The Generalized Pairs Plot

John W. Emerson, Walton A. Green, Barret Schloerke,
Dianne Cook, Heike Hofmann, and Hadley Wickham

Department of Statistics
Yale University, New Haven, CT 06520

Department of Organismic and Evolutionary Biology
Harvard University, Cambridge, MA 02138

3 * Department of Statistics
Iowa State, Ames, IA 50011

Department of Statistics
Rice University, Houston, TX 77251

February 6, 2011

Author's Footnote:

John W. Emerson (john.emerson@yale.edu) is Associate Professor, Department of Statistics, Yale University; P.O. Box 208290, New Haven, CT 06520-8290. Walton A. Green (wgreen@fas.harvard.edu) is Research Associate, Department of Organismic and Evolutionary Biology, Harvard University; 26 Oxford St, Cambridge, MA 02138. Barret Schloerke (schloerke@gmail.com) is an undergraduate student at Iowa State University; Ames, IA 50011-1210. Dianne Cook (dicook@iastate.edu) is Full Professor of Statistics at Iowa State University; Ames, IA 50011-1210. Heike Hofmann (hofmann@iastate.edu) is Associate Professor of Statistics at Iowa State University; Ames, IA 50011-1210. Hadley Wickham (hadley@rice.edu) is Assistant Professor of Statistics at Rice University; P.O. Box 1892, Houston, Texas 77251-1892.

Abstract

This paper develops a generalization of the *scatterplot matrix* (also called a *pairs plot* or a *generalized draftsman's display*) based on the recognition that most data sets include both categorical and quantitative information. Traditional grids of scatterplots often obscure important features of the data when one or more variables are categorical but coded as numerical. Instead, we offer a range of flexible options including variations on the mosaic display (where areas are proportional to counts in a contingency table), boxplots, stripplots, and other depictions of joint and conditional distributions of combinations of categorical and quantitative variables. The use of these features may reveal structure in multivariate data which otherwise might go unnoticed in the process of exploratory data analysis.

KEYWORDS: graphics, scatterplot matrix, generalized draftsman's display, mosaic plot, grammar of graphics, exploratory data analysis

1 Introduction

The practice of graphical representation of quantitative information dates to the 18th century work of Playfair (1786). Modern graphical exploration has grown largely from the original works of Chernoff (1973), Tukey (1977), Chambers, Cleveland, Kleiner and Tukey (1983), Tufte (1983), and Cleveland (1985). Since then, some major methodological contributions include Cleveland (1993), Becker, Cleveland and Shyu (1996) and Wilkinson (1999). There have been numerous other contributions, many cited in this paper. Most modern methods of graphical display have been implemented in the R language and environment for statistical computing (R Development Core Team 2005).

This paper contributes to the development of the *pairs plot*, first appearing, to the best of our knowledge, in Hartigan (1975). It is also referred to as the *generalized draftsman's display* by Tukey and Tukey (1981) and Chambers et al. (1983), and the *scatterplot matrix* (SPLOM) by Cleveland (1993) and Basford and Tukey (1999). The pairs plot is a grid of scatterplots showing the bivariate relationships between all pairs of variables in a multivariate data set. Although the authors of this paper (and many other academics and data analysts) regularly use this graphical display, it isn't clear that it is widely used in practice. Our informal survey of several statistics texts that include multiple regression revealed inconsistent use of pairs plots.

Most data sets consist of both quantitative and categorical variables. When all variables (or when all variable of interest) are quantitative, the scatterplot matrix is a natural tool for graphical exploration. Friendly (1994) proposed an alternative form of the plot for displaying pairwise relationships among a set of categorical variables based on the mosaic plot (Hartigan and Kleiner 1984). Emerson, Green and Hartigan (2006) presented the first known *generalized pairs plot*, addressing the need for a more flexible display of a mixture of quantitative and categorical variables. Though the use of *generalized* contrasts the original usage of Chambers et al. (1983), the authors agreed that the name seems most appropriate and should be adopted for this purpose.

