COGS 9 - Introduction to Data Science
Winter 24

Week 5 Discussion

Slides can be found here: https://github.com/PrasannaKumaran/COGS-9---WI24

Agenda for Today

- Quick overview of the Readings for this week
  - Tidy Data
  - Data organization in spreadsheet
- Questions

Reading: Tidy Data

- Focuses on important aspects of data cleaning
- Structuring datasets to facilitate analysis
- Organize data values within a dataset

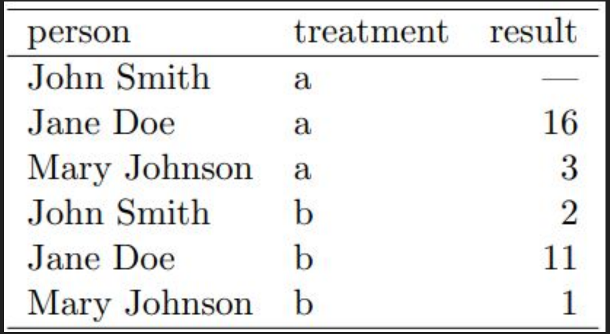
Reading: Tidy Data

- What is Tidy Data?
  - Provide a standardized way to link the structure of a dataset with its semantics
  - Standard vocabulary for describing the structure and semantics
- Data Structure?
  - Layout of rows and columns
  - Describe underlying semantics apart from appearance

Reading: Tidy Data

- Data Semantics
  - Dataset is a collection of values
  - Values belongs to a variable and an observation
  - Below shows variables as columns and observations as rows
  - Variable names are crucial
    - Height and weight is not same as

      Height and width



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

Reading: Tidy Data

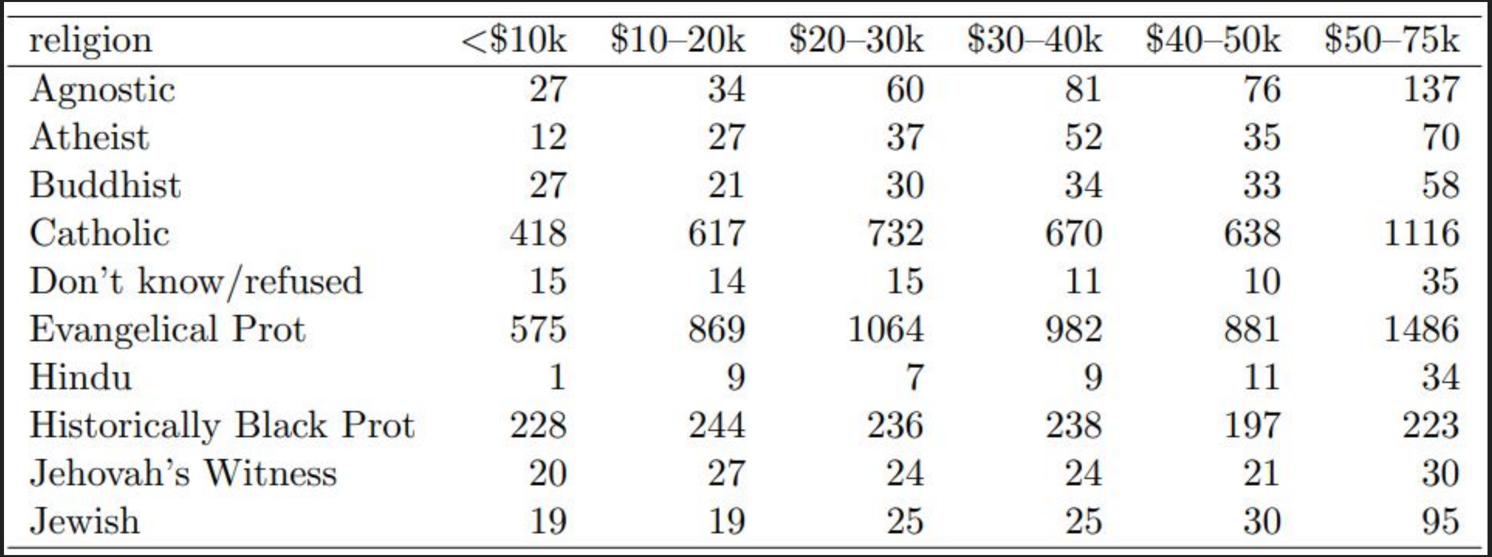
- Tidy data is a standard way of mapping the meaning of a dataset
  - Each variable forms a column
  - Each observation forms a row
  - Each type of observational unit forms a table
- Makes it easy for an analyst or computer to extract required information

