

A Brain-Friendly Guide

Head First Data Analysis



Predict
your raise
with linear
regression

**A learner's guide to
big numbers, statistics,
and good decisions**

Sell more toys by
optimizing your
business model



Experiment to
discover who your
customers *really* are



Overcome
your
cognitive
biases



Load important
statistical concepts
directly into your brain



Clean messy data
for efficient analysis



O'REILLY®

Michael Milton

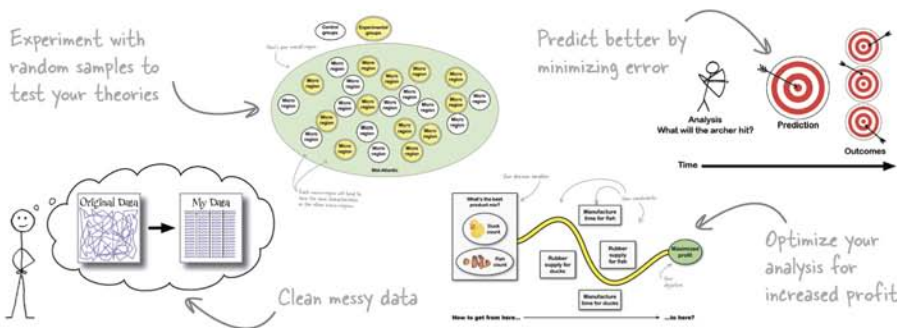
Head First Data Analysis

Information Theory/Data Analysis

What will you learn from this book?

There's a whole world of data out there, and it's your job to make sense of it all. Where to begin? *Head First Data Analysis* helps you organize your data in Excel or OpenOffice, take it further with R, find meaningful patterns with scatterplots and histograms, draw conclusions using heuristics, predict the future by experimenting and testing hypotheses, and display findings with clear visualizations.

Whether you're a product developer researching the viability of a new product, a marketing manager gauging the effectiveness of a campaign, a salesperson presenting data to clients, or a lone entrepreneur responsible for all these data-intensive functions and more, *Head First Data Analysis* is a complete learning experience for making data the most useful tool in your business.



What's so special about this book?

We think your time is too valuable to waste struggling with new concepts. Using the latest research in cognitive science and learning theory to craft a multi-sensory learning experience, *Head First Data Analysis* uses a visually rich format designed for the way your brain works, not a text-heavy approach that puts you to sleep.

US \$49.99

CAN \$62.99

ISBN: 978-0-596-15393-9



O'REILLY®

oreilly.com
headfirstlabs.com

Safari®
Books Online

Free online edition
for 45 days with
purchase of this book.
Details on last page.

“It’s about time a straightforward and comprehensive guide to analyzing data was written that makes learning the concepts simple and fun. Concepts are good in theory and even better in practicality.”

— Anthony Rose, President,
Support Analytics

“*Head First Data Analysis* shows how to find and unlock the power of data in everyday life and how systematic data analysis can improve decision making.”

— Eric Heilman,
Statistics teacher,
Georgetown
Preparatory School

“Buried under mountains of data? Fill your toolbox with the analytical skills that give you an edge and turn raw numbers into real knowledge.”

— Bill Mieltski,
Software engineer

Advance Praise for *Head First Data Analysis*

“It’s about time a straightforward and comprehensive guide to analyzing data was written that makes learning the concepts simple and fun. It will change the way you think and approach problems using proven techniques and free tools. Concepts are good in theory and even better in practicality.”

— **Anthony Rose, President, Support Analytics**

“*Head First Data Analysis* does a fantastic job of giving readers systematic methods to analyze real-world problems. From coffee, to rubber duckies, to asking for a raise, *Head First Data Analysis* shows the reader how to find and unlock the power of data in everyday life. Using everything from graphs and visual aides to computer programs like Excel and R, *Head First Data Analysis* gives readers at all levels accessible ways to understand how systematic data analysis can improve decision making both large and small.”

— **Eric Heilman, Statistics teacher, Georgetown Preparatory School**

“Buried under mountains of data? Let Michael Milton be your guide as you fill your toolbox with the analytical skills that give you an edge. In *Head First Data Analysis*, you’ll learn how to turn raw numbers into real knowledge. Put away your Ouija board and tarot cards; all you need to make good decisions is some software and a copy of this book.”

— **Bill Mietelski, Software engineer**

Praise for other *Head First* books

“Kathy and Bert’s *Head First Java* transforms the printed page into the closest thing to a GUI you’ve ever seen. In a wry, hip manner, the authors make learning Java an engaging ‘what’re they gonna do next?’ experience.”

—**Warren Keuffel, Software Development Magazine**

“Beyond the engaging style that drags you forward from know-nothing into exalted Java warrior status, *Head First Java* covers a huge amount of practical matters that other texts leave as the dreaded “exercise for the reader...” It’s clever, wry, hip and practical—there aren’t a lot of textbooks that can make that claim and live up to it while also teaching you about object serialization and network launch protocols.”

—**Dr. Dan Russell, Director of User Sciences and Experience Research
IBM Almaden Research Center (and teacher of Artificial Intelligence at
Stanford University)**

“It’s fast, irreverent, fun, and engaging. Be careful—you might actually learn something!”

