Paper 169-26

Outliers, Inliers, and Just Plain Liars -- New Graphical EDA+ (EDA Plus) Techniques for Understanding Data

David DesJardins, U.S. Bureau of the Census, Washington DC, 20233

ABSTRACT

Graphs, the natural language of mankind, offer even novice data analysts a quantum leap in their data editing/analysis capability. In our dynamic, rapidly changing world, we are faced with a virtual flood of data -- often from very vital/complex systems. Unfortunately, this flood of data can likewise be a real challenge to understand/analyze/edit. Many of our traditional/conventional data analysis/editing methods generate fixed-formula printouts requiring an analyst to review and correct the fields of records that are thought to be erroneous. There are many limitations to these conventional methods -- even when well designed. They sometimes overlook basic methodological problems, they typically channel the reviewers in a manner that either may not allow a number of the errors to be found, or they focus on traditional (outdated?) relationships, and they often do not account for changes in the data.

A special course has been designed by the author to teach a number of powerful graphical based methods to deal with these problems. This very popular course uses new, easy to learn (point-and-click), interactive Exploratory Data Analysis (EDA) software packages (SAS's JMP® and Insight® graphical data analysis software) -- and makes these techniques very straightforward to apply. These graphical methods can first be applied in an exploratory manner -- to discover nuances that conventional methods are likely to
miss. Then, special interactive EDA + graphic forms (creating "live" graphs) can then be used in a straightforward, highly productive manner to edit these data. . (NOTE: The "+" is used in EDA+ is used to signify the enhanced interactive EDA methodology developed by the author to enhance EDA methodology -- this is also called EDA PLUS.) Lastly, these new EDA+ graphical methods allow "inlier" detection and confirmatory review of data -- to not only find hidden relationships within these data but to also assure that corrections made during to the editing process have worked well.

NOTE: Although these interactive EDA techniques are also taught as part of a graduate level statistics course by the author, they offer even novice subject matter specialists a quantum leap in their data analysis capability. As such, it can be taught to a variety of audiences.

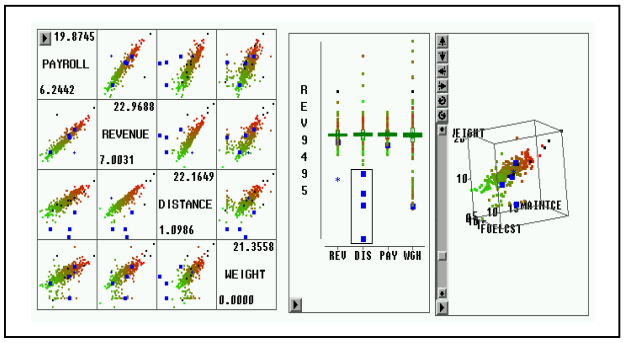
Keywords: Exploratory Data Analysis (EDA), EDA + (EDA plus), interactive graphics, industry profile and leverage plots, inliers

1. INTRODUCTION

A key focus of this paper is on ***"Dead" graphs*** -- may they rest in peace!! Four key factors contribute to a revolution in data analysis and make the introduction of these EDA+ methods (with their ***"live" graphs***) at this time a momentous opportunity. First, these graphics software packages provide our analysts with the ability to generate hundreds of graphs in a matter of mere minutes. Generating such a large number of graphs would have taken weeks just a few years ago. Second, by using key EDA+ techniques like brushing and animation in combination with specially designed graphic forms (like Industry Profile Plots), these ***interactive*** software packages provide powerful/sophisticated "live" graphical methods of looking at the data and reviewing subcomponents of it. If the analysts believe the data is in error, then they can easily discover/correct it in an interactive manner using these point-and-click tools. Third, some versions of this software are incredibly inexpensive. For instance, the comprehensive student version of SAS's JMP®, PC-based, software package is available for about $60. Fourth, and most important, individuals using these multi-purpose EDA techniques are not locked into custom designed software packages and fixed ways of looking at the data. (Many custom designed graphical data analysis packages have a 6-month lead/design window.) By using the above hardware and software tools, we have developed new general-purpose graphical forms and special techniques that greatly enhance the speed and efficiency of data editing/analysis tasks (see particularly DesJardins, 1998). Analysts no longer need spend time editing their data with fixed methods and cumbersome, boring, tabular printouts.

