# Storytelling with Data

## PSY 410: Data Science for Psychology

Dr. Sara Weston

2026-05-20

### Setup

## Why storytelling?

### Data alone isn't enough

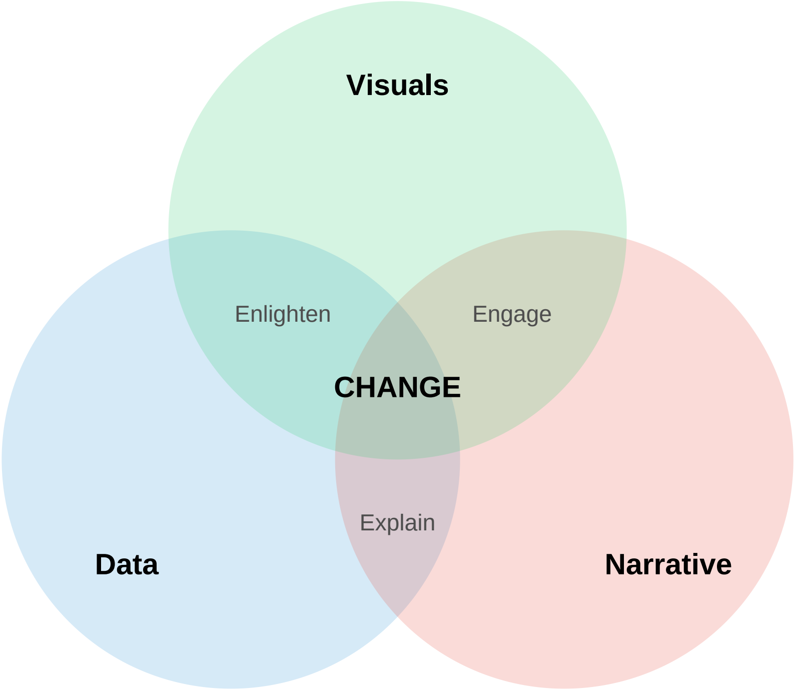
You've learned to:

- Import and clean data
- Transform and summarize
- Create visualizations
- Handle missing data

. . .

But **technical skills** **communication skills**

## The data storytelling triad



Adapted from Brent Dykes

## Why stories work

**Stories are memorable:**

- 63% of people remember stories
- Only 5% remember statistics

. . .

**Stories are persuasive:**

- Charity brochure study: Story about one child raised 2x more donations than statistics about millions

. . .

**Decisions are emotional:**

- We think we're rational
- But emotions drive decision-making
- Stories engage emotions

### Your goal as a data scientist

**Don't just show data — tell a story that:**

1. Answers a specific question
2. Provides context
3. Highlights what matters
4. Leads to action or understanding

## The storytelling framework

### Step 1: Understand the context

Before making any visualization, ask:

- **Who is the audience?**
  - Researchers? General public? Clinicians? Grant reviewers?
- **What do they care about?**
  - Effect sizes? Practical implications? Cost savings?
- **What action do you want them to take?**
  - Fund your research? Change clinical practice? Read your paper?

### Example: Different audiences, different stories

**Finding:** CBT reduces depression by 8 points on the BDI-II (d = 0.65)

. . .

**For researchers:**

- Effect size, confidence intervals, p-values
- Comparison to other interventions
- Limitations and future directions

. . .

**For clinicians:**

- Practical significance: “Patients move from moderate to mild depression”
- How to implement, training required
- Success rates, dropout rates

### Step 2: Choose appropriate visuals

Match your plot type to your message:

| Goal | Good choice | Bad choice |
|---|---|---|
| Show change over time | Line plot | Pie chart |
| Compare groups | Bar chart, boxplot | 3D pie chart |
| Show distribution | Histogram, density | Table |
| Show relationship | Scatterplot | Multiple pie charts |
| Show parts of whole | Stacked bar, treemap | 3D bar chart |

