Data Cleaning

UCSB SUMMER STATS WORKSHOP 2023

WEEK 3

Tidy Data

Hadley Wickham
RStudio

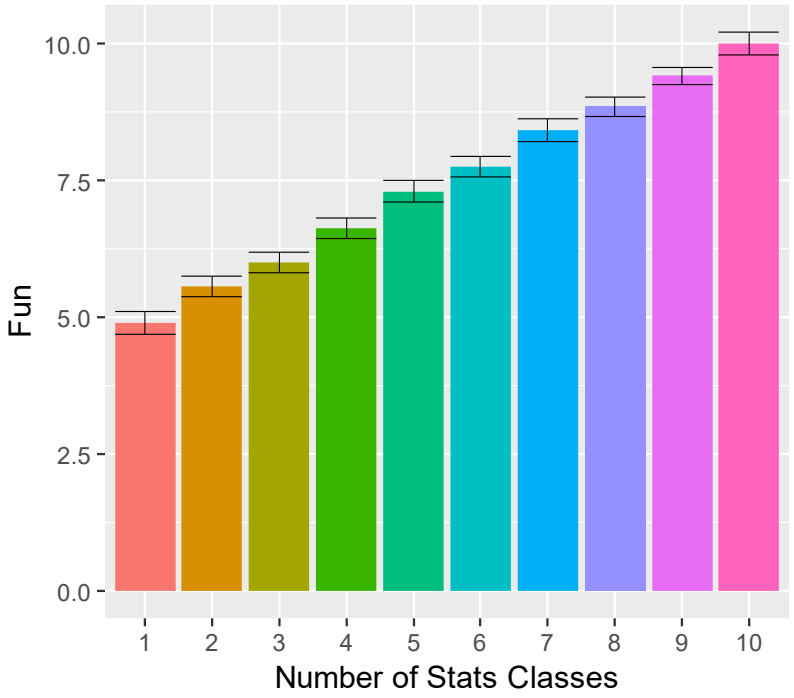
Martin Monkman

The Data Preparation Journey

Finding Your Way With R

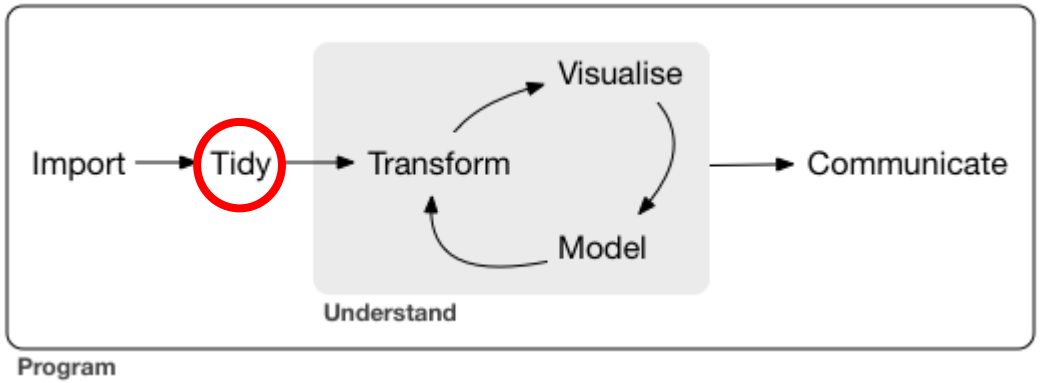
Trimming

	var1	var2	var3	var4	var5	var6
1	5.082520	5.907829	7.073734	7.884836	6.324818	8.437262
2	5.443151	7.327779	6.050779	5.991775	5.954487	8.307324
3	4.024609	4.881766	2.734992	7.532117	8.664693	5.746925
4	4.480013	5.121075	5.930708	8.619031	7.977853	7.893341
5	3.973761	6.783843	6.567947	7.843699	6.626595	8.080891
6	3.582721	5.856451	4.185818	4.937832	8.730098	7.508317
7	5.335790	3.517303	5.621268	7.377966	8.066959	7.014396
8	5.191268	4.818285	5.429966	7.181727	6.207386	7.291512
9	6.403909	5.418646	5.696535	6.638591	7.451374	8.104118
10	5.131530	6.384721	5.296592	6.585722	8.171418	7.812140
11	5.765037	6.683692	4.739766	7.428553	6.292216	6.575642
12	5.361895	4.621512	7.468145	7.290700	9.264005	8.643813
13	4.421865	5.503943	7.108561	7.322792	6.871867	7.731232
14	5.194853	6.172803	5.393268	6.906846	9.319122	8.511639
15	4.935410	4.590636	6.746634	5.931916	5.377686	7.150316



Raw Data

Analysis



Wickham and Grolemund, 2016

Tidy data:

- 1. Every variable has its own column
- 2. Every observation has its own row
- 3. Every type of data has its own dataframe

Every cell should tell you exactly one piece of information about exactly one observation of the world

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

Wickham, 2014

Data Cleaning Guidelines

Monkman (2023):

- Complete: ~represent all of your data
- Consistent: be systematic
  - Name your variables according to a consistent system
  - Keep your variables in consistent units
  - Structure your data systematically
- Accurate: free of errors
  - Data entry errors
  - Data processing errors

Data Cleaning Guidelines

Use an automated process

- **Process nothing by hand**
- Hand processing leads to errors
- Does not produce receipts

Should yield:

- Raw data (**completely** unedited)
- Processed data
- Data processing script
- Ideally: codebook

Hongbo Yu

Moral emotions (e.g., guilt, gratitude), affective neuroscience, social psychology

YES lab

Week 3 Example Data

N = 720 Participants (360 female)

Each participant paired with a sham partner

Learn one piece of information about partner. Either:

- Morally good: e.g., they volunteer their time at a soup kitchen
- Morally bad: e.g., they occasionally steal money from an elderly family member

Puzzle task:

- Must solve a puzzle in under 1 minute or partner receives electric shock
- Either participant or partner attempts puzzle
- Shock is either low, medium, or high voltage
- Puzzle is unsolvable (partner always gets shocked)

Week 3 Example Data

Measures:

Guilt:

- "How much guilt do you feel toward your partner?"
- 7-point Likert Scale
- Some missing data due to computer failure

Generosity:

- Participant asked how they would divide $10 between themselves and their partner
- Two time points: (1) start of experiment and (2) after shock

Personality:

- Ten-Item Personality Inventory (TIPI)
- 10 7-point Likert scales
- Measures Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
  - 2 items each
- Half of items reverse-coded

Ten-Item Personality Inventory-(TIPI)

Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
1	2	3	4	5	6	7

I see myself as:

1. _____ Extraverted, enthusiastic.
2. _____ Critical, quarrelsome.
3. _____ Dependable, self-disciplined.
4. _____ Anxious, easily upset.
5. _____ Open to new experiences, complex.
6. _____ Reserved, quiet.
7. _____ Sympathetic, warm.
8. _____ Disorganized, careless.
9. _____ Calm, emotionally stable.
10. _____ Conventional, uncreative.

TIPI scale scoring ("R" denotes reverse-scored items):

Extraversion: 1, 6R; Agreeableness: 2R, 7; Conscientiousness; 3, 8R; Emotional Stability: 4R, 9;

Openness to Experiences: 5, 10R.

Predictions

People will feel more guilty and be more generous when:

- They caused the shock (vs. partner)
- The partner is morally good (vs. morally bad)
- The shock is stronger (vs. weaker)

Women will report more guilt and be more generous than men

Agreeable people and neurotic people will be more guilty and more generous