Shown in figure 1 in this printed form is a "dead" Data Profile graph devised by the author. Although it is a quite powerful graph even in this "dead/printed" format, this static display is but a small fragment of the information it could reveal in a "live" interactive format -- where you could, for instance, brush across outlier points. This paper will attempt to explain the extent of the information that could be revealed in an interactive session designed for the analysis as well as the display of our data.

A Dead Graph -- RIP

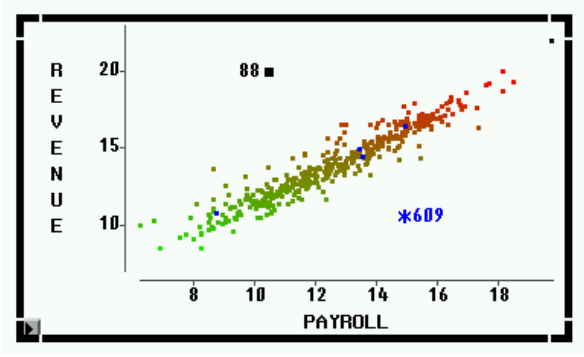


This paper shows how a series of well-designed ***interactive*** graphical methods can be developed and used to edit/analyze/explore data. There are many "blind", "black box", CPU algorithms that allow detection/printouts of outliers in distributions that may be in error (for instance, Granquist, 1997). However, these methods often cannot yield insight into situations in which more subtle distributional errors occur. There are also a number of new graphical packages that make it very easy to use to locate and correct errors (outliers) in data. Cleveland (1993) has also provided a variety of (static) methods for graphical data analysis. Granquist (1997) and Hogan (1996) have shown how to apply some of these methods to files of businesses. These methods are often found wanting, however, in the face of more subtle distributional errors/problems and "inliers" (defined below).

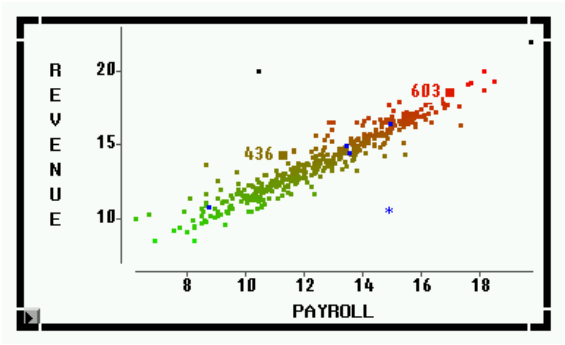
An *outlier* is a data value that lies in the tail of the statistical distribution of a set of data values. The intuition is that outliers in the distribution of uncorrected (raw) data are more likely to be incorrect. Examples are data values that lie in the tails of the distributions of ratios of two fields (ratio edits), weighted sums of fields (linear inequality edits), and Mahalanobis distributions (multivariate normal) or outlying points to point clouds with graphs. An *inlier* is a data value that lies in the interior of a statistical distribution and is in error. Because inliers are difficult to distinguish from good data values they are sometimes difficult to find and correct. A simple example of an inlier might be a value in a record reported in the wrong units, say degrees Fahrenheit instead of degrees Celsius.

Isolated inliers may not be a problem and may be almost impossible to distinguish from correct data. Sets of inliers of moderate size may seriously affect uses of microdata. In some situations, interactive graphical methods are required to discover these sets of inliers. These methods can also be used for finding erroneous mixture distributions. For instance, in more advanced situations, sets of inliers may arise when two or more administrative lists are linked and some of the identifying information is in error. Another example is to check when corrections are done. Then we can use these graphical methods to confirm the plausibility of the changes. This new methodology can be used to clean up data -- or determine that there are errors in data for which clean up methods need to be created.

Points # 88 and 609 are "outliers"



Points # 436 and 603 correlate well with the majority of these points and would be considered "inliers".



In this paper, we provide a brief overview of these new interactive graphical methods that are currently being taught by the author in his special EDA + course. This course has become so popular that it has gone from just being taught at the US Census Bureau, to being taught (by special invitation) at Statistics Sweden, Statistics Canada, Statistics Italy, BLS, the UN, the Washington Statistical Society, and as a graduate statistics course for USDA.

The outline of this paper is as follows. In the next section, we cover how exploratory data analysis (EDA) methods can be used to detect errors and to correct data. Although these simple methods are typically quite straightforward to learn and apply, they often yield information about serious errors that have previously gone undetected. ("In the land of the blind, a one eyed man is king!") These graphical methods provide an easy check on the efficacy of our fundamental applications and a cross-check on our corrections. In the third section, we show how more sophisticated EDA methods (EDA +) can detect the existence of hidden problems -- inliers. These inliers might arise, for instance, as a flagging error within the mixture of two distributions. The final section consists of concluding remarks.

