Examinação de Bancos de Dados

- What all practising data analysts agree on is that the proportion of project time spent on data cleaning is huge. Estimates of 75–90% have been suggested. (UTH, pg 20{33pdf}).

Examinação de Bancos de Dados

- An examination of tables from the 1950 U. S. Census of Population, and of the basic Persons punch card, shows that a few of the cards were punched one column to the right of the proper position in at least some columns. The result is that numbers reported in certain rare categories -- very young widowers and divorces, and male Indians 10-14 or 20-24-- were greatly exaggerated. These errors occurred in spite of a careful checking program, and illustrate the necessity for users to view data concerning rare categories with special caution.

Dados Categorizados

- Barcharts and Spineplots for Univariate Categorical Data
- Mosaic Plots for Multi-dimensional Categorical Data

Spine Plots vs. Bar Charts

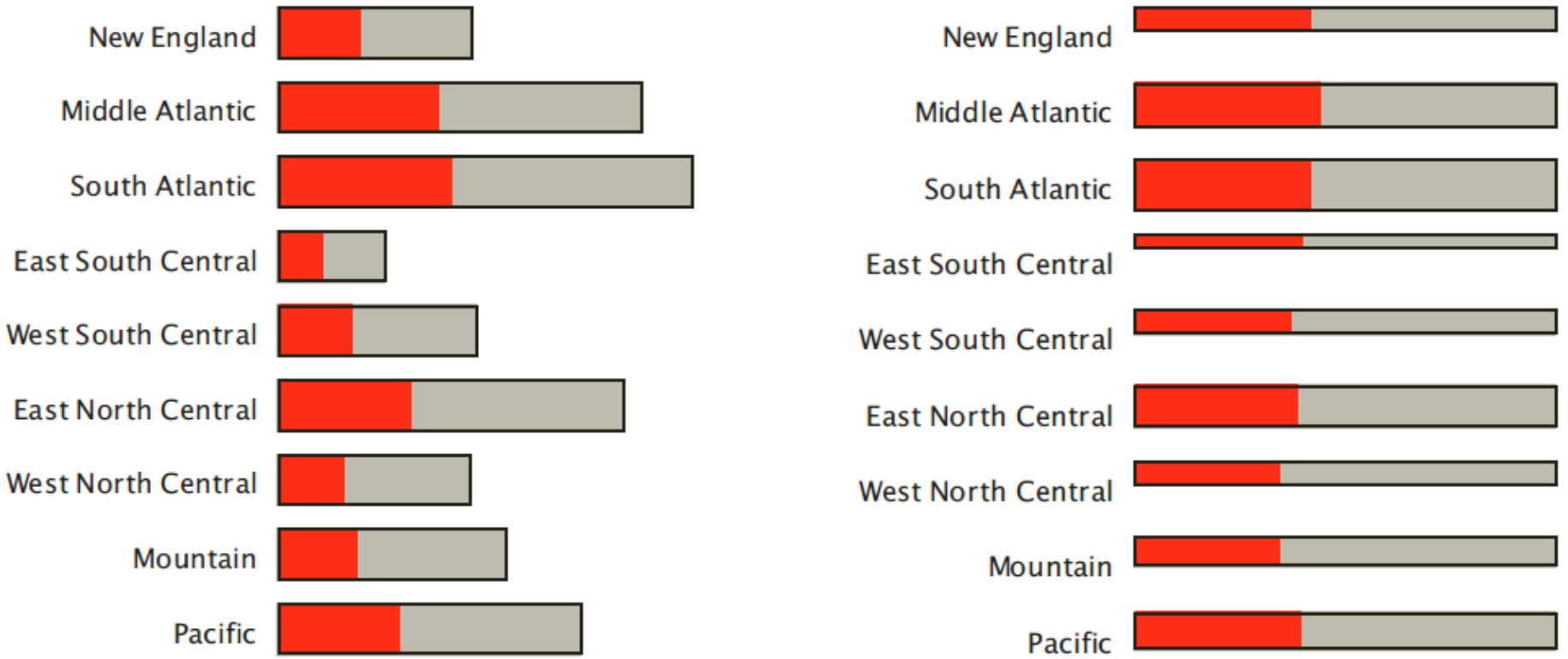


Fig. 2.3. *A spineplot (right) allows the comparison of proportions across categories.*

Mosaic Plots

- Are designed to show the dependencies and interactions between multiple categorical variables in one plot.
- Podem mostrar significância de diferenças.
- Spline Plots são Mosaic Plots

Mosaic Plots

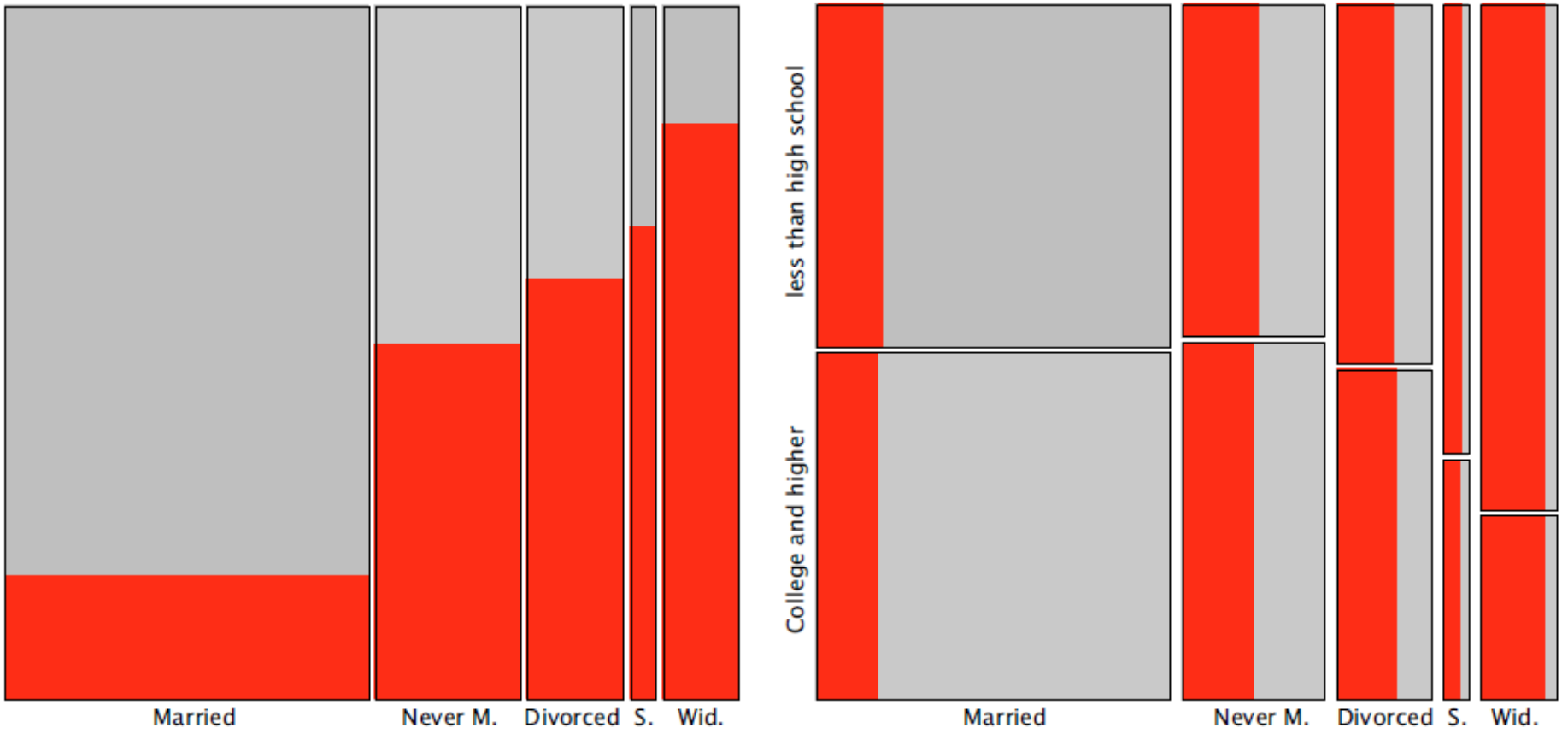


Fig. 2.6. *Development of a mosaic plot including* Marital Status *and* Education *(all females are highlighted).*

Dados Contínuos

- The most commonly used plots for continuous data are dotplots, boxplots, and histograms for one-dimensional data and scatterplots for twodimensional data. Methods and plots for higher dimensions of continuous data include parallel coordinates and the Grand Tour.

