Problemy zarządzania, vol. 7, nr 4 (26): ...–...
ISSN 1644-9584, © Wydział Zarządzania UW

Visualizing Data, Visualizing Models: Getting Priorities Right When Analyzing Data

Alex Chavez and Richard Gonzalez

We discuss methods of visualizing data and statistical models. The key innovation underlying these methods is that specialized graphs illustrate underlying patterns in data and they illustrate the properties of the statistical model that are used to analyze the data. These visualization techniques provide ways of understanding the underlying relation between variables, can illustrate how the statistical model works and can illustrate how the statistical model may not be capturing the underlying patterns among the variables.

1. Data analysis as conducted in the social sciences

Too often analyses are taught as a set of rules one needs to follow. Examples of such rules are "check that your data are normally distributed," "if you have tables of counts, use a Chi-Square contingency test, and "a factorial design should be tested with ANOVA." Most statistics classes are preoccupied with the details of conducting the statistical test so students naturally assume that data analysis is about those details. For example, students can come away from a statistics course believing that data analysis is exclusively about answering questions such as: Should I use a one or two tailed hypothesis test? How should I adjust the p-value because I'm conducting multiple tests? Am I using an unbiased estimate? Those students carry such an emphasis on to professional positions in academics, industry and government.

Of course these are all important aspects of data analysis, but the point of this paper is that they are not the only aspects of analyses, nor are they the most important ones. We argue that such emphasis on the "rules of statistical analysis" that is present in statistics classes and in our research journals is partly misplaced. Instead, we should be teaching other skills in statistics courses and demanding additional criteria in our published papers. For example, one important skill is how to convert a business question into an empirical question that can be assessed with data. This skill requires identifying the key aspects of the business question, identifying how to

address those aspects with variables that can be observed, deciding how to collect relevant data and what sources to use, deciding how to organize those data and analyze them in a manner that can inform practice, and making the proper inferences from those data in a way that informs the original business question. Likewise, an analogous example occurs in the research setting where skill is needed to convert a research question into an empirical question, and decisions such as how to operationalize theoretical variables are common.

Business reports and academic articles should emphasize the description of the findings and their status to the original business question or academic question that led to the result. That is, the results need to be presented clearly so that the reader can understand how the data inform the original business or research question. Our point is not that statistical inference or statistical estimation should be ignored in business and research reports. Our argument is that the description of the results should be the primary information that is emphasized in the report. For instance, a good sentence in a results section could be "The participants who received the persuasive message tended to purchase more items during the observation period (mean = 3.4, sd=1) than participants who received the neutral message (mean = 1.2, sd=1.1), $t(543) = 2$, $p < .001$." Note how the statement of the t-test and the p-value merely punctuate the sentence to give a "stamp of approval" that the observation is statistically significant, which merely means it has met the conventional criterion in one's field. Compare with the usual way that results sections are written: "An independent means two sample t test was conducted on the data. Results showed that the means in the two groups were statistically different, $t(543) = 2$, $p < .001$."

In this paper we present examples from relatively recent developments in the statistical literature surrounding the problem of visualization of data. Our own contribution is to add the visualization of the statistical model, and to superimpose the model in the same visualization as the data. Two simple examples of this idea are (1) the well-known practice of displaying the regression line (the model) along with the scatterplot (a representation of the raw data) and (2) the practice of displaying a normal distribution (the model) superimposed over a histogram (a representation of the raw data). We show that this type of model and data plot can be extended to more complicated cases and jointly can illustrate properties of model and data that are not easy to grasp with other methods.

In the following subsections we present a few examples of visualization techniques, which the reader can reproduce using the statistical program R (www.R-project.org) and the code contained in the appendix.

