Week 2: Data wrangling

ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du

Andrew.Du2@colostate.edu

Office Hours: Thursdays, 1:00–4:00pm in GSB 312

1

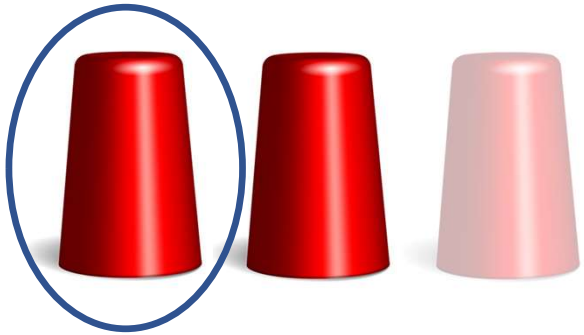
Statistics vignette

- Let's play a game…

2

Statistics vignette

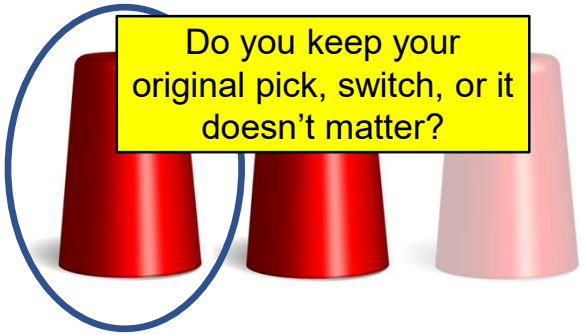
- Let's play a game…



3

Statistics vignette

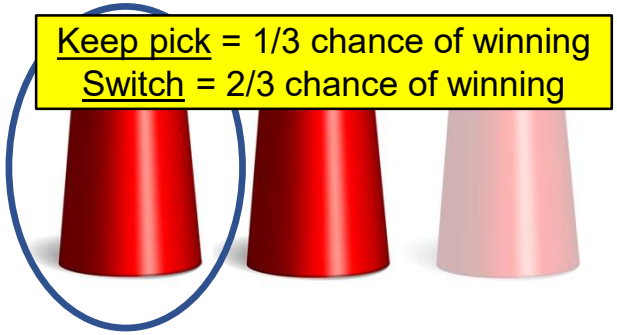
- Let's play a game…



4

Statistics vignette

- Let's play a game…



5

The Monty Hall problem

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?
—Craig F. Whitaker, Columbia, Md.

Ask Marilyn™
BY MARILYN VOS SAVANT

6

Lecture outline

1. Data wrangling
   - What is it, and why is it important?
2. General rules for data organization in spreadsheets
   - Emphasize the intimate connection with R
3. Data structures

7

Data wrangling

What is it, and why is it important?

8

What is data wrangling?

- Cleaning and organizing raw data to suit your research questions and analyses
- Important because data are rarely in a form ready to be analyzed *according to your goals*
- Incredibly marketable skill! Some data scientists only wrangle data

