



YOUR SAS TECHNOLOGY REPORT

FEBRUARY 2006

Dear SAS Technology Report Readers,

There are many ways you can put SAS to work for you. This newsletter, I hope, is one useful and informative vehicle. And our [customer Web site](#) offers technical support assistance, documentation, samples, communities, software downloads and training. In fact, training is an excellent way to learn about SAS® software. It can also be used to supplement or refresh your existing knowledge.

SAS President and CEO Jim Goodnight said recently, “For 30 years, our goal here at SAS has been to provide superior software that empowers you to take your business beyond expectations. And on that journey toward excellence, SAS training is the road map.”

To help you make the most of your SAS software, we offer instructor-led training, self-paced e-learning and SAS certification. Check out the [training Web site](#) for details. If you’re planning to attend the annual SAS Users Group International ([SUGI](#)) conference in March, you can begin your training in San Francisco. See the [Web site](#) for the training line-up.

Thanks for using SAS!

A handwritten signature in black ink that reads "Shelley Sessoms".

Shelley Sessoms
Editor, *Your SAS Technology Report*

Contact

We welcome feedback and article ideas.
Send e-mail to techeditor@sas.com.

Drop Variables from a SAS Data Set When All its Values are Missing

Use arrays and iterative DO loops to determine whether a variable's values are all missing. Create a macro variable whose value is the list of the variables to be dropped. Call the macro variable in a DROP statement in a subsequent DATA step.

```
/* Create sample data set */
```

```
data missing;
  input x1 x2 x3 x4 y1 y2 y3 y4;
datalines;
1 . 1 . 1 4
1 1 . . 2 . 5
1 . 1 . 3 . . 6
;
```

Put existing variables into one array and create another array to hold values "true" and "false". Since an initial value of "true" is given for all variables in the ALLMISS array, they are retained until a new value is assigned according to the IF statement. After all observations have been read and the values of ALLMISS1-ALLMISS8 have been assigned accordingly, any left as 'true' indicate a variable that needs to be dropped. These variable names are put into a macro variable, MLIST, and then used on a DROP statement in the next step.

```
options mlogic mprint symbolgen;
```

```
data _null_;
  array test(8) x1-x4 y1-y4;
  array allmiss (8) $ (8*'true');
  length list $ 50;
  set missing end=end;
  do i=1 to dim(test);
    if test(i) ne . then allmiss(i)='false';
  end;
  if end=1 then
  do i= 1 to dim(test);
    if allmiss(i) ='true' then list=trim(list)||' '||trim(vname(test(i)));
  end;
  call symput('mlist',list);
run;
```

```
data notmiss;
  set missing;
  drop &mlist;
run;
```

```
proc print;
run;
```

Measures to Successful Coding

By Stephanie R. Thompson, Associate Director for Institutional Research, University of Memphis

Where Do I Start?

Have you ever found yourself sitting in front of your computer staring at a blank Editor window wondering how to start coding for your latest assignment? Maybe you have brought up some code that you have used before or gotten from someone else and wondered, “How can I change this to do what I need to do now?” If this seems somewhat familiar, this four-part series of articles may be your solution. There are some measures you can take even before opening a SAS® session to make this scenario a thing of the past.

Background

People from all kinds of backgrounds are using SAS software to meet their business needs. Some have previous programming experience and some have never before typed a line of code. There is more to SAS programming than just knowing some syntax and launching a job. Being a successful SAS programmer means bringing yourself up to speed with some basic programming skills. Before entering even the first line of code into the Enhanced Editor, some preparation is needed – preparation that help you more than you can imagine. This series will provide the basics of programming that can help you write good code and ensure you are getting the answers you intended.

How Do I Get There?

Over the years, I have been asked by co-workers to “look at my code and tell me if this is right.” To their frustration, I ask a question in return: “What question were you trying to answer?” On the surface their question seems straightforward and my answer seems a bit unhelpful. Code without errors and warnings will produce output. But will it produce the output they intended and, more importantly, is it correct? Following a few simple steps up front will help ensure the answer to both questions is “yes”.

