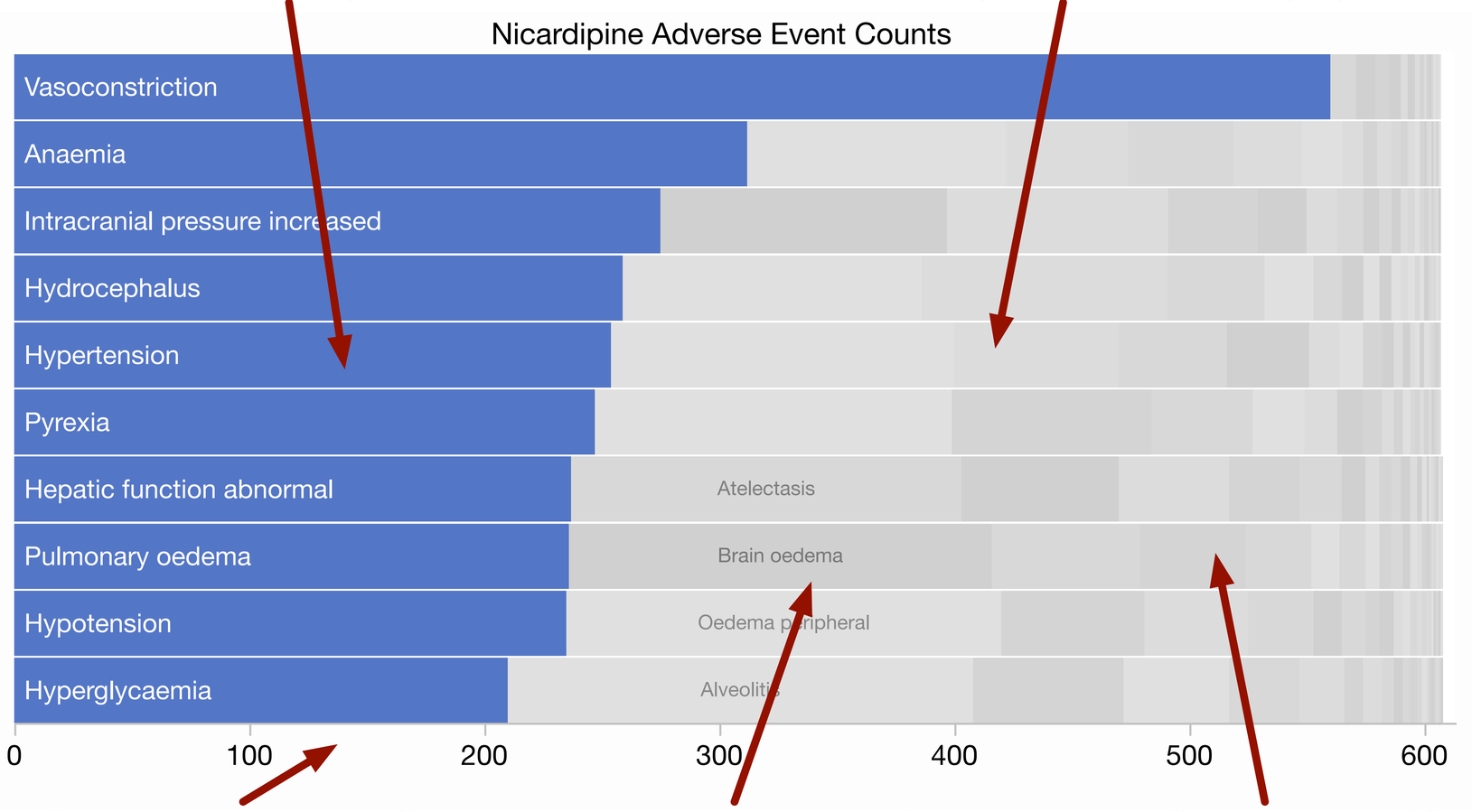
Introducing the Packed Bars Chart Type

Xan Gregg, SAS Institute, JMP Development, @xangregg



The *packed bars* chart type attempts to bring the **Focus+Context** principle to bar charts. Focus+Context is a powerful technique for seeing the important parts of a data set (the focus) with high fidelity while at the same time seeing the supporting data (the context) with a low-fidelity, details-on-demand view. Typically, the focus elements are overlaid in the foreground layer using prominent colors with the context elements in the background using muted colors. Simple layering doesn't work for bar charts. In a packed bars chart, an ordered bar chart of the top categories represents the focus component, and all other categories are *packed* into the space beyond the focus bars with de-emphasized coloring.

