Using SAS/INSIGHT ® Software as an Exploratory Data Mining Platform
Robin Way, SAS Institute Inc., Portland, OR

ABSTRACT

Data mining has captured the hearts and minds of business analysts seeking a solution for exploring and modeling vastly larger, more complex and less well-behaved datasets. Exploratory data analysis, typically consisting of activities like statistical visualization, hypothesis generation, and introductory model fitting is a vital first step in any successful data mining venture. Exploratory data analysis produces direct benefits for data miners in enhanced understanding of data, improved clarity and confidence of the modeling results, and avoidance of pitfalls early in the process. SAS/INSIGHT is SAS Institute's leading software for facilitating exploratory data analysis featuring a uniquely visual analytic toolset. This paper will review the usefulness of SAS/INSIGHT software for exploratory data analysis, interactive regression modeling, and advanced multidimensional data visualization. Along the way, we'll explore how to win baseball games and still save enough money to build the stadium's new luxury boxes!

INTRODUCTION

This paper introduces the usefulness of exploratory data analysis through techniques including single and multidimensional visualization, regression modeling and multivariate analysis. Using major league baseball batting and fielding statistics for actual players, the paper also demonstrates the use of SAS/INSIGHT software to facilitate the application of these exploratory data analysis and modeling techniques.

BASEBALL STATISTICS AS AN EXPLORATORY DATA ANALYSIS SCENARIO

Imagine you are the Director of Franchise Operations for the local major league baseball team: the invisible part of the team that decides how to balance the franchise's investments in recruiting and keeping talent, advertising and public relations, concessions, and stadium development. Marketing reports that fans are pretty happy with the beer and hot dogs, with the new retractable dome, and with television coverage. Attendance at games is strong, as is the pitching staff. You've got a manager that motivates and challenges the team. However, some of the recently hired free agents aren't producing as many home runs and runs batted in (RBIs) as expected, and your win/loss record is suffering. What's more, satisfaction is down among your season-ticket holders because all your existing luxury boxes are sold out for the next ten years, and funds are tight due to the free agent salaries.

In summary, you've got to win more games to keep the fans happy, but you can't go spending an arm and a leg for new talent. The data mining objectives are consequently:

- Hire more effective batting and fielding talent
- Win more games by bringing in more runs
- Keep new salary costs down

The objectives for this project are to select a subset of players to recruit who offer better than average value for their current salary. In order to accomplish this end result, we will have to first get a lay of the land, and understand what factors contribute to player salaries. Then, by comparing what players should be paid with their actual salaries, we can identify our set of recruits.

The dataset for this project is SAMPSIO.DMABASE, from the sample library shipped with Enterprise Miner™ software. A similar example is available in the dataset SASUSER.BASEBALL that is shipped with version 6.12 of the SAS ® System. This dataset contains 1986 single season and career statistics for selected Major League Baseball players (not including pitchers). A few simple adjustments were made to this dataset. First, I created new variables for current season batting average and career batting average (i.e., both are equal to number of hits divided by number of at-bats for the respective timeframe). Next, I added the team name to each observation to clarify the differences between multiple teams in the same city. Third, I've color-coded all players' observations depending on whether they play infield, outfield, or other (e.g., utility players and designated hitters).

EXPLORATORY DATA ANALYSIS

Exploratory analysis helps us get our footing in a new dataset by analyzing variable distributions and inter-variable relationships.

ONE-DIMENSIONAL ANALYSIS

Histograms and boxplots are a quick way of describing a continuous scale variable's distribution, and enable the analyst to quickly assess the distribution's moments and quantiles. Nominal scale variables can be analyzed using mosaic charts, which are essentially visual versions of a frequency table.

A SAS/INSIGHT software technique called "brushing" allows the analyst to select portions of one variable's distribution with the mouse, and instantly view in a different window how the observations underlying the selection are related to the distribution of another variable. The brush can be re-sized and moved across any variable's chart, enabling a real-time analysis of variable relationships.

