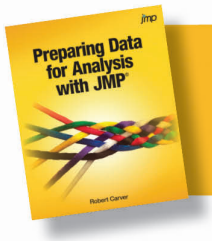
Preparing Data for Analysis with JMP®

Robert Carver



An Introduction to Preparing Data for Analysis with JMP®.
Full book available for purchase here.

Contents

About This Book

What Does This Book Cover?

In a 2008 interview, Google's chief economist, Hal Varian, remarked:

> *I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades…*[1]

Perhaps the very least attractive aspect of the "sexy job" is the work involved in assembling, reconciling, tidying, cleaning, and otherwise preparing data from various sources *before* the serious work of processing, extracting value, visualizing, and communicating. Although data preparation typically consumes an enormous share of the time in most projects, it receives comparably little attention in the data analytics literature.

It is as if data preparation is a dark art or a nasty family secret, widely acknowledged but not spoken about in polite company. This book is all about using the extensive capabilities of JMP to facilitate and regularize the phases of preparing data for analysis.

This book is entirely and exclusively about the stages that precede the actual analysis in a statistical investigation. It covers methods for extracting data from various sources and in different formats and converting them to JMP data tables. Because so many projects call for merging multiple data tables, we see how the powerful JMP Query Builder for Tables facilitates such operations, enabling the analyst to manage data consolidation at scale and at relatively high speed.

As practitioners know all too well, once the data are all in one place, the real adventure begins. JMP can also speed the work of "rounding up the usual suspects" of dirty data: missing observations, outliers, mismatched key fields, and implausibly perfect relationships. After identifying issues, we have an array of alternatives for resolving, mitigating, and managing them. Finally, the book also covers options for communicating data and results to non-users of JMP.

JMP typically offers multiple options to tackle a given issue, and the book presents both the alternatives and a sense of what to use when. For techniques that are out of the mainstream (for example, Principal Components Analysis for data reduction and mitigation of missing cases), chapters provide both introductions to the methods and reliable references to other sources.

Anticipating that some preparatory work might need to be done repeatedly, another recurring theme is reproducibility. Whether a reader is conversant with JMP Scripting Language, users can create a reproducible audit trail by pointing and clicking. The processes illustrated in the book can be preserved by simply saving the scripts that are being written in the background every step of the way.

Oddly enough for a book that prominently features statistical software, there is scarcely any coverage of analytic techniques here, except insofar as a technique helps identify or repair a data problem. The analytic techniques are always in mind, because data preparation must be informed by the varying requirements of different techniques. However, the analytic platforms remain off-stage, as it were.

The book also does not cover JMP fundamentals. There are numerous books, videos, and other training materials for the new user, and readers who are just encountering JMP for the first time are better served to start elsewhere.

Is This Book for You?

If you are a practitioner working with messy data from multiple sources, this book can help. In particular, this book is for practitioners with access to raw data and a pressing need to, in Varian's words, "understand it, to process it, to extract value from it, to visualize it, to communicate it."

What Are the Prerequisites for This Book?

You should have some grounding in statistical methods, and have a working acquaintance with JMP. Some of the features illustrated in the book are available only in JMP Pro, so ideally readers have a JMP Pro license.

What Should You Know about the Examples?

This book includes tutorials for you to follow to gain hands-on experience with JMP. After working through the varied examples within the chapters, you should be able to apply concepts and techniques to your own data. The topics for examples are drawn from various fields. If you need to learn data preparation skills for your own work, you are probably aware that analysis projects require a combination of subject-area knowledge, statistical thinking, and familiarity with data structures. Because each reader brings different domain expertise, the examples are intended to be accessible to most readers regardless of professional background.

Some use data tables that are included with your JMP installation Sample Data Library. Others come from public domain sources, and all of the data tables shown in the book are available for download at http://support.sas.com/publishing/authors/carver.html.

Software Used to Develop the Book's Content

This book was prepared using JMP Pro 13. Readers with earlier versions will find that some menus have changed and that some functionality is not available to them.

Example Data

As noted earlier you can access the example code and data for this book by linking to its author page at http://support.sas.com/publishing/authors/carver.html.

Output and Graphics

Most of the images in the book were captured from within JMP Pro 13, and a few others from websites. Printed editions of the book are in black-and-white, which plainly does not convey the effective use of color produced by JMP in some reports. Electronic editions do render graphics in color.

Where Are the Exercise Solutions?

Most chapters include extended hands-on examples with detailed instructions. To get the most from the book, please follow along and actually work through the exercises and examples. There are no "homework" exercises at the end of chapters, because the goal of the illustrations is to help readers with their own projects and data. Hence, the solutions appear right within the chapter, and your screen should match up with the many screen captures provided in the text.