Reading: Tidy Data

- Tidying messy datasets
  - 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

Reading: Tidy Data

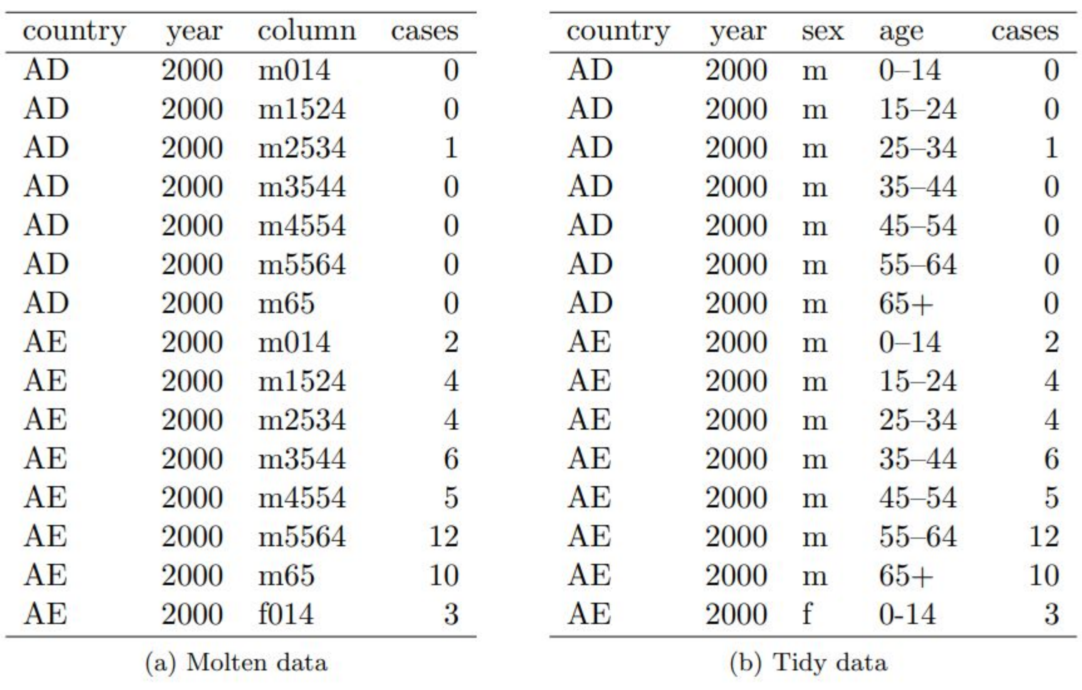
- Column headers are values not variable names