The featured example shows a typical packed bars chart. The data set contains 201 unique adverse events recorded from a clinical trial for the drug nicardipine. The 10 most common adverse events are represented by the primary bars, which are highlighted in blue. By themselves, the primary bars form a useful bar chart, sometimes called a "top 10" chart. The length of each bar corresponds to the response variable, which in this case is the count of the occurrences of each event. The remaining adverse events are each represented by a bar with a muted color. These secondary bars lengths are on the same scale as the primary bars, which allows them to share the x axis. The secondary bars are not necessarily labeled.

The packed bars form has all the features of an ordered bar chart plus the following information from the secondary bars in a compact form (barely any additional space in this case).

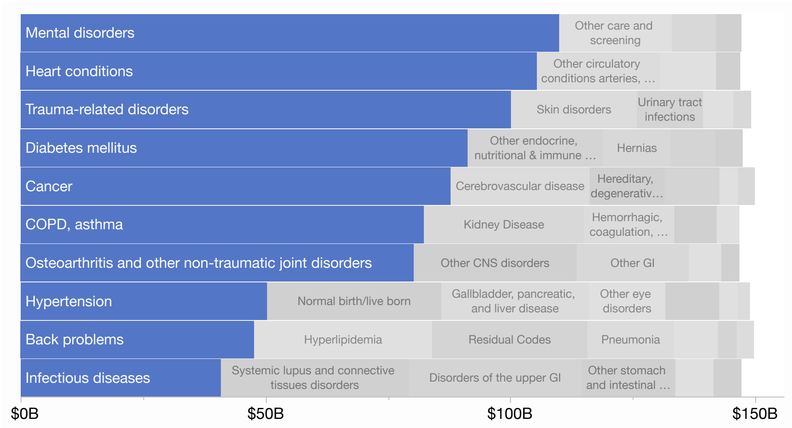
- An estimate of the relative areas of the primary and grand totals. Here we can estimate that the top 10 adverse events account for almost half of the total event occurrences.
- Values for the larger of the secondary categories. Though the secondary bars don't have the aligned baselines like the primary bars do, we can still reasonably estimate bar length in the presence of the x axis.
- A sense of the distribution of the secondary categories. Here we can see there are another 10 or so decent sized categories before the sizes drop off quickly.
- An estimate of the primary bars as percentages of the grand sum. With 10 primary bars, each row of bars represents 10% of the grand sum. Since the top category, vasoconstriction, takes up most of the first row, it represents almost 10% of all adverse events.
- An estimate of the grand sum itself. Especially when the number of primary categories is a round number like 10, we can multiply it by the axis extent to get an estimate of the grand sum. In this case it's 10 rows × 600+ occurrences = 6000+ occurrences.

The packing algorithm fills the space greedily from left to right with bars of descending size. Placing the smallest bars last usually results in a rectangular fill, but the right edge may be jagged when there are few small bars.

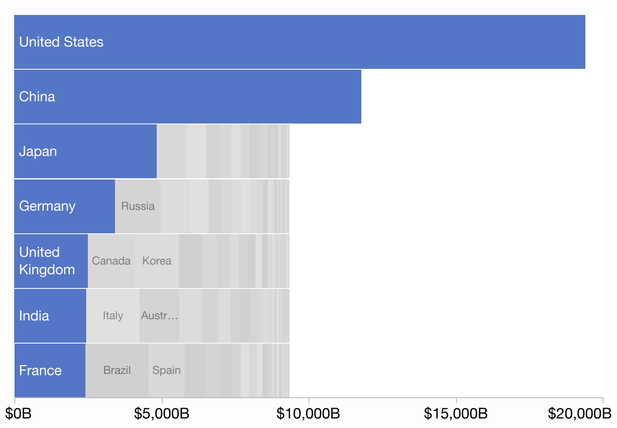
Implementation scripts are available at github.com/xangregg/packedbars.

Comparing cardinality

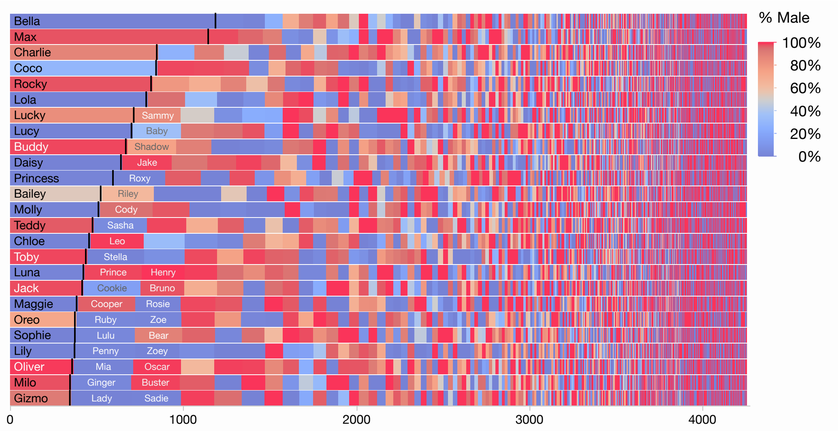
Packed bars scale from dozens to thousands of categorical levels.



