UCDAVIS DataLab
Data Science and Informatics

Excelling with Excel:

Best Practices for Keeping Your Data Tidy

February 8th, 2021

Dr. Pamela Reynolds, Victoria Farrar

plreynolds@ucdavis.edu
vsfarrar@ucdavis.edu

Please keep videos off and stay muted to reduce bandwidth.

Please note: Today's workshop will be recorded and posted to DataLab's website.

UC DAVIS
DataLab
Data Science and Informatics

Research
Training
Community

»» Increase UC Davis' research impact via data-driven expertise.

»» Support the next generation of data-capable researchers and students.

»» Foster and coordinate data-enabled researchers and university units.

Supporting innovation, accelerating research for the entire community.

http://datalab.ucdavis.edu

Introductory Poll

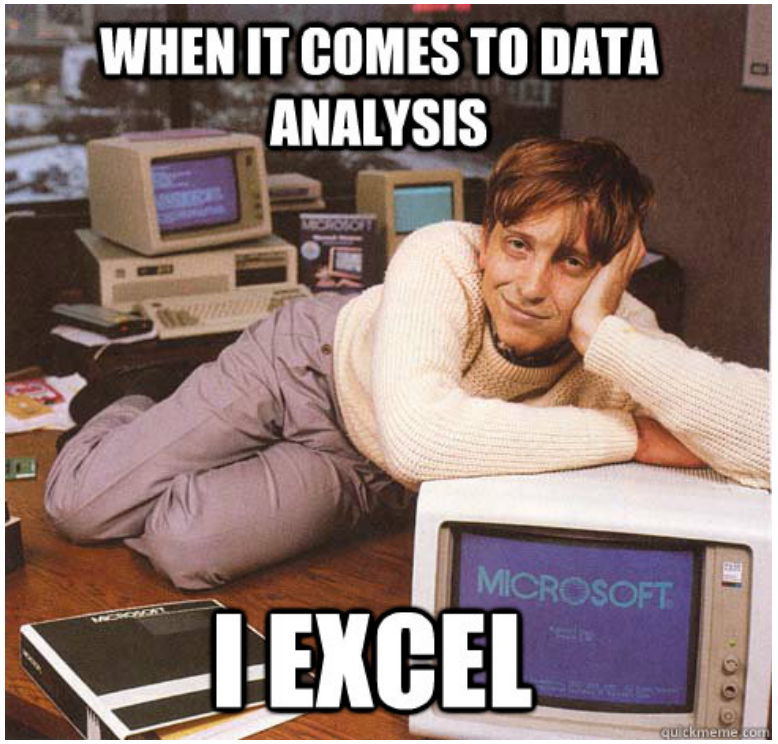
What do you, or will you, collect data on?
(i.e. What is your “study system”?)

Drop your responses in the chat.

UCDAVIS
DataLab
Data Science and Informatics

Workshop Goals

- Share data processor / spreadsheet manager tips & tricks to take away and use in your own research process
- Introduce data management best practices



Learning Objectives

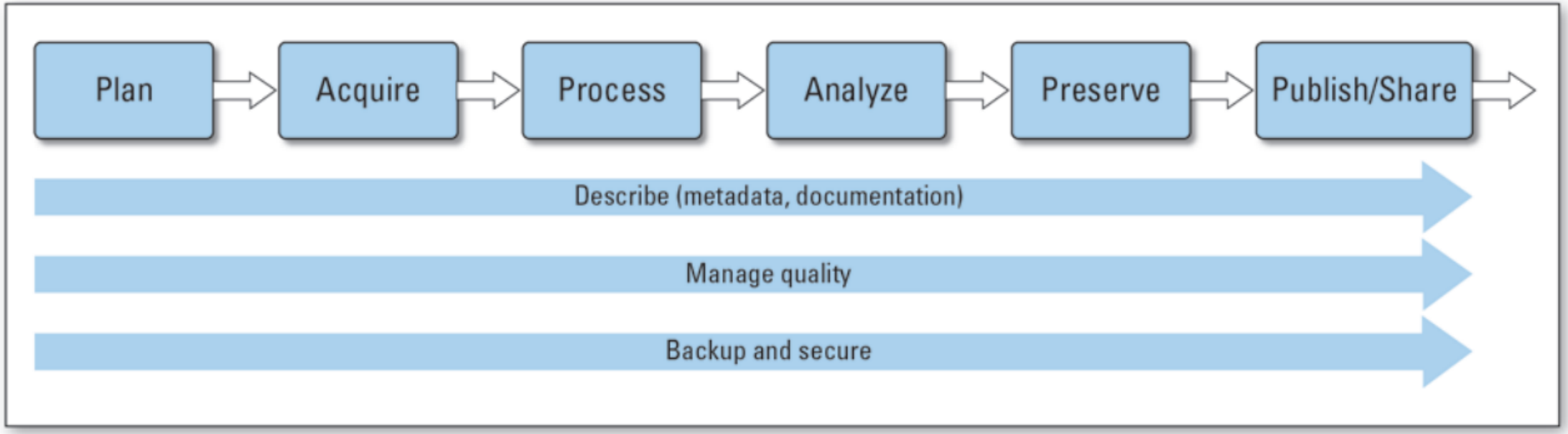
By the end of this workshop, you should be able to:

