Data Expo 2013: Data Visualization on the Soul of the Community

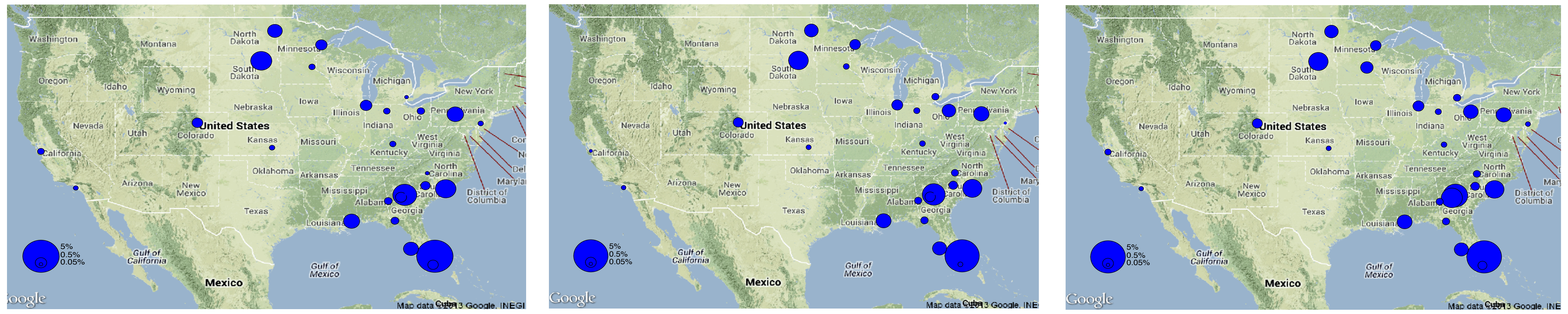
Amelia McNamara, University of California, Los Angeles

The problem:

The problem set forth by the 2013 Data Expo was to provide a graphical summary of important features of a given data set. The data provided were three years of survey data on "The Soul of the Community" collected by the Knight Foundation. The Knight Foundation surveyed members of 26 communities over the telephone, asking a variety of questions about community involvement.

Which communities were over-surveyed?

In most communities, approximately 400 people were interviewed, but certain communities were surveyed much more. It appears that the Knight Foundation was trying to survey places at an approximately similar rate, which is why Philadelphia (for example) was surveyed 1633 times in 2010. To see which places were over- or under-represented in the survey, I created maps showing the percentage of the community that was polled for each polling year.

