Untangling the Knots: Seeing Clearly in a World of Messy Data

Denise Bradford

2025-06-13

Table of contents

The Story We Want Data to Tell

At our core, we are pattern finders. Whether we're scientists searching for a breakthrough, business analysts seeking a competitive edge, or just curious individuals trying to make sense of the world, we turn to data to tell us a story. We hope for clarity, for that "aha!" moment when a complex set of numbers suddenly reveals a hidden truth.

We've developed incredible tools to assist us on this quest. One of the most powerful tools for examining multiple features simultaneously is the **Parallel Coordinate Plot (PCP)**.

Parallel Coordinates: A Simple Idea for a Complex World

Imagine you have a dataset about various types of cars, including information on their horsepower, fuel efficiency, weight, and price. How can you observe the relationships between all these factors simultaneously?

A Parallel Coordinate Plot takes a simple, elegant approach. Instead of positioning axes at right angles, like a standard chart, it aligns them one after another in parallel. Then, it draws a line for each car, connecting its specific values across all the axes.

When it works, it's remarkable. You can see clusters of lines, which represent groups of similar cars. You might identify a car whose line follows an unusual path, indicating that it's an outlier. You've transformed a sterile spreadsheet into a dynamic, visual narrative.

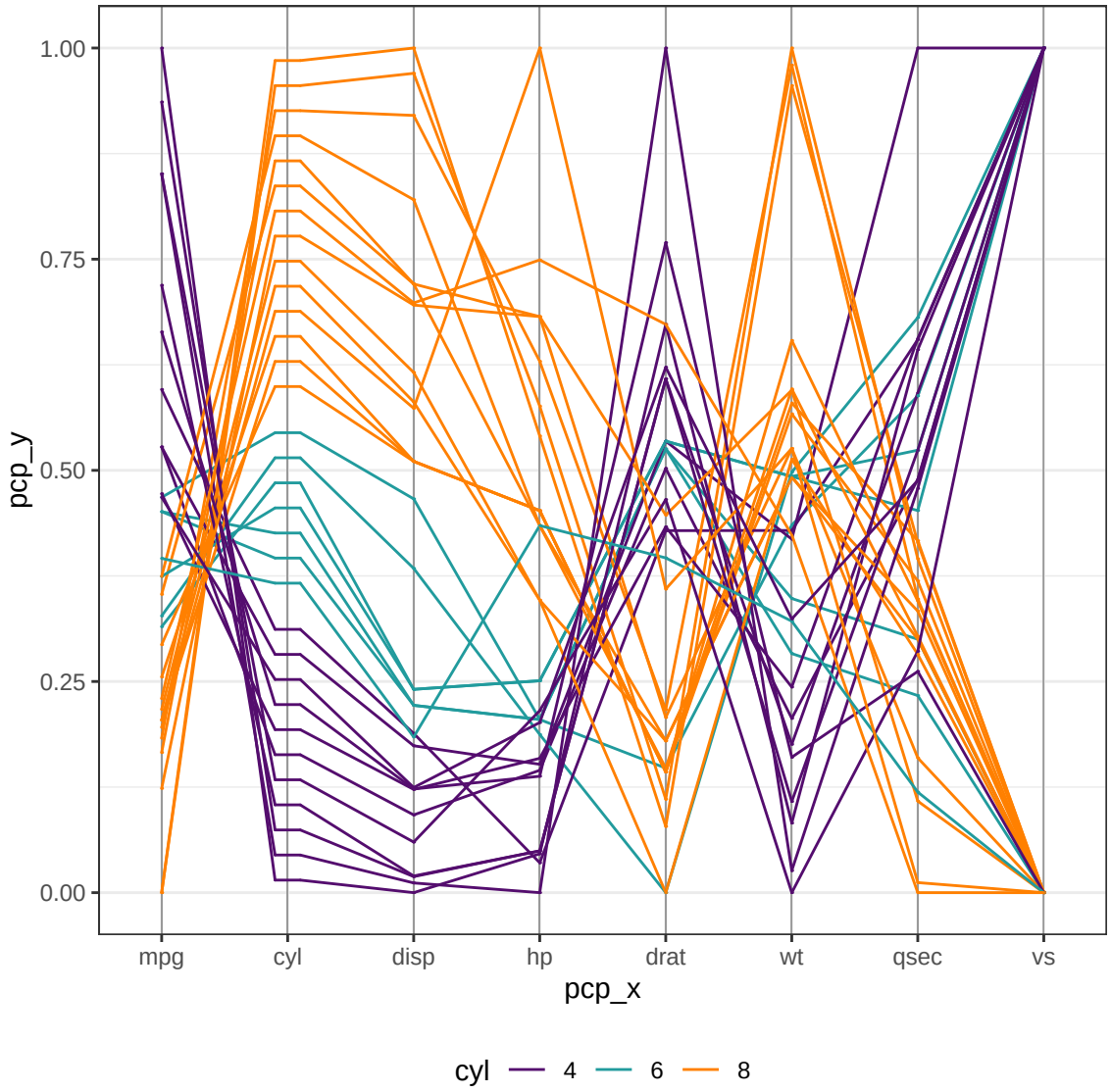

Figure 1: A clean and effective Parallel Coordinate Plot showing clear patterns in the carss dataset.

The Wall of Overlap: When the Story Gets Lost

But often, our data isn't as neat as we would like. The real world is filled with values that aren't unique. Consider data that comes from:

- **Survey responses:** (e.g., "On a scale of 1 to 5…")
- **Categorical labels:** (e.g., "Small," "Medium," "Large")
- **Rounded numbers:** (e.g., sensor data recorded to the nearest degree)

When many data points share the same exact value on an axis, their lines must all pass through the exact same point. This creates the "Wall of Overlap"—a visual traffic jam where dozens, or even thousands, of lines converge into an unreadable, solid block.