The following set measures have proven useful to me over the years:

1. What’s the question?
2. Know thy data
3. Draw it out
4. Add code elements
5. $E=MC^2$ (a.k.a. calculations)
6. Check logic
7. Start coding

Each of these measures will be covered over the course of the four articles. At the end, you will have a sound framework for developing your programs.

What’s the Question?

Many times we receive assignments that sound like commands. Run a sales report from last week. Find out which regions are behind plan. Tell me which product tier will sell best in each store. Straightforward? Maybe, or maybe not. Is the person making the request looking for net or gross sales? Dollars or units? Which metrics of the plan is she interested in – all or just one? Does she want the report to include all regions or just

those that have fallen behind? What does she mean by tiers that “sell best” – those that make the most profit or that sell the greatest quantity? These are just a few of the possible questions that come to mind.

Let me share an experience that will help put this into context. I was working on a team that was asked to determine product placement. The executive vice president of the department called my boss and me to a meeting. He told us to run a clustering analysis on a particular product line and present the results to him. We worked for the next week to do as we were asked. We generated so much information and output that when the report was printed it measured more than four feet in length. Together, with our output as backup support, we made our official presentation to the executive. After about five minutes, he said, “That’s not what I want. Go back and do a different analysis.” After a second week of analysis we went back in to present our results. Again, we were sent back with new instructions.

I was afraid it was starting to look to the VP as if we were incompetent, but my boss persisted and we again repeated the go-do-something-else assignment. These “go and do” missions were not simple tasks. Each job required a lot of coding and processor time, and it was beginning to become clear that the VP had something specific he was looking for. I asked my boss if we could just ask him what he wanted to know. My manager agreed and we returned to the VP’s office. “What are you trying to learn about this product line?” I asked. He said he wanted to know which product tier sold best in each store. We thanked him and set off in a new direction.

A clustering analysis was not the method to answer his product question – regardless of how many times we ran it. He had heard about this type of analysis and thought it sounded like a good approach. In the end, the analysis turned out to be much simpler. We could have saved a lot of time and effort had we just asked a few questions at the outset.

It is important to understand what needs to be learned (how many, how much, what percentage), over what time frame (last week, the first quarter, or year to date), and how to present the data (raw numbers, percentage change, or variance from a baseline). The type of analysis that you should eventually perform depends on the answers to these and other questions. You should also find out if summary data is sufficient or if more data needs to be presented. Ask what types of statistics or metrics should be calculated.

These are just a few examples of the types of questions pertinent to any assignment. Others will come to mind once you start to inquire. These questions also help the requester develop a stronger sense of the request and better define his needs. Sometimes requests are just as simple as they sounded, but won’t it be nice to determine that on the front end?

Information gathered by this process is fundamental to building the SAS program in a later step. The answers will help determine which data sources to use, what procedures are needed in the program, and what type of output to generate. Avoiding rework is just one benefit. Asking questions also shows that you understand the business and are taking an active role in its performance. This adds value and elevates you from being “just the person who does the reports.”

Conclusion

Understanding the request is the first and most important measure you take in developing any SAS program. It sets the stage for everything that comes after it – from choosing data sources to the look of your code. It saves you time and frustration. It also shows that you understand the business and are participating in its success.

Part two of this series will address the topics, “Know Thy Data” and “Draw It Out.” Knowing what to get is at least as important as knowing where to get it and how to get it efficiently. Drawing it out means visualizing your program to develop better code.

Reference Link:

SUGI 30 Paper 146-30: Stephanie R. Thompson

New to SAS and New to Programming? What You Need to Do Before Typing Code
(<http://www2.sas.com/proceedings/sugi30/146-30.pdf>)

About the Author

Stephanie Thompson is the associate director for institutional research at the University of Memphis. With more than 20 years of programming experience, she applies statistical and modeling techniques to solve business problems in various manufacturing, retail, and academic environments using SAS and other programming languages. Major projects include product sales forecasting, evaluation of salary equity, modeling student success factors, and using predictive modeling to identify sales anomalies. She conducts SAS training courses on campus and has presented at SUGI 30, the Southern Association of Institutional Research Annual Conference, the SCSUG Arkansas SAS Day and the Memphis Area SUG. She also is scheduled to present at SUGI 31. Stephanie holds a B.S. in industrial engineering from the Rochester Institute of Technology, an M.B.A. from St. Bonaventure University, and is pursuing a Ph.D. in business administration with a concentration in economics from the University of Memphis.