Section 2 presents the basic design of the generalized pairs plot. We then discuss two implementations available in R extension packages `gpairs` (Emerson and Green 2010) and `GGally` (Schloerke, Cook, Hofmann and Wickham 2010) in Sections 3 and 4, respectively. The former approach was a methodological development for exploratory data analysis, while the lat-

ter focuses on these plots as a contribution to the framework of Wilkinson's grammar of graphics as implemented by Wickham (2009). Section 5 concludes with a discussion. Supplementary materials available online provide extensive additional examples.[1]

2 The generalized pairs plot

The generalized pairs plot should not be confused with the generalized draftsman's display of Chambers et al. (1983); we regard the latter as a traditional pairs plot or scatterplot matrix of quantative information. Figure 1 shows an example of a scatterplot matrix of Fisher's iris data (Fisher 1936), originally collected by Anderson (1935). Here, the species is treated numerically (1 for Setosa, 2 for Versicolor, and 3 for Virginica). This plot could be improved by using color to identify the species instead of explicitly including the numerical representation of species as a quantitative variable. Doing so uncovers striking clusterings of petal and sepal meaurements by species, an exercise left to the reader.

When a data set (such as the iris data) includes one or more categorical variables the traditional display can be less than ideal. Friendly (1994) proposed a grid of mosaic tiles for displaying sets of entirely categorical variables. Our generalization takes this a step further, recognizing the need for different types of panels that together can display a rich set of features of a diverse collection of continuous and categorical variables. There are three general types of displays. A panel (or tile, or display) containing a graphic or other summary information corresponding to two quantative variables is called *purely quantative*. A panel for two categorical variables is called *purely categorical*. The last type corresponds to one categorical and one quantitative variable, called a *mixed* display.[2]

Scatterplots are naturally used with two quantitative variables, and various options may provide information on correlation, missing values, or linear or non-linear fits, for example. Mosaic plots (Hartigan and Kleiner 1984) provide a graphical display of counts in a contingency table for two categorical variables where areas are proportional to counts. There are several ways

[1] We should consider this last part on the nature of any supplementary materials.

[2] Walton isn't sure formal terminology is needed. He proposes quantitative-quantitative, quantitative-categorical, and categorical-categorical, perhaps as an alternative to standard phrasing. Worth debating further before we wordsmith extensively.