- Describe how to organize a data-driven project
- Identify best practices for spreadsheet formatting and data entry
- Use basic tools in Microsoft Excel for data validation and filtering
- Define restricted vocabularies and data dictionaries
- Compare the advantages of different data file formats

Before Excel:
Setting Up Your Data Project

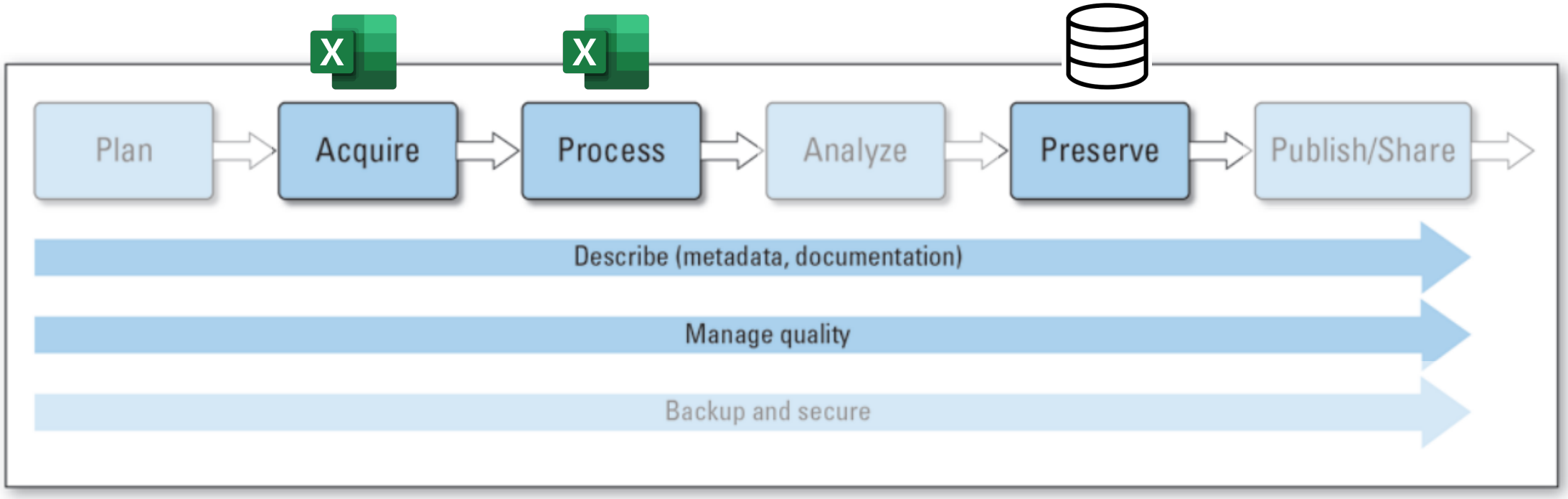
UCDAVIS
DataLab
Data Science and Informatics

The Data Lifecycle



Source: USGS

The Data Lifecycle



Source: USGS

Breakout Room Activity

Access and download the dataset
`woc_congress_messy.xlsx`
(see link in chat)

This data contains information on the history of women of color who have held national elected office in the United States Congress.

Wikimedia

nounproject.com

UCDAVIS DataLab
Data Science and Informatics

Breakout Room Activity

In a breakout room with your peers, imagine you want to analyze the data in this spreadsheet.

Introduce yourselves to each other, and then think of a question you would want to answer with this data.

What issues might these spreadsheets present when trying to answer your question?

Designate one person in the group to share your observations in the chat after.

15 minutes

UCDAVIS DataLab
Data Science and Informatics

Debrief: Spreadsheet activity

What did you or your group find challenging with the format?

What would you change?

Drop your responses in the chat.

