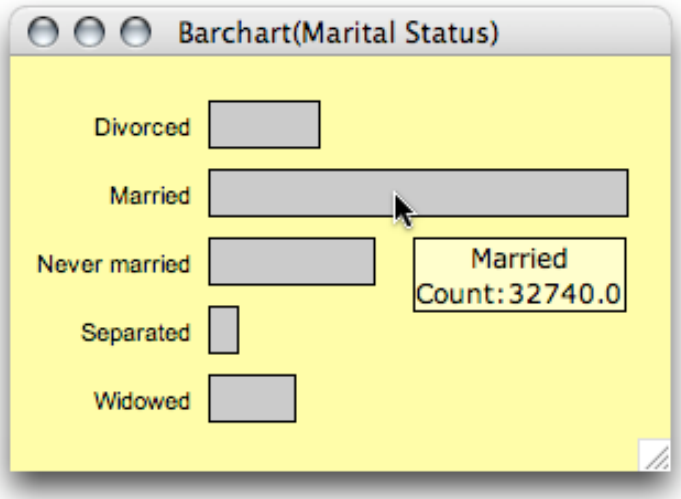
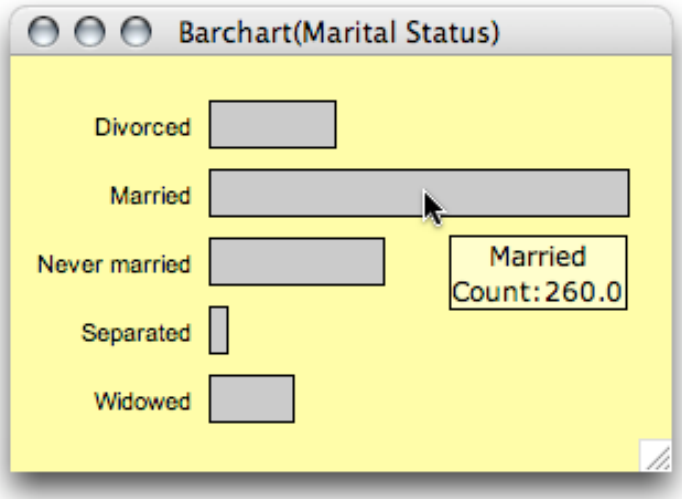
Large Data: Area Based Plots

- Area based plots are almost unaffected by the size of a dataset
- Only when we look at variables which get more and more categories as the data set size grows we will face the problem of managing very many categories/tiles.
- In this case logical zooming/conditioning and sorting mechanisms are indispensable

