Modeling Techniques in Predictive Analytics with Python and R

A Guide to Data Science

THOMAS W. MILLER

Associate Publisher: Amy Neidlinger
Executive Editor: Jeanne Glasser
Operations Specialist: Jodi Kemper
Cover Designer: Alan Clements
Managing Editor: Kristy Hart
Project Editor: Andy Beaster
Senior Compositor: Gloria Schurick
Manufacturing Buyer: Dan Uhrig

Published by Pearson Education, Inc.
Upper Saddle River, New Jersey 07458

Printed in the United States of America

First Printing October 2014

ISBN-10: 0-13-3892069
ISBN-13: 978-0-13-389206-2

Pearson Education LTD.
Pearson Education Australia PTY, Limited.
Pearson Education Singapore, Pte. Ltd.
Pearson Education Asia, Ltd.
Pearson Education Canada, Ltd.
Pearson Educacin de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.
Library of Congress Control Number: 2014948913

Contents

Preface v

Figures xi

Tables xv

Exhibits xvii

1 Analytics and Data Science 1

2 Advertising and Promotion 16

3 Preference and Choice 33

4 Market Basket Analysis 43

5 Economic Data Analysis 61

6 Operations Management 81

7 Text Analytics 103

8 Sentiment Analysis 135

9 Sports Analytics 187

10 Spatial Data Analysis 211

11 Brand and Price 239

12 The Big Little Data Game 273

A Data Science Methods 277
- A.1 Databases and Data Preparation 279
- A.2 Classical and Bayesian Statistics 281
- A.3 Regression and Classification 284
- A.4 Machine Learning 289
- A.5 Web and Social Network Analysis 291
- A.6 Recommender Systems 293
- A.7 Product Positioning 295
- A.8 Market Segmentation 297
- A.9 Site Selection 299
- A.10 Financial Data Science 300

B Measurement 301

C Case Studies 315
- C.1 Return of the Bobbleheads 315
- C.2 DriveTime Sedans 316
- C.3 Two Month's Salary 321
- C.4 Wisconsin Dells 325
- C.5 Computer Choice Study 330

D Code and Utilities 335

Bibliography 379

Index 413

1 Analytics and Data Science

Mr. Maguire: "I just want to say one word to you, just one word."

Ben: "Yes, sir."

Mr. Maguire: "Are you listening?"

Ben: "Yes, I am."

Mr. Maguire: "Plastics."

—WALTER BROOKE AS MR. MAGUIRE AND DUSTIN HOFFMAN AS BEN (BENJAMIN BRADDOCK) IN *The Graduate* (1967)

While earning a degree in philosophy may not be the best career move (unless a student plans to teach philosophy, and few of these positions are available), I greatly value my years as a student of philosophy and the liberal arts. For my bachelor's degree, I wrote an honors paper on Bertrand Russell. In graduate school at the University of Minnesota, I took courses from one of the truly great philosophers, Herbert Feigl. I read about science and the search for truth, otherwise known as epistemology. My favorite philosophy was logical empiricism.

Although my days of "thinking about thinking" (which is how Feigl defined philosophy) are far behind me, in those early years of academic training I was able to develop a keen sense for what is real and what is just talk.

A *model* is a representation of things, a rendering or description of reality. A typical model in data science is an attempt to relate one set of variables to another. Limited, imprecise, but useful, a model helps us to make sense of the world. A model is more than just talk because it is based on data.

Predictive analytics brings together management, information technology, and modeling. It is designed for today's data-intensive world. Predictive analytics is data science, a multidisciplinary skill set essential for success in business, nonprofit organizations, and government. Whether forecasting sales or market share, finding a good retail site or investment opportunity, identifying consumer segments and target markets, or assessing the potential of new products or risks associated with existing products, modeling methods in predictive analytics provide the key.

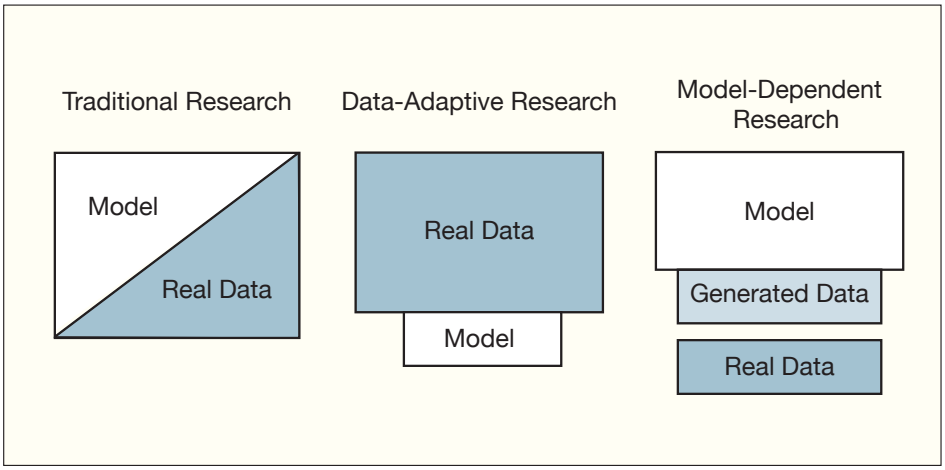
Data scientists, those working in the field of predictive analytics, speak the language of business—accounting, finance, marketing, and management. They know about information technology, including data structures, algorithms, and object-oriented programming. They understand statistical modeling, machine learning, and mathematical programming. Data scientists are methodological eclectics, drawing from many scientific disciplines and translating the results of empirical research into words and pictures that management can understand.

Predictive analytics, as with much of statistics, involves searching for meaningful relationships among variables and representing those relationships in models. There are response variables—things we are trying to predict. There are explanatory variables or predictors—things that we observe, manipulate, or control and might relate to the response.

Regression methods help us to predict a response with meaningful magnitude, such as quantity sold, stock price, or return on investment. Classification methods help us to predict a categorical response. Which brand will be purchased? Will the consumer buy the product or not? Will the account holder pay off or default on the loan? Is this bank transaction true or fraudulent?

Prediction problems are defined by their width or number of potential predictors and by their depth or number of observations in the data set. It is the number of potential predictors in business, marketing, and investment analysis that causes the most difficulty. There can be thousands of potential

Figure 1.1. *Data and models for research*



predictors with weak relationships to the response. With the aid of computers, hundreds or thousands of models can be fit to subsets of the data and tested on other subsets of the data, providing an evaluation of each predictor. Predictive modeling involves finding good subsets of predictors. Models that fit the data well are better than models that fit the data poorly. Simple models are better than complex models.

Consider three general approaches to research and modeling as employed in predictive analytics: traditional, data-adaptive, and model-dependent. See figure 1.1. The traditional approach to research, statistical inference, and modeling begins with the specification of a theory or model. Classical or Bayesian methods of statistical inference are employed. Traditional methods, such as linear regression and logistic regression, estimate parameters for linear predictors. Model building involves fitting models to data and checking them with diagnostics. We validate traditional models before using them to make predictions.

When we employ a data-adaptive approach, we begin with data and search through those data to find useful predictors. We give little thought to theories or hypotheses prior to running the analysis. This is the world of machine learning, sometimes called statistical learning or data mining. Data-adaptive methods adapt to the available data, representing nonlinear relationships and interactions among variables. The data determine the model.

Data-adaptive methods are data-driven. As with traditional models, we validate data-adaptive models before using them to make predictions.

Model-dependent research is the third approach. It begins with the specification of a model and uses that model to generate data, predictions, or recommendations. Simulations and mathematical programming methods, primary tools of operations research, are examples of model-dependent research. When employing a model-dependent or simulation approach, models are improved by comparing generated data with real data. We ask whether simulated consumers, firms, and markets behave like real consumers, firms, and markets. The comparison with real data serves as a form of validation.

It is often a combination of models and methods that works best. Consider an application from the field of financial research. The manager of a mutual fund is looking for additional stocks for a fund's portfolio. A financial engineer employs a data-adaptive model (perhaps a neural network) to search across thousands of performance indicators and stocks, identifying a subset of stocks for further analysis. Then, working with that subset of stocks, the financial engineer employs a theory-based approach (CAPM, the capital asset pricing model) to identify a smaller set of stocks to recommend to the fund manager. As a final step, using model-dependent research (mathematical programming), the engineer identifies the minimum-risk capital investment for each of the stocks in the portfolio.

Data may be organized by observational unit, time, and space. The observational or cross-sectional unit could be an individual consumer or business or any other basis for collecting and grouping data. Data are organized in time by seconds, minutes, hours, days, and so on. Space or location is often defined by longitude and latitude.

Consider numbers of customers entering grocery stores (units of analysis) in Glendale, California on Monday (one point in time), ignoring the spatial location of the stores—these are cross-sectional data. Suppose we work with one of those stores, looking at numbers of customers entering the store each day of the week for six months—these are time series data. Then we look at numbers of customers at all of the grocery stores in Glendale across six months—these are longitudinal or panel data. To complete our study, we locate these stores by longitude and latitude, so we have spatial

or spatio-temporal data. For any of these data structures we could consider measures in addition to the number of customers entering stores. We look at store sales, consumer or nearby resident demographics, traffic on Glendale streets, and so doing move to multiple time series and multivariate methods. The organization of the data we collect affects the structure of the models we employ.

As we consider business problems in this book, we touch on many types of models, including cross-sectional, time series, and spatial data models. Whatever the structure of the data and associated models, prediction is the unifying theme. We use the data we have to predict data we do not yet have, recognizing that prediction is a precarious enterprise. It is the process of extrapolating and forecasting. And model validation is essential to the process.

To make predictions, we may employ classical or Bayesian methods. Or we may dispense with traditional statistics entirely and rely upon machine learning algorithms. We do what works.[1] Our approach to predictive analytics is based upon a simple premise:

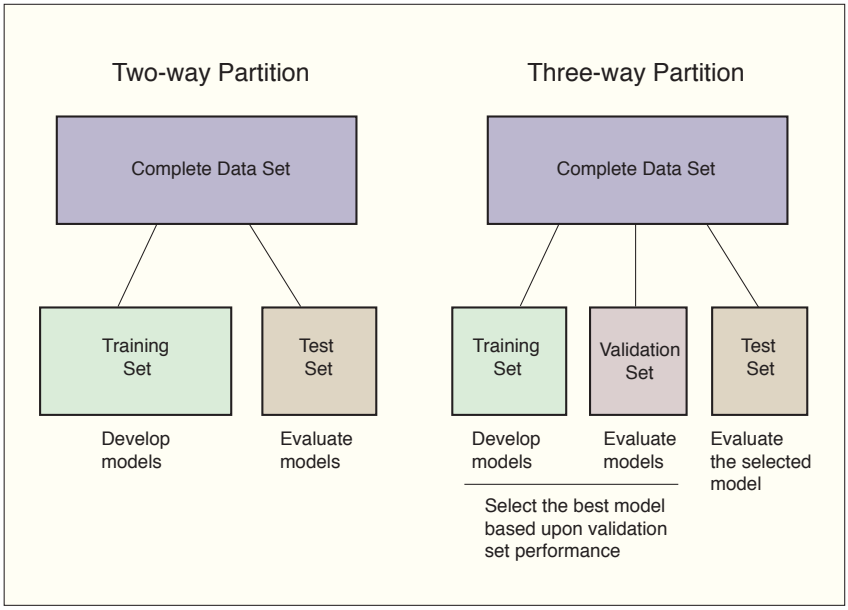
The value of a model lies in the quality of its predictions.

We learn from statistics that we should quantify our uncertainty. On the one hand, we have confidence intervals, point estimates with associated standard errors, significance tests, and *p*-values—that is the classical way. On the other hand, we have posterior probability distributions, probability intervals, prediction intervals, Bayes factors, and subjective (perhaps diffuse) priors—the path of Bayesian statistics. Indices such as the Akaike information criterion (AIC) or the Bayes information criterion (BIC) help us to to judge one model against another, providing a balance between goodness-of-fit and parsimony.

Central to our approach is a *training-and-test regimen*. We partition sample data into training and test sets. We build our model on the training set and

[1] Within the statistical literature, Seymour Geisser (1929–2004) introduced an approach best described as *Bayesian predictive inference* (Geisser 1993). Bayesian statistics is named after Reverend Thomas Bayes (1706–1761), the creator of Bayes Theorem. In our emphasis upon the success of predictions, we are in agreement with Geisser. Our approach, however, is purely empirical and in no way dependent upon classical or Bayesian thinking.

Figure 1.2. *Training-and-Test Regimen for Model Evaluation*



evaluate it on the test set. Simple two- and three-way data partitioning are shown in figure 1.2.

A random splitting of a sample into training and test sets could be fortuitous, especially when working with small data sets, so we sometimes conduct statistical experiments by executing a number of random splits and averaging performance indices from the resulting test sets. There are extensions to and variations on the training-and-test theme.

One variation on the training-and-test theme is multi-fold cross-validation, illustrated in figure 1.3. We partition the sample data into M folds of approximately equal size and conduct a series of tests. For the five-fold cross-validation shown in the figure, we would first train on sets B through E and test on set A. Then we would train on sets A and C through E, and test on B. We continue until each of the five folds has been utilized as a test set. We assess performance by averaging across the test sets. In leave-one-out cross-valuation, the logical extreme of multi-fold cross-validation, there are as many test sets as there are observations in the sample.