9

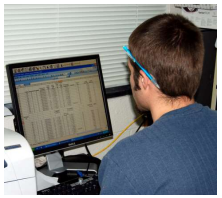
The data pipeline

Collect raw data

10

The data pipeline

Collect raw data → Enter data into computer spreadsheet



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	original.order	Transect	Occ	Relative age	Dental age	Species	Bone	Side	WS	Teeth	New Ind	Damage	Count	Articulated	Burial
2	1	T16-03	1	Adult	No Value	Zebra	Innominate	L	3		1	3,		No Value	No Value
3	2	T16-03	1	Adult	No Value	Zebra	Humerus	R	3		0	0		No Value	No Value
4	3	T16-03	1	Adult	No Value	Zebra	Tibia	L	3		0	0		No Value	No Value
5	4	T16-03	1	Adult	No Value	Zebra	Femur	R	3		0	1,		No Value	No Value
6	5	T16-03	2	Juvenile		2 Zebra	Cranium	L+R	3	dpM2-M1	1	4,		No Value	No Value
7	6	T16-03	2	Juvenile		2 Zebra	Innominate	R	3		0	1,		No Value	No Value
8	7	T16-03	2	Juvenile		2 Zebra	Cervical	NA	3		0			No Value	No Value
9	8	T16-03	2	Juvenile		2 Zebra	Femur	L	3		0			No Value	No Value
10	9	T16-03	2	Juvenile		2 Zebra	Axis	NA	3		0			No Value	No Value
11	10	T16-03	2	Juvenile		2 Zebra	Atlas	NA	3		0			No Value	No Value
12	11	T16-03	2	Juvenile		2 Zebra	Cervical	NA	3		0			No Value	No Value
13	12	T16-03	2	Juvenile		2 Zebra	Thoracic	NA	3		0			No Value	No Value
14	13	T16-03	2	Juvenile		2 Zebra	Thoracic	NA	3		0			No Value	>50%
15	14	T16-03	2	Juvenile		2 Zebra	Rib	Indt	3		0		3	No Value	No Value
16	15	T16-03	2	Juvenile		2 Zebra	Humerus	L	3		0			No Value	No Value

11

The data pipeline

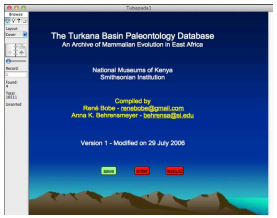
Collect raw data → Enter data into computer spreadsheet

- Also enter **metadata** – "data about data"
- For example:
  - Name of data collector
  - GPS coordinates
  - When data was collected
  - Funding support
  - Methods used to collect data
  - Units of measurement
  - Description of abbreviations
  - What the variables are

12

The data pipeline

Collect raw data → Enter data into computer spreadsheet



Published datasets/ databases

13

Where issues arise

Collect raw data → Enter data into computer spreadsheet

- Missing data
- Erroneous measurements (e.g., broken or miscalibrated instruments)

Published datasets/ databases

Subject number	Human height (inches)
1	64
2	70
3	
4	58
5	110
6	124
7	118

14

Where issues arise

Collect raw data → Enter data into computer spreadsheet

- Missing data
- Transcription errors
- Inconsistent data formats
- Combine data in one cell

Published datasets/ databases

Subject number	Human height (inches)	Sex	Date of birth
1	64	M	1970-02-24
22	70	m	1992-09-12
3		F	1977-07-11
4	58	F	1965-03-20
5	6	Fem	1982-05-25
6	73	M	06-10-1987
7; 8	68; 62	M; M	2000-12-13

15

Where issues arise

Collect raw data → Enter data into computer spreadsheet

- Missing data
- Transcription errors
- Inconsistent data formats
- Combine data in one cell

In 13 audits of real-world spreadsheets, an average of 88% contain errors (Panko 2008)

Published datasets/ databases

Subject number	Human height (inches)	Sex	Date of birth
1	64	M	1970-02-24
22	70	m	1992-09-12
3		F	1977-07-11
4	58	F	1965-03-20
5	6	Fem	1982-05-25
6	73	M	06-10-1987
7; 8	68; 62	M; M	2000-12-13

16

Where issues arise

Collect raw data → Enter data into computer spreadsheet

- Ideally 100% clean, but often not
- Missing data
- Inconsistent data formats
- Extra variables
- Extra data

Published datasets/ databases

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

- What if I'm interested only in petal lengths of individuals from *Iris virginica* species?

17

How to wrangle?

- Once research question & general analytical methodology is defined, it's clear how data should be collected and organized
- **ALWAYS** keep the raw "dirty" data file! Save cleaned data to new file
- **ALWAYS** back up your data (e.g., external hard drive, the cloud)!