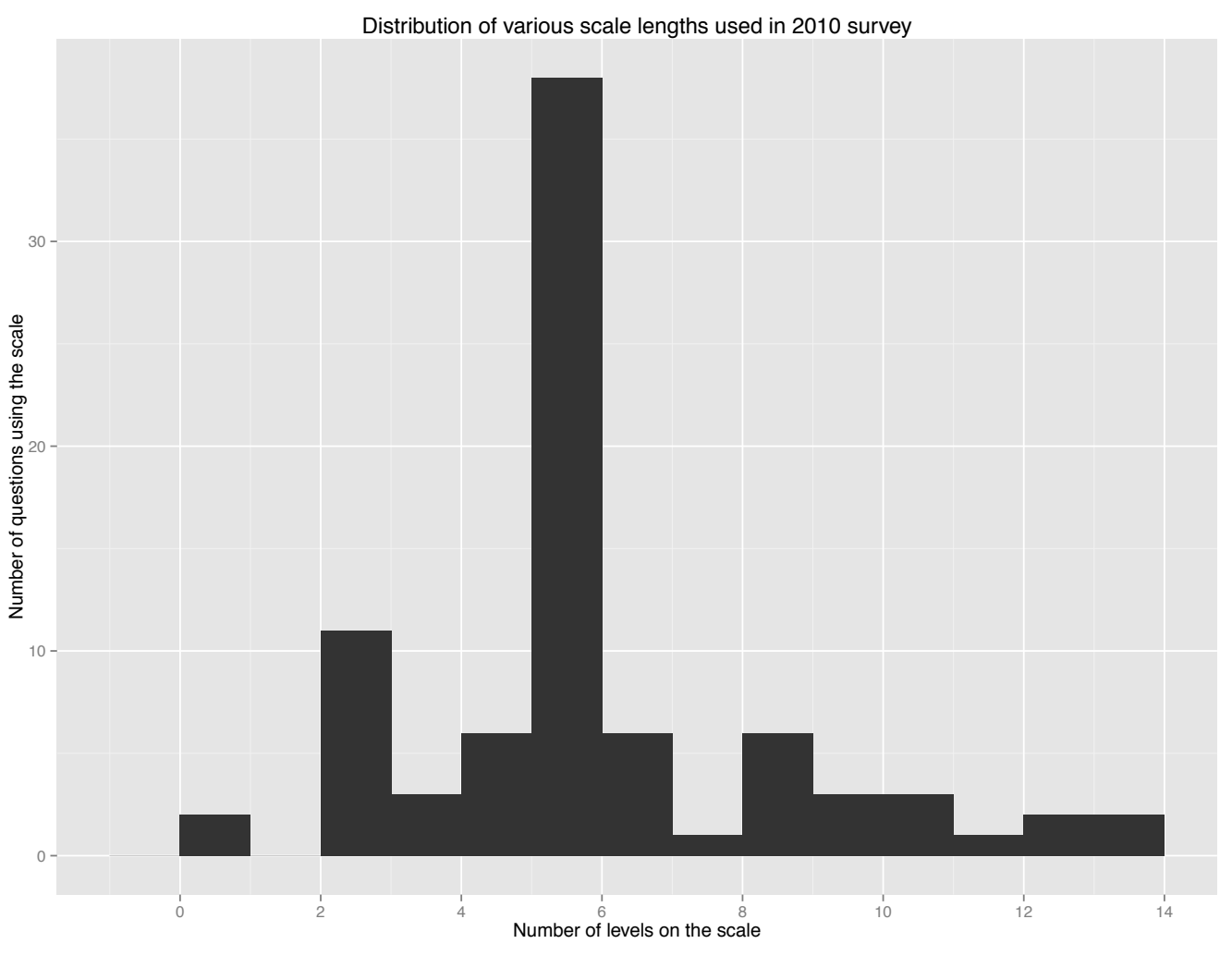


2008 2009 2010

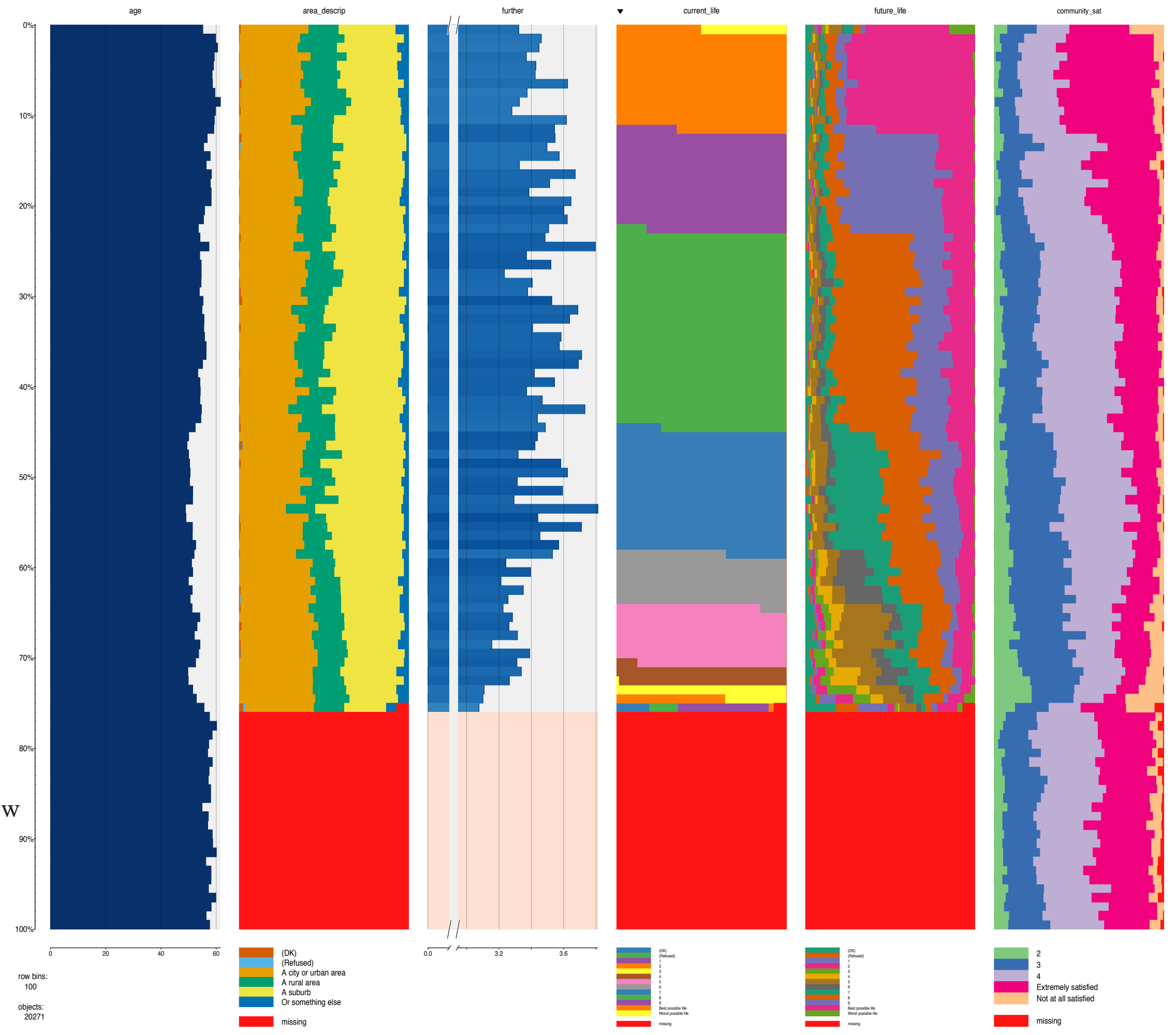
The population data for these maps came from the Intercensal population estimates compiled by the Census Bureau. Population estimates are calculated each year in between Census years. These maps show the percentage of the community that was polled, and percentages hover around a mean of 0.07%, with lots of variation. Palm Beach, FL always looks over-represented because the minimal sample size of 400 was always used, leading to a polling rate around 4%. Large communities like Philadelphia, PA, look underrepresented, with a rate around 0.01%. And there is some variation over time, especially on the East side of the US. For example, Akron, OH begins with a polling rate of 0.01%, which rises to 0.07% and then 0.09%, as a result of polling increasing from around 400 residents to more than 1700.