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

Reading: Tidy Data

- Multiple variables stored in one column



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

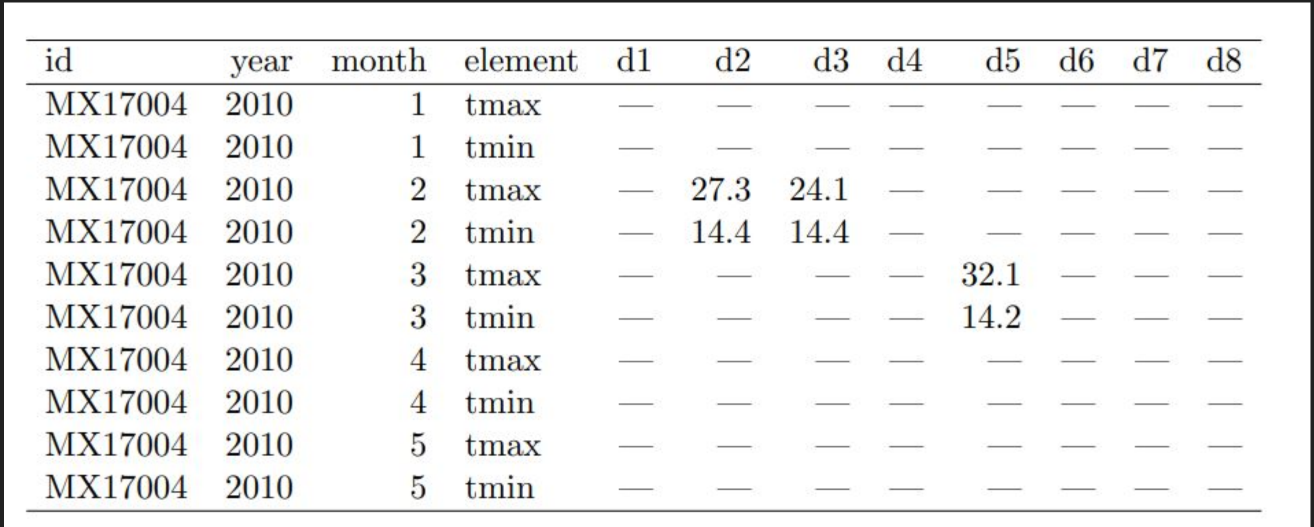
(a) Molten data

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

Reading: Tidy Data

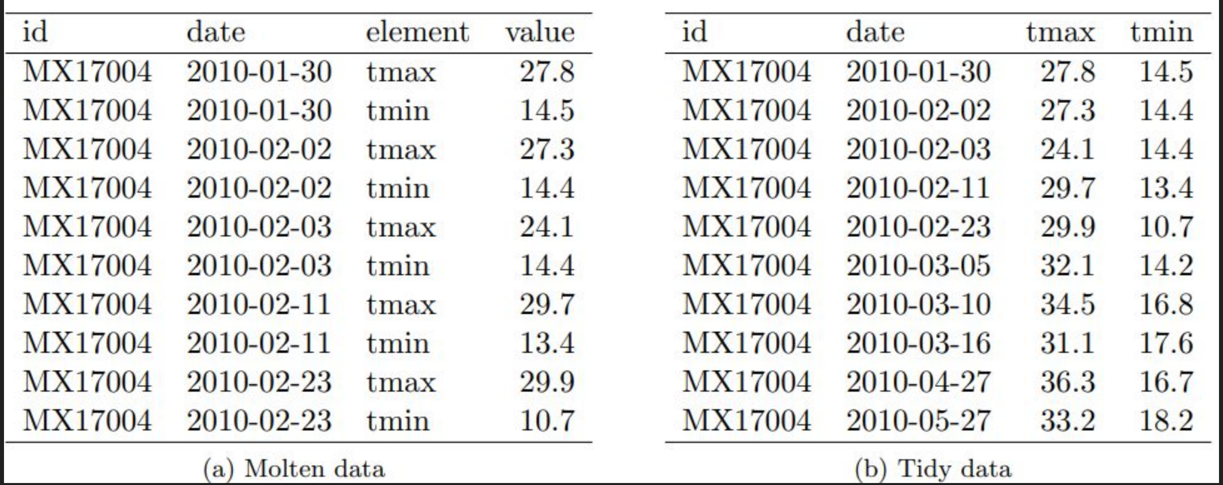
- Variables are stored in both rows and columns



