

The Data Detective's Toolkit

**Cutting-Edge Techniques and
SAS[®] Macros to Clean, Prepare,
and Manage Data**

A pair of black-rimmed glasses is shown from a top-down perspective, resting on a blue, textured surface that resembles a book cover or a piece of fabric. The glasses are the central focus of the lower half of the image.

Kim Chantala

The correct bibliographic citation for this manual is as follows: Chantala, Kim. 2020. *The Data Detective's Toolkit: Cutting-Edge Techniques and SAS® Macros to Clean, Prepare, and Manage Data*. Cary, NC: SAS Institute Inc.

The Data Detective's Toolkit: Cutting-Edge Techniques and SAS® Macros to Clean, Prepare, and Manage Data

Copyright © 2020, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-952363-04-7 (Hardcover)

ISBN 978-1-952363-00-9 (Paperback)

ISBN 978-1-952363-01-6 (Web PDF)

ISBN 978-1-952363-02-3 (EPUB)

ISBN 978-1-952363-03-0 (Kindle)

All Rights Reserved. Produced in the United States of America.

For a hard copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

December 2020

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

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

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

Contents

About This Book	v
About The Author	ix
Acknowledgments	xi
Chapter 1: Advantages of Using the Data Detective's Toolkit	1
Introduction.....	1
An Overview of the Data Detective's Toolkit.....	2
Summary	7
Chapter 2: The Data Detective's Toolkit and SAS	9
Introduction.....	9
Preparing Your SAS Data Set	9
Fundamental SAS Macro Concepts	15
The Output Delivery System.....	30
Summary	31
Chapter 3: Codebooks: A Roadmap to Your Data	33
Introduction.....	33
Understanding Codebooks	33
Using the %TK_codebook Macro	36
Example 3-1: Create a Codebook with Potential Problem Reports	40
Inside the Toolkit: %TK_codebook	52
Summary	53
Chapter 4: Customizing Codebooks	55
Introduction.....	55
Example 4-1: Embellishing Titles	55
Example 4-2: Add a Logo to Your Codebook	59
Example 4-3: Codebook Output Data Set and Default Design	61
Example 4-4: Create a Custom Design for Your Codebook	71
Summary	81
Chapter 5: Catalog Your Data	83
Introduction.....	83
Using the %TK_inventory Macro	84
Using the %TK_xwalk Macro.....	89
Summary	95

Chapter 6: Detecting and Correcting Data Errors	97
Introduction	97
Harmonizing Data Sets: Using the %TK_harmony Macro.....	98
Example 6-1: Harmonizing Two Data Sets.....	99
Inside the Toolkit: How %TK_harmony Works	103
Finding Duplicates: Using the %TK_find_dups Macro.....	108
Example 6-2: Identifying Duplicates Based on Multiple Variables.....	108
Inside the Toolkit: How %TK_find_dups Works.....	111
Summary.....	114
Chapter 7: Inspect and Edit Flow through Skip Patterns	115
Introduction	115
Understanding Skip Patterns.....	116
Identifying Skip Patterns in a Survey.....	117
Traditional Method of Auditing Skip Patterns	120
Example 7-1: Using the %TK_skip_edit Macro	123
A Blueprint to Using %TK_skip_edit.....	130
Example 7-2: Automated Method of Checking Skip Patterns.....	132
Inside the Toolkit: How %TK_skip_edit Works	148
Summary.....	154
Chapter 8: Create and Validate New Variables	157
Introduction	157
Coding Variables.....	157
Easy Ways to Check Variable Construction.....	162
Summary.....	174
Appendix A: Your Part in the Data Life Cycle	177
Introduction	177
Understanding the Data Life Cycle.....	177
Summary.....	185
Appendix B: Skip Pattern Data Codebook.....	187
Introduction	187
SAS Program to Create Codebook	187
Appendix C: Research Data Codebook	193
Introduction	193
SAS Program to Create Codebook	193
Index.....	197

About This Book

What Does This Book Cover?