Acknowledgments

I am very grateful to SAS Press for the invitation and encouragement to write this book, and particularly to Sian Roberts and Julie Platt for their patience throughout the process. Kathy Underwood speedily and expertly copy-edited the entire manuscript. Some chapters would truly have not materialized without expert consultations and suggestions from members of the JMP development team including Brady Brady, Michael Hecht, Eric Hill, Don McCormack, Heman Robinson, and Russ Wolfinger.

The JMP Early Adopter program was indispensable in preparing the manuscript. Thanks to Jeff Perkinson and Daniel Valente for access and updates. JMP Academic Ambassadors Curt Hinrichs, Mia Stevens, and Volker Kraft all assisted in more ways than they can imagine.

Eric Hill, Mike Vorburger, and Richard Zink caught errors, rescued me from rabbit holes, and recommended numerous improvements to earlier drafts. Thank you for your invaluable service as reviewers.

To friends and colleagues outside and inside the "JMPiverse" for data, for perspective, and for pedagogical recommendations: Max Harpers of GroupLens and the University of Minnesota, Professor Nick Horton of Amherst College, Rob Lievense of Perrigo, and Professor Michael Salé of Stonehill College.

Finally, to my wife (and sometimes technical consultant on database matters) Donna, from whom I've stolen too many Sundays together to write this book—Thank you, sweetie.

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 https://support.sas.com/publishing 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: https://support.sas.com/publishing.

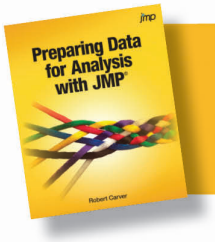
---

[1] McKinsey & Company. 2009. "Hal Varian on how the Web challenges managers." Available at http://www.mckinsey.com/industries/high-tech/our-insights/hal-varian-on-how-the-web-challenges-managers.

About The Author

Robert Carver is Professor of Business Administration at Stonehill College in Easton, Massachusetts, and Senior Lecturer at the International Business School at Brandeis University in Waltham, Massachusetts. At both institutions, he teaches courses on business analytics in addition to general management courses. He is the author of *Practical Data Analysis with JMP®, Second Edition*. His primary research interest is statistics education, and he is an Associate Editor for the Journal of Statistics Education. A JMP user since 2006, Carver holds an A.B. in political science from Amherst College in Amherst, Massachusetts, and an M.P.P. and Ph.D. in public policy from the University of Michigan at Ann Arbor.

Learn more about this author by visiting his author page at http://support.sas.com/publishing/authors/carver.html. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.



Chapter 1: Data Management in the Analytics Process

Introduction

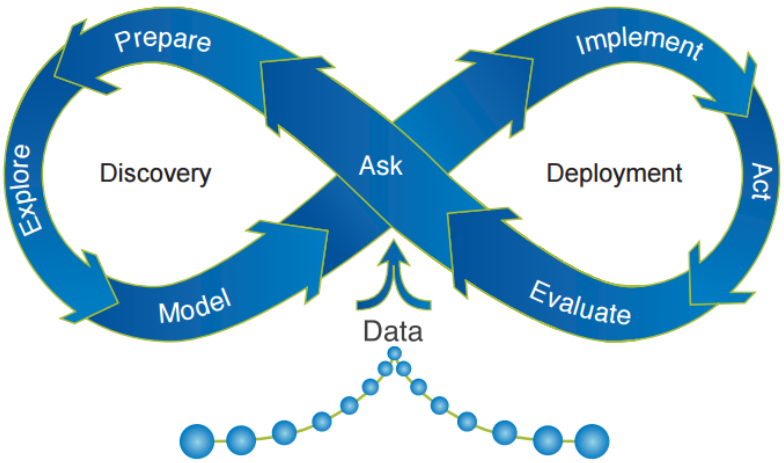
Although reliable estimates are difficult to come by, there seems to be consensus that data preparation—locating, assembling, reconciling, merging, cleaning, and so on—consumes something like 80% of the time required for a statistical project (Press 2016). In comparison to the literature about building statistical models and performing analysis, there are relatively few books written on the topic of data preparation. (Some good examples include McCallum 2013, Osborne 2013, and Svolba, 2006.) This book addresses just that—the unglamorous, time-consuming, laborious, and sometimes dreaded "dirty work" of statistical investigations. This work is variously known as data wrangling, data cleaning, the "janitorial work," or simply data management. Equally important, this book is about data management using JMP, so that JMP users can do nearly all of the wrangling tasks using one software environment, without having to hop off and onto different platforms.

This first chapter places data management and preparation in the context of a larger investigative process. In addition, it introduces some over-arching themes and assumptions that run through the eleven chapters to follow. As a starting point, let's understand that the work of data management occurs within a process and often within an organizational context.

