

"Using Excel with SAS® to Combine Customer Survey Data Sets," or "A Simple But Tedious Method for Combining Customer Survey Data Sets"

Cyndi V. Thompson and Mark E. Thompson, Forefront Economics

ABSTRACT AND INTRODUCTION

Companies routinely survey their customers on many issues including customer satisfaction, product awareness, company image, and demographics; however, due to changes in company needs, personnel, and external vendors, customer surveys tend to change in content and format from survey to survey. These changes make combining data sets from more than one survey difficult. However, the rewards from a richer customer information data set can pay large dividends through more informed market planning.

In an effort to reap these rewards, an approach for combining customer survey data sets from several years into one customer research data set was developed utilizing Excel to build SAS code for applying data conversion macros and consistent formatting and labeling. This method was used to successfully merge the results of five years of customer surveys for a Pacific Northwest company and is presented in this paper.

OBJECTIVE

The marketing department of a major Pacific Northwest company desired to use the results of five years of annual customer surveys for strategic planning and marketing purposes. Files containing the results of each of the annual surveys were available in ASCII format. However, the questions, order of questions, and data values of answers varied with each implementation of the annual survey. In order to exploit the data to the fullest, the survey results needed to be accessible in a common format across all years. Hence, the primary objective of this project was to combine each of five surveys into one data set with common variable names, data values, formats, and labels.

METHODOLOGY

The methodology developed for this project can be broken down into three major steps including:

1. Review and categorize every question in each survey.

2. Assign SAS data conversion macros and standard variable attributes.
3. Standardize each of the annual survey data sets and combine into one common SAS data set.

Each of the steps, along with an overview of the process and a description of the major elements used in the process, is shown in Chart 1.

The three-step process uses an Excel spreadsheet (ES) to summarize and store the results of steps one and two. SAS is used to standardize and merge the data sets using the information in the ES. In this sense, the ES controls the entire process, determining when and how data values are converted and which standard variable attributes are applied.

Construction of the ES is tedious work, reviewing the specifics of each question in each survey. Approximately 80 to 90 percent of the total effort involved constructing the ES. Each row in the ES corresponds to a survey question. The columns of the ES are grouped into three major sections, as shown in Table 1. Section 1 contains the standard variable attributes each question will be assigned, while Section 2 contains a set of columns for each survey year where information unique to that survey year is entered and stored. Finally, Section 3 contains SAS code automatically generated based on information entered in Sections 1 and 2 of the ES. In this form, the ES provides a convenient way to quickly see if a question was asked in any particular year and also provides the information needed by SAS to merge the survey responses across years.