Figure 1.3. *Training-and-Test Using Multi-fold Cross-validation*

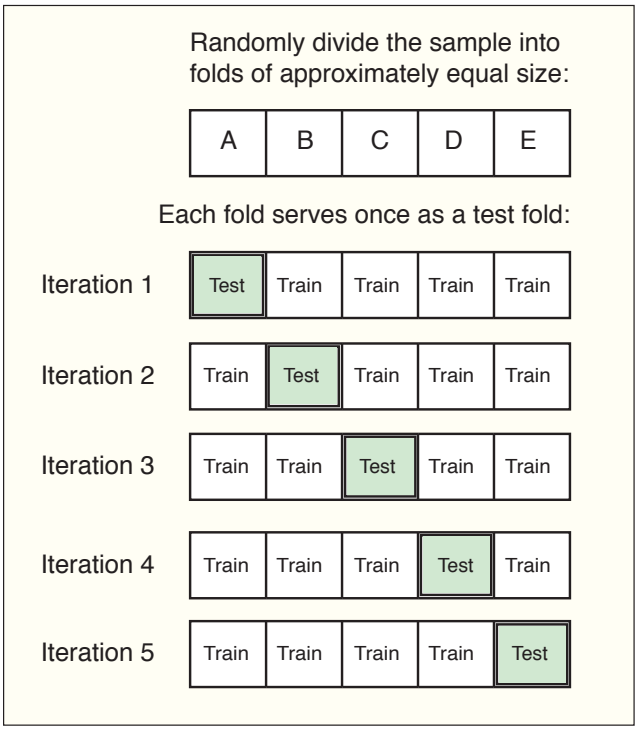
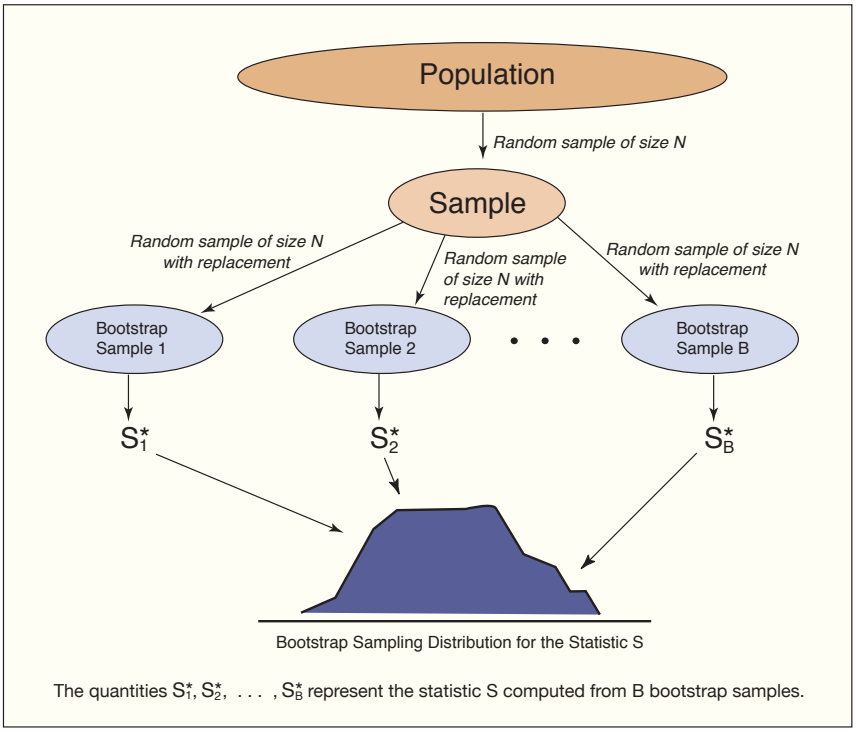


Figure 1.4. *Training-and-Test with Bootstrap Resampling*



Another variation on the training-and-test regimen is the class of bootstrap methods. If a sample approximates the population from which it was drawn, then a sample from the sample (what is known as a resample) also approximates the population. A bootstrap procedure, as illustrated in figure 1.4, involves repeated resampling with replacement. That is, we take many random samples with replacement from the sample, and for each of these resamples, we compute a statistic of interest. The bootstrap distribution of the statistic approximates the sampling distribution of that statistic. What is the value of the bootstrap? It frees us from having to make assumptions about the population distribution. We can estimate standard errors and make probability statements working from the sample data alone. The bootstrap may also be employed to improve estimates of prediction error within a leave-one-out cross-validation process. Cross-validation and bootstrap methods are reviewed in Davison and Hinkley (1997), Efron and Tibshirani (1993), and Hastie, Tibshirani, and Friedman (2009).

Table 1.1. *Data for the Anscombe Quartet*

Set I		*Set II*		*Set III*		*Set IV*	
x1	*y1*	*x2*	*y2*	*x3*	*y3*	*x4*	*y4*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

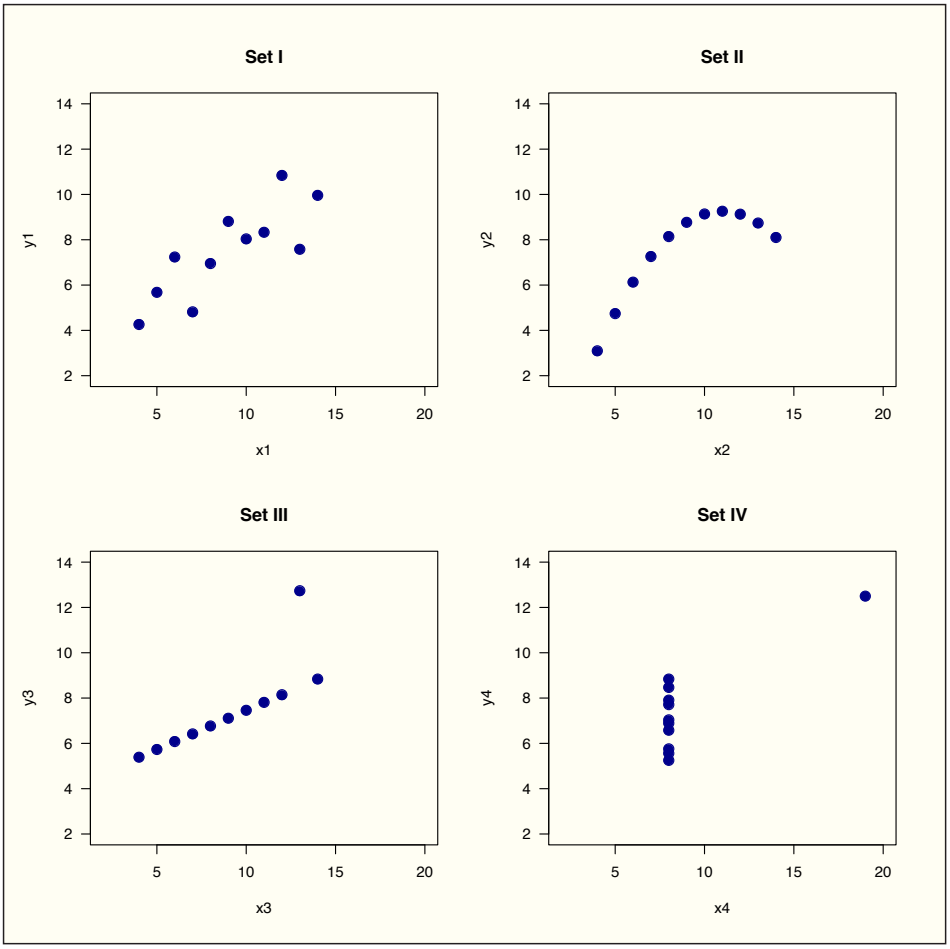
Data visualization is critical to the work of data science. Examples in this book demonstrate the importance of data visualization in discovery, diagnostics, and design. We employ tools of exploratory data analysis (discovery) and statistical modeling (diagnostics). In communicating results to management, we use presentation graphics (design).

There is no more telling demonstration of the importance of statistical graphics and data visualization than a demonstration that is affectionately known as the Anscombe Quartet. Consider the data sets in table 1.1, developed by Anscombe (1973). Looking at these tabulated data, the casual reader will note that the fourth data set is clearly different from the others. What about the first three data sets? Are there obvious differences in patterns of relationship between x and y?

When we regress y on x for the data sets, we see that the models provide similar statistical summaries. The mean of the response y is 7.5, the mean of the explanatory variable x is 9. The regression analyses for the four data sets are virtually identical. The fitted regression equation for each of the four sets is $\hat{y} = 3 + 0.5x$. The proportion of response variance accounted for is 0.67 for each of the four models.

Following Anscombe (1973), we would argue that statistical summaries fail to tell the story of data. We must look beyond data tables, regression coefficients, and the results of statistical tests. It is the plots in figure 1.5 that tell the story. The four Anscombe data sets are very different from one another.

Figure 1.5. *Importance of Data Visualization: The Anscombe Quartet*



The Anscombe Quartet shows that we must look at data to understand them. Python and R programs for the Anscombe Quartet are provided at the end of this chapter in exhibits 1.1 and 1.2, respectively.

Visualization tools help us learn from data. We explore data, discover patterns in data, identify groups of observations that go together and unusual observations or outliers. We note relationships among variables, sometimes detecting underlying dimensions in the data.

Graphics for exploratory data analysis are reviewed in classic references by Tukey (1977) and Tukey and Mosteller (1977). Regression graphics are covered by Cook (1998), Cook and Weisberg (1999), and Fox and Weisberg (2011). Statistical graphics and data visualization are illustrated in the works of Tufte (1990, 1997, 2004, 2006), Few (2009), and Yau (2011, 2013). Wilkinson (2005) presents a review of human perception and graphics, as well as a conceptual structure for understanding statistical graphics. Cairo (2013) provides a general review of information graphics. Heer, Bostock, and Ogievetsky (2010) demonstrate contemporary visualization techniques for web distribution. When working with very large data sets, special methods may be needed, such as partial transparency and hexbin plots (Unwin, Theus, and Hofmann 2006; Carr, Lewin-Koh, and Maechler 2014; Lewin-Koh 2014).

Python and R represent rich programming environments for data visualization, including interfaces to visualization applications on the World Wide Web. Chun (2007), Beazley (2009), and Beazley and Jones (2013) review the Python programming environment. Matloff (2011) and Lander (2014) provide useful introductions to R. An R graphics overview is provided by Murrell (2011). R lattice graphics, discussed by Sarkar (2008, 2014), build upon the conceptual structure of an earlier system called S-Plus Trellis™ (Cleveland 1993; Becker and Cleveland 1996). Wilkinson's (2005) "grammar of graphics" approach has been implemented in the Python ggplot package (Lamp 2014) and in the R ggplot2 package (Wickham and Chang 2014), with R programming examples provided by Chang (2013). Cairo (2013) and Zeileis, Hornik, and Murrell (2009, 2014) provide advice about colors for statistical graphics. Ihaka et al. (2014) show how to specify colors in R by hue, chroma, and luminance.

These are the things that data scientists do:

- **Finding out about.** This is the first thing we do—information search, finding what others have done before, learning from the literature. We draw on the work of academics and practitioners in many fields of study, contributors to predictive analytics and data science.
- **Preparing text and data.** Text is unstructured or partially structured. Data are often messy or missing. We extract features from text. We define measures. We prepare text and data for analysis and modeling.
- **Looking at data.** We do exploratory data analysis, data visualization for the purpose of discovery. We look for groups in data. We find outliers. We identify common dimensions, patterns, and trends.
- **Predicting how much.** We are often asked to predict how many units or dollars of product will be sold, the price of financial securities or real estate. Regression techniques are useful for making these predictions.
- **Predicting yes or no.** Many business problems are classification problems. We use classification methods to predict whether or not a person will buy a product, default on a loan, or access a web page.
- **Testing it out.** We examine models with diagnostic graphics. We see how well a model developed on one data set works on other data sets. We employ a training-and-test regimen with data partitioning, cross-validation, or bootstrap methods.
- **Playing what-if.** We manipulate key variables to see what happens to our predictions. We play what-if games in simulated marketplaces. We employ sensitivity or stress testing of mathematical programming models. We see how values of input variables affect outcomes, payoffs, and predictions. We assess uncertainty about forecasts.
- **Explaining it all.** Data and models help us understand the world. We turn what we have learned into an explanation that others can understand. We present project results in a clear and concise manner. These presentations benefit from well-constructed data visualizations.

Let us begin.

Exhibit 1.1. *Programming the Anscombe Quartet (Python)*

```
# The Anscombe Quartet (Python)
# demonstration data from
# Anscombe, F. J. 1973, February. Graphs in statistical analysis.
#  The American Statistician 27: 1721.

# prepare for Python version 3x features and functions
from __future__ import division, print_function

# import packages for Anscombe Quartet demonstration
import pandas as pd  # data frame operations
import numpy as np  # arrays and math functions
import statsmodels.api as sm  # statistical models (including regression)
import matplotlib.pyplot as plt  # 2D plotting

# define the anscombe data frame using dictionary of equal-length lists
anscombe = pd.DataFrame({'x1' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x2' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x3' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x4' : [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8],
    'y1' : [8.04, 6.95,  7.58, 8.81, 8.33, 9.96, 7.24, 4.26,10.84, 4.82, 5.68],
    'y2' : [9.14, 8.14,  8.74, 8.77, 9.26, 8.1, 6.13, 3.1,  9.13, 7.26, 4.74],
    'y3' : [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73],
    'y4' : [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]})

# fit linear regression models by ordinary least squares
set_I_design_matrix = sm.add_constant(anscombe['x1'])
set_I_model = sm.OLS(anscombe['y1'], set_I_design_matrix)
print(set_I_model.fit().summary())

set_II_design_matrix = sm.add_constant(anscombe['x2'])
set_II_model = sm.OLS(anscombe['y2'], set_II_design_matrix)
print(set_II_model.fit().summary())

set_III_design_matrix = sm.add_constant(anscombe['x3'])
set_III_model = sm.OLS(anscombe['y3'], set_III_design_matrix)
print(set_III_model.fit().summary())

set_IV_design_matrix = sm.add_constant(anscombe['x4'])
set_IV_model = sm.OLS(anscombe['y4'], set_IV_design_matrix)
print(set_IV_model.fit().summary())

# create scatter plots
fig = plt.figure()
set_I = fig.add_subplot(2, 2, 1)
set_I.scatter(anscombe['x1'],anscombe['y1'])
set_I.set_title('Set I')
set_I.set_xlabel('x1')
set_I.set_ylabel('y1')
set_I.set_xlim(2, 20)
set_I.set_ylim(2, 14)
```

```
set_II = fig.add_subplot(2, 2, 2)
set_II.scatter(anscombe['x2'],anscombe['y2'])
set_II.set_title('Set II')
set_II.set_xlabel('x2')
set_II.set_ylabel('y2')
set_II.set_xlim(2, 20)
set_II.set_ylim(2, 14)

set_III = fig.add_subplot(2, 2, 3)
set_III.scatter(anscombe['x3'],anscombe['y3'])
set_III.set_title('Set III')
set_III.set_xlabel('x3')
set_III.set_ylabel('y3')
set_III.set_xlim(2, 20)
set_III.set_ylim(2, 14)

set_IV = fig.add_subplot(2, 2, 4)
set_IV.scatter(anscombe['x4'],anscombe['y4'])
set_IV.set_title('Set IV')
set_IV.set_xlabel('x4')
set_IV.set_ylabel('y4')
set_IV.set_xlim(2, 20)
set_IV.set_ylim(2, 14)

plt.subplots_adjust(left=0.1, right=0.925, top=0.925, bottom=0.1,
    wspace = 0.3, hspace = 0.4)
plt.savefig('fig_anscombe_Python.pdf', bbox_inches = 'tight', dpi=None,
    facecolor='w', edgecolor='b', orientation='portrait', papertype=None,
    format=None, transparent=True, pad_inches=0.25, frameon=None)

# Suggestions for the student:
# See if you can develop a quartet of your own,
# or perhaps just a duet, two very different data sets
# with the same fitted model.
```

Exhibit 1.2. *Programming the Anscombe Quartet (R)*

```
# The Anscombe Quartet (R)

# demonstration data from
# Anscombe, F. J. 1973, February. Graphs in statistical analysis.
#  The American Statistician 27: 1721.

# define the anscombe data frame
anscombe <- data.frame(
    x1 = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
    x2 = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
    x3 = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5),
    x4 = c(8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8),
    y1 = c(8.04, 6.95,  7.58, 8.81, 8.33, 9.96, 7.24, 4.26,10.84, 4.82, 5.68),
    y2 = c(9.14, 8.14,  8.74, 8.77, 9.26, 8.1, 6.13, 3.1,  9.13, 7.26, 4.74),
    y3 = c(7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73),
    y4 = c(6.58, 5.76,  7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89))

# show results from four regression analyses
with(anscombe, print(summary(lm(y1 ~ x1, data = anscombe))))
with(anscombe, print(summary(lm(y2 ~ x2, data = anscombe))))
with(anscombe, print(summary(lm(y3 ~ x3, data = anscombe))))
with(anscombe, print(summary(lm(y4 ~ x4, data = anscombe))))

# place four plots on one page using standard R graphics
# ensuring that all have the same scales
# for horizontal and vertical axes
pdf(file = "fig_anscombe_R.pdf", width = 8.5, height = 8.5)
par(mfrow=c(2,2), mar=c(5.1, 4.1, 4.1, 2.1))
with(anscombe, plot(x1, y1, xlim=c(2,20), ylim=c(2,14), pch = 19,
    col = "darkblue", cex = 1.5, las = 1, xlab = "x1", ylab = "y1"))
title("Set I")
with(anscombe,plot(x2, y2, xlim=c(2,20), ylim=c(2,14), pch = 19,
    col = "darkblue", cex = 1.5, las = 1, xlab = "x2", ylab = "y2"))
title("Set II")
with(anscombe,plot(x3, y3, xlim=c(2,20), ylim=c(2,14), pch = 19,
    col = "darkblue", cex = 1.5, las = 1, xlab = "x3", ylab = "y3"))
title("Set III")
with(anscombe,plot(x4, y4, xlim=c(2,20), ylim=c(2,14), pch = 19,
    col = "darkblue", cex = 1.5, las = 1, xlab = "x4", ylab = "y4"))
title("Set IV")
dev.off()

# par(mfrow=c(1,1),mar=c(5.1, 4.1, 4.1, 2.1))  # return to plotting defaults
```

Bibliography

Ackland, R. 2013. *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. Los Angeles: Sage.