based on 63,756 observations

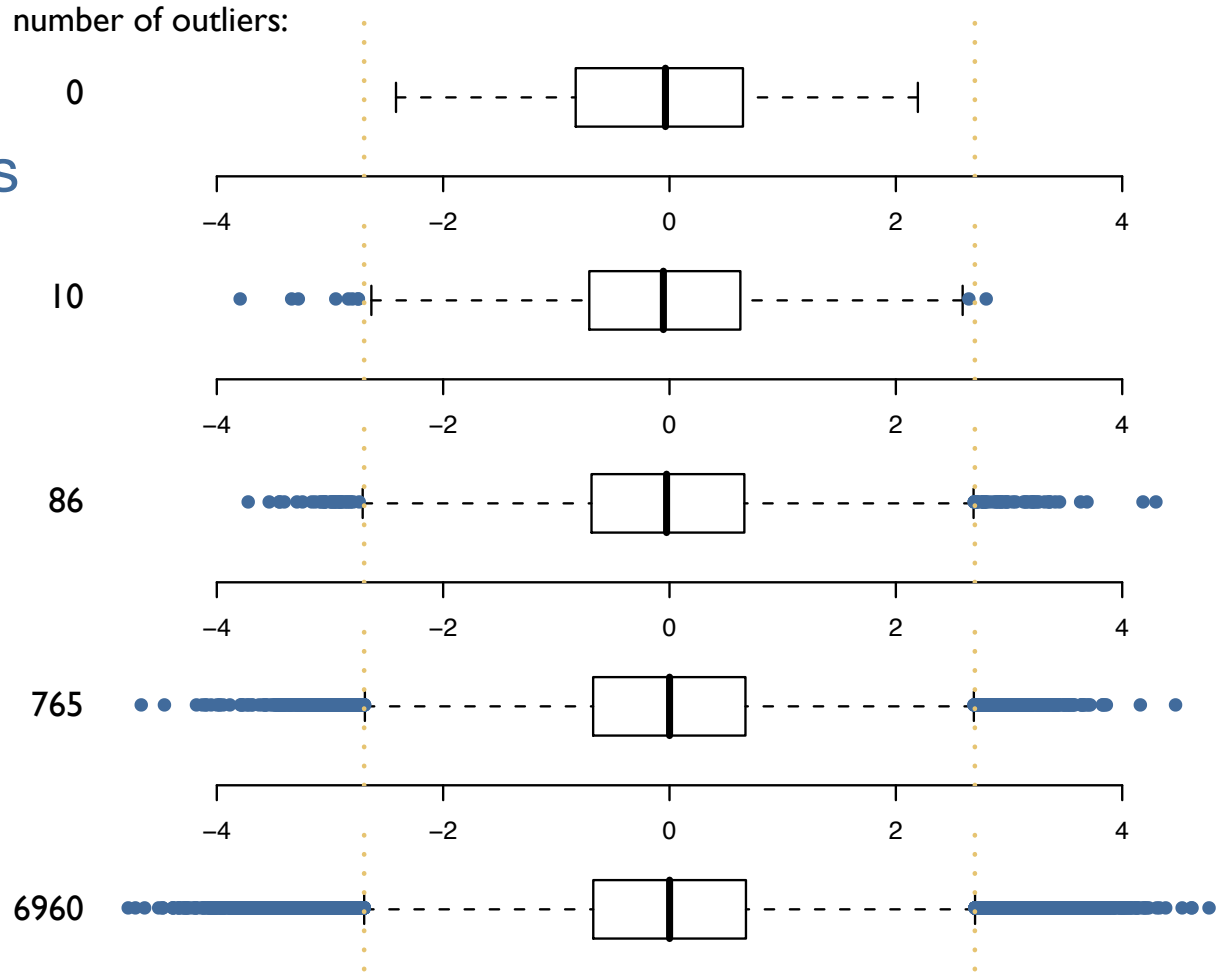


based on 510 observations



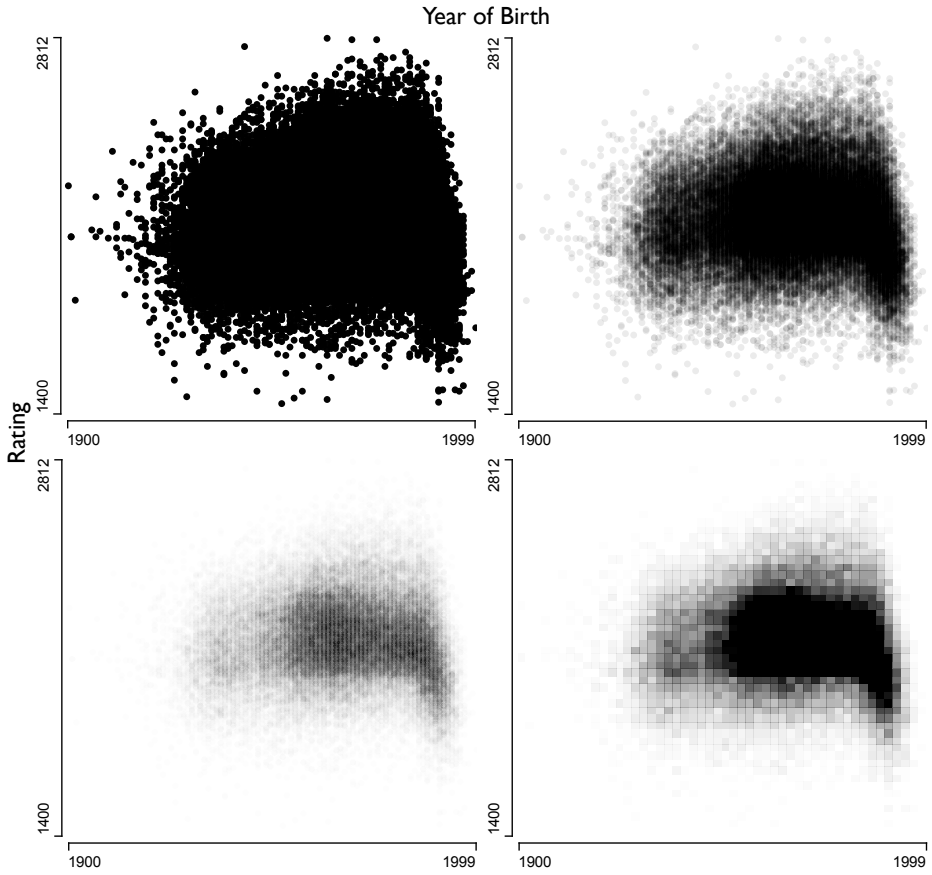
Glyph Based Plots: Boxplots

- Boxplots unfortunately suffer the problem that according to their definition the number of outliers grows linearly with the number of points displayed in the plot
- Although the statistics shown in the plot are robust and scale up without any problems we end up with plotting very many outliers which can not be interpreted sensibly



Glyph Based Plots: Scatterplots

- As already explained in section three, the variation of α-transparency and point size can help to visualize the density structure
- For large datasets these variations are indispensable for getting any information out of a scatterplot
- A far more efficient, though less precise, solution is the use of binning
- A binned scatterplot is nothing more than a 2-dim. histogram where cell counts are mapped to shades of gray



Glyph Based Plots: Parallel Coordinate Plots

- Parallel Coordinates suffer far more from overplotting than scatterplot as they plot a whole line per value and variable
- None of the parallel coordinate plots you saw in the course so far was plotted without alpha blending
- Unfortunately, for really large datasets an 8-bit (=256 levels) α-transparency vector is not sufficient