The Little SAS Book for Enterprise Guide 3.0

By: Lora Delwiche and Susan Slaughter

List price: 49.95 USD

384 pages

ISBN: 1-59047-786-3

Publisher: SAS Press

Publication Date: November 2005

Description:

Learning to use SAS Enterprise Guide has never been easier! With The Little SAS Book for Enterprise Guide 3.0, Susan Slaughter and Lora Delwiche help you quickly become familiar with the SAS Enterprise Guide point-and-click environment. A series of carefully designed tutorials help you master the basics of the tasks you'll want to do most frequently. The reference section of the book expands on the tutorial topics, focusing on specific features. If you are new to SAS or new to SAS Enterprise Guide, this book will be an invaluable tool for you to use on your way to becoming an expert.

SAS Products Addressed: SAS Enterprise Guide

Releases: 9.1.3, 9.1.2, 9.1

Operating Systems: Windows

[Order today!](#)

Beyond the Semicolon: 20 Years of Lessons Learned as a SAS® User

By: Bryan K. Beverly

LESSON #1 – Know Your Environment

It is essential to know that your SAS environment extends "beyond the semicolon." Your career path traverses three communities of interest and practice. To be successful, you must be familiar with and able to navigate comfortably across all three areas: the management domain, the technology domain and the end-user domain.

Management consists of the sponsors, primary stakeholders and ultimate authority within your organization. The concerns of these persons center on requirement scope, delivery schedules, costs, quality and risk. Managers are people-oriented who need good political survival skills, and can see technology as both a means to an end and as a business-support function.

The technology domain is comprised of the information technology professionals. The concerns of the technologists consist of such issues as elegance and efficiency of computing platforms, applications and database structures. Technologists see their skills, tools, products and services as the goal of business.

The end users are the people who use information technology products and services. Their chief concerns are for tools that are easy to use and that make them more productive.

As a SAS developer, it is very easy to develop tunnel vision by concentrating solely on honing the skills, knowledge base and abilities related to programming. However, to become well-rounded, you must understand your position relative to the other communities of interest and practice. Think of your organization as an automobile manufacturer. The managers would be the company executives and the end users would be the people who buy the vehicles; your role is that of the worker on the assembly floor. What you do has value to you and your peers, but for the most part, the specifics of what you do are not the concern of the executives or the customers. Actually, the executives are the producers, the customers are the consumers, and the floor workers are the invisible but vital organs that close the gap between the producers and the consumers. Hence, as a SAS developer, you must be able to understand the needs and mindsets of the people that you support.

Here are some tips that will increase your advantage beyond the semicolon:

Learn what your organization does. Even if you are a consultant, you should learn the organization's purpose because the purpose creates the context for what you develop. By learning the problems that need solving, it makes it easier to translate requirements into technical specifications. It also demonstrates to management that your level of interest extends beyond your paycheck. From a technical perspective, knowing the business makes it easier to anticipate change because you have a feel for the dynamics.

Determine the financial health of the company and how SAS resources are financially supported. Funding for SAS resources varies from the private sector, public sector and academia. Knowing whether SAS is a mission-critical investment or if it is

something licensed because of some extra budget money will determine how much support you can expect to do your job.

Speak to managers in terms that they understand. Managers are not concerned with DO LOOPS or the number of variables in the Program Data Vector. You cannot expect managers to share your passion for technology – even if they were technologists at one time in their careers. When you communicate with managers, you must express how SAS is adding value, helping to meet business schedules or creating profit. Explaining the benefits is especially important if you are asked to justify a request for more SAS products or training. Purchasing software or sending you to training represents an expense and lost days of productivity. Hence, you will have to master the language of management to make your requests palatable.

Listen to the end users and learn to be flexible with them. You may be the technologist, but the end users are the ones who determine your raises, promotions and dismissals. When developing applications, be prepared for a one-time request to become a standard application. Moreover, regardless of how well you define the requirements, the requirements will change because end users often refine what they want after you produce what they have requested. To that extent, it makes sense to make your applications macro-intensive so that you can move quickly through the iterations.