—**Ken Arnold, former Senior Engineer at Sun Microsystems
Coauthor (with James Gosling, creator of Java), *The Java Programming
Language***

“I feel like a thousand pounds of books have just been lifted off of my head.”

—**Ward Cunningham, inventor of the Wiki and founder of the Hillside Group**

“Just the right tone for the geeked-out, casual-cool guru coder in all of us. The right reference for practical development strategies—gets my brain going without having to slog through a bunch of tired stale professor-speak.”

—**Travis Kalanick, Founder of Scour and Red Swoosh
Member of the MIT TR100**

“There are books you buy, books you keep, books you keep on your desk, and thanks to O’Reilly and the *Head First* crew, there is the ultimate category, *Head First* books. They’re the ones that are dog-eared, mangled, and carried everywhere. *Head First SQL* is at the top of my stack. Heck, even the PDF I have for review is tattered and torn.”

— **Bill Sawyer, ATG Curriculum Manager, Oracle**

“This book’s admirable clarity, humor and substantial doses of clever make it the sort of book that helps even non-programmers think well about problem-solving.”

— **Cory Doctorow, co-editor of BoingBoing
Author, *Down and Out in the Magic Kingdom*
and *Someone Comes to Town, Someone Leaves Town***

Praise for other *Head First* books

“I received the book yesterday and started to read it...and I couldn't stop. This is definitely très 'cool.' It is fun, but they cover a lot of ground and they are right to the point. I'm really impressed.”

— **Erich Gamma, IBM Distinguished Engineer, and co-author of *Design Patterns***

“One of the funniest and smartest books on software design I've ever read.”

— **Aaron LaBerge, VP Technology, ESPN.com**

“What used to be a long trial and error learning process has now been reduced neatly into an engaging paperback.”

— **Mike Davidson, CEO, Newsvine, Inc.**

“Elegant design is at the core of every chapter here, each concept conveyed with equal doses of pragmatism and wit.”

— **Ken Goldstein, Executive Vice President, Disney Online**

“I ♥ *Head First HTML with CSS & XHTML*—it teaches you everything you need to learn in a 'fun coated' format.”

— **Sally Applin, UI Designer and Artist**

“Usually when reading through a book or article on design patterns, I'd have to occasionally stick myself in the eye with something just to make sure I was paying attention. Not with this book. Odd as it may sound, this book makes learning about design patterns fun.

“While other books on design patterns are saying 'Buehler... Buehler... Buehler...' this book is on the float belting out 'Shake it up, baby!'”

— **Eric Wuehler**

“I literally love this book. In fact, I kissed this book in front of my wife.”

— **Satish Kumar**

Other related books from O'Reilly

Analyzing Business Data with Excel

Excel Scientific and Engineering Cookbook

Access Data Analysis Cookbook

Other books in O'Reilly's *Head First* series

Head First Java

Head First Object-Oriented Analysis and Design (OOA&D)

Head First HTML with CSS and XHTML

Head First Design Patterns

Head First Servlets and JSP

Head First EJB

Head First PMP

Head First SQL

Head First Software Development

Head First JavaScript

Head First Ajax

Head First Physics

Head First Statistics

Head First Rails

Head First PHP & MySQL

Head First Algebra

Head First Web Design

Head First Networking

Head First Data Analysis

Wouldn't it be dreamy if there was a book on data analysis that wasn't just a glorified printout of Microsoft Excel help files? But it's probably just a fantasy...



Michael Milton

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Taipei • Tokyo

Head First Data Analysis

by Michael Milton

Copyright © 2009 Michael Milton. All rights reserved.

Printed in the United States of America.

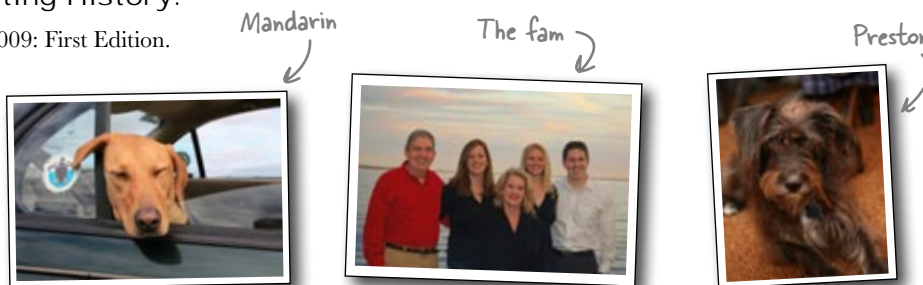
Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly Media books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*safari.oreilly.com*). For more information, contact our corporate/institutional sales department: (800) 998-9938 or *corporate@oreilly.com*.

Series Creators:	Kathy Sierra, Bert Bates
Series Editor:	Brett D. McLaughlin
Editor:	Brian Sawyer
Cover Designers:	Karen Montgomery
Production Editor:	Scott DeLugan
Proofreader:	Nancy Reinhardt
Indexer:	Jay Harward
Page Viewers:	Mandarin, the fam, and Preston

Printing History:

July 2009: First Edition.



The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. The *Head First* series designations, *Head First Data Analysis* and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and the authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

No data was harmed in the making of this book.



