# Section Four

**Class** COGS 9

**Teaching Assistant:** Matthew Feigelis

# Background

- Spreadsheets and data frames are widely used for data entry, storage, and analysis
- Good spreadsheets are less error-prone, easier for computers to process, and easier to share with collaborators
- Standardized data storage allow for open source **analysis** tools that work on your data straight out of the box. **Huge time saver.**

# Tidy Data

- **Tidy**: a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)
- This makes it easier to develop **tidy tools** for **data analysis,** regardless of technical domain, the application of the tools (e.g. plotting x vs y, or doing a regression) should be easy

# What is a dataset?

- A dataset is a collection of **values** that can include numbers or words
- Each value in a dataset belongs to an **observation, variable pair**.
- What are **variables**?
- What are **observations**?

| 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 |

# What about this dataset?

- How many variables are in the dataset?

| 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 |

Three! This is in **wide** form, it's good for visualizing sometimes, but it's not easy to work with. We call it **messy**, instead of **tidy**.

# Tidy Data

- If we make it Tidy. It looks like this. Now it's clear there is 3 variables. It's the **same information** but in a different form, and easier to do analysis on.

| 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 |

# Tidy Data

- **Tidy rules**:
  - Each variable forms a column
  - Each observation forms a row
  - Each type of observational unit forms a table

# What about this dataset?

- This example violates #1: the column headers are values, rather than variables.
- We need to explicitly create the variable column (income)
- Then *melt* those values in the header into that column, thus turning the wide dataset, long (here, *molten*)

| 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 |

| 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 |

# Another common example of messy data: Multiple variables are stored in one column

- Dataset: World Health Organization records on the counts of tuberculosis cases by country, year, and demographic

- Code Example of Melting and Tidying

| 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

Table 10: Tidying the TB dataset requires firs into two variables: `sex` and `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 |

(a) Molten data

Table 10: Tidying the TB dataset requires firs into two variables: `sex` and `age`.

# Discussion Question

Discuss amongst groups of 2-4 people the following. Take 5 minutes. Then we'll reconvene and make a list together.

1. What are 3 things we can do to clean this data up?

| 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 | — | — | — | — | — | — | — | — |

Table 11: Original weather dataset. There is a column for each possible day in the month. Columns d9 to d31 have been omitted to conserve space.

| 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

# Attendance

- Enter your number and today’s word into the attendance form

- Todays word:
- Form: https://forms.gle/tx9GcpANHEMwj8Jv7

- https://tinyurl.com/cog9-spring-23