;0;0;1,6;91;35,546875;60;I25100;6/9/16;0;0;0;1;0;0;0;0;0;0;1;0;70;52;0;0;;-42006;0;;0;0;0;0;EXITUS
1214070;79;1;;1;1;0;1,53;62;26,48553975;63;I350;7/9/16;1;0;0;0;0;0;0;0;0;1;0;0;55;45;0;0;;-42007;0;;0;0;0;0;
3080108;67;1;2;1;1;1;1,71;90;30,77870114;40;I25100;9/9/16;0;0;0;1;0;0;0;0;0;1;0;0;74;65;0;0;;-42009;0;;0;0;0;0;
1303055;58;0;1;0;1;1;1,51;53;23,24459454;30;I350;9/9/16;0;0;0;0;1;1;0;0;0;1;0;0;90;73;0;0;12/9/16;3;1;;0;0;1;1;
1160299;77;1;1;1;1;1;1,76;72;23,24380165;40;I2109;12/9/16;0;0;0;1;0;0;0;0;0;1;0;0;0;0;0;0;;-42012;0;;0;0;0;0;
3086046;84;0;0;1;0;0;1,55;87;36,21227888;60;I350;12/9/16;0;0;0;0;0;0;0;0;1;1;0;0;108;73;0;0;12/9/16;0;1;;0;0;1;0
902250;66;1;2;1;1;1;1,64;82;30,48780488;45;T827;12/9/16;0;0;0;0;0;0;0;1;0;1;0;0;0;0;0;0;;-42012;0;;0;0;0;0
3087969;63;1;2;1;1;1;1,65;80;29,38475666;50;I25100;14/9/16;0;0;0;1;0;0;0;0;0;1;0;0;41;35;0;0;16/9/16;2;1;;1;0;0;0
447304;85;0;0;1;1;0;1,54;94;39,63568899;67;I3511;14/9/16;1;0;0;0;0;0;0;0;0;1;0;0;47;40;0;0;16/9/16;2;1;;1;0;0;1
3086016;74;1;2;1;1;0;1,65;69,8;25,63820018;70;I350;15/9/16;0;0;0;0;1;1;0;0;0;1;0;0;100;95;1;0;;-42015;0;;0;0;0;0
898454;73;1;1;0;1;0;1,72;82;27,71768524;60;I080;15/9/16;1;1;0;0;0;0;0;0;0;1;0;0;141;128;0;0;;-42015;0;;0;0;0;0
3082340;57;1;1;1;1;0;1,63;67;25,21735858;30;I350;15/9/16;1;0;0;0;0;0;0;0;0;1;0;0;78;68;0;0;;-42015;0;;0;0;0;0
3082346;63;1;0;0;0;1;1,72;53;17,91508924;30;I350;16/9/16;1;0;0;0;0;0;0;0;0;1;0;0;80;67;0;0;27/9/16;11;1;;1;0;0;1
636551;57;1;0;1;1;0;1,76;92;29,70041322;60;I359;16/9/16;0;0;0;0;0;0;0;0;1;1;0;0;97;123;1;1;;-42016;0;;0;0;0;0
3088364;60;1;1;0;0;0;1,85;72;21,03725347;55;I25100;16/9/16;0;0;0;1;0;0;0;0;0;0;0;1;;;0;0;19/9/16;3;1;;1;0;0;0
1269415;80;1;2;1;0;0;1,67;85;30,47796622;30;G9751;19/9/16;0;0;0;1;0;0;0;0;0;1;0;0;;;1;1;21/9/16;2;1;;1;0;0;1
853918;80;0;0;1;0;0;1,63;86;32,36854981;55;I340;20/9/16;0;1;0;0;0;0;0;0;0;1;0;0;120;102;0;0;26/9/16;6;1;;1;0;0;1
1293271;80;1;0;1;0;0;1,66;75;27,21730295;60;I359;20/9/16;1;0;0;0;0;0;0;0;0;1;0;0;74;64;0;0;;-42020;0;;0;0;0;0
1359802;61;1;2;1;1;1;1,52;66;28,56648199;60;I350;22/9/16;1;0;0;0;0;0;0;0;0;1;0;0;78;69;0;0;24/9/16;2;1;;1;0;0;1
3088642;80;0;0;1;1;0;1,56;68;27,94214333;60;I340;22/9/16;0;1;0;0;0;0;0;0;0;1;0;0;62;54;0;0;;-42022;0;;0;0;0;0
561636;70;0;0;1;1;0;1,55;73;30,38501561;60;I710;22/9/16;0;0;0;0;0;0;0;1;0;0;1;0;110;68;0;0;22/9/16;0;1;;0;0;1;0
550478;67;1;2;1;1;0;1,62;76,8;29,26383173;60;I350;23/9/16;0;0;0;0;1;1;0;0;0;1;0;0;158;127;0;0;23/9/16;0;1;;0;1;0;0
3085608;62;1;2;1;1;0;1,69;84;29,41073492;60;I25100;23/9/16;0;0;0;1;0;0;0;0;0;1;0;0;72;63;0;0;;-42023;0;;0;0;0;0
3087806;62;0;2;1;1;0;1,55;55;22,89281998;60;I25100;26/9/16;0;0;0;1;0;0;0;0;0;1;0;0;102;25;0;0;;-42026;0;;0;0;0;0
1248468;75;1;0;1;1;1;1,57;66;26,77593411;55;I350;26/9/16;1;0;0;0;0;0;0;0;0;1;0;0;98;79;0;0;29/9/16;3;1;;1;0;0;0
3036322;67;1;2;1;1;0;1,69;79;27,66009594;65;I25100;26/9/16;0;0;0;1;0;0;0;0;0;1;0;0;55;48;0;0;;-42026;0;;0;0;0;0
3077591;78;0;0;1;1;1;1,55;85;35,3798127;59;I359;27/9/16;1;0;0;0;0;0;0;0;1;0;0;38;29;0;0;27/9/16;0;1;;1;0;0;0
462339;65;1;1;1;1;0;1,66;91;33,02366091;60;I350;27/9/16;0;0;1;0;0;0;0;0;0;1;0;0;83;70;0;0;;-42027;0;;0;0;0;0
Line 38, Column 76
Tab Size: 4
Plain Text

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

@patricsg

@patri_1871

www.ispasturias.es

Plataforma de Bioestadística y Epidemiología

Instituto de Investigación Sanitaria del Principado de Asturias (ISPA)

Fundación para la Investigación e Innovación Biosanitaria del Principado de Asturias (FINBA)

ISPA
Instituto de Investigación Sanitaria del Principado de Asturias

FINBA
Fundación para la Investigación y la Innovación Biosanitaria del Principado de Asturias