2. EDA GRAPHICS AND OUTLIER DETECTION

Conventional editing methods have often involved the development of if-then-else/ratio rules in computer software. These rules delineate records that may need editing. These methods have typically been used in computer environments in which single dimensioned, printouts were created for analyst review. Review and correction of these data in this format can often be time-consuming and incredibly boring/tedious. Worse, errors may not be located because edit rules are inflexible, are based only a few variables, or are not designed to detect certain classes of mistakes. In addition, over time, updated sets of questions may be asked -- and methods that are developed for one survey may not be exactly appropriate on other surveys. Perhaps the most serious problem (in view of our rapidly changing society), these fixed-formula basic editing ratios/algorithms may vary as our data/industries change.

Use of modern, graphical data software in an exploratory manner allows analysts to detect errors that cannot be detected by inflexible editing rules. These general-purpose software packages provide a quick, easy to learn methodology for updating algorithms/databases as errors are detected. Statistical agencies often do not know how easily the methods can be applied. They may not be aware the basic software can often be learned in only one day -- and that and that the general

purpose application of this software is now well developed. For instance, "cookbooks" have been created by the author for a number of types of data. They can provide users with 10 very powerful, special purpose data analysis graphs (for instance, a Data Profile and Industry Profile Plots).

2. 1 Interactive EDA Methodology:

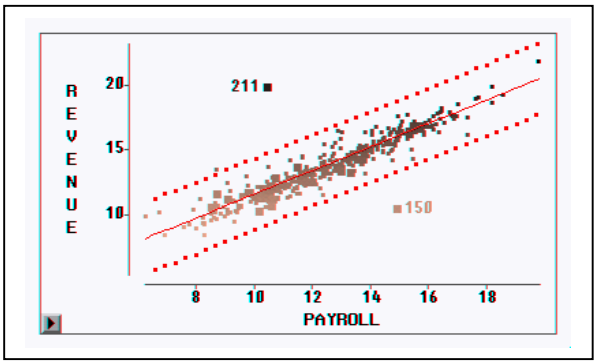
Graphs also communicate across a wide area of expertise. A properly chosen graph can make even sophisticated statistical concepts clear to laymen. Accordingly, Bureau statisticians can now more quickly/effectively explain to our subject matter specialists the fundamental concepts behind these new graphical data analysis techniques. Thus, these new EDA methods now allow our analysts to discover any number of errors in their data. However, an advantage of EDA+ (EDA PLUS) methodology is that it uses the full potential of modern computer technology (as opposed to "dead" graphs). This truly empowers our subject matter specialists and provides them with a quantum leap in our data analysis methodology. (Unfortunately, if the reader were to review a dozen current EDA texts, little in the way of this new interactive EDA + methodology would likely be found.)

As outlined earlier, the fixed computer algorithms of conventional edit methods have often missed unanticipated errors in the data. The fixed algorithms are blind in the sense that they cannot adapt to new situations. For instance, comparing variables and using ratios requires a very good understanding of the often unique/unpredictable relationships between each of the variables. These relationships can vary markedly for different point cloud clusters associated/corresponding to companies in different industrial codes. The variation can be substantial between companies in different size ranges or when survey forms are re-designed. Time series variances such as business cycles and periodic anomalies can affect these relationships as well. Basic relationships for data sets can also simply vary substantially over time as well. In addition to producing outliers, these factors often hide inliers. By using powerful EDA + graphics tools in combination with an individuals trained in this new methodology, these problems can be quickly identified and dealt with. In this section, we show how EDA methods provide a revolution in data analysis in statistical agencies. The next section highlights key examples of where EDA has shown itself to be significantly better at data editing, in the analysis of data, and for identifying outliers and inliers. Also noted is that it is also often much better at discovering undetected flaws in traditional data analysis methodologies in data at statistical agencies.