How many different scale lengths are represented in the survey?

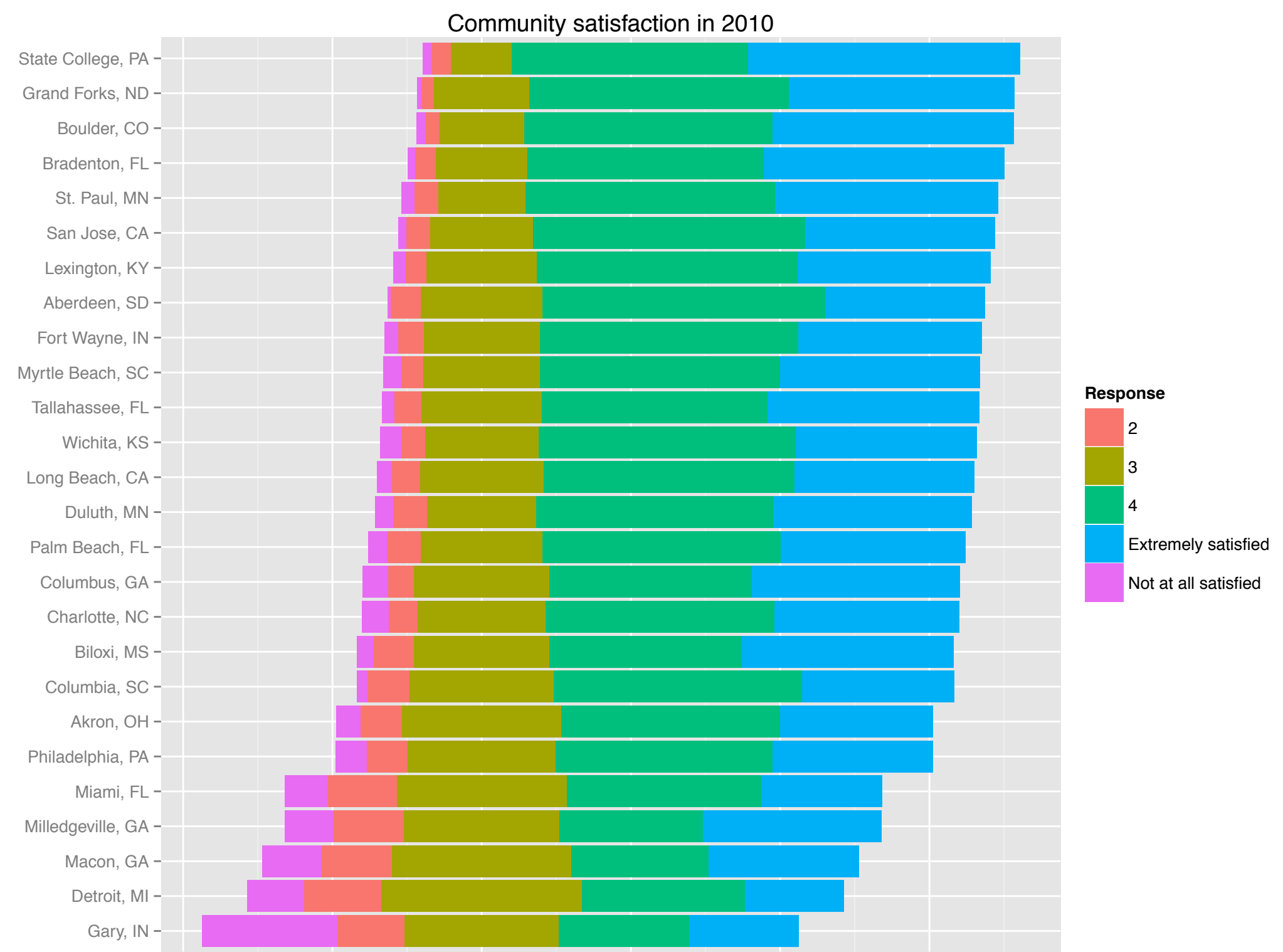
The data are comprised of many demographic variables identifying the respondent to the survey, as well as their responses to survey questions. Most survey questions were answered on a Likert scale, although there was little consistency in the number of levels for the scale. The most common scale was a five-point scale, as in "Not at all satisfied, 2, 3, 4, Extremely satisfied" or "Very bad, 2, 3, 4, Very good." However, many other scales were used. In fact, the second half of each data set is comprised of recoded variables, which have been normalized onto three-point scales with the levels high, medium, and low.



Do people answer all the questions?



While the 2008 and 2009 data sets are almost complete, the 2010 data set has about 25% missing data. This missing data is characterized by almost all the demographic information being present, but only one survey question answered (that being, "how satisfied are you with this community as a place to live"). Interestingly, with the 4880 incomplete responses removed, the 2010 dataset is reduced to 15,000 observations, which is much closer to the 14,000 observations the two prior years.



Which place reported the greatest level of community satisfaction?

Knowing that the question about community satisfaction was only survey question answered by all respondents, it made sense to see which communities reported the highest levels of community satisfaction. Because it does not make sense to compute a mean of a categorical variable, I researched visualizations of Likert scale data. One common way to display Likert scales is net stacked distribution graphs. These graphs show the percentage of positive and negative responses centered around a zero axis, to show the overall balance of responses.

