

Exploratory Data Analysis with Data Desk

Martin Theus

University of Augsburg, 86135 Augsburg, Germany
e-mail: Martin.Theus@Math.Uni-Augsburg.DE

Summary

Many statistical software packages claim to support exploratory data analysis (EDA). But most of them do not meet the demands of EDA, because their general concept contradicts the workflow of EDA. DataDesk has been developed in the past ten years, and was always based on an interactive, graphical interface.

This paper presents the basic concepts of EDA and shows how DataDesk is able to support the ideas of EDA.

Keywords: Exploratory Data Analysis, Interactive Graphics, Interactive Modelling, Extendability

1 The Philosophy of EDA

In 1962 John W. Tukey stated, that statistics was ignoring “real life problems”. This claim can have two consequences. The first one would be to argue, that statistics is not a “real life science”. But since this argument is not even true for mathematics — our modern life is dominated by techniques made possible by engineers using the integral and differential calculus explored in the 19th century — it can hardly be true for statistics. The other consequence of Tukey’s claim would be to invent a new discipline, besides statistics, dealing more with “real life problems”. One might think, that data analysis is the answer to this idea. But names alone do not generate new disciplines. I am not sure whether there is any one university, where

students can study data analysis as a subject today. And being a statistician is still a strange thing in Germany, because statistics is mostly hosted in mathematics departments and taught by mathematicians, who have never been confronted with data analysis problems.

What was Tukey's reaction on his claim. He still felt as a statistician, but started to look for new techniques, which he gathered under the name "Exploratory Data Analysis". To get an idea, what is hidden by this name, one can look at his book called "EDA" (Tukey 1977). Here are some of the most important headlines:

1. SCRATCHING DOWN NUMBERS (stem-and-leaf)
2. SCHEMATIC SUMMARIES (pictures and numbers)
3. EASY RE-EXPRESSION
4. EFFECTIVE COMPARISON (including well-chosen expression)
- ⋮
19. SHAPES OF DISTRIBUTION
20. MATHEMATICAL DISTRIBUTIONS
21. POSTSCRIPT
21. A Our relationship to the computer

Tukey formulated his ideas in times where computers were expensive, non-interactive and non-graphical. This makes him in some sense a visionary, because he could imagine tools and techniques, which were not possible to implement at that time.

But besides the visionaries of a new epoch, there must be a new technology, supporting the new visions of the new era. For EDA, computer science, as for many other fields, was the key technology to get the ball rolling. Although computers have been around in different forms for more than 40 years now, the invention of the concept of a graphical desktop was the most important step for applying EDA.

But even new visions and new technologies do not imply, that things change. Projects have to start, that incorporate the new ideas, using new technologies. DataDesk, initiated by Paul Velleman is an entirely new development to meet the requirements of EDA. In contrast to most other statistics packages, DataDesk never used a teletype text-interface for user interaction, but was always based on the Macintosh desktop concept.

The rest of this chapter shows how DataDesk meets the four main threads of Exploratory Data Analysis, which are:

- Visual exploration,

- Extensive user interaction,
- Re-expressing,
- Meta data support.

1.1 Visual exploration

As mentioned above, DataDesk uses the desktop concept to handle data, or more general objects. Figure 1 shows a sample desktop of a DataDesk session.

