Tidying and Cleaning Data

Instructors

Cole Crawford
Humanities Research Computing Specialist, DARTH

Jeremy Guillette
Instructional Technologist, Academic Technology for FAS

Agenda

- Principles of Tidy Data
- Hands-on exercise

Tidy Data

Tidy Data Overview

- Tidy, clean data is critical for analysis
- Tidy datasets share many characteristics
- Cardinal rules for organizing tidy, tabular data
- Recognizing (and avoiding) messy data mistakes
- Dates as data
- File formats

Organizing Research Data

- All research relies on data
- Data don’t need to be big to be useful
- Most data can be represented in a tabular format
- Python and tidy data: reproducible workflow
- Data organization is key for efficient research practices
- Tidy data is required for analysis

“Like families, **tidy datasets** are all alike, but every **messy dataset** is messy in its own way.”

-Hadley Wickham

Structuring Data

- Tidy dataset share key characteristics
- A dataset is a collection of values
- Every value belongs to a variable and an observation

	A	B	C	D
1	id	title	document_type	author
2	983	Occultation of the Twelfth Imam	Secondary Sources	Jassim Hussain
3	982	Mantle of the Prophet	Secondary Sources	Roy Mottahedeh
4	185	Kitāb al-Ghayba	Historical Primary Sources	Shaykh Ṭūṣī
5	986	Kashf al-Asrār	Contemporary Primary So	Ayatullāh Rūhullah Khumaynī
6	392	The Law of the Dubai International Financial Centre: (	Articles	Alejandro Carballo
7	399	Choice of Law and Islamic Finance	Articles	Julio C. Colón
8	400	A Murābaḥah Transaction in an English Court: The Lo	Articles	Kilian Bälz
9	438	Corporate Insolvency in Malaysia	Secondary Sources	Sonali Abeyratne
10	441	Foundations of Forgiveness in Islamic Bankruptcy Law	Secondary Sources	Jason J. Kilborn
11	439	The Surprising Irrelevance of Islamic Bankruptcy	Secondary Sources	Haider Ala Hamoudi
12	1252	The True Story of Sharia in American Courts	Secondary Sources	Abed Awad

Structuring Data

- In spreadsheet terms:
- Spreadsheet: Every cell belongs to a column and an row
- Each type of observation belongs on its own table / spreadsheet / tab

	A	B	C	D
1	id	title	document_type	author
2	983	Occultation of the Twelfth Imam	Secondary Sources	Jassim Hussain
3	982	Mantle of the Prophet	Secondary Sources	Roy Mottahedeh
4	185	Kitāb al-Ghayba	Historical Primary Sources	Shaykh Ṭūṣī
5	986	Kashf al-Asrār	Contemporary Primary So	Ayatullāh Rūhullah Khumaynī
6	392	The Law of the Dubai International Financial Centre: (	Articles	Alejandro Carballo
7	399	Choice of Law and Islamic Finance	Articles	Julio C. Colón
8	400	A Murābaḥah Transaction in an English Court: The Lo	Articles	Kilian Bälz
9	438	Corporate Insolvency in Malaysia	Secondary Sources	Sonali Abeyratne
10	441	Foundations of Forgiveness in Islamic Bankruptcy Law	Secondary Sources	Jason J. Kilborn
11	439	The Surprising Irrelevance of Islamic Bankruptcy	Secondary Sources	Haider Ala Hamoudi
12	1252	The True Story of Sharia in American Courts	Secondary Sources	Abed Awad

Structuring Data: Cardinal Rules

- Put all your variables (what you're measuring) in columns
- Put each observation in its own row
- Put each type of observation in its own table / file
- Don't combine multiple pieces of information in one cell

Structuring Data: Cardinal Rules

- Leave the “raw” data raw
- Don’t mix presentation with data
- Track your process and aim for a reproducible workflow
  - Git, OSF
  - Jupyter Notebooks
  - Notes!
- Export the cleaned data to an open-source, plain text format like CSV

How can we tell if data is messy?