18

Important not only for analyses

- Need to (should) publish raw data these days
  - Data collection supported by public funds *legally must* be published (granting agency technically owns the data)
  - Increases collegiality and rate of scientific progress (e.g., large-scale analyses)
  - Transparency & replicability of analyses
  - Clean data → increase your citation count!
- Be courteous & publish data in clean, analyzable format!

19

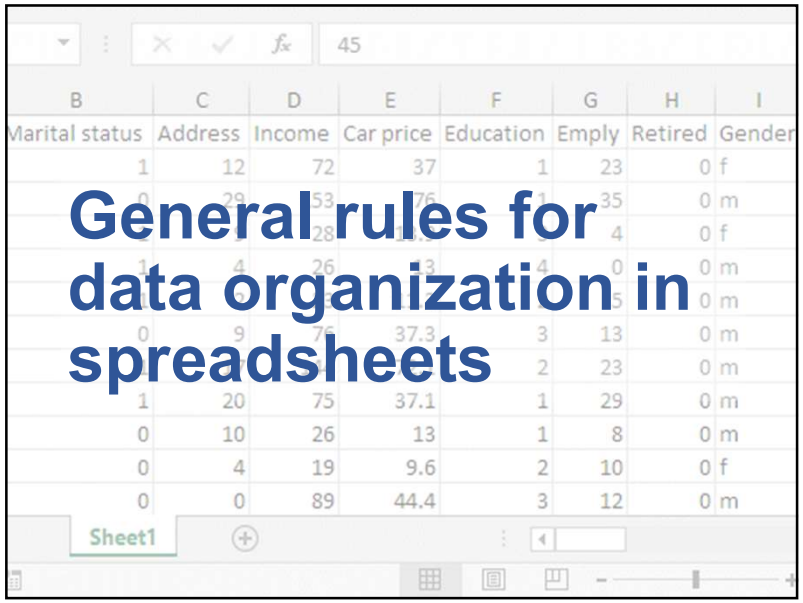
R makes wrangling (relatively) easy!

- Don't need to fix *every single* data entry by hand
- Leaves a record of what you did (your R script)

20

Questions?

21



22

What is a spreadsheet?

- An electronic page in which each row represents a single observation (i.e., unit of study), and each column represents a variable

Variables

	A	B	C	D
1	**ID Number**	**First**	**Last**	**Email**
2	5	Bob	Tester	Bob@gmail.com
3	3	Jane	Smith	Jane@gmail.com
4	8	Lazada	Inc	Lazada@gmail.com
5	103	Stuff	&nonsense	Stuff@gmail.com

Observations

Cell

23

What is a spreadsheet?

- An electronic page in which each row represents a single observation (i.e., unit of study), and each column represents a variable
- Used for entering, storing, analyzing (not anymore), and visualizing data (not anymore)
  - Analyzing & visualizing data in R ensures original dataset remains unchanged
- Most commonly used program is Microsoft Excel (it's what I use)
- For R, use comma-separated values (.csv) files (not proprietary & works in Excel)

24

CSV format

Plain text format

```
id,sex,glucose,insulin,triglyc
101,Male,134.1,0.60,273.4
102,Female,120.0,1.18,243.6
103,Male,124.8,1.23,297.6
104,Male,83.1,1.16,142.4
105,Male,105.2,0.73,215.7
```

- What it looks like in Excel

- Only **one** Excel sheet can be saved to **one** CSV file!

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

25

Rules for data organization

- In general, how R handles names, numbers, etc. is how data should be structured
- In fact, cleaning a dataset in R for analyses → dataset is publishable!
- Becoming proficient in R makes you better at organizing data!