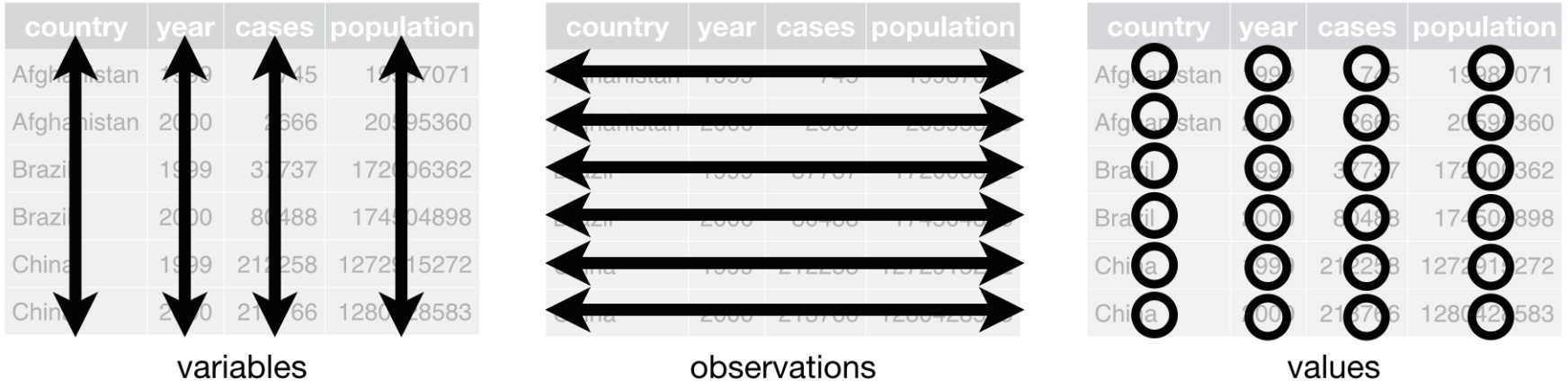
Spreadsheet Best Practices

UCDAVIS
DataLab
Data Science and Informatics

1. Keep variables in columns, observations in rows

Why?

- Consistency
- Facilitates simple import into programming languages



Source: R for Data Science

UCDAVIS DataLab
Data Science and Informatics

1. Keep variables in columns, observations in rows

Example of the issue:

	Senators				
1	Carol Moseley, Illinois	2	Mazie K. Hirono	3	Duckworth, Tammy
	1993-1999		HI		Illinois
	African-American		2007-		2013-present
	39.7817° N, 89.6501° W		Asian-American/Pacific-Islander		Asian-American/Pacific Islander
	Won 38.3% of the vote		21.315603, -157.858093		39° 47' 59.9958", -89° 39' 0.0072"
			Won 62.60% of the vote		Won 64.38 % of the vote
4	Kamala Harris, CA	5	Cortez-Masto, Catherine		
	2017-2021 (note: became Vice Presider		NV		
	African-American		2017-present		
	38° 34' 32.7504" N, 121° 28' 4		Hispanic / Latinx		
	Won 61.60 % of the vote		36.114647, -115.172813		
			Won 47.10% of the vote		

UCDAVIS DataLab
Data Science and Informatics

1. Keep variables in columns, observations in rows

Better:

Name	State	Term of Service	Ethnicity	State Capitol Coordinate	Percent of Vote Won
Carol Moseley	Illinois	1993-1999	African-American	39.7817° N, 89.6501°	0.38
Mazie K. Hirono	HI	2007-	Asian-American/Pacifi	21.315603, -157.8580	62.6
Duckworth, Tammy	Illinois	2013-present	Asian-American/Pacifi	39° 47' 59.9958", -89°	64.38
Kamala Harris	CA	2017-2021 (note: became Vice President of US in 20	African-American, Asia	38° 34' 32.7504" N, 12	0.62
Cortez-Masto, Catherir	NV	2017-present	Hispanic/Latinx	36.114647, -115.1728	47.1

2. Stick to one data type per column / variable.

Why?

- Facilitates analyses down the line (e.g. sort, summarize)
- Keeps data “clean” and less likely to be corrupted/lost

2. Stick to one data type per column / variable.

Example of the issue:

Name	State	Term of Service	Ethnicity	State Capitol Coordinates	Percent of Vote Won
Carol Moseley	Illinois	1993-1999	African-American	39.7817° N, 89.6501°	0.38
Mazie K. Hirono	HI	2007-	Asian-American/Pacif	21.315603, -157.8580	62.6
Duckworth, Tammy	Illinois	2013-present	Asian-American/Pacif	39° 47' 59.9958", -89°	64.38
Kamala Harris	CA	2017-2021 (note: became Vice President of US in 20	African-American, Asi	38° 34' 32.7504" N, 12	0.62
Cortez-Masto, Catherir	NV	2017-present	Hispanic/Latinx	36.114647, -115.1728	47.1

2. Stick to one data type per column / variable.

Better:

Name	State	Start of Term	End of Term	Ethnicity	State Capitol Latitude	State Capitol Longitude	Percent of Vote Won
Carol Moseley	Illinois	1993	1999	African-American	39.7817° N	89.6501° W	0.38
Mazie K. Hirono	HI	2007	NA	Asian-American/Pacific-	21.315603	-157.858093	62.6
Duckworth, Tammy	Illinois	2013	NA	Asian-American/Pacific	39° 47' 59.9958"	-89° 39' 0.0072"	64.38
Kamala Harris	CA	2017	2021 (note: became Vice President of the US)	African-American, Asian	38° 34' 32.7504" N	121° 28' 43.8636'' W.	0.62
Cortez-Masto, Catherine	NV	2017	NA	Hispanic/Latinx	36.114647	-115.172813	47.1

3. Add notes / comments in separate columns.

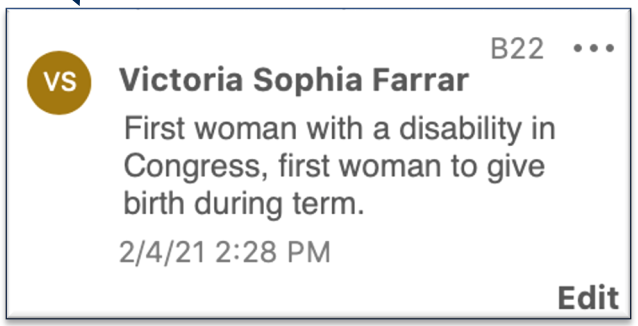
Why?

- Keeps data atomized and makes it easier to find information
- Facilitates downstream analyses *(Same as #2)*

3. Add notes and comments in separate columns.

Example of the issue:

Name	State	Start of Term	End of Term	Ethnicity	State Capitol Latitude	State Capitol Longitude	Percent of Vote Won
Carol Moseley	Illinois	1993	1999	African-American	39.7817° N	89.6501° W	0.38
Mazie K. Hirono	HI	2007	NA	Asian-American/Pacific-	21.315603	-157.858093	62.6
Duckworth, Tammy	Illinois	2013	NA	Asian-American/Pacific	39° 47' 59.9958"	-89° 39' 0.0072"	64.38
Kamala Harris	CA	2017	2021 (note: became Vice President of the US)	African-American, Asian	38° 34' 32.7504" N	121° 28' 43.8636'' W.	0.62
Cortez-Masto, Catherine	NV	2017	NA	Hispanic/Latinx	36.114647	-115.172813	47.1



UCDAVIS DataLab
Data Science and Informatics

3. Add notes and comments in separate columns.

Better:

Name	State	Start of Term	End of Term	Ethnicity	State Capitol Latitude	State Capitol Longitude	Percent of Vote Won	Notes
Carol Moseley	Illinois	1993	1999	African-American	39.7817° N	89.6501° W	0.38	
Mazie K. Hirono	HI	2007	NA	Asian-American/Pacific-	21.315603	-157.858093	62.6	
Duckworth, Tammy	Illinois	2013	NA	Asian-American/Pacific	39° 47' 59.9958"	-89° 39' 0.0072"	64.38	First woman with a disability in Congress, first woman to give birth during term in office
Kamala Harris	CA	2017	2021	African-American, Asian	38° 34' 32.7504" N	121° 28' 43.8636'' W.	0.62	Became Vice President of US
Cortez-Masto, Catherine	NV	2017	NA	Hispanic/Latinx	36.114647	-115.172813	47.1	

UCDAVIS DataLab
Data Science and Informatics

4. Code information explicitly in cells rather than in formatting.

Why?

- Facilitate downstream analyses
- Keep information accessible outside of Excel (formatting is often Excel-specific)

4. Code information explicitly in cells rather than in formatting.

Example of the issue:

Name	State	Start of Term	End of Term	Ethnicity	State Capitol Latitude	State Capitol Longitude	Percent of Vote Won	Notes
Carol Moseley	Illinois	1993	1999	African-American	39.7817° N	89.6501° W	0.38	
Mazie K. Hirono	HI	2007	NA	Asian-American/Pacific-	21.315603	-157.858093	62.6	
Duckworth, Tammy	Illinois	2013	NA	Asian-American/Pacific	39° 47' 59.9958"	-89° 39' 0.0072"	64.38	First woman with a disability in Congress, first woman to give birth during term in office
Kamala Harris	CA	2017	2021	African-American, Asian	38° 34' 32.7504" N	121° 28' 43.8636'' W.	0.62	Became Vice President of US
Cortez-Masto, Catherine	NV	2017	NA	Hispanic/Latinx	36.114647	-115.172813	47.1	