Dotplot

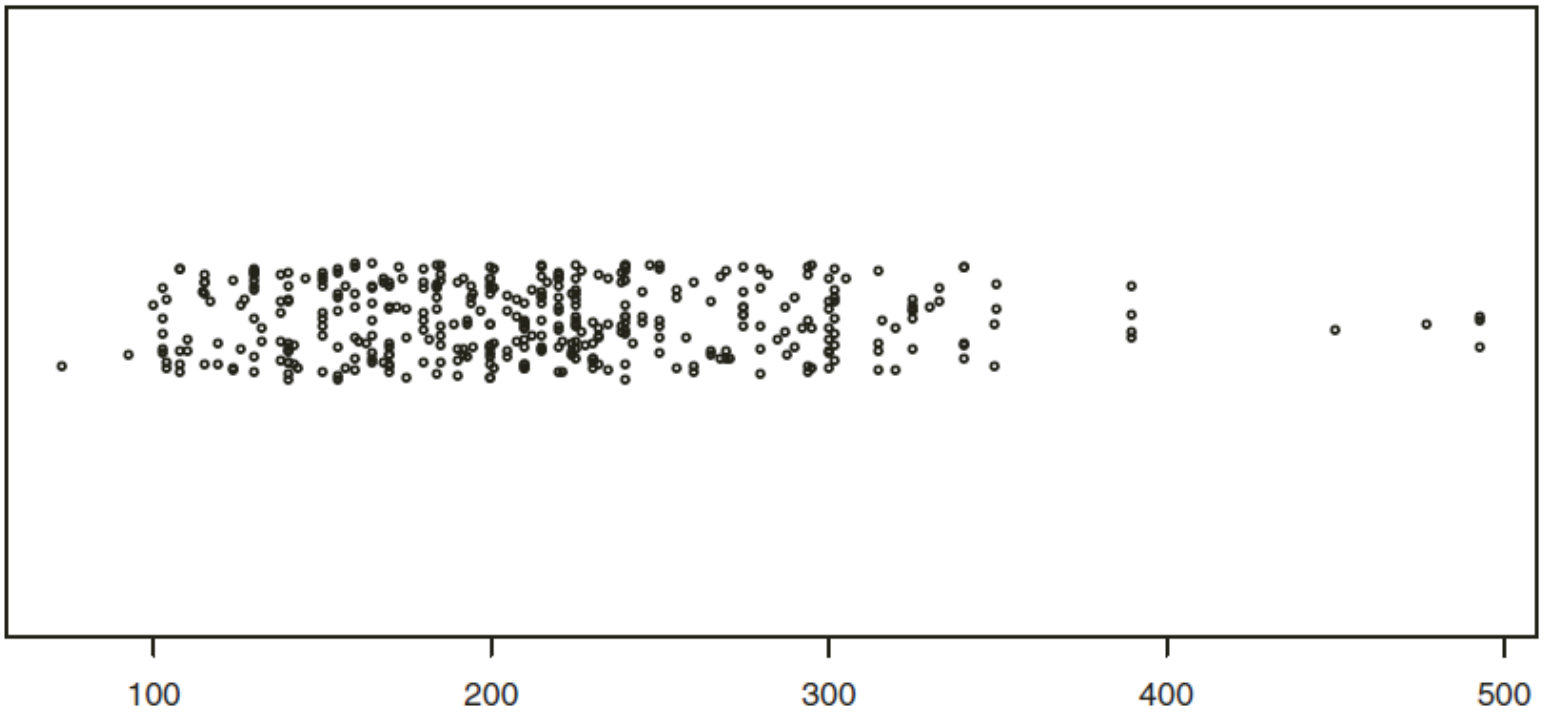


Fig. 2.7. *A jittered dotplot of* Horsepower *for the Cars2004 data.*

Histograma

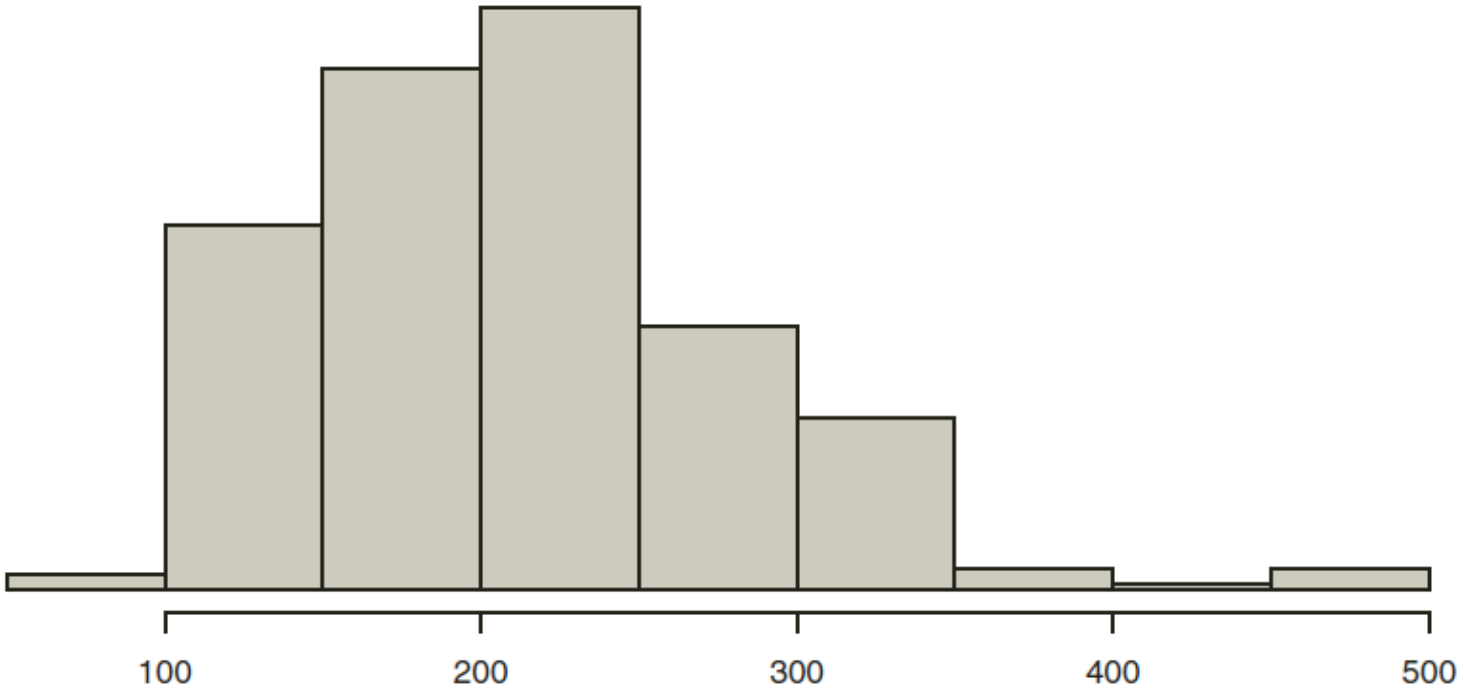


Fig. 2.8. *A default histogram of* Horsepower *for the Cars2004 data.*

Histograma

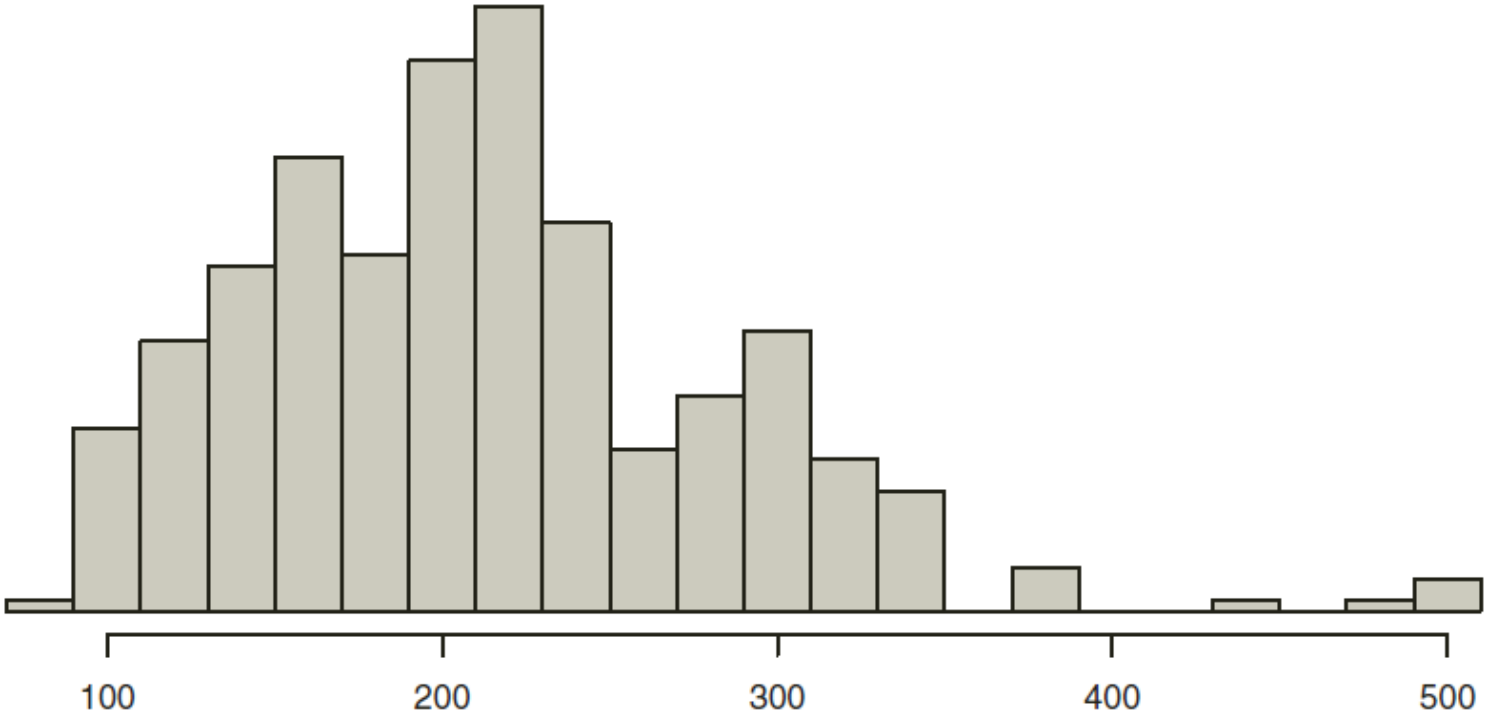


Fig. 2.9. *A histogram of* Horsepower *for the Cars2004 data, with anchorpoint 70 and binwidth 20.*