Develop acceptance criteria at the beginning of the development projects. The truth about applications is that they are never finished unless they failed to deliver any value. Unless there is an agreement with the project sponsors on what defines “done,” there will be frustration on both sides. Development efforts involve a lot of mental and emotional consumption; hence, a sense of closure is needed so that any changes can be addressed with a fresh perspective.

Help managers understand the total cost of ownership for SAS projects. If you are asked for a time and cost estimate for developing a SAS application, you need to consider: (1) if additional training is needed, (2) if the platform will accommodate the programs and data sets associated with the project, (3) if a back-up system or continuity of operations site is needed, (4) how many people are needed for performing development tasks and (5) what level of programmers are needed since senior-level programmers are more expensive. The bottom line is that SAS projects involve more than just licensing software; they also involve supporting the infrastructure and environment.

There are many more words of wisdom that could be added to this list. But suffice it to say that, as a SAS professional, you can enhance your position by seeing the entire field. It is not enough to be a great technologist; you must also look beyond the semicolon to understand the needs and mindsets of the managers and end users. By sharpening this perspective, you add value to your career and to the people you support.

Information Quality for Business Intelligence and Data Mining: Assuring Quality for Strategic Information Uses

By Larry P. English, President and Principal, Information Impact International, Inc.
© 2005 Information Impact International

During the summer of 2005, scientists confirmed that there is a real pattern of global warming when they discovered and resolved an information quality problem in the data capture of surface temperatures. Satellites collecting data at the equator had reported temperatures that over time were relatively stable or showed a possible cooling trend. However, the satellites collecting that data had drifted off course and were reporting as daytime temperatures readings that were actually taken at night. Corrections to the data confirmed that there is warming at the equator that is consistent with surface warming around the globe.

This example illustrates how important trend findings can be obscured, misidentified or interpreted incorrectly if there are information quality problems anywhere in the information value chain.

Introduction

Information problems in information definition, data content, data preparation and information presentation can cause business intelligence processes to fail.

Here, I identify some of the critical information quality (IQ) problems in collection, preparing and presenting information for business intelligence and data mining along with IQ principles for mitigation or prevention.

Information Quality Issues for Data Mining and Statistical Analysis

Problems that hamper effective statistical data analysis stem from many sources of error introduction. First, data may not be clearly or accurately defined, causing a mismatch in the definition and the actual facts collected. Data can be captured inaccurately, or samples can be biased in record selection. Information quality decay causes data to become inaccurate when the characteristic of a real-world object changes. For example, if the price of an item changes, updated price values must be captured to assure the integrity of the analysis.

It is vital for the analyst to have or to conduct an information quality assessment to assure accuracy – not just validity – and completeness of data early in data preparation to allow time for any correction initiatives and preparation for mining.

Data preparation failure occurs when data is transformed in a way that is not able to be analyzed correctly by the data mining tools.

Finally, presentation graphics or display may not clearly convey the significance in the discovered patterns. Some examples:

- ❑ **Clear, correct, complete information definition:** An example of poor data definition comes from a survey taken by university students about student cell phone use. One attribute, “Included Minutes per Month,” was defined as: “Monthly allowed calling minutes that is written in the contract between the cell phone service provider and the customer.” The sample of data collected included “400,” “5000,” “9999,” “500 plus night,” “Unlimited,” “625 (5000 n&w),” “500/free night weekend,” “50,000”

and “-” (missing data). Due to the lack of clarity of the definition and absence of any data formatting, this data had to be manipulated and transformed for proper analysis.

IQ principle: Define data with business subject matter experts. Develop a consensus standard for values or data format. Provide training to information producers. Assess IQ for conformance to standards.

- ❑ **Measurement or data collection errors:** The global warming study described at the beginning of this article represents one kind of measurement error. Others include incorrect calibration of the temperature measurement device or improper placement of the measurement device.

IQ principle: Verify the calibration of measurement devices periodically. Assure consistent placement to capture data at a time, place or conditions that enable the identification of meaningful trends, e.g., taking surface temperature at the same location, at the same time of day by all satellites.