This book uses RepKover™, a durable and flexible lay-flat binding.

ISBN: 978-0-596-15393-9

[M]

Dedicated to the memory of my grandmother, Jane Reese Gibbs.

Author of Head First Data Analysis



Michael Milton

Michael Milton has spent most of his career helping nonprofit organizations improve their fundraising by interpreting and acting on the data they collect from their donors.

He has a degree in philosophy from New College of Florida and one in religious ethics from Yale University. He found reading *Head First* to be a revelation after spending years reading *boring* books filled with terribly important stuff and is grateful to have the opportunity to write an *exciting* book filled with terribly important stuff.

When he's not in the library or the bookstore, you can find him running, taking pictures, and brewing beer.

Table of Contents (Summary)

	Intro	xxvii
1	Introduction to Data Analysis: <i>Break It Down</i>	1
2	Experiments: <i>Test Your Theories</i>	37
3	Optimization: <i>Take It to the Max</i>	75
4	Data Visualization: <i>Pictures Make You Smarter</i>	111
5	Hypothesis Testing: <i>Say It Ain't So</i>	139
6	Bayesian Statistics: <i>Get Past First Base</i>	169
7	Subjective Probabilities: <i>Numerical Belief</i>	191
8	Heuristics: <i>Analyze Like a Human</i>	225
9	Histograms: <i>The Shape of Numbers</i>	251
10	Regression: <i>Prediction</i>	279
11	Error: <i>Err Well</i>	315
12	Relational Databases: <i>Can You Relate?</i>	359
13	Cleaning Data: <i>Impose Order</i>	385
i	Leftovers: <i>The Top Ten Things (We Didn't Cover)</i>	417
ii	Install R: <i>Start R Up!</i>	427
iii	Install Excel Analysis Tools: <i>The ToolPak</i>	431

Table of Contents (the real thing)

Intro

Your brain on data analysis. Here *you* are trying to *learn* something, while here your *brain* is doing you a favor by making sure the learning doesn't *stick*. Your brain's thinking, "Better leave room for more important things, like which wild animals to avoid and whether naked snowboarding is a bad idea." So how *do* you trick your brain into thinking that your life depends on knowing data analysis?

Who is this book for?	xxviii
We know what you're thinking	xxix
Metacognition	xxxi
Bend your brain into submission	xxxiii
Read Me	xxxiv
The technical review team	xxxvi
Acknowledgments	xxxvii

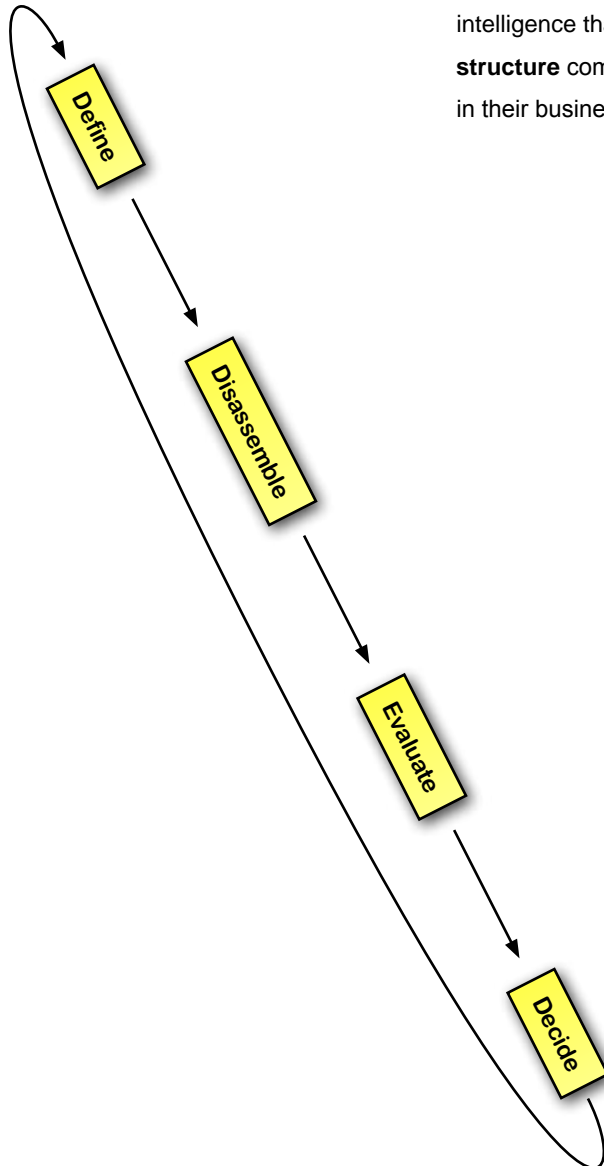
introduction to data analysis

Break it down

1

Data is everywhere.

Nowadays, everyone has to deal with mounds of data, whether they call themselves “data analysts” or not. But people who possess a toolbox of data analysis skills have a **massive edge** on everyone else, because they understand what to **do** with all that stuff. They know how to translate raw numbers into intelligence that **drives real-world action**. They know how to **break down and structure** complex problems and data sets to get right to the heart of the problems in their business.



