Data in Spreadsheets

Stat 133 with Gaston Sanchez

Leia Organa

Female

1.50m tall

Luke Skywalker

Male

1.72m tall

Han Solo

Male

1.80m tall

Working with spreadheets?

name	gender	height
Leia Organa	female	1.50
Luke Skywalker	male	1.72
Han Solo	male	1.80

Analyst /Scientist

Tables in Spreadsheets

	A	B	C
1	name	gender	height
2	Leia Organa	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

Storing a data table

Many people enter and store their data in spreadsheets

e.g. MS Excel, Google Sheets, Apple Numbers

Using spreadsheet provides a nice graphical display of a table's content

Using spreadsheet software brings (a deceptive) comfort

In my humble opinion

Spreadsheets do have a role and a place in the toolkit of a data scientist.

In fact, they could be used in any stage of the Data Analysis Cycle.

But keep in mind that they enormously **reduce reproducibility**. And they should not be used as your default data-storage option.

Data in spreadsheets ...

Are so ubiquitous

Can be easy to work with

But can be a sloppy mess

Let’s discuss Karl Broman’s proposed recommendations when organizing data in spreadsheets.

https://kbroman.org/dataorg/

Data organization in spreadsheets.

By Karl Broman, and Kara Woo (2018)

The American Statistician 78:2–10 (doi:10.1080/00031305.2017.1375989)

Be Consistent

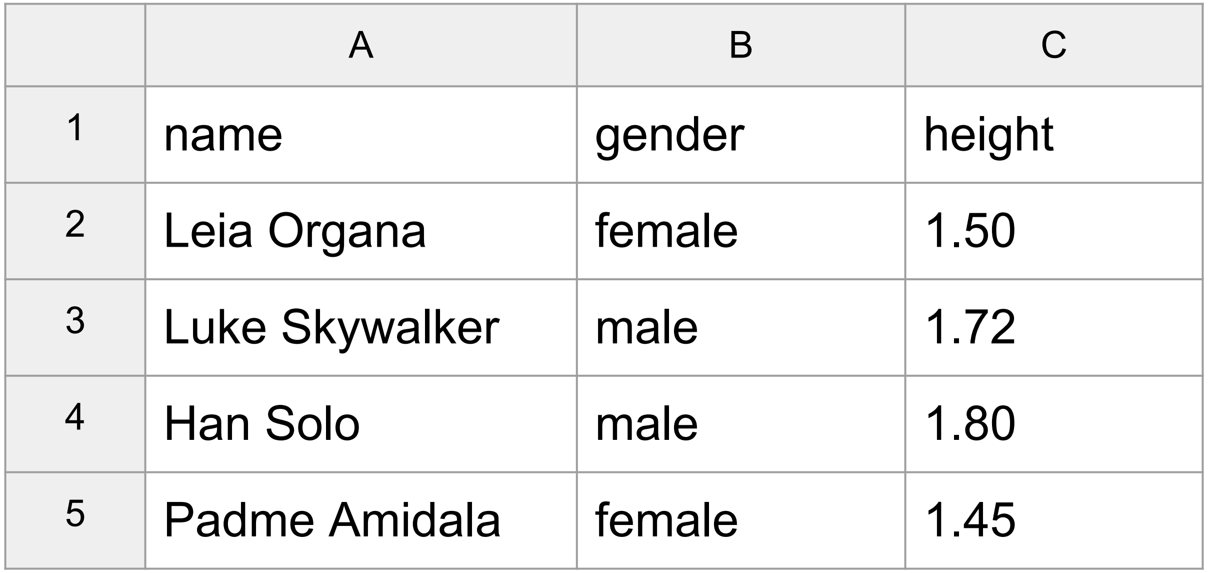
Avoid inconsistent codes for variables (don’t)

don’t

	A	B	C
1	name	gender	height
2	Leia Organa	female	1.50
3	Luke Skywalker	MALE	1m 72 cm
4	Han Solo	male	1.80
5	Padme Amidala	F	145 cm

Use consistent codes for variables (do)

do



	A	B	C
1	name	gender	height
2	Leia Organa	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80
5	Padme Amidala	female	1.45