- ❑ **Sampling bias:** Data must be representative of the population being studied. If there is undiscovered sample bias and the population is not proportionately represented, the discovered trends will not be representative.

IQ principle: If sampling is made at information collection, assure that items or objects are selected by statistical sampling techniques, so that each object has equal likelihood of being selected. If data is being sampled, the data set must have a representative sample to the real-world collection of objects or events it represents. Record samples must be made using the same statistical sampling. If there are different strata in the population, assure a proportionate representation of each stratum, such as among the different classifications of frequent flyers and non-frequent flyers.

This having been said, there are some cases where you need to develop a “biased” sample that includes a higher proportion of outliers in order to predict rare events, such as fraudulent transactions. The modeling tools can better predict and can correct for the over-sampled cases or objects. The IQ principle here is that if the purpose of the analysis is to detect a rare class, use a training set in which the rare case is *over-represented*. If you specify correct prior probabilities of the rare case, the data mining predictions (posterior probabilities) will be correctly adjusted no matter what the proportions in the training set. If no prior probabilities are specified, the estimated posterior probabilities for the rare case will be too high, and the data mining predictions will be biased.

- ❑ **Missing values:** Data sources often contain observations that have missing values for one or more variables. Missing values can result from data collection errors, incomplete customer responses, actual system and measurement failures, or from a revision of the data collection scope over time, such as tracking new variables that were not included in the previous data collection schema. If an observation contains a missing value, then by default that observation is not used for modeling methods like neural network or linear regression. However, rejecting all incomplete observations may ignore useful or important information still contained in the non-missing variables. Rejecting all incomplete observations may also bias the sample, since observations that have missing values may have other things in common as well.

IQ principle: How should we treat missing data values? While there is no single correct answer, there are guidelines.

The first and best choice is to go back to the original real-world object and collect the data if it is knowable, such as the birth date of a person, and if the time of collection does not conflict with the time of collection of the other data, such as temperature on a different day from the other data points. For events, such as measurements at a point in time, there must be a reliable recording of the event data to capture it with accuracy. Estimating the “best” missing value replacement technique requires assumptions about the true (missing) data. For example, if a variable’s data distribution follows a normal population response, you may replace a missing value with the mean of the variable. Be aware that replacing missing values with the mean, median, or another measure of central tendency is simple, *but* it can greatly affect a variable’s sample distribution. Use these replacement statistics carefully and only when the effect is minimal.

Another imputation technique replaces missing values with the mean of all other responses given by that data source, such as the exit poll responses at a specific precinct. This assumes that the input from that specific data source conforms to a normal distribution. Another technique studies the data to see if the missing values occur in only a few variables. If those variables are determined to be insignificant, the variables can be rejected from the analysis. However, the observations can still be used by the modeling nodes.

At a point there may be too much missing data for acceptable statistical analysis, and you may have to discard such attributes from the data. Another strategy is to use a modeling technique like decision trees, which automatically handle missing values. Finally, you may want to create a missing value indicator attribute and use it as candidate predictors in the model. The presence or absence of a value itself can be predictive.

- ❑ **Inaccurate values:** In most cases, inaccurate data can cause processes to fail. The higher the frequency the more severe the failure. Some errors in variables such as salary amount or age, can tolerate some precision error without significant trend discovery failure.

IQ principle: As with missing data, the best form of correction is to return to the real-world object to re-measure or discover the correct value.

- ❑ **Value synonyms:** Where data does not have a standardized value set, there may be different data values that represent the same characteristic, such as unit-of-measure synonyms “12,” “Doz,” and “Dz.” This causes the problem of dilution of patterns involving unit of measure of one dozen items in an order unit. If there were a relatively normal distribution of values among the three synonyms, the frequency of occurrence of unit of measure of one dozen will represent only about one third of all items with a real unit of measure of one dozen.

IQ principle: Identify and standardize the synonyms to a single value. This cannot be done arbitrarily; you must involve the business subject-matter experts. The real solution requires this to be standardized in the source processes and databases.