Adler, J. 2010. *R in a Nutshell: A Desktop Quick Reference*. Sebastopol, Calif.: O'Reilly.

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo 1996. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), *Handbook of Data Mining and Knowledge Discovery*, Chapter 12, pp. 307–328. Menlo Park, Calif. and Cambridge, Mass.: American Association for Artificial Intelligence and MIT Press.

Agresti, A. 2013. *Categorical Data Analysis* (third ed.). New York: Wiley.

Aizaki, H. 2012, September 22. Basic functions for supporting an implementation of choice experiments in R. *Journal of Statistical Software, Code Snippets* 50(2):1–24. `http://www.jstatsoft.org/v50/c02`.

Aizaki, H. 2014. *support.CEs: Basic Functions for Supporting an Implementation of Choice Experiments*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/packages/support.CEs/support.CEs.pdf`.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Aksin, Z., M. Armony, and V. Mehrotra 2007, November–December. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* 16(6):665–688.

Alamar, B. C. 2013. *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers*. New York: Columbia University Press. 205

Albert, J. 2009. *Bayesian Computation with R*. New York: Springer. 283

Albert, J. H. 2003. *Teaching Statistics Using Baseball*. Washington D.C.: The Mathematical Association of America. 207

Albert, J. H. and J. Bennett 2003. *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. New York: Springer. 207

Albert, J. H., J. Bennett, and J. J. Cochran (eds.) 2005. *Anthology of Statistics in Sports*. Alexandria, Va.: ASA-SIAM.

Alfons, A. 2014a. *cvTools: Cross-Validation Tools for Regression Models*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/cvTools/cvTools.pdf.

Alfons, A. 2014b. *simFrame: Simulation Framework*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/simFrame/simFrame.pdf. 221, 289

Alfons, A., M. Templ, and P. Filzmoser 2010, November 16. An object-oriented framework for statistical simulation: The R package simFrame. *Journal of Statistical Software* 37(3):1–36. http://www.jstatsoft.org/v37/i03.

Alfons, A., M. Templ, and P. Filzmoser 2014. *An Object-Oriented Framework for Statistical Simulation: The R Package simFrame*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/simFrame/vignettes/simFrame-intro.pdf. 221, 289

Allen, M. J. and W. M. Yen 2002. *Introduction to Measurement Theory*. Prospect Heights, Ill.: Waveland Press.

Allison, P. D. 2010. *Survival Analysis Using SAS: A Practical Guide* (second ed.). Cary, N.C.: SAS Institute Inc.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding 1993. *Statistical Models Based on Counting Processes*. New York: Springer.

Anscombe, F. J. 1973, February. Graphs in statistical analysis. *The American Statistician* 27: 17–21.

Armstrong, J. S. (ed.) 2001. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Boston: Kluwer.

Asur, S. and B. A. Huberman 2010. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pp. 492–499. Washington, DC, USA: IEEE Computer Society. http://dx.doi.org/10.1109/WI-IAT.2010.63. 150

Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman 2009. Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting* 25:146–166.

Avery, C. and J. Chevalier 1999. Identifying investor sentiment from price paths: The case of football betting. *Journal of Business* 72(4):493–521.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics with R*. Cambridge, UK: Cambridge University Press. 119

Bacon, L. D. 2002. Marketing. In W. Klösgen and J. M. Żytkow (eds.), *Handbook of Data Mining and Knowledge Discovery*, Chapter 34, pp. 715–725. Oxford: Oxford University Press.

Baeza-Yates, R. and B. Ribeiro-Neto 1999. *Modern Information Retrieval*. New York: ACM Press.

Bagozzi, R. P. and Y. Yi 1991. Multitrait-multimethod matrices in consumer research. *Journal of Consumer Research* 17:426–439. 302

Bar-Eli, M., S. Avugos, and M. Raab 2006. Twenty years of "hot hand" research: Review and critique. *Journal of Sport and Exercise* 7:525–553. 208

Barbera, P. 2014. *streamR: Access to Twitter Streaming API via R*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/streamR/streamR.pdf. 291

Barndorff-Nielsen, O. E., J. L. Jensen, and W. S. Kendall (eds.) 1993. *Networks and Chaos—Statistical Procedures and Probabilistic Aspects*. London: Chapman and Hall.

Bates, D. and M. Maechler 2014. *Matrix: Sparse and Dense Matrix Classes and Methods*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/support.CEs/support.CEs.pdf. 294

Bates, D. M. and D. G. Watts 2007. *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Baumohl, B. 2008. *The Secrets of Economic Indicators: Hidden Clues to Future Economic Trends and Investment Opportunities* (second ed.). Upper Saddle River, N.J.: Pearson.

Beazley, D. M. 2009. *Python Essential Reference* (fourth ed.). Upper Saddle River, N.J.: Pearson Education.

Beazley, D. M. and B. K. Jones 2013. *Python Cookbook* (third ed.). Sebastopol, Calif.: O'Reilly.

Becker, R. A. and J. M. Chambers 1984. *S: An Interactive Environment for Data Analysis and Graphics*. Belmont, CA: Wadsworth. Champions of S, S-Plus, and R call this *the brown book*.

Becker, R. A., J. M. Chambers, and A. R. Wilks 1988. *S: An Interactive Environment for Data Analysis and Graphics*. Pacific Grove, Calif.: Wadsworth & Brooks/Cole. Champions of S, S-Plus, and R call this *the blue book*.

Becker, R. A. and W. S. Cleveland 1996. *S-Plus Trellis^{TM} Graphics User's Manual*. Seattle: MathSoft, Inc. 11

Becker, R. A., A. R. Wilks, R. Brownrigg, and T. P. Minka 2014. *maps: Draw Geographical Maps*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/maps/maps.pdf.

Beleites, C. 2014. *arrayhelpers: Convenience Functions for Arrays*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/arrayhelpers/arrayhelpers.pdf.

Belew, R. K. 2000. *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge: Cambridge University Press.

Belsley, D. A., E. Kuh, and R. E. Welsch 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley. 214

Ben-Akiva, M. and S. R. Lerman 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge: MIT Press.

Benninga, S. 2008. *Financial Modeling* (third ed.). Cambridge, Mass.: MIT Press.

Berk, R. A. 2008. *Statistical Learning from a Regression Perspective*. New York: Springer.

Berkelaar, M. 2014. *lpSolve: Interface to Lp_solve v.5.5 to solve linear/integer programs*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/lpSolve/lpSolve.pdf.

Berndt, E. R. 1991. *The Practice of Econometrics*. Reading, Mass.: Addison-Wesley.

Berri, D. J. and M. B. Schmidt 2010. *Stumbling on Wins: Two Economists Expose the Pitfalls on the Road to Victory in Professional Sports*. Upper Saddle River, N.J.: FT Press/Pearson. 205

Berry, D. A. 1996. *Statistics: A Bayesian Perspective*. Belmont, Calif.: Duxbury.

Berry, M. W. and M. Browne 2005. *Google's Page Rank and Beyond: The Science of Search Engine Rankings* (second ed.). Philadelphia: SIAM.

Berthold, M. R., N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel 2007. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. New York: Springer. ISSN 1431-8814. ISBN 978-3-540-78239-1.

Betz, N. E. and D. J. Weiss 2001. Validity. In B. Bolton (ed.), *Handbook of Measurement and Evaluation in Rehabilitation* (third ed.)., pp. 49–73. Gaithersburg, Md.: Aspen Publishers.

Bird, S., E. Klein, and E. Loper 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, Calif.: O'Reilly. http://www.nltk.org/book/. 314

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.

Bivand, R. 2014. *Geographically Weighted Regression*. Comprehensive R Archive Network. 2014. http://cran.at.r-project.org/web/packages/spgwr/vignettes/GWR.pdf.

Bivand, R. and D. Yu 2014. *spgwr: Geographically Weighted Regression*. Comprehensive R Archive Network. 2014. http://cran.at.r-project.org/web/packages/spgwr/spgwr.pdf.

Bivand, R. A., E. J. Pebesma, and V. Gómez-Rubio 2008. *Applied Spatial Data Analysis with R*. New York: Springer. 299

Blei, D., A. Ng, and M. Jordan 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022. 290

Bollen, J., H. Mao, and X. Zeng 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2:1–8.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein 2009. *Introduction to Meta-Analysis*. New York: Wiley. 275

Borg, I. and P. J. F. Groenen 2010. *Modern Multidimensional Scaling: Theory and Applications* (second ed.). New York: Springer. 296

Boser, B. E., I. M. Guyon, and V. N. Vapnik 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Conference on Computational Learning Theory*, pp. 144–152. Association for Computing Machinery Press. 144

Bowman, A. W. and A. Azzalini 1997. *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Box, G. E. P. and D. R. Cox 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)* 26(2):211–252.

Box, G. E. P., W. G. Hunter, and J. S. Hunter 2005. *Statistics for Experimenters: Design, Innovation, and Discovery* (second ed.). New York: Wiley.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel 2008. *Time Series Analysis: Forecasting and Control* (fourth ed.). New York: Wiley. 64

Bozdogan, H. (ed.) 2004. *Statistical Data Mining and Knowledge Discovery*. Boca Raton, Fla.: CRC Press.

Boztug, Y. and T. Reutterer 2008. A combined approach for segment-specific market basket analysis. *European Journal of Operational Research* 187(1):294–312. 55

Breiman, L. 2001a. Random forests. *Machine Learning* 45(1):5–32.

Breiman, L. 2001b. Statistical modeling: The two cultures. *Statistical Science* 16(3):199–215.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone 1984. *Classification and Regression Trees*. New York: Chapman & Hall.

Broughton, D. 2012, November 12–18. Everybody loves bobbleheads. *Sports Business Journal*:9. Retrieved from the World Wide Web at http://www.sportsbusinessdaily.com/Journal/Issues/2012/11/12/Research-and-Ratings/Bobbleheads.aspx. 315

Brown, F. G. 1976. *Principles of Educational and Psychological Testing* (Second ed.). New York: Holt, Rinehart, and Winston. 314

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* 100(469):36–50.

Brownlee, J. 2011. *Clever Algorithms: Nature-Inspired Programming Recipies*. Melbourne, Australia: Creative Commons. http://www.CleverAlgorithms.com. 290

Bruzzese, D. and C. Davino 2008. Visual mining of association rules. In S. Simoff, M. H. Böhlen, and A. Mazeika (eds.), *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, pp. 103–122. New York: Springer.

Buchta, C. and M. Hahsler 2014. *arulesSequences: Mining Frequent Sequences*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/arulesSequences/arulesSequences.pdf.

Bühlmann, P. and S. van de Geer 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York: Springer. 288

Burnham, K. P. and D. R. Anderson 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (second ed.). New York: Springer-Verlag.

Burns, P. 2011, April. *The R Inferno*. http://www.burns-stat.com/pages/Tutor/R_inferno.pdf.

Butts, C. T. 2014. *sna: Tools for Social Network Analysis*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/sna/sna.pdf. 291

Buvač, V. and P. J. Stone 2001, April 2. The General Inquirer user's guide. Software developed with the support of Harvard University and The Gallup Organization.

Cairo, A. 2013. *The Functional Art: An Introduction to Information Graphics and Visualization*. Berkeley, Calif: New Riders.

Cameron, A. C. and P. K. Trivedi 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Campbell, D. T. and D. W. Fiske 1959. Convergent validity and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin* 56:81–105.

Campbell, D. T. and J. C. Stanley 1963. *Experimental and Quasi-Experimental Designs for Research*. Skokie, IL: Rand McNally.

Canadilla, P. 2014. *queueing: Analysis of Queueing Networks and Models*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/queueing/queueing.pdf.

Carlin, B. P. 1996, February. Improved NCAA basketball tournament modeling via point spread and team strength information. *The American Statistician* 50(1):39–43. 205

Carlin, B. P. and T. A. Louis 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall. 283

Carr, D., N. Lewin-Koh, and M. Maechler 2014. *hexbin: Hexagonal Binning Routines*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/hexbin/hexbin.pdf. 11

Carr, D. B. 1991. Looking at large data sets using binned data plots. In A. Buja and P. Tukey (eds.), *Computing and Graphics in Statistics*, pp. 7–39. New York: Springer-Verlag.

Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield 1987. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association* 83:424–436.

Carroll, B., P. Palmer, J. Thorn, and D. Pietrusza (eds.) 1998. *The Hidden Game of Football: The Next Edition*. Kingston, N.Y.: Total Sports.

Carroll, J. D. and P. E. Green 1995. Psychometric methods in marketing research: Part I, conjoint analysis. *Journal of Marketing Research* 32:385–391.

Carroll, J. D. and P. E. Green 1997. Psychometric methods in marketing research: Part II, multidimensional scaling. *Journal of Marketing Research* 34:193–204. 296

Caudill, M. and C. Butler 1990. *Naturally Intelligent Systems*. Cambridge: MIT Press.

Cespedes, F. V., J. P. Dougherty, and B. S. Skinner, III 2013, Winter. How to identify the best customers for your business. *MIT Sloan Management Review* 54(2):53–59.

Chakrabarti, S. 2003. *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco: Morgan Kaufmann.

Chambers, J. M. 1998. *Programming with Data: A Guide to the S Language*. New York: Springer-Verlag. We could call this "the green book." Original documentation for S4 classes in S, S-Plus, and R.

