

Table of Contents

Preface	vii
PART I - Data Ingestion	1
Chapter 1: Tabular Formats	3
Tidying Up	4
CSV	9
Sanity Checks	10
The Good, the Bad, and the Textual Data	13
The Bad	13
The Good	18
Spreadsheets Considered Harmful	21
SQL RDBMS	29
Massaging Data Types	30
Repeating in R	34
Where SQL Goes Wrong (and How to Notice It)	36
Other Formats	42
HDF5 and NetCDF-4	44
Tools and Libraries	45
SQLite	50
Apache Parquet	52
Data Frames	54
Spark/Scala	56
Pandas and Derived Wrappers	58
Vaex	59
Data Frames in R (Tidyverse)	61
Data Frames in R (data.table)	63
Bash for Fun	64
Exercises	65
Tidy Data from Excel	65
Tidy Data from SQL	67
Denouement	68

Chapter 2: Hierarchical Formats	71
JSON	72
What JSON Looks Like	74
NaN Handling and Data Types	78
JSON Lines	82
GeoJSON	85
Tidy Geography	88
JSON Schema	92
XML	99
User Records	100
Keyhole Markup Language	102
Configuration Files	108
INI and Flat Custom Formats	109
TOML	110
Yet Another Markup Language	114
NoSQL Databases	119
Document-Oriented Databases	121
Missing Fields	123
Denormalization and Its Discontents	125
Key/Value Stores	127
Exercises	130
Exploring Filled Area	130
Create a Relational Model	131
Denouement	133
Chapter 3: Repurposing Data Sources	135
Web Scraping	136
HTML Tables	137
Non-Tabular Data	140
Command-Line Scraping	146
Portable Document Format	148
Image Formats	153
Pixel Statistics	156
Channel Manipulation	159
Metadata	161
Binary Serialized Data Structures	165
Custom Text Formats	170
A Structured Log	171
Character Encodings	175
Exercises	182
Enhancing the NPY Parser	182
Scraping Web Traffic	183

Denouement	185
PART II - The Vicissitudes of Error	187
Chapter 4: Anomaly Detection	189
Missing Data	191
SQL	192
Hierarchical Formats	196
Sentinels	197
Miscoded Data	201
Fixed Bounds	205
Outliers	210
Z-Score	211
Interquartile Range	216
Multivariate Outliers	219
Exercises	221
A Famous Experiment	221
Misspelled Words	223
Denouement	225
Chapter 5: Data Quality	227
Missing Data	228
Biasing Trends	232
Understanding Bias	233
Detecting Bias	236
Comparison to Baselines	240
Benford's Law	244
Class Imbalance	246
Normalization and Scaling	253
Applying a Machine Learning Model	256
Scaling Techniques	257
Factor and Sample Weighting	262
Cyclicity and Autocorrelation	267
Domain Knowledge Trends	271
Discovered Cycles	278
Bespoke Validation	282
Collation Validation	283
Transcription Validation	287
Exercises	291
Data Characterization	291
Oversampled Polls	294
Denouement	296

PART III - Rectification and Creation	297
Chapter 6: Value Imputation	299
Typical-Value Imputation	301
Typical Tabular Data	302
Locality Imputation	309
Trend Imputation	313
Types of Trends	314
A Larger Coarse Time Series	317
Understanding the Data	318
Removing Unusable Data	321
Imputing Consistency	322
Interpolation	325
Non-Temporal Trends	327
Sampling	332
Undersampling	335
Oversampling	339
Exercises	345
Alternate Trend Imputation	345
Balancing Multiple Features	346
Denouement	348
Chapter 7: Feature Engineering	351
Date/Time Fields	352
Creating Datetimes	354
Imposing Regularity	355
Duplicated Timestamps	358
Adding Timestamps	359
String Fields	364
Fuzzy Matching	367
Explicit Categories	372
String Vectors	379
Decompositions	388
Rotation and Whitening	389
Dimensionality Reduction	391
Visualization	394
Quantization and Binarization	398
One-Hot Encoding	406
Polynomial Features	409
Generating Synthetic Features	411
Feature Selection	413

Exercises	417
Intermittent Occurrences	417
Characterizing Levels	418
Denouement	419
PART IV - Ancillary Matters	421
Closure	423
What You Know	423
What You Don't Know (Yet)	424
Glossary	427
Other Books You May Enjoy	463
Index	467

Doing the Other 80% of the Work

It is something of a truism in data science, data analysis, or machine learning that most of the effort needed to achieve your actual purpose lies in cleaning your data. The subtitle of this work alludes to a commonly assigned percentage. A keynote speaker I listened to at a data science conference a few years ago made a joke—perhaps one already widely repeated by the time he told it—about talking with a colleague of his. The colleague complained of data cleaning taking up half of her time, in response to which the speaker expressed astonishment that it could be so little as 50%.

Without worrying too much about assigning a precise percentage, in my experience working as a technologist and data scientist, I have found that the bulk of what I do is preparing my data for the statistical analyses, machine learning models, or nuanced visualizations that I would like to utilize it for. Although hopeful executives, or technical managers a bit removed from the daily work, tend to have an eternal optimism that the next set of data the organization acquires will be clean and easy to work with, I have yet to find that to be true in my concrete experience.

Certainly, some data is better and some is worse. But *all data is dirty*, at least within a very small margin of error in the tally. Even datasets that have been published, carefully studied, and that are widely distributed as canonical examples for statistics textbooks or software libraries, generally have a moderate number of data integrity problems. Even after our best pre-processing, a more attainable goal should be to make our data *less dirty*; making it *clean* remains unduly utopian in aspiration.