Data professionals who survived deep cuts in funding during the financial crisis of 2007–2008 had to develop innovative methods of data preparation. This book presents innovative data tools and techniques that helped data managers, practitioners, and programmers survive these challenges by reducing the cost and time needed for data management while improving the quality of data prepared with their use. These tools include SAS macros as well as ingenious ways of using SAS procedures and functions.

Is This Book for You?

This book is designed to help automate many of the tasks performed to turn raw data into analysis-friendly data. These tasks are often filled with a mix of irksome and strenuous activities that stand between you and data that can be used. This book will help preparers of the data in different ways:

Intermediate and Advanced users:	You will reduce your workload and improve the quality of your data by using the SAS macro programs included with this book to automate error-checking and create documentation for your project data. Using these programs included with this book will alleviate the tedious nature of data preparation by automating the identification of inconsistencies and anomalies in raw data.
-------------------------------------	---

- Novice users:** If you are not familiar with SAS and are just starting to work with data, you will need to get help from a more experienced programmer to use the SAS macro programs that automatically produce codebooks, reports highlighting problems in the data, inventories of available data sets, and crosswalks showing commonalities of multiple data sets. These are covered in Chapters 3 through 6. Once the SAS statements are set up to run the SAS programs producing these reports, you will find it easy to assist in the detective work of data preparation. Examining these reports will really help you get to know your data, and you can help to solve problems identified in the data. Focusing on the discussion of the output in examples of this book will help you learn to interpret these reports and lead to a better understanding of your data. Skip the sections in each chapter titled “Inside the Toolkit” that discuss the macro program statements in detail.
- Data managers and Research staff:** You will be able to choose from the many automated reports that function as roadmaps into your data, snapshots of data quality and monitoring, and use these reports to improve communication between your programmer, practitioners, and the data collection sponsors.
- All users:** No matter what your level of experience, you should read Chapter 1, “Advantages of Using the Data Detective’s Toolkit” and Appendix A, “Your Part in the Data Life Cycle.”

What Are the Prerequisites for This Book?

Familiarity with SAS programming (the DATA step and basic rules of the SAS language) as well as manipulating SAS data with procedures such as PROC CONTENTS, PROC MEANS, and PROC FREQ provide adequate prerequisites for working with the SAS programs and techniques discussed in this book. Familiarity with basic features of the SAS macro language would be useful to run the SAS macro programs that accompany this book. For programmers new to the SAS macro language, detailed instruction is provided in Chapter 2 with information about using SAS.

What Should You Know about the Examples?

Software Used to Develop the Book's Content

The output in this book was created with SAS 9.4. Most programs in this book can be run using BASE SAS on the platform that you typically use. A few of the examples use procedures found in the SAS/STAT software.

Example Code and Data

All data used in the examples in this book was simulated. Any resemblance to actual data sets is purely coincidental. Errors and other anomalies were purposely added to the data to illustrate special features described in this book to clean, prepare, and perform quality control checks on your data.

You can access the example code and data for this book by linking to its author page at <https://support.sas.com/chantala>.

Output and Graphics

All output in this book was created with the SAS Output Delivery System. Your output might look slightly different because changes in the appearance of some tables have occurred during the formatting of this book.

We Want to Hear from You

SAS Press books are written *by SAS Users for SAS Users*. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books to do the following:

- Sign up to review a book
- Recommend a topic
- Request information about how to become a SAS Press author
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through saspress@sas.com or https://support.sas.com/author_feedback.

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: sas.com/books.

Learn more about this author by visiting her author page at <http://support.sas.com/chantala>. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

About the Author



Kim Chantala is a Programmer Analyst in the Research Computing Division at RTI International with over 25 years of experience in managing and analyzing research data. Before joining RTI International, she was a data analyst at the University of North Carolina at Chapel Hill. In addition to providing data management and analytical services at the University, she taught workshops on analyzing survey data, focusing on the problems of sample weights and design effects. Kim believes that the real challenge in data analysis is bridging the gap between raw or acquired data and data that is ready to analyze. This inspired her to develop computerized data management tools revolutionizing the way data is prepared, allowing users to improve the quality of their data while lowering the cost of data preparation.

Kim earned a BS in Engineering Physics from the Colorado School of Mines and an MS in Biometrics from the University of Colorado.