A Continuous Process

Successful analyses require a disciplined process similar to the one shown here in Figure 1.1 (SAS, 2016). In such a process, a question gives rise to fact-gathering, analysis, and implementation of a solution. The solution, in turn, is monitored and gives rise to further questions. Though process is often portrayed as circular, this particular depiction has the added appeal of connoting an infinite loop with data playing a central role. However, as we'll see in the pages that follow, the step identified as "Prepare" is not drawn remotely to scale. That's the bit that reportedly can require 80% of the time in a project, and will consume an even larger proportion of the pages in this book.

Figure 1.1: One Model of the Analytics Life Cycle (SAS)



For the most part, this book breaks down that single step into component parts, explaining obstacles that arise and offering techniques to address them. JMP is particularly well suited to expediting the smaller steps that make up the work of preparation. In addition, we'll step into the exploratory stage in later chapters, cycling back as exploration reveals the need for additional pre-processing and preparation. In any event, though, the process properly starts with questions that can be addressed with data.

Asking Questions That Data Can Help to Answer

One underlying assumption in this book is that we undertake statistical investigations because someone has questions (a) that might be answered via empirical investigation and (b) whose answer potentially has measurable value to an organization, society, or an individual. In other words, we'll assume that the questions *matter* to someone. That is, there are benefits to be reaped from finding answers. This also implies that those asking the questions are likely to be attuned to the costs of securing the data and performing relevant analysis.

If you are reading this book, you probably have questions that are important to your work or areas of interest. Which environmental hazards have the greatest impact on respiratory health? Which customers will most likely prefer product A over product B? Do tax cuts encourage employers to create new jobs?

Although the questions driving a study often arise within an organization, the data most germane to the question might reside inside or outside the organization. In this discussion, let's refer to the organization asking the questions as the client. The client might have some of the needed data within its own files and databases (you hope the data are in digital form). Other data might be freely available in the public domain, while other data might be available for a price. Yet other data could be proprietary belonging to competitors, and still other information might not yet have been gathered or curated by anyone.

Early in the process, the analyst or analysis team need to determine the specific data that will serve to address the questions posed or build the models that will have value. This requires domain knowledge, an understanding of what is organizationally feasible and an awareness of what data are available. Perhaps hardest of all, it can require a knowledge of what the team does not know. For some problems, the point of the analysis is *feature selection*—that is, identifying the variables that matter most, that explain or predict a dependent variable or outcome. These problems cannot be resolved through software, but the search for relevant data certainly can be either facilitated or impeded by the software.

Sourcing Relevant Data

Once the analysts have initial ideas about the types of data to obtain, the challenges of locating pertinent data can be considerable. As noted, the organization's own data stores might not have exactly what is needed. Sometimes, organizational dynamics or legal or ethical considerations can impede access.

If we're lucky, external data reside in the public domain and are easily accessible. In other cases, the data are proprietary and belong to a competitor or to an entity that will make them available at a price. Worst of all, sometimes the data simply do not exist in any accessible form, so the analytics team will either need to gather data or use proxy variables, "near enough," if you will, to represent the constructs of interest in a study.

This book has little to say about the many hurdles of data procurement. Chapter 3 discusses some of them, but in general this is an area where domain knowledge is key. The analytics team and the client organization need to know where to look for relevant data.

Reproducibility

As you begin to extract data, it is quite important to document the process in detail for the sake of reproducibility. Some projects are one-time, unique tasks, but for others there is value in being able to reproduce all of the steps taken. Where there are considerations of intellectual property rights, this is critical. If you simply want the ability to audit the task for completeness and critical evaluation, a full record and audit trail is indispensable.

The individual most likely to want to reproduce the process in the future is *you*. Whether others will one day retrace your footsteps, it is altogether likely that you or your team will go back to fetch additional variables, or to build another project on the foundation of this one. So, as the saying goes, be kind to your future self and document (Gandrud 2015, p. 7).

As we'll see throughout the book, documentation is quite straightforward within a JMP session. What's more, we can document selectively preserving results and scripts that we ultimately judge worth saving without having to enshrine every error and false start.

Combining and Reconciling Multiple Sources

Many data modeling projects call for data from multiple sources. In addition to locating all of the variables that you might use, analysts often confront the need to reconcile disparities across data sources. Nonstandard abbreviations or coding schemes, different representations of times and dates, and varying units of measurement abound. Before all of the data can be assembled into a single, well-organized table suitable for analysis, the differences need to be ironed out.

Because the irregularities can present themselves at different stages of the preparation phase, they are discussed in several chapters. Chapter 5 through 8 cover many of the aspects of combining and reconciling data.