Messy Data

	A	B	C
1	name	treatmenta	treatmentb
2	John Smith	NA	18
3	Jane Doe	4	1
4	Mary Johnson	6	7

	A	B	C	D
1	treatment	John.Smith	Jane.Doe	Mary.Johnson
2	a	NA	4	6
3	b	18	1	7

	A	B	C
1	name	treatment	n
2	Jane Doe	a	4
3	Jane Doe	b	1
4	John Smith	a	NA
5	John Smith	b	18
6	Mary Johnson	a	6
7	Mary Johnson	b	7

Messy Data

id	title	document_type	processed	publisher_and_location	published
983	Occultation of the Twelfth I	Secondary Sources	TRUE		1982
982	Mantle of the Prophet	Secondary Sources	TRUE		2002
185	Kitāb al-Ghayba	Historical Primary S	FALSE	Intisharāt-i Masjid-i Moqaddas-i Jamkarān | Tehran	
986	Kashf al-Asrār	Contemporary Prim	FALSE		1944
400	A Murābaḥah Transaction ir	Articles	FALSE		2004
438	Corporate Insolvency in Ma	Secondary Sources		International Insolvency Review	2000
441	Foundations of Forgiveness	Secondary Sources	FALSE	American Bankruptcy Law Journal	2011
896	Indonesian Supreme Court I	Court Cases	FALSE	Mahkamah Agung Republik Indonesia | Jakarta	3/30/15
1396	Ambon Religious Court Deci	Court Cases	FALSE	Mahkamah Agung Republik Indonesia | Ambon	6/30/11
1007	Kendari Religious Court Dec	Court Cases	FALSE	Mahkamah Agung Republik Indonesia | Kendari	8/25/15
				Not yet published on site	

Messy Data Mistakes

Most Common Problems: Messy Data

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

Messy Data

- Treating data like a notebook
  - Mixing context (presentation) with content (data)

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		

Messy Data

- Column headers are variables, not values

religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Messy Data: Stacked (Melted)

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Messy Data

- Multiple variables in one column
- Watch for split names (eg **gender-age**)

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2
AE	2000	m	15–24	4
AE	2000	m	25–34	4
AE	2000	m	35–44	6
AE	2000	m	45–54	5
AE	2000	m	55–64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Messy Data

- Different types of observations in one table

year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

Table 7: The first eight Billboard top hits for 2000. Other columns not shown are `wk4`, `wk5`, ..., `wk75`.

Messy Data

- Different types of observations in one table

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66

Table 8: First fifteen rows of the tidied Billboard dataset. The `date` column does not appear in the original table, but can be computed from `date.entered` and `week`.

Messy Data

- Different types of observations in one table

id	artist	track	time
1	2 Pac	Baby Don't Cry	4:22
2	2Ge+her	The Hardest Part Of ...	3:15
3	3 Doors Down	Kryptonite	3:53
4	3 Doors Down	Loser	4:24
5	504 Boyz	Wobble Wobble	3:35
6	98^0	Give Me Just One Nig...	3:24
7	A*Teens	Dancing Queen	3:44
8	Aaliyah	I Don't Wanna	4:15
9	Aaliyah	Try Again	4:03
10	Adams, Yolanda	Open My Heart	5:30
11	Adkins, Trace	More	3:05
12	Aguilera, Christina	Come On Over Baby	3:38
13	Aguilera, Christina	I Turn To You	4:00
14	Aguilera, Christina	What A Girl Wants	3:18
15	Alice Deejay	Better Off Alone	6:50

id	date	rank
1	2000-02-26	87
1	2000-03-04	82
1	2000-03-11	72
1	2000-03-18	77
1	2000-03-25	87
1	2000-04-01	94
1	2000-04-08	99
2	2000-09-02	91
2	2000-09-09	87
2	2000-09-16	92
3	2000-04-08	81
3	2000-04-15	70
3	2000-04-22	68
3	2000-04-29	67
3	2000-05-06	66

Table 13: Normalized Billboard dataset split up into song dataset (left) and rank dataset (right). First 15 rows of each dataset shown; `genre` omitted from song dataset, `week` omitted from rank dataset.