Chart 1. Three Step Process to Standardizing Yearly Surveys

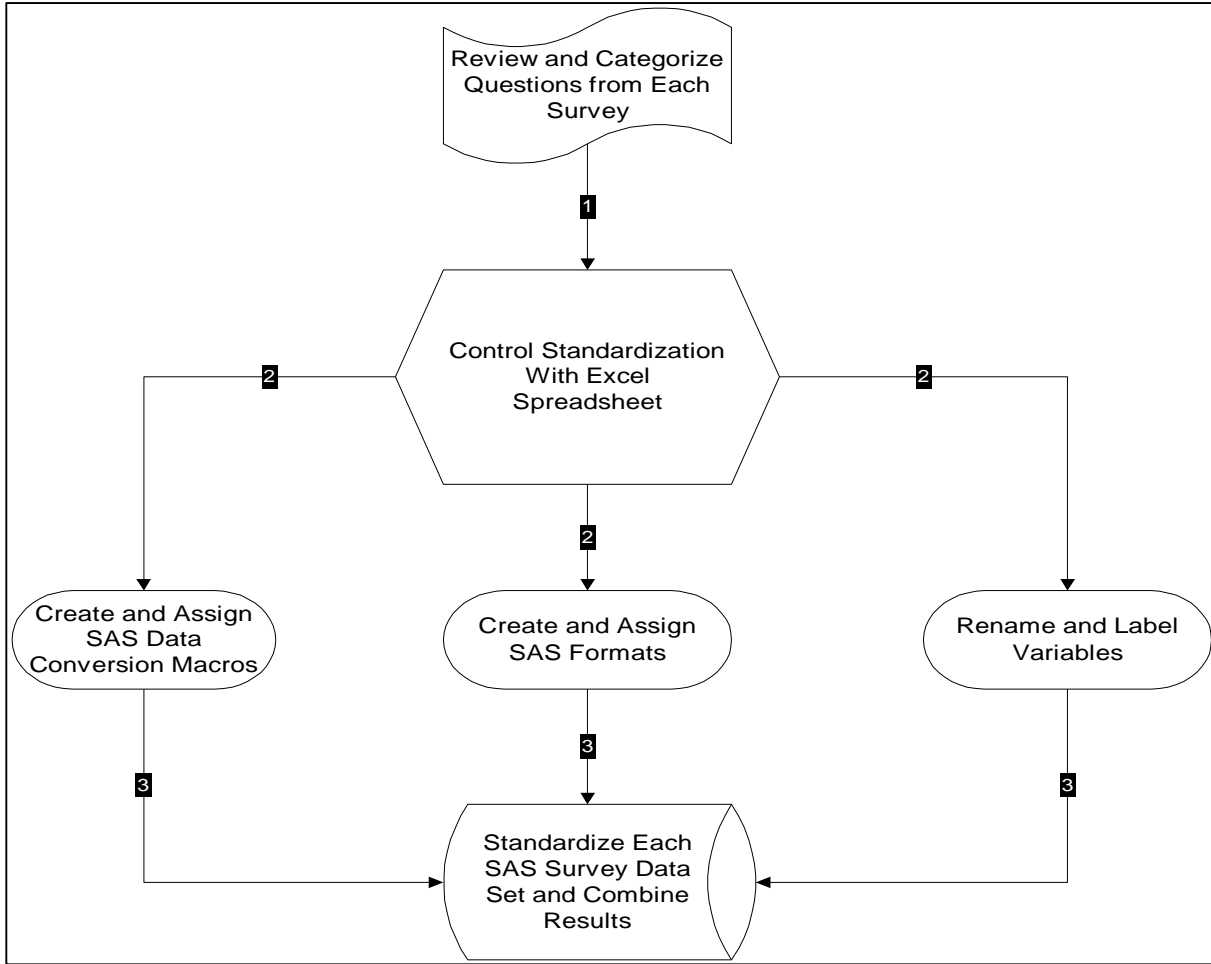


Table 1. Overview of the Excel Spreadsheet

Section 1: Standard SAS Variable Attributes	Section 2: Survey Specific Information	Section 3: SAS Code Generated from Sections 1 & 2
<ul style="list-style-type: none"> • Name • Label • Format 	<ul style="list-style-type: none"> • Name in Survey Data Set • Data Conversion Required 	<ul style="list-style-type: none"> • Data Conversion Macro • Renaming • Formatting • Labeling

**Table 2. Specific Example for Sections 1 & 2 of the Excel Spreadsheet
“What Year was Your Home Built?” (Row 376)**

Section 1: Standard SAS Variable Attributes			Section 2: Survey Specific Information			
All Years			1996 Survey		1995 Survey	
Variable Name	SAS Label	SAS Format	Variable Name	Conversion Macro	Variable Name	Conversion Macro
d_yrblt	Year Home Built	dk	d6	yr2to4	d12	miss4

EXAMPLE

Perhaps the best way to understand the process is to see a specific example. The question “What year was your home built?” will be used to illustrate the data-standardization process. Table 2 shows the first two ES sections, the standard SAS variable attributes and the survey specific variable information, for the year-built question (row 376 of the ES). In Section 1, the year-built question was assigned the variable name, *d_yrblt*, the label, “Year Home Built”, and the format, *dk*, for all years.

As shown in Section 2 of Table 2, the year-built question was asked in both 1995 and 1996 (as well as other years not shown). The original surveys listed the year-built question as d6 and d12 in the 1996 and 1995 surveys, respectively. Responses in 1996 were provided as two digits (e.g. “74”) while responses in all other years were four digits (e.g. “1974”).

Many instances of variations in data values across survey years, such as in this example, were encountered in the project. Data conversion macros were developed to standardize data values across years and assigned to the appropriate variables in Section 2 of the ES. To convert question d6 data values of the 1996 survey, the data conversion macro, YR2TO4, was developed and assigned to variables in Section 2 of the ES:

```
%macro yr2to4(var);
if &var lt 97 then &var=1900+&var;
else if &var=97 then &var='d';
else if &var>97 then &var='x';
%mend yr2to4;
```

In the original 1996 data set, numeric values of “97” were used to code a response of “don’t know”, while values greater than “97” were considered illogical based on the topic of the question. To standardize these responses, “d” and “x” were one of four values converted to designate missing value types. Note that the year-built question was assigned the format, *dk*, in Section 1 of the ES. The *dk* format statements are presented below:

```
proc format library=library;
value dk
. n = 'Not Applicable'
. d = 'Dont Know'
. r = 'Refused'
. x = 'Missing-Other';
```

For the 1995 survey, the year-built question was recorded as d12 and assigned the MISS4 macro to convert the value “9997” to the special SAS missing value code of “d”. In this way all numeric values designating responses other than valid numeric values were converted to missing values. Several variations of the missing value conversion macro were used to convert numeric values to missing. The MISS4 macro is shown below:

```
%macro miss4(var);
if &var=9997 then &var='d';
%mend miss4;
```

Columns from Section 3 of the ES, representing the automatically generated SAS code, are shown in Table 3. Specifically, Table 3 lists the SAS code for the year-built question of the 1996 survey generated from the information shown in Table 2. SAS code to convert data and rename, label and format variables is generated in Section 3 of the ES.