Identifying and Addressing Data Issues

Once data sources have been identified and targeted, you must consider issues of data integrity. It is the rare data source that supplies complete, accurate, timely raw tabulated data. For many analytical purposes, we'll want (or require) data tables that are in third normal form – variables in columns and observations in rows, but some data sources won't be organized that way. Tables will have missing cell, sparse arrays (mostly zeros), or erroneous data values. There will be outliers, skewed distributions, non-linear relationships, and so on.

The later chapters devote considerable attention to (a) ways of detecting such data problems and (b) addressing them. Here again, JMP has extensive functionality to ease the way forward. Note also, that our goal is to "address" the issues. That is, we cannot always resolve or eliminate the problems. There are generally ways to mitigate data issues when they cannot be directly cured. These are among the matters taken up in Chapters 9 through 11.

Data Requirements Shaped by Modeling Strategies

The analysis plan for a project influences, or should influence, the data management and preparation activities. Modeling methodologies have their own requirements for the organization of a data table, for units of analysis, and for data types. As a simple and familiar example, models built on paired observations will expect to find data pairs in separate columns.

Hence, you need to be constantly mindful of the stages to follow when preparing a set of data for analysis. Data preparation happens within the context of the full investigative cycle for a reason, and that goes beyond variable selection. Chapters 7 and Chapters 9 through 11 touch on these issues.

Plan of the Book

The investigative process is neither linear nor purely sequential, though we depict it as a logical sequence. That notwithstanding, analysts regularly need to loop back to an earlier phase along the way to a successful conclusion.

In a similar way, books do need to present information in sequence. But readers are free to depart from the plan, to iterate, and to bypass materials that are already familiar. Still, it's helpful to have a mental map of the plan before deviating from it.

I've developed the book in three main sections. Part I consists of Chapters 1 through 3. This part builds a foundation that explains the investigative cycle, introduces common methods used to organize raw data, and reviews the array of common challenges arising in different data sources.

Part II gets down to the "how" of acquiring and structuring a collection of variables for building models. Chapters 4 through 8 present some background concepts and theory, and work through several examples of importing data into JMP from foreign sources and then combining disparate elements into single JMP data tables. Running through Part II is a comprehensive illustrative example about the competitive success of world nations in the summer Olympic Games.

Lastly, Part III (Chapters 9 through 12) covers approaches to some of the thorniest data preparation problems: how to detect problematic data, how to deal with missing data, and how to transform and otherwise modify variables for analysis. Chapter 12 reverses, in a sense, the work of data acquisition by demonstrating ways to share or export data and results to platforms other than JMP.

Conclusion

The main take-away from this chapter is that data management is a large and complex piece of a larger and more complex process. It might not be the most glamorous part of modeling and data analytics, but it is as essential to success as proper soil preparation is to abundant crop yields in agriculture. This chapter has described the process with broad strokes and outlined the tasks the lie ahead.

Before we begin the search for data to wrangle and manage, it's useful to understand some things about the underlying structure of data that we might find. The next chapter reviews and explains the most common ways of organizing and representing raw data.

References

Asay, Matt. 2016 "NoSQL keeps rising, but relational databases still dominate big data." *TechRepublic.com*, April 5, 2016. Available at http://www.techrepublic.com/article/nosql-keeps-rising-but-relational-databases-still-dominate-big-data/.

Carver, Robert., Michelle Everson, John Gabrosek, Nicholas Horton, Robin Lock, Megan Mocko, Allan Rossman, Ginger Holmes Rowell, Paul Velleman, Jeffrey Witmer, and Beverly Wood. 2016. *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016*. Alexandria VA: American Statistical Association.

Gandrud, Christopher. 2015. *Reproducible Research with R and RStudio, 2nd Edition.* Boca Raton, FL: CRC Press.

McCallum, Q. Ethan, ed. 2013. *Bad Data Handbook*. Sebastopol, CA: O'Reilly Media.

Osborne, Jason W. 2013. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Thousand Oaks CA: Sage.

Press, Gil. 2016. "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says." *Forbes*, March 23, 2016. Available online at http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#4fe4cf1e7f75.

SAS Institute Inc. 2016. SAS Institute white paper. "Managing the Analytical Life Cycle for Decisions at Scale: How to Go From Data to Decisions as Quickly as Possible." Cary, NC: SAS Institute Inc.

Svolba, Gerhard. 2006. *Data Preparation for Analytics Using SAS*. Cary, NC: SAS Institute Inc.

Tintle, Nathan., Beth L. Chance, George W. Cobb, Allan J. Rossman, Soma Roy, and Jill VanderStoep. 2014. *Introduction to Statistical Investigations.* Hoboken, NJ, Wiley.

Wild, C. J., and M. Pfannkuch. 1999. "Statistical Thinking in Empirical Enquiry." *International Statistical Review*, 67(3), 223-265.