

COGS 9 – Discussion Section A01 and A02

Kunal Rustagi (TA): Mon 9AM ([Zoom](#))
Connor McManigal (IA)
Yupei Sun(IA)

Tidying data

Why Tidy Data?

Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table.

What is dataset?

- A dataset is a collection of values, usually either numbers (if quantitative) or strings (if qualitative)
- Values are organized in two ways. Every value belongs to a variable and an observation.

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Problems with Messy Dataset

- 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.

Column headers are values, not variable names

Definition: Melting

- Turning columns into rows
- Parametrizing a list of columns that are already variables and convert the other columns into variables containing repeated column headings and the concatenated data values from the previous separate columns

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

Multiple variables stored in one column

After melting, we need to split the column column into columns each containing one kind of data.

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

(a) Molten data

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

(b) Tidy data

Variables are stored in both rows and columns

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

(a) Molten data

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

(b) Tidy data

Multiple types in one table

Datasets often involve values collected at multiple levels, on different types of observational units.

During tidying, each type of observational unit should be stored in its own table.

This is closely related to the idea of database normalization, where each fact is expressed in only one place. (could lead to potential inconsistencies within the df)

id	artist	track	time	id	date	rank
1	2 Pac	Baby Don't Cry	4:22	1	2000-02-26	87
2	2Ge+her	The Hardest Part Of ...	3:15	1	2000-03-04	82
3	3 Doors Down	Kryptonite	3:53	1	2000-03-11	72
4	3 Doors Down	Loser	4:24	1	2000-03-18	77
5	504 Boyz	Wobble Wobble	3:35	1	2000-03-25	87
6	98~0	Give Me Just One Nig...	3:24	1	2000-04-01	94
7	A*Teens	Dancing Queen	3:44	1	2000-04-08	99
8	Aaliyah	I Don't Wanna	4:15	2	2000-09-02	91
9	Aaliyah	Try Again	4:03	2	2000-09-09	87
10	Adams, Yolanda	Open My Heart	5:30	2	2000-09-16	92
11	Adkins, Trace	More	3:05	3	2000-04-08	81
12	Aguilera, Christina	Come On Over Baby	3:38	3	2000-04-15	70
13	Aguilera, Christina	I Turn To You	4:00	3	2000-04-22	68
14	Aguilera, Christina	What A Girl Wants	3:18	3	2000-04-29	67
15	Alice Deejay	Better Off Alone	6:50	3	2000-05-06	66

