

Largely about Largeness: Graphics of Large Datasets

Antony Unwin¹

University of Augsburg, e-mail: unwin@uni-augsburg.de

Abstract: Potentially there is a great deal of structure in high dimensional data and neither statistical models nor graphical displays can hope to do more than identify some of the features. Quite often there will be insufficient data to draw formal conclusions, even though it is usually possible to draw insightful pictures. If there is a lot of data, then they are likely to be heterogeneous, and better investigated with a collection of local analyses than with a single all-embracing global analysis. In this overview talk we look at strategies for visualizing high dimensional datasets and describe some graphical tools for this kind of problem, both variations of well-known tools adapted for the purpose and some newer graphical ideas.

Keywords: Large datasets, Visualization

1. Visualizing large datasets

Nowadays it is possible to work with datasets of sizes we could only have dreamt about a few years ago. For instance, this year's dataset visualization contest for the US Joint Statistical Meetings Conference involved a dataset with some one hundred and twenty million cases. More importantly, it involved not only the basic dataset of about thirty variables per case, but also the possibility of incorporating additional information from other sources. This is typical of modern large datasets: largeness implies not just more cases and more variables, but also more complex structures (Unwin *et al.*, 2006).

Just having more cases to deal with is not a major issue. Some existing graphics can readily be directly applied (generally area-based graphics like histograms and barcharts). Figure 1 shows an early example from the 1890 US Census. Others have to be adapted (generally, case-based graphics like scatterplots or parallel coordinate plots). Having many variables raises more serious problems. Even displaying univariate graphics for all variables can become awkward when there as few as twenty-five, as Figure 2 demonstrates. Multivariate displays are also limited. Mosaicplots can handle only a few categorical variables (sometimes as many as ten binary variables, but that is still not a lot for a large survey). Parallel coordinate plots can be drawn for over a hundred continuous variables, but as with the mosaicplots they are only useful if an insightful picture emerges. Parallel coordinate plots of just two variables can be unhelpful if there is no structure to be seen. What this means is that a different approach is needed.

¹Address of correspondence: Dept. of Computer Oriented Statistics and Data Analysis, Augsburg University, 86135, Augsburg, Germany

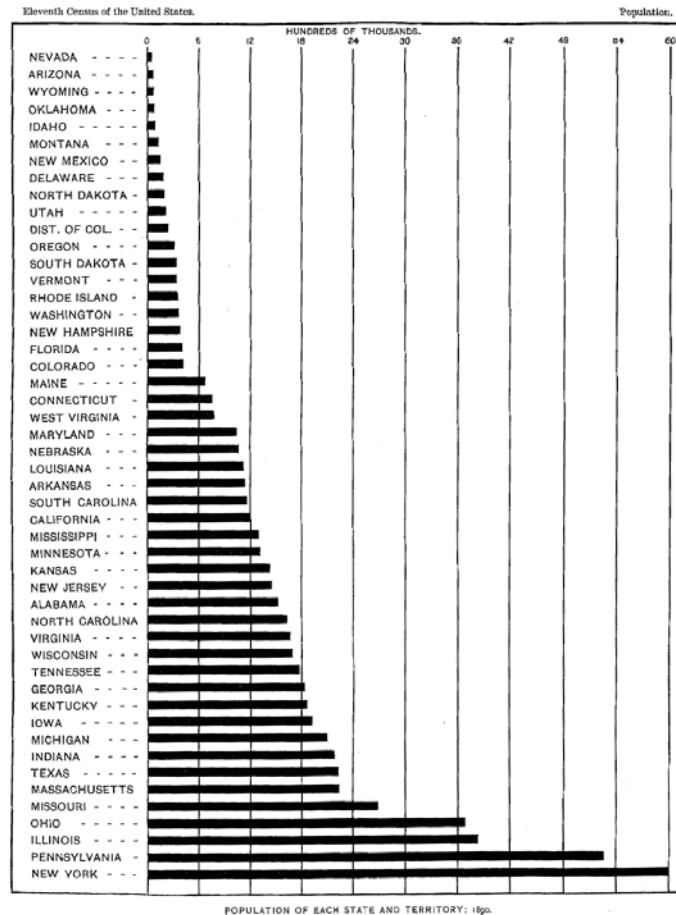
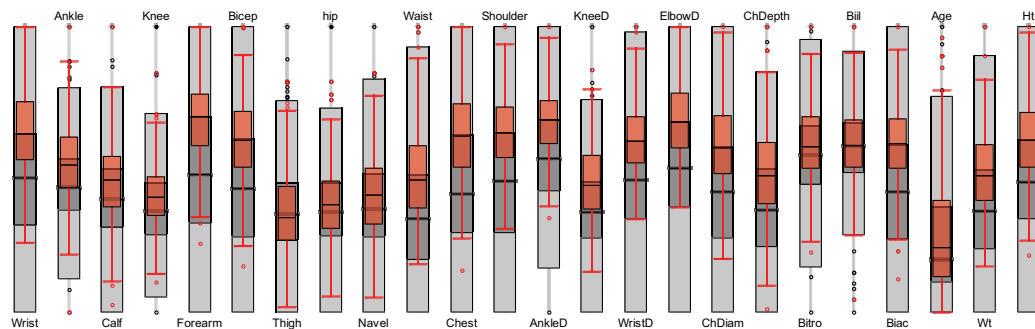


Figure 1: An early graphic of a big dataset: the US Population in 1890 from the official census report.