Learn more about this author by visiting her author page at <http://support.sas.com/chantala>. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

Chapter 1: Advantages of Using the Data Detective’s Toolkit

Introduction	1
An Overview of the <i>Data Detective’s Toolkit</i>	2
%TK_codebook	3
%TK_inventory.....	4
%TK_xwalk.....	5
%TK_find_dups	5
%TK_harmony.....	5
%TK_skip_edit	6
%TK_max_length	6
Summary.....	7

Introduction

You will find the right data tools in this book for creating project data that is ready for exploration and analysis. Using these tools will reduce the amount of time needed to clean, edit, validate, and document your data. Advantages of using the techniques in this book include:

- Accomplishing more while doing less by automating and modernizing the typical data preparation activities
- Beginning at the end by creating research-ready data sets and documentation early in the project with continual updates and improvements throughout collection and preparation
- Keeping the sponsor or lead research investigators engaged by providing codebooks, crosswalks, and data catalogs for review early in the project, thus including them as part of quality control surveillance for the data

This book includes a set of SAS macro programs that automate many of the labor-intensive tasks that you perform during data preparation. Using these macro programs will help guard against compromising quality control and documentation efforts due to rigid project budgets and timelines. You will be able to automate producing codebooks, crosswalks, and data catalogs. Innovative logic built into these macro programs computerizes monitoring the quality of your data using information from the formats and labels created for the variables in your data set. You will receive concise reports identifying invalid data – such as out of range values, missing data, redundant, or contradictory data.

You only need to create a SAS data set with labels and formats assigned to each variable to use these macro programs. It could not be easier or faster to create data that you can trust. The SAS macro programs accompanying this book are available at no charge and can be downloaded from the author page for this book at support.sas.com/chantala.

In the following chapters, you will learn how to use these macro programs to make your job easier and create higher quality data. This chapter introduces you to the macro programs accompanying this book and highlights how they can help solve many of the problems that you face in data preparation.

An Overview of the *Data Detective's Toolkit*

Data preparation is a heroic task, often with inconsistencies and anomalies in raw data that you must resolve to make the data usable. Your job will include:

- Investigating unexpected or missing values
- Resolving conflicting information across variables
- Mitigating incorrect flow through skip patterns
- Examining incomplete data
- Combining multiple data sets with different attributes
- Documenting changes in data collection methods or instruments during collection

Reconciling these issues requires careful investigation and alleviation during data cleaning and preparation. Rapid advancement in software for both data collection and analysis has encouraged more complex data to be collected. This has caused greater challenges for you as the programmer responsible for turning it into high-quality, research-friendly data. Advances in software to help you solve these issues has progressed at a slower pace than advances in software for analysis or collecting data. This lag in development of computerized tools for data preparation has motivated the development of the macro programs included with this book.

These macro programs have been developed to help you work more efficiently when preparing data and automate much of the tedious work in identifying and correcting problems in your data. Table 1-1 lists the macro programs provided with this book and what they will do for you.

Table 1-1: List of Macro Programs in the *Data Detective's Toolkit*

Macro Data Tool	Function
%TK_Codebook	Create codebook, monitor data quality, and identify problems in variables that need further investigation
%TK_inventory	Create a catalog showing the inventory of all SAS data sets in a folder

Macro Data Tool	Function
%TK_xwalk	Create a crosswalk table used to show the relationship of variables across a group of data sets
%TK_find_dups	Find records with duplicate identification variables in a data set
%TK_harmony	Harmonize two data sets with varying file formats, variable naming conventions, data types, labels, and formats
%TK_skip_edit	Analyze skip patterns to identify contradictory responses, incorrect flow through prescribed skip pattern, and recode incorrect responses
%TK_max_length	Dynamically create the list of variables with maximum storage lengths needed for the LENGTH statement when you are merging or concatenating two data sets

The only requirement for using these data tools is creating SAS data sets with formats and labels assigned to each variable. Once you have the SAS data set created, you will only need a simple line of SAS code to invoke each of the data tools. The first three macro programs create useful documentation for your data sets. You can create them at the beginning of the project and benefit by having them available for everyone in your team.