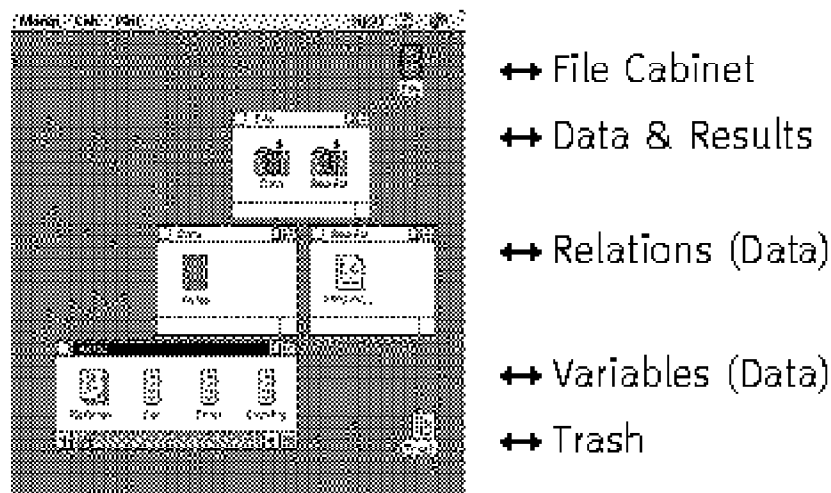


Figure 1: A sample desktop in DataDesk

Every object, which can be a whole relation of data, a graph, a comment, a formula etc. is represented by its icon. These icons can be moved to nearly any place on the desktop, only restricted by the relational constraints of the data. This is especially important when grouping data or other objects hierarchically in folders. Any result of an analysis is stored automatically in a results folder, and remains part of the session, until it is deleted explicitly.

Data are stored in relations. This enables the analyst to load as many data sets as useful. DataDesk also offers special functions, to work with relations. In contrast to other packages, which use a numerical data matrix as the standard data structure, data in DataDesk can be numbers, text, and even infinite values (" ∞ ") or missing values (" \bullet ").

The visual exploration of data is mainly based on graphical representations of the data. DataDesk supports several interactive graphs:

- Dotplots, Boxplots

- Scatterplots, Rotating Plot
- Barcharts, Piecharts
- Histograms
- (multiple) Lineplots

But besides graphs, simple interactive tables such as

- Frequency breakdowns
- Contingency tables

can be found, too.

1.2 Extensive user interaction

User interactions can be subdivided into two parts. The first part, the desktop concept, is described above. The second part is the direct interaction with graphs and tables. Nearly all graphs and tables inside DataDesk are linked, i.e. selections in one plot or graph initiate highlightings in others. DataDesk supports three stages of linking

- **Cold Linking**
Changes in one output do not affect others
(Default for most other packages!)
- **Warm Linking**
Changes can be propagated on demand
(DataDesk default)
- **Hot Linking**
Changes are automatically propagated to all other outputs

It is not clear which stage should be the default, but the hot link mode seems to be the best choice, since it is worse to rely on an output, that is not up to date, than to lose intermediate results by an automatic update.

The three pillars of the paradigm of linked highlighting are:

1. Selection

DataDesk offers several tools and modes to select data. Figure 2 shows the tools and modes pallet inside DataDesk. In principle a single static tool (pointer, dragbox, lasso or slicer) and the dynamic brush should enable the user to perform any selection. But if a selection is too complicated with a rudimentary tool, an analyst will leave out this step in the analysis — thus it is obvious, that we need a lot of different tools, to enable an easy and smooth analysis process.

The same holds true for the selection modes. Although every mode could be derived from “and” and “not”, it is much more efficient to use the other modes directly.

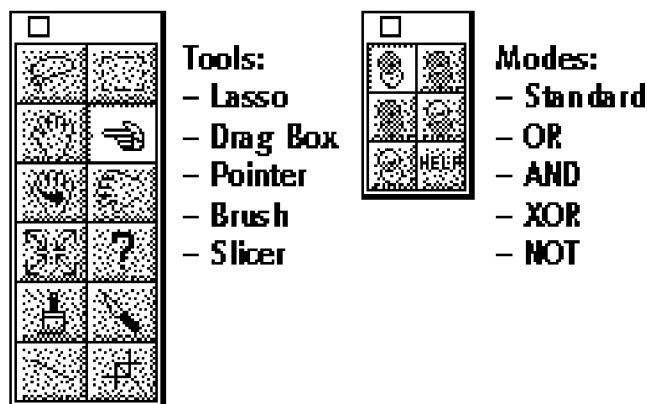


Figure 2: The selection and mode toolbox in DataDesk

2. Highlighting

Highlighting seems to be defined very clearly. This is true for any kind of scatterplot. Difficulties only arise for areal plots, and “special plots” like boxplots. DataDesk supports highlighting for areal plot, i.e. for barcharts. Highlighting in boxplots is only supported partly. Another