### Step 3: Eliminate clutter

**Clutter** is anything that doesn’t help your audience understand the message.

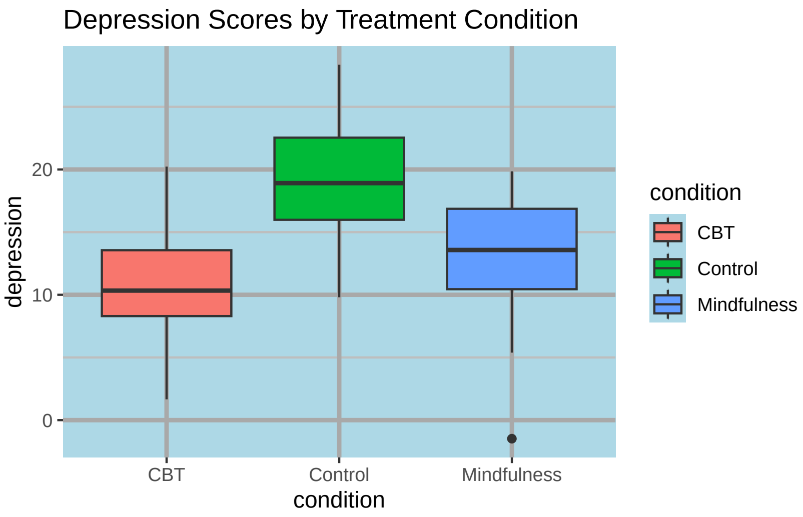
. . .

Common clutter:

- Unnecessary gridlines
- Heavy borders and backgrounds
- Too many colors
- Redundant labels
- Chart junk (3D effects, shadows, unnecessary decorations)

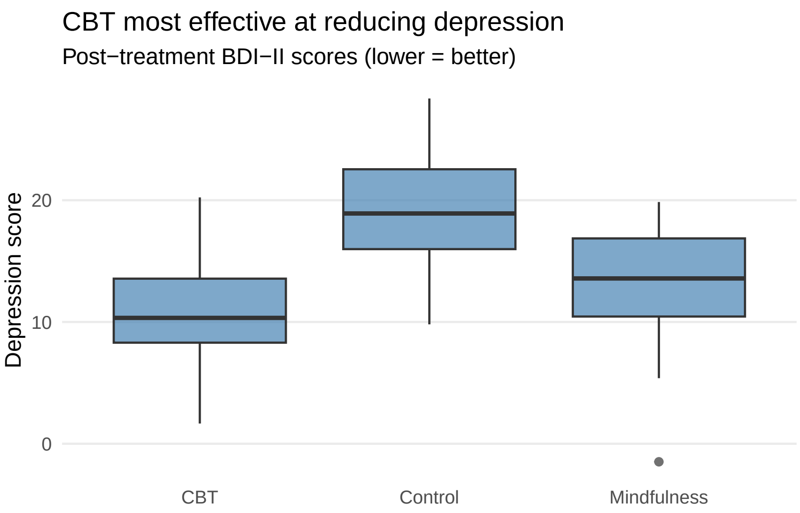
### Example: Cluttered figure

```
therapy_data <- tibble(
  condition = rep(c("Control", "CBT", "Mindfulness"), each = 30),
  depression = c(rnorm(30, 18, 5), rnorm(30, 12, 5), rnorm(30, 14, 5))
)

ggplot(therapy_data, aes(x = condition, y = depression, fill = condition)) +
  geom_boxplot() +
  labs(title = "Depression Scores by Treatment Condition") +
  theme_gray() +
  theme(
    panel.background = element_rect(fill = "lightblue"),
    panel.grid.major = element_line(color = "darkgray", size = 1),
    panel.grid.minor = element_line(color = "gray", size = 0.5)
  )
```



### Example: Decluttered figure

```
ggplot(therapy_data, aes(x = condition, y = depression)) +
  geom_boxplot(fill = "steelblue", alpha = 0.7) +
  labs(
    title = "CBT most effective at reducing depression",
    subtitle = "Post-treatment BDI-II scores (lower = better)",
    x = NULL,
    y = "Depression score"
  ) +
  theme_minimal() +
  theme(
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank()
  )
```



## Gestalt principles of design

Your brain groups things automatically based on:

1. **Proximity** — things close together are related
2. **Similarity** — things that look similar are related
3. **Enclosure** — things inside a boundary are related
4. **Connection** — things connected by lines are related

. . .