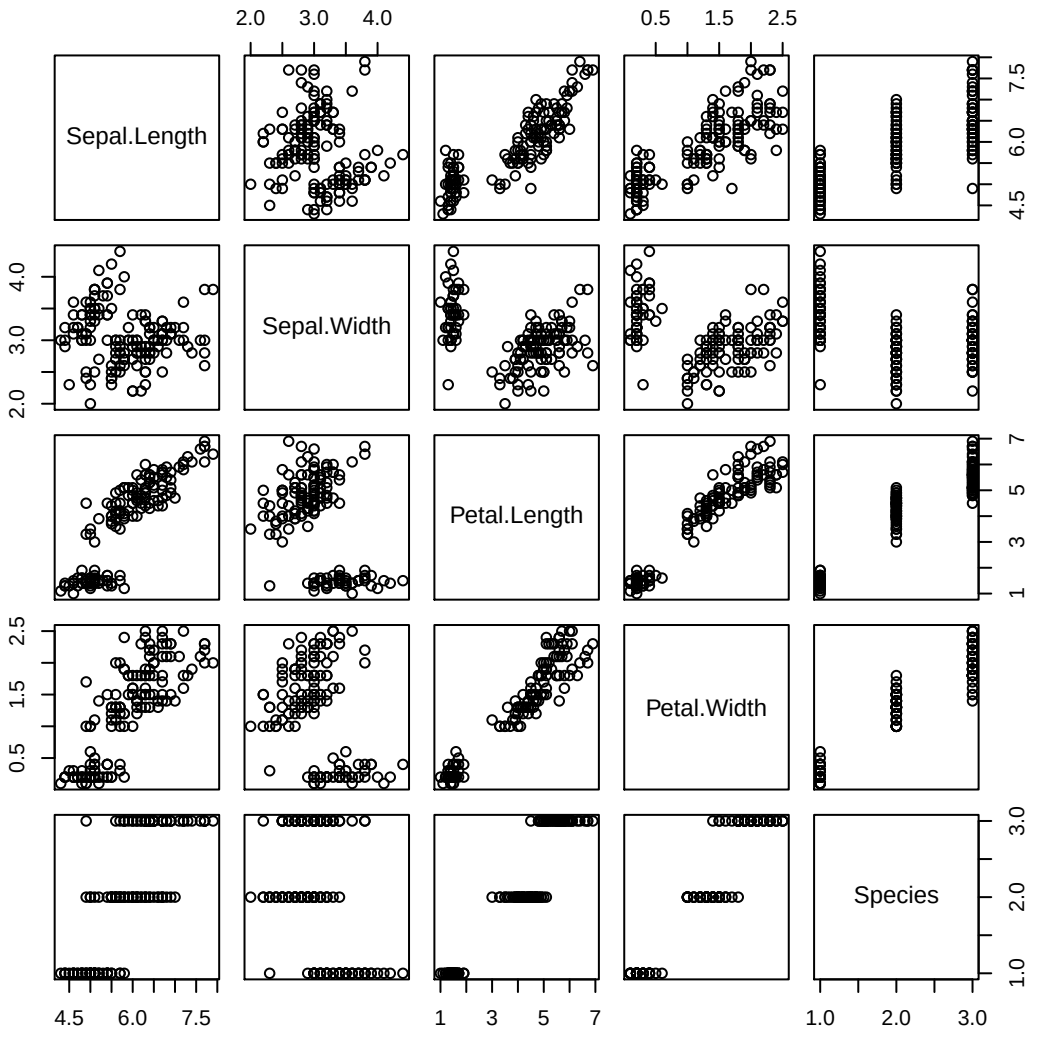

Figure 1: A plain old pairs plot of the iris data.

these counts can be displayed, however. In some cases, it is helpful to visually emphasize the two conditional distributions. In other cases, a display better reflecting the joint distribution may be preferable. The association between a categorical and a quantitative variable may be depicted using a box-and-whisker plot (Tukey 1977) or some variation thereof showing the conditional distribution.[3]

Figure 2 gives a first simple example.[4] For pairs of variables leading to scatterplots or boxplots, the information in the upper and lower diagonals of this particular plot is redundant. However, the mosaic tiles between `sex` and `day` show both of the conditional distributions; the tile in row three, column four gives the distribution of `day` conditional on `sex`, for example. Histograms and bar charts on the diagonal reflect the marginal distributions of the variables. Examples in Sections 3 and 4 illustrate a wide range of refinements of this basic generalized pairs plot; other variations on the theme may be envisioned and implemented by interested researchers.

3 Exploratory Data Analysis and `gpairs`

Our development of the generalized pairs plot follows in the exploratory data analysis (EDA) tradition of John Tukey and John Hartigan. EDA includes a wide range of activities. At the most basic level, what is a data set? In the vast majority of cases, the answer includes a description of the contents of "rows" (cases, observations, subjects, ...) and "columns" (variables, characteristics, measurements, ...) as typically arranged in a spreadsheet. Are there missing values? Are there both quantitative and categorical variables? Such simple descriptions often reveal important features and surprises that may demand attention prior to further analyses.

A summary such as that shown in Table 1 is a good starting point; these data are from the 2010 Environmental Performance Index (Emerson, Esty, Levy, Kim, Mara, de Sherbinin and Srebotnjak 2010). Each of 231 countries may be classified as being landlocked (`LandLock`, having no direct access to an ocean) and as having a high population density (`HighPopDens`). Indices were constructed to reflect overall environmental performance (`EPI`) as

[3] Yes, we could introduce the term fluctuation plot in this paragraph, but that seems implementation-specific and is a subset of the more general term mosaic plot.