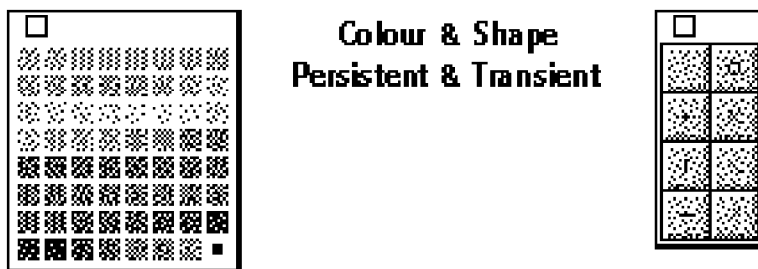


Figure 3: The colour and marker palette in DataDesk

problem arises with the assignment of colour for different groups. Figure 3 shows the two palettes for the assignment of colours and markers. As long as the assignment is disjunctive, one does not have to deal with the colouring of intersections. Since the number of possible intersections grows dramatically to $2^k - k - 1$ with the number k of colours, it is hard to find suitable colour-assignments for the intersections. This is mainly due to the fact that all colour-models (eg. RGB, HSL, CMYK etc.) are of fixed dimensionality. Thus DataDesk only offers a disjunctive assignment of colours.

A much tighter limitation is the fact that areal plots do not offer any colouring, although groups in DataDesk are always disjunct.

3. Interrogation

Geographers tend to put as much information into a graphics at a time as possible. In contrast to this approach, Tufte (1983) invented the *Data-ink-ratio*. Tufte demands a maximisation of the data-ink-ratio in order to increase the readability of plots. In this context interrogation become very important. Figure 4 shows a sample interrogation.

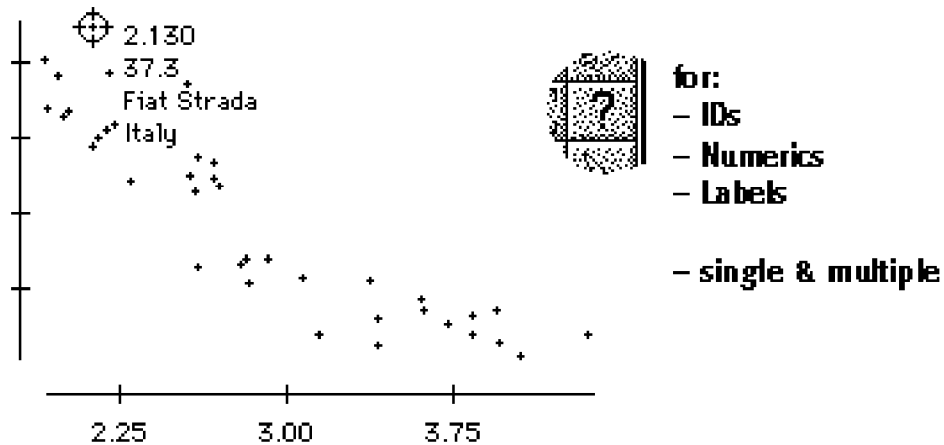


Figure 4: Sample interrogation of text and numbers

DataDesk offers the interrogation of text and numbers for all selected variables in all point plots. Although this is much more than other packages can offer, a generalisation of interrogations would be desirable. For more details on context sensitive interrogation — eg. the interrogation of bars in barchart or the interrogation of boxplots — see Theus (1996).

In general most tools, modes, colours and markers can be used more or less orthogonally, which is very important for efficient and smooth analyses.

1.3 Re-expressing

DataDesk supports the re-expressing of variables in two ways:

1. Derived Variables

Derived variables remain hot linked to their source variables. They can be derived by using any algebraic term including if/then/else-statements. Any change of the source data is reflected by a change of the derived data. The same three stages of linking apply for derived variables.

2. Sliders

Sliders provide dynamic transformations linked to any dependent output. Since sliders can be seen as parameters, which can be used to form new derived variables, a dependent output can be a summary, as well as a plot or a linear model. Figure 5 shows a slider, controlling the