Avoid several codes for missing values (don’t)

don’t

	A	B	C
1	name	gender	height
2	Leia Organa	female	1.50
3	Luke Skywalker	male	?
4	Han Solo	male	1.80
5	Padme Amidala	female	9999

Use a single fixed code for missing values (do)

do

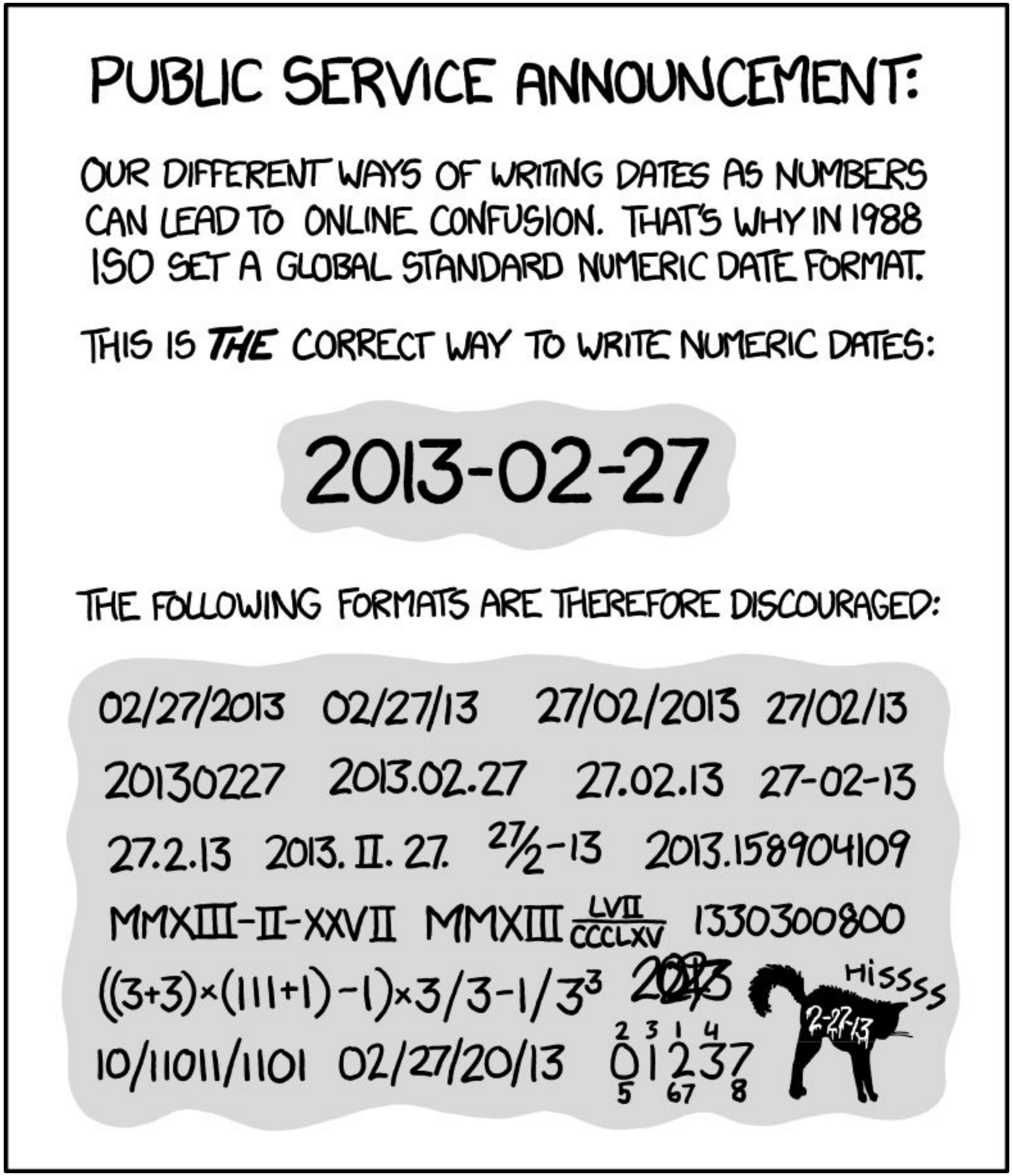
	A	B	C
1	name	gender	height
2	Leia Organa	female	1.50
3	Luke Skywalker	male	NA
4	Han Solo	male	1.80
5	Padme Amidala	female	NA