Use these principles intentionally!

### Step 4: Focus attention

**Preattentive attributes** are processed by the brain in < 500ms:

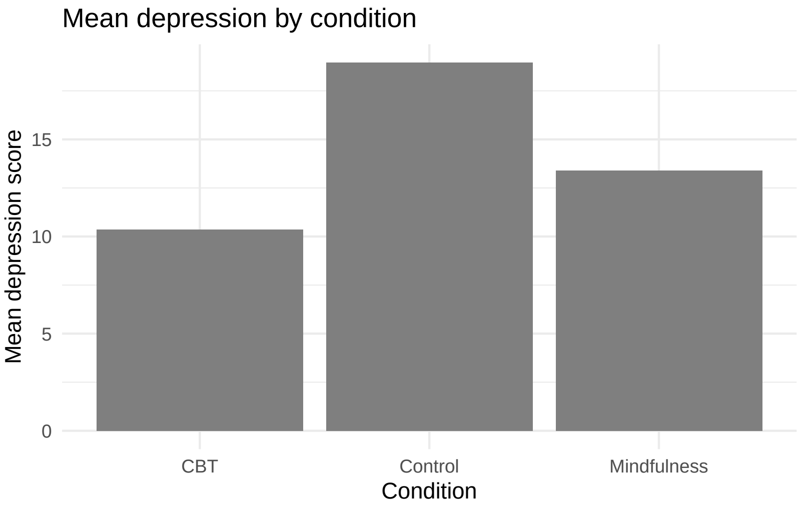
- **Position** (most powerful)
- **Size**
- **Color** (especially contrast)
- **Shape**

. . .

Use these to **direct attention to what matters**

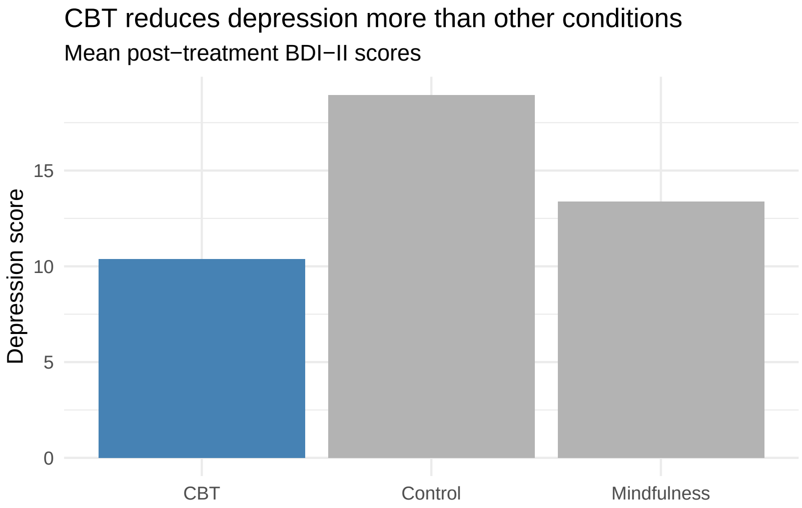
### Example: Without focus

```r
therapy_summary <- therapy_data |>
  group_by(condition) |>
  summarize(mean_depression = mean(depression))

ggplot(therapy_summary, aes(x = condition, y = mean_depression)) +
  geom_col(fill = "gray50") +
  labs(
    title = "Mean depression by condition",
    x = "Condition",
    y = "Mean depression score"
  ) +
  theme_minimal()
```



