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. 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.

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. The estimates don't vary too much from the Census count, but I decided to use the yearly estimates for the sake of having slightly different populations from year to year. For example, the 2010 Census count of the population of Palm Beach, FL was 8,348. In 2009, the estimate was 8,456, and in 2008 it was 8,631.

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 under-represented, 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. It's not clear why decisions like these were made.

The data are comprised of many demographic variables identifying the respondant 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.

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 observatsions, which is much closer to the 14,000 observations the two prior years.

Because of this, it appears that this question was the crucial one.

In the demographic section dealing with race, the possible choices from the data set were: [1] " " "4" [3] "Asian" "Black or African-American" [5] "Don't Know" "Hispanic" [7] "Native Hawaiian or other Pacific Islander" "Refused" [9] "Some other race (list)" "White"

"4" doesn't even show up in the codebook.