75 disease categories of US medical spending. Relatively low cardinality allows for more labeling of secondary categories but leaves a more jagged right edge.



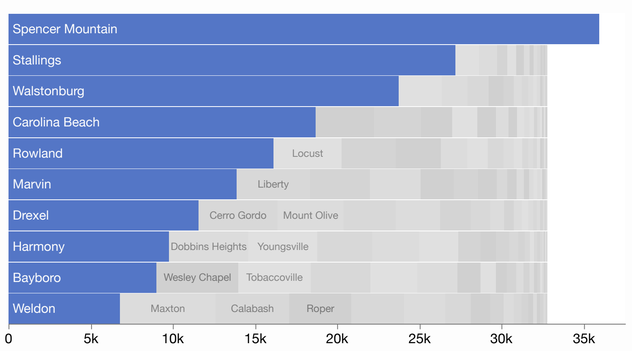
200 world economies sized by nominal GDP. For highly skewed data, the top bar or two will extend beyond the packed area.



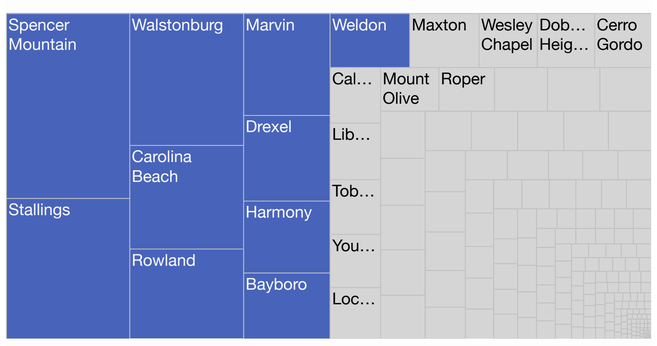
16,000 New York City dog names, colored by gender ratio. When color is important for secondary bars, the primary bars are separated with bold end-caps.

Comparing chart forms

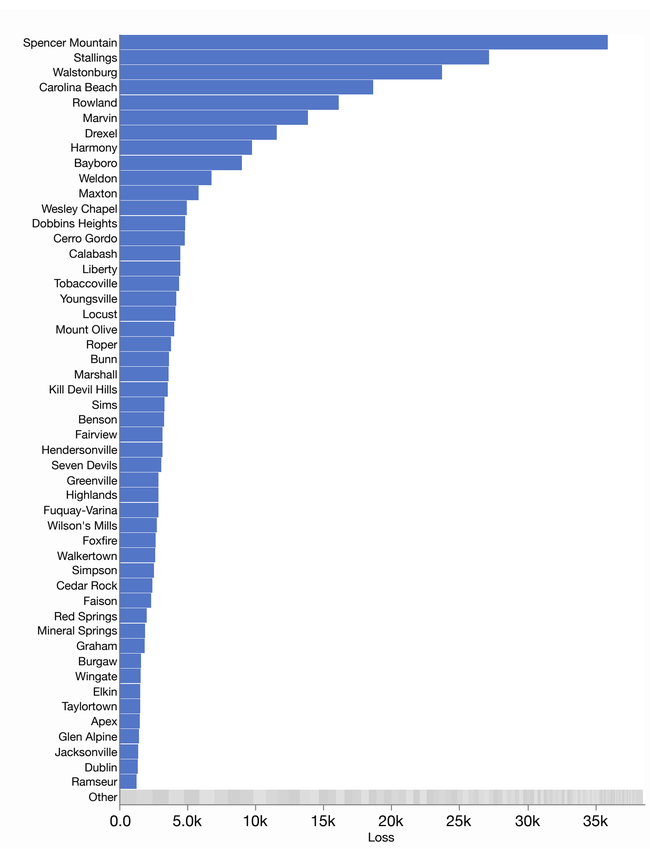
Five views of the same quality control data set containing production loss values for 200 industrial sites (names anonymized).



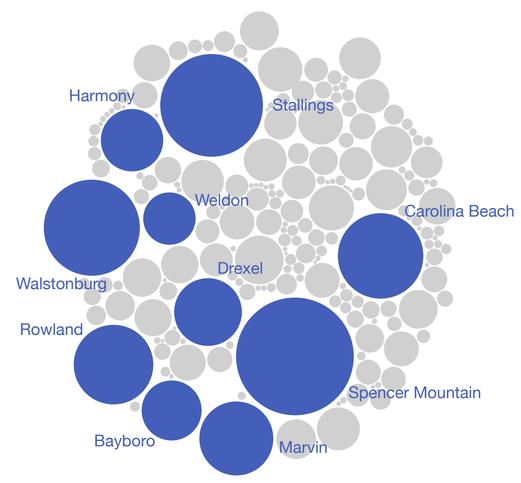
A **packed bars** chart effectively shows the values of the top 10 sites and their relation to the grand sum, accounting for about half of the total production loss. The top site accounts for over 10%.