Messy Data

- Multiple tables in one spreadsheet

	A	B	C	D	E	F	G	H	I	J
10			Dwelling							
11		key_id	roof_type	wall type	floor type	rooms				
12		1	grass	muddaub	errth	1				
13		2	grass	muddaub	earth	1			includes barn	
14		3	mabati_sloping	burntbricks	cement	-99				
15		4	mabatisloping	burntbricks	earth	1				
16		5	grass	burntbricks	earth	1				
17		6	grass	muddaub	earth	1				
18		7	grass	muddaub	earth	1				
19		8	mabatisloping	burntbricks	cement	3				
20		9	grass	burntbricks	earth	1				
21		10	mabatisloping	burntbricks	cement	5				
22										
23									Plots	
24			Livestock					key_id	plots	water use
25		key_id	livestock_owned_and_numbers	poultry	look_after_cows			1	2	no
26		1	1, poultry	yes	no			2	3	yes (only in sur
27		2	3, (oxen , cows, goats)	yes	no			3	1	no
28		3	1, (none)	yes	no			4	3	no
29		4	2, (oxen , cows)	yes	no			5		N

Messy Data

- Bad null values

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

Messy Data: Field Names

Good Name	Good Alternative	Avoid
wall_type	WallType	wall type
longitude	GpsLongitude	gps:Longitude
gender	gender	M/F
Informant_01	first_informant	1st Inf

Messy Data: More Mistakes

- Multiple tabs
  - Data inconsistency
  - Extra steps
- Not filling in zeroes
- Putting units in cells
- Special characters

Dates as Data

Dates in Spreadsheets

- Stored in one column
- This can be problematic
- How else could we store dates?

Date Formats in Spreadsheets

A	B	C	D	E	F	G	H	I
Typed	**day-month**	**DOW-month-day-year**	**initial-year**	**month-year**	**value_automatic**	**plus**	**add**	**add_number**
2-jul	2-Jul	Thursday, July 2, 2020	7/2/2020	Jul-20	44014	90	9/30/2020	44,104.00
1-jan-1900	1-Jan	Monday, January 1, 1900	1/1/1900	Jan-00	2	90	4/1/1900	92.00

- There are many different ways to handle dates
- These will often break when you try to export data from Excel / Sheets
- Ambiguity in data entry is bad
- Spreadsheet programs store dates as ints (see last cell)

Preferred Date Formats

- MONTH DAY YEAR
  01 20 2021
- YEAR DAY_OF_YEAR
  2021 20
- Single string
  - YYYYMMDDhhmmss
  - January 20, 2021 09:21:30 = 20210120092130
- ISO-8601
  - UTC Offset for EST: -05
  - 2021-01-20T09:21:30 -0500
  - Or, in UTC: 2021-01-20T14:21:30Z