[4] Need some background and a proper citation of the tips data set here I think, perhaps with a comment that the plot has shortcomings, including obscuring the rounding of tips?

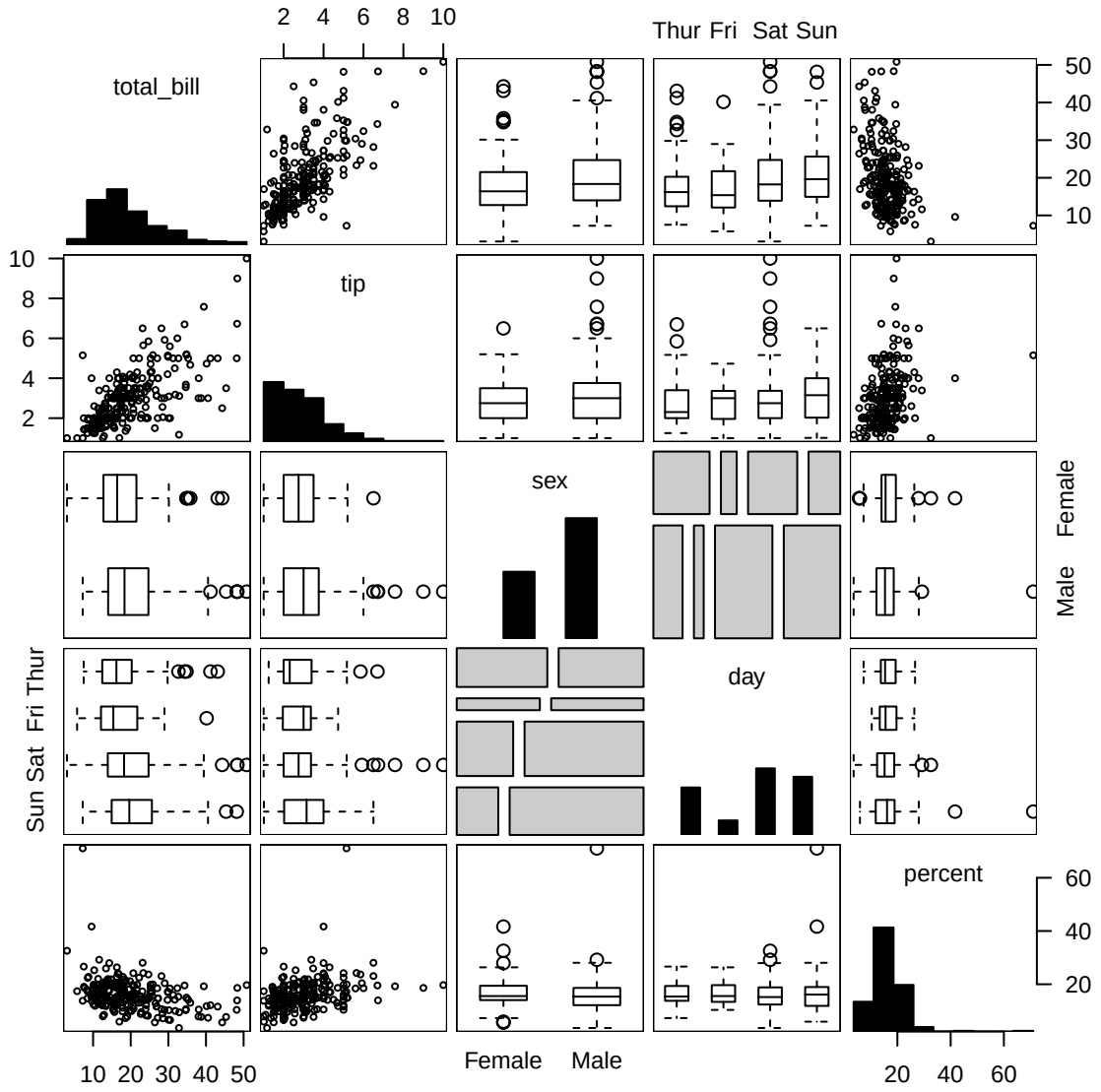

Figure 2: Most basic example?

