

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	What Is Data Science?	1
1.2	Diabetes in America	3
1.3	Authors of the Federalist Papers	5
1.4	Forecasting NASDAQ Stock Prices	6
1.5	Remarks	8
1.6	The Book	8
1.7	Algorithms	11
1.8	Python	12
1.9	R	13
1.10	Terminology and Notation	14
	1.10.1 Matrices and Vectors	14
1.11	Book Website	16

## Part I Data Reduction

<b>2</b>	<b>Data Mapping and Data Dictionaries</b>	19
2.1	Data Reduction	19
2.2	Political Contributions	20
2.3	Dictionaries	22
2.4	Tutorial: Big Contributors	22
2.5	Data Reduction	27
	2.5.1 Notation and Terminology	28
	2.5.2 The Political Contributions Example	29
	2.5.3 Mappings	30
2.6	Tutorial: Election Cycle Contributions	31
2.7	Similarity Measures	38
	2.7.1 Computation	41
2.8	Tutorial: Computing Similarity	43
2.9	Concluding Remarks About Dictionaries	47
2.10	Exercises	48

2.10.1	Conceptual .....	48
2.10.2	Computational .....	49
<b>3</b>	<b>Scalable Algorithms and Associative Statistics .....</b>	<b>51</b>
3.1	Introduction .....	51
3.2	Example: Obesity in the United States .....	53
3.3	Associative Statistics .....	54
3.4	Univariate Observations .....	55
3.4.1	Histograms .....	57
3.4.2	Histogram Construction .....	58
3.5	Functions .....	60
3.6	Tutorial: Histogram Construction .....	61
3.6.1	Synopsis .....	74
3.7	Multivariate Data .....	74
3.7.1	Notation and Terminology .....	75
3.7.2	Estimators .....	76
3.7.3	The Augmented Moment Matrix .....	79
3.7.4	Synopsis .....	80
3.8	Tutorial: Computing the Correlation Matrix .....	80
3.8.1	Conclusion .....	87
3.9	Introduction to Linear Regression .....	88
3.9.1	The Linear Regression Model .....	89
3.9.2	The Estimator of $\beta$ .....	90
3.9.3	Accuracy Assessment .....	93
3.9.4	Computing $R^2_{\text{adjusted}}$ .....	94
3.10	Tutorial: Computing $\hat{\beta}$ .....	95
3.10.1	Conclusion .....	101
3.11	Exercises .....	102
3.11.1	Conceptual .....	102
3.11.2	Computational .....	103
<b>4</b>	<b>Hadoop and MapReduce .....</b>	<b>105</b>
4.1	Introduction .....	105
4.2	The Hadoop Ecosystem .....	106
4.2.1	The Hadoop Distributed File System .....	106
4.2.2	MapReduce .....	108
4.2.3	Mapping .....	108
4.2.4	Reduction .....	110
4.3	Developing a Hadoop Application .....	111
4.4	Medicare Payments .....	111
4.5	The Command Line Environment .....	113
4.6	Tutorial: Programming a MapReduce Algorithm .....	113
4.6.1	The Mapper .....	116
4.6.2	The Reducer .....	120
4.6.3	Synopsis .....	123

- 4.7 Tutorial: Using Amazon Web Services ..... 124
  - 4.7.1 Closing Remarks ..... 128
- 4.8 Exercises ..... 128
  - 4.8.1 Conceptual ..... 128
  - 4.8.2 Computational ..... 128