Acme Cosmetics needs your help	2
The CEO wants data analysis to help increase sales	3
Data analysis is careful thinking about evidence	4
Define the problem	5
Your client will help you define your problem	6
Acme’s CEO has some feedback for you	8
Break the problem and data into smaller pieces	9
Now take another look at what you know	10
Evaluate the pieces	13
Analysis begins when you insert yourself	14
Make a recommendation	15
Your report is ready	16
The CEO likes your work	17
An article just came across the wire	18
You let the CEO’s beliefs take you down the wrong path	20
Your assumptions and beliefs about the world are your mental model	21
Your statistical model depends on your mental model	22
Mental models should always include what you don’t know	25
The CEO tells you what he doesn’t know	26
Acme just sent you a huge list of raw data	28
Time to drill further into the data	31
General American Wholesalers confirms your impression	32
Here’s what you did	35
Your analysis led your client to a brilliant decision	36

experiments

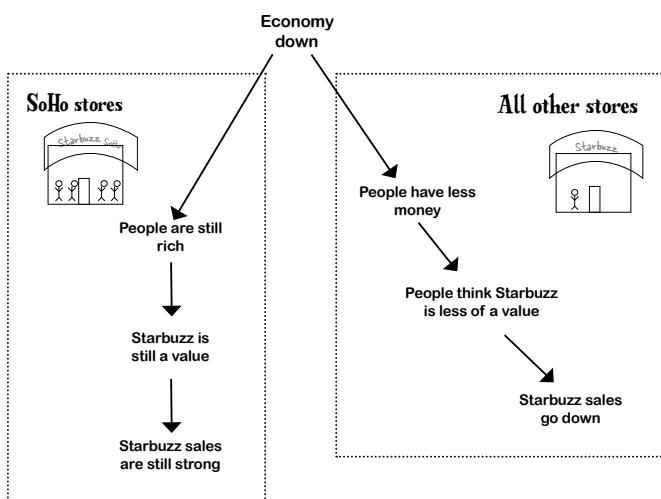
Test your theories

2

Can you show what you believe?

In a real **empirical** test? There's nothing like a good experiment to solve your problems and show you the way the world really works. Instead of having to rely exclusively on your **observational data**, a well-executed experiment can often help you make **causal connections**. Strong empirical data will make your analytical judgments all the more powerful.

It's a coffee recession!	38
The Starbuzz board meeting is in three months	39
The Starbuzz Survey	41
Always use the method of comparison	42
Comparisons are key for observational data	43
Could value perception be causing the revenue decline?	44
A typical customer's thinking	46
Observational studies are full of confounders	47
How location might be confounding your results	48
Manage confounders by breaking the data into chunks	50
It's worse than we thought!	53
You need an experiment to say which strategy will work best	54
The Starbuzz CEO is in a big hurry	55
Starbuzz drops its prices	56
One month later...	57
Control groups give you a baseline	58
Not getting fired 101	61
Let's experiment again for real!	62
One month later...	63
Confounders also plague experiments	64
Avoid confounders by selecting groups carefully	65
Randomization selects similar groups	67
Randomness Exposed	68
Your experiment is ready to go	71
The results are in	72
Starbuzz has an empirically tested sales strategy	73



optimization

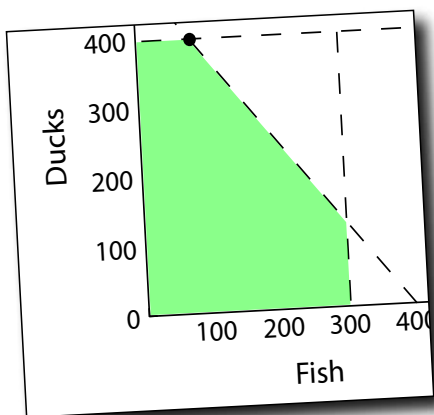
Take it to the max

3

We all want more of something.

And we're always trying to figure out how to get it. *If* the things we want more of—profit, money, efficiency, speed—can be represented numerically, then chances are, there's an tool of data analysis to help us tweak our *decision variables*, which will help us find the **solution** or *optimal point* where we get the most of what we want. In this chapter, you'll be using one of those tools and the powerful spreadsheet **Solver** package that implements it.

You're now in the bath toy game	76
Constraints limit the variables you control	79
Decision variables are things you can control	79
You have an optimization problem	80
Find your objective with the objective function	81
Your objective function	82
Show product mixes with your other constraints	83
Plot multiple constraints on the same chart	84
Your good options are all in the feasible region	85
Your new constraint changed the feasible region	87
Your spreadsheet does optimization	90
Solver crunched your optimization problem in a snap	94
Profits fell through the floor	97
Your model only describes what you put into it	98
Calibrate your assumptions to your analytical objectives	99
Watch out for negatively linked variables	103
Your new plan is working like a charm	108
Your assumptions are based on an ever-changing reality	109