A histogram of player salaries indicates a distribution skewed to the right; a few players receive high salaries while most (72%) make less than $1 million annually. Another histogram indicates career length (i.e., years in the major league) is also skewed to the right, but to a less severe degree. There is a distinct decline in the distribution after seven years (see Figure 1).
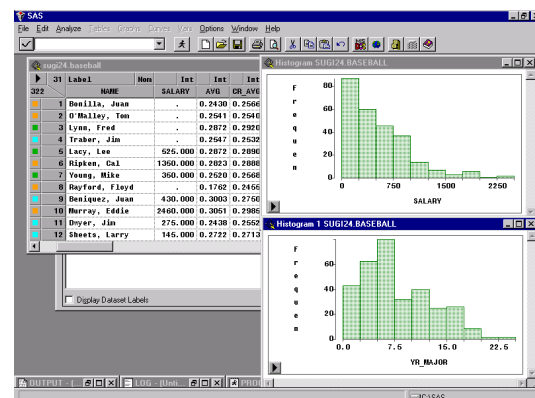

Figure 1

It is not necessary to request one histogram at a time. We have constructed two more histogram charts, each containing four variables in their own section of the window. This allows the

analyst to view distributions of multiple variables quickly and in a single window, which is useful when in exploratory mode. The first histogram window shows variable distributions for 1986 single-season batting statistics (i.e., batting average, runs, home runs, and runs batted in). A second histogram window shows distributions for the same variables representing career statistics for the same measurements. Distributions of corresponding variables are similar between single-season and career statistics. Batting averages follow a roughly normal distribution, while other batting statistics (counts, rather than ratios) are skewed to the right. The career statistics for runs, homers, and RBIs are more sharply skewed, suggesting players with distinctively successful batting performance over their career number relatively fewer than those who hit well in a single season (see Figure 2).
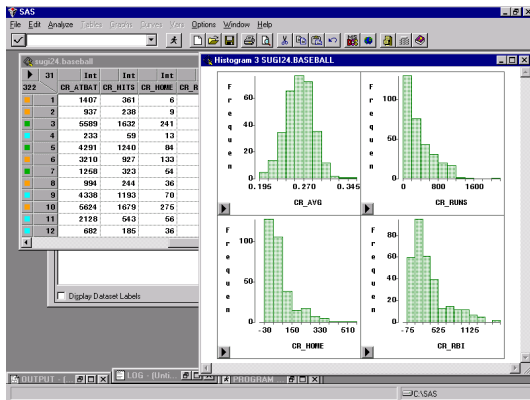

Figure 2

By brushing the salary distribution, we can see players with high salaries tend to have high numbers of RBIs but don't stand out in terms of batting average. However, career statistics do identify a distinct group of highly paid players who have high career batting averages.

TWO-DIMENSIONAL ANALYSIS

Two-dimensional charts allow the analyst to visualize inter-variable relationships in a single chart, instead of brushing one chart and viewing relationships in another window. There are a variety of two-dimensional charting techniques in SAS/INSIGHT software, including scatter and line plots, as well as two-dimensional boxplots of a continuous variable compared against classes of a categorical variable. Mosaic charts can also be used to compare two categorical variables.