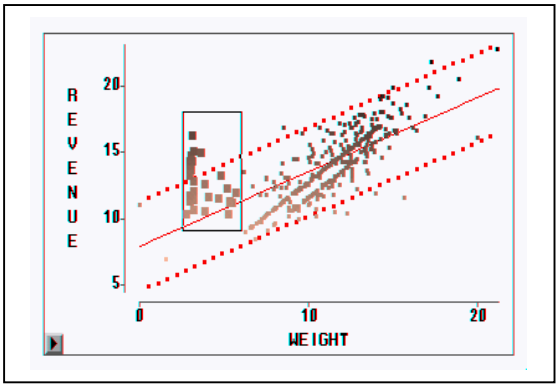
NOTE: For confidentiality reasons, all the examples used in this paper use fictional data or artificial data that are similar to real data.

2.2 Production Control Charts

In some areas of the Bureau, Production Control Charts (PPCs) are now automatically produced as part of an enhancement to the user interface. Using the best X/Y correlations of the variables in our survey data sets, PPCs can be created -- including editing aids like the 80% confidence lines shown on the left below. Analysts can thus quickly highlight/mark the points that need editing (like companies #211 and 150) -- and reference these points in other related graphs. In most situations, this makes the data very straightforward to edit.



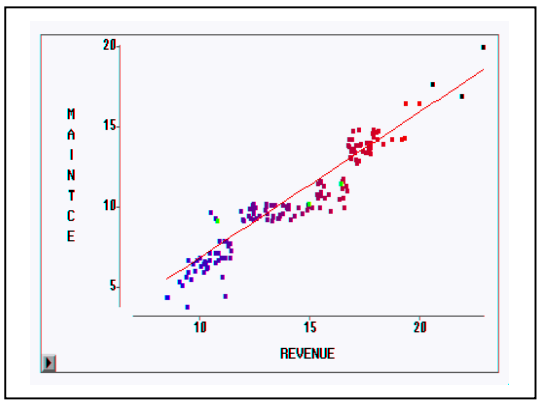
In the example shown below, however, there was a serious problem with a subset of the Bureau's Transportation Survey data. In this graph, the highlighted points (with unusually high Revenue and low Shipping Weight) are simply out of place. Even an untrained eye can easily detect that there is a serious problem with this "fit".



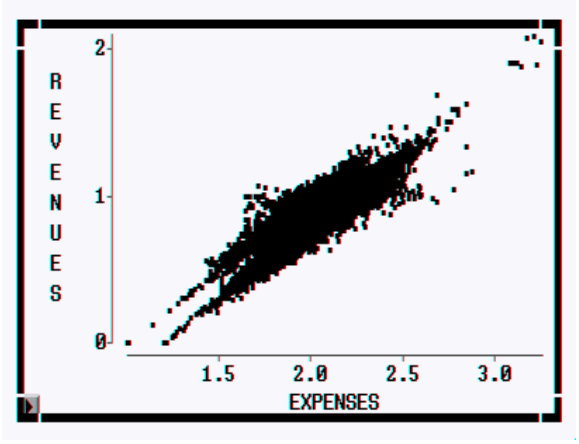
This is an excellent illustration of an elementary use of EDA methodology. Once seen, further investigation disclosed that these companies were reporting Shipping Weight in tons instead of the unit of measure requested on the questionnaire -- pounds. It is noted that this problem had been going on for years. Given their lower values, our conventional fixed ratio methods of editing had not detected this problem. The fixed methods were "blind" to the unforeseen anomalies in the data -- that was easily determined with a proper graphical view.

2.3 Finding Previously Undetected Errors

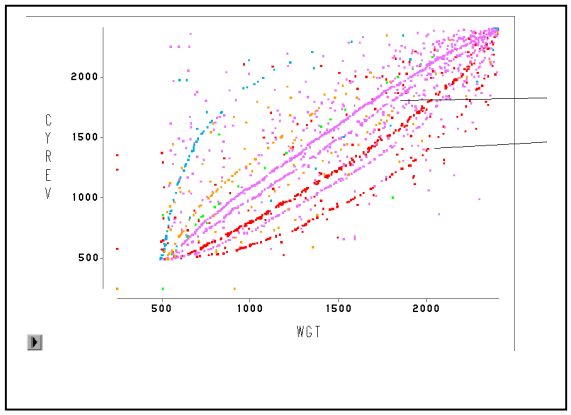
In some areas of the Bureau, even the simplest graphic techniques have often proved to be revolutionary. The graphical methods offer key insights into potential problems in the data. Humans can easily see a distinct clustering of the points in the next graph. Many "blind" computer algorithms would simply report a good linear relationship between reported Revenues and Maintenance Costs (Expenses) for this industry group.