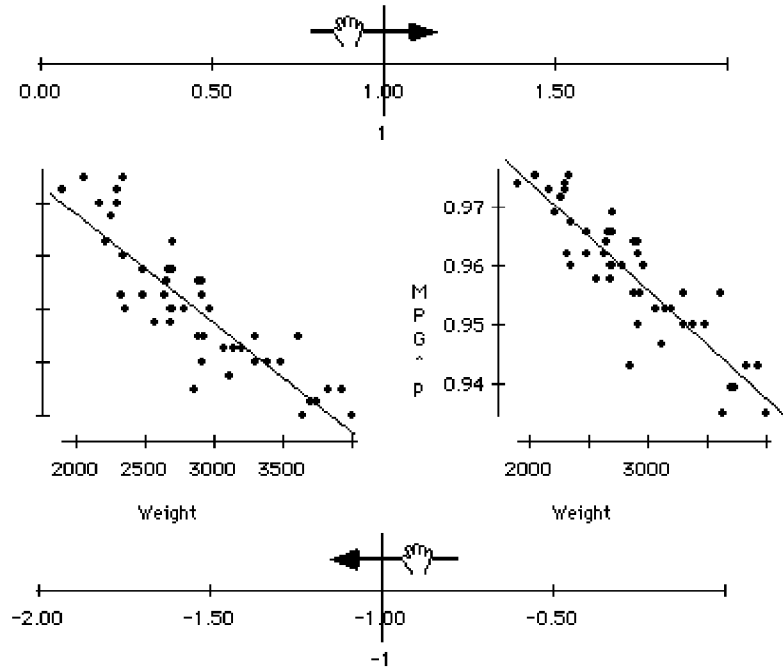


Figure 5: Sample Box-Cox-Transformation of MPG (Miles per Gallon)

parameter λ of a Box-Cox-Transformation, defined as follows

$$x_{BC}^{\lambda} := \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{for } x \neq 0 \\ \ln(x) & \text{for } x = 0 \end{cases}.$$

The corresponding derived variable is given by:

```

IF ('MPGp' <> 0) THEN
  ((exp(ln('MPG'))*'MPGp')-1) / 'MPGp'
ELSE
  ln('MPG')

```

(where “MPG” is the transformed variable and “MPGp” is the transformation parameter) and can be generated automatically by a single DataDesk command.

Again the advantage of the interaction, arises in the instantaneous update of the outputs.

1.4 Meta Data

The term “Meta Data” has been around for some time. One person may say that the use of meta data is just like using your common sense, others try to formalize, how to handle all the information, that is available besides the pure data matrix.

Many statistical packages just use a data matrix as a representation of a dataset. Adding the names of the variables is already some sort of basic meta information. DataDesk can hold different data sets, as well as any textual or graphical information, which is related to the data set in any way.

Imagine Fishers famous Iris data. Everyone of us has used it, mostly without having the faintest idea, what the species look like, or where the measurements have been taken exactly. It is really easy in DataDesk to add pictures of the species, and a sketch of locations of the two measures to the dataset.

2 Interactivity is the key to EDA

Interactivity is often understood in different ways. Many developers of software regard a system, which allows the user to type in commands or specify options in dialog boxes as interactive.

A user of a command line system would argue, that he works interactively, if he looks at four histograms with different binsizes, issuing four successive commands. Users of DataDesk, JMP (JMP (1995)) or MANET, who can change the binwidth by just dragging the mouse, will not share this interpretation of interactivity. To make it a bit clearer, no one would speak of a interactive 3d-plot, where the user has to specify the projection angle, but would expect to have interactive control over the projection, resp. the rotation.

In many situations, information can be revealed much better by observing the change of a scenario, than it can be done from a static view.

DataDesk offers some more interactive features, which are common to all outputs.

- **Drag & Drop**
Variables in all output windows can be added or substituted via drag & drop.
- **Pop-Up-Menus**
Although it would be canonical to remove a variable from a list, by

dragging it into the trash, DataDesk uses pop-up-menus to remove a variable from an output. In most cases, this solution is much more efficient.