data visualization

Pictures make you smarter

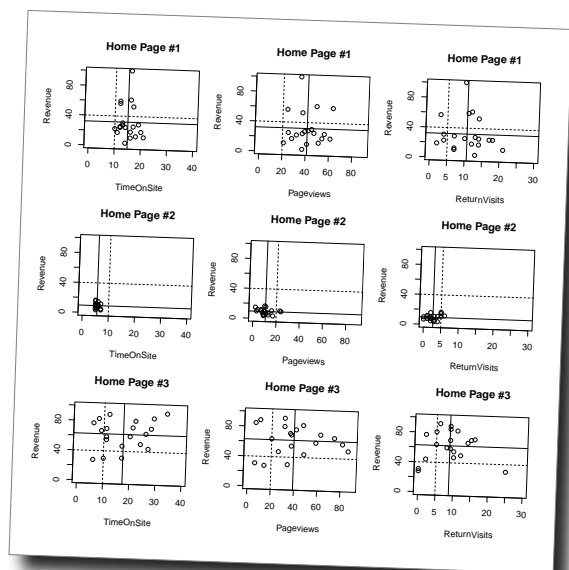
4

You need more than a table of numbers.

Your data is brilliantly complex, with more variables than you can shake a stick at.

Mulling over mounds and mounds of spreadsheets isn't just boring; it can actually be a waste of your time. A clear, highly multivariate visualization can, in a small space, show you the forest that you'd miss for the trees if you were just looking at spreadsheets all the time.

New Army needs to optimize their website	112
The results are in, but the information designer is out	113
The last information designer submitted these three infographics	114
What data is behind the visualizations?	115
Show the data!	116
Here's some unsolicited advice from the last designer	117
Too much data is never your problem	118
Making the data pretty isn't your problem either	119
Data visualization is all about making the right comparisons	120
Your visualization is already more useful than the rejected ones	123
Use scatterplots to explore causes	124
The best visualizations are highly multivariate	125
Show more variables by looking at charts together	126
The visualization is great, but the web guru's not satisfied yet	130
Good visual designs help you think about causes	131
The experiment designers weigh in	132
The experiment designers have some hypotheses of their own	135
The client is pleased with your work	136
Orders are coming in from everywhere!	137



hypothesis testing

Say it ain't so

5

The world can be tricky to explain.

And it can be fiendishly difficult when you have to deal with complex, heterogeneous data to anticipate future events. This is why analysts don't just take the obvious explanations and assume them to be true: the careful reasoning of data analysis enables you to meticulously evaluate a bunch of options so that you can incorporate all the information you have into your models. You're about to learn about **falsification**, an unintuitive but powerful way to do just that.

Gimme some skin...	140
When do we start making new phone skins?	141
PodPhone doesn't want you to predict their next move	142
Here's everything we know	143
ElectroSkinny's analysis does fit the data	144
ElectroSkinny obtained this confidential strategy memo	145
Variables can be negatively or positively linked	146
Causes in the real world are networked, not linear	149
Hypothesize PodPhone's options	150
You have what you need to run a hypothesis test	151
Falsification is the heart of hypothesis testing	152
Diagnosticity helps you find the hypothesis with the least disconfirmation	160
You can't rule out all the hypotheses, but you can say which is strongest	163
You just got a picture message...	164
It's a launch!	167



bayesian statistics

Get past first base

6

You'll always be collecting new data.

And you need to make sure that every analysis you do incorporates the data you have that's relevant to your problem. You've learned how *falsification* can be used to deal with heterogeneous data sources, but what about **straight up probabilities**? The answer involves an extremely handy analytic tool called **Bayes' rule**, which will help you incorporate your **base rates** to uncover not-so-obvious insights with ever-changing data.

The doctor has disturbing news	170
Let's take the accuracy analysis one claim at a time	173
How common is lizard flu really?	174
You've been counting false positives	175
All these terms describe conditional probabilities	176
You need to count false positives, true positives, false negatives, and true negatives	177
1 percent of people have lizard flu	178
Your chances of having lizard flu are still pretty low	181
Do complex probabilistic thinking with simple whole numbers	182
Bayes' rule manages your base rates when you get new data	182
You can use Bayes' rule over and over	183
Your second test result is negative	184
The new test has different accuracy statistics	185
New information can change your base rate	186
What a relief!	189

Cough



7

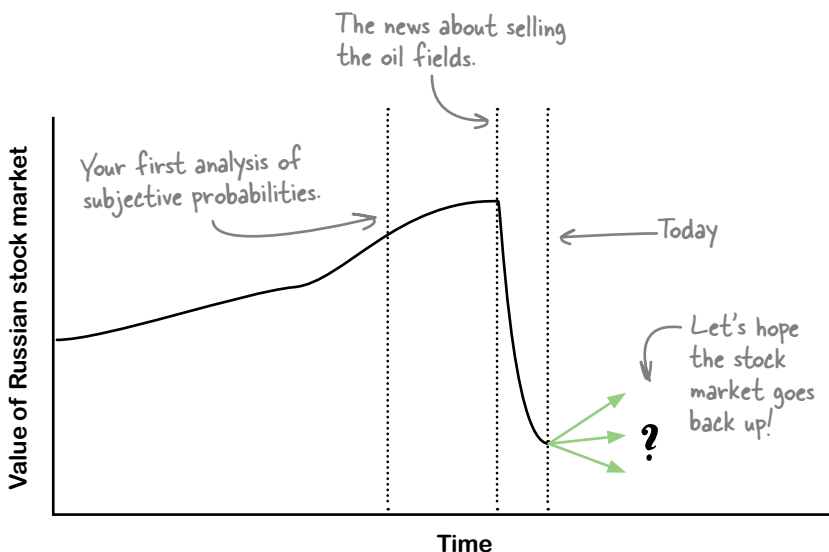
subjective probabilities

Numerical belief

Sometimes, it's a good idea to make up numbers.