%TK_codebook

The first tool, %TK_codebook, creates a codebook. This macro uses one statement requiring only that you provide the name and location of your SAS data set, the library for the formats assigned to the variables, and a name for your codebook as shown below:

```
%TK_codebook(lib=work,
             file1=test,
             fmtlib=library,
             cb_type=XL SX,
             cb_file=&WorkFolder./Test_CodeBook.xlsx ,
             var_order=internal,
             cb_output = my_codebook,
             cb_size=BRIEF,
             organization = One record per CASEID,
             include_warn=YES);
```

It could not be easier to create a codebook for your data set. But the best feature is yet to come! %TK_codebook will also examine each variable and print informative reports about potential

problems. Using information from the label and format assigned to each variable, the %TK_codebook macro warns your data team about variables having the following problems:

- Values missing from the assigned format
- Out of range values
- Missing labels
- No assigned format
- Having 100% missing values
- No variation in the response value

For each variable automatically examined, you would have to write several SAS statements and examine multiple tables to figure out which variables need further examination. If your data set has 1000 variables, you will write SAS statements to create over 2000 tables, examine each table manually to identify problems, then summarize the problems that need investigation. With the reports from %TK_codebook, you are presented with a concise summary of only those variables needing close examination and why they need examination. You will spend your time correcting problems rather than writing repetitive SAS code and examining piles of SAS output. Chapter 3 teaches you how to use %TK_codebook to create a codebook and potential problem reports. These reports identify variables having the problems listed earlier in this section. Chapter 4 teaches you how to customize your codebook in both appearance and adding additional information about variables to the data used to create a codebook.

%TK_inventory

A catalog of all the SAS data sets for your project can be created at any time during the data life cycle with %TK_inventory by simply providing the full path name of the folder where the data sets reside:

```
libname SAS_data "/Data_Detective/Book/SAS_Datasets";
```

```
%TK_inventory(libref=SAS_data);
```

For each data set in the folder associated with libref SAS_data, %TK_inventory will provide information about the following characteristics:

- Data set name
- Data set label
- Creation date
- Number of observations
- Number of variables

This catalog provides a concise summary of the data sets and where they are located, providing an ideal document for communicating a listing of available data. It makes it easier for you and

your team to track the progression of developing your data sets. Chapter 5 teaches you how to use the %TK_inventory macro tool.

%TK_xwalk

The %TK_xwalk tool creates a data crosswalk to help you identify equivalent variables in multiple data sets as well as differences in the attributes of variables having the same name in more than one data set. Again, you only need to use one short statement with a list of data files for %TK_xwalk to create your crosswalk.

```
%TK_xwalk(SetList = SAS_Data.studya SAS_Data.demog SAS_Data.health);
```

This statement creates a mapping of variables across two or more distinct data sets. Reviewing the crosswalk will help you identify variables used to merge the data as well as avoid truncating values when merging or concatenating data sets. You will learn to use %TK_xwalk in Chapter 5.

%TK_find_dups

You will need to examine each data set verifying that variables uniquely identifying an observation occur only on one observation. You will need to do this on every data set that is created, possibly each time changes are made to program creating your data set. With just a few strokes of the keyboard %TK_find_dups will easily do this for you:

```
%TK_find_dups(dataset=work.STUDY, one_rec_per=CASEID*WAVE,  
              up_output=STUDY_DUPS);
```

The output from %TK_find_dups includes the following:

- Table showing the number of observations having identical values of the unique identification variables (CASEID*WAVE)
- Table showing the values of the identification variables that are duplicated across observations.
- Output data set with values of duplicated identification variables that you can use to extract the duplicated observations from your data set.

Chapter 6 teaches you how to use %TK_find_dups.

%TK_harmony

The %TK_harmony macro can identify possible problems with merging or concatenating two data sets. It is very simple to use, requiring only one statement providing the names of the data sets being harmonized, and nicknames for each data set used in the harmony report created by the %TK_harmony.

```
%TK_harmony(set1= SAS_data.demography_a1,  
            set1_id=Web,  
            set2= SAS_data.demography_a2,  
            set2_id=Paper,  
            out=harmony_results);
```

%TK_harmony compares the two data sets and creates a report with the following information:

- Variables unique to each set
- Variables with the same name having different labels
- Variables with the same name having different data types or lengths

You will learn to use the %TK_harmony macro and the output tables in Chapter 6.