Boxplot

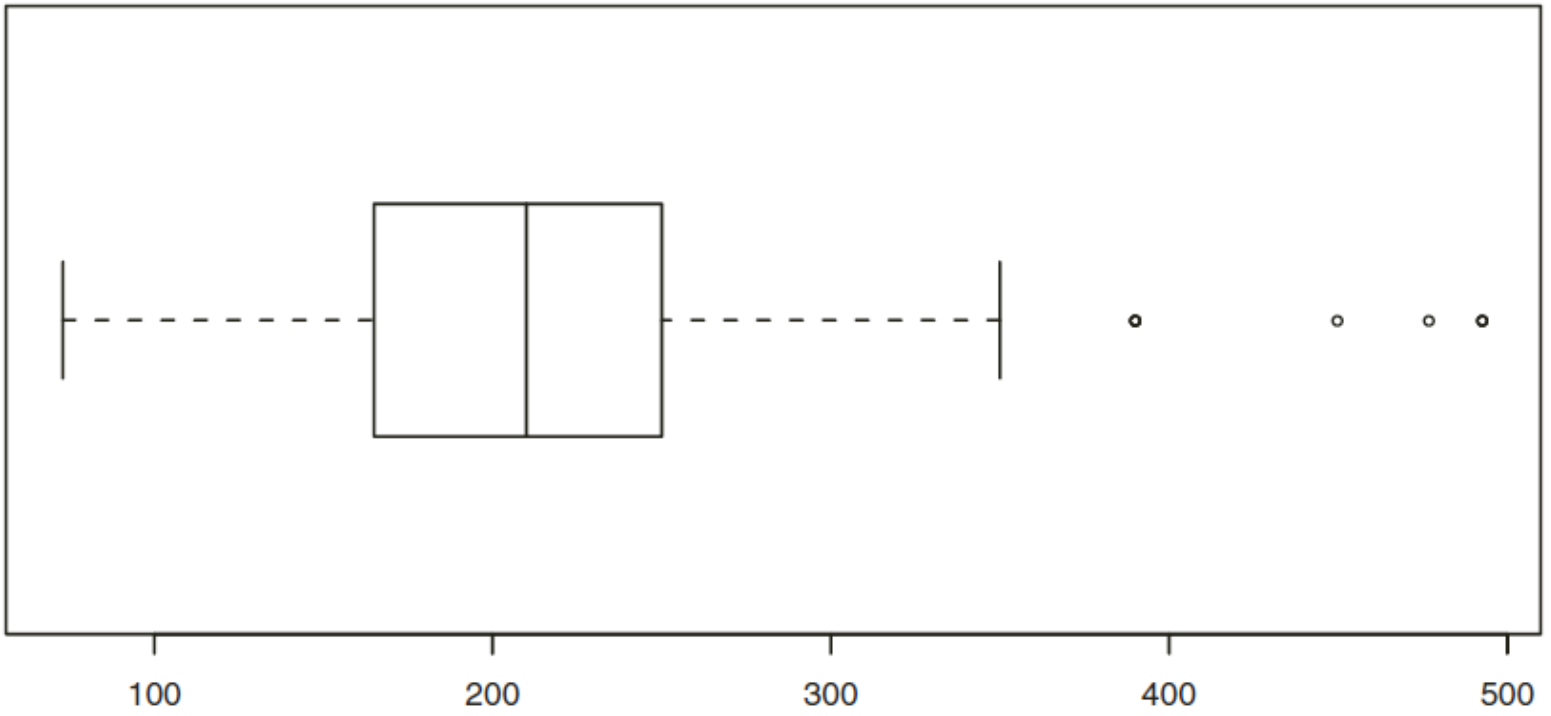


Fig. 2.10. *A boxplot of* Horsepower *for the Cars2004 data.*

Comparação

Table 2.1. *A Comparison of the Strengths (+) and Weakness (–) of Plots for Univariate Continuous Data ('○' Means Neither Strength nor Weakness)*

	Dotplot	**Histogram**	**Boxplot**
Visualizing the shape of a distribution	–	+	○
Detection of outlier	+	–	+
Inspection of gaps, discreteness	+	○	–
Size of the sample	○	○	–
Comparison of distributions	–	○	+