- *HyperView*[®]*menus*

HyperView[®]*menus* provide guidance through potential analyses. They are available at the little triangle of any window, or are indicated by a change of the cursor to the shape of a pointing hand. Figure 6 shows

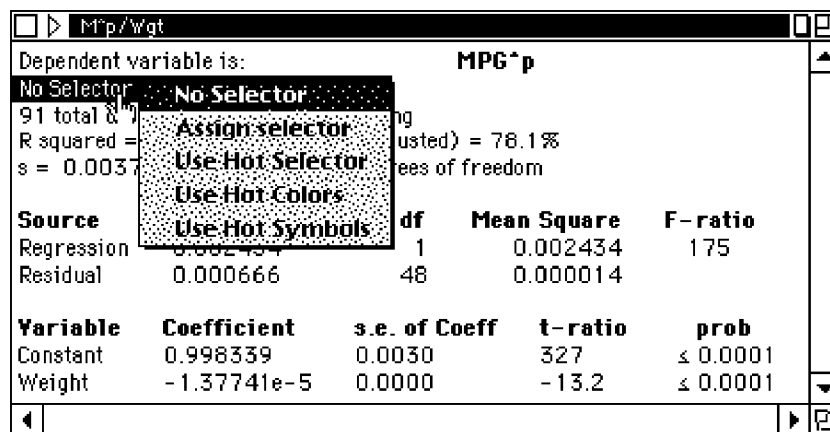


Figure 6: A *HyperView*[®]*menu* inside a regression window.

an example of a *HyperView*[®]*menu* inside a regression window.

- **Selectors**

Selectors are variables, which determine which subset to include in an analysis. There are three kinds of selectors:

- Standard Selectors

All cases with corresponding non zero cases in the selector variable are included. Unless this selector-variable is not based on a derived variable, this kind of selector is static.

- Colour & Marker Selectors

The subsets are determined by the cases attributes, i.e. only cases with non-standard colour or marker are used in an analysis.

- Hot Selectors

Hot selectors form the most efficient way to restrict an analysis to a certain subgroup. When defining the hot selector mode for an output window, only currently selected cases are included in an analysis. Thus hot selectors, which are at the same time a sixth selection mode, enable dynamic analyses to be performed.

The most important difference of DataDesk to other packages is the fact that interactivity is not restricted to graphs solely, but can be found inside tables and output windows of statistical models as well.

3 The Role of Statistical Models

Often exploratory-graphical approaches and mathematical statistical approaches are seen as concurrent methods or strategies. This is mostly due to the fact, that backers of the one or the other approach do not know the other in depth, but is not true at all.

In general the following strategy could be advised. From every dataset specific questions arise, and the analysis should answer these questions. These two points, the questions and solutions, should always frame the whole analysis.

Exploratory methods lead to an understanding of the structures of the data. In most situations, this is the first step in an analysis, to extract the qualitative information in the data. The step may be also used to redefine the questions, which have been formulated regarding the dataset.

In a next step, statistical models can be used to look at the data to validate the structure and obtain quantitative information. The results of the models

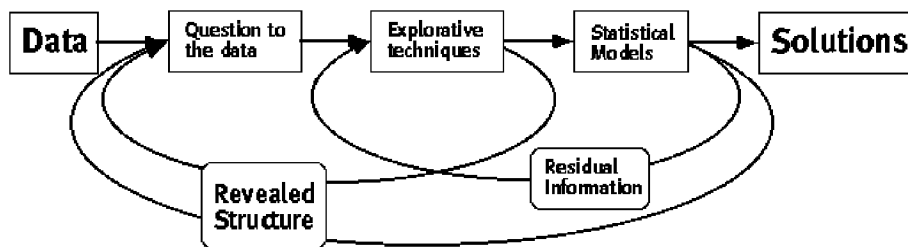


Figure 7: An analysis strategy, which includes both, exploratory and mathematical statistical techniques.

can reveal new structures in the data, and thus redefine the initial questions, or restart an exploratory analysis, using the residual output of the model. Figure 7 depicts the different steps of an analysis.

DataDesk offers the following statistical procedures.

- summary statistics
- hypothesis tests (parametric and nonparametric)
- correlation
- contingency tables

- cluster analysis
- principal components
- regression
 - linear
 - nonlinear
 - logistic
 - nonparametric
- General Linear Model
 - ANOVA
 - ANCOVA
 - MANOVA, covering
 - * repeated measures
 - * unbalanced designs
 - * nested designs
 - * designs with random terms
 - * designs that have cells with no cases

The general linear model, is covered under a consistent, easy-to-use interface.

4 Batteries included?

Systems like S-Plus (Venables & Ripley 1994) or XploRe (Härdle et al. 1995) are extendable systems from their basic design of being a programming language, with special build-in functions, which support the handling of data matrices and arrays.