Chambers, J. M. 2008. *Software for Data Analysis: Programming in R*. New York: Springer.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey 1983. *Graphical Methods for Data Analysis*. Belmont, Calif.: Wadsworth.

Chambers, J. M. and T. J. Hastie (eds.) 1992. *Statistical Models in S*. Pacific Grove, Calif.: Wadsworth & Brooks/Cole. Champions of S, S-Plus, and R call this "the white book." It introduced statistical modeling syntax using S3 classes.

Chang, W. 2013. *R Graphics Cookbook*. Sebastopol, Calif.: O'Reilly.

Chang, W. 2014. *ggvis: Interactive Grammar of Graphics*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/ggvis/ggvis.pdf.

Charniak, E. 1993. *Statistical Language Learning*. Cambridge: MIT Press.

Chatfield, C. (ed.) 2003. *The Analysis of Time Series: An Introduction* (sixth ed.). New York: CRC Press.

Chatterjee, S. and A. S. Hadi 2012. *Regression Analysis by Example* (fifth ed.). New York: Wiley.

Chau, M. and H. Chen 2003. Personalized and focused Web spiders. In N. Zhong, J. Liu, and Y. Yao (eds.), *Web Intelligence*, Chapter 10, pp. 198–217. New York: Springer.

Chen, D.-G. and K. E. Peace 2013. *Applied Meta-Analysis with R*. Boca Raton, Fla.: Chapman & Hall/CRC.

Chen, H. and M. Chau 2004. Web mining: Machine learning for Web applications. In B. Cronin (ed.), *Annual Review of Information Science and Technology*, Volume 38, Chapter 6, pp. 289–329. Medford, N.J.: Information Today.

Chen, H., M. Chau, and D. Zeng 2002. CI Spider: A tool for competitive intelligence on the Web. *Decision Support Systems* 34:1–17.

Cherkassky, V. and F. Mulier 1998. *Learning from Data: Concepts, Theory, and Methods*. New York: Wiley.

Chihara, L. and T. Hesterberg 2011. *Mathematical Statistics with Resampling and R*. New York: Wiley.

Chodorow, K. 2013. *MongoDB: The Definitive Guide* (second ed.). Sebastopol, Calif.: O'Reilly.

Christensen, R. 1997. *Log-Linear Models and Logistic-Regression* (second ed.). New York: Springer.

Christianini, N. and J. Shawe-Taylor 2000. *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press.

Chun, W. J. 2007. *Core Python Programming* (second ed.). Upper Saddle River, N.J.: Pearson Education.

Chun, W. J. 2012. *Core Python Applications Programming* (third ed.). Upper Saddle River, N.J.: Pearson Education.

Cleveland, W. S. 1993. *Visualizing Data*. Murray Hill, N.J.: AT&T Bell Laboratories. Initial documentation for trellis graphics in S-Plus. 11

Cleveland, W. S. 1994. *The Elements of Graphing Data*. Murray Hill, N.J.: AT&T Bell Laboratories.

Clifton, B. 2012. *Advanced Web Metrics with Google Analytics* (third ed.). New York: Wiley.

Cochran, W. G. 1977. *Sampling Techniques*. New York: Wiley.

Cochran, W. G. and G. M. Cox 1957. *Experimental Designs* (Second ed.). New York: Wiley.

Cohen, J. 1960, April. A coefficient of agreement for nominal data. *Educational and Psychological Measurement* 20(1):37–46. 287

Commandeur, J. J. F. and S. J. Koopman (eds.) 2007. *An Introduction to State Space Time Series Analysis*. Oxford: Oxford University Press. 66

Congdon, P. 2001. *Bayesian Statistical Modeling*. New York: Wiley.

Congdon, P. 2003. *Applied Bayesian Modeling*. New York: Wiley.

Conway, D. and J. M. White 2012. *Machine Learning for Hackers*. Sebastopol, Calif.: O'Reilly.

Cook, R. D. 1998. *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cook, R. D. 2007. Fisher lecture: Dimension reduction in regression. *Statistical Science* 22: 1–26.

Cook, R. D. and S. Weisberg 1999. *Applied Regression Including Computing and Graphics*. New York: Wiley.

Cook, T. D. and D. T. Campbell 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Cooper, L. G. 1999. Market share models. In J. Eliashberg and G. L. Lilien (eds.), *Handbook of Operations Research and Management Science: Vol. 5, Marketing*, Chapter 1, pp. 259–314. New York: Elsevier North Holland. 254

Cooper, L. G. and M. Nakanishi 1988. *Market-Share Analysis*. Norwell, Mass.: Kluwer. 254

Cowpertwait, P. S. P. and A. V. Metcalfe 2009. *Introductory Time Series with R*. New York: Springer.

Cox, D. R. 1958. *Planning of Experiments*. New York: Wiley.

Cox, D. R. 1970. *Analysis of Binary Data*. London: Chapman and Hall.

Cox, T. F. and M. A. A. Cox 1994. *Multidimensional Scaling*. London: Chapman & Hall. 296

Craddock, J. (ed.) 2012. *VideoHound's Golden Movie Retriever 2013: The Complete Guide to Movies on All Home Entertainment Formats*. Farmington Hills, Mich.: Gale.

Cranor, L. F. and B. A. LaMacchia 1998. Spam! *Communications of the ACM* 41(8):74–83.

Cressie, N. 1993. *Statistics for Spatial Data* (revised ed.). New York: Wiley. 299

Cressie, N. and C. K. Wikle 2011. *Statistics for Spatio-Temporal Data*. New York: Wiley.

Cristianini, N. and J. Shawe-Taylor 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

Croissant, Y. 2014. *mlogit: Multinomial Logit Model*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/packages/mlogit/mlogit.pdf`. 254

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297–334. 314

Cronbach, L. J. 1995. Giving method variance its due. In P. E. Shrout and S. T. Fiske (eds.), *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*, Chapter 10, pp. 145–157. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cryer, J. D. and K.-S. Chan 2008. *Time Series Analysis with Applications in R* (second ed.). New York: Springer.

Cumming, G. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Daconta, M. C., L. J. Obrst, and K. T. Smith 2003. *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*. New York: Wiley.

Dalgaard, P. 2002. *Introductory Statistics with R*. New York: Springer-Verlag.

Dantzig, G. B. 1954. A comment on Edie's *Traffic Delays at Toll Booths*. *Operations Research* 2(2):107–108.

Davenport, T. H. and J. G. Harris 2007. *Competing on Analytics: The New Science of Winning*. Boston: Harvard Business School Press. 204, 278

Davenport, T. H., J. G. Harris, and R. Morison 2010. *Analytics at Work: Smarter Decisions, Better Results*. Boston: Harvard Business School Press. 278

Davies, R. and D. Rogers (eds.) 1984. *Store Location and Store Assessment Research*. New York: Wiley. 299

Davison, A. C. and D. V. Hinkley 1997. *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

Davison, M. L. 1992. *Multidimensional Scaling*. Melbourne, Fla.: Krieger. 296

Dean, J. and S. Ghemawat 2004. MapReduce: Simplifed Data Processing on Large Clusters. Retrieved from the World Wide Web at `http://static.usenix.org/event/osdi04/tech/full_papers/dean/dean.pdf`.

Dehuri, S., M. Patra, B. B. Misra, and A. K. Jagadev (eds.) 2012. *Intelligent Techniques in Recommendation Systems*. Hershey, Pa.: IGI Global.

Delen, D., R. Sharda, and P. Kumar 2007. Movie forecast guru: A Web-based DSS for Hollywood managers. *Decision Support Systems* 43(4):1151–1170. 150

Deming, W. E. 1950. *Some Theory of Sampling*. New York: Wiley. Republished in 1966 by Dover Publications, New York.

Demšar, J. and B. Zupan 2013. Orange: Data mining fruitful and fun—A historical perspective. *Informatica* 37:55–60. 289, 337

Dickson, P. R. 1997. *Marketing Management* (second ed.). Orlando, Fla.: Harcourt Brace & Company. 26

Diggle, P. J., K.-Y. Liang, and S. L. Zeger 1994. *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

DiPierro, M. 2013. *Annotated Algorithms in Python with Applications in Physics, Biology, and Finance*. Chicago: experts4solutions.

Dippold, K. and H. Hruschka 2013. Variable selection for market basket analysis. *Computational Statistics* 28(2):519–539. `http://dx.doi.org/10.1007/s00180-012-0315-3`. ISSN 0943-4062. 55

Downey, A. B. 2012. *Think Python: How to Think Like a Computer Scientist*. Sebastopol, Calif.: O'Reilly.

Draper, N. R. and H. Smith 1998. *Applied Regression Analysis* (third ed.). New York: Wiley.

Dubins, L. E. and L. J. Savage 1965. *Inequalities for Stochastic Processes: How to Gamble If You Must*. New York: Dover. 206

Duda, R. O., P. E. Hart, and D. G. Stork 2001. *Pattern Classification* (second ed.). New York: Wiley.

Dumais, S. T. 2004. Latent semantic analysis. In B. Cronin (ed.), *Annual Review of Information Science and Technology*, Volume 38, Chapter 4, pp. 189–230. Medford, N.J.: Information Today.

Durbin, J. and S. J. Koopman 2012. *Time Series Analysis by State Space Methods* (second ed.). New York: Oxford University Press. 66

Easley, D. and J. Kleinberg 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge, UK: Cambridge University Press.

Eddelbuettel, D. 2014. *CRAN Task View: Empirical Finance*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/views/Finance.html`. 300

Eden, S. 2013, March 4. Meet the world's top NBA gambler. *ESPN The Magazine (Analytics Issue)*. `http://espn.go.com/blog/playbook/dollars/post/_/id/2935/meet-the-worlds-top-nba-gambler`. 207

Efron, B. 1986. Why isn't everyone a Bayesian (with commentary). *The American Statistician* 40(1):1–11.

Efron, B. 2012. *Large-Scale Inference: Empirical Bayes Methods of Estimation, Testing, and Prediction* (reprint ed.). Cambridge, UK: Cambridge University Press. 275

Efron, B. and R. Tibshirani 1993. *An Introduction to the Bootstrap*. London: Chapman and Hall.

Elff, M. 2014. *mclogit: Mixed Conditional Logit*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/mclogit/mclogit.pdf. 254

Eliashberg, J. and G. L. Lilien (eds.) 1993. *Handbooks in Operations Research and Management Science: Volume 5 Marketing*. New York: North Holland.

Enders, W. 2010. *Applied Econometric Time Series* (third ed.). New York: Wiley.

Engelbrecht, A. P. 2007. *Computational Intelligence: An Introduction* (second ed.). New York: Wiley. 290

Ernst, A. T., H. Jiang, M. Krishnamoorthy, and D. Sier 2004. Staff scheduling and rostering: A review of applications, methods, and models. *European Journal of Operations Research* 153:3–27.

Everitt, B. 2005. *R and S-Plus Companion to Multivariate Analysis*. New York: Springer.

Everitt, B. and G. Dunn 2001. *Applied Multivariate Data Analysis* (second ed.). New York: Wiley. 119

Everitt, B. and S. Rabe-Hesketh 1997. *The Analysis of Proximity Data*. London: Arnold.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl 2011. *Cluster Analysis* (fifth ed.). New York: Wiley.

Fader, P. S. and B. G. S. Hardie 1996, November. Modeling consumer choice among SKUs. *Journal of Marketing Research* 33:442–452.

Fader, P. S. and B. G. S. Hardie 2002. A note on an integrated model of consumer buying behavior. *European Journal of Operational Research* 139(3):682–687.

Faraway, J. J. 2004. *Linear Models with R*. Boca Raton, Fla.: Chapman & Hall/CRC.

Fawcett, T. 2003, January 7. ROC graphs: Notes and practical considerations for researchers. http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.) 1996. *Advances in Knowledge Discovery and Data Mining*. Cambridge: MIT Press.

Feinerer, I. 2012. *Introduction to the wordnet Package*. Comprehensive R Archive Network. 2012. http://cran.r-project.org/web/packages/wordnet/vignettes/wordnet.pdf. 119

Feinerer, I. 2014. *Introduction to the tm Package*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf.

Feinerer, I. and K. Hornik 2014a. *tm: Text Mining Package*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/tm/tm.pdf.

Feinerer, I. and K. Hornik 2014b. *wordnet: WordNet Interface*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/wordnet/wordnet.pdf. 119

Feinerer, I., K. Hornik, and D. Meyer 2008, 3 31. Text mining infrastructure in R. *Journal of Statistical Software* 25(5):1–54. http://www.jstatsoft.org/v25/i05. ISSN 1548-7660.

Feldman, D. M. 2002a, Winter. The pricing puzzle. *Marketing Research*:14–19.

Feldman, R. 1999. Mining unstructured data. In *Tutorial Notes of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 182–236. ACM Press. ISBN 1-58113-171-2.

Feldman, R. 2002b. Text mining. In W. Klösgen and J. M. Żytkow (eds.), *Handbook of Data Mining and Knowledge Discovery*, Chapter 38, pp. 749–757. Oxford: Oxford University Press.

Feldman, R., Y. Aumann, Y. Liberzon, K. Ankori, J. Schler, and B. Rosenfeld 2001. A domain independent environment for creating information extraction modules. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pp. 586–588. ACM Press. ISBN 1-58113-436-3.

Feldman, R. and H. Hirsh 1997. Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems* 9(1):83–97. ISSN 0925-9902.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press. 119

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications* (third ed.), Volume I. New York: Wiley. 206

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications*, Volume 2. New York: Wiley. 206

Fellows, I. 2014a. wordcloud makes words less cloudy. Retrieved from the World Wide Web at `http://blog.fellstat.com/`. 119, 337

Fellows, I. 2014b. *wordcloud: Word Clouds*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf`. 119, 337

Fenzel, D., J. Hendler, H. Lieberman, and W. Wahlster (eds.) 2003. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. Cambridge: MIT Press.

Few, S. 2009. *Now You See It: Simple Visualization Techniques and Quantitative Analysis*. Oakland, Calif.: Analytics Press.

Fienberg, S. E. 2007. *Analysis of Cross-Classified Categorical Data* (second ed.). New York: Springer.

Firth, D. 1991. Generalized linear models. In D. Hinkley and E. Snell (eds.), *Statistical Theory and Modeling: In Honour of Sir David Cox, FRS*, Chapter 3, pp. 55–82. London: Chapman and Hall.

Fishelson-Holstine, H. 2004, February. The role of credit scoring in increasing homeownership for underserved populations. Joint Center for Housing Studies Working Paper. Retrieved from the World Wide Web at `http://jchs.harvard.edu/sites/jchs.harvard.edu/files/babc_04-12.pdf`. 300

Fisher, R. A. 1970. *Statistical Methods for Research Workers* (fourteenth ed.). Edinburgh: Oliver and Boyd. First edition published in 1925. 283

Fisher, R. A. 1971. *Design of Experiments* (ninth ed.). New York: Macmillan. First edition published in 1935. 283

Fiske, D. W. 1971. *Measuring the Concepts of Personality*. Chicago: Aldine. 314

Fogel, D. B. 2001. *Blondie24: Playing at the Edge of AI*. San Francisco: Morgan Kaufmann. 206

Foster, G. 2013, August 26. Behold the power of a free bobblehead doll. *The Wall Street Journal*. Retrieved from the World Wide Web at `http://online.wsj.com/article/SB10001424127887323407104579037182599696524.html`. 315

Fotheringham, A. S., C. Brunsdon, and M. Charlton 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New York: Wiley. 221

Fox, J. 2002, January. Robust regression: Appendix to an R and S-PLUS companion to applied regression. Retrieved from the World Wide Web at http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-robust-regression.pdf. 288

Fox, J. 2014. *car: Companion to Applied Regression*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/car/car.pdf.