Scatter plots identify the relationship of salary with both single-season and career batting averages (see Figure 3). It appears career batting averages have a stronger correlation with salaries than single-season averages, although in both cases the correlation is clearly positive and non-zero. Scatter plots can also be used to compare a nominal-level variable with several levels against a continuous variable, such as the scatter plot comparison of salaries by team. Quickly a number of very highly paid players are identified, such as Eddie Murray, Jim Rice, Don Mattingly, Gary Carter, Ozzie Smith, Mike Schmidt and Dale Murphy (see Figure 4). A similar result comes from a boxplot of salary statistics by team; the New York Yankees paid the highest mean annual salary ($991,000) (see Figure 5).
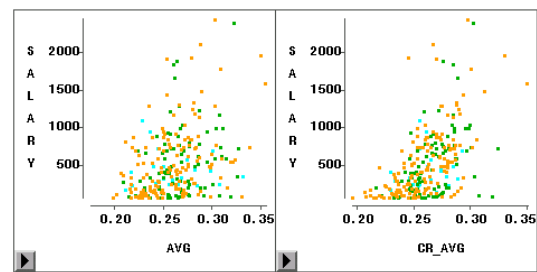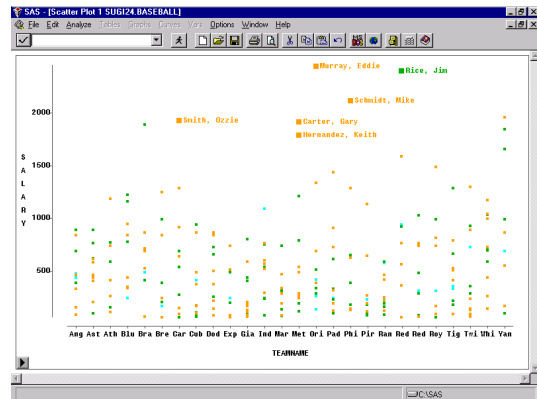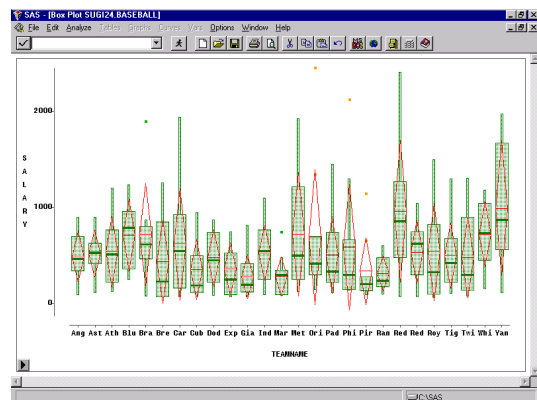

Figure 3

Figure 4

Figure 5

THREE-DIMENSIONAL ANALYSIS

Even more powerful visualization is available by using three-dimensional charts. Multivariate relationships can be explored and analyzed, usually as a precursor to regression analysis and multivariate procedures such as principal components and factor analysis.

Multiple scatter plots can be used to quickly assess relationships among many variables at once. These multiple scatter plots bear a lot in common with crosstabulations, where the shape of the scatter has a direct relationship with measures of association like the Pearson product-moment correlation coefficient. The first multiple scatter plot compares salary with single-season fielding statistics (see Figure 6). It is difficult to conclude from these plots that salary has a strong linear relationship with any fielding statistics; one interesting trend is between put outs and assists. (It turns out this relationship is an artifact of the way baseball fielding statistics are tabulated; infielders like first and second basemen acquire many more putouts than other players, while shortstops and outfielders have many more assists.)
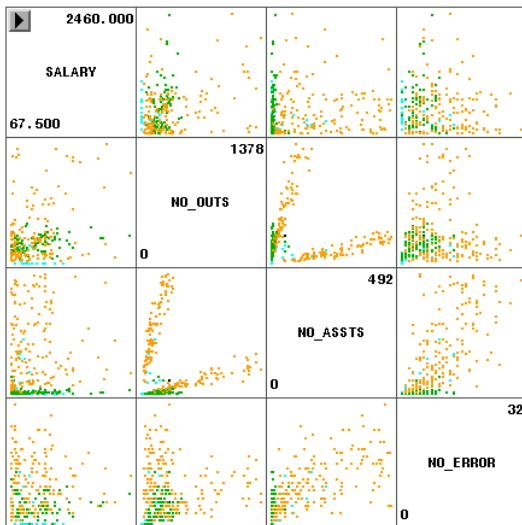
Figure 6