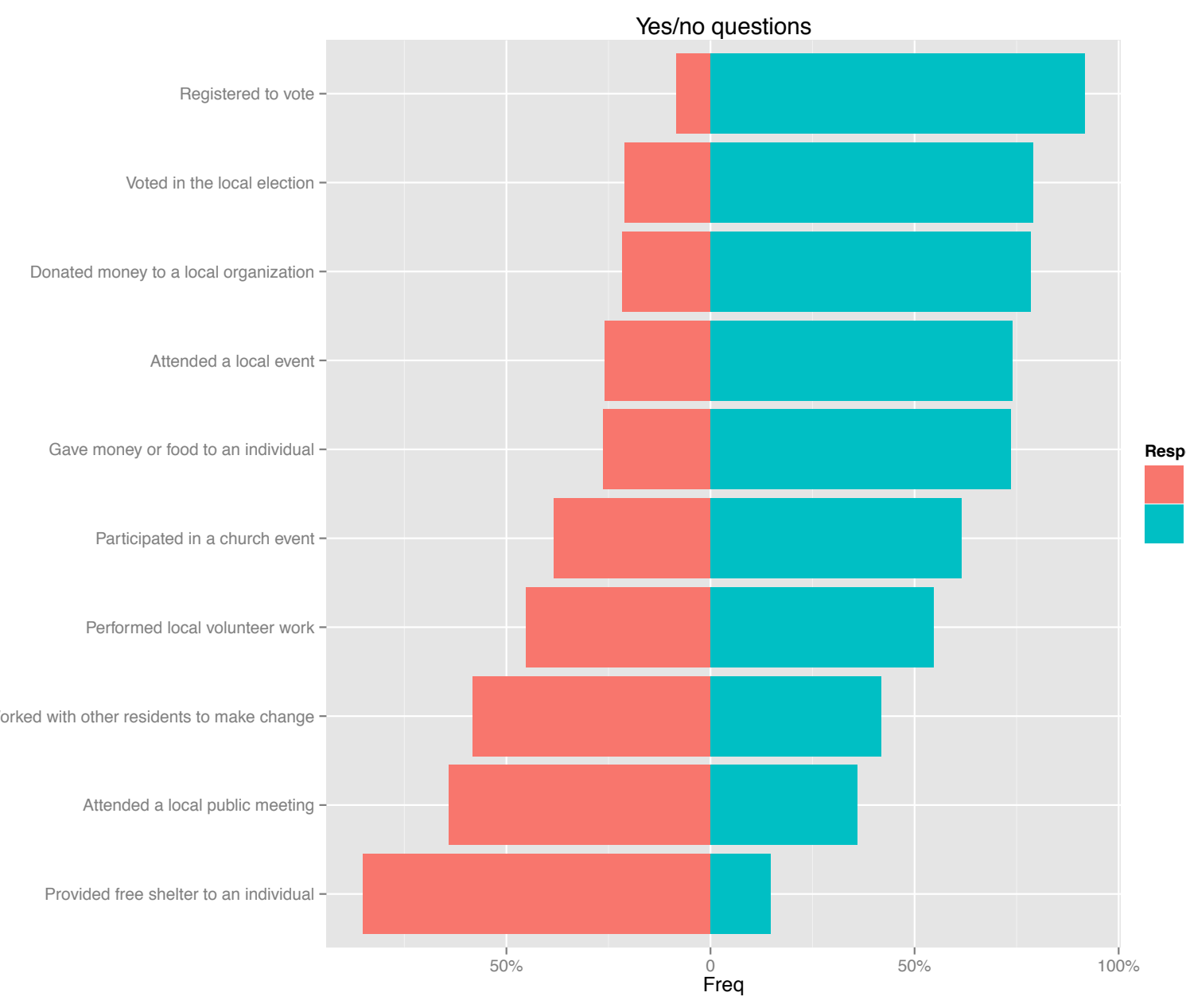
My code was cribbed heavily from a blog post I read on Likert scale visualizations. The main problem with these visualizations is that for scales with an odd number of levels, it makes sense to center the middle value around the zero axis. Because of this, you must split the data into two pieces and plot them separately, which means that legend ordering is lost.