Fox, J. and S. Weisberg 2011. *An R Companion to Applied Regression* (second ed.). Thousand Oaks, Calif.: Sage. 143

Franks, B. 2012. *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. Hoboken, N.J.: Wiley. 278

Frees, E. W. and T. W. Miller 2004. Sales forecasting with longitudinal data models. *International Journal of Forecasting* 20:99–114.

Friedl, J. E. F. 2006. *Mastering Regular Expressions* (third ed.). Sebastopol, Calif.: O'Reilly. 119

Friedman, J. H. 1991. Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* 19(1):1–141.

Friendly, M. 1994. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association* 89:17–23.

Friendly, M. 2000. *Visualizing Categorical Data*. Cary, N.C.: SAS Institute.

Gabriel, K. R. 1971. The biplot graphical display of matrices with application to principal component analysis. *Biometrika* 58:453–467. 107, 296

Garcia-Molina, H., J. D. Ullman, and J. Widom 2009. *Database Systems: The Complete Book* (second ed.). Upper Saddle River, N.J.: Prentice-Hall.

Garey, M. R. and D. S. Johnson 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W. H. Freeman.

Geisler, C. 2004. *Analyzing Streams of Language: Twelve Steps to the Systematic Coding of Text, Talk, and Other Verbal Data*. New York: Pearson Education.

Geisser, S. 1993. *Predictive Inference: An Introduction*. New York: Chapman & Hall. 5, 275, 283

Geisser, S. and W. O. Johnson 2006. *Modes of Parametric Statistical Inference*. New York: Wiley. 275

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin 1995. *Bayesian Data Analysis*. London: Chapman & Hall. 283

Gelman, A. and J. Hill 2007. *Data Analysis Using Regression and Mulitlevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press. 275

Gelman, A., J. Hill, Y.-S. Su, M. Yajima, and M. G. Pittau 2014. *mi: Missing Data Imputation and Model Checking*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/mi/mi.pdf.

Gentle, J. E. 2002. *Elements of Computational Statistics*. New York: Springer.

Gentle, J. E. 2003. *Random Number Generation and Monte Carlo Methods* (second ed.). New York: Springer.

Gentry, J. 2014a. *Twitter Client for R*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/twitteR/vignettes/twitteR.pdf. 291

Gentry, J. 2014b. *twitteR: R based Twitter Client*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/twitteR/twitteR.pdf. 291

Ghiselli, E. E. 1964. *Theory of Psychological Measurement*. New York: McGraw-Hill. 314

Glickman, M. E. and H. S. Stern 1998. A state-space model for national football league scores. *Journal of the American Statistical Association* 93(441):25–35. 205

Gnanadesikan, R. 1997. *Methods for Statistical Data Analysis of Multivariate Observations* (second ed.). New York: Wiley. 119

Goldman, S. (ed.) 2005. *Mind Game: How the Boston Red Sox got Smart, Won a World Series, and Created a New Blueprint for Winning*. New York: Workman Publishing. 204

Gordon, A. D. 1999. *Classification* (second ed.). Boca Raton, Fla.: Chapman & Hall/CRC.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27:857–871.

Gower, J. C. and D. J. Hand 1996. *Biplots*. London: Chapman & Hall. 107, 296

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37:424–438.

Graybill, F. A. 1961. *Introduction to Linear Statistical Models, Volume 1*. New York: McGraw-Hill.

Graybill, F. A. 2000. *Theory and Application of the Linear Model*. Stamford, Conn.: Cengage Learning.

Greene, W. H. 2012. *Econometric Analysis* (seventh ed.). Upper Saddle River, N.J.: Pearson Prentice Hall. 253

Gries, S. T. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge.

Gries, S. T. 2013. *Statistics for Linguistics with R: A Pratical Introduction* (second revised ed.). Berlin: De Gruyter Mouton. 119

Grolemund, G. 2014. *Hands-On Programming with R: Write Your Own Functions and Simulations*. Sebastopol, Calif.: O'Reilly.

Grolemund, G. and H. Wickham 2011, April 7. Dates and times made easy with lubridate. *Journal of Statistical Software* 40(3):1–25. http://www.jstatsoft.org/v40/i03.

Grolemund, G. and H. Wickham 2014. *lubridate: Make Dealing with Dates a Little Easier*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/lubridate/lubridate.pdf.

Gross, D., J. F. Shortle, J. M. Thompson, and C. M. Harris 2008. *Fundamentals of Queueing Theory* (fourth ed.). New York: Wiley.

Grothendieck, G. 2014a. *gsubfn: Utilities for Strings and Function Arguments*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/gsubfn/gsubfn.pdf.

Grothendieck, G. 2014b. *gsubfn: Utilities for Strings and Function Arguments (Vignette)*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/gsubfn/vignettes/gsubfn.pdf.

Groves, R. M., F. J. Fowler, Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau 2009. *Survey Methodology* (second ed.). New York: Wiley.

Guilford, J. P. 1954. *Psychometric Methods* (second ed.). New York: McGraw-Hill. First edition published in 1936. 314

Gulliksen, H. 1950. *Theory of Mental Tests*. New York: Wiley. 111, 314

Gustafsson, A., A. Herrmann, and F. Huber (eds.) 2000. *Conjoint Measurement: Methods and Applications*. New York: Springer-Verlag.

Hahsler, M. 2014a. *arulesNBMiner: Mining NB-Frequent Itemsets and NB-Precise Rules*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/arulesNBMiner/arulesNBMiner.pdf.

Hahsler, M. 2014b. *recommenderlab: A Framework for Developing and Testing Recommendation Algorithms*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf.

Hahsler, M. 2014c. *recommenderlab: Lab for Developing and Testing Recommender Algorithms*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/recommenderlab/recommenderlab.pdf.

Hahsler, M., C. Buchta, B. Grün, and K. Hornik 2014a. *arules: Mining Association Rules and Frequent Itemsets*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/arules/arules.pdf.

Hahsler, M., C. Buchta, B. Grün, and K. Hornik 2014b. *Introduction to arules: A Computational Environment for Mining Association Rules and Frequent Itemsets*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf.

Hahsler, M., C. Buchta, and K. Hornik 2008. Selective association rule generation. *Computational Statistics* 23:303–315. 55

Hahsler, M. and S. Chelluboina 2014a. *arulesViz: Visualizing Association Rules and Frequent Itemsets*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/arulesViz/arulesViz.pdf.

Hahsler, M. and S. Chelluboina 2014b. *Visualizing Association Rules: Introduction to the R-extension Package arulesViz*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf.

Hahsler, M., S. Chelluboina, K. Hornik, and C. Buchta 2011. The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets. *Journal of Machine Learning Research* 12:2021–2025.

Hahsler, M., B. Grün, and K. Hornik 2005, September 29. arules: A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software* 14(15):1–25. http://www.jstatsoft.org/v14/i15.

Hahsler, M., K. Hornik, and T. Reutterer 2006. Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul (eds.), *Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 598–605. New York: Springer.

Hakes, J. K. and R. D. Sauer 2006. An economic evaluation of the "Moneyball" hypothesis. *Journal of Economic Perspectives* 20(3):173–185. 204

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton, N.J.: Princeton University Press.

Han, J., M. Kamber, and J. Pei 2011. *Data Mining: Concepts and Techniques* (third ed.). San Francisco: Morgan Kaufmann.

Hand, D., H. Mannila, and P. Smyth 2001. *Principles of Data Mining*. Cambridge: MIT Press.

Hand, D. J. 1997. *Construction and Assessment of Classification Rules*. New York: Wiley.

Hanssens, D. M., L. J. Parsons, and R. L. Schultz 2001. *Market Response Models: Econometric and Time Series Analysis* (second ed.). Boston: Kluwer.

Harrell, Jr., F. E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Harrington, P. 2012. *Machine Learning in Action*. Shelter Island, N.Y.: Manning. 55

Hart, R. P. 2000a. *Campaign Talk: Why Elections Are Good for Us*. Princeton, N.J.: Princeton University Press.

Hart, R. P. 2000b. *DICTION 5.0: The Text Analysis Program*. Thousand Oaks, Calif.: Sage.

Hart, R. P. 2001. Redeveloping Diction: theoretical considerations. In M. D. West (ed.), *Theory, Method, and Practice in Computer Content Analysis*, Chapter 3, pp. 43–60. Westport, Conn.: Ablex.

Hartigan, J. A. and B. Kleiner 1984. A mosaic of television ratings. *The American Statistician* 38(1):32–35.

Hastie, T., R. Tibshirani, and J. Friedman 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (second ed.). New York: Springer. 275

Hastie, T. J. 1992a. Generalized additive models. In J. M. Chambers and T. J. Hastie (eds.), *Statistical Models in S*, Chapter 7, pp. 249–307. Pacific Grove, Calif.: Wadsworth & Brooks/Cole.

Hastie, T. J. 1992b. Generalized linear models. In J. M. Chambers and T. J. Hastie (eds.), *Statistical Models in S*, Chapter 6, pp. 195–247. Pacific Grove, Calif.: Wadsworth & Brooks/Cole.

Hastie, T. J. and R. Tibshirani 1990. *Generalized Additive Models*. London: Chapman and Hall.

Hausser, R. 2001. *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language* (second ed.). New York: Springer-Verlag.

Haykin, S. 2008. *Neural Networks and Learning Machines* (third ed.). Upper Saddle River, N.J.: Prentice Hall.

Hearst, M. A. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64. 118

Hearst, M. A. 1999, June 20–26. Untangling text data mining. In *Proceedings of ACL'99: The 37th Annual Meeting of the Association for Computational Linguistics*. Retrieved from the World Wide Web on March 20, 2004, at: http://www.sims.berkeley.edu/hearst.

Hearst, M. A. 2003, October 17. What is text mining? Retrieved from the World Wide Web on March 20, 2004, at: http://www.sims.berkeley.edu/hearst.

Heer, J., M. Bostock, and V. Ogievetsky 2010, May 1. A tour through the visualization zoo: A survey of powerful visualization techniques, from the obvious to the obscure. *acmqueue: Association for Computing Machinery*:1–22. Retrieved from the World Wide Web at `http://queue.acm.org/detail.cfm?id=1805128`.

Heer, J., N. Kong, and M. Agrawala 2009. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1303–1312. New York: ACM. `http://doi.acm.org/10.1145/1518701.1518897`.

Heiberger, R. M. and B. Holland 2004. *Statistical Analysis and Data Display: An Intermediate Course*. New York: Springer.

Hemenway, K. and T. Calishain 2004. *Spidering Hacks: 100 Industrial-Strength Tips & Tools*. Sabastopol, Calif.: O'Reilly.

Hensher, D. A. and L. W. Johnson 1981. *Applied Discrete-Choice Modeling*. New York: Wiley.

Hensher, D. A., J. M. Rose, and W. H. Greene 2005. *Applied Choice Analysis: A Primer*. Cambridge: Cambridge University Press.

Hinkelmann, K. and O. Kempthorne 1994. *Design and Analysis of Experiments: Volume I. Introduction to Experimental Design*. New York: Wiley. Revision of Kempthorne (1952).

Hinkley, D. V., N. Reid, and E. J. Snell (eds.) 1991. *Statistical Theory and Modeling*. London: Chapman and Hall. 275, 283

Hlavac, M. 2014. *stargazer: LaTeX Code for Well Formatted Regression and Summary Statistics Tables*. Cambridge, USA: Comprehensive R Archive Network and Harvard University. 2014. http://CRAN.R-project.org/package=stargazer.

Hoerl, A. E. and R. W. Kennard 2000. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 42(1):80–86. Reprinted from *Technometrics*, volume 12. 288

Hoff, P. D. 2009. *A First Course in Bayesian Statistical Methods*. New York: Springer.

Hoffmann, T. 2014. *batch: Batching Routines in Parallel and Passing Command-Line Arguments to R*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/batch/batch.pdf.

Holden, K., D. A. Peel, and J. L. Thompson 1990. *Economic Forecasting: An Introduction*. Cambridge, UK: Cambridge University Press.

Hollis, N. 2005, Fall. Branding unmasked: Expose the mysterious consumer purchase process. *Marketing Research* 17(3):24–29.

Honaker, J., G. King, and M. Blackwell 2014. *Amelia II: A Program for Missing Data*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/Amelia/Amelia.pdf.

Hornik, K. 2014a. *RWeka Odds and Ends*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/RWeka/vignettes/RWeka.pdf.

Hornik, K. 2014b. *RWeka: R/Weka Interface*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/RWeka/RWeka.pdf.

Hosmer, D. W., S. Lemeshow, and S. May 2013. *Applied Survival Analysis: Regression Modeling of Time to Event Data* (second ed.). New York: Wiley.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant 2013. *Applied Logistic Regression* (third ed.). New York: Wiley. 143

Hothorn, Leisch, Zeileis, and Hornik 2005, September. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics* 14(3):675–699.

Hothorn, T., K. Hornik, and A. Zeileis 2014. *party: A Laboratory for Recursive Partytioning*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/party/vignettes/party.pdf.

Hothorn, T. and A. Zeileis 2014a. *partykit vignette: A Toolkit for Recursive Partytioning*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/partykit/vignettes/partykit.pdf.

Hothorn, T. and A. Zeileis 2014b. *partykit: A Toolkit for Recursive Partytioning*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/partykit/partykit.pdf.

Hothorn, T., A. Zeileis, R. W. Farebrother, C. Cummins, G. Millo, and D. Mitchell 2014. *lmtest: Testing Linear Regression Models*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/lmtest/lmtest.pdf.

Hu, M. and B. Liu 2004, August 22–25. Mining and summarizing customer reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*. Full paper available from the World Wide Web at http://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf Original source for opinion and sentiment lexicon, available from the World Wide Web at http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon.

Huber, J. and K. Zwerina 1996. The importance of utility balance in efficient choice designs. *Journal of Marketing Research* 33:307–317.

Huberman, B. A. 2001. *The Laws of the Web: Patterns in the Ecology of Information*. Cambridge: MIT Press.

Huet, S., A. Bouvier, M.-A. Poursat, and E. Jolivet 2004. *Statistical Tools for Nonlinear Regression: A Practical Guide with S-Plus and R Examples* (second ed.). New York: Springer.

Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang 2011. Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis* 55:2579–2589.

Hyndman, R. J. and G. Athanasopoulos 2014. *Forecasting: Principles and Practice*. Online: OTexts. https://www.otexts.org/fpp.