A second set of scatter plots compares salary with single-season and career batting statistics (see Figure 7). It is less difficult to draw conclusions from comparisons within like variables than with salary. Salary appears to be most strongly correlated with career runs and career batting average. Strong correlations appear between career homers, runs and RBIs. A similar but less pronounced correlation is present for the same measurements at the single-season level.
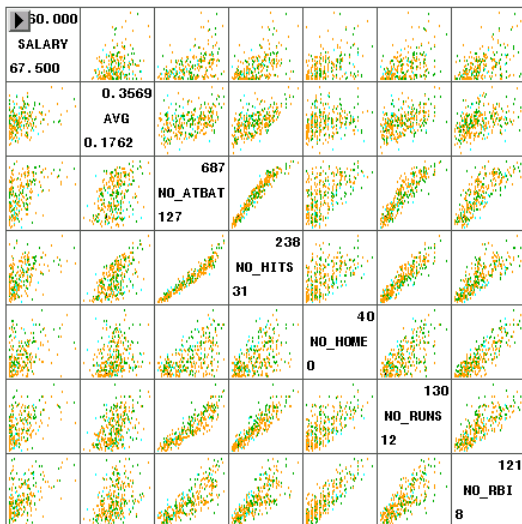

Figure 7

We will explore rotating three-dimensional plots in the subsequent discussion of multi-dimensional visualization.

FOUR-DIMENSIONAL ANALYSIS (AND BEYOND)

Formerly the hallowed ground of rocket scientists, SAS/INSIGHT software gives any analyst the ability to simulate and explore the mysterious "fourth dimension" and beyond. The application of color and symbology explores extra-dimensional patterns indirectly, while animation provides a more direct analysis, particularly when applied to time and date variables.

The business simulation used in this paper does take direct advantage of animation, but the use of color is employed to

designate the field position of each player. Orange is used for infield players (catcher, first, second and third base, and shortstop), green for outfield players (left, center and right field), and blue for other positions (utility players and designated hitters).

If we had multiple years of player statistics, or monthly statistics, it might be appropriate to use animation to view the trends in hitting or fielding over the course of a season or across seasons. The choice of animation versus a multi-dimensional chart depends on the context. Generally it is useful to use color or symbology to indicate multiple time periods within a standard chart without animation. However, as a single chart becomes more complex and thus harder to interpret, the judicious use of animation has payoffs.

LINEAR REGRESSION MODELING AND MODIFICATION

The process of building a regression model in SAS/INSIGHT software is inherently interactive and offers specific advantages over interactive model building using traditional regression procedures in SAS/STAT ® software. Diagnostics are plentiful and easy to access, making model modification quick and painless.

BUILDING A MODEL

Generally regression models are developed using the "Fit" technique included in SAS/INSIGHT software. The Fit Menu allows the analyst to specify dependent and independent variables, to interactively cross and nest independent variables, and to specify model weights and by-variables. Dependent variables can be continuous or categorical and there are a variety of model-fitting options, including response distribution, link function, and scaling parameters. The menu also allows the analyst to specify new variables to be output from the model, such as predicted and residual values of the dependent variable, and to select from a variety of tabular and graphic fit diagnostics.

DIAGNOSTICS

Some of the diagnostics that are more useful for multiple linear regression include the traditional r-squared, F, t, and variance inflation statistics. I also find very useful the following charts:

- Residual versus predicted dependent variable plots for testing error variance for heteroscedasticity (see Figure 8's left-hand chart)
- Residual normal quantile-quantile plots for testing normal distribution of residuals (see Figure 8's right-hand chart)
- Partial leverage plots for examining the partial correlations of each independent variable and the likely effects of outlying observations (see Figure 10 in following section of this paper for an example)
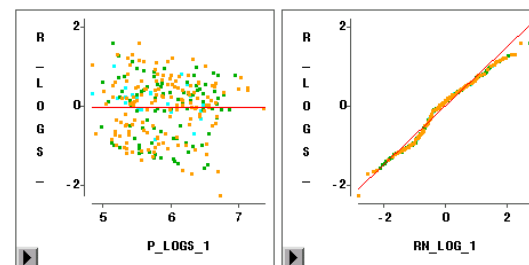