well as performance on two subcategories, environmental health (`ENVHEALTH`) and ecosystem vitality (`ECOSYSTEM`). The range of observations of the subcategories were coordinated because standardizing the variances led to undesirable features of the aggregate index; the range does not span the interval from 0 to 100 because even the highest-performing country, for example, still has room for improving many aspects of the underlying performance metrics. Unfortunately, missing data prevented the construction of the ecosystem vitality index (and hence the overall environmental performance index) for 68 of the countries; there was less missing information for environmental health (49 countries).

	variable.name	type	missing	distinct.values	precision	min	max
1	`Country`	character	0	231	NA	AFG	ZWE
2	`EPI`	numeric	68	163	1e-08	32.12	93.48
3	`Landlock`	pure factor	0	2	NA	No	Yes
4	`HighPopDens`	pure factor	0	2	NA	No	Yes
5	`ENVHEALTH`	numeric	49	173	1e-08	0.06	95.09
6	`ECOSYSTEM`	numeric	68	163	1e-08	0.06	95.09

Table 1: A summary of a subset of the 2010 Environmental Performance data using the `whatis()` function of R package `YaleToolkit`.

EDA almost always proceeds with tables of categorical variables and univariate graphical displays such as histograms. Bivariate associates are often explored with scatterplots and side-by-side boxplots, as appropriate, with two-way tables and mosaic plots used for pairs of categorical variables. The boxplot shown in Figure 3 provides a standard graphical exploration of the bivariate association between a categorical variable (landlocked status, in this case) and a continuous variable (the environmental health index). Another popular alternative is a pair of stacked histograms, though the choice of histogram bin widths may reveal or obscure information in the conditional distributions.

Emerson et al. (2006) presented the *barcode plot*, originally developed by Hartigan in the spirit of rugs and stripplots (see Chambers and Hastie (1992), for example) and named because of its similarity to the Universal Product Code (UCP) on commercial packaging. Figure 4 shows how the barcode plot can reveal an interesting aspect of the data not evident in the boxplot

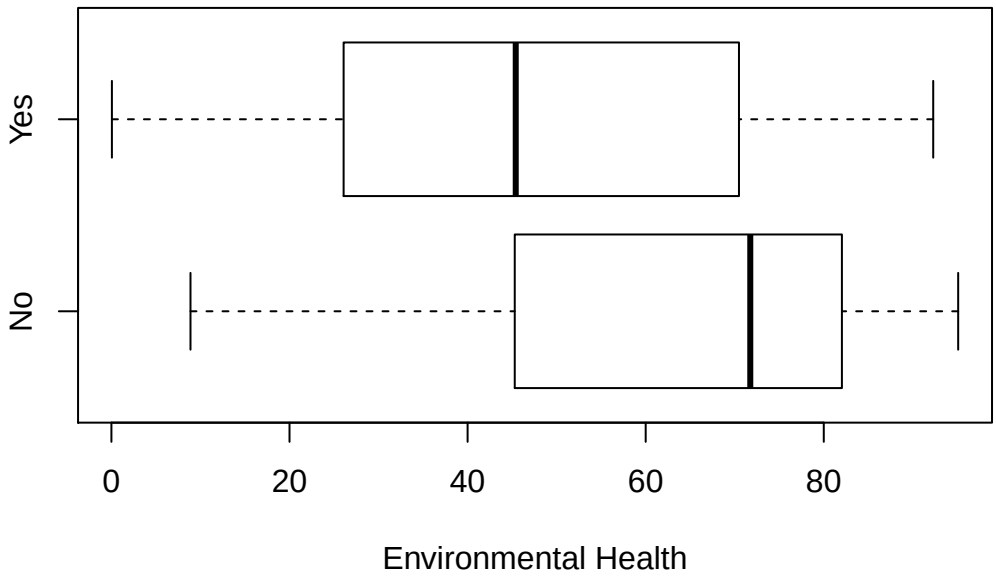

Figure 3: A boxplot example.