2. Plotting symbols

Interactions involving categorical variables that are difficult to detect through regression analyses often can be detected easily through the manipulation of plotting symbols. Chu (2001) analyzed a data set (http://www.amstat.org/publications/jse/v9n2/4C.data) that included 308 diamond prices by carat weight (numeric), color (ordinal), clarity (ordinal), and certifying agency (nominal). Analysis began with a scatterplot of price vs. carat weight and was followed by a linear regression of log-price on the four predictors mentioned above. Although the successful pricing models were developed following regression diagnostics, additional data visualization at any stage of the analysis would have yielded a qualitatively different conclusion about the role of the certifying agency. Figure 1 uses different plotting symbols to represent the third dimension of certifying agency on the bivariate scatterplot of log-prices vs. carat weight. Strong clustering patterns emerge, indicating an interaction between certifying agency and carat weight.* The 40 diamonds weighing less than 0.3 carats were all certified by the IGI, and had prices that rose steeply with carat size. On

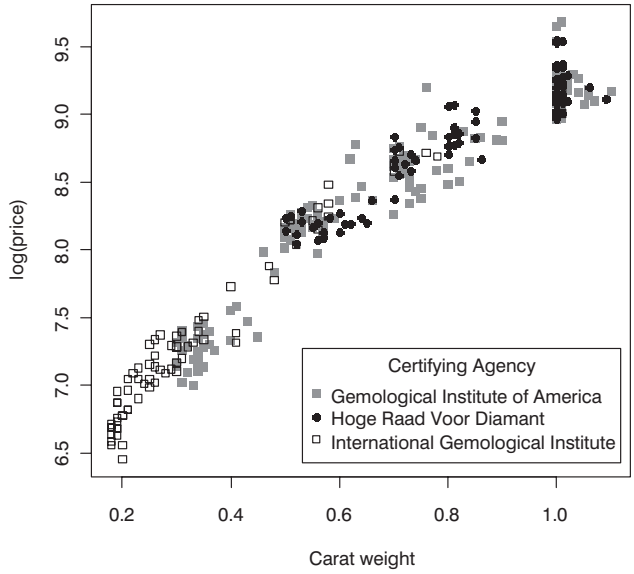

Figure 1. Log-prices by carat weight and certifying agency.

* $F(2) = 81.11$, $p < .0001$. Adjusted-R^2 increased from .97 to .98 when adding Certifying Agency x Carat Weight to the initial model, implying that a relatively large portion of the remaining variance was explained by this interaction.

the other hand, the 79 diamonds that Hoge Raad Voor Diamant certified weighed a minimum of 0.5 carats, and had prices that rose less steeply with carat weight. Related techniques include replacing plotting symbols with numbers corresponding to the levels of an ordinal variable, and using color shades, grayscales, or symbol sizes proportional to the values of a numerical variable. These techniques can be used in situations ranging from regression diagnostics (e.g., residual or leverage plots) to visually representing clustering solutions.

3. A richer scatterplot/correlation display

It is common for datasets to have many variables. Typically a data analyst will (1) examine a histogram to check the shape of the distribution of each variable, (2) examine the correlations, (3) examine the scatterplots to check for linearity, outliers and patterns suggesting violations of the equality of variance assumption, and (4) run regressions (sometimes nonparametric or splines) in order to model the relation between pairwise variables. The next plot we review puts all that information in one display. While to the untrained eye the display may be dense and confusing, it does not take more than several seconds to navigate through the display and immediately see the value of putting all that information into one bundle. To illustrate we use a real estate data set of 225 commercial properties for lease (http://www-stat.wharton.upenn.edu/~buja/STAT-541/real-estate-data.txt). There are 21 variables, of which we focus on annual rent per square foot, years since renovation, building age in years, distances to city center and airport in miles, the fraction of office units that are rented in the building, location (city, old suburb, new suburb), and the length of the lease contract in years.** Figure 2 displays a subset of these variables with histograms and density fits along the diagonal, correlations above the diagonal, scatterplots with regression fits below the diagonal, and color-coding of the plotting symbols to represent location. Rent prices increase with occupancy rates, building age, and years since the last renovation. The latter two findings can be explained by the fact that newer or recently renovated properties tend to be in the suburb, whereas older properties are in the city. The plot reveals other interesting findings, such as the presence of at least two airports (as distance to the airport increases, proximity to the city either increases or decreases based on which suburb the property is in), and again highlights the usefulness of representing an additional dimension using plotting symbols.