Figure 8

MODIFYING THE MODEL

After an initial model has been specified and run, the analyst traditionally relies on interpretation of estimated parameters and

fit diagnostics to assess model performance. Using SAS/INSIGHT software, the analyst can additionally use the Fit menu to do the following:

- Add new independent variables to the model to capture unexplained variance in the dependent variable
- Remove independent variables from the model that don't appear to explain any variance in the independent variable
- Transform independent variables that appear to share a curvilinear or non-linear relationship with the dependent variable and add them to the model

Outlying observations that affect model performance can be examined and dropped, or a new variable that explicitly captures the effects of an outlier can be created and added to the model. A very handy part of SAS/INSIGHT software's fit charting capabilities is that observations dropped from a model can be removed from calculations but still displayed and marked accordingly. This permits the analyst to evaluate in real time the impact of the outlying observation on the model's goodness of fit.

Model modification is faster when the analyst uses the Fit menu's "Apply" button rather than the "OK" button to run a new model—this leaves the menu intact rather than closing it before the fit results are displayed.

USING REGRESSION MODELING TO DEVELOP THE BASEBALL RECRUITING PROFILE

We used SAS/INSIGHT software to fit four multiple linear regression models. The first three models examined separate sets of independent variables, and the final model brought together the best combination of independent variables from the previous specifications.

The first model examined the relationship of single-season fielding statistics with the log of salary. We employed the log transform on the dependent variable to mitigate the highly skewed distribution of the original variable. While the F statistic indicates the model explains more variance in the dependent variables than its expected (i.e, mean) value, the adjusted r-squared statistic (0.561) indicates that fielding statistics alone do not explain enough variation to be satisfactory. All variables were statistically significant or nearly so (i.e., number of errors had a t-statistic of –1.89, which is significant at the alpha=0.10 level), and all had the expected sign.

The second model examined the log of salary as a function of single-season batting statistics. The adjusted r-squared statistic (0.317) suggests an improvement over the single-season fielding model. Batting average, hits and bases on balls were statistically significant and had the correct sign. Home runs were not statistically significant, and runs, while significant, had the unexpected sign (i.e., negative relationship with salary). We chose to leave the number of runs in the model on theoretical grounds. Multicollinearity was not an issue in this model. Error variance appeared to be relatively constant, resulting from the use of the log of salary instead of its "raw," original value. This ensures that statistical tests and confidence intervals resulting from this model are correct.

The third model examined the log of salary as a function of career batting statistics. This model improved yet again over the single-season batting statistics model with an adjusted r-squared of 0.463. We adjusted each of the count-based independent variables in this model (hits, home runs, runs, RBIs, and bases on balls) by dividing each by the number of years in each player's career. This adjustment generates batting statistics that are more normally distributed and reduces the collinearity among independent variables in the model. Career batting average and the average bases on balls per year over the career were both statistically significant and had the expected sign. Multicollinearity was mild compared with a version of the model that used raw counts of batting statistics.

The fourth model brings together each previous individual model. The overall model goodness of fit improved slightly to 0.485. However, only three variables (number of errors, single-season bases on balls, and career batting average) were significant and each exhibited the expected sign. The remaining variables were left in the model on theoretical grounds.

The interpretation of the parameter estimates in the fourth model still left us feeling like this model did not perform well based on theoretical assumptions about the components of player salaries. It is possible that effects such as salary negotiations and other unavailable data might help capture more variance in the salary variable and develop a more satisfactory model. However, another modeling technique might reduce the natural covariance among the large set of independent variables into a smaller, more manageable set of underlying constructs. For this reason, we pursued a principal components analysis, described in the next section.

MULTIVARIATE ANALYSIS & VISUALIZATION

Multivariate analyses are appropriate when an analyst is faced with a sizable set of measured variables that capture underlying or "innate" constructs (the basis for which is theoretical or conceptual). Approaches like principal components and factor analysis are sometimes referred to as "data reduction" techniques, which is technically accurate but perhaps an oversimplification of what actually happens in the course of applying these approaches. SAS/INSIGHT software includes a "Multivariate" analysis menu that includes multi-way scatter plots, confidence interval ellipses and principal components analysis. However, SAS/INSIGHT software can also aid in the visualization of output from other multivariate procedures, as we shall see in our continuing baseball analysis.