UCDAVIS
DataLab
Data Science and Informatics

4. Code information explicitly in cells rather than in formatting.

Better:

Name	State	Party	Start of Term	End of Term	Ethnicity	State Capitol Latitude	State Capitol Longitude	Percent of Vote Wor	Notes
Carol Moseley	Illinois	Democrat	1993	1999	African-American	39.7817° N	89.6501° W	0.38	
Mazie K. Hirono	HI	Dem	2007	NA	Asian-American/Pacific-	21.315603	-157.858093	62.6	
Duckworth, Tammy	Illinois	D	2013	NA	Asian-American/Pacific	39° 47' 59.9958"	-89° 39' 0.0072"	64.38	First woman with a disability in Congress, first woman to give birth during term in office
Kamala Harris	CA	D	2017	2021	African-American, Asian	38° 34' 32.7504" N	121° 28' 43.8636'' W.	0.62	Became Vice President of US
Cortez-Masto, Catherine	NV	D	2017	NA	Hispanic/Latinx	36.114647	-115.172813	47.1	

UCDAVIS DataLab
Data Science and Informatics

5. Be consistent with variable codes, formats, and units.

Why?

- Facilitate downstream analyses
- Increase interpretability of data for future use

5. Be consistent with variable codes, formats, and units.

Example of the issue:

Name	State	Party	Start of Term	End of Term	Ethnicity	State Capitol Latitude	State Capitol Longitude	Percent of Vote Wo	Notes
Carol Moseley	Illinois	Democrat	1993	1999	African-American	39.7817° N	89.6501° W	0.38	
Mazie K. Hirono	HI	Dem	2007	NA	Asian-American/Pacific-	21.315603	-157.858093	62.6	
Duckworth, Tammy	Illinois	D	2013	NA	Asian-American/Pacific	39° 47' 59.9958"	-89° 39' 0.0072"	64.38	First woman with a disability in Congress, first woman to give birth during term in office
Kamala Harris	CA	D	2017	2021	African-American, Asian	38° 34' 32.7504" N	121° 28' 43.8636'' W.	0.62	Became Vice President of US
Cortez-Masto, Catherine	NV	D	2017	NA	Hispanic/Latinx	36.114647	-115.172813	47.1	

5. Be consistent with variable codes, formats, and units.

Better:

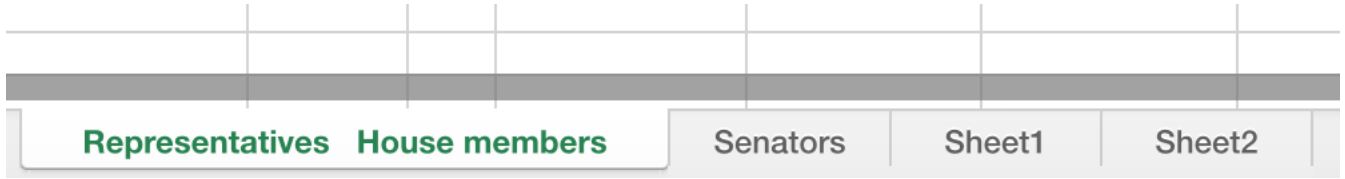
Name	State	Party	Start of Term	End of Term	Ethnicity	State Capitol Latitude	State Capitol Longitude	Percent of Vote Wor	Notes
Mosely, Carol	IL	D	1993	1999	African-American	39.799999	-89.650002	38	
Hirono, Mazie K.	HI	D	2007	NA	Asian-American/Pacific	21.315603	-157.858093	62.6	
Duckworth, Tammy	IL	D	2013	NA	Asian-American/Pacific	39.799999	-89.650002	64.38	First woman with a disability i
Harris, Kamala	CA	D	2017	2021	African-American, Asian	38.575764	-121.478851	62	Became Vice President of US
Cortez-Masto, Catherine	NV	D	2017	NA	Hispanic/Latinx	36.114647	-115.172813	47.1	

6. Store one dataset per file.

Why?

- Separate tabs are data processor-specific and will be lost in other formats
- Keeps data easy to search and find

Example of the issue:



6. Store one dataset per file.

Better:

One single file

`woc_congress.xls`

OR

Two separate files with similar formatting

`woc_senators.xls`

`woc_representatives.xls`

Poll Time!

Which of these have you come across “in the wild”?