This type of plot is easy to produce in R and provides an excellent way of initially visualizing a data set. A variant of the plot was proposed by Friendly in his corrgram plot (2002), which uses ellipses in a correlation matrix where rows and columns have been reordered to have similar correlations in adjacent cells.

** We omitted 3 outliers that had extreme values of rent per square foot.

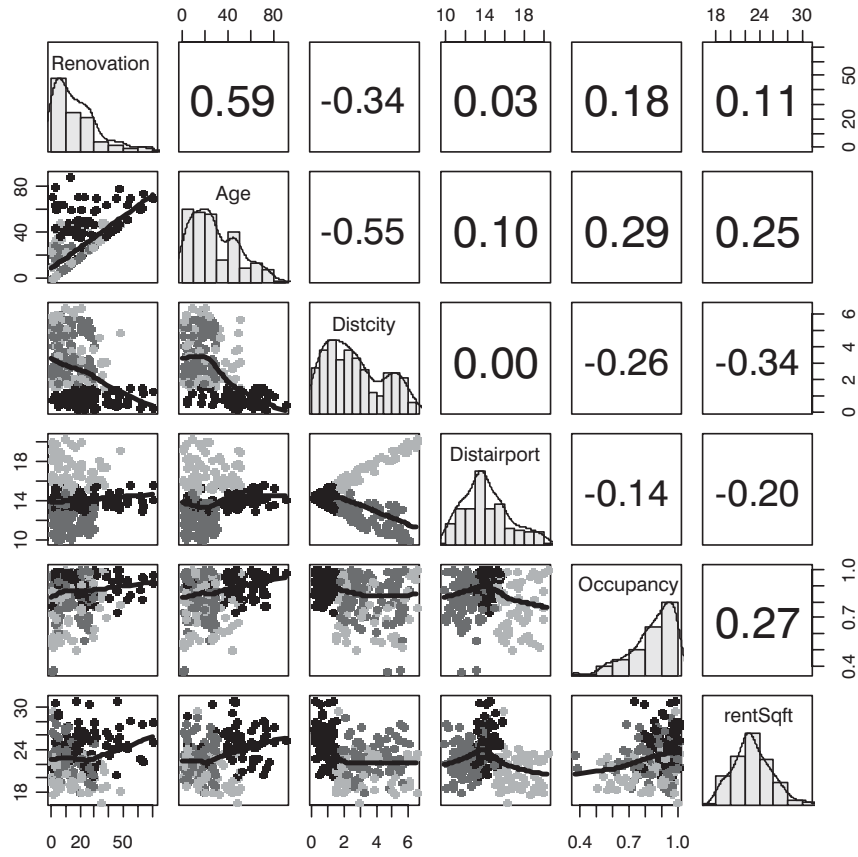

Figure 2. A scatterplot/correlation matrix with univariate summaries along the diagonal. Grey scales represent the location (light-gray = old suburb, medium-gray = new-suburb, black = city).

4. Lattice plots

All too often statistical analysts tend to cut a continuous variable into discrete categories to aid in statistical analysis and interpretation. For example, a continuous variable is dichotomized into a low and a high group using, say, a median split. There has been some attention given to the effects of catergorization of continuous data in the literature, but the issue is far from resolved. The lattice, or trellis, system of graphing data allows the analyst to maintain the original continuous form but also make the interpretation more straightforward by using a particular form of cut. The lattice plot in Figure 3 illustrates a clever way of exploring the relation between rent and lease length as a function of the conditioning variable

of occupancy rate. Note that the conditioning variable is "cut" with overlapping regions, which are displayed as rectangles that appear at the top of the Figure. A scatterplot is presented in each panel that describes the raw data. Each scatterplot also includes a model fit, which in this example is a nonparametric regression curve using a loess estimator, thus showing the model and data together.