%TK_skip_edit

Skip patterns are used in data collection to ensure that only relevant questions are asked each person participating in the survey. For example, your study might have a set of questions that are asked only of female participants. Male participants would have missing values for all of these questions.

The %TK_skip_edit macro can be used to validate skip patterns as follows:

- Validate that a variable follows the expected pattern of nonmissing/missing values when the variable is part of the skip pattern logic
- Handle special recoding to correct inconsistencies in skip patterns and help users understand why a variable is missing

For example, suppose question PG1 asks women the number of pregnancies they have had in their lifetime. This would not be asked if the participant was male. Question DEM2 in the survey asks each participant their sex (1=female, 2=male). %TK_skip_edit uses this information to examine this skip pattern for you and change the value of PG1 to missing if a male responded to that question. You only need to set up a format identifying the values of a variable that cause a SKIP, and then pass this information to TK_skip_edit:

```
proc format;  
  value SKIP2f 2='2=SKIP';  
run;  
%TK_skip_edit(check_var = PG1,  
  skip_vars = DEM2,  
  skip_fmts = DEM2 skip2f.);
```

%TK_skip_edit produces an annotated table reporting results from analyzing data flow through the skip pattern and any edits that were made to the data to resolve inconsistencies in the data flow. You will learn more about skip patterns and how to use the %TK_skip_edit macro in Chapter 7.

%TK_max_length

SAS prints the following message in your log file to warn you that there is a mismatch in the storage length of variables in the data sets being combined in a DATA step:

```
WARNING: Multiple lengths were specified for the variable VAR_NAME by  
input data set(s). This can cause truncation of data.
```


When you see this message, it means that the values stored in VAR_NAME were possibly truncated when the data sets were combined with a MERGE or SET statement. To prevent this from happening, you can use the %TK_max_length macro to create a macro variable named &MAX_LENGTHS that contains information about the variables common to two data sets but have different storage lengths. This list includes the name and the longest defined length of each variable. Macro variable &MAX_LENGTHS can be used in the LENGTH statement in the DATA step to prevent truncation of data values when two data sets are combined. The SAS statements below show how easy it is to use %TK_max_length and a LENGTH statement to prevent truncating data values:

```
%TK_max_length(set1=My_Data.teleform_data, set2=My_data.web_data);

data survey_v2;
length &max_lengths;
set My_Data.teleform_data My_Data.web_data;
run;
```

You will learn more about using the %TK_max_length in Chapter 2.

Summary

This chapter explained the benefits of using this book for data cleaning, preparation, and management. Using these macro programs reduces the time needed to prepare data that you can trust. You will automate creating documentation for your data by easily creating codebooks, crosswalks, and data catalogs with just a few strokes on the keyboard. The way you clean data will be modernized enabling you to easily to detect, investigate, and correct inaccurate data values in your data set.

The strength of using these macro programs to automate cleaning data and creating documentation lies in their general applicability and simplicity of use. The only requirement for you to use them is having a SAS data set with labels and formats assigned to the variables.

You will use these tools in every stage of the life cycle of your data. Read Appendix A to understand more about the data life cycle. You will read about the common activities in every stage of the data life cycle, learning how your data flows through each stage from inception of the idea to acquire your data through archival at project end. You will find useful checklists showing recommended tasks for cleaning, using, distributing, and archiving your data.

Ready to take your SAS[®] and JMP[®] skills up a notch?



Be among the first to know about new books,
special events, and exclusive discounts.


support.sas.com/newbooks

Share your expertise. Write a book with SAS.

support.sas.com/publish

Continue your skills development with free online learning.

www.sas.com/free-training

 sas.com/books
for additional books and resources.


THE POWER TO KNOW.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. * indicates USA registration.
Other brand and product names are trademarks of their respective companies. © 2020 SAS Institute Inc. All rights reserved. M2063821 US.1120