Seriously. But only if those numbers describe your own mental states, expressing your beliefs. **Subjective probability** is a straightforward way of injecting some real *rigor* into your hunches, and you're about to see how. Along the way, you are going to learn how to evaluate the spread of data using **standard deviation** and enjoy a special guest appearance from one of the more powerful analytic tools you've learned.

Backwater Investments needs your help	192
Their analysts are at each other's throats	193
Subjective probabilities describe expert beliefs	198
Subjective probabilities might show no real disagreement after all	199
The analysts responded with their subjective probabilities	201
The CEO doesn't see what you're up to	202
The CEO loves your work	207
The standard deviation measures how far points are from the average	208
You were totally blindsided by this news	213
Bayes' rule is great for revising subjective probabilities	217
The CEO knows exactly what to do with this new information	223
Russian stock owners rejoice!	224



heuristics

8

Analyze like a human

The real world has more variables than you can handle.

There is always going to be data that you can't have. And even when you do have data on most of the things you want to understand, *optimizing* methods are often **elusive** and **time consuming**. Fortunately, most of the actual thinking you do in life is not “rational maximizing”—it's processing incomplete and uncertain information with rules of thumb so that you can make decisions quickly. What is really cool is that these rules can **actually work** and are important (and necessary) tools for data analysts.

LitterGitters submitted their report to the city council	226
The LitterGitters have really cleaned up this town	227
The LitterGitters have been measuring their campaign's effectiveness	228
The mandate is to reduce the tonnage of litter	229
Tonnage is unfeasible to measure	230
Give people a hard question, and they'll answer an easier one instead	231
Littering in Dataville is a complex system	232
You can't build and implement a unified litter-measuring model	233
Heuristics are a middle ground between going with your gut and optimization	236
Use a fast and frugal tree	239
Is there a simpler way to assess LitterGitters' success?	240
Stereotypes are heuristics	244
Your analysis is ready to present	246
Looks like your analysis impressed the city council members	249



histograms

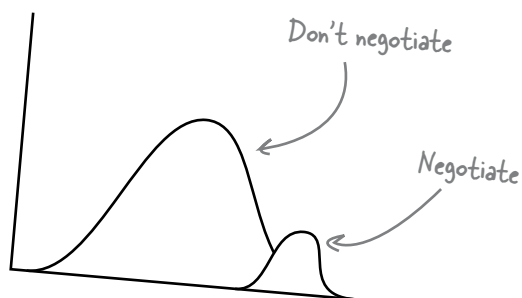
The shape of numbers

9

How much can a bar graph tell you?

There are about a zillion ways of **showing data with pictures**, but one of them is special. **Histograms**, which are kind of similar to bar graphs, are a super-fast and easy way to summarize data. You're about to use these powerful little charts to measure your data's **spread, variability, central tendency**, and more. No matter how large your data set is, if you draw a histogram with it, you'll be able to "see" what's happening inside of it. And you're about to do it with a new, free, crazy-powerful **software tool**.

Your annual review is coming up	252
Going for more cash could play out in a bunch of different ways	254
Here's some data on raises	255
Histograms show frequencies of groups of numbers	262
Gaps between bars in a histogram mean gaps among the data points	263
Install and run R	264
Load data into R	265
R creates beautiful histograms	266
Make histograms from subsets of your data	271
Negotiation pays	276
What will negotiation mean for you?	277



10

regression
Prediction**Predict it.**

Regression is an incredibly powerful statistical tool that, when used correctly, has the ability to help you predict certain values. When used with a controlled experiment, regression can actually help you predict the future. Businesses use it like crazy to help them build models to explain customer behavior. You're about to see that the judicious use of regression can be very profitable indeed.

What are you going to do with all this money?	280
An analysis that tells people what to ask for could be huge	283
Behold... the Raise Reckoner!	284
Inside the algorithm will be a method to predict raises	286
Scatterplots compare two variables	292
A line could tell your clients where to aim	294
Predict values in each strip with the graph of averages	297
The regression line predicts what raises people will receive	298
The line is useful if your data shows a linear correlation	300
You need an equation to make your predictions precise	304
Tell R to create a regression object	306
The regression equation goes hand in hand with your scatterplot	309
The regression equation is the Raise Reckoner algorithm	310
Your raise predictor didn't work out as planned...	313



THE RAISE RECKONER

What will happen if we request a certain amount of money? Find out with this equation:

$$y = 2.3 + 0.7x$$

Where x is the amount requested, and y is the amount we can expect to receive.