The plot reveals that properties with longer leases offer a per-annum discount, but that this effect depends on the occupancy rate of the building. Buildings whose offices are nearly all rented offer the steepest discounts, albeit at higher average rents. Having an abundant supply, low occupancy buildings might be offering their properties at bottom line prices. Moreover, whereas Figure 2 suggested that many of the relations between rent and the other variables were explained by location, this does not appear to be the case in the present analysis. The horizontal homogeneity of gray scale in Figure 3 within each scatterplot and across scatterplots suggests that location respectively explains neither the relationship between rent and lease length nor the interactive effect of occupancy rate and lease length on rent.

The complicated and systematic relations obvious from this rich display would not necessarily emerge from more traditional reporting mechanisms that appear in business reports and academic journals. For instance, the analyst might report the results of a regression using lease length and occupancy as two covariates, thus "controlling" for the additive effects of those covariates. There has been much critique in the recent statistics literature about such simple approaches to statistical adjustment of covariates. The lattice plot allows a more straightforward way to assess the relation between variables, including covariates and confounders. For more information on the lattice plot see Sarkar (2009).

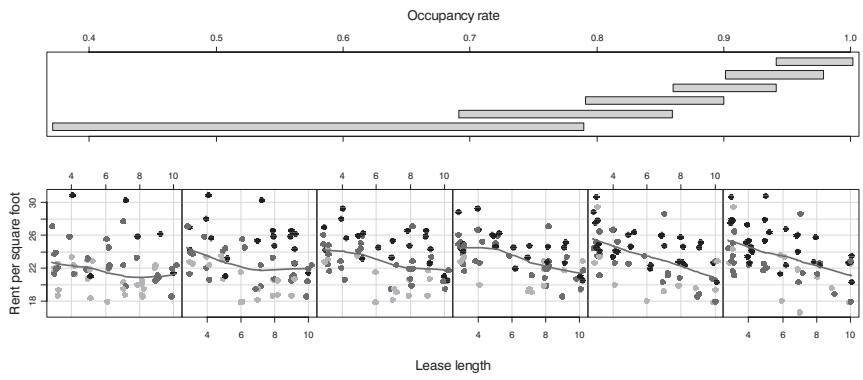

Figure 3. A lattice plot for the rent per square foot by lease length, percentage of building occupancy, and location (light-gray = old suburb, medium-gray = new-suburb, black = city). Points are moved slightly vertically and horizontally at random for ease of visualization.

5. Clustering and PCA

Data analysis sometimes progresses into the realm of reducing the dimensionality of a large number of variables in a dataset. This can be accomplished through principal components analysis (PCA) and related factor analytic models, through clustering of variables, or through clustering of subjects (as in research questions surrounding market segmentation).

Visualization can do much by way of informing the standard PCA and clustering reporting. We illustrate with a marketing data set (http://www-stat.wharton.upenn.edu/~buja/STAT-541/mktg.dat) analyzed by Lang & Swayne (2001) of approximately 2000 households surveyed by AT&T about their telecommunication needs on 9 variables such as their needs for certain product classes, the durations of incoming and outgoing calls, and demographic variables. The primary research goal of this analysis was to provide information about market segmentation – that is, to find any identifiable clusters of consumers. Identifying such clusters is important for marketing because advertisements can be tailored specifically for a particular segment. The modern approach to clustering and market segmentation analysis is to cluster based on patterns in the data rather than say cluster according to "known groups," such as males between the ages of 18 and 29. Moreover, it is not clear that clustering according to demographics always provides clear market segments.