## Example: With focus

```r
therapy_summary <- therapy_summary |>
  mutate(highlight = if_else(condition == "CBT", "Highlight", "Normal"))

ggplot(therapy_summary, aes(x = condition, y = mean_depression, fill = highlight)) +
  geom_col() +
  scale_fill_manual(values = c("Highlight" = "steelblue", "Normal" = "gray70")) +
  labs(
    title = "CBT reduces depression more than other conditions",
    subtitle = "Mean post-treatment BDI-II scores",
    x = NULL,
    y = "Depression score"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



## Step 5: Think like a designer

**Visual hierarchy** guides the eye:

1. **Title** — what should they remember?
2. **Main visual** — the data
3. **Supporting elements** — axes, labels, legend
4. **Context** — subtitle, caption, notes

. . .

**Size matters:**

- Important = bigger
- Secondary = smaller

## Effective titles

**Bad title:** "Depression scores by condition"

. . .

**Better title:** "CBT most effective at reducing depression"

. . .

**Even better (with context):**

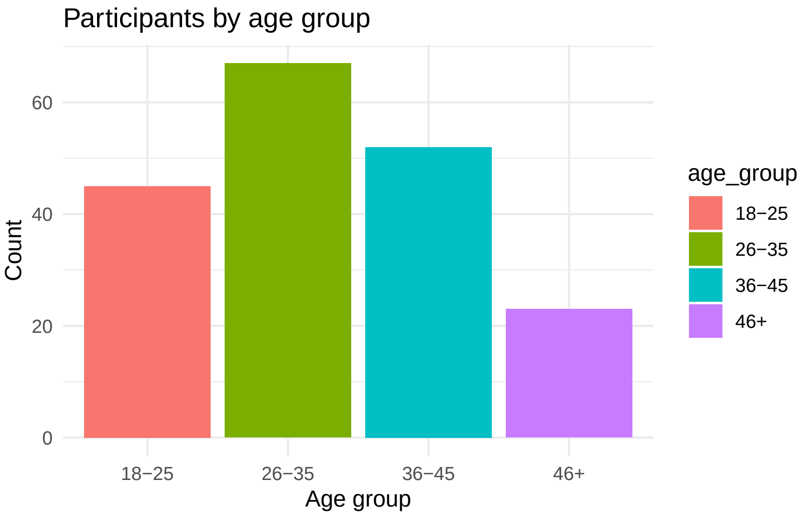
- Title: “CBT reduces depression by 8 points”
- Subtitle: “Compared to control (2 points) and mindfulness (4 points)”

### Color best practices

1. **Use color purposefully** — to highlight, not decorate
2. **Be colorblind-friendly** — use `viridis` or `ColorBrewer`
3. **Limit your palette** — 3-5 colors maximum
4. **Consider meaning** — red = danger/bad, green = good, blue = neutral

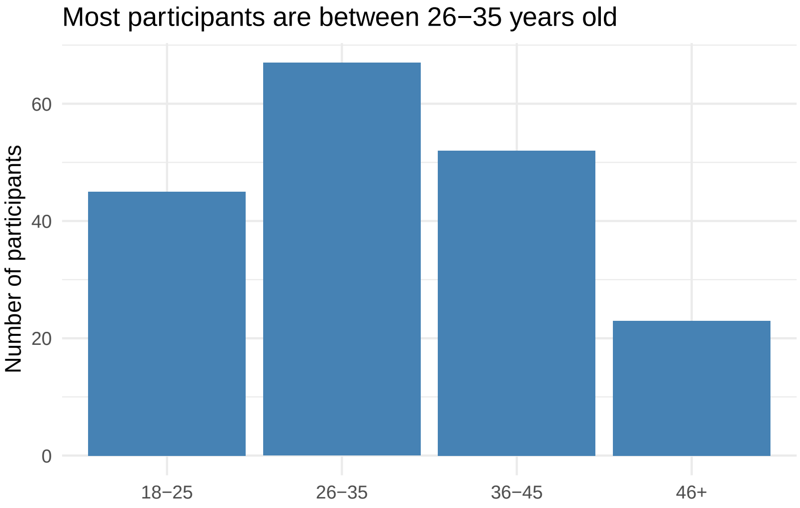
### Color example: Before

```r
demographics <- tibble(
  age_group = c("18-25", "26-35", "36-45", "46+"),
  count = c(45, 67, 52, 23)
)

ggplot(demographics, aes(x = age_group, y = count, fill = age_group)) +
  geom_col() +
  labs(title = "Participants by age group", x = "Age group", y = "Count") +
  theme_minimal()
```



### Color example: After