Scatterplots and Scatterplot Matrices

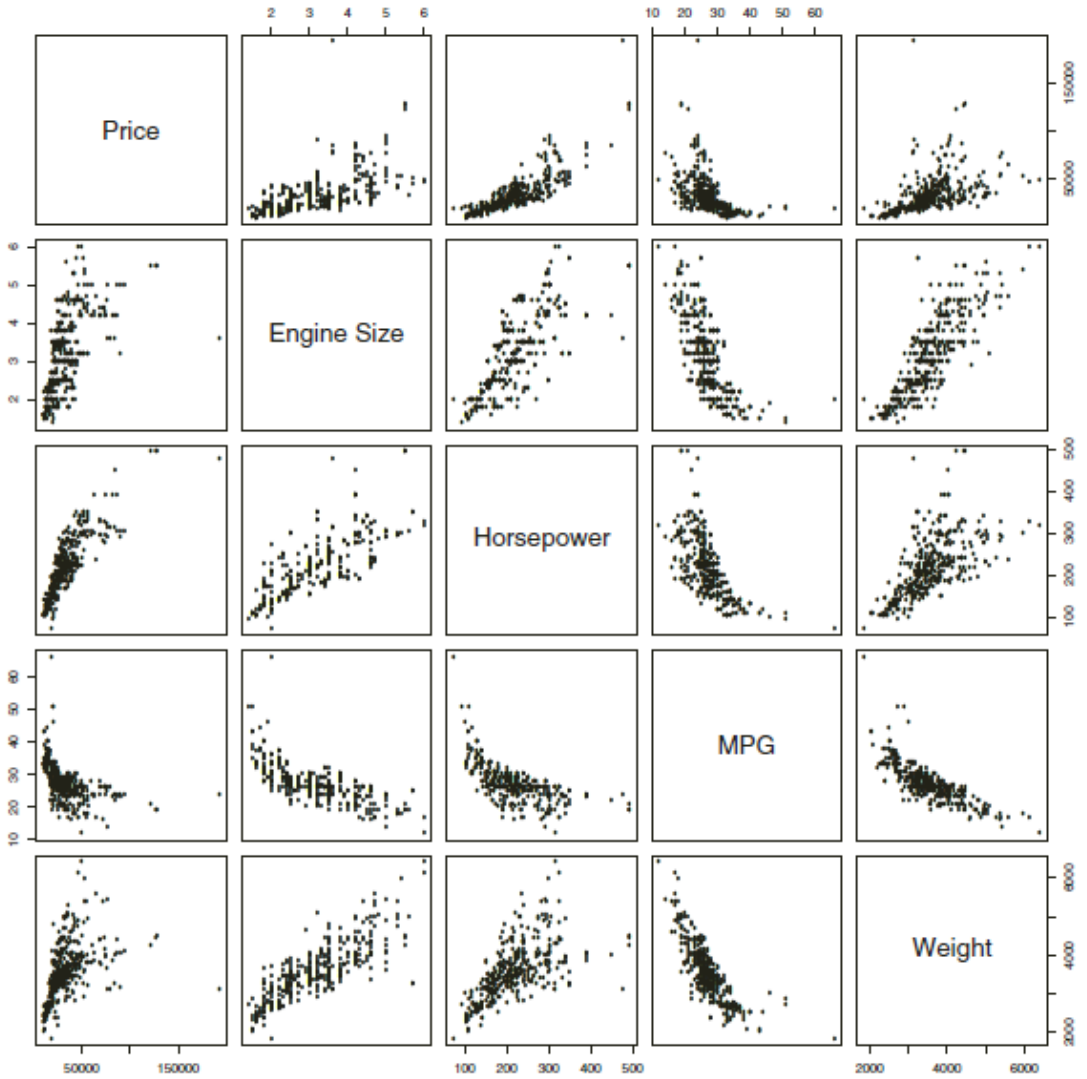


Fig. 2.12. *A full scatterplot matrix for 5 variables from the cars dataset.*

Grand Tour

- http://www.jstatsoft.org/v40/i02/

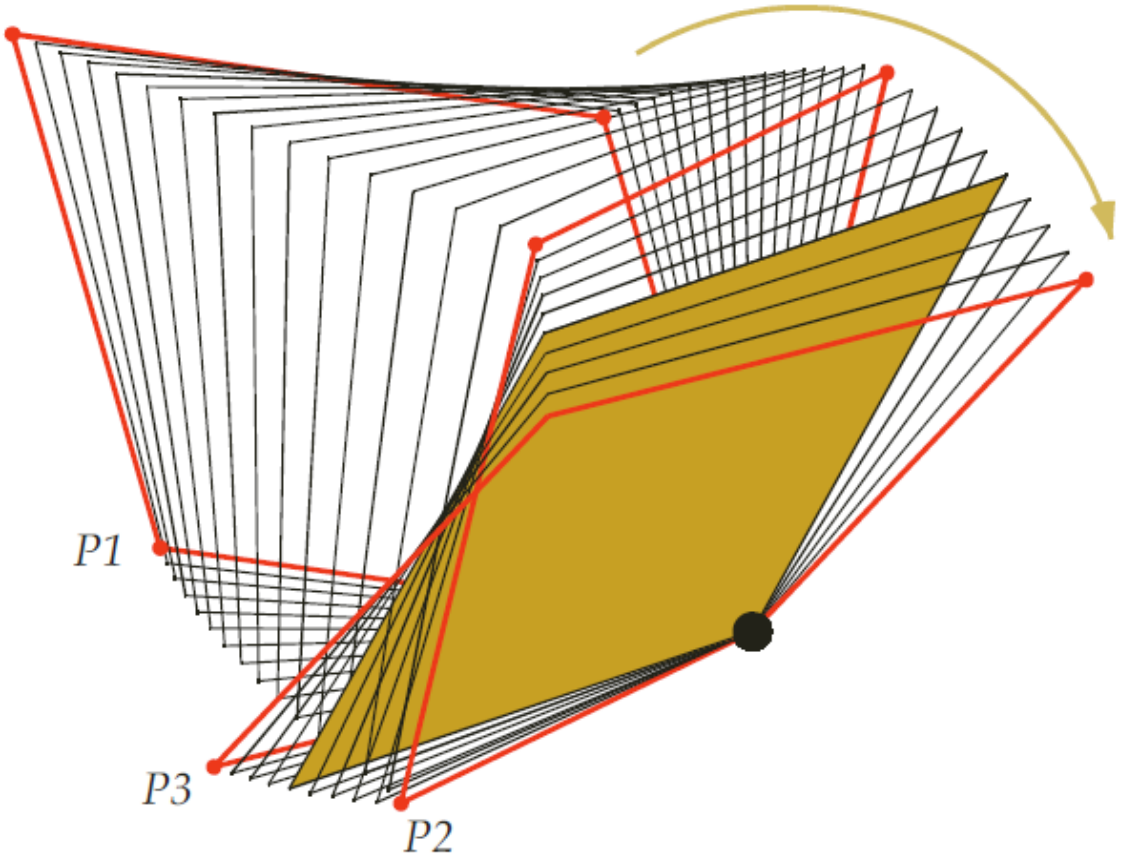


Fig. 2.13. *Sample path of a Grand Tour. The three base frames P1 to P3 are interpolated by intermediate projection planes to generate a smooth transition.*

Parallel Coordinates

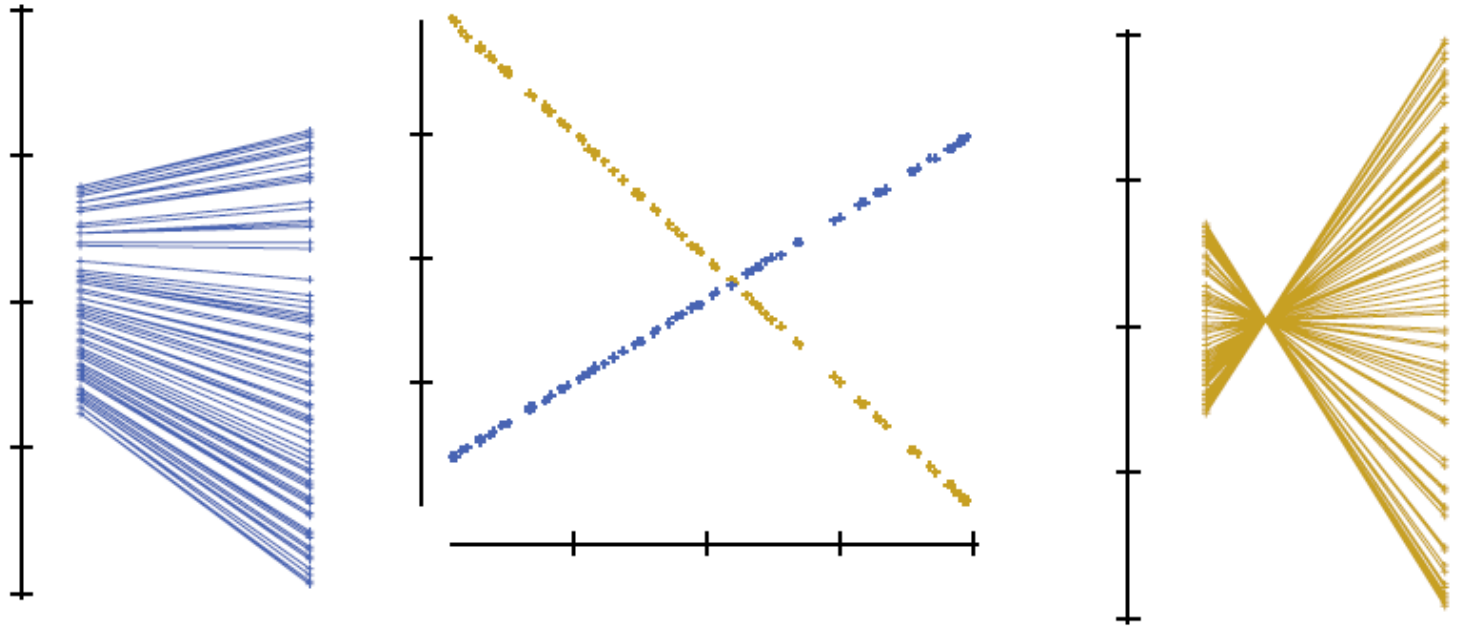


Fig. 2.15. *How lines in two dimensions translate into parallel coordinates.*

Parallel Coordinates

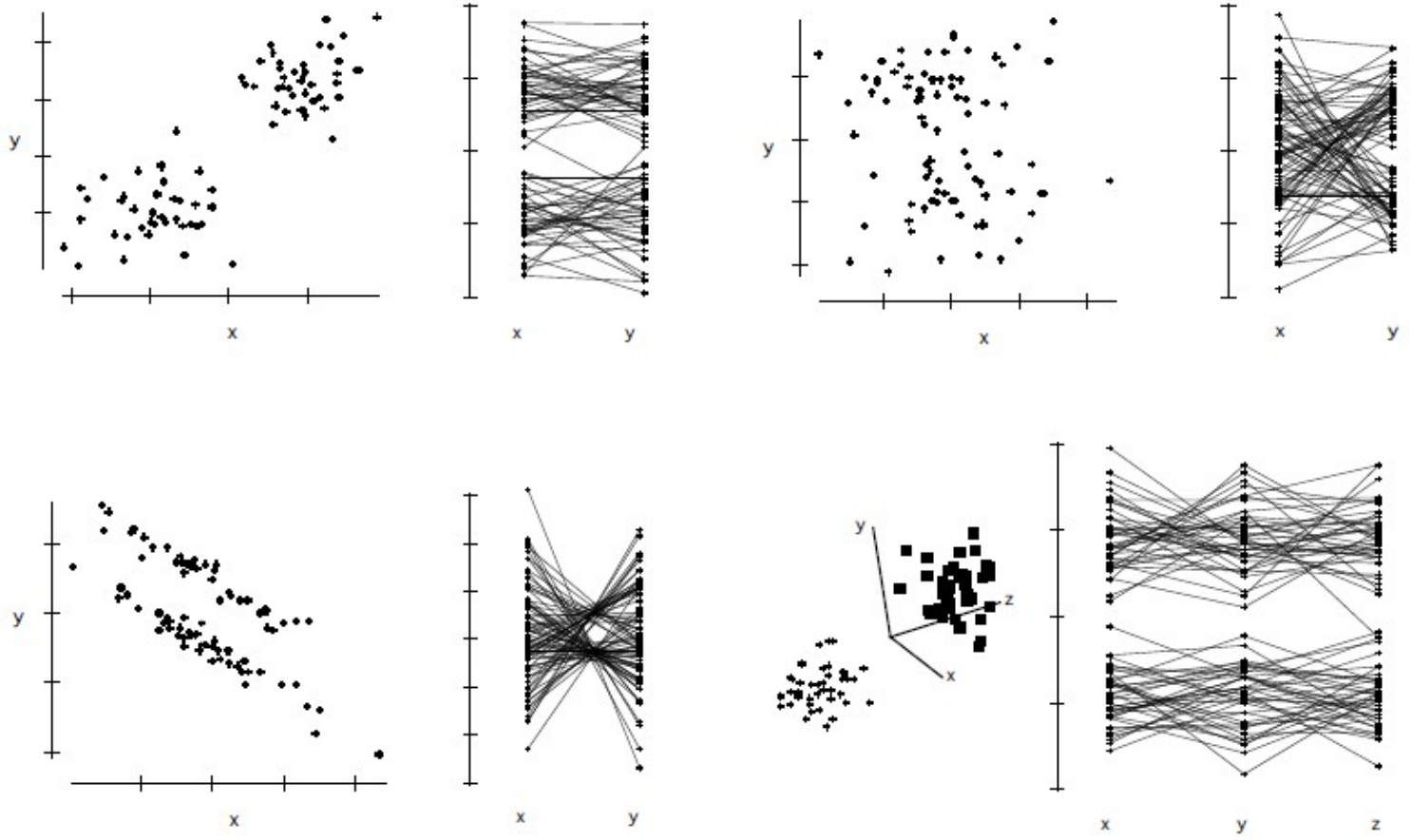


Fig. 2.16. *How groups can be identified in parallel coordinates.*

Dados Mistos (Lattice/Trellis)