Following Lang & Swayne (2001), we used a K-means clustering approach. Figure 4 presents a clear segmentation in the space of first two principal components. The first principal component loaded on primarily on call volume, followed by product need, youth, and education, whereas the second principal component loaded primarily on youth, followed by product need and inverse call volume. Although four clusters were used, clear segmentation is achieved with two or three clusters as well. Marketers would be interested in selling the products whose needs were measured to the rightmost and top clusters in the four-cluster solution, which correspond to respectively high and low call volume users. These clusters could be examined against other demographics such as gender, age, and education level. This type of analysis extends more traditional analyses in which demographic variables are entered as predictors in a regression and the analysis "rejects" or "fails to reject" (i.e., according to the statistical test) whether that variable predicts the dependent variable controlling for other predictors in the regression. Clustering techniques can be particularly useful when actual market segments do not correspond to any known market segments, and this could provide useful information for the market researcher who is seeking to find comparative advantage over a competitor's marketing campaign.

There are various types of clustering techniques that are available in R, and many packages include corresponding visualization functions. Some useful R packages include cluster, Mclust, flexmix, flexclus, and homals. Some of these algorithms impose more structure on the data than others. Research in the general problem of clustering and what are called "mixture models" is currently very active so we expect to see many new techniques in the short term.

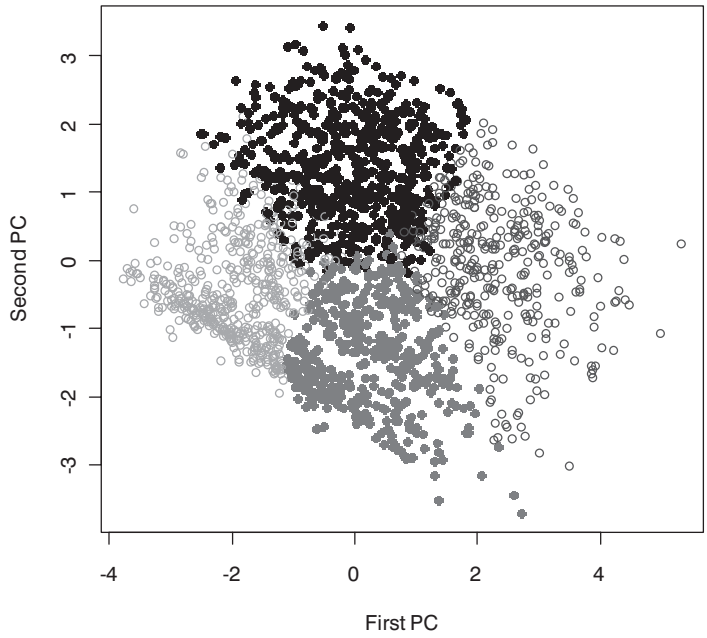

Figure 4. A four-means clustering solution for the marketing data.

6. Three Dimensional Visualization

One advantage offered by the power of computing is making three-dimensional (3d) visualization more feasible. At this point complete mouse-interactive 3d data plots require special programs to be installed on one's computer, but it is possible to use computer languages such as Java to develop specialized interactive plots that display within traditional web browsers such as Internet Explorer and Mozilla Firefox without requiring specialized software. Other ways of visualizing more than two dimensions can make use of animated gif files or movie files where a three dimensional plot can be rotated through the animation so the third dimension can be visualized.

An advantage of higher dimensional plots is that they permit the joint display of the model and data when there are several predictors. We have argued throughout this paper that it is informative to plot simultaneously the data and the model because it illustrates the reduction in dimensionality offered by the model and also how well the data conform to that reduction.

We illustrate with a relatively complicated example involving a seemingly unrelated regression, also known as an actor-partner interaction model in the psychological literature. This type of regression can arise when, say, data from two employees working together are jointly modeled, such as in an organizational psychology or management setting when collecting worker satisfaction from members of a work team or when collecting data from a worker and their supervisor. Let's imagine we want to understand supervisor and employee satisfaction at time t given the same two variables at time (t-1). There could be actor effects (e.g., supervisor's data at the earlier time predicts the supervisor's data at the next time point) and partner effects (e.g., employee's data at the earlier time predicts the supervisor's data at the next point). The regression paths have useful interpretations because they permit assessment of how an individual's satisfaction is predicted both by his own as well as his "partner's" previous level of satisfaction. What makes this model special and complicated is that there are in a sense two regressions that are computed (one for the supervisor's time t data and another for the employee's time t data), but these two regressions are related because they come from a matched pair of individuals–supervisor and employee. Correlated residuals are added to the model to account for such dependencies in what would otherwise be two separate and independent regression models. Such a model can highlight important processes in the dyadic relation because it allows a decomposition of the actor effects (self data predicting self data) and the partner effects (partner data predicting self data).