Consistency in general

Naming style:

- Variables
- IDs
- NAs
- Files
- Dates
- Layouts
- Annotations

Dates


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
10/11011/1101 02/27/20/13 01237
HISSSSS
2-27-13

https://xkcd.com/1179/

No empty cells

Avoid blank/empty cells (don’t)

don’t

	A	B	C
1	name	gender	titles
2	Leia Organa	female	princess
3			senator
4			general
5	Luke Skywalker	male	knight
6			master
7	Han Solo	male	captain

Fill in all cells (do)

do

	A	B	C
1	name	gender	titles
2	Leia Organa	female	princess
3	Leia Organa	female	senator
4	Leia Organa	female	general
5	Luke Skywalker	male	knight
6	Luke Skywalker	male	master
7	Han Solo	male	captain

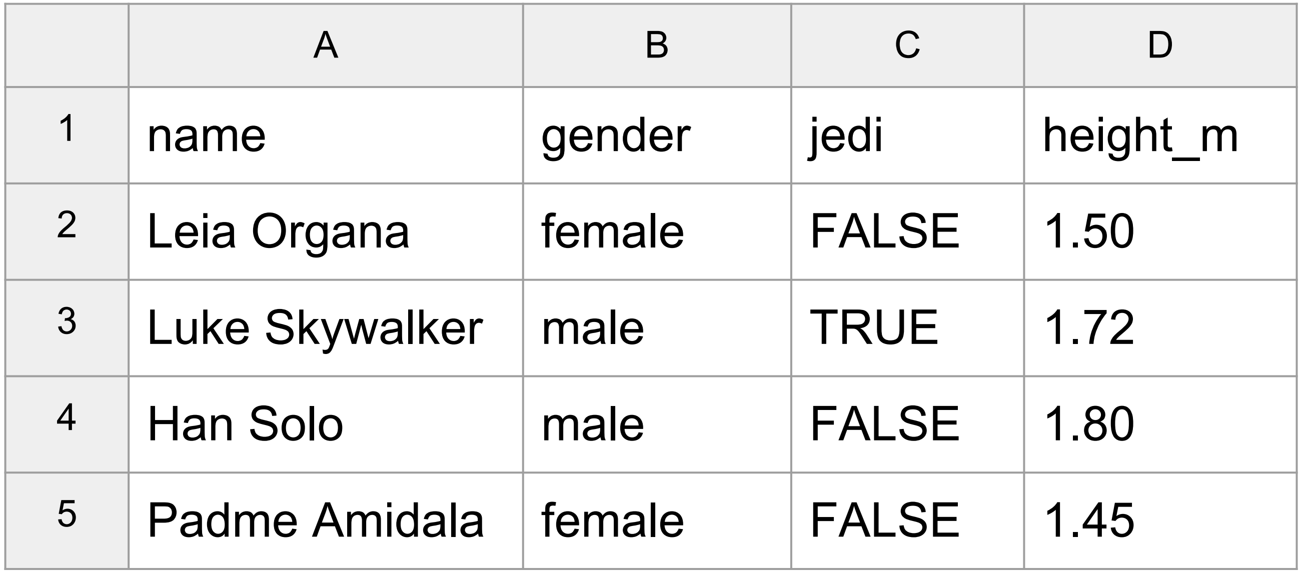
Various things in a cell (don't)

don't

	A	B	C
1	name	gender-jedi	height
2	Leia Organa	female (no)	1.50 m
3	Luke Skywalker	male (yes)	1.72 m
4	Han Solo	male (no)	1.80 m
5	Padme Amidala	female (no)	1.45 m

Put just one thing in a cell (do)

do



	A	B	C	D
1	name	gender	jedi	height_m
2	Leia Organa	female	FALSE	1.50
3	Luke Skywalker	male	TRUE	1.72
4	Han Solo	male	FALSE	1.80
5	Padme Amidala	female	FALSE	1.45