- ❑ **Overloaded variable values:** Often, data that is not controlled contains values that do not represent the characteristic the variable was designed for. Knowledge workers, in the absence of a well-designed database may have to “force” new facts

into existing data elements. These overloaded fields create problems in trend correlation because they represent a different characteristic about the object or event that may bias the correlation of the original characteristic. For example, a Gender Code data element contained supposed “valid values” of “male,” “female,” “initials,” “ambiguous,” and “unknown.” The last three values did not represent gender; they represented why a gender-assignment routine was not able to determine gender by looking at the first name of the person.

IQ principle: For overloaded variables that represent multiple characteristics important to trend identification, break them out into separate variables and assure you have the correct definition. However, if the overloaded values are mutually exclusive, this will introduce missing data in both variables. In the Gender Code example above, if you were not able to contact the persons, you would have to provide a value of “unknown” and determine the impact of the missing data on your trend analysis.

- ❑ **Currency:** Currency represents the age of the data. Different trend analyses may require different ages of information. Identifying meaningful patterns requires having data of a common time period. For example, insurance policies have changes in business rules over time. You would not take policies in force 10 years ago and analyze them against the features of the comparable policies being sold today.

IQ principle: Understand the currency of all data required for a given model and assure data selection fits the age requirements.

- ❑ **Concurrency:** Concurrency is the timing difference of equivalence of data in one data store to another data store based on movement of data from one store to another. Records should be equivalent in content once the records reach a downstream data store. Data that is extracted from different databases may reach a given data set at different times. For example, orders reported today in the order fulfillment database will not be found in the historical order ODS (Operational Data Store) until tomorrow because they are extracted and loaded nightly. Shipments are not loaded until the end of the week. Returns are processed and loaded only after end of month. This causes problems in bringing data together to study patterns when the time periods of the transactions are different. To handle concurrency issues you must assure that the data extracted from multiple data sets represents a single time period.

IQ principle: Establish extract schedules (or extract transactions based on dates) from the various databases that will assure that transactions represent events or objects at a single point in time or time period. Solve the root causes by minimizing unnecessary redundant databases and information float. Eliminate the need for moving data to another database if that data can support all processes across the life cycle, such as persons or organizations that may be in a state of “prospect,” “active customer,” “preferred customer” or “inactive customer.”

Maintain appropriate date and time stamps and relationships of events to assure correlation of returns to the orders for which they are returned.

- ❑ **Outliers and anomalies:** Outliers are values that do not fit the expected set or range of valid values. Outliers may be errors or they may be accurate values but are beyond the realm of reasonability. A client of mine bought an item from a home improvement store for one penny (\$0.01). This item price is an outlier and is an

inaccurate price. With the cell phone study mentioned above, the 50,000 “Included Minutes per Month” was both an outlier and an error, inasmuch as 50,000 minutes represents 34 days, 17 hours and 20 minutes, impossible to cram into even a 31-day month. It apparently was an arbitrary number given for “unlimited minutes.”

Outliers, even when correct, cause problems in mining, for they can often skew the statistical analysis. Figure 1 shows the bias in the frequency distribution of the “Included minutes” with the 50,000 minute outlier.

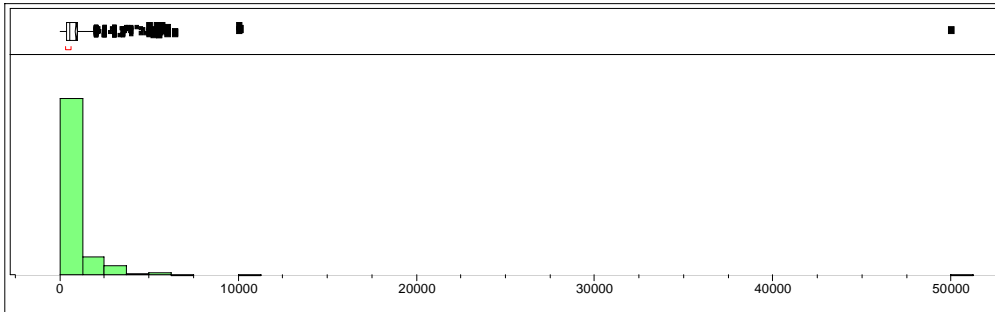


Figure 1 Outlier Bias in "Included Minutes" in Cell Phone Plans

IQ principle: Outliers that are in fact errors should be corrected to the valid value. The 50,000 minute value, created to estimate a plan with “unlimited” minutes, had the negative side effect of skewing the valid minute “maximums.” To handle this problem you might (1) drop the outlier(s), or (2) estimate a realistic maximum, derived from analysis of actual minutes used by customers with “unlimited” minute plans. See Figure 2.