Figure 2: Boxplots of 24 body measurements on 507 people, men selected

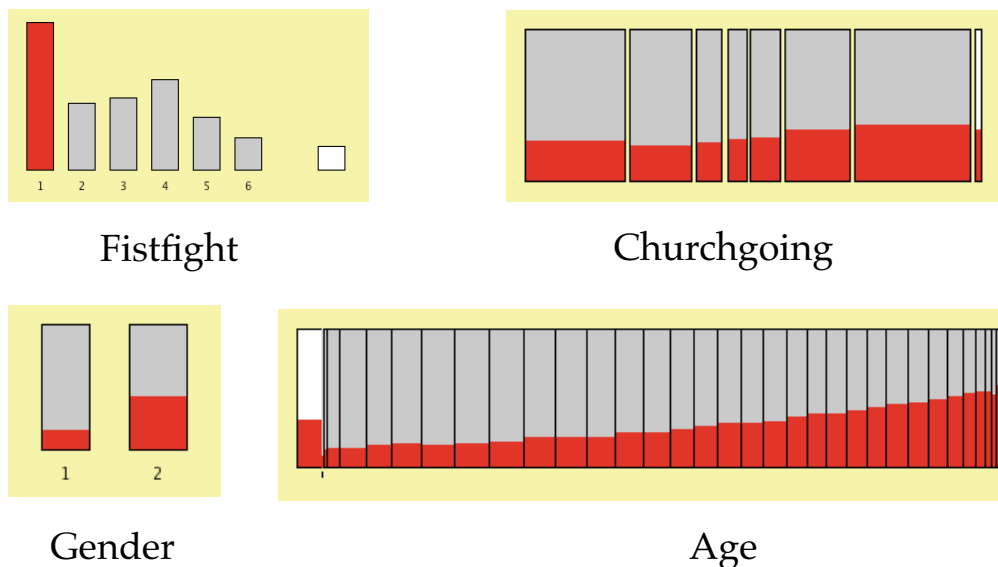


2. Strategies

In practice, there is little to be gained from displaying all variables of a large dataset simultaneously or in the same plot. There are usually groups of variables which should be considered together and particular hypotheses that are to be investigated involving

them. Whereas analysing a small dataset can lead to some interesting results, even if little information is available on the background to the data, this is extremely unlikely for large datasets. There may be a lot more information to be found, but only someone with the necessary background information can say what is worth having. A typical example is the DDB Life Style Survey, which was discussed in the book *Bowling Alone* (Putnam, 2000). There are a few hundred pieces of information recorded on each of some 85,000 individuals: demographics, social habits, political opinions, and so on. For example, consider Figure 3. People who think they would not do better than average in a fistfight are more likely to be female, more likely to be older and go to church a little more often. Other variables could be considered (e.g., marital status, number of small children, income) and the associations between the potentially explanatory variables (e.g., there are more older women and older people go to church more often).

Figure 3: *People who definitely disagree they would do better than average in a fistfight have been selected.*



Analyses concentrating on subsets of variables concerning particular topics carried out by domain experts are likely to produce interesting and worthwhile results. Graphics help considerably both in eliciting the information and presenting the findings. Attempts to analyse all the variables simultaneously may uncover some potentially interesting relationships, which might usefully be checked, but are more likely to produce far too many irrelevant results, which obscure anything of value.

3. Conclusions

There has been quite a lot of work on dimension reduction methods for multivariate datasets, leading to graphics of low-dimensional projections of the data. This approach is powerful when the variables are of similar type and the data are homogeneous. Here we have concentrated on analysing the raw data in a dataset without transformations, primar-

ily using interactive tools to highlight structure.

The paper is illustrated with examples of visualizing large datasets. A few individual graphics can only partially convey the application of graphical methods in practice. They show some final results, but little of the process by which the results were reached. An actual analysis may involve many, many graphics, drawn, discarded, redrawn, revised, and amended.

References

Putnam R.D. (2000) *Bowling Alone*, Touchstone, New York.

Unwin A.R., Theus M. and Hofmann H. (2006) *Graphics of Large Datasets*, Springer.