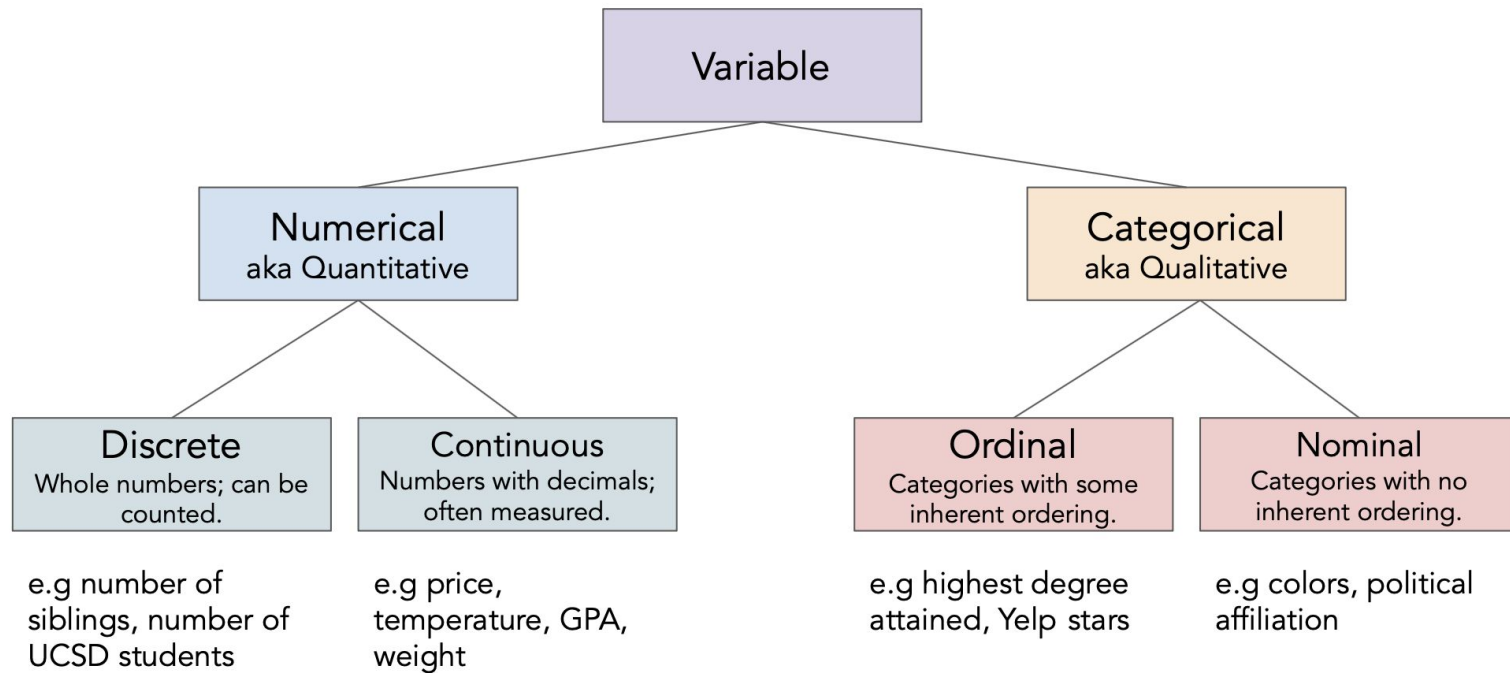
Tidy Tools

Tidying data makes it easier to maintain and do analysis with.

Manipulation functions:

- Filter: subsetting or removing observations based on some condition.
- Transform: adding or modifying variables. These modifications can involve either a single variable (e.g., log-transformation), or multiple variables (e.g., computing density from weight and volume).
- Aggregate: collapsing multiple values into a single value (e.g., by summing or taking means).
- Sort: changing the order of observations.

Types of Data



Note that numerical variables can be stored as strings, and categorical variables can be stored as numbers!

	Student ID	Student Name	Month	Day	Year	2021 tuition	2022 tuition	Percent Growth	Paid	DSC 80 Final Grade
0	A20104523	John Black	10	12	2020	\$40000.00	\$50000.00	25.00%	N	89
1	A20345992	Mark White	4	15	2019	\$9200.00	\$10120.00	10.00%	Y	90
2	A21942188	Amy Red	5	14	2021	\$50000.00	\$62500.00	25.00%	N	97
3	A28049910	Tom Green	7	10	2020	\$7000.00	\$9800.00	40.00%	Y	54
4	A27456704	Rose Pink	3	3	2021	\$10000.00	\$5000.00	-50.00%	Y	Pass

```

Student ID      object
Student Name    object
Month           int64
Day             int64
Year            int64
2021 tuition    object
2022 tuition    object
Percent Growth  object
Paid            object
DSC 80 Final Grade object
dtype: object

```

	Student ID	Student Name	Month	Day	Year	2021 tuition	2022 tuition	Percent Growth	Paid	DSC 80 Final Grade
0	A20104523	John Black	10	12	2020	\$40000.00	\$50000.00	25.00%	N	89
1	A20345992	Mark White	4	15	2019	\$9200.00	\$10120.00	10.00%	Y	90
2	A21942188	Amy Red	5	14	2021	\$50000.00	\$62500.00	25.00%	N	97
3	A28049910	Tom Green	7	10	2020	\$7000.00	\$9800.00	40.00%	Y	54
4	A27456704	Rose Pink	3	3	2021	\$10000.00	\$5000.00	-50.00%	Y	Pass

1. Clean student name: transform current format to last name, first name (White, Mark)
2. Clean Month, Day, Year: combine the three columns to Year-Month-Day (2019-04-15)
3. Clean tuition: strip '\$' and convert to float type
4. Clean paid: replace N, Y to False, True or 0, 1
5. Clean final grade: convert to int or float type

Missing Values

	First Name	Gender	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	97308.0	6.945	True	Marketing
1	Thomas	Male	61933.0	NaN	True	NaN
2	Jerry	Male	NaN	9.340	True	Finance
3	Dennis	NaN	115163.0	10.125	False	Legal
4	NaN	Female	NaN	11.598	NaN	Finance
5	Angela	NaN	NaN	18.523	True	Engineering
6	Shawn	Male	111737.0	6.414	False	NaN
7	Rachel	Female	142032.0	12.599	False	Business Development
8	Linda	Female	57427.0	9.557	True	Client Services
9	Stephanie	Female	36844.0	5.574	True	Business Development
10	NaN	NaN	NaN	NaN	NaN	NaN

1. Drop missing values
2. Impute missing values
 - a. Mean
 - b. Median
 - c. Mode
 - d. Use machine learning algorithm to predict missing values
3. Interpolation: estimating missing values based on the values of other variables in the dataset
 - a. Linear
 - b. Polynomial...