Hyndman, R. J., G. Athanasopoulos, S. Razbash, D. Schmidt, Z. Zhou, and Y. Khan 2014. *forecast: Forecasting Functions for Time Series and Linear Models*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/forecast/forecast.pdf.

Hyndman, R. J., A. B. Koehler, J. K. Ord, and R. D. Snyder 2008. *Forecasting with Exponential Smoothing: The State Space Approach*. New York: Springer. 66

Ihaka, R., P. Murrell, K. Hornik, J. C. Fisher, and A. Zeileis 2014. *colorspace: Color Space Manipulation*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/colorspace/colorspace.pdf.

Indurkhya, N. and F. J. Damerau (eds.) 2010. *Handbook of Natural Language Processing* (second ed.). Boca Raton, Fla.: Chapman and Hall/CRC.

Ingersoll, G. S., T. S. Morton, and A. L. Farris 2013. *Taming Text: How to Find, Organize, and Manipulate It*. Shelter Island, N.Y.: Manning. 290

Inselberg, A. 1985. The plane with parallel coordinates. *The Visual Computer* 1(4):69–91. 246

Ivezic, Z., A. J. Connolly, J. T. VanderPlas, and A. Gray 2014. *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton, N.J.: Princeton University Press.

Izenman, A. J. 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. New York: Springer. 119

Jackson, M. O. 2008. *Social and Economic Networks*. Princeton, N.J.: Princeton University Press.

James, B. 2010. *The New Bill James Historical Baseball Abstract*. New York: Free Press. 205

James, G., D. Witten, T. Hastie, and R. Tibshirani 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Janert, P. K. 2011. *Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists*. Sebastopol, Calif.: O'Reilly.

Janssen, A. J. E. M., J. S. H. van Leeuwaarden, and B. Zwart 2011, November–December. Refining square-root safety staffing by expanding Erlang C. *Operations Research* 59 (6):1512–1522.

Johnson, K. 2008. *Quantitative Methods in Linguistics*. Malden, Mass.: Blackwell Publishing. 119

Johnson, M. E. 1987. *Multivariate Statistical Simulation*. New York: Wiley.

Johnson, R. A. and D. W. Wichern 1998. *Applied Multivariate Statistical Analysis* (fourth ed.). Upper Saddle River, N.J.: Prentice Hall. 119

Jones, Maillardet, and Robinson 2014. *Introduction to Scientific Programming and Simulation Using R* (second ed.). Boca Raton, Fla.: Chapman & Hall/CRC.

Joula, P. 2008. *Authorship Attribution*. Hanover, Mass.: Now Publishers. 118

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl, and T.-C. Lee 1985. *The Theory and Practice of Econometrics* (second ed.). New York: Wiley.

Jurafsky, D. and J. H. Martin 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (second ed.). Upper Saddle River, N.J.: Prentice Hall.

Kabacoff, R. 2011. *R in Action*. Shelter Island, N.J.: Manning.

Kadushin, C. 2012. *Understanding Social Networks*. New York: Oxford University Press.

Kahin, B. and H. R. Varian (eds.) 2000. *Internet Publishing: The Economics of Digital Information and Intellectual Property*. Cambridge, Mass.: MIT Press.

Kahn, L. M. 2000. The sports business as a labor market laboratory. *Journal of Economic Perspectives* 14(3):75–94. 204

Kaluzny, S. P., S. C. Vega, T. P. Cardoso, and A. A. Shelly 1998. *S+ Spatial Statistics: User's Manual for Windows and UNIX*. New York: Springer-Verlag.

Kaplan, D. T. 2012. *Statistical Modeling: A Fresh Approach* (second ed.). St. Paul, Minn.: Project Mosaic.

Kaplan, E. H. and S. J. Garstka 2001. March madness and the office pool. *Management Science* 47(3):369–382. 205

Kass, R. E. and A. E. Raftery 1995. Bayes factors. *Journal of the American Statistical Association* 90:773–795.

Kaufman, L. and P. J. Rousseeuw 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Keller, J. B. 1994. A characterization of the Poisson distribution and the probability of winning a game. *The American Statistician* 48(4):294–298.

Kelly, E. F. and P. J. Stone (eds.) 1975. *Computer Recognition of English Word Senses*. Amsterdam: North-Holland.

Kempthorne, O. 1952. *The Design and Analysis of Experiments*. New York: Wiley. Also see Hinkelmann and Kempthorne (1994).

Kennedy, P. 2008. *A Guide to Econometrics* (sixth ed.). New York: Wiley.

Keppel, G. and T. D. Wickens 2004. *Design and Analysis: A Researcher's Handbook* (fourth ed.). Upper Saddle River: N.J.: Pearson.

Keri, J. (ed.) 2006. *Baseball Between the Numbers: Why Everything You Know About the Game is Wrong*. New York: Basic Books. 205

Kirk, R. E. 2013. *Experimental Design: Procedures for the Behavioral Sciences* (fourth ed.). Thousand Oaks, Calif.: Sage.

Kleiber, C. and A. Zeileis 2008. *Applied Econometrics with R*. New York: Springer.

Klein, P. N. 2013. *Coding the Matrix: Linear Algebra through Applications to Computer Science*. Providence, R.I.: Newtonian Press.

Kleinrock, L. 2009. *Queueing Systems: Computer Systems Modeling Fundamentals Volume 1*. New York: Wiley.

Klösgen, W. and J. M. Żytkow (eds.) 2002. *Handbook of Data Mining and Knowledge Discovery*. Oxford: Oxford University Press.

Kohonen, T. 2008. *Self-Organizing Maps* (third ed.). New York: Springer-Verlag.

Kolaczyk, E. D. 2009. *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.

Kolaczyk, E. D. and G. Csárdi 2014. *Statistical Analysis of Network Data with R*. New York: Springer.

Kolari, P. and A. Joshi 2004. Web mining research and practice. *IEEE Computing Science and Engineering* 6(4):49–53.

Koller, M. 2014. *Simulations for Sharpening Wald-type Inference in Robust Regression for Small Samples*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/robustbase/vignettes/lmrob_simulation.pdf. 288

Koller, M. and W. A. Stahel 2011. Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics and Data Analysis* 55(8):2504–2515. 288

Konik, M. 2006. *The Smart Money: How the World's Best Sports Bettors Beat the Bookies Out of Millions*. New York: Simon & Schuster.

Kotler, P. and K. L. Keller 2012. *Marketing Management* (fourteenth ed.). Upper Saddle River, N.J.: Prentice Hall. 26

Kotsiantis, S. and D. Kanellopoulos 2006. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering* 32(1):71–82. Retrieved from the World Wide Web at http://www.csis.pace.edu/~ctappert/dps/d861-13/session2-p1.pdf.

Krippendorff, K. H. 2012. *Content Analysis: An Introduction to Its Methodology* (third ed.). Thousand Oaks, Calif.: Sage. 148

Krishnamurthi, L. 2001. Pricing strategies and tactics. In D. Iacobucci (ed.), *Kellogg on Marketing*, Chapter 12, pp. 279–301. Wiley.

Kuhfeld, W. F., R. D. Tobias, and M. Garratt 1994. Efficient experimental design with marketing research applications. *Journal of Marketing Research* 31:545–557.

Kuhn, M. 2014. *caret: Classification and Regression Training*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/caret/caret.pdf.

Kuhn, M. and K. Johnson 2013. *Applied Predictive Modeling*. New York: Springer.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li 2004. *Applied Linear Statistical Models* (fifth ed.). Boston: McGraw-Hill.

Lam, W. and K.-Y. Lai 2001. A meta-learning approach for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 303–309. ACM Press. ISBN 1-58113-331-6.

Lamp, G. 2014. *ggplot for Python*. GitHub. 2014. https://github.com/yhat/ggplot. 11

Lander, J. P. 2014. *R for Everyone: Advanced Analytics and Graphics*. Upper Saddle River, N.J.: Pearson Education.

Langley, P. 1996. *Elements of Machine Learning*. San Francisco: Morgan Kaufmann.

Langtangen, H. P. 2009. *Python Scripting for Computational Science* (third ed.). New York: Springer.

Langtangen, H. P. 2012. *A Primer on Scientific Programming with Python* (third ed.). New York: Springer.

Langville, A. N. and C. D. Meyer 2006. *Google's Page Rank and Beyond: The Science of Search Engine Rankings*. Princeton, N.J.: Princeton University Press.

Langville, A. N. and C. D. Meyer 2012. *Who's 1?: The Science of Rating and Ranking*. Princeton, N.J.: Princeton University Press.

Lantz, B. 2013. *Machine Learning with R*. Birmingham, U.K.: Packt Publishing.

Larsen, B. and C. Aone 1999. Fast and effective text mining using linear-time document clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 16–22. ACM Press. ISBN 1-58113-143-7.

Laursen, G. H. N. and J. Thorlund 2010. *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting*. Hoboken, N.J.: Wiley. 278

Lawrence, S. and C. L. Giles 1998. Searching the World Wide Web. *Science* 280(3):98–100.

Le, C. T. 1997. *Applied Survival Analysis*. New York: Wiley.

Le, C. T. 1998. *Applied Categorical Data Analysis*. New York: Wiley.

Lebart, L. 1998. Visualizations of textual data. In J. Blasius and M. Greenacre (eds.), *Visualizing of Categorical Data*, Chapter 11, pp. 133–147. San Diego: Academic Press.

Ledoiter, J. 2013. *Data Mining and Business Analytics with R*. New York: Wiley.

Lee, C.-H. and H.-C. Yang 1999. A Web text mining approach based on self-organizing map. In *Proceedings of the second international workshop on Web information and data management*, pp. 59–62. ACM Press. ISBN 1-58113-221-2.

Leeflang, P. S. H., D. R. Wittink, M. Wedel, and P. A. Naert 2000. *Building Models for Marketing Decisions*. Boston: Kluwer.

Leetaru, K. 2011. *Data Mining Methods for Content Analysis: An Introduction to the Computational Analysis of Content*. New York: Routledge. 148

Leisch, F. and B. Gruen 2014. *CRAN Task View: Cluster Analysis & Finite Mixture Models*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/views/Cluster.html.

Lemke, R. J., M. Leonard, and K. Tlhokwane 2010. Estimating attendance at major league baseball games for the 2007 season. *Journal of Sports Economics* 11(3):316–348. 315

Leonard, A. 2013, February 1. How Netflix is turning viewers into puppets. *Salon*. Retrieved from the World Wide Web at: http://www.salon.com/2013/02/01/how_netflix_is_turning_viewers_into_puppets/.

Levy, P. S. and S. Lemeshow 2008. *Sampling of Populations: Methods and Applications* (fourth ed.). New York: Wiley.

Lewin-Koh, N. 2014. *Hexagon Binning: an Overview*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/hexbin/vignettes/hexagon_binning.pdf. 11

Lewis, M. 2003. *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton & Company. 204

Lewis, T. G. 2009. *Network Science: Theory and Applications*. New York: Wiley.

Liaw, A. and M. Wiener 2014. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

Ligges, U. 2005. *Programmieren mit R*. Berlin and Heidelberg: Springer.

Lilien, G. L., P. Kotler, and K. S. Moorthy 1992. *Marketing Models*. Englewood Cliffs, N.J.: Prentice-Hall.

Lilien, G. L. and A. Rangaswamy 2003. *Marketing Engineering: Computer-Assisted Marketing Analysis and Planning* (second ed.). Upper Saddle River, N.J.: Prentice Hall. 299

Lindsey, J. K. 1997. *Applying Generalized Linear Models*. New York: Springer.

Little, J. D. C. 1970. Models and managers: The concept of a decision calculus. *Management Science* 16(8):B466–B485.

Little, R. J. A. and D. B. Rubin 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Liu, B. 2010. Sentiment analysis and subjectivity. In N. Indurkhya and F. J. Damerau (eds.), *Handbook of Natural Language Processing* (second ed.)., pp. 627–665. Boca Raton, Fla.: Chapman and Hall/CRC.

Liu, B. 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. New York: Springer. 290

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. San Rafael, Calif.: Morgan & Claypool.

Lloyd, C. D. 2010. *Spatial Data Analysis: An Introduction for GIS Users*. Oxford, UK: Oxford University Press.

Lloyd, C. J. 1999. *Statistical Analysis of Categorical Data*. New York: Wiley.

Lord, F. M. and M. R. Novick 1968. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley. 111, 314

Louviere, J. J. 1993. The best-worst or maximum difference measurement model: Applications to behavioral research in marketing. Paper presented at the American Marketing Association Behavioral Research Conference, Phoenix.

Louviere, J. J., D. A. Hensher, and J. D. Swait 2000. *Stated Choice Methods: Analysis and Application*. Cambridge: Cambridge University Press.

Luce, D. and J. Tukey 1964. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology* 1:1–27.

Luger, G. F. 2008. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving* (sixth ed.). Boston: Addison-Wesley. 290

Lumley, T. 2004, June. Programmers' Niche: A simple class, in S3 and S4. *R News* 4(1): 33–36. http://CRAN.R-project.org/doc/Rnews/.

Lumley, T. 2010. *Complex Surveys: A Guide to Analysis Using R*. New York: Wiley.

Lumley, T. 2014. *mitools: Tools for Multiple Imputation of Missing Data*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/mitools/mitools.pdf.

Lyon, D. W. 2000. Pricing research. In C. Chakrapani (ed.), *Marketing Research: State-of-the-Art Perspectives*, Chapter 19, pp. 551–582. Chicago: American Marketing Association.

Lyon, D. W. 2002, Winter. The price is right (or is it)? *Marketing Research*:8–13.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts 2011, June. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Portland, Ore.: Association for Computational Linguistics. Retrieved from the World Wide Web at http://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf.

Maechler, M. 2014a. *Package cluster*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/cluster/cluster.pdf.

Maechler, M. 2014b. *robustbase: Basic Robust Statistics*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/robustbase/robustbase.pdf. 288

Maindonald, J. and J. Braun 2003. *Data Analysis and Graphics Using R: An Example-Based Approach*. Cambridge: Cambridge University Press.

Makridakis, S., S. C. Wheelwright, and R. J. Hyndman 2005. *Forecasting Methods and Applications* (third ed.). New York: Wiley.

Mallows, C. L. 1973. Some comments on C_p. *Technometrics* 15:661–675.

Mani, I. and M. T. Maybury (eds.) 1999. *Advances in Automatic Text Summarization*. Cambridge: MIT Press.

Manly, B. F. J. 1992. *The Design and Analysis of Research Studies*. Cambridge: Cambridge University Press.

Manly, B. F. J. 1994. *Multivariate Statistical Methods: A Primer* (second ed.). London: Chapman & Hall. 119

Manning, C. D. and H. Schütze 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marchi, M. and J. Albert 2014. *Analyzing Baseball Data with R*. Boca Raton, Fla.: Chapman & Hall/CRC. 207

Marder, E. 1997. *The Laws of Choice: Predicting Consumer Behavior*. New York: Free Press.

Maronna, R. A., D. R. Martin, and V. J. Yohai 2006. *Robust Statistics Theory and Methods*. New York: Wiley. 288

Marshall, P. and E. T. Bradlow 2002. A unified approach to conjoint analysis methods. *Journal of the American Statistical Association* 97(459):674–682.