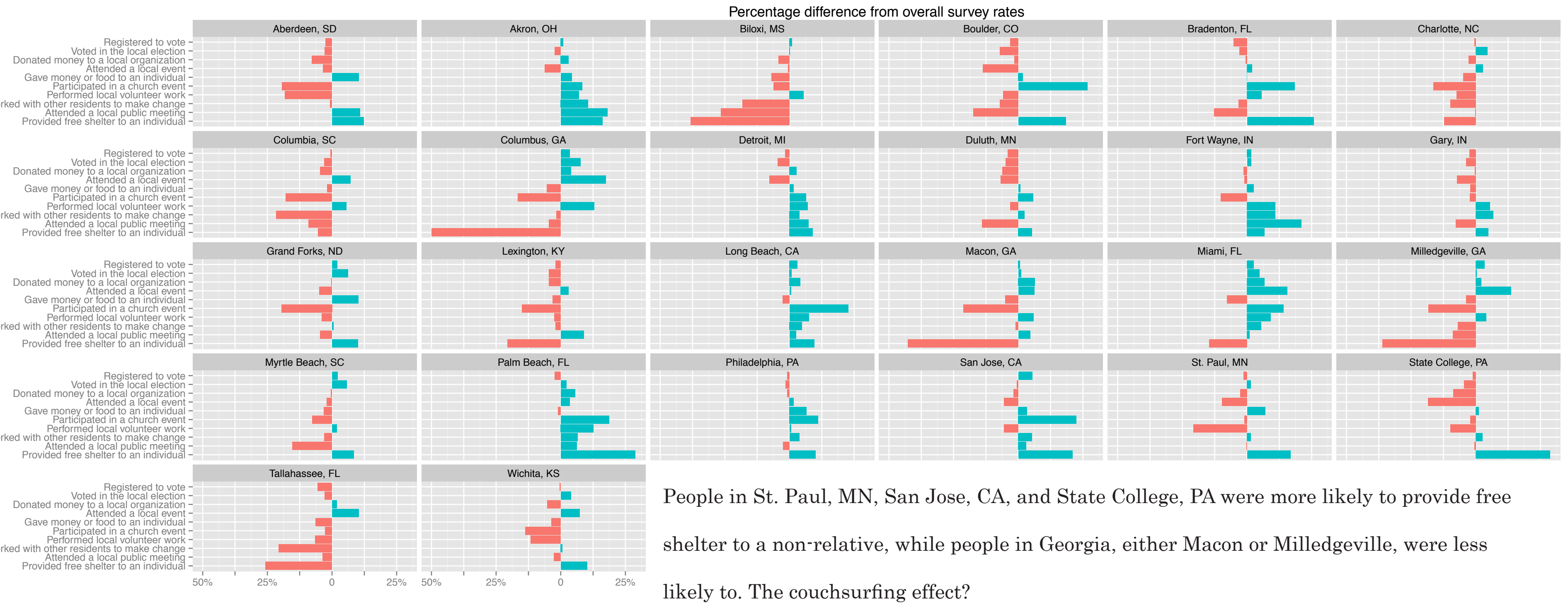
From this plot, we can see that people in State College, PA reported much greater levels of community satisfaction than people in Detroit, MI or Gary, IN. I noted with satisfaction that two of my favorite communities, St. Paul, MN and San Jose, CA, were rated quite similarly high.

What behaviors do survey-takers generally engage in?

In 2010, the most common behavior was being registered to vote, and the least common was providing free shelter to a non-relative. While the surveys in 2008 and 2009 did not contain the full set of yes/no questions, the faceted plot showing the responses over time did not show any clear trend.

Similarly, the faceted plot showing the responses from 2010 by community just looked like noise. The plot on the right shows the percent difference of the community rate from the overall rate.

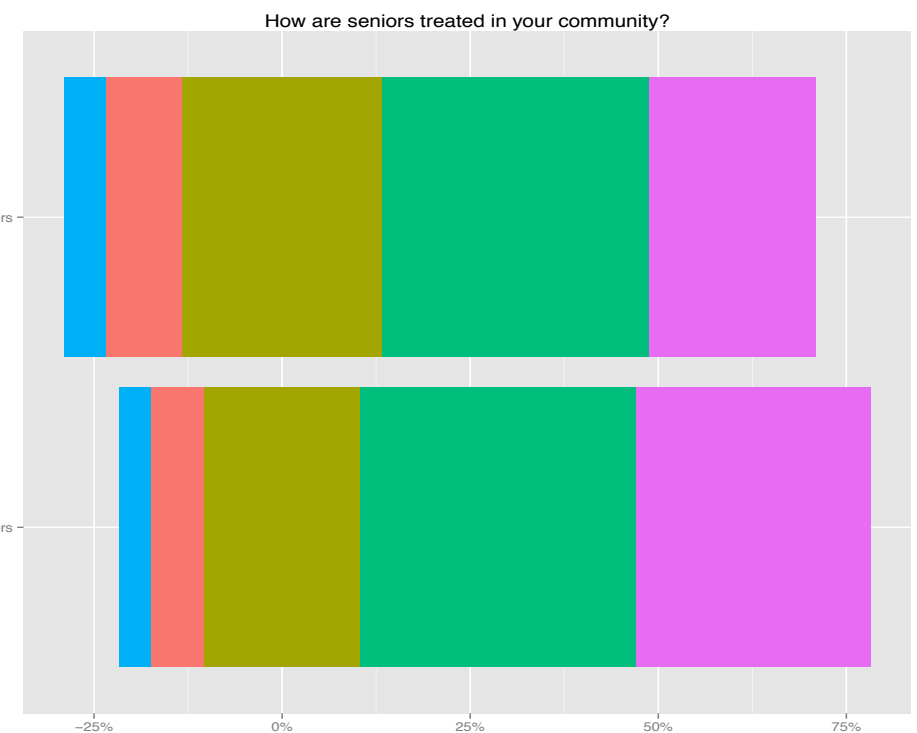




People in St. Paul, MN, San Jose, CA, and State College, PA were more likely to provide free shelter to a non-relative, while people in Georgia, either Macon or Milledgeville, were less likely to. The couchsurfing effect?

Do residents have meta-knowledge about their community?

That is, does the general population's opinion on the community as a place for seniors conform to seniors' own feelings about the community as a place to live? Do non-parents know how the community is for people with children? Do non-minorities know how the community is for racial and ethnic minorities?