26

Rules for data organization

1. Make sure your names are always ***consistent***
   1. R will treat "M", "`male`", and "`Male`" as completely different. Stick to one!
   2. Likewise, be careful of extra spaces! "`Male`" is treated differently than "`Male `"
   3. Be consistent with your formatting (e.g., don't use both `2020-08-29` and `08-29-2020`)

27

Rules for data organization

28

Rules for data organization

1. Make sure your names are always ***consistent***
2. Use NA to represent missing data
   1. Best to not leave spreadsheet cell blank (even though R will automatically replace it w/ NA)
   2. R is great at dealing with NAs, so don't use other symbols (e.g., ".", "-")
   3. Can add notes about missing data in another column

29

Rules for data organization

1. Make sure your names are always ***consistent***
2. Use NA to represent missing data
3. Avoid spaces. Use underscores ("snake case"), periods, or camel case instead
   - `human_height`, `human.height`, `humanHeight`

30

Rules for data organization

1. Make sure your names are always ***consistent***
2. Use NA to represent missing data
3. Avoid spaces. Use underscores ("snake case"), periods, or camel case instead
4. Avoid special characters (e.g., $, @, %, !, #, &, *)
5. Use short, informative variable names (e.g., `HumanHeight_in`)

31

Why are these bad names?

good name	good alternative	avoid
`Max_temp_C`	`MaxTemp`	`Maximum Temp (°C)`
`Precipitation_mm`	`Precipitation`	`precmm`
`Mean_year_growth`	`MeanYearGrowth`	`Mean growth/year`
`sex`	`sex`	`M/F`
`weight`	`weight`	`w.`
`cell_type`	`CellType`	`Cell type`
`Observation_01`	`first_observation`	`1st Obs.`

32

Rules for data organization

6. Put only one piece of data in each cell
   1. E.g., don't input:
      `1. 11,40` for lat/long
      `2. 75kg` for mass
      `3. ~10?` (uncertain measurement)
   2. When in doubt, put data in separate columns

33

Rules for data organization

6. Put only one piece of data in each cell
7. Do not use font color or highlighting
   1. R cannot interpret font colors/highlighting
   2. Won't be saved in a CSV file anyway

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102	2015–06–14	95.3
4	103	2015–06–18	97.5
5	104	2015–06–18	1.1
6	105	2015–06–18	108.0
7	106	2015–06–20	149.0
8	107	2015–06–20	169.4

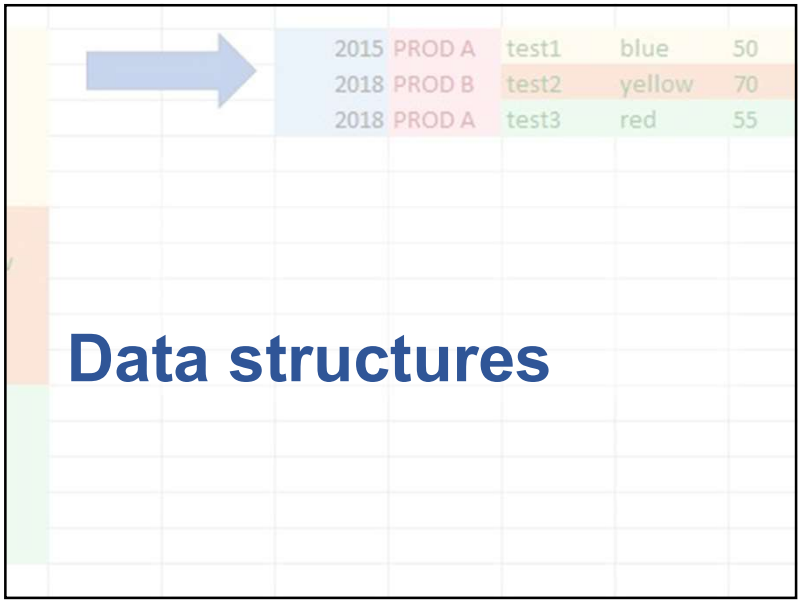
	A	B	C	D
1	id	date	glucose	outlier
2	101	2015–06–14	149.3	FALSE
3	102	2015–06–14	95.3	FALSE
4	103	2015–06–18	97.5	FALSE
5	104	2015–06–18	1.1	TRUE
6	105	2015–06–18	108.0	FALSE
7	106	2015–06–20	149.0	FALSE
8	107	2015–06–20	169.4	FALSE

Broman & Woo 2018

34

Questions?