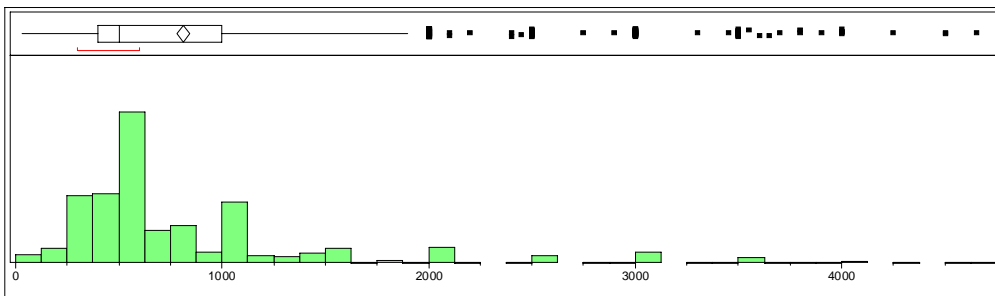


Figure 2 "Included Minutes" Without the 50,000 Minute Outlier

The data (Figure 2) still looks skewed, with the right half of the chart falling beyond three standard deviations, so you might explore methods to “normalize” the data, such as using a logarithmic scale or by standardizing the data from 0 to 1 (or -1 to 1). See the discussion about mapping attribute data to numeric values, below.

- ❑ **Mapping categorical data to numeric values.** Categorical data or attribute codes are not easily interpreted as to the relative relationship in trend analysis. In order to communicate to a statistical analysis tool, the codes or data must be analyzed to its relative position on a continuum. For example, a mining tool cannot differentiate between “Gold,” “Platinum,” or “Executive Platinum” frequent flyers by the text. This prevents such data from being useful in trend analysis.

IQ principle: Because most modeling techniques cannot interpret alphabetic codes, categorical or alphabetic information must be translated to ordered numeric values that can be interpreted for correlation. For example the “Included Minutes” values in the cell phone study above need to be translated to numbers between 0 and 1 that represents the relative degree of time between them because the modeling tool cannot interpret “unlimited” and “50,000” to be the same, nor “625 (5000 n&w)” and “500/free night weekend,” to be roughly the same. For example, the mapping might look something like this:

<i>Actual Data Value</i>	<i>Numerical “Equivalent”</i>
200	0.1
400	0.15
500 plus night	0.4
500 / free night & weekend	0.5
625 (5,000 night & weekend)	0.5
10,000	0.8
Unlimited	1.0

Figure 3 Mapping Categorical Data to Numerical Values for Analysis

- ❑ **Modeling errors (correlated attributes):** Some data elements may be redundant in that they tell you the same information about an object, such as “birth date” and “age,” “gender” and “personal title,” or “frequent flyer status” and “miles flown.” When multiple correlated attributes are included they will skew the analysis.

IQ principle: Identify pairs of attributes that have a direct or closely direct (or indirect) correlation and eliminate one of the attributes, generally the one derivable, such as “age” versus “birth date.”

- ❑ **Data preparation errors (enhancement):** Often, trends and patterns cannot be determined with much precision without having external data representing real factors that influence behavior. Data preparation includes not only internal data, but also external data. For example, unexpected cold spells or warm spells influence behavior of purchases. Interest rate fluctuations, political events and other external factors can provide critical variables useful for predicting behavior.

IQ principle: You must understand your predictive model and determine whether internally known data are sufficient or whether external data can provide variables that are required to correctly interpret customer or product behavior.

Once you determine that external data is needed, you must acquire it, *and* you must assess its quality. Find out what IQ processes the information provider uses to assure quality of their information.

Conclusion

Assuring the quality of information for effective data mining and business intelligence begins long before the extraction and preparation of the data for mining. It begins with clear, accurate and complete definition of the data itself (the information product

specification), and with error-proofing and controlling the processes that capture the data (for completeness, accuracy and precision), and with maintaining the quality as it may be subject to change.