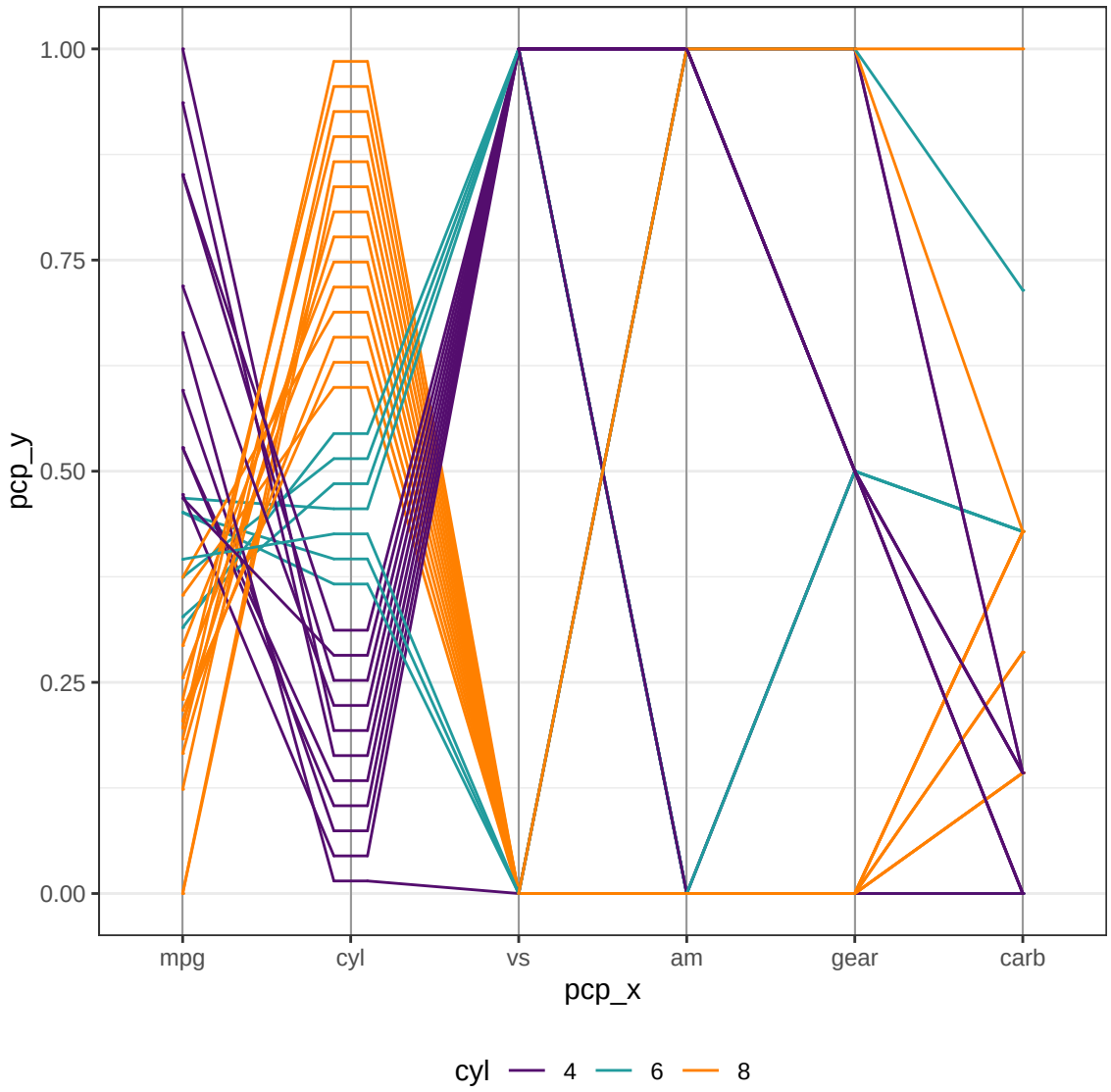

Figure 2: An unusable PCP where numerical ties have created a dense, uninformative mesh of lines, completely hiding any underlying patterns.

At this moment, our beautiful story is lost. The plot, which was meant to clarify things, now creates confusion. It's frustrating because we know patterns are hidden somewhere, but the visualization itself obstructs our view.

A Tale of Two Philosophies

How we should address this problem is a genuine debate among data visualization experts. The discussion is framed by the "Grammar of Graphics," a foundational set of principles for creating charts.

The Purist's View: "Show the Data, Warts and All"

One side of the argument is that a visualization's primary duty is to be truthful to the raw data. This viewpoint is rooted in the formal system proposed by Leland Wilkinson in his foundational book, "The Grammar of Graphics," [@wilkinson2005]. His work treats graphics as the direct result of mapping data to geometry, meaning the resulting plot should show overlapping lines if the data has ties. Similarly, the influential work of Edward Tufte in "The Visual Display of Quantitative Information" argues against any visual element that distorts the data, championing graphical integrity above all [@tufte1983visual]. From this perspective, spreading lines apart is a form of manipulation—it's portraying something that isn't literally present in the data, which could be misleading. They argue that the responsibility lies with the person creating the chart to acknowledge their data's complexity.

The Pragmatist's View: "A Chart Should Provide Insight"

The other side argues that the ultimate goal of a visualization is to help a person understand something. If a chart is technically "truthful" but practically useless, has it really succeeded? This philosophy is championed by Hadley Wickham's implementation of a "Layered Grammar of Graphics" in the widely used `ggplot2` package, which includes practical tools like `position_jitter()` specifically designed to solve overplotting. Following this tradition, recent work by @vanderplas2023 on parallel coordinate plots directly confronts the issue, arguing that we need thoughtful, principled ways to untangle these knots. The idea isn't to lie, but to gently pull the threads apart so we can follow them again. Spreading the lines at a tie-point isn't hiding the truth; it's revealing it by showing us how many lines were tied up in that knot.

