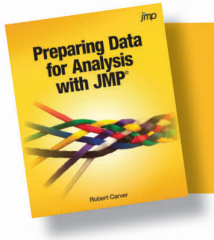

An Introduction to Preparing Data for Analysis with JMP®.
Full book available for purchase here.

Contents

About This Book ix
About The Author xiii
Chapter 1: Data Management in the Analytics Process 1
Introduction 1
A Continuous Process 2
Asking Questions That Data Can Help to Answer 2
Sourcing Relevant Data 3
Reproducibility 3
Combining and Reconciling Multiple Sources 4
Identifying and Addressing Data Issues 4
Data Requirements Shaped by Modeling Strategies 4
Plan of the Book 5
Conclusion 5
References 5
Chapter 2: Data Management Foundations 7
Introduction 7
Matching Form to Function 8
JMP Data Tables 9
Data Types and Modeling Types 10
Data Types 10
Modeling Types 10
Basics of Relational Databases 12
Conclusion 13
References 14
Chapter 3: Sources of Data and Their Challenges 15
Introduction 15
Internal Data in Flat Files 15
Relational Databases 16

External Data on the World Wide Web 16
User-Facing Query Interfaces 16
Tabular Data Pages 19
Evolving WWW Data Standards 19
Ethical and Legal Considerations 19
Conclusion 20
References 21
Chapter 4: Single Files 23
Introduction 23
Review of JMP File Types 23
Common Formats Other than JMP 25
MS Excel 25
Text Files 32
SAS Files 39
Other Data File Formats 41
Conclusion 42
References 42
Chapter 5: Database Queries 43
Introduction 43
Sample Databases in This Chapter 44
Connecting to a Database 44
Extracting Data from One Table in a Database 48
Import an Entire Table 48
Import a Subset of a Table 49
Querying a Database from JMP 52
Query Builder 52
An Illustrative Scenario: Bicycle Parts 55
Designing a Query with Query Builder 57
Query Builder for SAS Server Data 64
Conclusion 66
References 67
Chapter 6: Importing Data from Websites 69
Introduction 69
Variety of Web Formats 70
Internet Open 70
Common Issues to Anticipate 72
Conclusion 74
References 75

Chapter 7: Reshaping a Data Table 77
Introduction 77
What Shape Is a Data Table? 78
Wide versus Long Format 78
Reasons for Wide and Long Formats 79
Stacking Wide Data 79
Unstacking Narrow Data 82
Additional Examples 83
Stacking Wide Data 83
Scripting for Reproducibility 85
Splitting Long Data 86
Transposing Rows and Columns 90
Reshaping the WDI Data 91
Conclusion 94
References 94
Chapter 8: Joining, Subsetting, and Filtering 97
Introduction 97
Combining Data from Multiple Tables with Join 98
Saving Memory with a Virtual Join 102
Why and How to Select a Subset 103
A Brief Detour: Creating a New Column from an Existing Column 104
Row Filters: Global and Local 107
Global Filter 107
Local Filter 109
A More Durable Subset 110
Combining Rows with Concatenate 111
Query Builder for Tables 113
Back to the Movies 113
Olympic Medals and Development Indicators 114
Conclusion 121
References 122
Chapter 9: Data Exploration: Visual and Automated Tools to Detect Problems 123
Introduction 123
Common Issues to Anticipate 124
On the Hunt for Dirty Data 125
Distribution 126
Columns Viewer 126

Multivariate (Correlations and Scatterplot Matrix) 128
More Tools within the Multivariate Platform 129
Principal Components 129
Outlier Analysis 130
Item Reliability 130
Explore Outliers 130
Quantile Range Outliers 132
Robust Fit Outliers 133
Multivariate Robust Outliers 133
Multivariate k-Nearest Neighbors Outliers 134
Explore Missing 135
Conclusion 136
References 137
Chapter 10: Missing Data Strategies 139
Introduction 139
Much Ado about Nothing? 140
Four Basic Approaches 142
Working with Complete Cases 142
Analysis with Sampling Weights 142
Imputation-based Methods 144
Recode 144
Informative Missing 145
Multivariate Normal Imputation 147
Multivariate SVD Imputation 149
Special Considerations for Time Series 151
Conclusion and a Note of Caution 153
References 153
Chapter 11: Data Preparation for Analysis 155
Introduction 155
Common Issues and Appropriate Strategies 156
Distribution of Observations 157
Noisy Data 157
Skewness or Outliers 160
Scale Differences among Model Variables 162
Too Many Levels of a Categorical Variable 163

High Dimensionality: Abundance of Columns 167
Correlated or Redundant Variables 167
Missing or Sparse Observations across Columns 168
A PCA Example 168
Abundance of Rows 173
Partitioning into Training, Validation, and Test Sets 173
Aggregating Rows with Summary Tables 176
Oversampling Rare Events 177
Date and Time-Related Issues 179
Formatting Dates and Times 179
Some Date Functions: Extracting Parts 180
Aggregation 181
Row Functions Especially Useful in Time-Ordered Data 181
Elapsed Time and Date Arithmetic 182
Conclusion 183
References 183
Chapter 12: Exporting Work to Other Platforms 185
Introduction 185
Why Export or Exchange Data? 185
Fit the Method to the Purpose 186
Save As 186
Export to a Database 187
Export to a SAS Library 188
Exporting Reports 189
Interactive Graphics 190
Static Images: Graphics Formats, PowerPoint, and Word 192
Conclusion 193
References 193
Index 195