```
ggplot(demographics, aes(x = age_group, y = count)) +
  geom_col(fill = "steelblue") +
  labs(
    title = "Most participants are between 26-35 years old",
    x = NULL,
    y = "Number of participants"
  ) +
  theme_minimal()
```



## Critical evaluation of figures

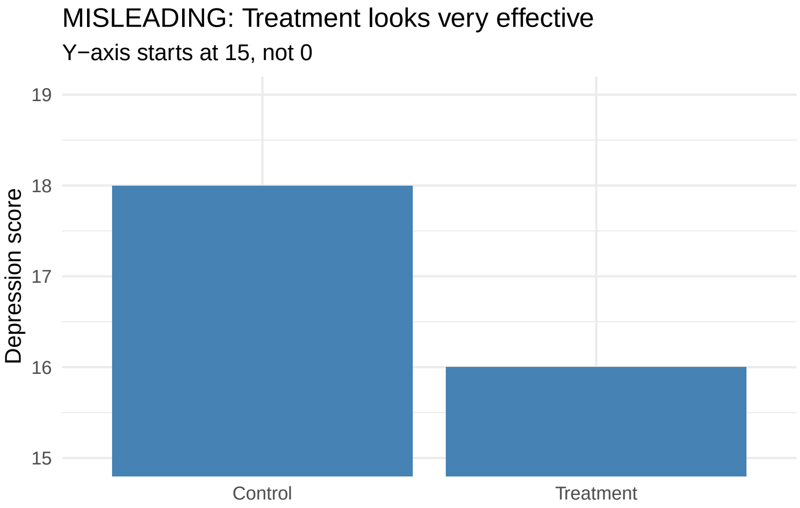
### Misleading figures

Visualizations can deceive (intentionally or not):

1. **Truncated y-axes** — exaggerate small differences
2. **Dual axes with different scales** — imply false relationships
3. **Cherry-picked time ranges** — hide broader trends
4. **3D charts** — distort perception of size
5. **Area/bubble charts** — hard to compare sizes accurately

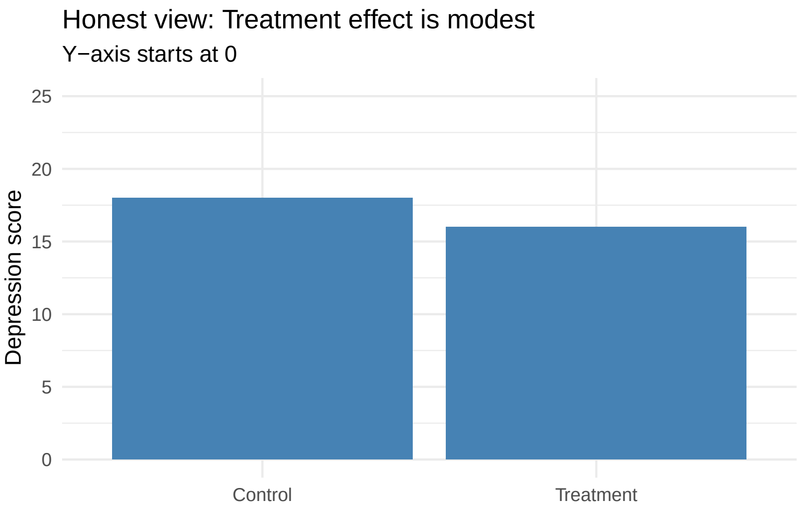
### Example: Truncated y-axis

```r
treatment_effect <- tibble(
  condition = c("Control", "Treatment"),
  score = c(18, 16)
)

ggplot(treatment_effect, aes(x = condition, y = score)) +
  geom_col(fill = "steelblue") +
  coord_cartesian(ylim = c(15, 19)) +  # Truncated!
  labs(
    title = "MISLEADING: Treatment looks very effective",
    subtitle = "Y-axis starts at 15, not 0",
    x = NULL,
    y = "Depression score"
  ) +
  theme_minimal()
```



### Fixed: Full y-axis