Extendability can be desirable for two reasons:

- Systems, which do not offer a special analysis technique, can be extended, as long as the user is able to implement the new technique. E.g. for S-plus there exist a lot of libraries, and most of them are gathered at the statlib-index (<http://www.statlib.com>).
- Successive steps in an often recurrent analysis can be collected in a procedure. In contrast to the implementation of new procedures, this is just an assembly of already existing elements.

With graphical, desktop based systems, extendability seems to be a much harder problem than for programming languages. LispStat (Tierney 1990) tries to incorporate both aspects, which are extendability based on a programming language (Lisp) and high user interaction inside graphs. To achieve this,

LispStat offers a lot of functions which support the assembly of interactive graphs.

DataDesk uses a slightly different approach. The build-in *Action Language* is a visual programming language. Since commands and operators are entered by pop-up-menus and drag & drop, syntax errors are impossible. The Action Language serves for implementing new techniques. Figure 8 shows the complete reference of all commands of the Action Language.

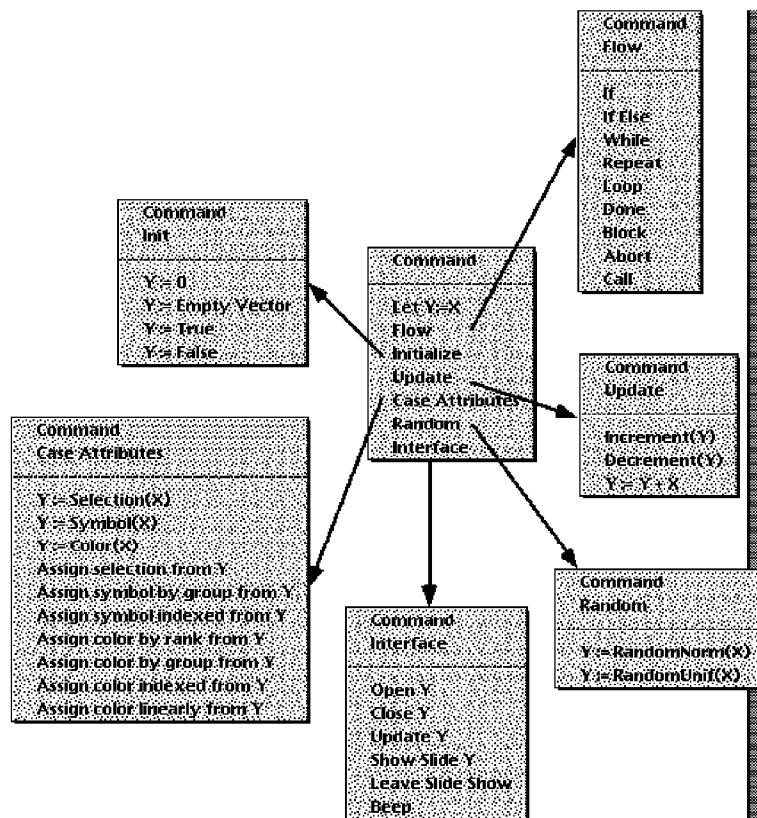


Figure 8: The complete reference of the Action Language

Whereas the use of the Action Language seems to be hard even for experienced programmers, the so called *Templates* in DataDesk are easy to use. Templates consist of a container, the corkboard, which can hold

- Variable sockets, which are placeholders for the input variables
- Action commands
- Buttons, which could start specific tasks

- Pictures

Figure 9 shows a sample template to draw a parallel coordinate plot. In this example only the variable slots and the three action buttons are placed on the corkboard.