Without first assuring the quality of the information definition, the data content the data preparation and the information presentation, the business intelligence conclusions will be “neither good business nor intelligent.”

About the Author

Larry English, President and Principal of Information Impact International, is an internationally recognized authority in information management and information quality. He has consulted in 29 countries on five continents. His TIQM[®] methodology applies quality principles to information quality management and has been implemented in many organizations worldwide. English was featured as one of the “21 Voices for the 21st Century” in the American Society for Quality's *Quality Progress* journal. His book, *Improving Data Warehouse and Business Information Quality*, was hailed as “the Information Quality Bible for the Information Age” by Masaaki Imai, the creator of the Kaizen quality system. It has been translated into Japanese by the first information services organization to win the Deming Prize for Quality. English writes the “Plain English about Information Quality” column in *DM Review* magazine. Now in its 10th year, the column is consistently one of the magazine's most popular. English serves as an editorial adviser for *DM Review* and the *Quality Assurance Journal*. He chairs The Information Quality Conferences in the United States and London and is co-founder of the International Association for Information and Data Quality (IAIDQ). English can be reached at Larry.English@infoimpact.com.

Test Drive Self-Paced e-Learning at SUGI

Sign up at the self-paced e-learning booth located in the demo area for free access to award-winning lessons in programming, certification training and business intelligence.

Read more <http://support.sas.com/selfpaced>

New Green Belt Training from the Design Institute for Six Sigma

Transform your process into profit. Using JMP software, you can learn about Six Sigma philosophy, tools and the DMAIC methodology, and apply it to your manufacturing or transactional breakthrough projects. Two new courses are now available.

Read more <http://support.sas.com/difss>

Perform a Fuzzy Merge Within a Range Using DATA Step Component Objects

Use the hash iterator to search a hash object for a value that is within a range from a second data set.

Note: For detailed information regarding object dot programming in the DATA step, please refer to SAS 9.1 Language Reference: Concepts, Using DATA Step Component Objects.

```
data one;
  input lastname: $15. typeofcar: $15. mileage;
datalines;
Jones Toyota 7435
Smith Toyota 13001
Jones2 Ford 3433
Smith2 Toyota 15032
Shepherd Nissan 4300
Shepherd2 Honda 5582
Williams Ford 10532
;

data two;
  input startrange endrange typeofservice & $35.;
datalines;
3000 5000 oil change
5001 6000 overdue oil change
6001 8000 oil change and tire rotation
8001 9000 overdue oil change
9001 11000 oil change
11001 12000 overdue oil change
12001 14000 oil change and tire rotation
14001 14999 overdue oil change
15000 15999 15000 mile check
;

data out(keep=lastname typeofcar mileage typeofservice);
  length startrange endrange 8 typeofservice $35;

  if _n_=1 then do;
    declare hash h(dataset:'two',hashexp:4);
    h.definekey('startrange');
    h.definedata('startrange','endrange','typeofservice');
    declare hiter hiter('h');
    h.definedone();
    call missing(startrange, endrange, typeofservice);
  end;
  set one;
  rc=hiter.first();
  do while (rc=0);
```

```
    if startrange le mileage le endrange then leave;
    rc1=hiter.next();
end;
run;

proc print;
run;
```

Webcasts and Events

Focus on BI Server Enhancements

Thursday, Feb. 16

1:00-2:30 p.m. ET

Tune into this demo for a sneak peek at new enhancements for SAS Enterprise BI Server that will spotlight the essentials for driving business innovation.

SAS Press Webinar Series

Tuesday, Feb. 21

12:30-1:15 p.m. ET

Topic: programmatically measure SAS application performance on any platform with the new LOGPARSE SAS macro. Join us for this free Webinar!

SUGI 31

March 26-29

San Francisco

This popular user event is only 40 days away! Be sure to attend and take home useful tips and techniques.

JMP User Conference

June 20-21

Cary, NC

Attend exciting and insightful sessions, plus roundtable discussions, a Scripting workshop, a Genomics Discovery event and exclusive training courses.