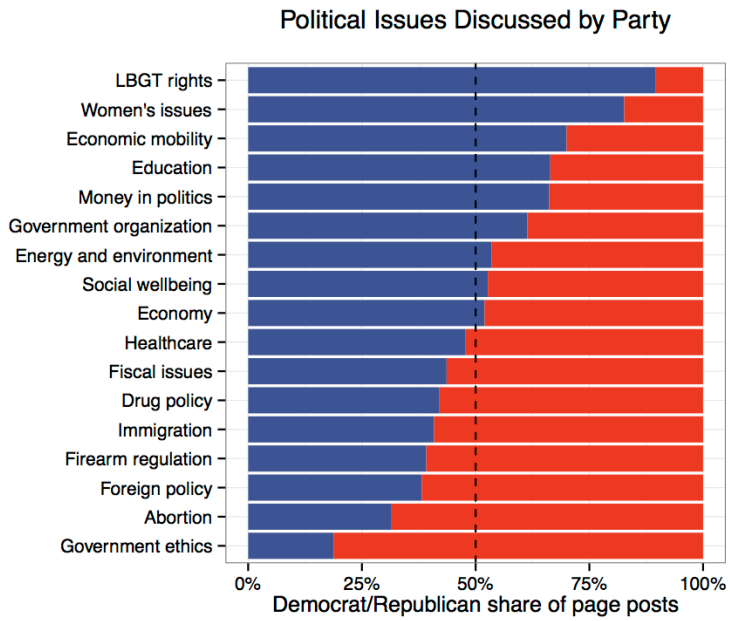
Solomon Messing

When to Use Stacked Barcharts?

Posted on October 11, 2014

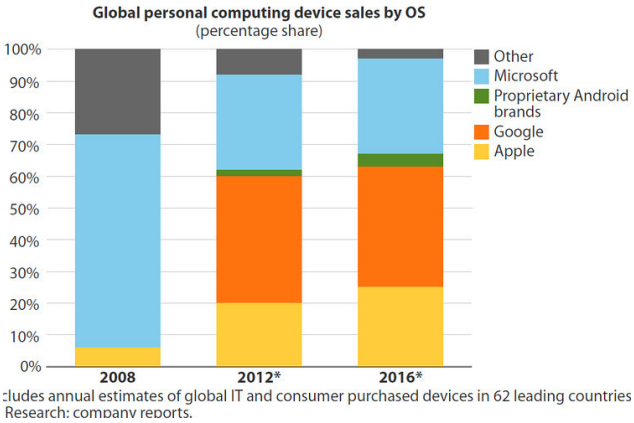


Yesterday a few of us on Facebook's Data Science Team released a blogpost showing how candidates are campaigning on Facebook in the 2014 U.S. midterm elections. It was picked up in the Washington Post, in which Reid Wilson calls us "data wizards." Outstanding.

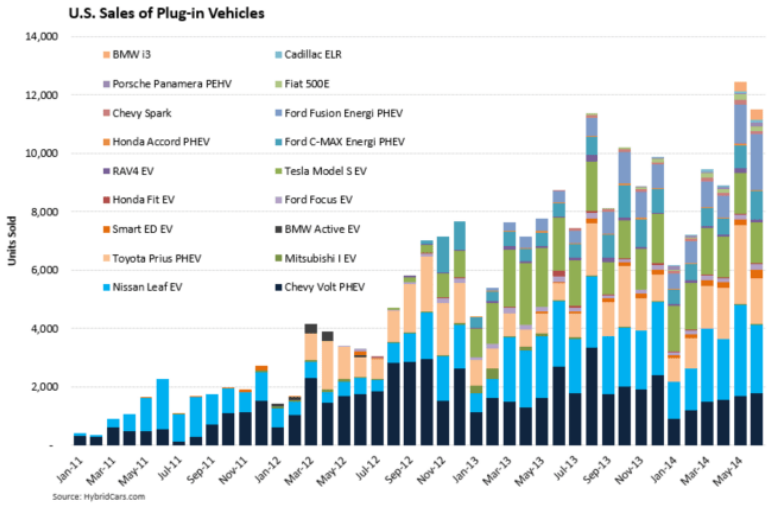
I used Hadly Wickham's ggplot2 for every visualization in the post except a map that Arjun Wilkins produced using D3, and for the first time I used stacked bar charts. Now as I've stated previously, one should generally avoid bar charts, and especially stacked bar charts, except in a few specific circumstances.