The graph below illustrates an example of a problem that had likewise gone undetected until a review using graphical methods. There are two rather distinct clusters of data. Because of a flagging error, companies that were tax exempt were accidentally intermixed with companies that were not tax exempt. By graphing the data, the analysts were able to detect how our "blind" editing software did not detect the problem with the flags in the data.



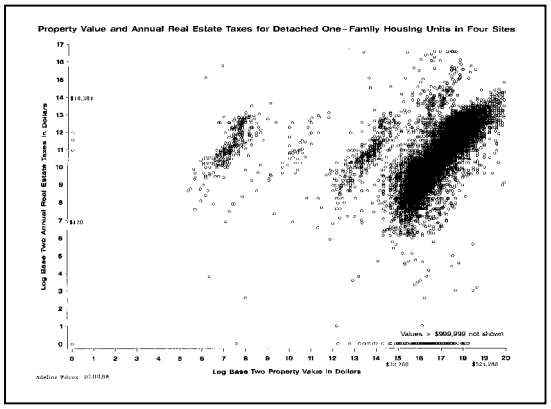
Exploration is a key aspect of these interactive graphical methods. EDA+ methods not only offer us a global perspective of the data -- but also the ability to explore any number of relationships between many variables. These perspectives can lead to the discovery of other possible problems with some of our traditional/fixed ways of editing. At the Bureau, the graph below is now referred to as a "football graph" (see DesJardins, 1998). During one EDA session, this unusual pattern in data became evident. The plot shows the reported Revenues vs. Shipping Weight for all of the companies in our Transportation Survey (ranked by size). Shown below is a copy of the graph as seen on the computer screen. Each of the distinct lines shown was in a different color (representing each of the different SIC codes assigned to our Transportation Survey -- for instance: "long distance shipping with storage", "short distance without storage", etc.). Further investigation showed that each of these lines was actually created by our imputation methodology. (If a Shipping Weight is not reported, then the reported Revenue and the regression equation is used to assign a Weight.) In the first days of this preliminary survey, a large number of points needed to be imputed. The other, rather randomly distributed, points in this graph are companies who reported both Revenue and Weight. An obvious question is: "Using the imputation formulas that are fixed a priori, are we perhaps over-imputing Weight values for SIC # 4213 and under-imputing Weight values for SIC # 4212?"



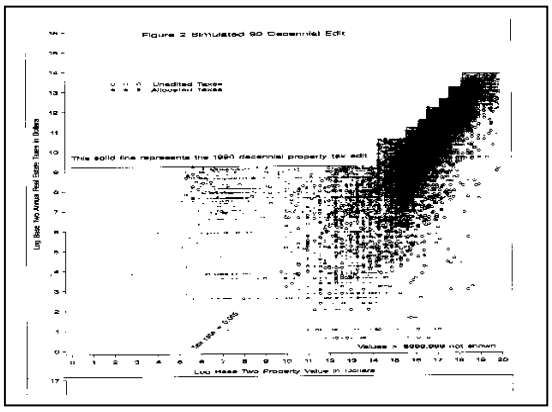
2.4 Spotting Common Errors

The error shown below is an example the "common errors" curriculum that is part of the EDA class taught by the author. The real goal of the EDA class is to teach an efficient (non labor-intensive) ***visual methodology*** that will help the analyst quickly understand their data. Plotting examples of common errors trains the eyes of analysts to recognize typical error patterns. The analyst can then quickly recognize those patterns in their surveys -- isolating those that are valid and those that do not make sense.

Fixed methods of editing can often be “blind” to some of the errors in data. This example shows how the application of these fixed methods (that have not been reviewed carefully) can massacre data. On the left we see a scatterplot of taxes paid (Y variable) versus assessed property values (X variable) for a number of homes in a county in the USA -- thousands of points (data are in log values). In our EDA class, one task assigned to students is to plot the errors that could commonly occur with our data. If the students have seen examples of this type, then the eye can usually pick out similar errors/patterns.



As can be seen, the data all fall within a rather well defined range of Y values (taxes paid). However, the smaller clusters of outlier points show inordinately high taxes for lower and lower property values.