The Path Forward: A Call for Clarity and Evidence

The goal isn't to pick a side in the debate, but to build a bridge between them. This work acknowledges that the pragmatists are correct—we need to see our data—but that the purists have a valid point—we must do it principally and honestly. The research plan is based on a simple idea: let's figure out the best way to untangle these lines and demonstrate that it helps people see better, more clearly, and confidently.

This research plan progresses from theory to practice in three distinct stages:

1. Theorize: Building a Principled Foundation: The first step is to formally define how to address visual ties within the respected Grammar of Graphics framework. Instead of an ad-hoc fix,

this approach will create a rigorous, predictable, and theoretically sound foundation for the solution.

2. Innovate: Designing a Smarter Solution Next, the research will focus on creating a novel algorithm that can separate the tangled lines. The goal is to develop a method that is more intelligent than simple jittering—one that is clear, informative, and preserves the essential characteristics of the original data.

3. Validate: Proving the Human Impact Finally, and most importantly, the new method must be tested with real users. Through controlled studies, we can answer critical questions:

- Does this method help analysts find patterns faster?
- Does it lead to more accurate conclusions?
- Does it increase confidence in the insights drawn from the chart?

Ultimately, this work is about empowering people. It's about transforming those frustrating "walls of overlap" back into the rich, meaningful stories we knew were there all along. It's about making our data visualization tools smarter, more honest, and more helpful in our universal quest to make sense of a complex, multidimensional world.