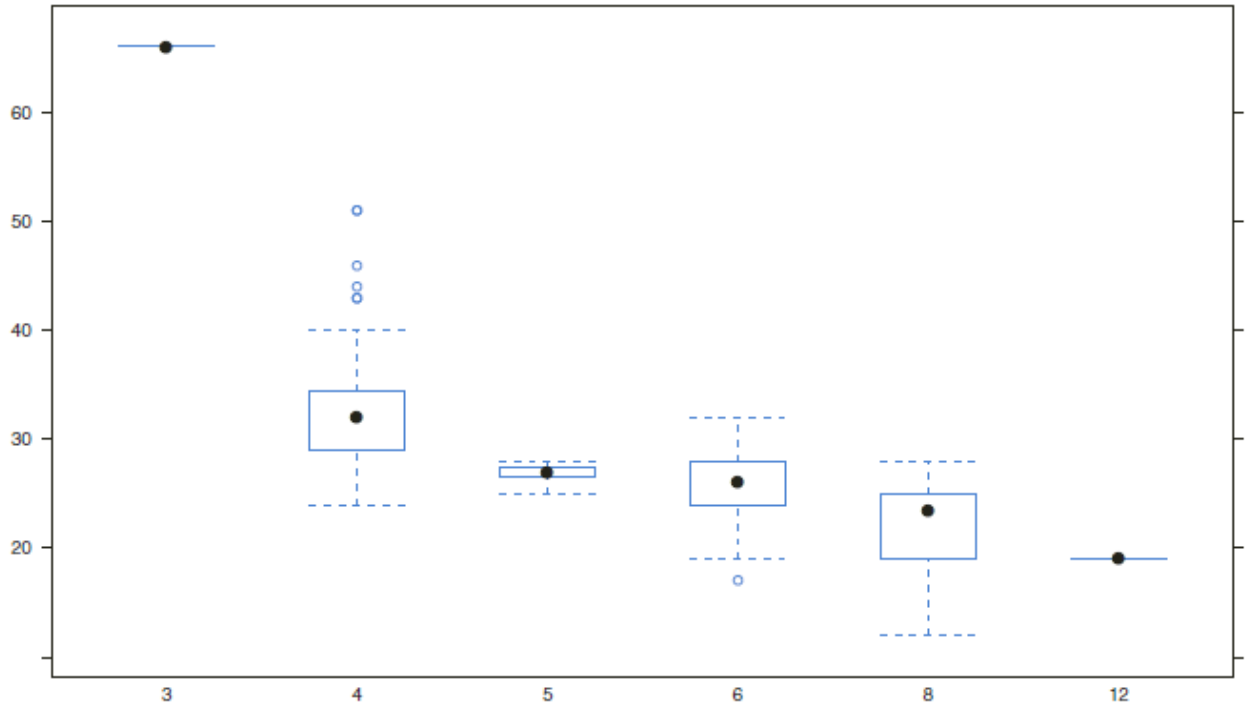


Fig. 2.18. *Boxplot* MPG *by* Cylinder *as simple form of a trellis display.*

Trellis (conditioning plots)

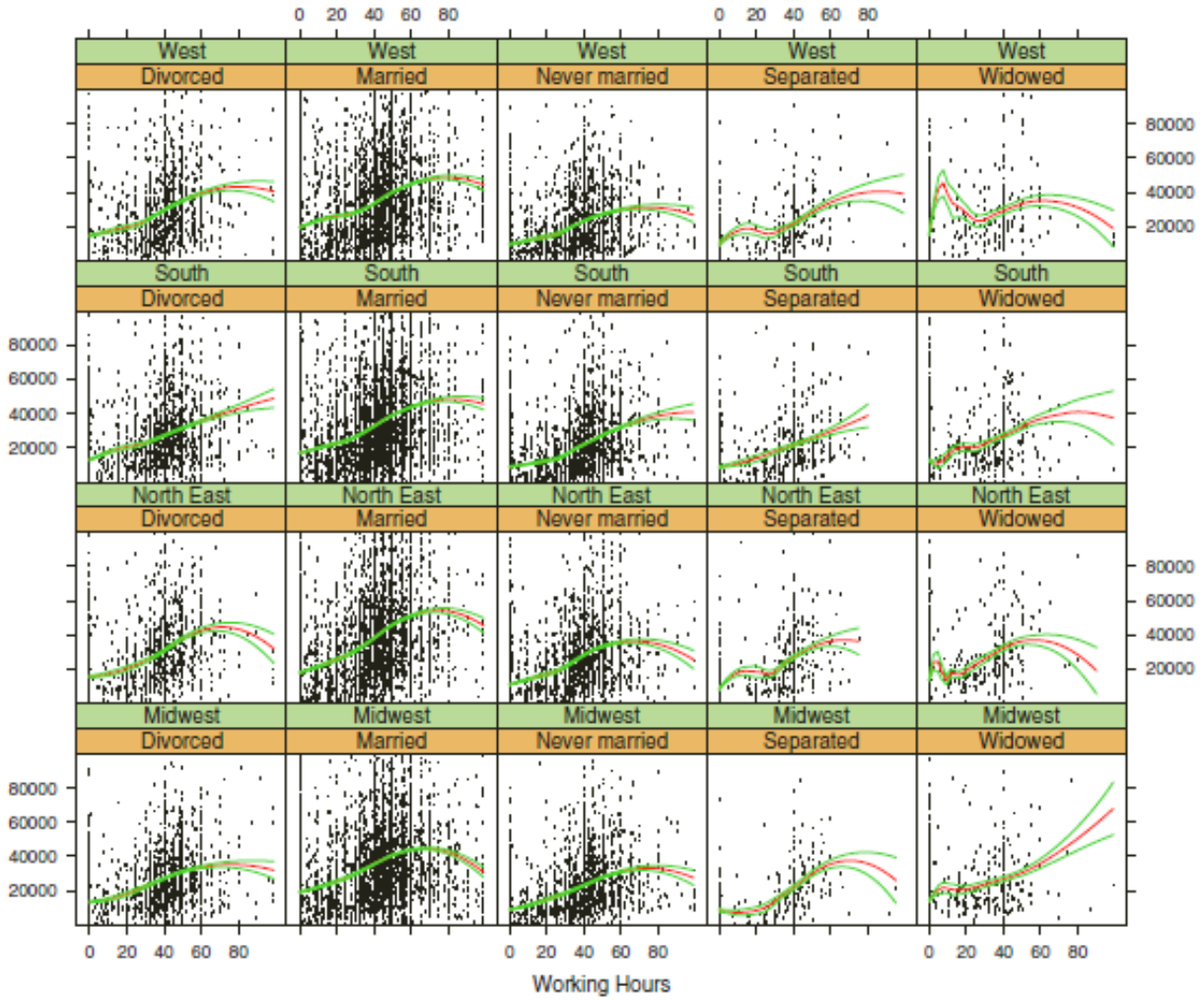
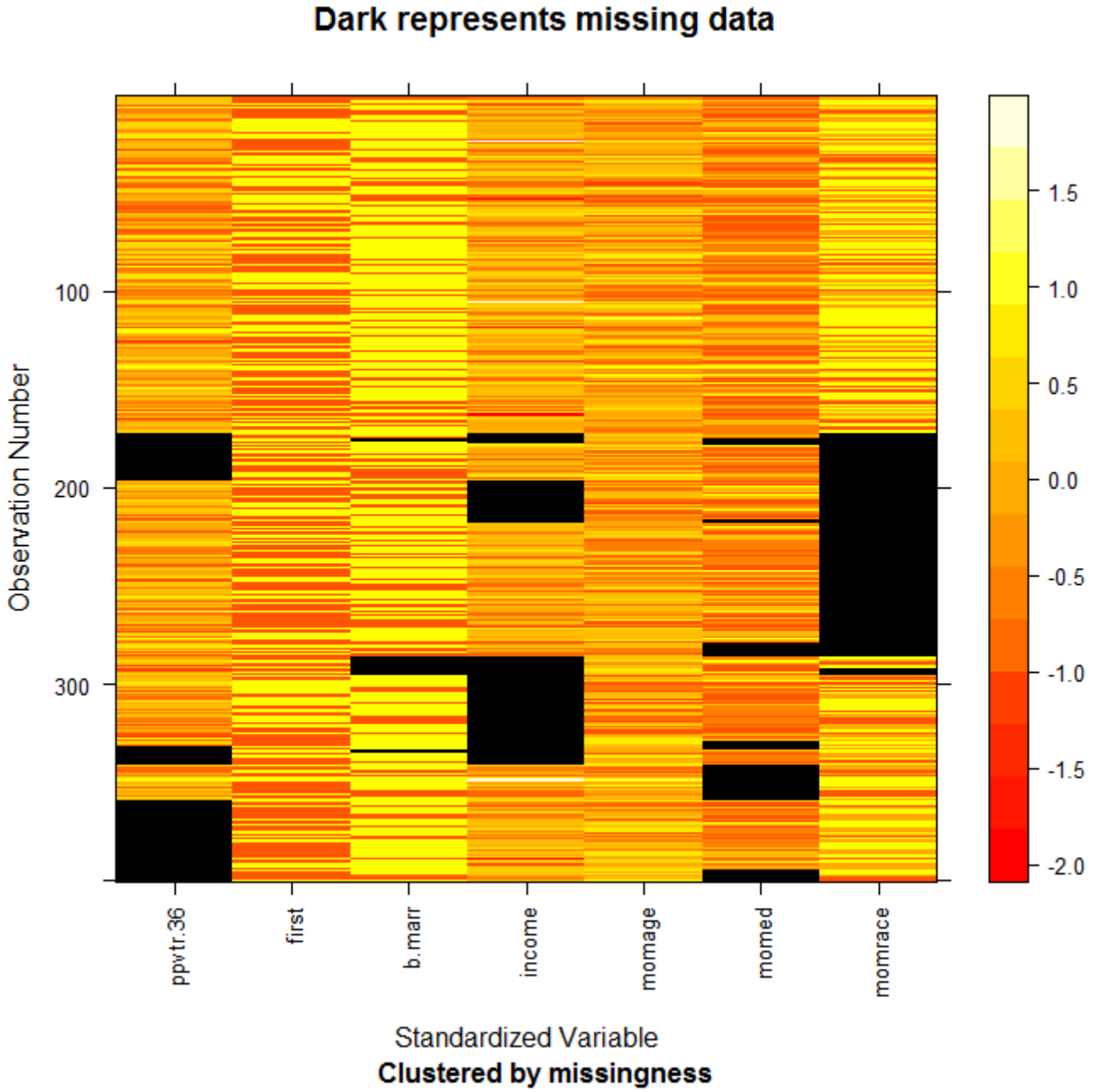