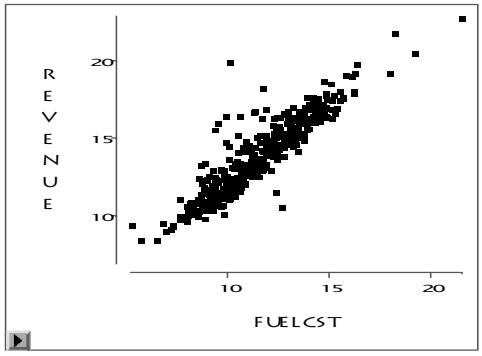
Above we see the results of one of our old (blind) editing technique -- as it tries to account for the smaller clusters of outlier data to the left of the main point cloud of these data. The darker points are values imputed by this algorithm. The error: A common error in recording or transcribing data is dropping zeros -- showing, for instance, an actual property value of $130,000 as $13,000. If we look closely at these data, we can see that these clusters represent errors of 10's, 100's, and 1000's. Note how the blind algorithm took "bites" out of successive steps of "good" values as it tried to adjust these data for these outlier points!

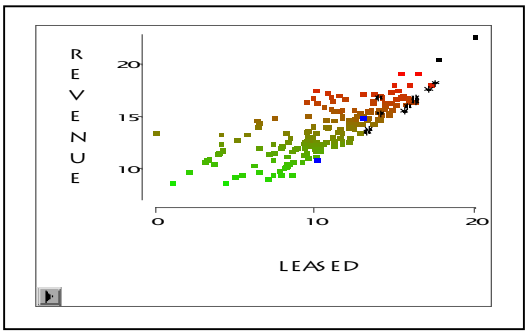
2.5 Advanced EDA+ Techniques -- Leverage plots

Leverage plots are also very helpful in identifying inliers. The partial leverage plot for each explanatory variable is used as an indicator of the relative influence of each observation on the parameter estimates. For a given explanatory variable, the partial leverage plot is the plot of the response variable and the explanatory variable after they have been made orthogonal to the other explanatory variables in the model.

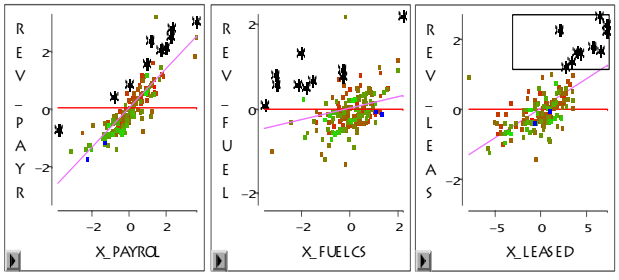
For linear models, the partial leverage plot for a selected explanatory variable can be obtained by plotting the residuals for the response variable against the residuals for the selected explanatory variable. The residuals for the response variable are calculated from a model having the selected explanatory variable omitted, and the residuals for the selected explanatory variable are calculated from a model where the selected explanatory variable is regressed on the remaining explanatory variables.). Let ry[j] and rx[j] be the residuals that result from regressing y and X(j) on X[j]. Then a partial leverage plot is a scatter plot of ry[j] against rx[j].

Two reference lines are displayed in the plot. One is the horizontal line of Y=0, and the other is the fitted regression of ry(j) against rx(j). The latter has an intercept of zero and a slope equal to the parameter estimate associated with the explanatory variable in the model. The leverage plot shows the changes in the residuals for the model with and without the explanatory variable. For a given data point in the plot, its residual without the explanatory variable is the vertical distance between the point and the horizontal line; its residual with the explanatory variable is the vertical distance between the point and the fitted line. For instance, below we see another two scatterplots of our trucking company data -- for reported revenues and reported fuel costs -- and for revenues and reported leasing costs. Aside from the fact that the lease cost data is a lot less correlated with revenues (than fuel costs), no alarm bells ring when we first look at these graphs.





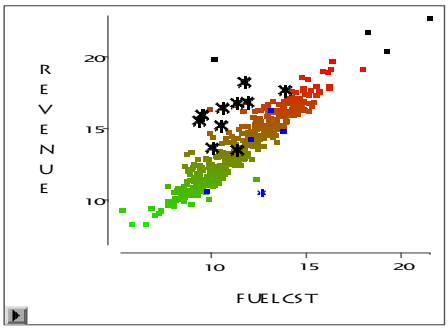
After we plot a fit (and the leverage plots) of these reported revenue versus these related variables, we begin to suspect something. As can be seen, there is a subset of very suspicious "outlier" points in the lease cost plot (highlighted within the square with a *). These points also show themselves to be the highest values in the leverage plots of payroll and fuel costs.