**Part II Extracting Information from Data**

- 5 Data Visualization** ..... 133
  - 5.1 Introduction ..... 133
  - 5.2 Principles of Data Visualization ..... 135
  - 5.3 Making Good Choices ..... 138
    - 5.3.1 Univariate Data ..... 139
    - 5.3.2 Bivariate and Multivariate Data ..... 142
  - 5.4 Harnessing the Machine ..... 148
    - 5.4.1 Building Fig. 5.2 ..... 151
    - 5.4.2 Building Fig. 5.3 ..... 152
    - 5.4.3 Building Fig. 5.4 ..... 153
    - 5.4.4 Building Fig. 5.5 ..... 154
    - 5.4.5 Building Fig. 5.8 ..... 155
    - 5.4.6 Building Fig. 5.10 ..... 156
    - 5.4.7 Building Fig. 5.11 ..... 157
  - 5.5 Exercises ..... 158
- 6 Linear Regression Methods** ..... 161
  - 6.1 Introduction ..... 161
  - 6.2 The Linear Regression Model ..... 162
    - 6.2.1 Example: Depression, Fatalism, and Simplicity ..... 164
    - 6.2.2 Least Squares ..... 166
    - 6.2.3 Confidence Intervals ..... 168
    - 6.2.4 Distributional Conditions ..... 170
    - 6.2.5 Hypothesis Testing ..... 171
    - 6.2.6 Cautionary Remarks ..... 175
  - 6.3 Introduction to R ..... 176
  - 6.4 Tutorial: R ..... 177
    - 6.4.1 Remark ..... 181
  - 6.5 Tutorial: Large Data Sets and R ..... 181
  - 6.6 Factors ..... 187
    - 6.6.1 Interaction ..... 189
    - 6.6.2 The Extra Sums-of-Squares  $F$ -test ..... 192
  - 6.7 Tutorial: Bike Share ..... 195
    - 6.7.1 An Incongruous Result ..... 200
  - 6.8 Analysis of Residuals ..... 200
    - 6.8.1 Linearity ..... 201

- 6.8.2 Example: The Bike Share Problem . . . . . 202
- 6.8.3 Independence . . . . . 204
- 6.9 Tutorial: Residual Analysis . . . . . 208
  - 6.9.1 Final Remarks . . . . . 210
- 6.10 Exercises . . . . . 211
  - 6.10.1 Conceptual . . . . . 211
  - 6.10.2 Computational . . . . . 212
- 7 Healthcare Analytics . . . . . 217**
  - 7.1 Introduction . . . . . 217
  - 7.2 The Behavioral Risk Factor Surveillance System . . . . . 219
    - 7.2.1 Estimation of Prevalence . . . . . 220
    - 7.2.2 Estimation of Incidence . . . . . 221
  - 7.3 Tutorial: Diabetes Prevalence and Incidence . . . . . 222
  - 7.4 Predicting At-Risk Individuals . . . . . 231
    - 7.4.1 Sensitivity and Specificity . . . . . 234
  - 7.5 Tutorial: Identifying At-Risk Individuals . . . . . 236
  - 7.6 Unusual Demographic Attribute Vectors . . . . . 243
  - 7.7 Tutorial: Building Neighborhood Sets . . . . . 245
    - 7.7.1 Synopsis . . . . . 247
  - 7.8 Exercises . . . . . 249
    - 7.8.1 Conceptual . . . . . 249
    - 7.8.2 Computational . . . . . 250
- 8 Cluster Analysis . . . . . 253**
  - 8.1 Introduction . . . . . 253
  - 8.2 Hierarchical Agglomerative Clustering . . . . . 254
  - 8.3 Comparison of States . . . . . 255
  - 8.4 Tutorial: Hierarchical Clustering of States . . . . . 258
    - 8.4.1 Synopsis . . . . . 264
  - 8.5 The  $k$ -Means Algorithm . . . . . 266
  - 8.6 Tutorial: The  $k$ -Means Algorithm . . . . . 268
    - 8.6.1 Synopsis . . . . . 273
  - 8.7 Exercises . . . . . 274
    - 8.7.1 Conceptual . . . . . 274
    - 8.7.2 Computational . . . . . 274

**Part III Predictive Analytics**

- 9  $k$ -Nearest Neighbor Prediction Functions . . . . . 279**
  - 9.1 Introduction . . . . . 279
    - 9.1.1 The Prediction Task . . . . . 280
  - 9.2 Notation and Terminology . . . . . 282
  - 9.3 Distance Metrics . . . . . 283
  - 9.4 The  $k$ -Nearest Neighbor Prediction Function . . . . . 284