Marsland, S. 2009. *Machine Learning: An Algorithmic Perspective*. Boca Raton, Fla.: Chapman & Hall/CRC.

Matloff, N. 2011. *The Art of R Programming*. San Francisco: no starch press.

Maybury, M. T. (ed.) 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park, Calif./ Cambridge: AAAI Press / MIT Press.

McCallum, Q. E. (ed.) 2013. *Bad Data Handbook*. Sebastopol, Calif.: O'Reilly.

McCullagh, P. and J. A. Nelder 1989. *Generalized Linear Models* (second ed.). New York: Chapman and Hall.

McFadden, D. 2001. Economic choices. *American Economic Review* 91:351–378.

McGrayne, S. B. 2011. *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines and Emerged Triumphant from Two Centuries of Controversy*. New Haven, Conn.: Yale University Press. 283

McIlroy, D., R. Brownrigg, T. P. Minka, and R. Bivand 2014. *mapproj: Map Projections*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/mapproj/mapproj.pdf.

McKinney, W. 2012. *Python for Data Analysis*. Sebastopol, Calif.: O'Reilly.

Meadow, C. T., B. R. Boyce, and D. H. Kraft 2000. *Text Information Retrieval Systems* (second ed.). San Diego: Academic Press.

Merkl, D. 2002. Text mining with self-organizing maps. In W. Klösgen and J. M. Żytkow (eds.), *Handbook of Data Mining and Knowledge Discovery*, Chapter 46.9, pp. 903–910. Oxford: Oxford University Press.

Meyer, D. 2014a. *Proximity Measures in the proxy Package for R*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/proxy/vignettes/overview.pdf.

Meyer, D. 2014b. *proxy: Distance and Similarity Measures*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/proxy/proxy.pdf.

Meyer, D. 2014c. *Support Vector Machines*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch 2014. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/e1071/e1071.pdf.

Meyer, D., A. Zeileis, and K. Hornik 2006, October 19. The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software* 17(3):1–48. http://www.jstatsoft.org/v17/i03.

Meyer, D., A. Zeileis, K. Hornik, and M. Friendly 2014a. *Residual-Based Shadings in vcd*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/vcd/vignettes/residual-shadings.pdf.

Meyer, D., A. Zeileis, K. Hornik, and M. Friendly 2014b. *vcd: Visualizing Categorical Data*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/vcd/vcd.pdf.

Michalawicz and Fogel 2004. *How to Solve It: Modern Heuristics*. New York: Springer. 290

Milborrow, S. 2014. *rpart.plot: Plot rpart models. An Enhanced Version of plot.rpart*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf.

Miller, G. A. 1995. Wordnet: A lexical database for English. *Communications of the ACM* 38 (11):39–41. 119

Miller, T. W. 2005. *Data and Text Mining: A Business Applications Approach*. Upper Saddle River, N.J.: Pearson Prentice Hall. http://www.pearsonhighered.com/educator/product/Data-and-Text-Mining-A-Business-Applications-Approach/9780131400856.page.

Miller, T. W. 2008a. *Research and Information Services: An Integrated Approach to Business*. Manhattan Beach, Calif.: Research Publishers LLC. http://research-publishers.com/rp/rais.htm.

Miller, T. W. 2008b. *Without a Tout: How to Pick a Winning Team*. Manhattan Beach, Calif.: Research Publishers LLC. http://research-publishers.com/rp/wat.html. 187, 204, 205, 206

Miller, T. W. 2015. *Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R* (revised and expanded ed.). Upper Saddle River, N.J.: Pearson Education. http://www.ftpress.com/miller.

Mitchell, M. 1996. *An Introduction to Genetic Algorithms*. Cambridge: MIT Press. 290

Mitchell, T. M. 1997. *Machine Learning*. New York: McGraw-Hill.

Moreno, J. L. 1934. Who shall survive?: Foundations of sociometry, group psychotherapy, and sociodrama. Reprinted in 1953 (second edition) and in 1978 (third edition) by Beacon House, Inc., Beacon, N.Y. 292

Moskowitz, T. J. and L. J. Wertheim 2011. *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. New York: Crown Archetype. 205

Mosteller, F. 1965. *Fifty Challenging Problems in Probability with Solutions*. Reading, Mass.: Addison-Wesley.

Mosteller, F. 1997. Lessons from sports statistics. *The American Statistician* 51(4):305–310.

Mosteller, F., R. E. K. Rourke, and G. B. Thomas, Jr. 1970. *Probability with Statistical Applications* (second ed.). Reading, Mass.: Addison-Wesley.

Mosteller, F. and J. W. Tukey 1977. *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley.

Mosteller, F. and D. L. Wallace 1984. *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers* (second ed.). New York: Springer. Earlier edition published in 1964 by Addison-Wesley, Reading, Mass. The previous title was *Inference and Disputed Authorship: The Federalist*.

Moustafa, R. and E. Wegman 2006. Multivariate continuous data—parallel coordinates. In A. Unwin, M. Theus, and H. Hoffman (eds.), *Graphics of Large Databases: Visualizing a Million*, Chapter 7, pp. 143–155. New York: Springer. 246

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, Mass.: MIT Press. 283, 290

Murrell, P. 2011. *R Graphics* (second ed.). Boca Raton, Fla.: CRC Press.

Nagle, T. T. and J. Hogan 2005. *The Strategy and Tactics of Pricing: A Guide to Growing More Profitably* (fourth ed.). Upper Saddle River, N.J.: Prentice Hall.

Nair, V. G. 2014. *Getting Started with Beautiful Soup: Build Your Own Web Scraper and Learn All About Web Scraping with Beautiful Soup*. Birmingham, UK: PACKT Publishing.

National Bureau of Economic Research 2010, September 20. Business cycle dating committee report. Available at http://www.nber.org/cycles/sept2010.html.

Neal, W. D. 2000. Market segmentation. In C. Chakrapani (ed.), *Marketing Research: State-of-the-Art Perspectives*, Chapter 1, pp. 375–399. American Marketing Association.

Nelson, W. B. 2003. *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*. Series on Statistics and Applied Probability. Philadelphia and Alexandria, Va.: ASA-SIAM.

Neuendorf, K. A. 2002. *The Content Analysis Guidebook*. Thousand Oaks, Calif.: Sage.

Newman, M. E. J. 2010. *Networks: An Introduction*. Oxford, UK: Oxford University Press.

Nielsen, A. 1979. Marketing research at the checkout. *Marketing Trends* 1:1–3.

Nunnally, J. C. 1967. *Psychometric Theory*. New York: McGraw-Hill. 111, 314

Nunnally, J. C. and I. H. Bernstein 1994. *Psychometric Theory* (third ed.). New York: McGraw-Hill. 314

O'Hagan, A. 2010. *Kendall's Advanced Theory of Statistics: Bayesian Inference*, Volume 2B. New York: Wiley. 275, 283

Oliver, D. 2004. *Basketball on Paper: Rules and Tools of Performance Analysis*. Dulles, Va.: Brassey's Press. 205

O'Neil, C. and R. Schutt 2014. *Doing Data Science: Straight Talk from the Frontline*. Sebastopol, Calif.: O'Reilly.

Orme, B. K. 2013. *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research* (third ed.). Glendale, Calif.: Research Publishers LLC. `http://research-publishers.com/rp/gsca.htm`. 296

Osborne, J. W. 2013. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Los Angeles: Sage.

Osgood, C. 1962. Studies in the generality of affective meaning systems. *American Psychologist* 17:10–28. 148

Osgood, C., G. Suci, and P. Tannenbaum (eds.) 1957. *The Measurement of Meaning*. Urbana, Ill.: University of Illinois Press. 148

Pace, R. K. and R. Barry 1997. Sparse spatial autoregressions. *Statistics and Probability Letters* 33:291–297.

Pang, B. and L. Lee 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2):1–135.

Pebesma, E. 2014. *CRAN Task View: Handling and Analyzing Spatio-Temporal Data*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/views/SpatioTemporal.html`.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Peppers, D. and M. Rogers 1993. *The One to One Future: Building Relationships One Customer at a Time*. New York: Doubleday. 298

Peta, J. 2013. *Trading Bases: How a Wall Street Trader Made a Fortune Betting on Baseball*. New York: Penguin. 206

Petris, G. 2010, October 13. An R package for dynamic linear models. *Journal of Statistical Software* 36(12):1–16. `http://www.jstatsoft.org/v36/i12`.

Petris, G. and W. Gilks 2014. *dlm: Bayesian and Likelihood Analysis of Dynamic Linear Models*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/dlm/dlm.pdf.

Petris, G., S. Petrone, and P. Campagnoli 2009. *Dynamic Linear Models with R*. New York: Springer.

Pfaff, B. 2013. *Financial Risk Modeling and Portfolio Optimization with R*. New York: Wiley.

Piatetsky-Shapiro, G. and W. Frawley (eds.) 1991. *Knowledge Discovery in Databases*. Menlo Park, Calif.: AAAI Press.

Pindyck, R. and D. Rubinfeld 2012. *Microeconomics* (eighth ed.). Upper Saddle River, N.J.: Pearson. 253

Pinheiro, J. C. and D. M. Bates 2009. *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag. 275

Pinker, S. 1994. *The Language Instinct*. New York: W. Morrow and Co.

Pinker, S. 1997. *How the Mind Works*. New York: W.W. Norton & Company.

Pinker, S. 1999. *Words and Rules: The Ingredients of Language*. New York: HarperCollins.

Pollard, R. 1973, June. Collegiate football scores and the negative binomial distribution. *Journal of the American Statistical Association* 68(342):351–352. 199

Popping, R. 2000. *Computer-Assisted Text Analysis*. Thousand Oaks, Calif.: Sage. 148

Potts, C. 2011. On the negativity of negation. In *Proceedings of Semantics and Linguistic Theory 20*, pp. 636–659. CLC Publications. Retrieved from the World Wide Web at http://elanguage.net/journals/salt/article/view/20.636/1414.

Press, S. J. 2004. The role of Bayesian and frequentist multivariate modeling in statistical data mining. In H. Bozdogan (ed.), *Statistical Data Mining and Knowledge Discovery*, Chapter 1, pp. 1–14. Boca Raton, Fla.: CRC Press.

Provost, F. and T. Fawcett 2014. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. Sebastopol, Calif.: O'Reilly. 278

Pustejovsky, J. and A. Stubbs 2013. *Natural Language Annotation for Machine Learning*. Sebastopol, Calif.: O'Reilly. 314

Putler, D. S. and R. E. Krider 2012. *Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R*. Boca Raton, Fla: Chapman & Hall/CRC.

Pyle, D. 1999. *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, Calif.: Morgan Kaufmann.

Radcliffe-Brown, A. R. 1940. On social structure. *Journal of the Royal Anthropological Society of Great Britain and Ireland* 70:1–12. 292

Radev, D., W. Fan, H. Qi, H. Wu, and A. Grewal 2002. Probabilistic question answering on the Web. In *Proceedings of the Eleventh International Conference on World Wide Web*, pp. 408–419. Washington, D.C.: ACM Press. ISBN 1-58113-449-5.

Rajaraman, A. and J. D. Ullman 2012. *Mining of Massive Datasets*. Cambridge, UK: Cambridge University Press. 55

Rastogi, R. and K. Shim 1999. Scalable algorithms for mining large databases. In *Tutorial notes of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 73–140. Washington, D.C.: ACM Press. ISBN 1-58113-171-2.

Ratner, B. 2011. *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data* (second ed.). Boca Raton, Fla.: CRC Press.

Reed, R. D. and R. J. Marks, II 1999. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. Cambridge: MIT Press.

Reep, C., R. Pollard, and B. Benjamin 1971. Skill and chance in ball games. *Journal of the Royal Statistical Society, Series A (General)* 134(4):623–629. 199

Reis, A. and J. Trout 2001. *Positioning: The Battle for Your Mind*. New York: McGraw-Hill.

Rencher, A. C. and G. B. Schaalje 2008. *Linear Models in Statistics* (second ed.). New York: Wiley.

Revelle, W. 2014. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/packages/psych/psych.pdf`.

Ricci, F., L. Rokach, B. Shapira, and P. B. Kantor (eds.) 2011. *Recommender Systems Handbook*. New York: Springer.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Robert, C. P. 2007. *The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation* (second ed.). New York: Springer. 275, 283

Robert, C. P. and G. Casella 2009. *Introducing Monte Carlo Methods with R*. New York: Springer. 283

Roberts, C. W. (ed.) 1997. *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, N.J.: Lawrence Erlbaum Associates. 148

Robinson, I., J. Webber, and E. Eifrem 2013. *Graph Databases*. Sebastopol, Calif.: O'Reilly.

Rogers, H. . J., H. Swaminathan, and R. K. Hambleton 1991. *Fundamentals of Item Response Theory*. Newbury Park, Calif.: Sage.

Romer, D. 2006. Do firms maximize? Evidence from professional football. *Journal of Political Economy* 114(2):340–365. 205

Rosen, J. 2001. *The Unwanted Gaze: The Destruction of Privacy in America*. New York: Vintage. 292

Rosenbaum, P. R. 1995. *Observational Studies*. New York: Springer.

Ross, S. M. 2006. *Introduction to Probability Models* (tenth ed.). New York: Academic Press.

Rossi, P. 2014. *bayesm: Bayesian Inference for Marketing/Micro-econometrics*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/packages/bayesm/bayesm.pdf`.

Rossi, P. E. and G. M. Allenby 2003. Bayesian statistics and marketing. *Marketing Science* 22(3):304–328.

Rossi, P. E., G. M. Allenby, and R. McCulloch 2005. *Bayesian Statistics and Marketing*. New York: Wiley.

Rounds, J. B., T. W. Miller, and R. V. Dawis 1978. Comparability of multiple rank order and paired comparison methods. *Applied Psychological Measurement* 2(3):415–422.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Ruppert, D. 2011. *Statistics and Data Analysis for Financial Engineering*. New York: Springer.

Ruppert, D. 2014. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Sebastopol, Calif.: O'Reilly.

Russell, S. and P. Norvig 2009. *Artificial Intelligence: A Modern Approach* (third ed.). Upper Saddle River, N.J.: Prentice Hall. 290

Ryan, J. A. 2014. *quantmod: Quantitative Financial Modelling Framework*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/quantmod/quantmod.pdf.

Ryan, T. P. 2008. *Modern Regression Methods* (second ed.). New York: Wiley. 143

Salsburg, D. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: Henry Holt and Company.

Sarkar, D. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Sarkar, D. 2014. *lattice: Lattice Graphics*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/lattice/lattice.pdf.

Sarkar, D. and F. Andrews 2014. *latticeExtra: Extra Graphical Utilities Based on Lattice*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/latticeExtra/latticeExtra.pdf.

Sauer, R. D. 1998, December. The economics of wagering markets. *Journal of Economic Literature* 36:2021–2064.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. 2000. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schauerhuber, M., A. Zeileis, D. Meyer, and K. Hornik 2008. Benchmarking open-source tree learners in R/RWeka. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (eds.), *Data Analysis, Machine Learning, and Applications*, pp. 389–396. New York: Springer.