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	—	—	—	—	—	—	—	—

Reading: Tidy Data

- Variables are stored in both rows and columns



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

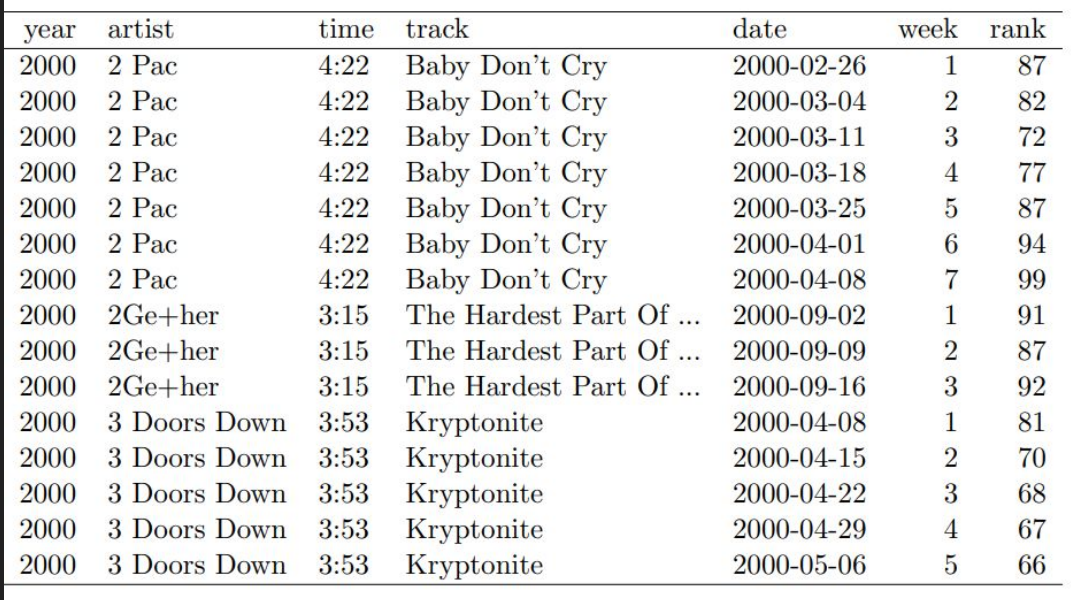
(a) Molten data

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

Reading: Tidy Data

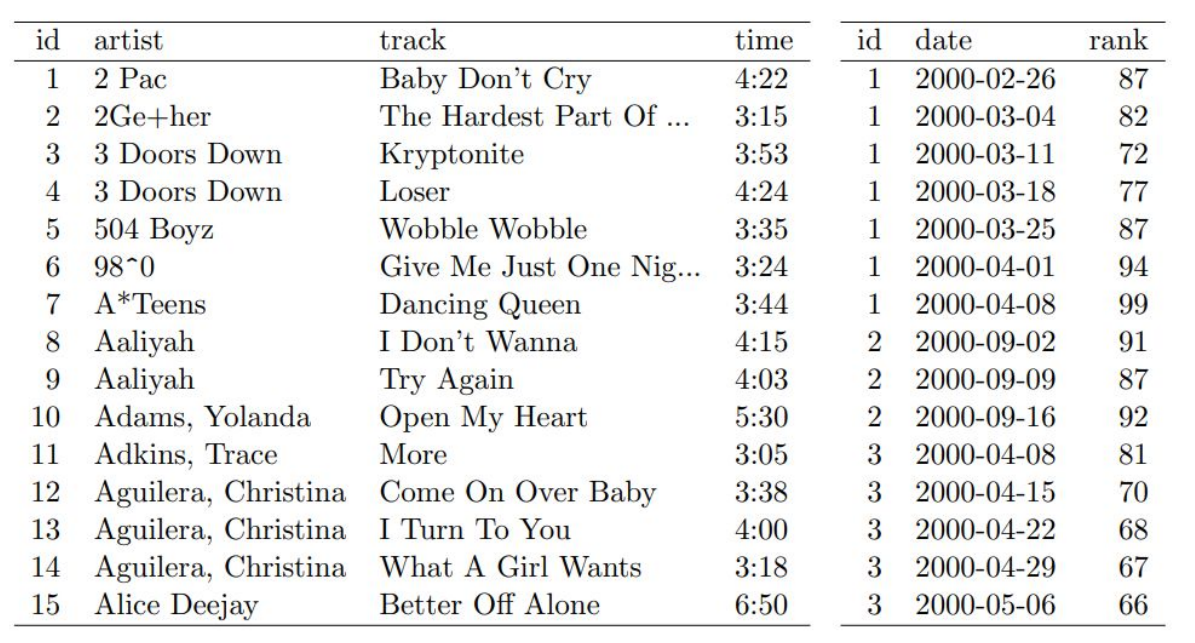
- Multiple types in one table



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

Reading: Tidy Data

- Multiple types 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	98ˆ0	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