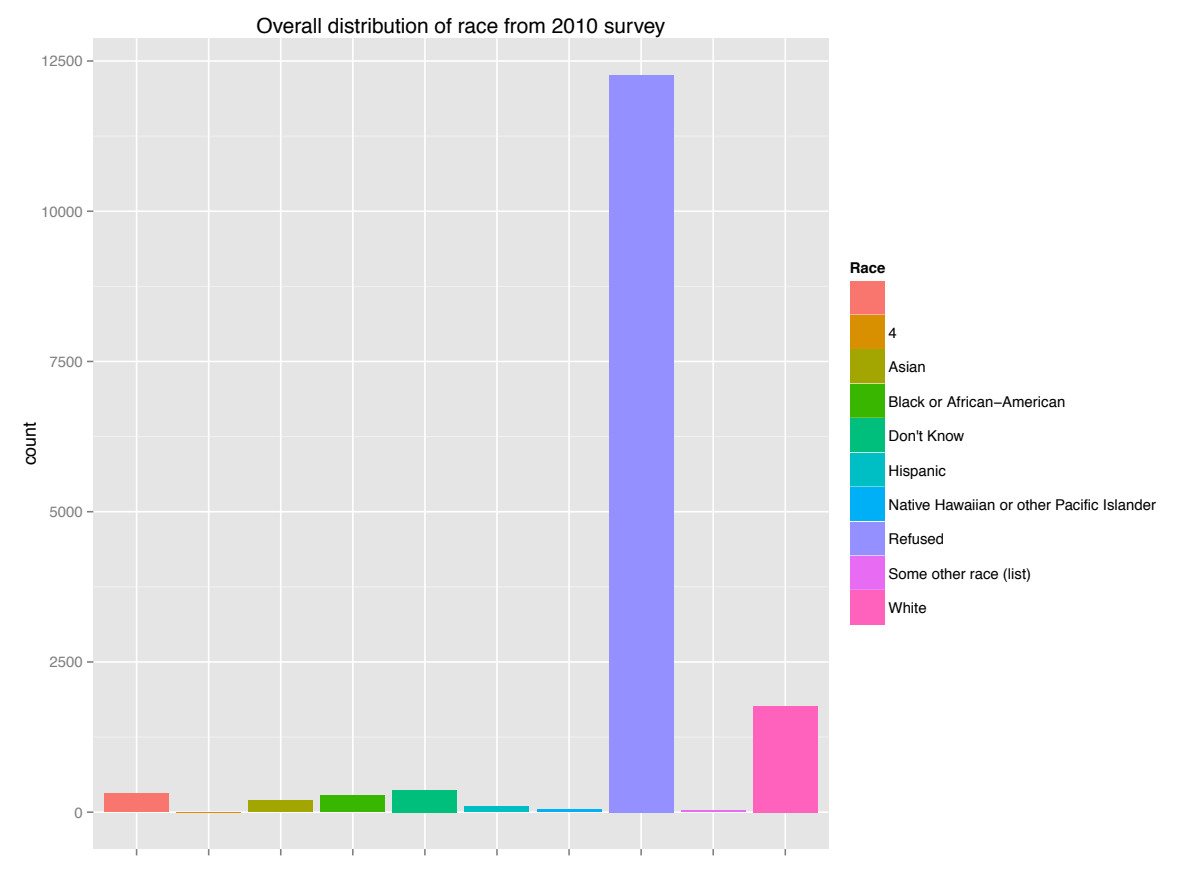


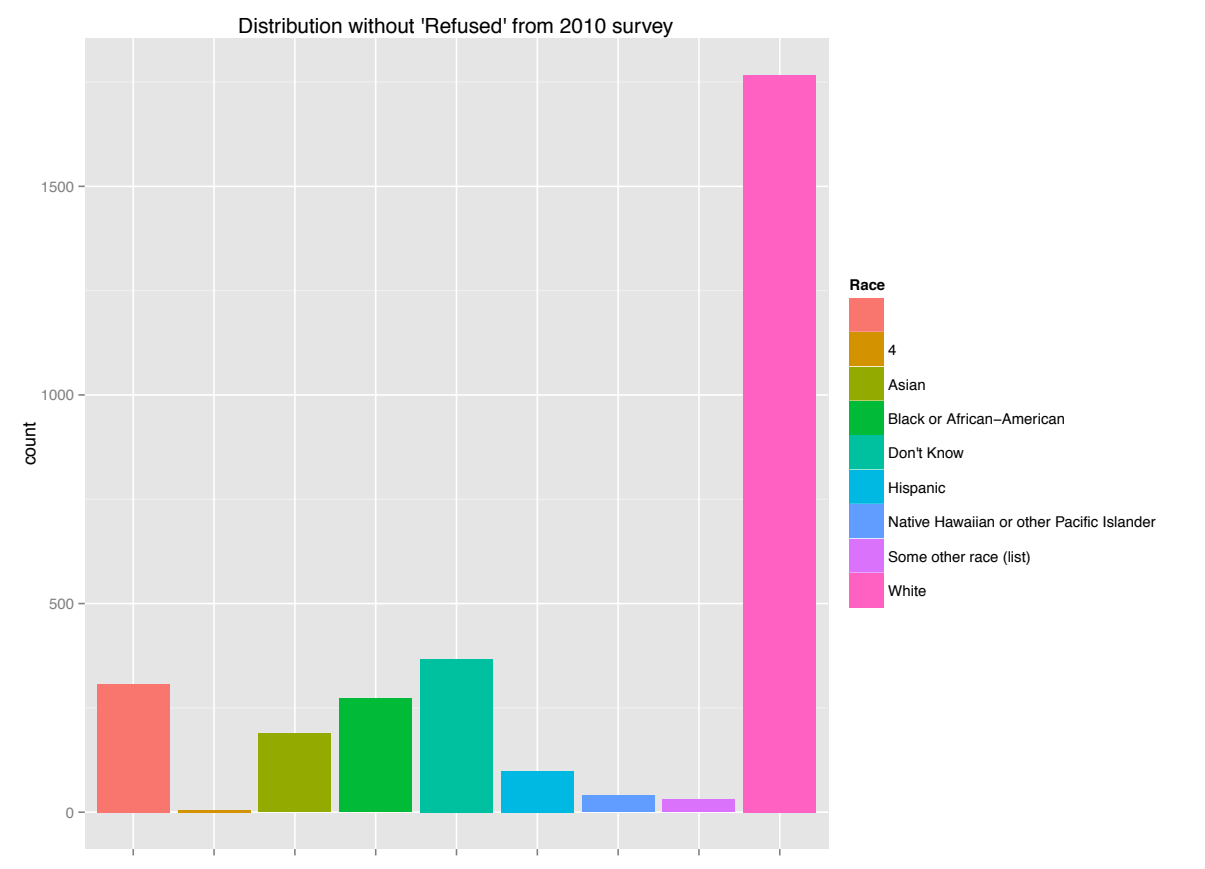
Is the community a good place for seniors?