Schrott, P. R. and D. J. Lanoue 1994. Trends and perspectives in content analysis. In I. Borg and P. Mohler (eds.), *Trends and Perspectives in Empirical Social Research*, pp. 327–345. Berlin: Walter de Gruyter.

Schwarz, A. 2004. *The Numbers Game: Baseball's Lifelong Fascination with Statistics*. New York: St. Martin's Griffin. 205

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6:461–464.

Schwertman, N. C., T. A. McCready, and L. Howard 1991, February. Probability models for the NCAA regional basketball tournaments. *The American Statistician* 45(1):35–38. 205

Schwertman, N. C., K. L. Schenk, and B. C. Holbrook 1996, February. More probability models for the NCAA regional basketball tournaments. *The American Statistician* 50 (1):34–38. 205

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.

Seber, G. A. F. 2000. *Multivariate Observations*. New York: Wiley. Originally published in 1984. 119

Segaran, T. 2007. *Collective Intelligence: Building Smart Web 2.0 Applications*. Sebastopol, Calif.: O'Reilly.

Sermas, R. 2014. *ChoiceModelR: Choice Modeling in R*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/ChoiceModelR/ChoiceModelR.pdf.

Shapiro, C. and H. R. Varian 1999. *Information Rules: A Strategic Guide to the New Economy*. Boston: Harvard Business School Press.

Sharda, R. and D. Delen 2006. Predicting box office success of motion pictures with neural networks. *Expert Systems with Applications* 30:243–254. 150

Sharma, S. 1996. *Applied Multivariate Techniques*. New York: Wiley. 119

Shmueli, G. 2010. To explain or predict? *Statistical Science* 25(3):289–310.

Shrout, P. E. and S. T. Fiske (eds.) 1995. *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Shumway, R. H. and D. S. Stoffer 2011. *Time Series Analysis and Its Applications with R* (third ed.). New York: Springer.

Siegel, E. 2013. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, N.J.: Wiley. 278

Silver, N. 2012. *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*. New York: The Penguin Press. 204

Simon, H. A. 2002. Foreward: Enhancing the intelligence of discovery systems. In W. Klösgen and J. M. Żytkow (eds.), *Handbook of Data Mining and Knowledge Discovery*, p. xvii. Oxford: Oxford University Press.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer 2005. ROCR: Visualizing classifier performance in R. *Bioinformatics* 21(20):3940–3941.

Snedecor, G. W. and W. G. Cochran 1989. *Statistical Methods* (eighth ed.). Ames, Iowa: Iowa State University Press. First edition published by Snedecor in 1937. 275, 283

Socher, R., J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Spector, P. 2008. *Data Management with R*. New York: Springer.

Srivastava, A. N. and M. Sahami (eds.) 2009. *Text Mining: Classification, Clustering, and Applications*. Boca Raton, Fla.: CRC Press.

Stahel, W. and S. Weisberg (eds.) 1991. *Directions in Robust Statistics and Diagnostics*, Volume 34 of *IMA Volumes in Mathematics and Its Applications*. New York: Springer-Verlag.

Stern, H. S. 1991, December. On the probability of winning a football game. *The American Statistician* 45(3):179–183. 205

Sternthal, B. and A. M. Tybout 2001. Segmentation and targeting. In D. Iacobucci (ed.), *Kellogg on Marketing*, Chapter 1, pp. 3–30. New York: Wiley.

Stevens, S. S. 1946, June 7. On the theory of scales of measurement. *Science* 103(2684): 677–680. http://www.jstor.org/stable/1671815. 314

Stigler, G. J. 1987. *The Theory of Price* (fourth ed.). New York: Macmillan. 253

Stone, P. J. 1997. Thematic text analysis: New agendas for analyzing text content. In C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*, Chapter 2, pp. 35–54. Mahwah, N.J.: Lawrence Erlbaum Associates.

Stone, P. J., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press.

Stuart, A., K. Ord, and S. Arnold 2010. *Kendall's Advanced Theory of Statistics: Classical Inference and the Linear Model*, Volume 2A. New York: Wiley. 275, 283

Suess, E. A. and B. E. Trumbo 2010. *Introduction to Probability Simulation and Gibbs Sampling with R*. New York: Springer. 283

Sullivan, D. 2001. *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*. New York: Wiley.

Szymanski, C. 2014. *dlmodeler: Generalized Dynamic Linear Modeler*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/dlmodeler/dlmodeler.pdf.

Taddy, M. 2013a. Measuring political sentiment on Twitter: factor-optimal design for multinomial inverse regression. Retrieved from the World Wide Web at http://arxiv.org/pdf/1206.3776v5.pdf.

Taddy, M. 2013b. Multinomial inverse regression for text analysis. Retrieved from the World Wide Web at http://arxiv.org/pdf/1012.2098v6.pdf.

Taddy, M. 2014. *textir: Inverse Regression for Text Analysis*. 2014. http://cran.r-project.org/web/packages/textir/textir.pdf.

Tan, P.-N., M. Steinbach, and V. Kumar 2006. *Introduction to Data Mining*. Boston: Addison-Wesley. 55

Tang, W., H. He, and X. M. Tu 2012. *Applied Categorical and Count Data Analysis*. Boca Raton, Fla.: Chapman & Hall/CRC.

Tango, T. M., M. G. Lichtman, and A. E. Dolphin 2007. *The Book: Playing the Percentages in Baseball*. Dulles, Va.: Potomac Books. 205

Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (third ed.). New York: Springer. 283

Tayman, J. and L. Pol 1995, Spring. Retail site selection and geographical information systems. *Journal of Applied Business Research* 11(2):46–54. 299

Therneau, T. 2014. *survival: Survival Analysis*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/survival/survival.pdf. 254

Therneau, T., B. Atkinson, and B. Ripley 2014. *rpart: Recursive Partitioning*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/rpart/rpart.pdf.

Therneau, T. and C. Crowson 2014. *Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf.

Therneau, T. M. and P. M. Grambsch 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Thompson, M. 1975. On any given Sunday: Fair competitor orderings with maximum likelihood methods. *Journal of the American Statistical Association* 70(351):536–541. 205

Thorn, J. and P. Palmer 1985. *The Hidden Game of Baseball: A Revolutionary Approach to Baseball and Its Statistics* (revised and updated ed.). New York: Doubleday. 205

Thurman, W. N. and M. E. Fisher 1988. Chickens, eggs, and causality, or which came first? *American Journal of Agricultural Economics* 70(2):237–238. 69

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological Review* 34:273–286.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288. 288

Tong, S. and D. Koller 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 2:45–66.

Torgerson, W. S. 1958. *Theory and Methods of Scaling*. New York: Wiley. 314

Train, K. and Y. Croissant 2014. *Kenneth Train's exercises using the mlogit Package for R*. Comprehensive R Archive Network. 2014. `http://cran.r-project.org/web/packages/mlogit/vignettes/Exercises.pdf`. 254

Train, K. E. 1985. *Qualitative Choice Analysis*. Cambridge, Mass.: MIT Press.

Train, K. E. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Trybula, W. J. 1999. Text mining. In M. E. Williams (ed.), *Annual Review of Information Science and Technology*, Volume 34, Chapter 7, pp. 385–420. Medford, N.J.: Information Today, Inc.

Tsay, R. S. 2013. *An Introduction to Analysis of Financial Data with R*. New York: Wiley.

Tsvetovat, M. and A. Kouznetsov 2011. *Social Network Analysis for Startups: Finding Connections on the Social Web*. Sebastopol, Calif.: O'Reilly.

Tufte, E. R. 1990. *Envisioning Information*. Cheshire, Conn.: Graphics Press.

Tufte, E. R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn.: Graphics Press.

Tufte, E. R. 2004. *The Visual Display of Quantitative Information* (second ed.). Cheshire, Conn.: Graphics Press.

Tufte, E. R. 2006. *Beautiful Evidence*. Cheshire, Conn.: Graphics Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.

Tukey, J. W. and F. Mosteller 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, Mass.: Addison-Wesley.

Turney, P. D. 2002, July 8–10. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*:417–424. Available from the National Research Council Canada publications archive. 150

Turow, J. 2013. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven, Conn.: Yale University Press. 292

Tybout, A. M. and B. Sternthal 2001. Brand positioning. In D. Iacobucci (ed.), *Kellogg on Marketing*, Chapter 2, pp. 31–57. New York: Wiley.

Unwin, A., M. Theus, and H. Hofmann (eds.) 2006. *Graphics of Large Datasets: Visualizing a Million*. New York: Springer. 11

Vapnik, V. N. 1998. *Statistical Learning Theory*. New York: Wiley. 144

Vapnik, V. N. 2000. *The Nature of Statistical Learning Theory* (second ed.). New York: Springer. 144

Varian, H. R. 2005. *Intermediate Microeconomics: A Modern Approach* (seventh ed.). New York: Norton. 253

Velleman, P. F. and L. Wilkinson 1993, February. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician* 47(1):65–72.

Venables, W. N. and B. D. Ripley 2000. *S Programming*. New York: Springer-Verlag.

Venables, W. N. and B. D. Ripley 2002. *Modern Applied Statistics with S* (fourth ed.). New York: Springer-Verlag. Champions of S, S-Plus, and R call this *the mustard book*.

Venables, W. N., D. M. Smith, and R Development Core Team 2001. *An Introduction to R*. Bristol, UK: Network Theory Limited.

Wainer, H. 1997. *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. New York: Springer-Verlag.

Walker, S. 2007, January 5. The man who shook up Vegas. *The Wall Street Journal*:W1, W10. Available from the World Wide Web at `http://online.wsj.com/news/articles/SB116796079037267731`. 187

Wasserman, L. 2010. *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer. 275, 283

Wasserman, S. and K. Faust 1994. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Wasserman, S. and D. Iacobucci 1986. Statistical analysis of discrete relational data. *British Journal of Mathematical and Statistical Psychology* 39:41–64.

Wassertheil-Smoller, S. 1990. *Biostatistics and Epidemiology: A Primer for Health Professionals*. New York: Springer.

Watts, D. J. 2003. *Six Degrees: The Science of a Connected Age*. New York: W.W. Norton.

Wedel, M. and W. Kamakura 2000. *Market Segmentation: Conceptual and Methodological Foundations* (second ed.). Boston: Kluwer.

Wegman, E. J. 1990. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association* 85:664–675. 246

Wei, Y. 2013. Colors in R. Retrieved from the World Wide Web at `http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf`.

Weisberg, S. 2005. *Applied Linear Regression* (third ed.). New York: Wiley.

Weiss, S. M., N. Indurkhya, and T. Zhang 2010. *Fundamentals of Predictive Text Mining*. New York: Springer.

West, B. T. 2006. A simple and flexible rating method for predicting success in the NCAA basketball tournament. *Journal of Quantitative Analysis in Sports* 2(3):1–14. 205

West, M. D. (ed.) 2001. *Theory, Method, and Practice in Computer Content Analysis*. Westport, Conn.: Ablex. 148

White, T. 2011. *Hadoop: The Definitive Guide* (second ed.). Sebastopol, Calif.: O'Reilly.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, H. 2010. stringr: Modern, consistent string processing. *The R Journal* 2(2):38–40.

Wickham, H. 2011, April 7. The split-apply-combine strategy for data analysis. *Journal of Statistical Software* 40(1):1–29. http://www.jstatsoft.org/v40/i01.

Wickham, H. 2014a. *plyr: Tools for Splitting, Applying and Combining Data*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/plyr/plyr.pdf.

Wickham, H. 2014b. *stringr: Make It Easier to Work with Strings*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/stringr/stringr.pdf. 119

Wickham, H. and W. Chang 2014. *ggplot2: An Implementation of the Grammar of Graphics*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. 11

Wilkinson, L. 2005. *The Grammar of Graphics* (second ed.). New York: Springer.

Williams, H. P. 2013. *Model Building in Mathematical Programming* (fifth ed.). New York: Wiley.

Winer, B. J., D. R. Brown, and K. M. Michels 1991. *Statistical Principles in Experimental Design* (third ed.). New York: McGraw-Hill.

Winston, W. L. 2009. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football*. Princeton, N.J.: Princeton University Press. 205

Witten, I. H., E. Frank, and M. A. Hall 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, Mass.: Morgan Kaufmann. 55, 289

Witten, I. H., A. Moffat, and T. C. Bell 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images* (second ed.). San Francisco: Morgan Kaufmann.

Wrigley, N. (ed.) 1988. *Store Choice, Store Location, and Market Analysis*. London, UK: Routledge. 299

Yao, K. 2007. *Weighing the Odds in Sports Betting*. Las Vegas, Nev.: Pi Yee Press. 206

Yates, F. 1980. *Sampling Methods for Censuses and Surveys* (fourth ed.). New York: Macmillan. First edition published by Griffin in London in 1949.

Yau, N. 2011. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. New York: Wiley.

Yau, N. 2013. *Data Points: Visualization That Means Something*. New York: Wiley.

Ye, N. (ed.) 2003. *The Handbook of Data Mining*. Mahwah, N.J.: Lawrence Erlbaum.

Youmans, G. 1990. Measuring lexical style and competence: The type-token vocabulary curve. *Style* 24(4):584–599. 118

Youmans, G. 1991. A new tool for discourse analysis: The vocabulary management profile. *Language* 67(4):763–789. 118

ZaÔane, O. R. and M.-L. Antonie 2002. Classifying text documents by associating terms with text categories. In *Proceedings of the Thirteenth Australasian Conference on Database Technologies*, pp. 215–222. Sydney, Australia: Australian Computer Society, Inc. ISBN 0-909925-83-6.

Zeileis, A., K. Hornik, and P. Murrell 2009, July. Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics and Data Analysis* 53(9):3259–3270.

Zeileis, A., K. Hornik, and P. Murrell 2014. *HCL-Based Color Palettes in R*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/colorspace/vignettes/hcl-colors.pdf.

Zeileis, A., T. Hothorn, and K. Hornik 2014. *party with the mob: Model-Based Recursive Partitioning in R*. Comprehensive R Archive Network. 2014. http://cran.r-project.org/web/packages/party/vignettes/MOB.pdf.

Zhang, H. and B. Singer 1999. *Recursive Partitioning in the Health Sciences*. New York: Springer-Verlag.

Zhao, Y. 2013. *R and Data Mining: Examples and Case Studies*. San Diego: Academic Press/Elsevier.

Zhong, N., J. Liu, and Y. Yao (eds.) 2003. *Web Intelligence*. New York: Springer-Verlag.

Zipf, H. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley.

Zivot, E. and J. Wang 2003. *Modeling Financial Time Series with S-PLUS*. Seattle: Insightful Corporation.

Zumel, N. and J. Mount 2014. *Practical Data Science with R*. Shelter Island, N.Y.: Manning.

Zwerina, K. 1997. *Discrete Choice Experiments in Marketing: Use of Priors in Efficient Choice Designs and Their Application to Individual Preference Measurement*. New York: Physica-Verlag.