```r
ggplot(treatment_effect, aes(x = condition, y = score)) +
  geom_col(fill = "steelblue") +
  coord_cartesian(ylim = c(0, 25)) +  # Full scale
  labs(
    title = "Honest view: Treatment effect is modest",
    subtitle = "Y-axis starts at 0",
    x = NULL,
    y = "Depression score"
  ) +
  theme_minimal()
```



## When truncated axes are okay

### Truncation is fine when:

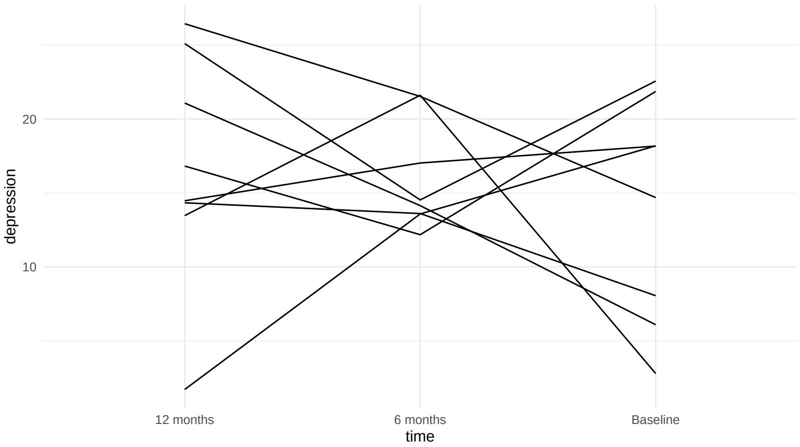
- The baseline is non-zero (e.g., human body temperature)
- You’re showing change over time (line plot)
- You explicitly note it in the caption

. . .

### Never truncate:

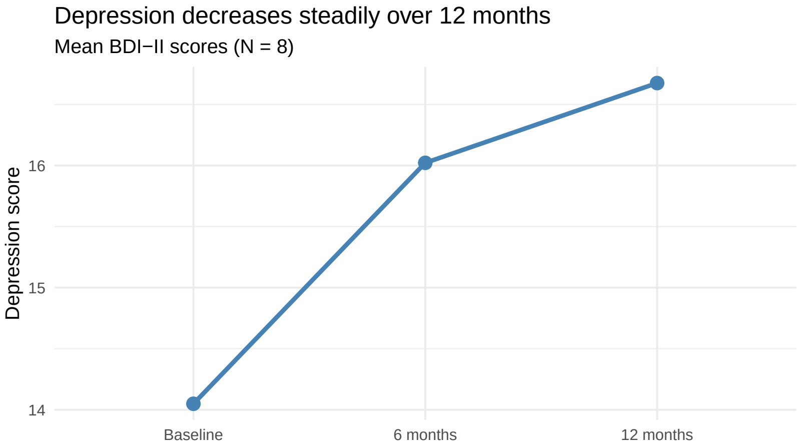
- Bar charts (bars must start at zero)
- When comparing magnitudes

## Boring: Spaghetti plot with no message

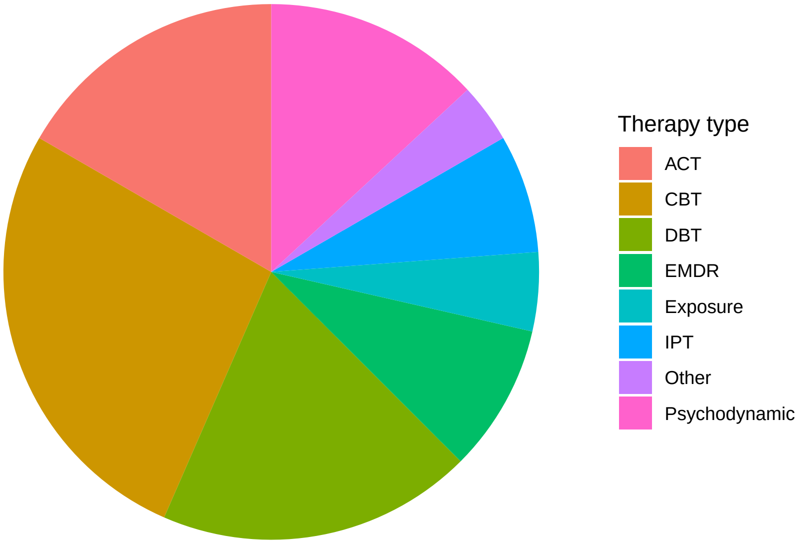