Reading: Tidy Data

- One type in multiple tables
  - Single type of observational unit spread over multiple tables or files
- Manipulation
  - Filter: subsetting or removing observations
  - Transform: adding or modifying variables
  - Aggregate: collapsing multiple values into a single value
  - Sort: changing the order of observations
- Visualization
- Modelling

Reading: Data Organization in Spreadsheet

- What is Spreadsheet?
  - Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization
- Focus of the Paper
  - Organizing spreadsheet data to reduce errors and ease later analyses
  - Both humans and computer programs can read

Reading: Data Organization in Spreadsheet

- Consistency
  - Use consistent codes for categorical variables.
  - Single column variables
    - Eg. male/female vs. M/F
  - Use a consistent fixed code for any missing values.
  - Every cell needs to be filled
    - Use “-” or “NA”, don’t use “-999”

Reading: Data Organization in Spreadsheet

- Consistency continued …
    - Use consistent variable names
        - Eg. “Glucose_10wk” vs “gluc_10weeks”
        - Capitalization matters
    - Use consistent subject identifiers
        - “Mouse153” vs “153”
    - Use a consistent data layout in multiple files
        - a consistent structure in multiple files

Reading: Data Organization in Spreadsheet

- Consistency continued …
    - Use consistent file names
        - Eg. “Serum batch1 2015-01-30.csv” vs “batch2 serum 52915.csv”
    - Use a consistent format for all dates
        - YYYY-MM-DD
    - Use consistent phrases in your notes
        - “dead” vs “Dead”
    - Be careful about extra spaces within cells
        - “ male ” vs “male”
        - Blank cell vs a cell with space

Reading: Data Organization in Spreadsheet

- Good names

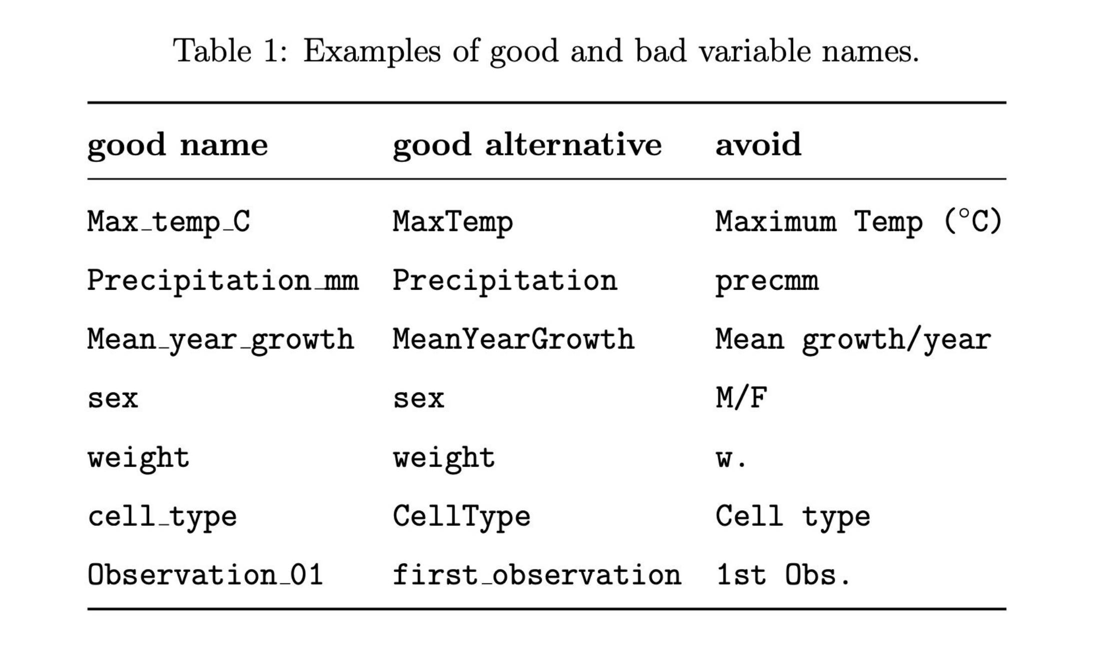


Table 1: Examples of good and bad variable names.