- 9.5 Exponentially Weighted  $k$ -Nearest Neighbors . . . . . 286
- 9.6 Tutorial: Digit Recognition . . . . . 287
  - 9.6.1 Remarks . . . . . 294
- 9.7 Accuracy Assessment . . . . . 295
  - 9.7.1 Confusion Matrices . . . . . 297
- 9.8  $k$ -Nearest Neighbor Regression . . . . . 298
- 9.9 Forecasting the S&P 500 . . . . . 299
- 9.10 Tutorial: Forecasting by Pattern Recognition . . . . . 300
  - 9.10.1 Remark . . . . . 307
- 9.11 Cross-Validation . . . . . 308
- 9.12 Exercises . . . . . 310
  - 9.12.1 Conceptual . . . . . 310
  - 9.12.2 Computational . . . . . 310
- 10 The Multinomial Naïve Bayes Prediction Function . . . . . 313**
  - 10.1 Introduction . . . . . 313
  - 10.2 The Federalist Papers . . . . . 314
  - 10.3 The Multinomial Naïve Bayes Prediction Function . . . . . 315
    - 10.3.1 Posterior Probabilities . . . . . 317
  - 10.4 Tutorial: Reducing the Federalist Papers . . . . . 319
    - 10.4.1 Summary . . . . . 325
  - 10.5 Tutorial: Predicting Authorship of the Disputed Federalist Papers . . . . . 325
    - 10.5.1 Remark . . . . . 329
  - 10.6 Tutorial: Customer Segmentation . . . . . 329
    - 10.6.1 Additive Smoothing . . . . . 330
    - 10.6.2 The Data . . . . . 332
    - 10.6.3 Remarks . . . . . 337
  - 10.7 Exercises . . . . . 338
    - 10.7.1 Conceptual . . . . . 338
    - 10.7.2 Computational . . . . . 339
- 11 Forecasting . . . . . 343**
  - 11.1 Introduction . . . . . 343
  - 11.2 Tutorial: Working with Time . . . . . 345
  - 11.3 Analytical Methods . . . . . 350
    - 11.3.1 Notation . . . . . 350
    - 11.3.2 Estimation of the Mean and Variance . . . . . 350
    - 11.3.3 Exponential Forecasting . . . . . 352
    - 11.3.4 Autocorrelation . . . . . 353
  - 11.4 Tutorial: Computing  $\hat{\rho}_\tau$  . . . . . 354
    - 11.4.1 Remarks . . . . . 359
  - 11.5 Drift and Forecasting . . . . . 359
  - 11.6 Holt-Winters Exponential Forecasting . . . . . 360
    - 11.6.1 Forecasting Error . . . . . 362

- 11.7 Tutorial: Holt-Winters Forecasting . . . . . 363
- 11.8 Regression-Based Forecasting of Stock Prices . . . . . 367
- 11.9 Tutorial: Regression-Based Forecasting . . . . . 368
  - 11.9.1 Remarks . . . . . 373
- 11.10 Time-Varying Regression Estimators . . . . . 374
- 11.11 Tutorial: Time-Varying Regression Estimators . . . . . 375
  - 11.11.1 Remarks . . . . . 377
- 11.12 Exercises . . . . . 377
  - 11.12.1 Conceptual . . . . . 377
  - 11.12.2 Computational . . . . . 378
- 12 Real-time Analytics . . . . . 381**
  - 12.1 Introduction . . . . . 381
  - 12.2 Forecasting with a NASDAQ Quotation Stream . . . . . 382
    - 12.2.1 Forecasting Algorithms . . . . . 383
  - 12.3 Tutorial: Forecasting the Apple Inc. Stream . . . . . 384
    - 12.3.1 Remarks . . . . . 389
  - 12.4 The Twitter Streaming API . . . . . 390
  - 12.5 Tutorial: Tapping the Twitter Stream . . . . . 391
    - 12.5.1 Remarks . . . . . 395
  - 12.6 Sentiment Analysis . . . . . 396
  - 12.7 Tutorial: Sentiment Analysis of Hashtag Groups . . . . . 398
  - 12.8 Exercises . . . . . 400
- A Solutions to Exercises . . . . . 403**
- B Accessing the Twitter API . . . . . 417**
- References . . . . . 419**
- Index . . . . . 423**

# List of Figures

1.1	Relative frequency of occurrence of the most common 20 words in Hamilton's undisputed papers . . . . .	6
1.2	Observed prices (points) and time-varying linear regression forecasts of Apple, Inc . . . . .	7
2.1	Donation totals reported to the Federal Election Commission by Congressional candidates and Political Action Committees plotted against reporting date . . . . .	21
2.2	Contributions to committees by individual contributors aggregated by employer . . . . .	30
3.1	Histograms of body mass index constructed from two samples of U.S. residents . . . . .	58
4.1	Flow chart showing the transfer of data from the NameNode to the DataNodes in the Hadoop ecosystem . . . . .	107
4.2	The distribution of average medicare payments for 5 three-digit zip codes . . . . .	112
5.1	A pie chart makes patterns in the data difficult to decode; the dotchart is an improvement . . . . .	136
5.2	Two views of monthly sales for four departments . . . . .	137
5.3	Three different ways of looking at monthly sales by department in the grocery store data . . . . .	139
5.4	Two different ways of visualizing the distribution of monthly sales numbers, the boxplot and the violin plot . . . . .	141
5.5	A dotchart of spend by month by department, with bars indicating the range of the data . . . . .	142
5.6	A mosaic plot showing the relationship between customer segments and departments shopped . . . . .	144