Eight lines going everywhere. What's the takeaway?

## Fixed: One clear message

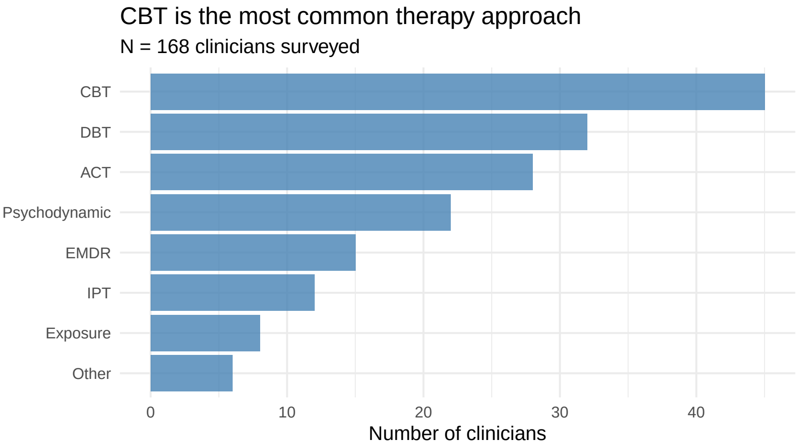


## Boring: Default pie chart



Eight slices. Which is biggest? By how much?

## Fixed: Ranked bar chart tells the story



## The “so what?” test

Every figure should answer a question:

- “Depression scores by condition”

- "CBT reduces depression more than control"

. . .

- "Correlation between age and reaction time"
- "Older adults respond 50ms slower per decade"

. . .

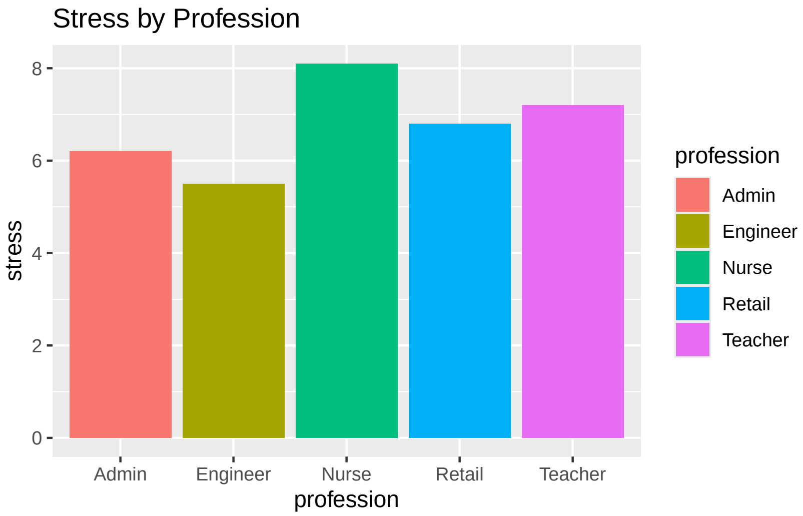
**Ask yourself:** If someone only sees this figure for 5 seconds, what should they remember?

## Pair coding break

### Your turn: Improve a figure

Here's a messy figure:

```r
stress_data <- tibble(
  profession = c("Teacher", "Nurse", "Engineer", "Retail", "Admin"),
  stress = c(7.2, 8.1, 5.5, 6.8, 6.2),
  burnout = c(6.8, 7.9, 4.8, 6.5, 5.9)
)

ggplot(stress_data, aes(x = profession, y = stress, fill = profession)) +
  geom_col() +
  labs(title = "Stress by Profession") +
  theme_gray()
```



1. Remove the unnecessary legend
2. Reorder professions by stress level
3. Highlight the profession with highest stress
4. Add a clear, message-driven title
5. Clean up the theme

**Time: 10 minutes**

## Applying to your final project

### Your final project narrative

Your final project should tell a story with three acts:

1. **Setup (Introduction)**
   - What's the question?
   - Why does it matter?
   - What's your hypothesis?
2. **Conflict (Results)**
   - What did you find?
   - Show with visualizations
   - Highlight surprising or important patterns
3. **Resolution (Discussion)**

- What does it mean?
- How does it answer your question?
- What should we do with this information?

## Building a narrative arc

**Weak narrative:**

> "I looked at depression and anxiety. Here's a histogram. Here's a scatterplot. Here's a boxplot. The correlation was 0.65."

. . .

**Strong narrative:**

> "Depression and anxiety often co-occur, but we don't know how strongly they're related in college students. I analyzed 200 student surveys and found a strong correlation (r = .65). This suggests these conditions may share underlying mechanisms and should be treated together."

## One thing to remember

At the end of your presentation, your audience should remember **one key takeaway**.

. . .

**What's yours?**

- "Social media use predicts anxiety in teens"
- "Mindfulness training reduces stress in nurses"
- "Memory declines linearly after age 50"
- "Treatment dropout is higher in low-income participants"

. . .

Every figure, every sentence should support that one key message.

### Practical tips for final projects

1. **Start with your key finding** — then work backward
2. **One main point per figure** — don't try to show everything at once
3. **Order matters** — build up to your main finding
4. **Edit ruthlessly** — remove anything that doesn't support your story
5. **Get feedback** — show your figures to someone outside the class

## APA figure formatting

### A note on APA formatting

You'll receive a handout on APA figure formatting guidelines.

. . .

**Key points:**

- Journals have **different** requirements
- APA is a starting point, not gospel
- Many journals want editable figures (not embedded in Word)
- Online journals have more flexibility than print

. . .

> ℹ Note
>
> Focus on **clarity and communication** first, then adjust formatting as needed for specific journals.

### Basic APA figure elements

1. **Figure number** — "Figure 1"
2. **Title** — brief and descriptive
3. **Image** — the actual plot
4. **Note** — additional context, definitions, copyright

. . .

Not included in the figure itself:

- Legend goes in figure note (if needed)
- No borders around the figure
- Black and white or grayscale (for print journals)

## End-of-deck exercise

### Critique these figures

For each figure, identify:

1. What story is it trying to tell?
2. What works well?
3. What could be improved?
4. How would you redesign it?

Look at 2-3 example figures from published psychology papers (I'll show them in class).

Then, apply these principles to **your own final project draft**.

## Wrapping up

### The storytelling checklist

Before finalizing any figure, ask:

- □ **Is the message clear?** Can someone understand it in 5 seconds?
- □ **Is there clutter?** Can I remove anything without losing meaning?
- □ **Does color help or distract?** Am I using it purposefully?
- □ **Is it accessible?** Colorblind-friendly? Good contrast?
- □ **Does it answer a question?** Can I state the "so what"?
- □ **Is it honest?** Am I representing the data fairly?

## Key takeaways

1. **Data + Visuals + Narrative = Change**
2. **Know your audience** — tailor your message to who you're speaking to
3. **Eliminate clutter** — less is more
4. **Focus attention** — use preattentive attributes strategically
5. **Be honest** — don't mislead with truncated axes or cherry-picked data
6. **Pass the "so what?" test** — every figure should answer a question
7. **One key message** — what do you want them to remember?

## Resources for continued learning

- **Knaflic (2015).** *Storytelling with Data* (on reserve)
- **Dykes (2016).** "Data Storytelling: The Essential Data Science Skill"
- **APA Style:** Figure formatting guidelines (handout provided)
- **ColorBrewer:** colorbrewer2.org for colorblind-safe palettes

## Before next class

**Read:**

- R4DS Ch 28: Quarto

**Do:**

- **Submit Assignment 7** (due today)
- **Submit Final Project Draft** (due today)
- Revise your figures based on today's principles
- Review feedback on your draft

### The one thing to remember

A figure without a story is just a picture. Ask “so what?” until you have the answer.

See you next week for Quarto and reproducible reports!