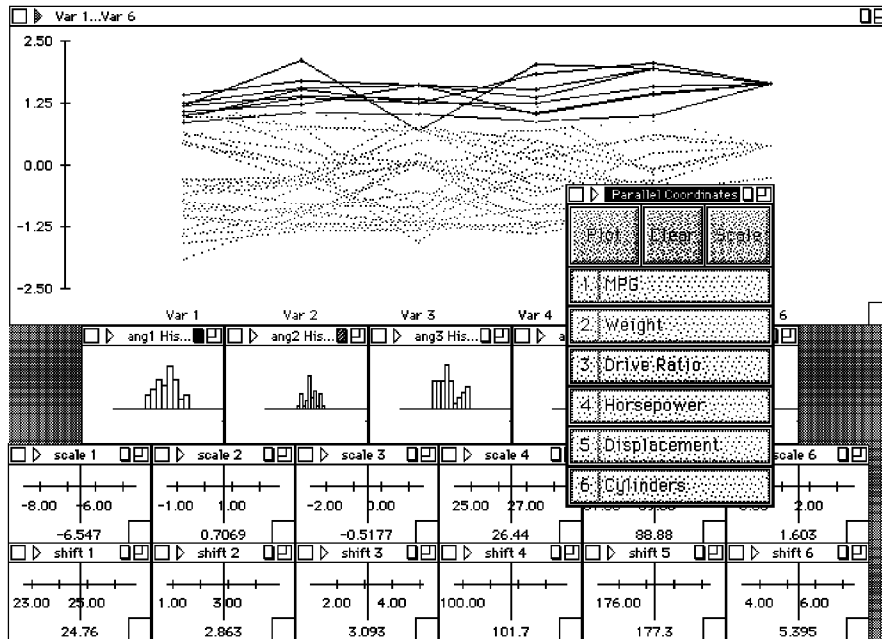


Figure 9: A template to draw an interactive parallel coordinate plot

Similar to the StatLib archive, Data Description Inc. runs a library of DataDesk templates at <http://www.datadesk.com/Templates.html>

All interactive exploratory systems seem to have much tighter limits concerning their extendability, as it is the case for systems based on a programming language. For all systems it is easy to point out, why this is the case. LispStat is a Lisp programming language, supporting interactive graphics. Thus the user must be very acquainted with the programming language as well as with the build-in objects. DataDesk was designed as a closed system, which offers only very basic extendability of functions and objects.

An interactive graphical system which is extendable from its basics, should be a mixture of the LispStat approach, i.e. offer graphic primitives, and DataDesk, i.e. offer a lot of build-in functions (based on the primitives) and offer a graphical interface for extending (programming) the system.

5 Where do we go from here?

DataDesk version 5 has been ten years in development. It is by far the most complete and efficient tool to work in an exploratory fashion with data sets. Since DataDesk was only developed for Macintosh compatible computers up till now, it proved popular only in a small circle. This will hopefully change with the new Windows version, which shipped in winter '96/'97.

There is the saying

If all you have is a hammer, every problem looks like a nail

This holds true for data analysis, too. If students do not learn to use new and modern tools, which support EDA, they will not start to work more flexibly. DataDesk is an excellent tool to serve both, analysis as well as teaching statistics and data analysis.

Paul Velleman's new interactive system "ActiveStats" (cf. Velleman (1997)), which is designed for teaching basic statistics, makes extensive use of DataDesk. Although this basic course only makes use a basic DataDesk functionality yet, it leads students to the right direction.

References

- Härdle, W., Klinke, S. & Turlach, B.A. (1995), *XploRe: An Interactive Statistical Computing Environment*. Springer, New York.
- JMP (1995). *JMP Statistical & Graphical Guide*. SAS Inst. Inc., Cary, NC.
- Theus, Martin (1996). MANET — Interactive Graphics for Missing Values. In: *New Techniques and Technologies for Statistics II*, IOS Press, Amsterdam.
- Tierney, Luke (1990). *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York.
- Tufte, Edward R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut.
- Tukey, John W. (1977). *EDA*. Reading MA, Addison-Wesley.
- Velleman, Paul F. (1995) *Data Desk 5.0, Data Description* Ithaca, New York.
- Velleman, Paul F. (1997) Multimedia for teaching statistics, In: *Advances in Statistical Software 6, Softstat 97*, ed. F. Faulbaum, Lucius & Lucius, Stuttgart, 509.
- Venables, W.N. & Ripley, B.D. (1994). *Modern applied statistics in S-PLUS*. Springer, New York.

Wilhelm, A. F. X., Unwin A. R. & Theus, Martin (1996) Software for Interactive Statistical Graphics — A Review, In: *Advances in Statistical Software* 5, *Softstat 95*, ed. F. Faulbaum, Lucius & Lucius, Stuttgart, 3–12.