Fig. 2.19. *A trellis display showing scatterplots for* Income *vs.* Working Hours *conditioned by* Region *and* Marital Status *for the* Census *data. Each plot panel has a local regression smoother superimposed. Confidence bands have been added to illustrate the variability of the estimate.*

Missing Data



Missing Data

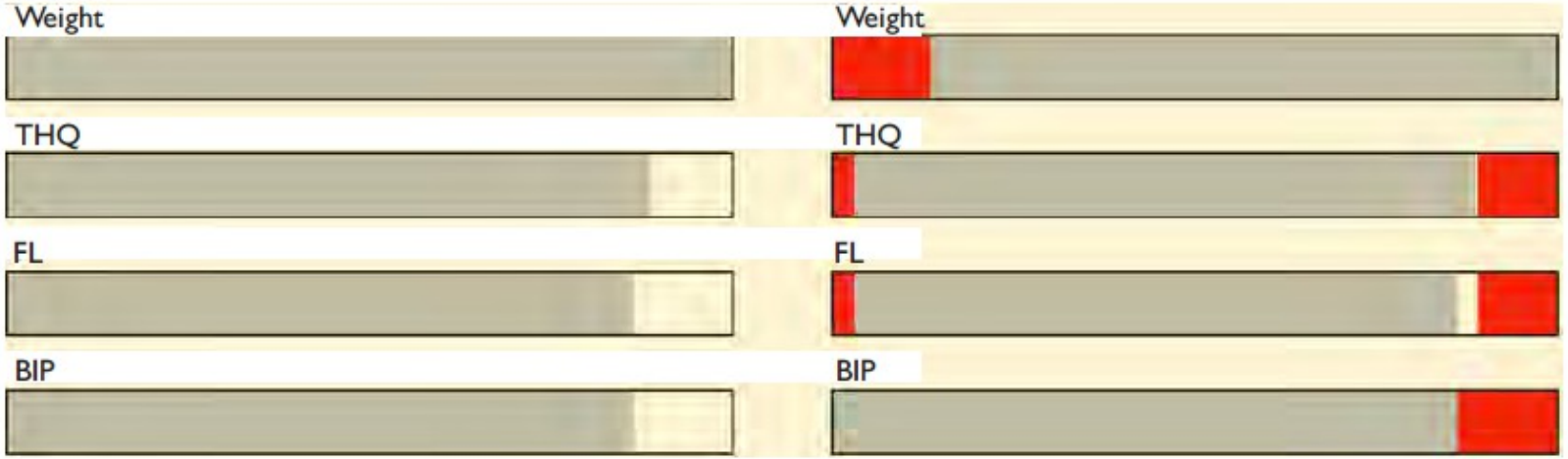


Fig. 2.26. *A missing value plot for the ultrasound dataset (left). All cases with missing values in* BIP *are selected (right).*

Missing Data

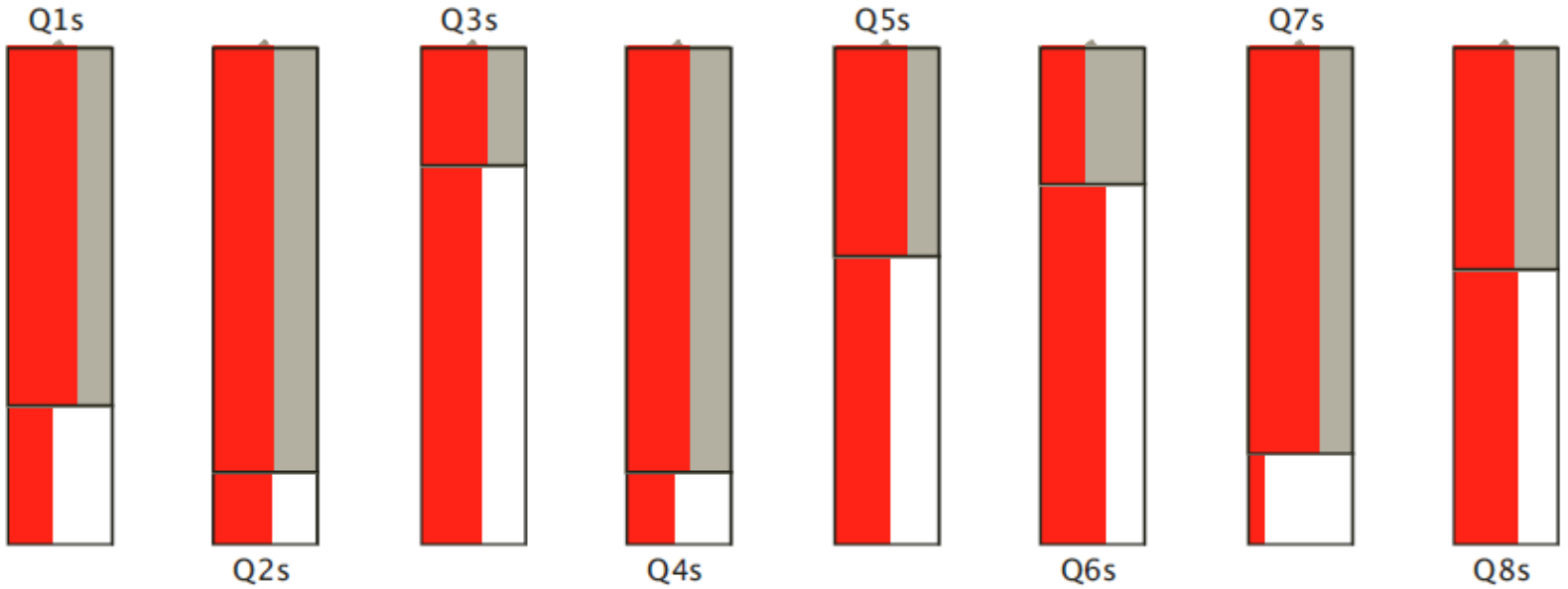


Fig. 2.27. *A missing value plot for the exams dataset. Students who passed the exam are highlighted.*

Missing Data

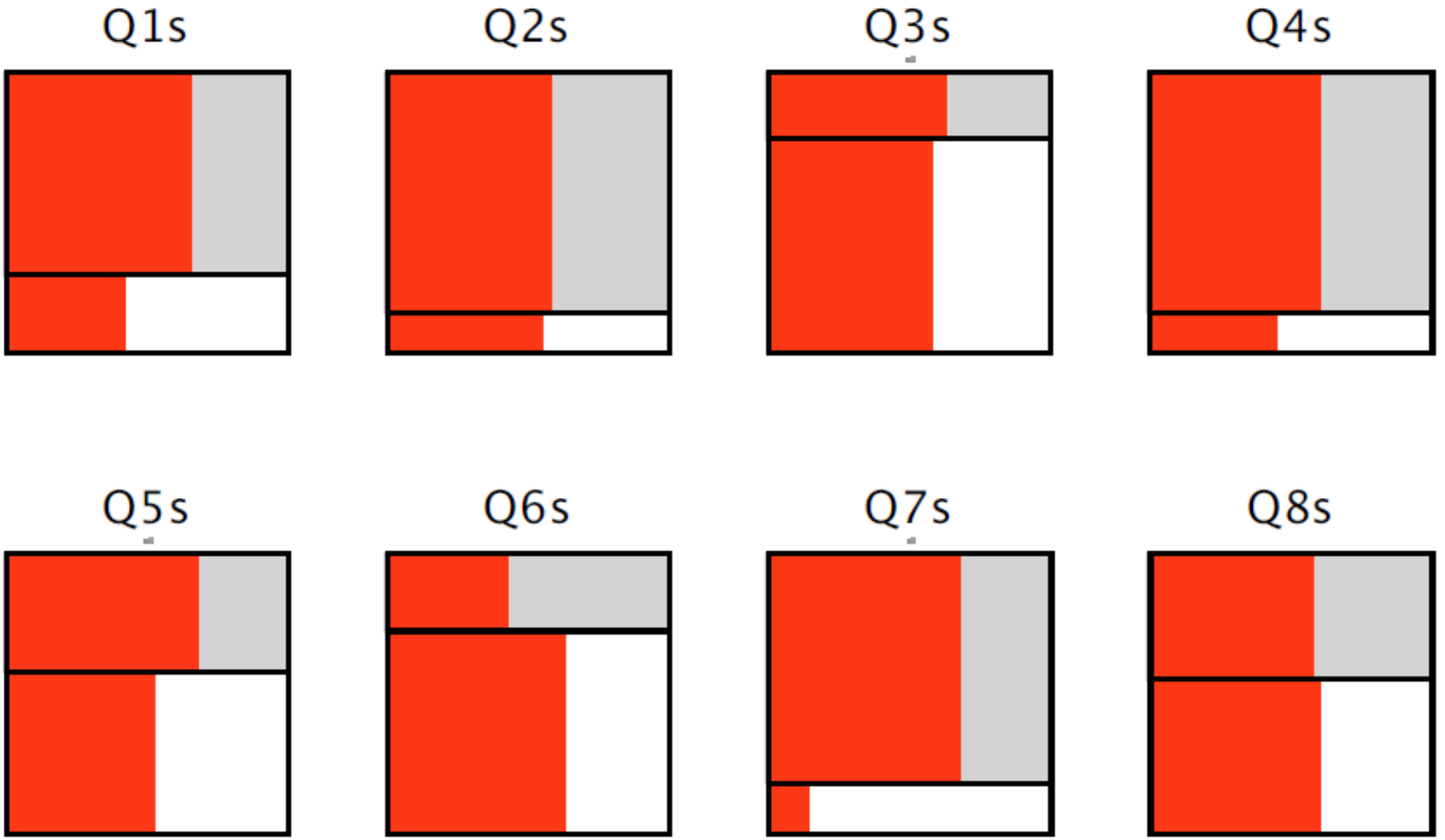


Fig. 2.28. *The same data as in Figure 2.27 with optimized aspect ratio.*

Graphical Data Analysis with R

- http://www.gradaanwr.net/
- Vai para "Content" para ver os gráficos e códigos de cada capítulo
- Os pacotes necessários são: ggplot2, gridExtra, ggthemes, dplyr, GGally, vcd, extracat
- Na aba "Supporting Materials" tem um pacote de dados e correções para os códigos

Grandes Bancos de Dados

- Gráficos de "área" (histograma, mosaic) não sofrem tanto com mais observações
- Gráficos de "pontos" sofrem, sim.
- Boxplots talvez não sejam tão interessantes, apesar de não sofrerem.
- Qui-quadrado sempre significante...