But let's talk about when not to use stacked bar charts first—I had the pleasure of chatting with Kaiser Fung of JunkCharts fame the other day, and I think what makes his site so compelling is the mix of schadenfreude and Fremdscham that makes taking apart someone else's mistake such an effective teaching strategy and such a memorable read. I also appreciate the subtle nod to junk art.

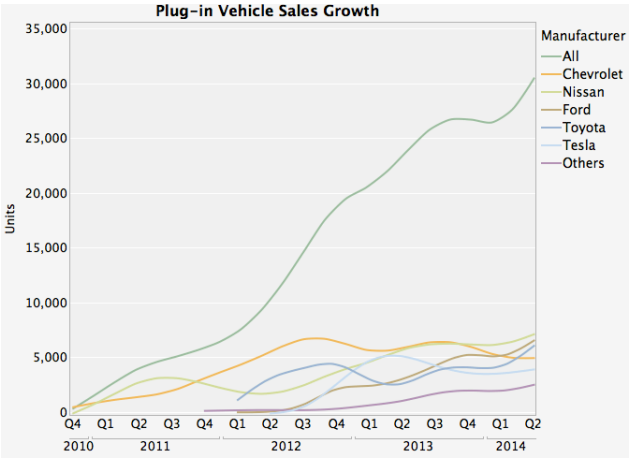
Here's a typical, terrible stacked bar chart, which I found on http://www.storytellingwithdata.com/ and originally published on a Wall Street Journal blogpost. It shows the share of the personal computing device market by operating system, over time. The problem with using a stacked bar chart is that there are only two common baselines for comparison (the top and bottom of the plotting area), but we are interested in the relative share for more than two OS brands. The post is really concerned with Microsoft, so one solution would be to plot Microsoft versus the rest, or perhaps Microsoft on top versus Apple on the bottom with "Other" in the middle. Then we'd be able to compare the over time market share for Apple and Microsoft. As the author points out, an over time trend can also be visualized with line plots.



By far the worst offender I found in my 5 minute Google search was [from junkcharts] and originally published on [Vox]. These cumulative sum plots are so bad I was surprised to see them still up. The first problem is that the plots represent an attempt to convey way too much information—either plot total sales or pick a few key brands that are most interesting and plot them on a multi-line chart or set of faceted time series plots. The only brand for which you can quickly get a sense of sales over time is the Chevy Volt because it's on the baseline. I'm sure the authors wanted to also convey the proportion of sales each year, but if you want to do that just plot the relative sales. Of course, the order in which the bars appear on the plot has no organizing principle, and you need to constantly move your eyes back and forth from the legend to the plot when trying to make sense of this monstrosity.