Tidy Data

Data tidying

Tidiness: *“The state or quality of being arranged neatly and in order”*

Tidy Data:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

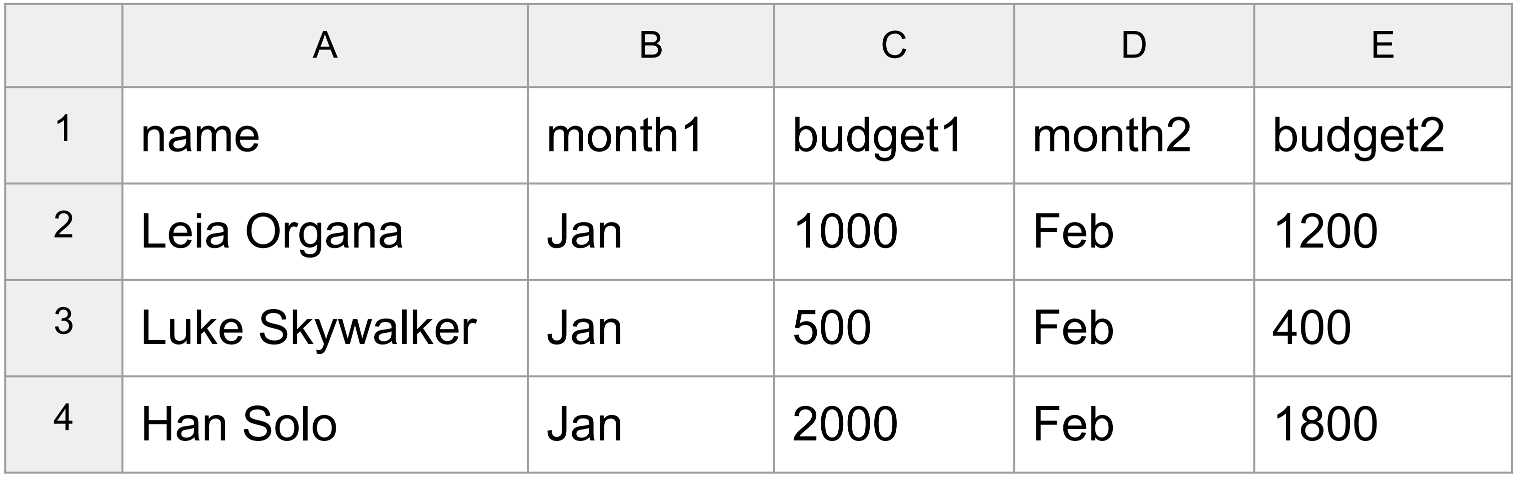
Messy

	A	B	C
1	name	gender-jedi	height
2	Leia Organa	female (no)	1.50 m
3	Luke Skywalker	male (yes)	1.72 m
4	Han Solo	male (no)	1.80 m
5	Padme Amidala	female (no)	1.45 m