5.7	A mosaic plot showing no relationship between two categorical variables . . . . .	144
5.8	A second example of a dotchart . . . . .	145
5.9	A basic scatterplot, showing monthly sales for our two largest departments across 6 years . . . . .	146
5.10	An example of multivariate data . . . . .	148
5.11	A scatterplot, faceted by department, of spend versus items, with a loess smoother added to each panel . . . . .	149
6.1	The fitted model of depression showing the estimated expected value of depression score given fatalism and simplicity scores . . . . .	166
6.2	Percent body fat plotted against skinfold thickness for 202 Australian athletes . . . . .	188
6.3	The distribution of counts by hour for registered and casual users . . . . .	197
6.4	Residuals plotted against the fitted values obtained from the regression of registered counts against hour of the day, holiday, and workingday . . . . .	203
6.5	Sample autocorrelation coefficients $\hat{\rho}_r$ , $r = 0, 1, \dots, 30$ , plotted against lag ( $r$ ) . . . . .	206
6.6	Residuals from the regression of registered counts against hour of the day, holiday, and working day plotted against day since January 1, 2011. A smooth is plotted to summarize trend . . . . .	207
6.7	A quantile-quantile plot comparing the distribution of the residuals to the standard normal distribution . . . . .	207
7.1	Estimated diabetes incidence plotted against estimated prevalence by state and U.S. territory . . . . .	231
7.2	Sensitivity and specificity plotted against the threshold $p$ . . . . .	237
8.1	Empirical body mass index distributions for five states . . . . .	256
8.2	Estimated distributions of body mass index for five clusters of states . . . . .	265
8.3	Estimated incidence and prevalence of diabetes . . . . .	266
9.1	Observations on carapace length and frontal lobe size measured on two color forms of the species <i>Leptograpsus variegatus</i> . . . . .	281
9.2	Weights assigned to neighbors by the conventional $k$ -nearest neighbor and exponentially-weighted $k$ -nearest neighbor prediction functions . . . . .	287
9.3	Number of reported measles cases by month in California . . . . .	299
9.4	S&P 500 indexes and the exponentially weighted $k$ -nearest neighbor regression predictions plotted against date . . . . .	308



11.1	Number of consumer complaints about mortgages plotted against date . . . . .	345
11.2	Exponential weights $w_{n-t}$ plotted against $n - t$ . . . . .	352
11.3	Estimates of the autocorrelation coefficient for lags 1, 2, . . . , 20 . . . . .	355
11.4	Apple stock prices circa 2013 . . . . .	362
11.5	Forecasting errors for a sequence of 10,000 time steps obtained from two linear regression prediction functions . . . . .	363
12.1	Observed and forecasted prices of Apple Inc. stock . . . . .	385
12.2	Frequencies of the 20 most common hashtags collected from a stream of 50,000 tweets, December 9, 2015 . . . . .	395
A.1	Dotchart of monthly sales by department . . . . .	406
A.2	A faceted graphic showing the empirical density of sales per month by department . . . . .	406
A.3	Monthly sales by department . . . . .	407
A.4	Percent body fat plotted against skinfold thickness for 202 Australian athletes . . . . .	408
A.5	Pre- and post-experiment weights for $n = 72$ anorexia patients. Points are identified by treatment group . . . . .	409
A.6	Pre- and post-experiment weights for $n = 72$ anorexia patients. Separate regression lines are shown for each treatment group . . . . .	409
A.7	Pre- and post-experiment weights for $n = 72$ anorexia patients. Data are plotted by treatment group along with a regression line . . . . .	410