- ❑ Variables and observations not in columns and rows
- ❑ Inconsistent codes or units within a single variable
- ❑ Stored information in colors without hard-coding it
- ❑ Put comments or notes in cells with data
- ❑ Multiple spreadsheets/tabs with separate datasets

7. Store data in a “raw” format.

Freepik, Flaticon.com

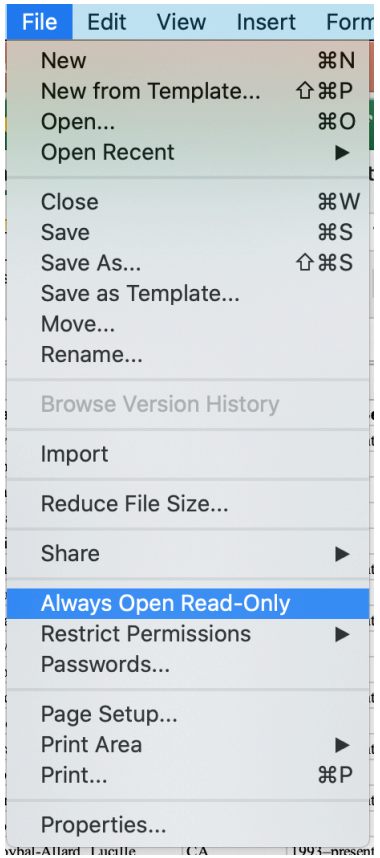
`data_original_raw.xls`

`data_cleaned_raw.xls`

UCDAVIS DataLab
Data Science and Informatics

7. Store data in a “raw” format.

Tip: Set files to always open Read-Only



Why?

Reduces risk of accidental edits, data erasures, or manipulations.

Keeps processed / clean data clean.

UCDAVIS DataLab
Data Science and Informatics

Activity

Try it:
What happens when you open an .xlsx file in a word editor, like TextEdit or WordPad / NotePad?

Tell us about it in the chat!

8. Save your files in a non-proprietary format.

UCDAVIS
DataLab
Data Science and Informatics

8. Save your files in a non-proprietary format.

- .xls/x formats are limited
- For long-term and collaborative use, use formats like
  - **.csv** : comma-separated values
  - **.txt** : tab-delimited text

woc_congress_clean.csv

```
1  Last Name,First (Middle) Name,Position,State,Party,Year Term Began,Year Term Ended,Term
   Number,Ethnicity,Notes
2  Lee,Barbara,U.S. Representative,CA,D,1997,NA,1,African-American,
3  Watson,Diane,U.S. Representative,CA,D,2001,2011,1,African-American,
4  Holmes Norton,Eleanor,U.S. Representative,DC,D,1991,NA,1,African-American,
5  Meek,Carrie,U.S. Representative,FL,D,1993,2003,1,African-American,
```

woc_congress_clean.txt

```
1  Last Name   First (Middle) Name Position    State   Party   Year Term Began Year Term Ended Term
   Number Ethnicity   Notes
2  Lee Barbara U.S. Representative CA  D   1997    NA  1   African-American
3  Watson  Diane   U.S. Representative CA  D   2001    2011    1   African-American
4  Holmes Norton   Eleanor U.S. Representative DC  D   1991    NA  1   African-American
5  Meek    Carrie  U.S. Representative FL  D   1993    2003    1   African-American
```

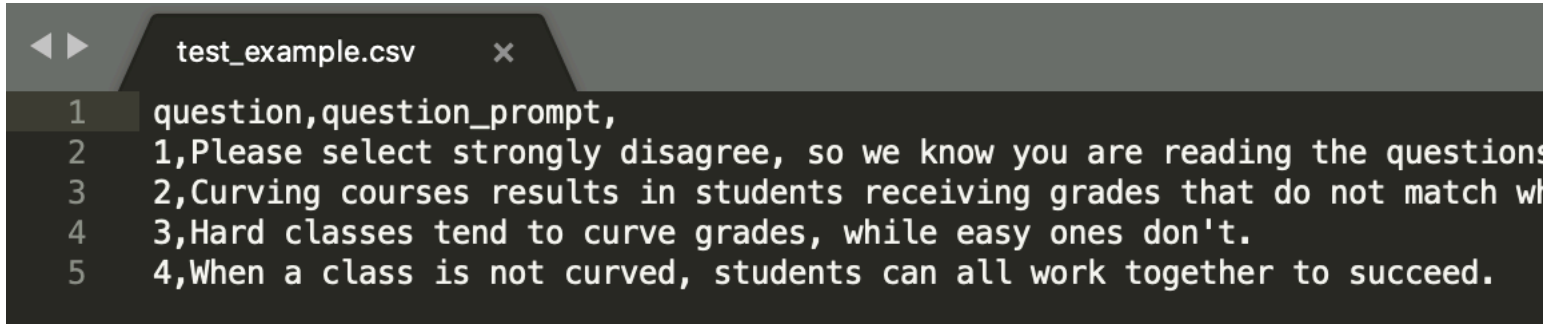
UCDAVIS DataLab
Data Science and Informatics

8. Save your files in a non-proprietary format.

Watch out! .csv may not be the best format for text data.