11

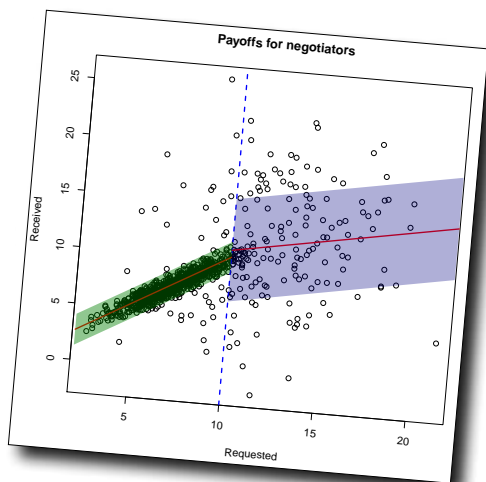
error

Err well

The world is messy.

So it should be no surprise that your predictions rarely hit the target squarely. But if you offer a prediction with an **error range**, you and your clients will know not only the average predicted value, but also how far you expect typical deviations from that error to be. Every time you express error, you offer a much richer perspective on your predictions and beliefs. And with the tools in this chapter, you'll also learn about how to get error under control, getting it as low as possible to increase confidence.

Your clients are pretty ticked off	316
What did your raise prediction algorithm do?	317
The segments of customers	318
The guy who asked for 25% went outside the model	321
How to handle the client who wants a prediction outside the data range	322
The guy who got fired because of extrapolation has cooled off	327
You've only solved part of the problem	328
What does the data for the screwy outcomes look like?	329
Chance errors are deviations from what your model predicts	330
Error is good for you and your client	334
Chance Error Exposed	335
Specify error quantitatively	336
Quantify your residual distribution with Root Mean Squared error	337
Your model in R already knows the R.M.S. error	338
R's summary of your linear model shows your R.M.S. error	340
Segmentation is all about managing error	346
Good regressions balance explanation and prediction	350
Your segmented models manage error better than the original model	352
Your clients are returning in droves	357



12

relational databases

Can you relate?

How do you structure really, really multivariate data?

A spreadsheet has only *two dimensions*: rows and columns. And if you have a bunch of dimensions of data, the **tabular format** gets old really quickly. In this chapter, you're about to see firsthand where spreadsheets make it really hard to manage multivariate data and learn **how relational database management systems** make it easy to store and retrieve countless permutations of multivariate data.

The Dataville Dispatch wants to analyze sales	360
Here's the data they keep to track their operations	361
You need to know how the data tables relate to each other	362
A database is a collection of data with well-specified relations to each other	365
Trace a path through the relations to make the comparison you need	366
Create a spreadsheet that goes across that path	366
Your summary ties article count and sales together	371
Looks like your scatterplot is going over really well	374
Copying and pasting all that data was a pain	375
Relational databases manage relations for you	376
Dataville Dispatch built an RDBMS with your relationship diagram	377
Dataville Dispatch extracted your data using the SQL language	379
Comparison possibilities are endless if your data is in a RDBMS	382
You're on the cover	383



13

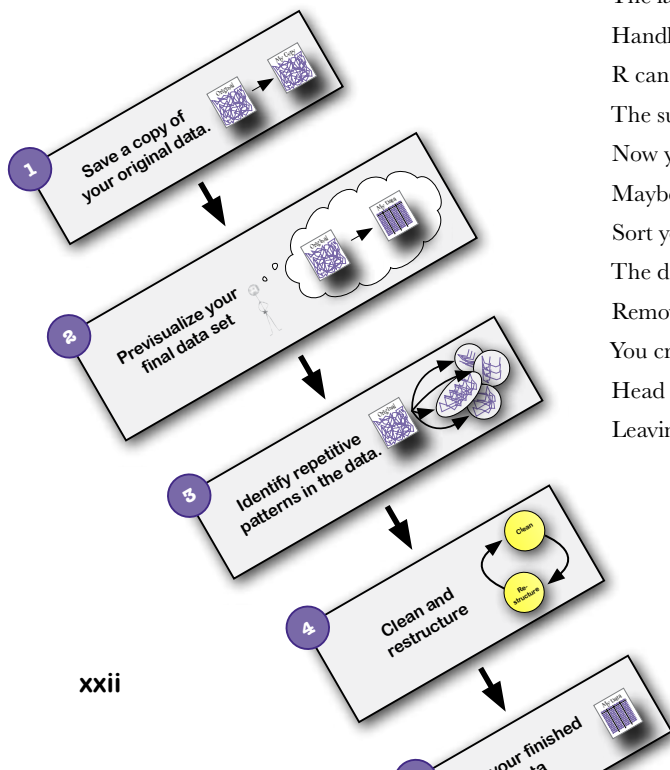
cleaning data

Impose order

Your data is useless...

...if it has messy structure. And a lot of people who **collect** data do a crummy job of maintaining a neat structure. If your data's not neat, you can't slice it or dice it, run formulas on it, or even really **see** it. You might as well just ignore it completely, right? Actually, you can do better. With a **clear vision** of how you need it to look and a few **text manipulation tools**, you can take the funkiest, craziest mess of data and **whip** it into something useful.