Gráficos de Barras?

- Três tipos de variáveis:
- Número de categorias fixas a priori; (estado civil)
- Número de categorias limitadas a priori; (tipos de carros; país de nascimento)
- Número de categorias "illimitadas"; (cidade de nascimento; data/hora/minuto de nascimento)

  nesse caso pode ordenar e agregar interativamente

Mosaic com muitas categorias

- Scroll ou Zoom

Binning de Gráficos de Dispersão

- Quando temos muitos dados, podemos usar a resolução do gráfico para escolher o tamanho dos intervalos a serem agrupados.

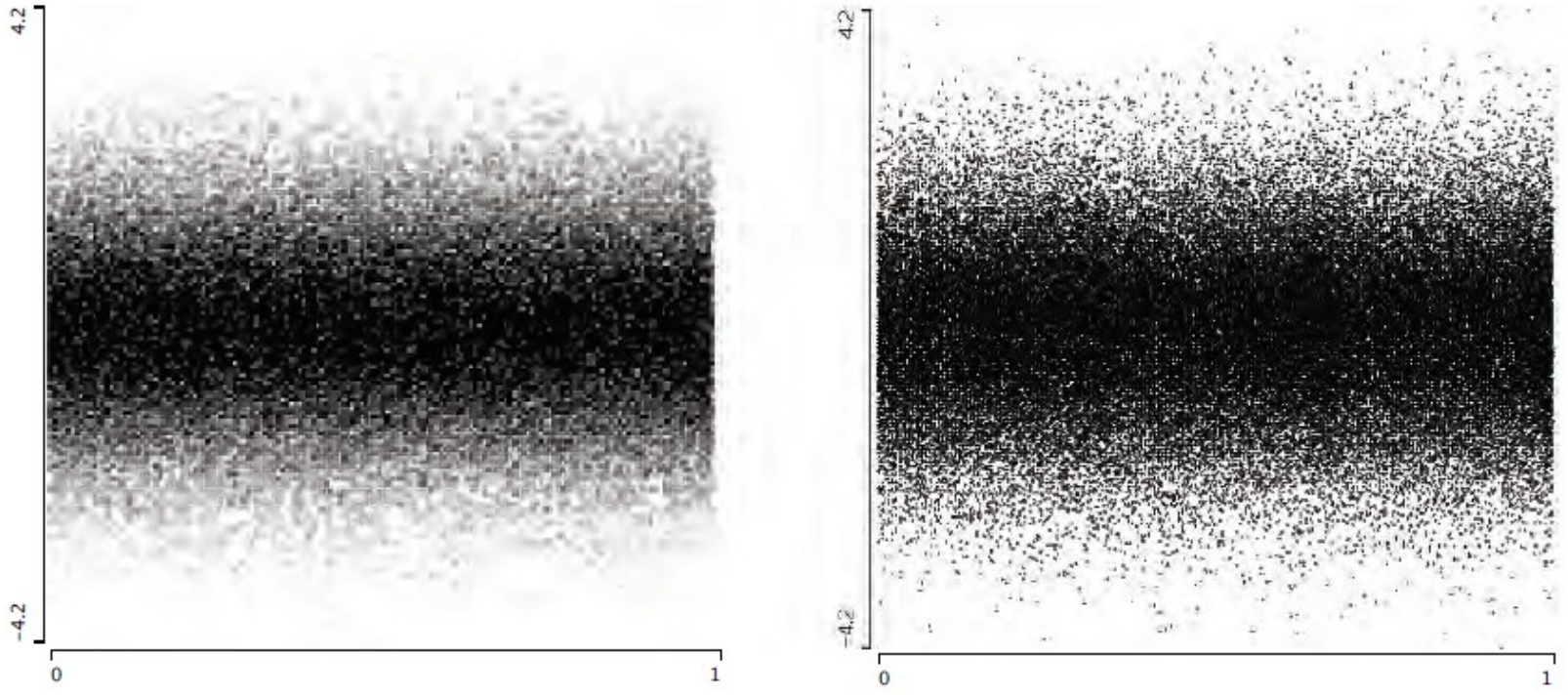


Fig. 3.10. *The sample data from Figure 3.6. Raw scatterplot (left) and binned to a 256 × 256 binning grid (right). Both displays render the structure of the data sufficiently well.*

Parallel Coordinates

- Muda a transparência dos pontos.

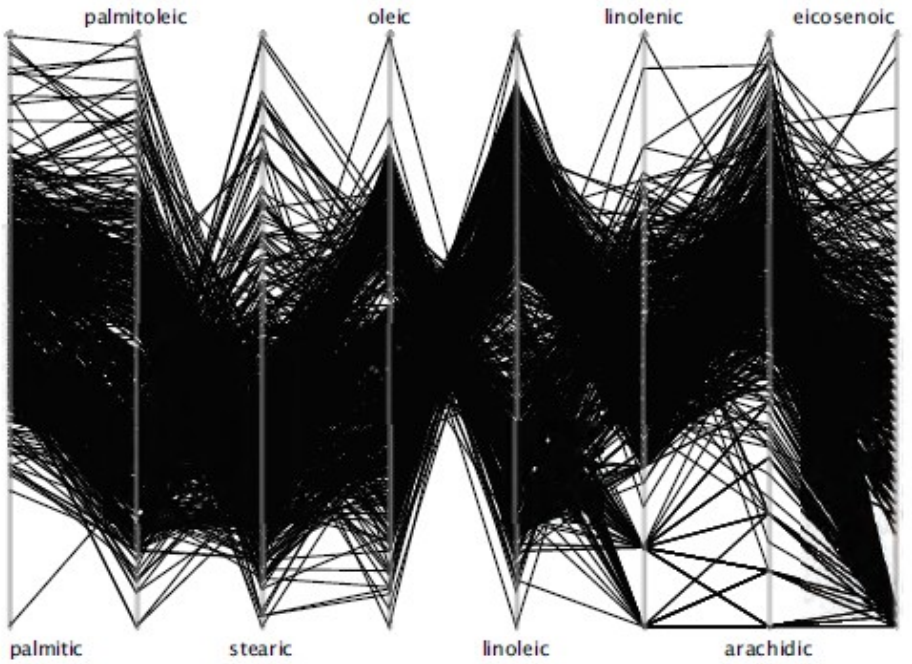


Fig. 3.8. *A parallel coordinate plot of 572 measurements on Italian olive oils. The heavy overplotting obscures the group structure in the data.*

Parallel Coordinates

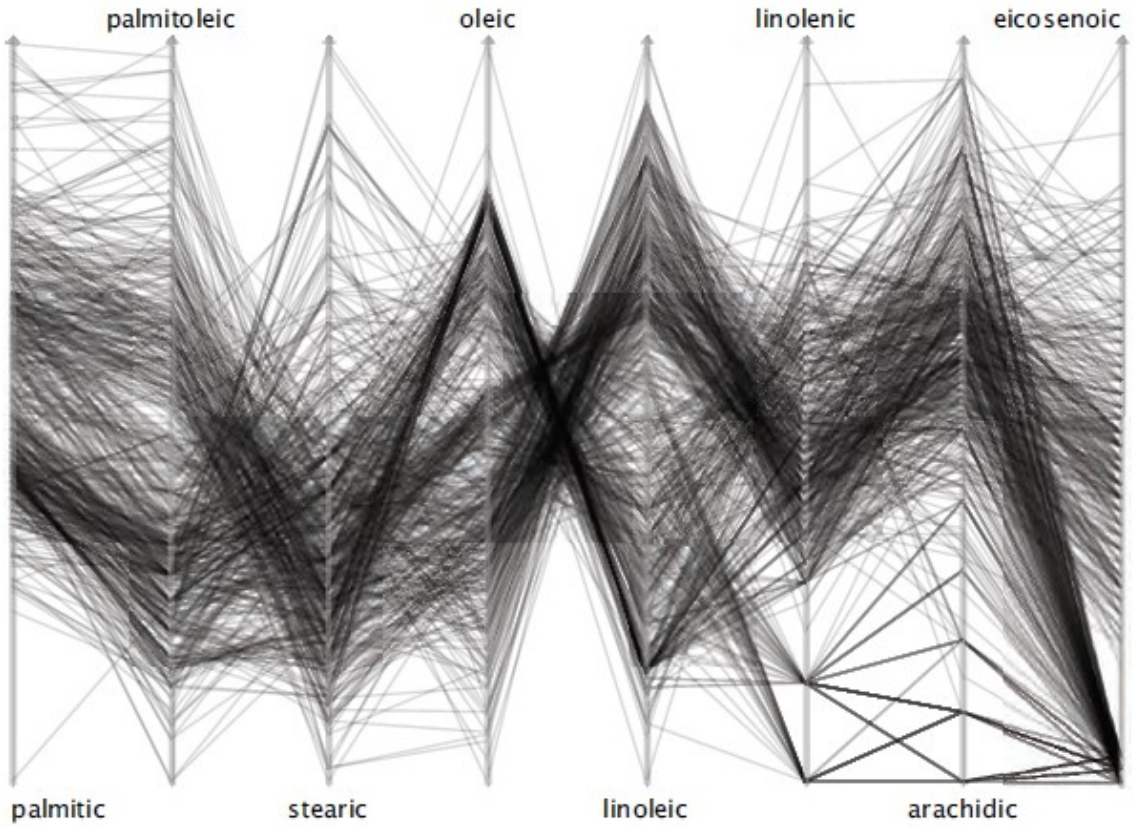


Fig. 3.9. *The same data as used in Figure 3.8 with α-transparency. Several of the 9 Regions can be seen.*