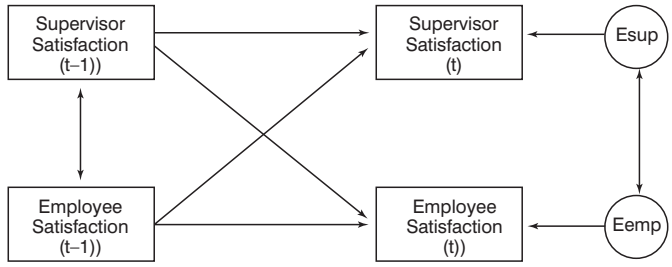

Figure 5. Actor-partner interaction model, also known as a seemingly unrelated regression model. There are two residuals, residuals for the supervisor and residuals for the employee, which are correlated.

Figure 6 shows a three dimensional plot of hypothetical data for this supervisor/employee satisfaction example. Note that there are four observations per dyad (the four rectangles in Figure 5). We display these four observations in a 3d scatterplot by defining two dimensions (X and Y axes) to correspond to supervisor and employee time (t-1) data and the third dimension (the vertical Z axis) to correspond to the time t data. We code the supervisor and employee with different gray scale values. Further, to illustrate which data come from a supervisor-employee dyad we connect the two time t data points from the same dyad with a vertical line segment. This allows a visual connection that two time t observations arose from the same dyad. Note that once the X and Y coordinates for time (t-1) are known for a given dyad the two corresponding points must appear in a vertical column above the X-Y point. In this way we can display 4 observations in three dimensions. The model in Figure 6 implies two planes, one for the supervisor and one for the employee, showing the role of the two additive time (t-1) predictors.

Information becomes clearer by including both the data and the model in the same 3d graph. One sees that the data may suggest a curvilinear

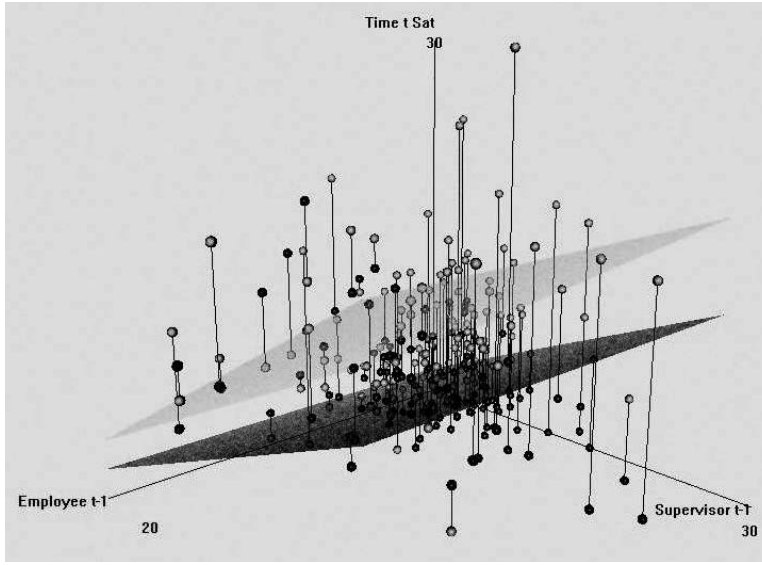

Figure 6: Seemingly unrelated regression (also known as the actor-partner interaction model). Two regression surfaces are plotted, the employee (light gray) and the supervisor (dark gray). Likewise, the X and Y dimensions represent employee and supervisor data at time (t -1) where as the satisfaction from employee and supervisor at time t appear on the Z axis. Note that vertical pairs of points are joined by a line segment and represent an employee-supervisor pair.

relation so the planes from the simple actor partner model may not be appropriate. But this is relatively easy to fix by including, say, polynomials in the regression equations. There are also other patterns that this representation shows, such as the data suggest that both the satisfaction of the employee and the supervisor are relatively low given the clustering of data near the low back corner corresponding to the vertex (0,0,0).