It appears that non-seniors underestimate how good a place it is for seniors. People 55 and older rated their community as a better place for seniors than did people under 55. This was true across every community surveyed.

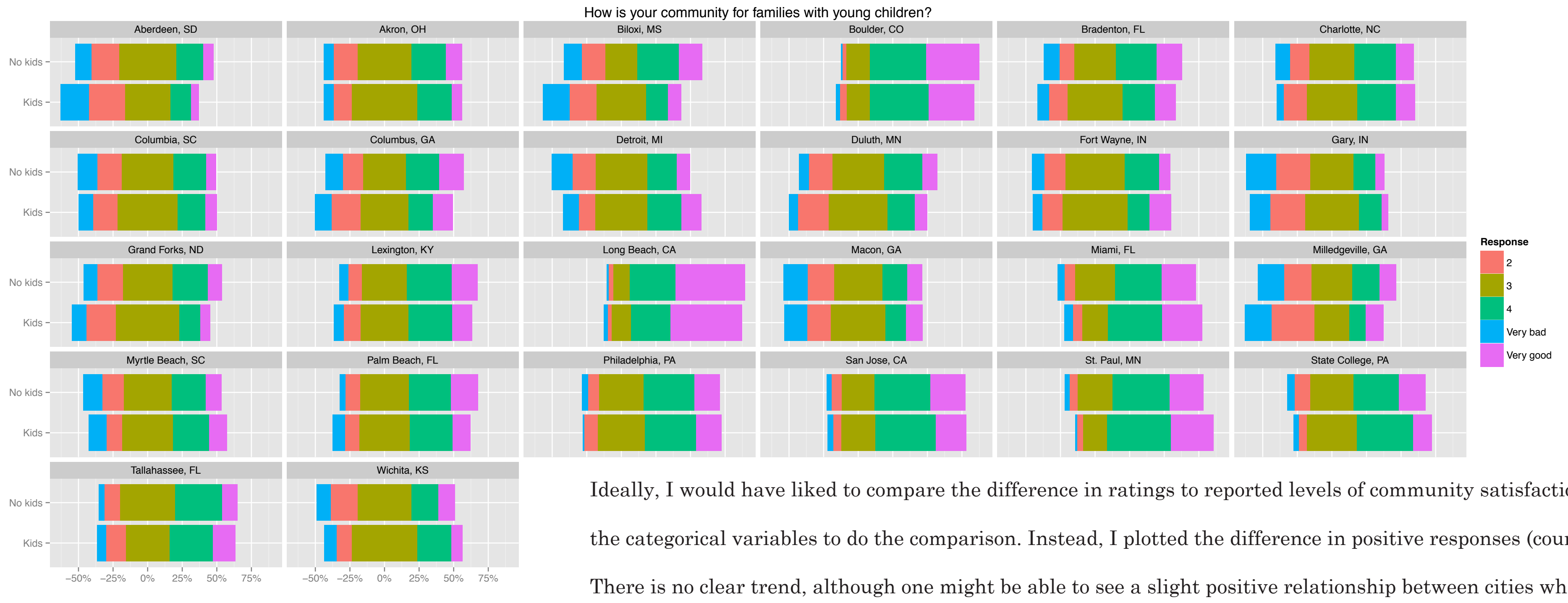
Is the community a good place for minorities?

For racial and ethnic minorities, the plots initially looked incredibly strange, which turned out to be because of the small sample sizes in each community. Looking back at the demographic breakdown for race, I noticed that the vast majority of respondents refused to answer. And those who did answer were predominantly white. Almost all the variation in plots could be explained by small sample sizes. This was disappointing, as I was interested to see the difference between the two groups.