Reading 2: Data Organization in Spreadsheets

Karl W. Broman and Kara H. Woo

Be consistent

1. Consistent codes for categorical variables
2. Consistent fixed code for any missing values
3. Consistent variable names
4. and so on...



Choose good names for things

Table 1: Examples of good and bad variable 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.

Write dates as YYYY-MM-DD and No empty cells

1. ISO 8601 standard
2. YYYY-MM-DD format

	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4

Put just one thing in a cell

	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
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

Figure 3: A tidy version of the data in Figure 2B.

Create a data dictionary

	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high-fat diet

No calculations in the raw data files



Don't use font color or highlighting as data

A

	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

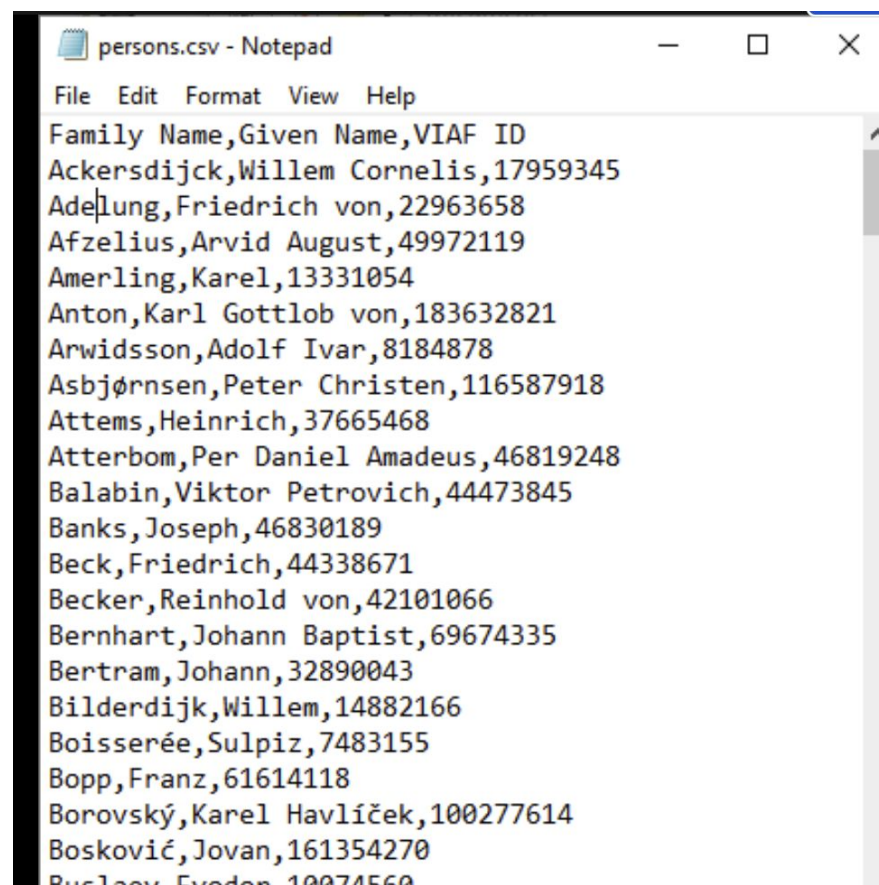
B

	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

Make backups



Save the data in plain text files



A screenshot of a Notepad window titled 'persons.csv - Notepad'. The window displays a list of names and their corresponding VIAF IDs, separated by commas. The text is as follows:

```
File Edit Format View Help
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
Arwidsson,Adolf Ivar,8184878
Asbjørnsen,Peter Christen,116587918
Attems,Heinrich,37665468
Atterbom,Per Daniel Amadeus,46819248
Balabin,Viktor Petrovich,44473845
Banks,Joseph,46830189
Beck,Friedrich,44338671
Becker,Reinhold von,42101066
Bernhart,Johann Baptist,69674335
Bertram,Johann,32890043
Bilderdijk,Willem,14882166
Boisserée,Sulpiz,7483155
Bopp,Franz,61614118
Borovský,Karel Havlíček,100277614
Bosković,Jovan,161354270
Busby,Evelyn,10074560
```