Just got a client list from a defunct competitor	386
The dirty secret of data analysis	387
Head First Head Hunters wants the list for their sales team	388
Cleaning messy data is all about preparation	392
Once you're organized, you can fix the data itself	393
Use the # sign as a delimiter	394
Excel split your data into columns using the delimiter	395
Use SUBSTITUTE to replace the carat character	399
You cleaned up all the first names	400
The last name pattern is too complex for SUBSTITUTE	402
Handle complex patterns with nested text formulas	403
R can use regular expressions to crunch complex data patterns	404
The sub command fixed your last names	406
Now you can ship the data to your client	407
Maybe you're not quite done yet...	408
Sort your data to show duplicate values together	409
The data is probably from a relational database	412
Remove duplicate names	413
You created nice, clean, unique records	414
Head First Head Hunters is recruiting like gangbusters!	415
Leaving town...	416



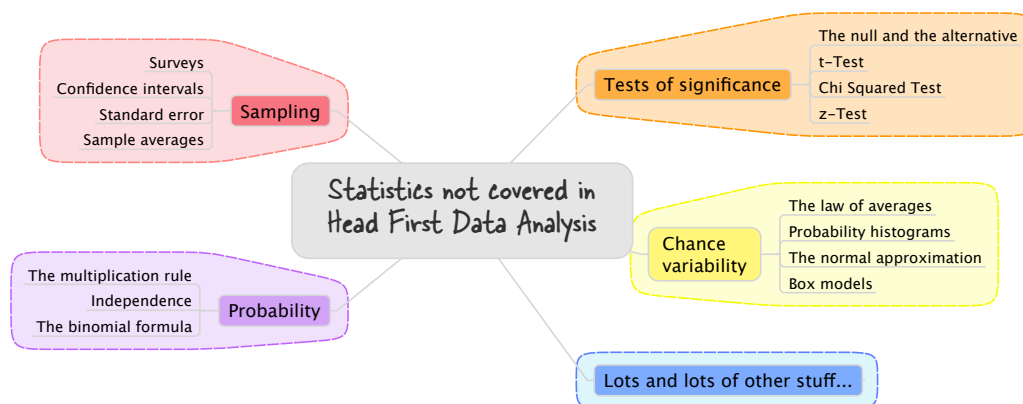
leftovers

The Top Ten Things (we didn't cover)

**You've come a long way.**

But data analysis is a vast and constantly evolving field, and there's so much left to learn. In this appendix, we'll go over ten items that there wasn't enough room to cover in this book but should be high on your list of topics to learn about next.

#1: Everything else in statistics	418
#2: Excel skills	419
#3: Edward Tufte and his principles of visualization	420
#4: PivotTables	421
#5: The R community	422
#6: Nonlinear and multiple regression	423
#7: Null-alternative hypothesis testing	424
#8: Randomness	424
#9: Google Docs	425
#10: Your expertise	426



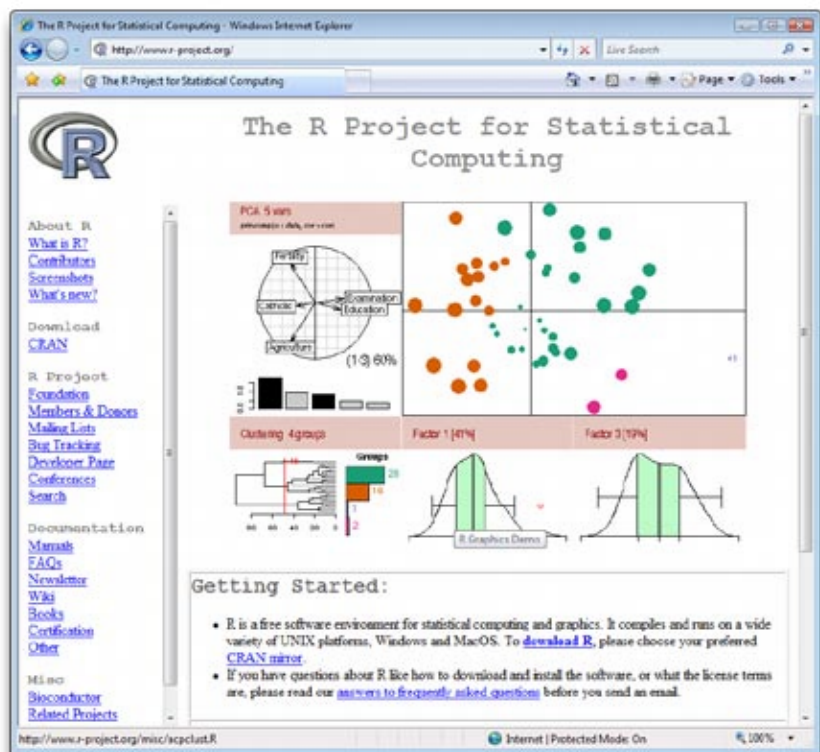


install r

Start R up!

Behind all that data-crunching power is enormous complexity.

But fortunately, getting R installed and **started** is something you can accomplish in just a few minutes, and this appendix is about to show you how to pull off your R install without a hitch.



install excel analysis tools

The ToolPak



Some of the best features of Excel aren't installed by default.

That's right, in order to run the optimization from Chapter 3 and the histograms from Chapter 9, you need to activate the **Solver** and the **Analysis ToolPak**, two extensions that are included in Excel by default but not activated without your initiative.

Install the data analysis tools in Excel

432