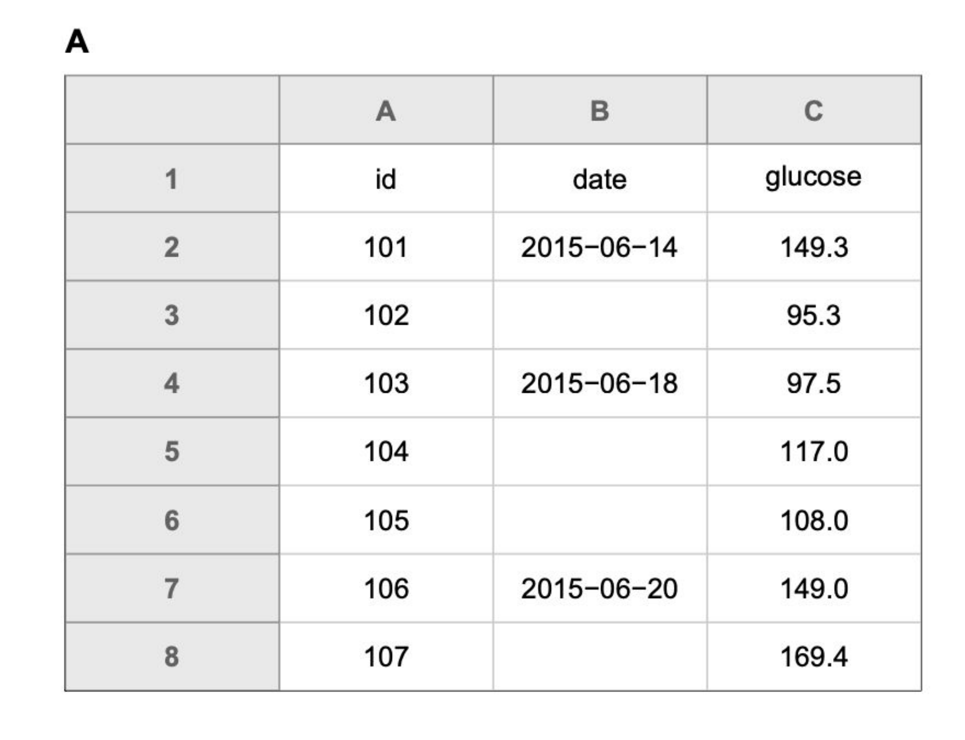
good name	good alternative	avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

Reading: Data Organization in Spreadsheet

- Dates
    - “ISO 8601”
    - YYYY-MM-DD
    - prefer to use a plain text format for columns in an Excel worksheet that are going to contain dates
    - to begin the date with an apostrophe
        - '2014-06-14