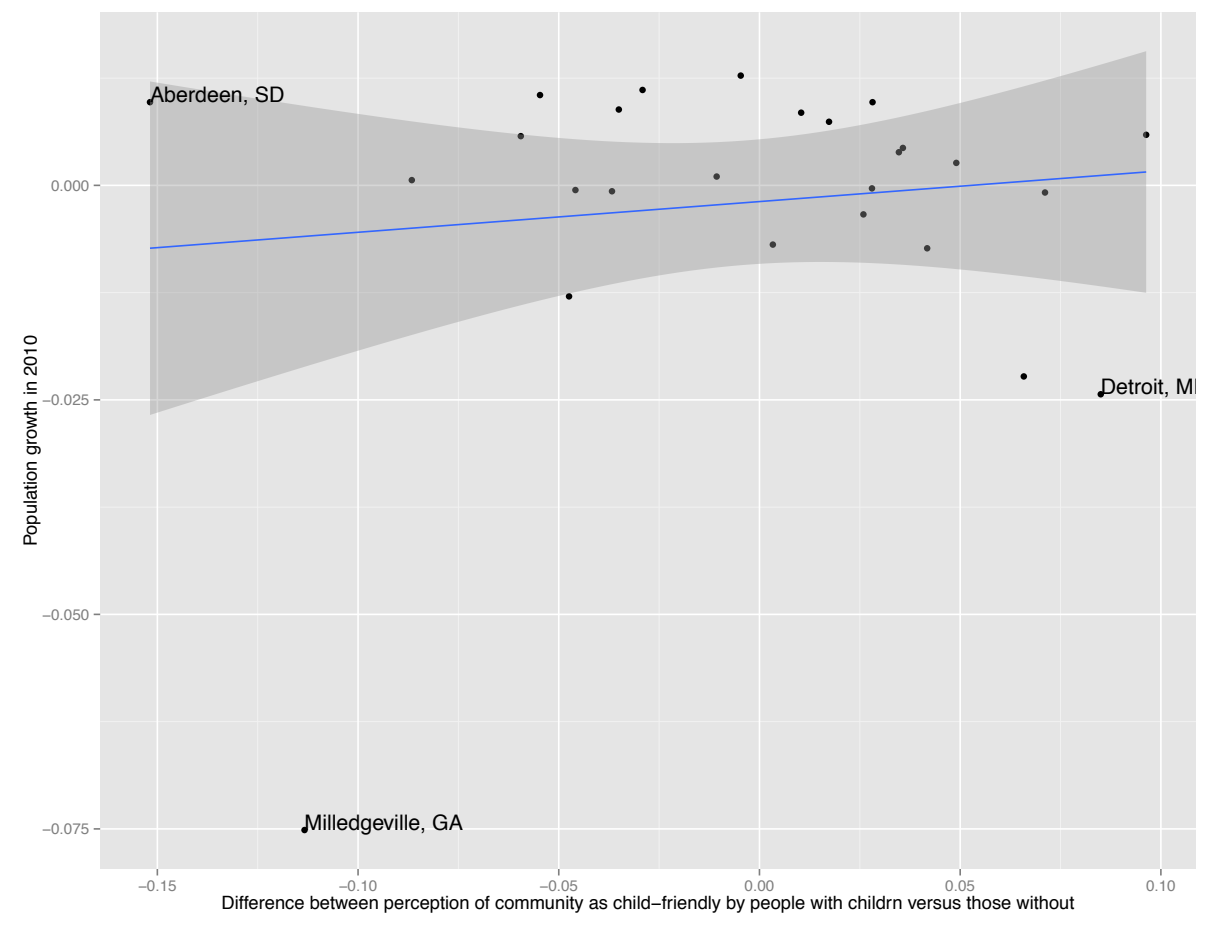




Is the community a good place for kids?



Luckily for me, there was some more interesting variation in the responses to the question about how good the community was for families with young children. Again, I struggled with small sample sizes, but by defining the group of those with children as anyone with a child under the age of 18 living on the premises, I was able to get around 100 respondents in each city, or about 25% of larger samples. Finally, there were communities where parents said it was worse for families with children than non-parents said, and vice versa. Interestingly, everyone seems to know that Long Beach, CA is good for children.



Ideally, I would have liked to compare the difference in ratings to reported levels of community satisfaction, but it was hard to know how to summarize the categorical variables to do the comparison. Instead, I plotted the difference in positive responses (counting neutral) against the rate of growth in 2010. There is no clear trend, although one might be able to see a slight positive relationship between cities where people with children believe it is better for kids and cities that saw growth in 2010. Certainly, Milledgeville, GA (where people without kids thought it was better for children than those with kids themselves) saw a steep drop in population.

Conclusions

My main conclusion is that it is incredibly difficult to deal with data that is largely categorical. It was hard to understand where numeric data came from (even the index variables weren't well-documented, and there was a whole category of measures like "Thriving" that had no explanation at all), and the sheer volume of differently-scaled categorical responses was daunting. Even converting all the questions into ordered factors seemed out of the question, because the initial ordering did not have enough consistency for a programmatic approach.

It was interesting to see which communities reported the highest levels of satisfaction with their community, and try to make visual correlations with other variables. I also enjoyed seeing how many people reported being registered to vote in the survey overall, and seeing the consistency of that percentage over the three years of the survey and the 26 communities surveyed.

References and resources

R packages: ggplot2, plyr, reshape, scales, tabplot, MobilizeSimple (github)

Population data: http://www.census.gov/popest/data/intercensal/cities/cities2010.html

Likert scale plots: Ethan Brown, http://statisfactions.com/2012/improved-net-stacked-distribution-graphs-via-ggplot2-trickery/

My code is available at: github.com/AmeliaMN/SoCDataViz

Thank you

To Terri Johnson and Mitch Halverson, who listened to me talk about this data for months