Use .txt instead.

Looks normal in .csv …

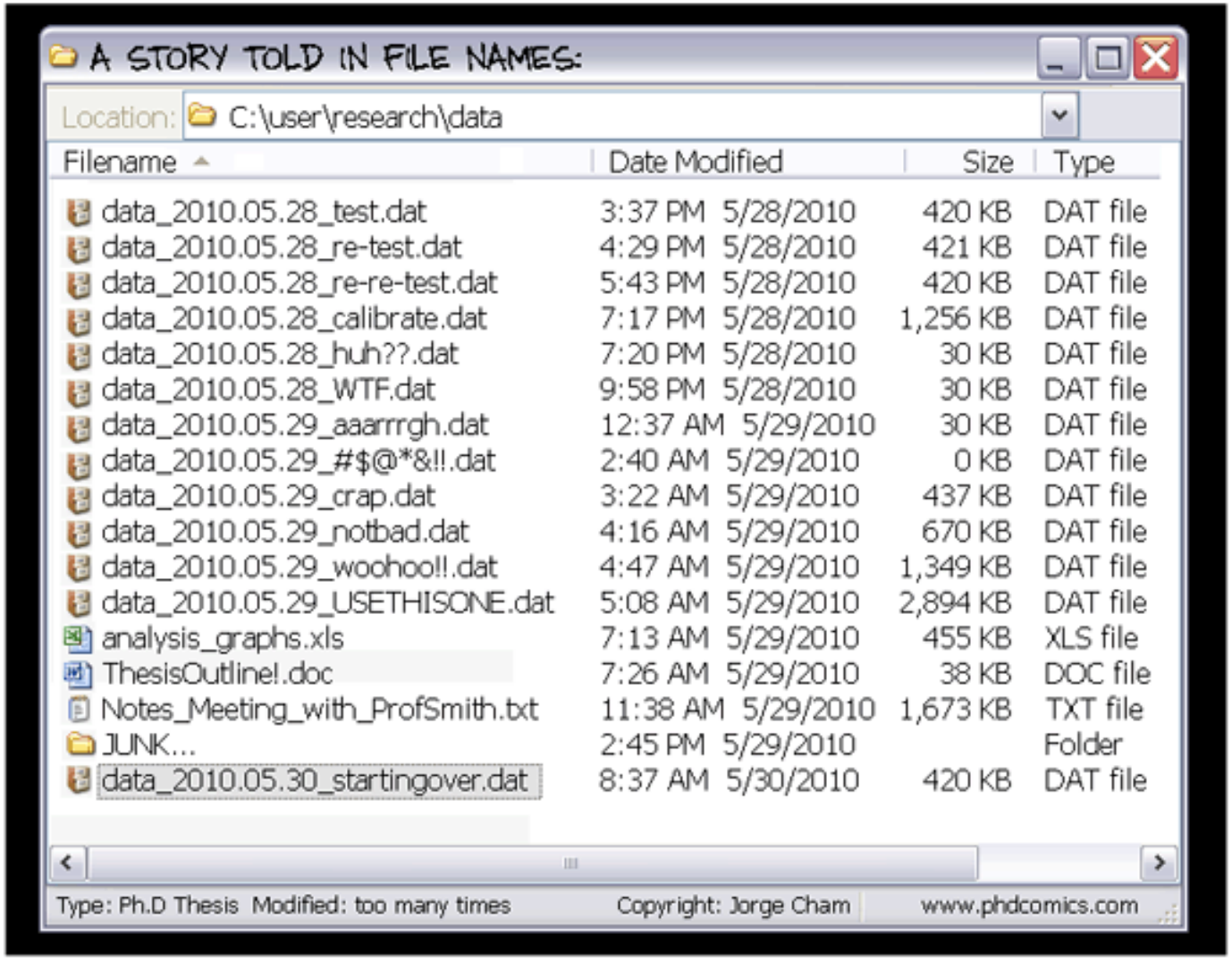


In R programming language:

	question	question_prompt	X
1	1	Please select strongly disagree	so we know you are reading the questions.
2	2	Curving courses results in students receiving grades t…	
3	3	Hard classes tend to curve grades	while easy ones don't.
4	4	When a class is not curved	students can all work together to succeed.