(Figure 3) and often obscured by histograms. In this case the tall spike in the bottom-right part of the display leads to the discovery that Germany, Finland, France, Luxembourg, Norway, and New Zealand have identical values of the environmental health index (only Luxembourg is landlocked). In addition, several pairs of countries were tied with similarly strong records of environmental health. The barcode plot exhibits a histogram-like appearance in the presence of ties in quantitative variables, and discovery of such ties often lead to further data exploration.

Our initial development of the generalized pairs plot combined scatterplots, mosaic plots, and the detailed barcode plots with the higher-level summary of traditional boxplots. Figure 5 shows a generalized pairs plot of selected variables from the 2010 Environmental Performance Index using the `gpairs` function of R extension package `gpairs` (Emerson and Green 2010). This particular plot avoids redundant displays, and highlights the United

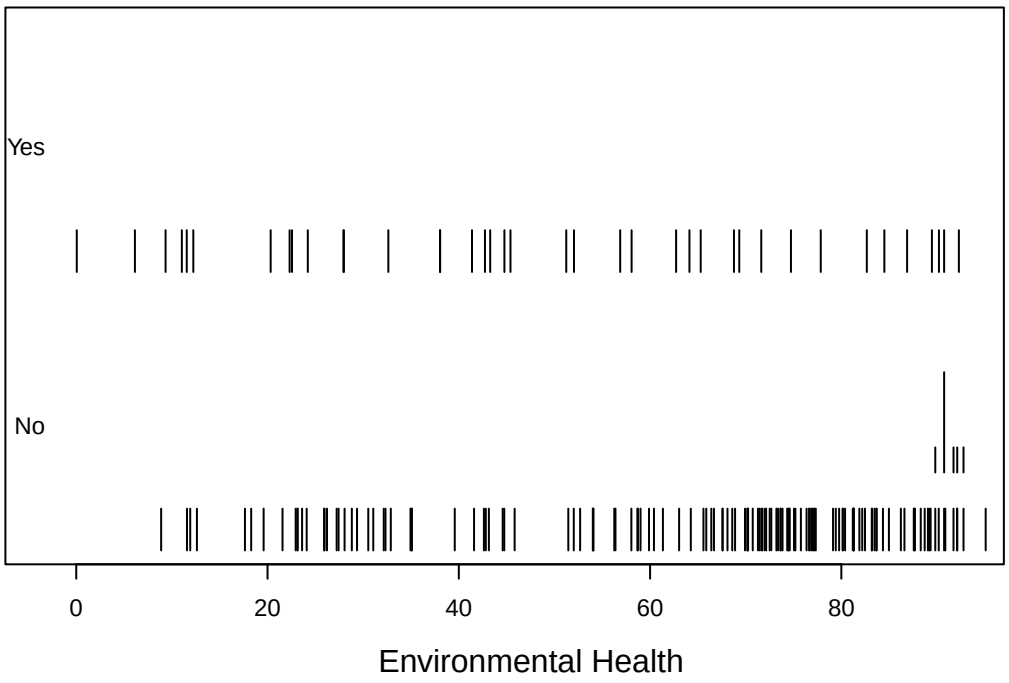

Figure 4: A barcode example. Jay and Walton need to consider adding a `ylab=` option which in this case would be "Landlocked".