Parallel Coordinates

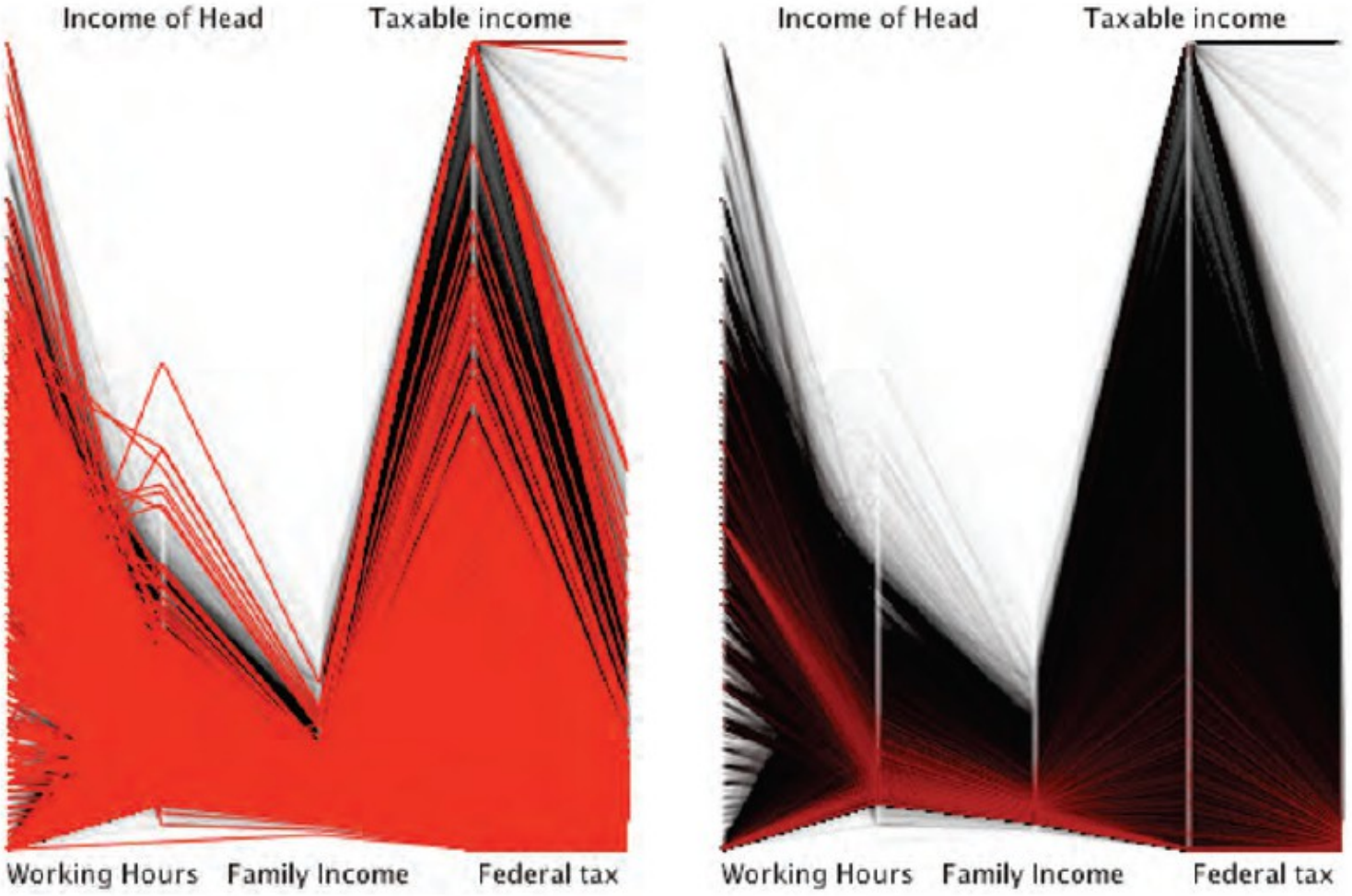


Fig. 4.15. *Parallel coordinate plots without α-blending on the highlighted cases (left) and with α-blending.*

Testes Robustos

- Se os dados não são normais, qual o problema?
- $H(x) = 0.9\Phi(x) + 0.1\Phi(x/10),$

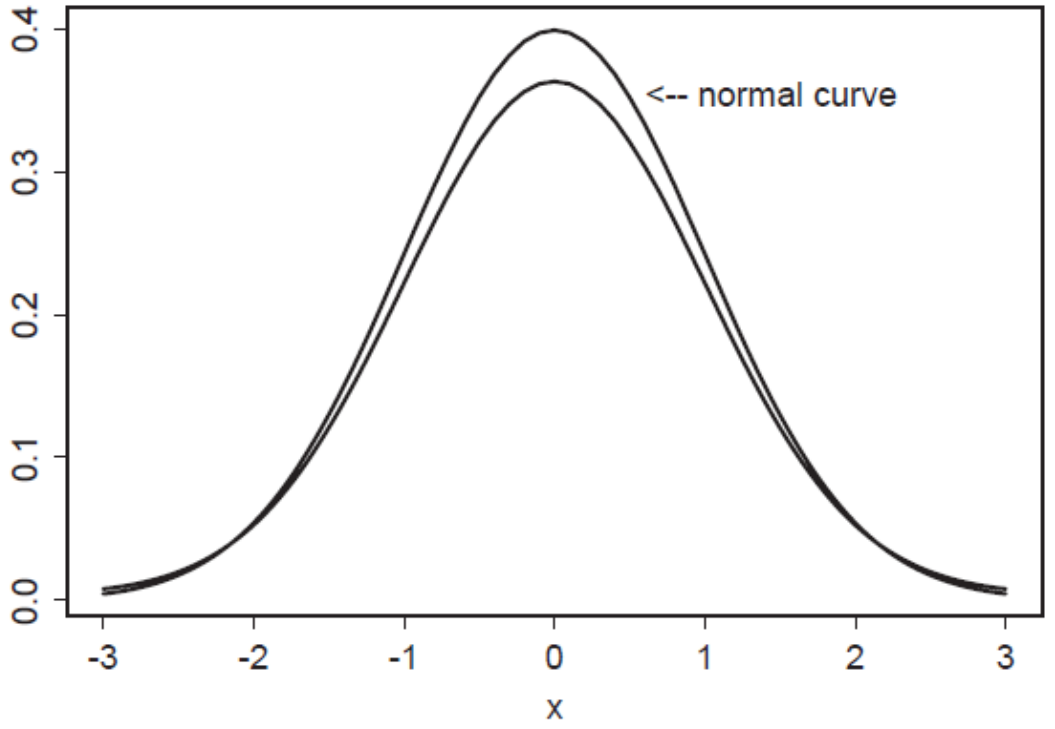


Figure 1.1 Normal and contaminated normal distributions.

Normal Contaminado

- O Normal Contaminado tem variância 10,9 em comparação com a variância do Normal.
- Sensivel às caudas.
- Intervalo de confiancia mais que três vezes maior.
- Seria melhor não assumir normalidade e perder um pouco de eficiência quando os dados são normais.
- Existem várias testes que fazem isso.

O que é o poder de um teste?

- A probabilidade de rejeitar a hipótese nula se a hipótese alternativa for verdadeira.
- Os testes que se baseiam em normalidade tem poder baixo com dados não-normais (27%) enquanto testes robustos podem chegar a poderes de 70% em situações semelhantes.
- Mas, isso importa menos quando temos muitos dados.

Medidas de Localização

- Definição Formal (média, mediana, ...)

$$\theta(X + b) = \theta(X) + b$$

$$\theta(-X) = -\theta(X)$$

$$X \geq 0 \text{ implies } \theta(X) \geq 0$$

$$\theta(aX) = a\theta(X).$$

Trimmed Distributions

$$\frac{1}{1-2\gamma} f(x), \qquad x_\gamma \le x \le x_{1-\gamma},$$

$$\frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \qquad -\infty \le x \le \infty,$$

$$f(x) = \frac{1}{.6}\frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \qquad -.84 \le x \le .84.$$

Comportamento sob contaminação (Função Influência)

$$F_{x,\epsilon} = (1-\epsilon)F + \epsilon\,\Delta_x.$$

$$\mathrm{IF}(x) = \lim \frac{T(F_{x,\epsilon}) - T(F)}{\epsilon},$$

$$T(F) = E(X),$$

$$\mathrm{IF}(x) = x - \mu,$$

Quanta contaminação para quebrar o estimador?

- A média pode ser "quebrada" por qualquer nível de contaminação.
- A médiana precisa de contaminação de 50%
- A média truncada precisa de contaminação maior do que foi excluído.
- O estimador do quantil q precisa de

  min(q,1-q) de contaminação para "quebrar".

Qual minha sugestão?

- Regressão quantílica
- Pacote quantreg no R