35



Data structures

36

Data structure

- ***Always*** rectangular!

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

37

Data structure

- ***Always*** rectangular!
- What **NOT** to do:

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

- Will give R fits & difficult to work with!

38

What to do instead

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

Each row = unique combination of variables

First row **ONLY** for variable names

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354

39

Relating different spreadsheets

- To keep spreadsheets rectangular, may need to keep different rectangles in different files
- Relate rectangles to each other using consistent variable names
  - e.g., don't use `HumanHeight` in one and `Human_Height` in another

40

Relating different spreadsheets

	A	B	C	D	E	F	G	H	I	J
1	Species	Apak	Aralee_Issie	Aramis	Asa_Issie	Asa_Koma	Asbole	Bed_I	Bed_III	Bed_IV
2	Aepyceros_afarensis	0	1	0	0	0	0	0	0	0
3	Aepyceros_aff_dietrchi	0	0	0	0	0	0	0	0	0
4	Aepyceros_aff_premelampus	0	0	0	0	0	0	0	0	0
5	Aepyceros_cf_afarensis	0	0	0	0	0	0	0	0	0
6	Aepyceros_cf_melampus	0	0	0	0	0	0	0	0	0
7	Aepyceros_cf_premelampus	0	0	0	0	0	0	0	0	0
8	Aepyceros_datoadeni	0	0	0	0	0	0	0	0	0
9	Aepyceros_dietrchi	0	0	0	0	0	0	0	0	0
10	Aepyceros_melampus	0	0	0	0	0	0	1	0	0

	A	B	C	D	E	F	G	H	I	J
1	Country	Basin	Site	Lat	Long	Formation_Sequence	Min. Age	Max. Age	Mean Age	Duration
2	Kenya	Turkana	Apak	2.9	36.05	Nachukui Formation	4.2	5	4.6	0.8
3	Ethiopia	Awash	Aralee_Issie			Woranso-Mille	3.57	3.76	3.67	0.19
4	Ethiopia	Awash	Aramis	10.3	40.3	Sagantole Formation	4.4	4.4	4.4	0
5	Ethiopia	Awash	Asa_Issie			Sagantole Formation			4.12	0
6	Ethiopia	Awash	Asa_Koma	10.3	40.26	Adu-Asa Formation	5.54	5.77	5.66	0.23
7	Ethiopia	Awash	Asbole	11.05	40.67	Busidima Formation	0.6	0.8	0.7	0.2
8	Tanzania	-	Bed_I	-2.98	35.3	Olduvai Gorge	1.8	2.038	1.919	0.238
9	Tanzania	-	Bed_III	-2.98	35.3	Olduvai Gorge	0.8	1.2	1	0.4
10	Tanzania	-	Bed_IV	-2.98	35.3	Olduvai Gorge	0.6	0.8	0.7	0.2

41

Questions?

42

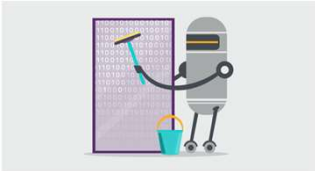
Summary

- The importance of data wrangling
- Rules for data in spreadsheets
  - Keep names consistent
  - Use NA to represent missing data
  - Avoid spaces & special characters
  - Use short, informative names
  - One piece of data per cell
  - Don't use highlighting or font coloring
- Keep spreadsheets rectangular!
  - Use different spreadsheets if necessary
- When in doubt, refer to Broman & Woo, 2018
- If you follow these rules when entering data, the less wrangling you will need to do later

43

But how to clean data?

- Thus far, you learned what good data practices are
- But what to do if you are dealing with bad, dirty data (***very*** common)?
- This week's R tutorial will teach you how to fix these data issues using R



44