States with a larger red point in the scatterplot panels. For pairs of quantitative variables, scatterplots appear above the diagonal. Below the diagonal, correlations (with significance at the 5% level marked with an asterix by default) and numbers of missing values provide additional information, with the shading providing visual reinforcement of the associations. For the two categorical variables (`Landlock` and `HighPopDens`), the corresponding mosaic tiles depict the two conditional distributions; `Landlock` conditional on `HighPopDens` is shown in the third row, second column, for example.

Other options are supported but not shown here. Stripplots may be used in place of boxplots or barcode plots, for example. Points may be customized in scatterplot panels using alternative symbols, sizes and colors for the exploration of high-dimensional patterns. A companion function, `corrgram`, is provided for convenience (see Friendly (2002) for a nice discussion of these plots).

4 An Extension of the Grammar of Graphics and ggpairs

5 Discussion

References

Anderson, E. (1935), "The irises of the Gaspe Peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

Basford, K. E., and Tukey, J. W. (1999), *Graphical analysis of multiresponse data : illustrated with a plant breeding trial*, Boca Raton, Fla.: Chapman & Hall/CRC.

Becker, R. A., Cleveland, W. S., and Shyu, M. J. (1996), "The Visual Design and Control of Trellis Display," *Journal of Computational and Graphical Statistics*, 5(2), 123–155.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, California: Wadsworth International Group.