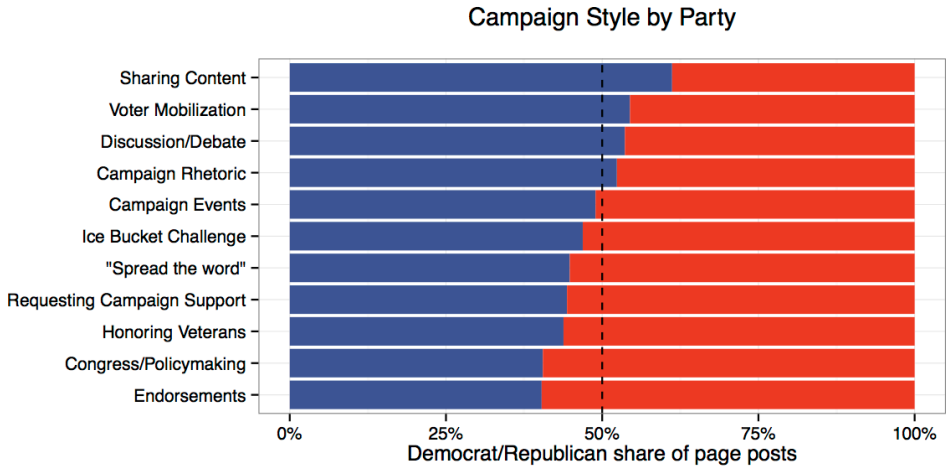


As Kaiser notes in his post, less is often more. Here's his redux, which uses lines and aggregates by both quarter and brand, resulting in a far superior visualization:

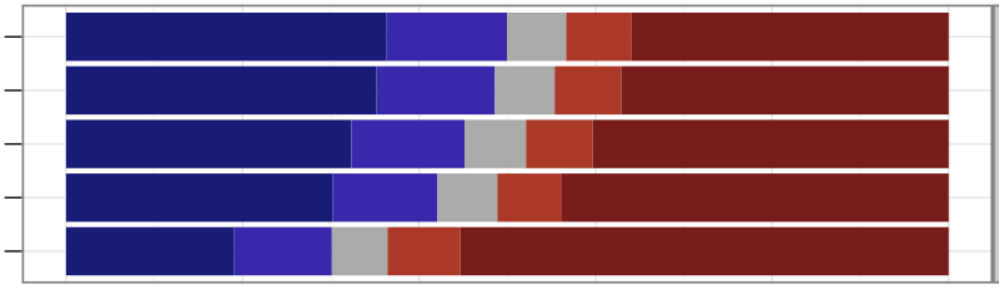


So when *should* you use a stacked bar chart? Here are a two scenarios with examples, inspired by work with Eytan Bakshy and conversations with Ta Chiraphadhanakul and John Myles White.

1. You care about comparing the proportion of two things, in this case the share of posts by Democrats and Republicans, along a variety of dimensions. In this case those dimensions consist of keyword (dictionary-based) categories (above) and LDA topics (below). When these are sorted by relative proportion, the reader gains insight into which campaign strategies and issues are used more by Republican or Democratic candidates.



2. You care about comparing proportions along an ordinal, additive variable such as 5-point party identification, along a set of dimensions. I provide an example from a forthcoming paper below (I'll re-insert the axis labels once it's published). Notice that it draws the reader toward two sets of comparisons across dimensions — one for strong democrats and republicans, the other for the set of *all* Democrats and *all* Republicans.



Of course, R code to produce these plots follows:

```
# Uncomment these lines and install if necessary:
#install.packages('ggplot2')
#install.packages('dplyr')
#install.packages('scales')

library(ggplot2)
library(dplyr)
library(scales)

# We start with the raw number of posts for each party for
# each candidate. Then we compute the total by party and
# category.

catsByParty %>% group_by(party, all_cats) %>%
summarise(tot = summ(posts))

# Next, compute the proportion by party for each category
# using dplyr::mutate
catsByParty <- catsByParty %>%
group_by(all_cats) %>%
mutate(prop = tot/sum(tot))

# Now compute the difference by category and order the
# categories by that difference:
catsByParty <- catsByParty %>% group_by(all_cats) %>%
    mutate(pdiff = diff(prop))

catsByParty$all_cats <- reorder(catsByParty$all_cats, -catsByParty$pdiff)

# And plot:
ggplot(catsByParty, aes(x=all_cats, y=prop, fill=party)) +
scale_y_continuous(labels = percent_format()) +
geom_bar(stat='identity') +
geom_hline(yintercept=.5, linetype = 'dashed') +
coord_flip() +
theme_bw() +
ylab('Democrat/Republican share of page posts') +
xlab('') +
scale_fill_manual(values=c('blue', 'red')) +
theme(legend.position='none') +
ggtitle('Political Issues Discussed by Party\n')
```

Share this:

Twitter Facebook 3

Like

One blogger likes this.