# List of Tables

- 1.1 A few profiles and estimated diabetes prevalence. Data from the Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System surveys . . . . . 4
- 2.1 Files and dictionaries used in Tutorial 2.6 . . . . . 32
- 2.2 The five major committee pairs from the 2012 election cycle with the largest Jaccard similarity. Also shown are the conditional probabilities  $\Pr(A|B)$  and  $\Pr(B|A)$  . . . . . 48
- 2.3 The top eight recipients of contributions from the Bachmann for Congress PAC during the 2010–2012 election cycle . . . . . 50
- 3.1 BRFSS data file names and sub-string positions of body mass index, sampling weight, and gender . . . . . 63
- 3.2 BRFSS data file names and field locations of the income, education, and age variables . . . . . 82
- 3.3 The sample correlation matrix between income, body mass index, and education computed from BRFSS data files . . . . . 87
- 3.4 Possible answers and codes to the question *would you say that in general your health is:* . . . . . 95
- 3.5 BRFSS data file names and field positions of the general health variable . . . . . 96
- 4.1 Some well-known three-digit zip code prefixes and the name of the USPS Sectional Center Facility that serves the zip code area . . . . . 114
- 5.1 Number of receipts cross-classified by department and the three largest customer segments, light, secondary, and primary 143
- 5.2 The first few rows of the data frame `month.summary` . . . . . 151
- 5.3 The first five rows of the `dept.summary` data.frame . . . . . 154

6.1	Parameter estimates, standard errors, and approximate 95% confidence intervals for the parameters of model (6.3) . . . . .	165
6.2	Parameter estimates and standard errors obtained from the linear regression of depression score on fatalism and simplicity . . . . .	173
6.3	Distribution of consumer complaint types obtained from $n = 269,064$ complaints lodged with the Consumer Financial Protection Bureau between January 2012 and July 2014 . . . . .	187
6.4	Parameter estimates and standard errors obtained from the linear regression of skinfold thickness on percent body fat . . . . .	189
6.5	Parameter estimates and standard errors obtained for the interaction model (formula (6.10)) . . . . .	190
6.6	Details of the extra-sums-of-squares $F$ -test for sport . . . . .	193
6.7	The extra-sums-of-squares $F$ -test for interaction between sport and gender . . . . .	194
6.8	Summary statistics from the models of user counts as a function of hour of the day and the working day indicator variable . . . . .	200
6.9	Model 6.16 for specific combinations of hour of day and holiday . . . . .	204
7.1	BRFSS data file field positions for sampling weight, gender, income, education, age class, body mass index (BMI), and diabetes . . . . .	223
7.2	BRFSS codes for diabetes . . . . .	223
7.3	Functions, where they were developed, and their purpose . . . . .	224
7.4	Data sets for the analysis of diabetes prevalence and incidence . . . . .	224
7.5	Ordinal variables and the number of levels of each . . . . .	233
7.6	A confusion matrix showing the classification of risk prediction outcomes of $n_{++}$ individuals . . . . .	235
9.1	A confusion matrix showing the results of predicting the group memberships of a set of test observations . . . . .	297
9.2	Estimated conditional accuracies obtained from the conventional eight-nearest neighbor prediction function using a training set of 37,800 observations and a test set of 4200 observations . . . . .	298
9.3	Apparent and cross-validation accuracy estimates for the $k$ -nearest-neighbor prediction function . . . . .	312
10.1	Authors of the Federalist papers . . . . .	314
10.2	A confusion matrix showing the results of predicting the authors of the Federalist papers . . . . .	329

10.3 A partial record from the data file . . . . . 332

10.4 Predicted customer segments for non-members . . . . . 338

11.1 Contents of the past-values storage list for  $\tau = 5$   
 for time steps  $t, t + 1$  and  $t + 4$  when  $t$  is a multiple of  $\tau$   
 (and hence,  $t \bmod \tau = 0$ ) . . . . . 356

12.1 A few dictionary entries showing polarity strength  
 and direction . . . . . 397

12.2 Numerical scores assigned to sentiment classes . . . . . 398

A.1 Fitted models for males and females . . . . . 408

A.2 Confidence intervals for  $\beta_1$  for males and females . . . . . 408

A.3 Confidence intervals for the centered intercepts . . . . . 409

A.4 Values of sensitivity and specificity for five choices of the  
 threshold  $p$  . . . . . 411

A.5 Estimated probabilities of group membership from the  
 conventional  $k$ -nearest-neighbor prediction function . . . . . 412

A.6 Estimates of root mean square prediction error  $\widehat{\sigma}_{\text{kNN}}$   
 as a function of  $d$  and  $\alpha$  . . . . . 412

A.7 Estimates of  $\sigma_{\text{reg}}^2$  for three choices of  $\alpha$  and predictor variable . 413

A.8 Mean sentiment of tweets containing a particular emotion . . . . 415





<http://www.springer.com/978-3-319-45795-6>

Algorithms for Data Science

Steele, B.; Chandler, J.; Reddy, S.

2016, XXIII, 430 p. 48 illus., 30 illus. in color., Hardcover

ISBN: 978-3-319-45795-6