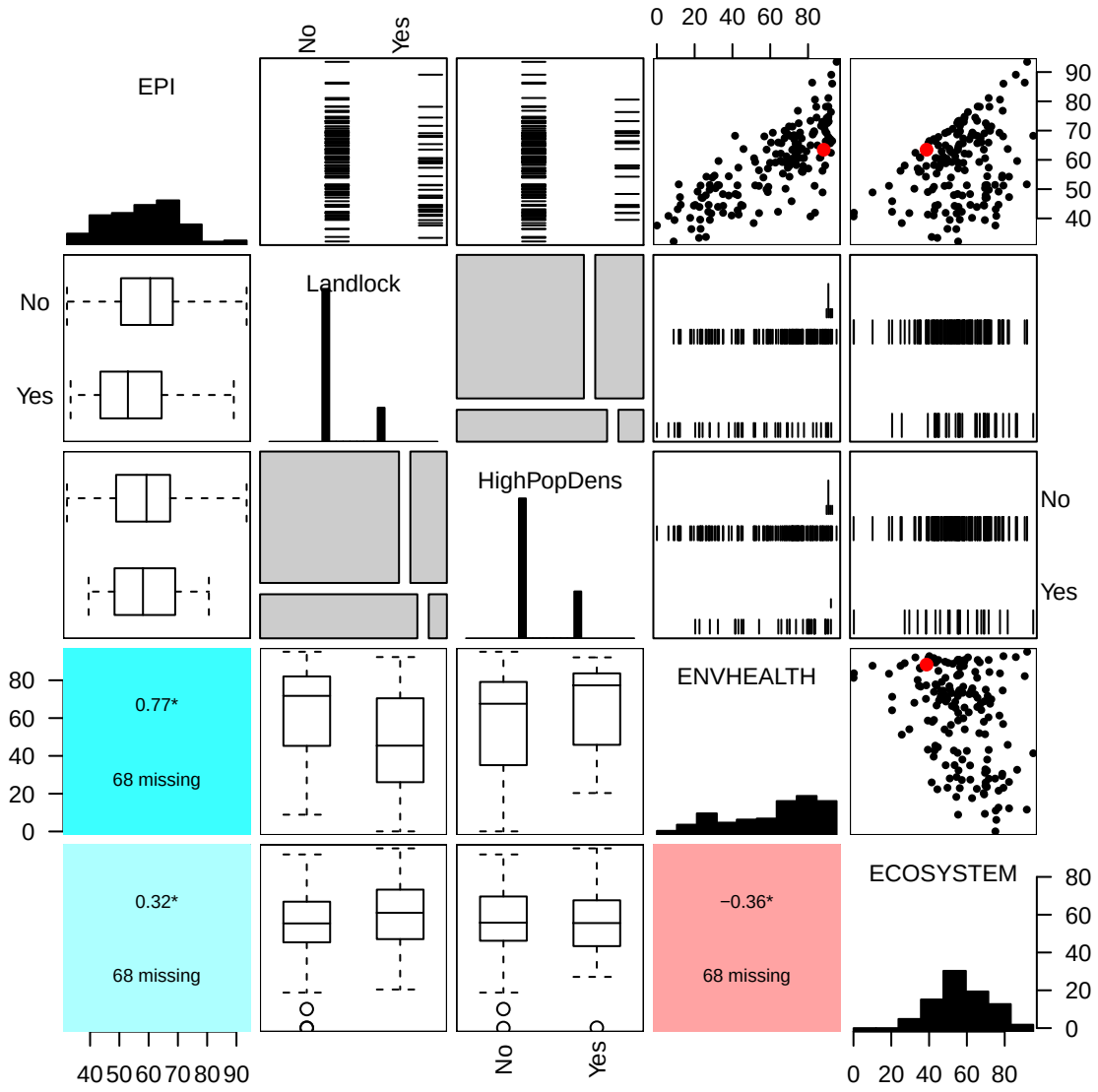

Figure 5: Here's a nice gpairs caption.

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, Pacific Grove, California: Wadsworth & Brooks/Cole Advanced Books & Software.

Chernoff, H. (1973), "The uses of faces to represent points in k-dimensional space graphically," *Journal of the American Statistical Association*, 68(???), 361–368.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Monterey, California: Wadsworth Advanced Books and Software.

Cleveland, W. S. (1993), *Visualizing Data*, Summit, New Jersey: Hobart Press.

Emerson, J. W., Esty, D. C., Levy, M. A., Kim, C. H., Mara, V., de Sherbinin, A., and Srebotnjak, T. (2010), *2010 Environmental Performance Index*, New Haven, Connecticut: Yale Center for Environmental Law and Policy.

Emerson, J. W., and Green, W. (2010), *YaleToolkit: Data exploration tools from Yale University.* R package version 3.2.
URL: *http://CRAN.R-project.org/package=YaleToolkit*

Emerson, J. W., Green, W. A., and Hartigan, J. A. (2006), "Barcodes, Generalized Pairs Plots, and Sparkmats,", UseR! Conference presentation, Vienna.

Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7.

Friendly, M. (1994), "Mosaic displays for multi-way contingency tables," *Journal of the American Statistical Association*, 89, 190–200.

Friendly, M. (2002), "Corrgrams: Exploratory displays for correlation matrices," *American Statistician*, 56(4), 316–324.

Hartigan, J. A. (1975), "Printer graphics for clustering," *Journal of Statistical Computation and Simulation*, 4, 187–213.

Hartigan, J., and Kleiner, B. (1984), "A mosaic of television ratings," *American Statistician*, 38(???), 32–35.

Playfair, W. (1786), *The commercial and political atlas; representing, by means of stained copper-plate charts, the exports, imports, and general trade of England, at a single view. To which are added, Charts of the revenue and debts of Ireland, ... by James Corry*, London: Debrett, Robinson, and Sewell.

R Development Core Team (2005), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Schloerke, B., Cook, D., Hofmann, H., and Wickham, H. (2010), *GGally: Extension to ggplot2.* R package version 0.2.2.1.
URL: *http://CRAN.R-project.org/package=GGally*

Tufte, E. (1983), *The Visual Display of Quantitative Information*, Cheshire, Conn.: The Graphics Press.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison Wesley.

Tukey, P. A., and Tukey, J. W. (1981), "Graphical Display of Data Sets in Three Or More Dimensions," in *Interpreting Multivariate Data*, ed. V. Barnett, Chichester, United Kingdom: Wiley and Sons, pp. 189–275.

Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, New York: Springer.
URL: *http://had.co.nz/ggplot2/book*

Wilkinson, L. (1999), *The Grammar of Graphics*, New York: Springer-Verlag.