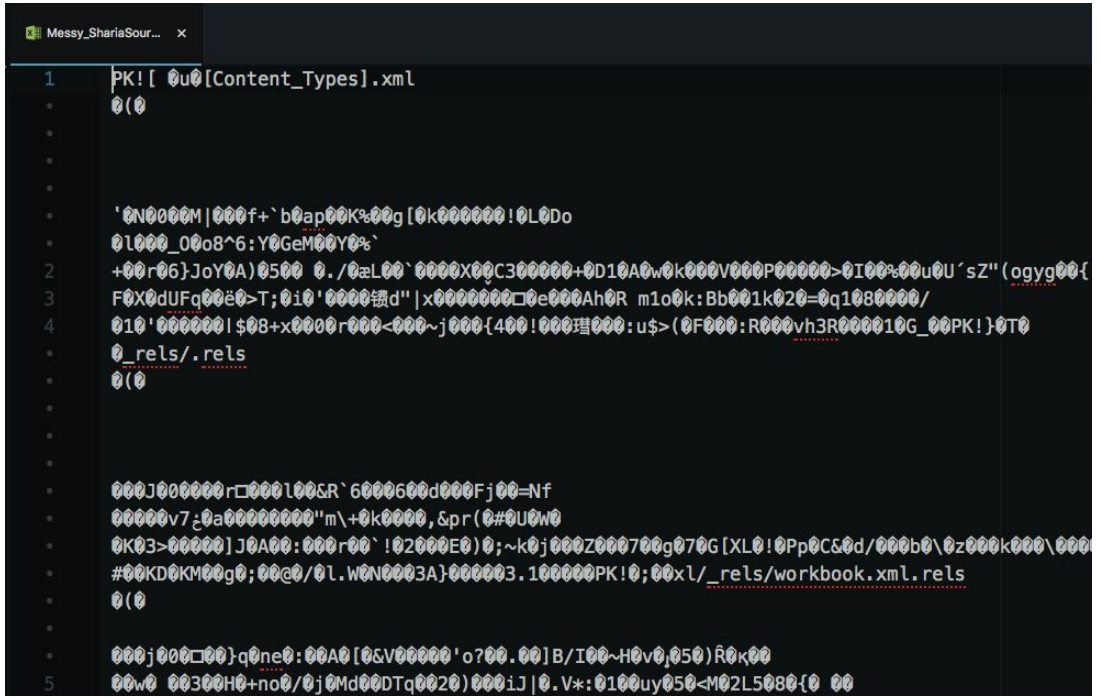
Historical Dates

- Spreadsheets struggle with historical (pre-1900) dates
- Spreadsheets struggle with non-Gregorian date systems
- Date formats vary by region
  - Is 11/12/20 November 12 or December 11?
  - Who collected the data? Where was the data collected? Ambiguous

File Format Options:
Keeping Your Data Free and Open

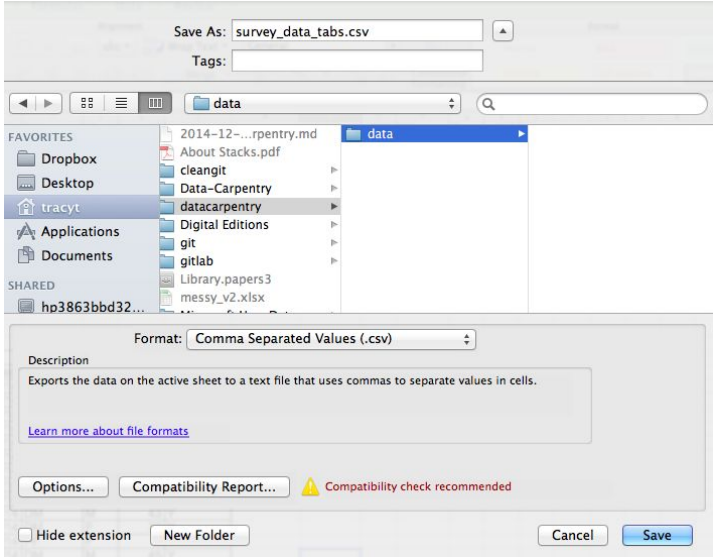
Getting and Keeping Data Clean

- Long term storage of data in proprietary formats (Excel) is a **bad idea**
  - Proprietary
  - Hard to open
  - Not future-compatible
  - Not always computable
  - Version problems



CSV to the Rescue

- Universal, open, static
  - Comma separated value (tab separated, etc)
  - Plaintext
  - Easy to open - even in a text editor
- How to save from Excel
  - File -> Save As
  - Format: CSV



CSV Compatibility

- Windows compatibility
  - Line ending problems
  - `\n vs \r\n`
  - “Windows comma separated file”
  - dos2Unix

```
data1,data2\r\n1,2\r\n4,5\r\n…
```

becomes

```
data1
data2\r
1
2\r
…
```

File Encodings

- Words and sentences are made up of individual **characters**: a, ā, غ
- There are numerous different **character encodings**
- Not all fonts have all encodings for all characters
- Default: Use Unicode's UTF-8!

Author: Guðrún Guðmundsdóttir. Title: Introduction to character encoding (文字符号化入門). Copyright © 2004-2007 W3C® (MIT, ERCIM, Keio).

Author: GuÃ°rÃºn GuÃ°mundsdÃ³ttir. Title: Introduction to character encoding (æ–‡å—ç¬¦å·åŒ–å…¥é–€). Copyright Â© 2004-2007 W3CÂ® (MIT, ERCIM, Keio).

Key Concepts

- Tidy, organized data is necessary for data analysis
  - Columns: single variable, rows: single observation, cells: single value
- Document your data organization and cleaning processes, either with code, a notebook, or notes
- Watch out for common messy data errors
- Keep a copy of raw data
- Keep presentation out of your data
- Beware of compound dates
- Use open file formats and UTF-8 for file encoding

Questions?

To the notebook!

Cleaning and Reshaping Data with Python + Pandas