UCDAVIS DataLab
Data Science and Informatics

9. Control your versions (or they will control you)!



UCDAVIS DataLab
Data Science and Informatics

9. Control your versions (or they will control you)!

- **File names should be**:
  - Human-Readable
  - Machine-Readable
  - Works well with default ordering

9. Control your versions (or they will control you)!

- **File names should be**:
  - Human-Readable
    - Origin date or date file was processed
    - Study system
    - Author name
    - Very brief description

OK:

`woc_congress.csv`

Better:

`woc_congress original uncleaned data VSF Feb.3.2021.csv`

UCDAVIS DataLab
Data Science and Informatics

9. Control your versions (or they will control you)!

- **File names should be**:
  - Human-Readable
  - Machine-Readable
    - Avoid spaces, punctuation, special characters in names
    - Use dashes, underscores, or CamelCase

OK:

`woc_congress original uncleaned data VSF Feb.3.2021.csv`

Better:

`woc_congress_original_uncleaned_data_VSF_Feb-3-2021.csv`

9. Control your versions (or they will control you)!

- **File names should be**:
  - Human-Readable
  - Machine-Readable
  - Works well with default ordering
    - Use ISO 8601 standard dates
      - YYYY-MM-DD

OK:

`woc_congress_original_uncleaned_data_VSF_ Feb-3-2021.csv`

Better:

`2021-02-03_woc_congress_original_uncleaned_VSF.csv`

UCDAVIS DataLab
Data Science and Informatics

9. Control your versions (or they will control you)!

- For more information on version-control software, like **Git**, check out upcoming workshops at DataLab
- **Introduction to Version Control with Git**
  - February 24th at 3:00 pm

UCDAVIS
DataLab
Data Science and Informatics

10. Document your data.

Important to record the decisions you make during data processing.

- Column names and datatypes
- Variable codes
- Datasets

metadata

data about the data

Describe (metadata, documentation)

UCDAVIS DataLab
Data Science and Informatics

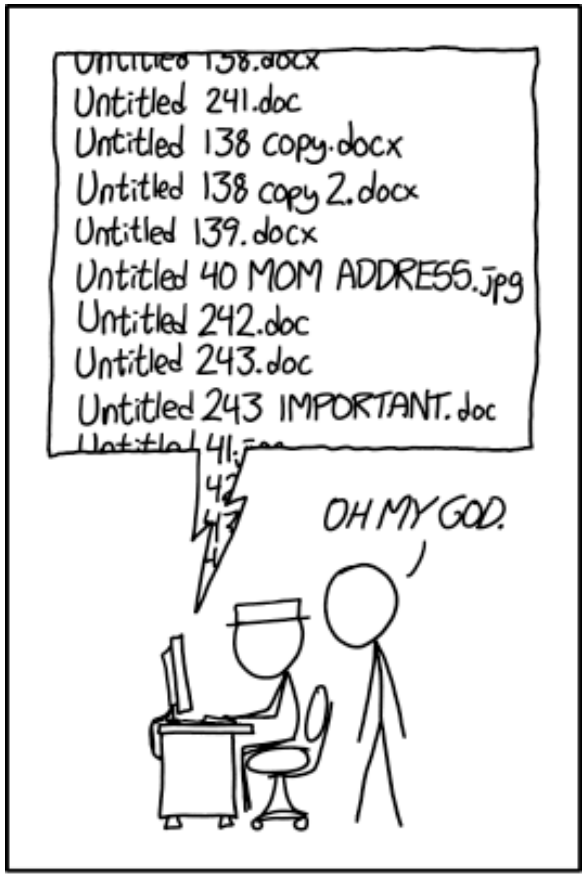
10. Document your data.

You might also want to consider project- or directory-level documentation too, which describes how datasets are processed, analyzed, and relate to each other.

Describe (metadata, documentation)

UCDAVIS DataLab
Data Science and Informatics

10. Document your data.



To learn more about metadata and data documentation, check out our upcoming workshop on

Thursday, March 4th

Register here:

https://datalab.ucdavis.edu/eventscalendar/readme-write-me/

UCDAVIS DataLab
Data Science and Informatics

Summary: Spreadsheet Best Practices

1. Keep variables in columns, observations in rows.
2. Stick to one datatype per variable/column.
3. Add notes/comments in separate columns.
4. Code information explicitly in cells rather than in formatting.
5. Be consistent with variable codes, formats and units.
6. Store one dataset per file.
7. Keep your data “raw”.
8. Save your files in a non-proprietary format.
9. Control your versions (or they will control you)!
10. Document your data.

Reader notes can be found here:

https://ucdavisdatalab.github.io/workshop_keeping_data_tidy

UCDAVIS DataLab
Data Science and Informatics

Feedback

Please complete the post-feedback survey!

https://ucdavis.co1.qualtrics.com/jfe/form/SV_eVgG4yvAKWW9fOm

UCDAVIS

DataLab

Data Science and Informatics