7. Conclusions

Our main point is that there is now a new emphasis on proper visualization of data and model together. Given that the primary objective of data analysis is to provide an answer to the original research question that motivated the data collection, we believe that such developments in visualization are a step in the right direction. Visualization helps one assess the data and model together in a manner that keeps in focus the original research question. This contrasts with the standard approach of placing concerns of statistical inference and estimation front and center. Our position is that the concerns of statistical inference and estimation should take on the role of a supporting actor rather than the lead. Visualization is one tool that helps us get our priorities right. The internet makes these tools accessible and there are wonderful open source tools available in the very sophisticated statistical package R. These new visualization techniques highlight patterns in data, making discovery of underlying relations between variables and groups more likely. The techniques can also provide more information about the relation of data to model and provide a dramatic way to communicate scientific findings that may be more accessible to a broader audience.

Authors
Alex Chavez and **Richard Gonzalez** – Department of Psychology, University of Michigan, USA.

References

Chu, S. 2001. Pricing the C's of diamond stones. *Journal of Statistics Education*, Vol. 9, Number 2.
Available online: http://amstat.org/publications/jse/v9n2/datasets.chu.html.
Friendly, M. 2002. Corrgrams: Exploratory displays for correlation matrices. *American Statistician*, Vol. 56, p. 316–324.
Lang, D.T. & Swayne, D.F. 2001. *GGobi meets R: an extensible environment for interactive dynamic data visualization*. Proceedings of the 2nd International Workshop on Distributed Statistical Computing.
Sarkar, D. 2009. *Lattice: Multivariate data visualization with R*, Springer.