PRINCIPAL COMPONENTS

Principal components analysis reduces the covariance among a set of n measurable variables into one or more linear equations, each of which explains the largest possible amount of covariance in n-dimensional space. Principal components by definition are uncorrelated, such that the first principal component explains the most covariance in the data matrix, then the second explains the most covariance of what is left subject to bring uncorrelated with the first component, and so on. There are as many principal components as variables, but the approach (well-applied) allows the analyst to explain a substantial degree of the total covariance with a smaller number of components than the number of original variables.

An advantage of principal components over factor analysis is that the system of equations is solvable—there is no need for estimation, and therefore results can be applied back to the original observations to assist in prediction. However, the notion that all principal components are uncorrelated leaves some analysts uneasy—they reason instead that most sets of real-world variables, such as baseball statistics, must be correlated in some way, and the orthogonal nature of principal components unnecessarily simplifies the nature of real things.

We used the SAS/INSIGHT software's multivariate analysis features to build principal components from single-season batting, single-season fielding, and career batting statistics. An examination of eigenvalues of the correlation matrix suggests four principal components explain a larger proportion of the correlation among these variables than the variables themselves (i.e., the first four eigenvalues are greater than 1). However, the pattern matrix for the first four principal components fails to deliver an easily interpretable profile, what statisticians call "simple structure." Simple structure means that each component exhibits a cluster of high loadings on a set of variables and low loadings on the remaining variables. It also means that no one variable has high loadings on more than one variable. This analysis fails the

test for simple structure and we conclude that principal components may not be an appropriate approach for this data. Instead we pursue factor analysis.

FACTOR ANALYSIS

Factor analysis is not as straightforward as principal components analysis, but permits (through a technique called oblique rotation) each underlying construct to be correlated. Given our earlier analyses with the baseball data, it seems intuitive that great athletes will have both good fielding statistics and good batting statistics. Using SAS/STAT software's PROC FACTOR, we ran a factor analysis on the same set of variables as in the principal components analysis (i.e., for the statisticians out there, we used iterated principal factor extraction with promax rotation on a correlation matrix with communalities on the diagonal estimated via the squared multiple correlation method). We also used the factor analysis results to create a set of output observations (appended to the original input dataset) that places each observation in the p-dimensional space (given p factors retained by the analysis).

```
proc factor data=sugi24.baseball method=prinit
  priors=smc nfactors=4 rotate=promax scree
  reorder out=sugi24.bbfactrc;
var no_outs no_assts no_error no_hits no_home
  no_runs no_rbi no_bb cr_hits cr_home cr_runs
  cr_rbi cr_bb avg cr_avg;
title 'Factor analysis on baseball data using
  career counts, not avgs';
run;
```

The results of this analysis again suggest that four factors should be retained, but this time simple structure (following oblique rotation) suggests clearly that the first factor loads most strongly on career batting counts (hits, runs, home runs, RBIs, and bases on balls). The second factor loads on single-season batting counts, the third on career and single-season batting averages, and the fourth on fielding statistics.

Using SAS/INSIGHT software's rotating three-dimensional scatter plot feature to visualize the results of the factor analysis, we see a large contingent of players with relatively low scores on the career batting count factor, due mostly to their short careers to date (see Figure 9).
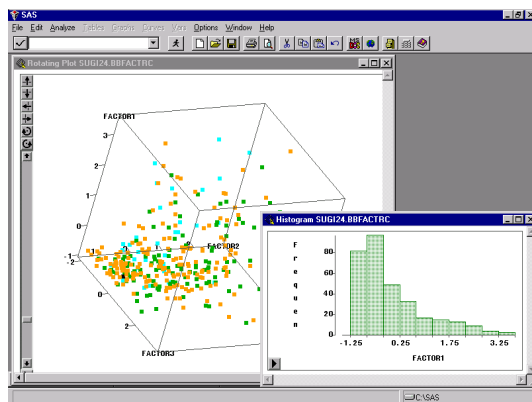
Figure 9