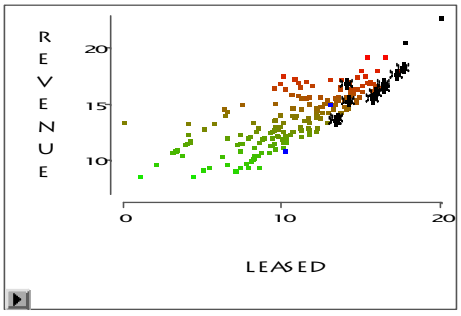


Returning now to the original scatterplots of these data (below), we can see that these highlighted points show up as a special cluster of points in the revenue versus fuel costs data (still marked with a *). These points stand out as a "bump" of companies with higher revenues for lower fuel costs than the

majority of these other companies. How are these trucking companies able to obtain higher revenues while burning less fuel? Looking again at these points in the scatterplot of revenues and leasing costs, we see that most of these points are clearly inliers -- they only discretely show up along the lowest edge of these data. For these companies, lower levels of revenues correspond to higher lease costs. In this example, it is clearly important that the analyst gain a better understand the trucking industry. It would appear that some leasing contracts provide trucks that come with a driver and fuel.

As can be seen, our EDA classes gives analysts (subject matter specialists) new tools that promote a better understanding of their data. These anomalies would certainly affect the imputation methodology that is used. These new tools not only quickly isolate data that may be in error, but offer an opportunity to truly understand it -- to show us subsets of our data that needs to be treated differently





3. CONCLUDING REMARKS

In addition to help in identifying outliers/inliers, we have found these EDA+ techniques to be very helpful in the following additional areas: (1) data analysis such as modeling and identifying clusters/unique subsets, (2) checking basic methodology such as imputation and sample weighting, and (3) multivariate analysis -- as was illustrated with leverage plots.

Using this EDA+ methodology, the Census Bureau is entering a whole new world of data analysis capability. This is made possible by new, very fast hardware (i.e., Pentiums and Unix workstations) and powerful, easy to use, point-and-click software (JMP® and INSIGHT® from SAS Institute). Formerly, software that they needed for their data analysis tasks. for systems development efforts to produce the custom our analysts had to learn the intricacies of programming or wait Instead, in conjunction with a quick, 40 hour, EDA course taught by the author, analysts are taught a variety of powerful EDA techniques using this easy to learn (basically point-and-click) software. The design of this courseware is revolutionary in two other ways as well. First, it stresses an interactive multivariate analysis -- for all of the variables on the survey form -- allowing for comparisons between variables in these data sets that, until now, had not been compared. The result is that we gain a real understanding of these data. Second, it is designed to be used in a highly interactive manner by our subject matter specialists -- who have only a moderate statistical background -- to give them a very powerful tool/understanding based on these key insights into their data.

The ability of the eye-mind combination to discern subtle and complicated relationships during review of graphs has long been known (see e.g., Cleveland 1993). This paper has demonstrated that is it possible to use new, powerful, user-friendly graphical software and enhanced EDA+ methodology to explore and correct data. The use of these EDA+ techniques has become a key part of our mainstream statistical methods for exploring and reviewing complicated statistical models. Further, through his very popular EDA/graphics course, the author has been particularly effective in moving these ideas in to day-to-day practice not only at the Census Bureau, but also in a number of national and international statistical agencies.

REFERENCES

Cleveland, William (1993), *Visualizing Data*, Hobart Press: Summit, NJ

DesJardins, David (1998*), "A New Graphical Techniques for the Analysis of Census Data",* Statistics Canada Conference Proceedings.

Granquist, L. (1997), "Macro-Editing - The Aggregate Method," *Statistical Data Editing, UN Conference of European Statisticians Statistical Standards and Studies, Geneva (Switzerland).*

Hogan, Howard (1995), "How Exploratory Data Analysis is Improving the Way We Collect Business Statistics," *American Statistical Association, Proceedings of the Section on Survey Research Methods,* pp. 102-107.

Sall, John (1990), "Leverage Plots for General Linear Hypotheses", *The American Statistician*, Volume 44, No. 4.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions,"* John Wiley: New York.

Winkler, W. E. (1997), "Problems with Inliers," paper presented at the European Conference of Statisticians, October 14-17, 1997, Prague, Czech Republic.

Author Information:

David DesJardins,
U.S. Bureau of the Census, SRD, Rm. 3000-4,
Washington, DC 20233-9100
Tel: 301-457-4863
david.l.desjardins@census.gov