Reading: Data Organization in Spreadsheet

- No empty cells
    - “NA” or a hyphen
    - Don’t let people assume repetition

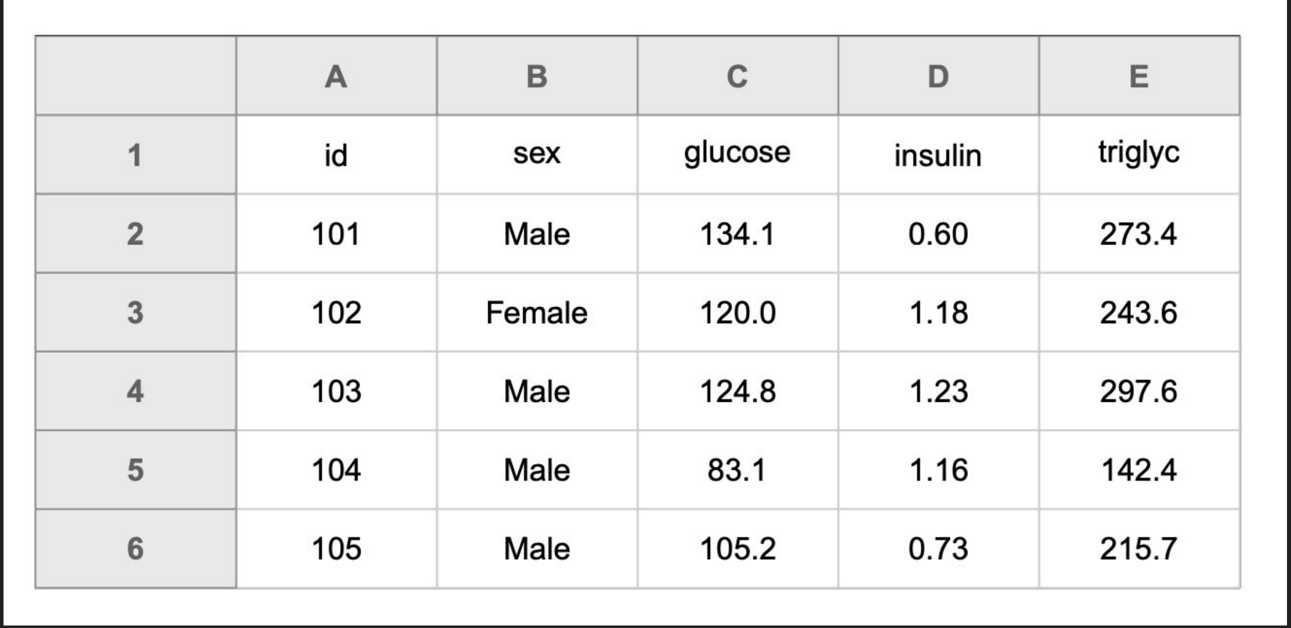


A

	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

Reading: Data Organization in Spreadsheet

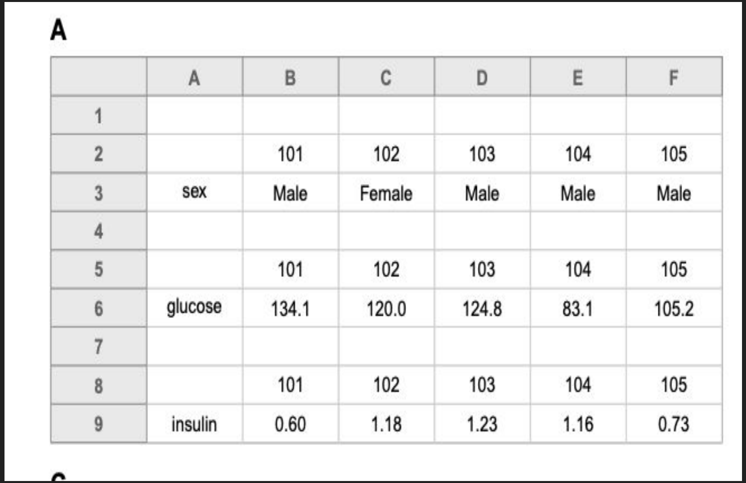
- One thing in a cell



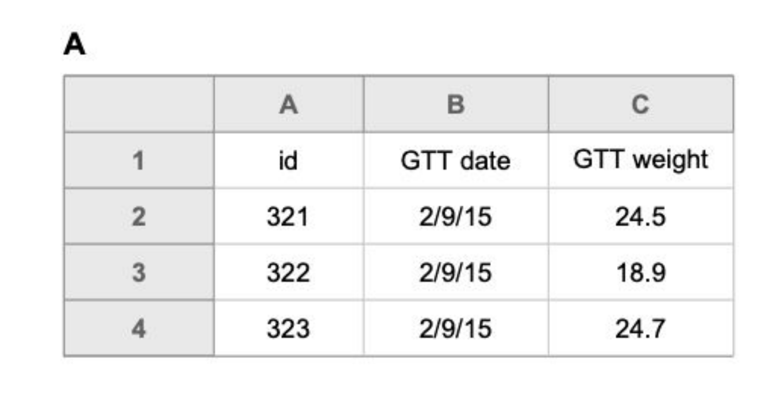
	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Reading: Data Organization in Spreadsheet