Tidy

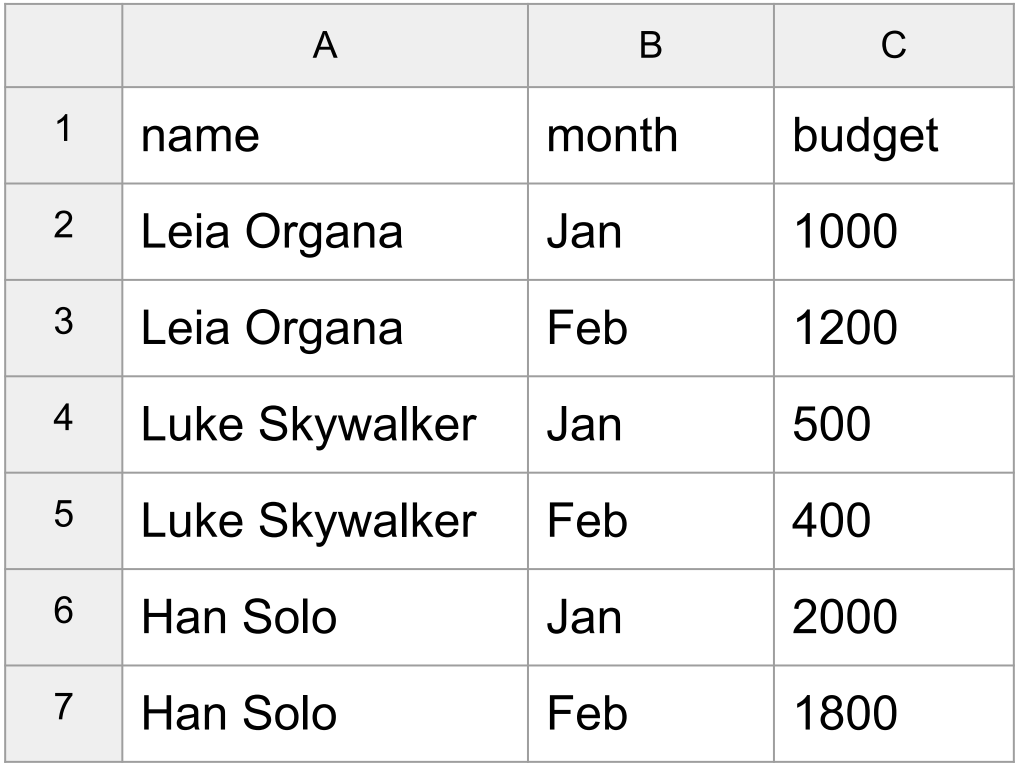
	A	B	C	D
1	name	gender	jedi	height_m
2	Leia Organa	female	FALSE	1.50
3	Luke Skywalker	male	TRUE	1.72
4	Han Solo	male	FALSE	1.80
5	Padme Amidala	female	FALSE	1.45

Messy



	A	B	C	D	E
1	name	month1	budget1	month2	budget2
2	Leia Organa	Jan	1000	Feb	1200
3	Luke Skywalker	Jan	500	Feb	400
4	Han Solo	Jan	2000	Feb	1800

Tidy



	A	B	C
1	name	month	budget
2	Leia Organa	Jan	1000
3	Leia Organa	Feb	1200
4	Luke Skywalker	Jan	500
5	Luke Skywalker	Feb	400
6	Han Solo	Jan	2000
7	Han Solo	Feb	1800

More on Tidy Data ...

Tidy Data paper

https://vita.had.co.nz/papers/tidy-data.pdf

Tidy data vignette

https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html

Chapter 12: Tidy Data (R4DS)

https://r4ds.had.co.nz/tidy-data.html

Create a Data Dictionary

Data Dictionary File

Use a separate file that explains what all the variables are.

This file is what some authors call **metadata** (information about your data).

I recommend using a plain text file (`.txt` or `.md`) to create a data dictionary.

Data Dictionary contents

The exact variable name as in the data file

A longer explanation about what the variable means

Suggested data type (e.g. int, real, boolean)

The measurement units

How missing values are codified (if any)

Expected minimum and maximum, perhaps

Say you have some data like this

	A	B	C	D
1	name	gender	jedi	height_m
2	Leia Organa	female	FALSE	1.50
3	Luke Skywalker	male	TRUE	1.72
4	Han Solo	male	FALSE	1.80
5	Padme Amidala	female	FALSE	NA

Data Dictionary: example

name: first and last name of an individual (character)

gender: reported gender “female”, “male” (character)

jedi: is the individual a jedi knight? TRUE, FALSE (logical)

height: height in meters; missing values as NA (real or double)

No calculations

No calculations in the raw data files

Many users include calculations and graphs in spreadsheet files.