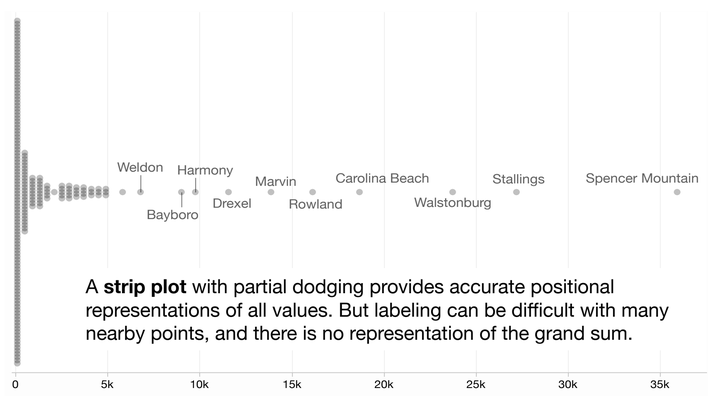
A squarified **treemap** provides a more regular representation of the part-to-whole proportion for the top 10 sites. But areas are less precise than anchored lengths, and the actual values aren't available without labels.



A traditional **bar chart** doesn't scale well to 200 categories. Often the small categories are omitted entirely or aggregated in an "Other" bar as done here. More of the distribution can be clearly seen at the cost of greater space consumption. There is little sense of aggregate proportions.



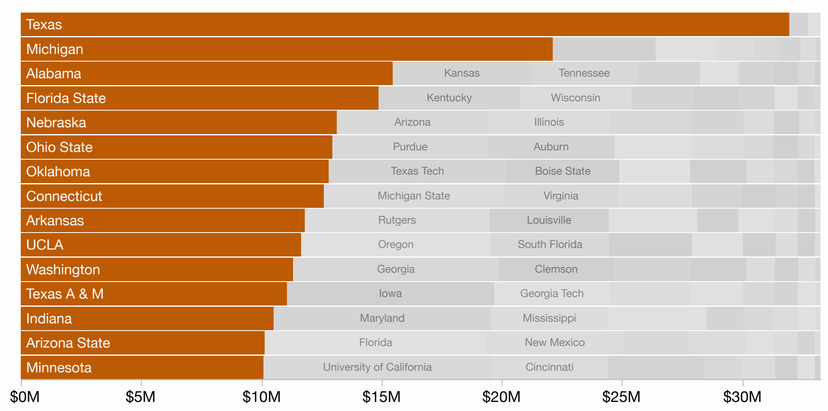
Packed bubbles show all the sites in a compact view while highlighting the top values. Size comparisons based on circle area are more difficult, and the aggregate area is obscured by packing gaps.



A **strip plot** with partial dodging provides accurate positional representations of all values. But labeling can be difficult with many nearby points, and there is no representation of the grand sum.

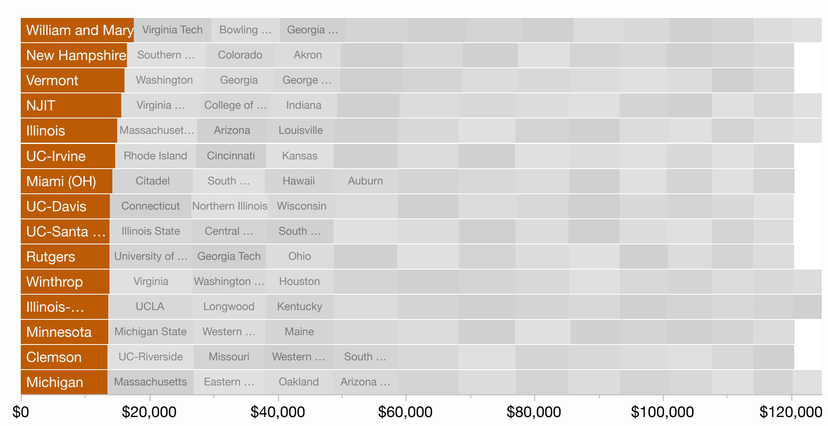
Comparing value distributions

Three variables for 200 public universities: athletic royalties, enrollment and in-state tuitions.



Packed bars excel at **exponential** distributions, such as athletic royalties. The top bars take up a sizable portion of the total, and the smallest bars provide a smooth right edge.

Ordered enrollment values decrease **linearly**. A steady decrease can be seen in the bar length from left to right.



The distribution of in-state tuition values is roughly **uniform**. There is little variation in the secondary bar sizes, and the right edge is noticeably chunky.