Related

Visualization series: Insight from Cleveland and Tufte on plotting numeric data by groups
In "R"

Visualization Series: Using Scatterplots and Models to Understand the Diamond Market
In "R"

Putting it all together: concise code to make dotplots with weighted bootstrapped standard errors
In "R"

About Solomon
Political Scientist, Facebook Data Science
View all posts by Solomon →

This entry was posted in R. Bookmark the permalink.

12 Responses to *When to Use Stacked Barcharts?*

Andrew Clark (@pssGuy) *says:*
October 12, 2014 at 9:51 am

TBH even these examples require more to illuminate. The campaign style chart gives the impression that all methods are equally important whilst 'Endorsements' might still be 50x more important than 'Saring Content' even for Democrats

Reply

Solomon *says:*
October 12, 2014 at 10:26 pm

Hey Andrew, right, these are purely descriptive. You should not conclude that one campaign style is more effective based on this visual. I'd encourage you to have a look at the full post with context if you haven't already at https://www.facebook.com/notes/facebook-data-science/campaign-rhetoric-and-style-on-facebook-in-the-2014-us-midterms/10152581594083859

Reply

John Blommers *says:*
October 12, 2014 at 7:34 pm

The R code does not run. The first error is in line 12. Where does the data come from that drives the code?

Reply

Andrew Clark (@pssGuy) *says:*

October 13, 2014 at 6:02 pm

I have had a look now. As you say, gives context. Some elegant combo of the two tables in that section would be interesting

Reply

edomaniac *says:*

October 14, 2014 at 11:40 am

The last one has me thinking a bit. It seems it would be of interest in a traditional 5 point rating scale (agree/disagree) situation if you were interested in polarity (or lack thereof) in responses. It seems that this information could be informative where just plotting the means (probably via dot plot as you advocate) might obscure some differences in response styles to questions. Am I on the wrong track? And are there any issues with such an approach?

Reply

Solomon *says:*

October 15, 2014 at 2:04 pm

I think the use case you describe is great. But you might run into problems if you're interested in the difference in the distribution across subgroups. If you were to examine the proportion of responses across each the 5 levels of agreement when you look at each of those contrasts you can quickly run into multiple comparisons issues, especially when it comes to post-hoc analyses (e.g., if you were to look at every possible subgroup in your data set). This is less of an issue when your contrasting just the mean of (ideally a battery of) 5 point responses. Note that it's also less of an issue when you are working with larger data sets that yield more precise estimates.

Reply

edomaniac *says:*

October 15, 2014 at 4:35 pm

Awesome. Your point is well taken. I was thinking of it more as a descriptive tool than anything. I work as a consultant and I'm trying to get my colleagues to use R and ggplot more often in their reports, because as you've noted elsewhere, it helps explain data to non-experts (I'm a relative newcomer to R having been trained primarily on SPSS). The standard now seems to be to hit clients over the head with frequencies or means until their eyes glaze over, so I'm always looking for new potential graphs to make things more intelligible.

edomaniac *says:*

October 15, 2014 at 10:22 am

One other question. Should there be a corresponding dataset for the code? Perhaps that should be included in the "party" object?

Reply

Solomon *says:*

October 15, 2014 at 2:05 pm

I just posted the code as an example.

Reply

edomaniac *says:*

October 15, 2014 at 4:39 pm

Awesome, no worries. It just helps a bit for me to see what the data look like before I get to manipulating. I'll play around with the code and make something work. Thanks for posting this stuff, I'll keep reading.

Reply

Pingback: *Dime, ¿qué quieres comparar con qué? – datanalytics*

Pingback: *Notes on Data Visualization – D3.js – data | poly*