Appendix: R Code

```
#Data set source (Singfat Chu): http://www.amstat.org/publications/jse/v9n2/4C.dat
#Read the data in from the web, name the variables
diamond <- read.table("http://www.amstat.org/publications/jse/v9n2/4C.dat", header=F)
names(diamond) <- c("carat", "color", "clarity", "certification", "price")
#Ensure that the factor levels are arranged from worst to best, but treat them nominally
diamond$color <- factor(diamond$color, levels=c("I","H","G","F","E","D"), ordered=F)
diamond$clarity <- factor(diamond$clarity, levels=c("VS2","VS1","VVS2","VVS1","IF"),
ordered=F)
#Plot the data
windows(6,6)
plot(
    log(price) ~ carat,
    xlab="Carat weight",
    data=diamond,
    type="p",
    pch=c(15,16,22)[as.numeric(factor(diamond$certification, labels = 1:3))],
    col=c(gray(.5), gray(0), gray(0))[factor(diamond$certification, labels = 1:3)]
)
#Add a legend
legend(
    "bottomright",
    inset=.03,
    title="Certifying Agency",
    pch = c(15,16,22),
    col = c(gray(.5), gray(0), gray(0)),
    legend=c("Gemological Institute of America", "Hoge Raad Voor Diamant",
"International Gemological Institute")
)
#Check the certifying agencies of diamonds weighing less than .3 carats
diamond[diamond$carat < .3, "certification"]
#Count the number of diamonds weighing less than .3 carats
sum(diamond$carat < .3)
#Count the number of diamonds certified by HRD
```

```
sum(diamond$cert == "HRD")
#Check the minimum carat weight certified by HRD
min(diamond[diamond$cert == "HRD", "carat"])
#Linear regression models, F-test, and model summaries
mod1 <- lm(log(price) ~ color + clarity + certification + carat, data=diamond)
mod2 <- lm(log(price) ~ color + clarity + certification * carat, data=diamond)
anova(mod1,mod2)
summary(mod1)
summary(mod2)


#Data set source (Andreas Buja): http://www-stat.wharton.upenn.edu/~buja/STAT-541/real-estate-data.txt
rest <- read.table("http://www-stat.wharton.upenn.edu/~buja/STAT-541/real-estate-data.txt", header=T)
rest$rentSqft <- rest$Renttotal/rest$Sqft
summary(rest)
#Omit a few properties with extreme rent per square footage values
rest <- rest[rest$rentSqft <= quantile(rest$rentSqft, .99),]
#Customized version of the scatterplot/correlation matrix for the real estate data set.
Redefines panel.smooth, panel.hist.density, and panel.cor to chage plotting symbols and
line width.
panel.smooth.fixed <- function (x, y, col = par("col"), bg = NA, pch = par("pch"), cex =
1, col.smooth = "black", span = 2/3, iter = 3, ...)
{
    points(x, y, pch = pch, col = col, bg = bg, cex = cex)
    ok <- is.finite(x) & is.finite(y)
    if (any(ok))
        lines(stats::lowess(x[ok], y[ok], f = span, iter = iter), lwd = 3,
            col = col.smooth, ...)
}
panel.hist.density.fixed <- function (x, col = NA, ...)
{
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5))
    h <- hist(x, plot = FALSE)
```

```
    breaks <- h$breaks
    nB <- length(breaks)
    y <- h$counts
    y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = gray(.9), ...)
    d <- density(x, na.rm = TRUE)
    d$y <- d$y/max(d$y)
    lines(d)
}
panel.cor.fixed <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r = (cor(x, y,use="pairwise"))
    txt <- format(c(round(r,digits), 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex )
}
windows(12,12)
pairs(
    rest[,c(3,5:7,10,22)],
    lower.panel = panel.smooth.fixed,
    diag.panel = panel.hist.density.fixed,
    upper.panel = panel.cor.fixed,
    col = c(gray(0), gray(.3), gray(.65))[1*(rest$Location == «CITY») + 2*(rest$Location
== «SUBNEW») + 3*(rest$Location == «SUBOLD»)],
    pch = 16
)
#A lattice/conditioning plot
windows(12,5)
coplot(
    jitter(rentSqft) ~ jitter(Leaselength) | jitter(Occupancy),
    xlab = c("Lease length", "Occupancy rate"), ylab = "Rent per square foot",
```

```
    panel = panel.smooth,
    #We use three shades of gray to represent location
    col = c(gray(0), gray(.3), gray(.65))[1*(rest$Location == "CITY") + 2*(rest$Location
== "SUBNEW") + 3*(rest$Location == "SUBOLD")],
    pch = 16,
    cex = 1.5,
    lwd = 2,
    rows = 1,
    data = rest
)
#Data set source (Andreas Buja): http://www-stat.wharton.upenn.edu/~buja/STAT-541/
mktg.dat
#Analyzed by Lang, D. T. & Swayne, D. F. (2001).
#Note: The clustering solution is sensitive to initial conditions. Run the code several times
if you initially do not achieve a satisfactory solution.
mktg <- read.table("http://www-stat.wharton.upenn.edu/~buja/STAT-541/mktg.dat",
header=T)
p <- dim(mktg[,-9])[2]
mktg.scale <- scale(mktg[,-9],scale=T,center=T)
rn <- matrix(rnorm(p*4), ncol=p)
mktg.kmeans <- kmeans(x = mktg.scale, centers=rn)
mktg.clusers <- mktg.kmeans$cluster
mktg.prcomp <- prcomp(mktg.scale)
mktg.prcomp
X <- mktg.prcomp$x
plot(
    X[,1], X[,2],
    xlab="First PC",
    ylab="Second PC",
    pch=c(16,21,16,21)[mktg.kmeans$cluster],
    col=c(gray(0), gray(.2), gray(.4), gray(.6))[mktg.kmeans$cluster]
```