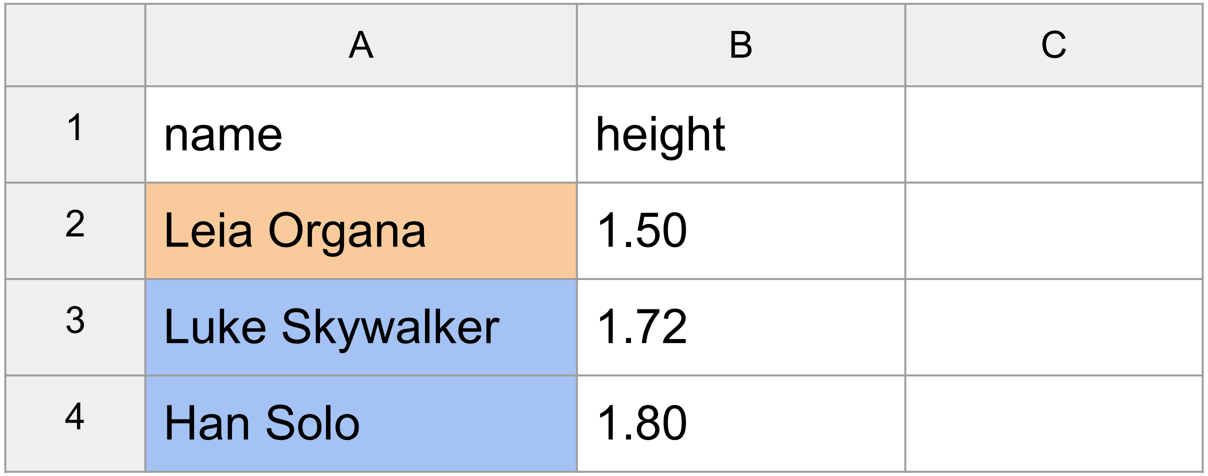
Doing calculations imply opening a file and typing things (running the risk of typing junk).

Your primary data file should contain just the data and nothing else (no calculations, no graphs).

Enriched Formatting

Avoid color/highlighting as data

don't

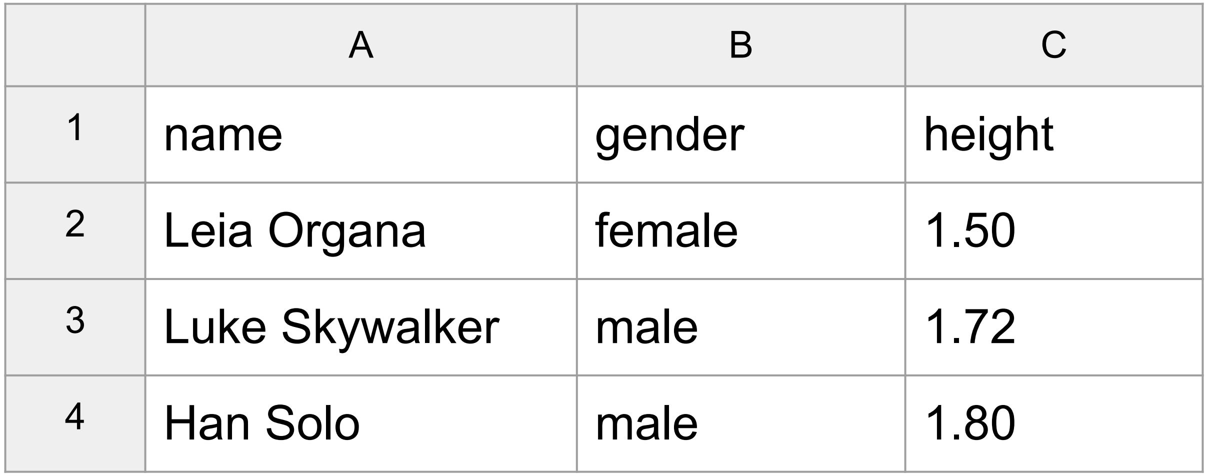


	A	B	C
1	name	height	
2	Leia Organa	1.50	
3	Luke Skywalker	1.72	
4	Han Solo	1.80	

female *male*

Avoid color/highlighting as data

do



	A	B	C
1	name	gender	height
2	Leia Organa	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

Save data in plain text files

Spreadsheet inconveniences

Excel files (.xls) are NOT text files

They are **enriched** files with added format elements

Cannot be opened with a text editor

You typically depend on ***proprietary*** software, and commercial fees perhaps

What happens when you try to open a spreadsheet file in a text editor?

Still want to save tables as **`.xlsx`** (or **`.xls`**)?

Every time you save a data file in **.xlsx** format ...

God kills a kitten

Good Practice

Whenever you work with some source of data stored in a native spreadsheet format (e.g. .xls, .xlsx, .numbers), always generate a text version (e.g. **`.csv, .txt, .dat`**)