The factor scores for each observation can be used in a regression analysis to capture the variance in salaries explained by these underlying constructs of player performance. It turns out that the fielding construct has little to do with salary variations, and that all three other factors (career batting, single-season batting, and batting averages) explain a significant portion of salary variance (see Figure 10).
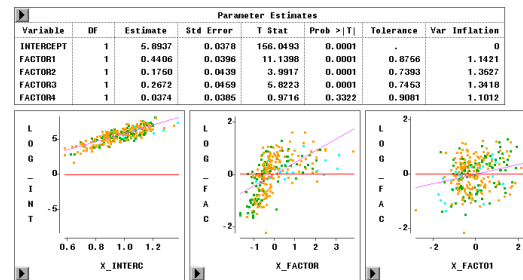
Figure 10

The regression model with only the first three factor scores as independent variables has an adjusted r-squared statistic of 0.536, a further improvement over the previous model. All parameter estimates are statistically significant with the expected sign. Multicollinearity is no longer a problem in this model (which we already know from the factor analysis inter-factor correlation table, where correlations ranged between 0.15 and 0.4).

Finally, we can select players with strong batting statistics who are paid less than their peers by comparing scatter plots of the log of salary and the first three factor scores. By focusing on the single-season batting count factor and the batting average factor, we can identify a set of approximately 20 players whose salaries are in the lower portion of the distribution and whose batting performance is in the upper portion of the distribution (see Figure 11). It stands to reason these are the top players for recruitment, since they are relatively low-paid but have excellent performance behind the plate. This list can be extracted to a separate dataset and distributed to field scouts and the rest of the franchise management.
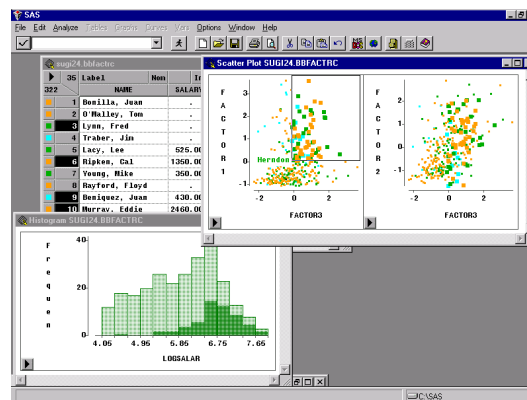
Figure 11

CONCLUSIONS

This paper has reviewed the usefulness of visualization and modeling techniques available in SAS/INSIGHT software for exploratory data analysis. There are many ways to expand the use of SAS/INSIGHT software beyond those described in this paper. The role of exploratory data analysis in data mining ventures is critical. As the applications of data mining grow in number and sophistication, data miners will find ways of using SAS/INSIGHT software in new and exciting directions.

REFERENCES

SAS Institute, Inc. (1995) SAS/INSIGHT User's Guide, Version 6, Third Edition. Cary, NC: SAS Institute Inc.

SAS Institute, Inc. (1989) SAS/STAT User's Guide, Version 6, Fourth Edition, Volumes 1 & 2. Cary, NC: SAS Institute Inc.

RECOMMENDED READING

Friedhoff, Richard M. and Benzon, William (1989) *Visualization: The Second Computer Revolution*. New York: Harry N. Abrams, Inc.

Friendly, Michael (1991). SAS System for Statistical Graphics. Cary, NC: SAS Institute Inc.

Hatcher, Larry (1994). A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. Cary, NC: SAS Institute Inc.

Tufte, Edward R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

    Robin E. Way, Jr.
    SAS Institute Inc.
    614 SW 11th Avenue, Suite 400
    Portland, Oregon 97205
    Work Phone: 503-243-4630
    Fax: 503-243-4631
    Email: sasrnw@wnt.sas.com