- Make it a rectangle



	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Male
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73



	A	B	C
1	id	GTT date	GTT weight
2	321	2/9/15	24.5
3	322	2/9/15	18.9
4	323	2/9/15	24.7

Reading: Data Organization in Spreadsheet

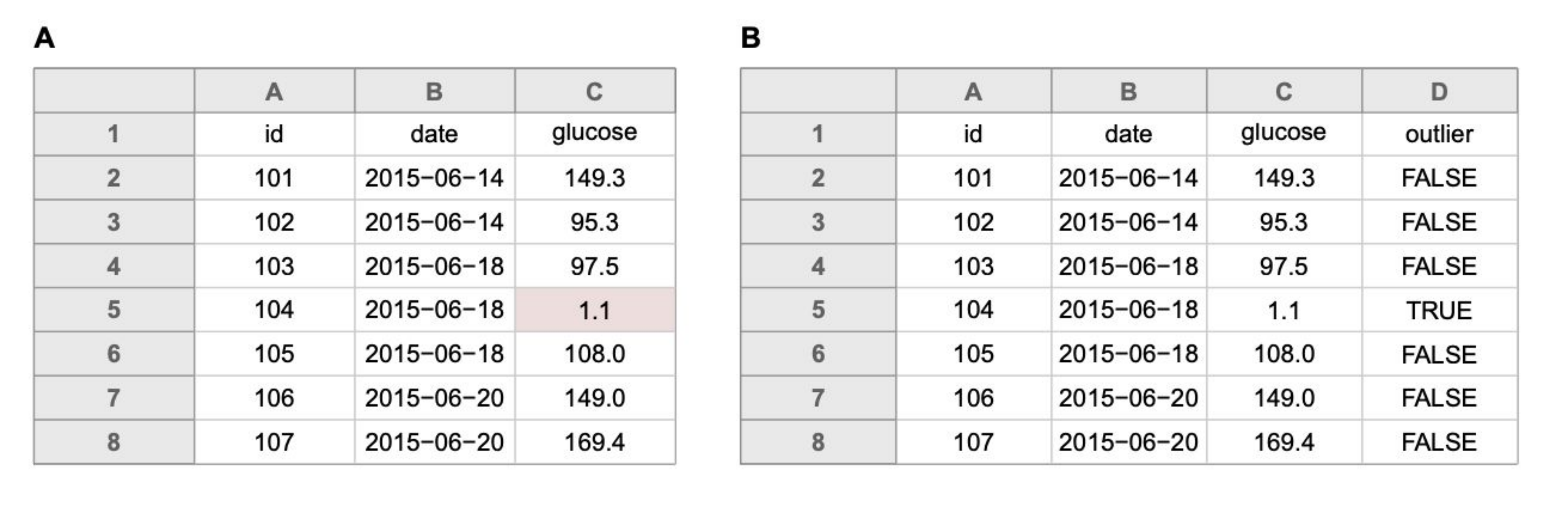
- Data dictionary
    - The exact variable name as in the data file
    - A version of the variable name that might be used in data visualizations
    - A longer explanation of what the variable means
    - The measurement units
    - Expected minimum and maximum values

Reading: Data Organization in Spreadsheet

- No calculations in the raw data files
    - Primary data file should contain just the data and nothing else: no calculations, no graphs
    - If you want to do some analyses in Excel, make a copy of the file and do your calculations and graphs in the copy

Reading: Data Organization in Spreadsheet

- 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

Reading: Data Organization in Spreadsheet

- Make backups
- Use data validation to avoid errors
  - In excel
    - Select a column
    - In the menu bar, choose Data → Validation
    - Choose appropriate validation criteria
      - A whole number in some range
      - A decimal number in some range
      - A list of possible values
      - Text, but with a limit on length
- Save the data in plain text files

Questions?

Make sure you filled out the attendance form

Attendance