**Table 3. Specific Example of Section 3 of the Excel Spreadsheet for the 1996 Survey
“What Year was Your Home Built?” (Row 376)**

Section 3: SAS Code Generated from Sections 1 & 2			
Data Conversion Macro	Rename	Format	Label
%yr2to4(d6);	Rename d6=d_yrblt;	Format d_yrblt dk.;	Label d_yrblt='Year Home Built';

In the final step of the process, the SAS code contained in the ES is read by SAS using Dynamic Data Exchange (DDE). The SAS program used to read the information in the ES and standardize the data sets is shown below:

```
filename macro96 dde
'Excel |[Var_Compare_Rename.xls]Sheet1!R4C30:R414C30';

filename code96 dde
'Excel |[Var_Compare_Rename.xls]Sheet1!R4C31:R414C33';

* STEP1: Revise data values using the predefined data conv macros;

%include 'd:\fe\example\sas\data conv macros.sas' / source2;

data cts96.cus96f;
set cts96.cussur96;
%include macro96 / source2; * reads macro statements from Excel;
run;

* STEP2: Rename variables and assign predefined formats;

proc datasets library=cts96;
modify cus96f(label='Survey, 1996 - Formatted');
%include code96 /source2; * reads in SAS code from Excel;
run;
```

The previous program is run while the ES is open in memory. As shown in the program, the data conversion macros are applied to the variables in the unformatted survey data set. Next, the variables are renamed, labeled, assigned formats and stored in the formatted data set for that format year.¹

¹ If desired, both steps could be completed in one data step by adding the “%include code96 / source2;” line of

The last part of the process involves merging all survey year data sets using the following program:

```
filename varlist dde
'Excel |[Var_Compare_Rename.xls]Sheet1!R4C29:R414C29';

Data library.allcts(label='Customer Surveys, 1992- 1996');
set cts96.cus96f (in=c96)
cts95.cus95f (in=c95)
cts94.cus94f (in=c94)
cts93.cus93f (in=c93)
cts92.cus92f (in=c92);
if c96 then suryear=1996;
if c95 then suryear=1995;
if c94 then suryear=1994;
if c93 then suryear=1993;
if c92 then suryear=1992;
label suryear='Survey Year';
keep
%include varlist /source2;
suryear;
run;
```

Only variables assigned a common SAS name in Section 1 of the ES are kept in the combined database. This allows for a simple method of excluding variables from the combined data set due to insufficient data, lack of presence over time, or any reason the analyst believes to be appropriate.

RESULTS

The final result of this project was a SAS database with nearly 10,000 customer responses to 255 questions. Although several questions were not present in all of the survey years, many questions were present in most years, providing a rich basis for market analysis and research. Working with the client, Forefront Economics has assisted with the following uses of the database:

Step 2 after the “% include macro96 / source2;” line in Step 1. Having two separate steps allowed for easier data checks on the data conversion results.

1. Developed a targeted marketing campaign from the results of a CHAID based decision tree. The resulting tree was used to score the client's customer database in terms of their predicted response to a specific product offer.
2. Customer segmentation and sales forecasting using detailed information from the database to define purchasing patterns by segments.
3. Answers to a variety of questions from internal users of marketing information.

The uses listed above, especially the data mining algorithms employed in item one and the data requirements of the segmentation strategy in two, would not have been possible without the extensive data organizing work discussed in this paper.

To most of us, working with decision trees and customer segmentation strategies to address strategically important marketing objectives is far more stimulating than the tedious work involved in gathering, cleaning, and standardizing data. However, the data organization work is clearly the underpinning that supports any analysis of the data for strategic insights. And, as analysis unearths valuable nuggets of information, the sweat from preparing the data, that made the discovery possible, becomes much more bearable.

AUTHORS

The authors welcome any questions or comments.

Cyndi V. Thompson and Mark E. Thompson
Forefront Economics
3800 SW Cedar Hills Blvd., Suite 299
Beaverton, OR 97005
Phone: (503) 626-1657
Fax: (503) 626-6320
E-mail: cyndi@forecon.com, mark@forecon.com

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.