

CONTENTS IN DETAIL

ACKNOWLEDGMENTS	xvii
------------------------	-------------

INTRODUCTION	xix
---------------------	------------

Who Is This Book For?	xx
About This Book.	xxi
Setting Up the Environment	xxii
Windows	xxii
macOS	xxii
Linux	xxiii
Installing Packages with Python	xxiv
Other Tools	xxv
Summary	xxv

1	
EXPLORATORY DATA ANALYSIS	1

Your First Day as CEO	2
Finding Patterns in Datasets	2
Using .csv Files to Review and Store Data	4
Displaying Data with Python	5
Calculating Summary Statistics	8
Analyzing Subsets of Data	10
Nighttime Data	10
Seasonal Data	11
Visualizing Data with Matplotlib	13
Drawing and Displaying a Simple Plot.	13
Clarifying Plots with Titles and Labels.	14
Plotting Subsets of Data	15
Testing Different Plot Types	16
Exploring Correlations	21
Calculating Correlations	21
Understanding Strong vs. Weak Correlations	22
Finding Correlations Between Variables	25
Creating Heat Maps	26
Exploring Further	29
Summary	29

2	
FORECASTING	31

Predicting Customer Demand.	32
Cleaning Erroneous Data	32
Plotting Data to Find Trends.	34

Performing Linear Regression	35
Applying Algebra to the Regression Line	38
Calculating Error Measurements	40
Using Regression to Forecast Future Trends	43
Trying More Regression Models	45
Multivariate Linear Regression to Predict Sales	45
Trigonometry to Capture Variations	47
Choosing the Best Regression to Use for Forecasting	50
Exploring Further	54
Summary	55

3
GROUP COMPARISONS **57**

Reading Population Data	58
Summary Statistics	58
Random Samples	59
Differences Between Sample Data	62
Performing Hypothesis Testing	66
The t-Test	67
Nuances of Hypothesis Testing	68
Comparing Groups in a Practical Context	70
Summary	74

4
A/B TESTING **75**

The Need for Experimentation	76
Running Experiments to Test New Hypotheses	77
Understanding the Math of A/B Testing	80
Translating the Math into Practice	81
Optimizing with the Champion/Challenger Framework	83
Preventing Mistakes with Twyman’s Law and A/A Testing	84
Understanding Effect Sizes	86
Calculating the Significance of Data	88
Applications and Advanced Considerations	90
The Ethics of A/B Testing	91
Summary	93

5
BINARY CLASSIFICATION **95**

Minimizing Customer Attrition	96
Using Linear Probability Models to Find High-Risk Customers	97
Plotting Attrition Risk	98
Confirming Relationships with Linear Regression	99
Predicting the Future	102
Making Business Recommendations	103
Measuring Prediction Accuracy	104
Using Multivariate LPMs	106
Creating New Metrics	108
Considering the Weaknesses of LPMs	109

Predicting Binary Outcomes with Logistic Regression	110
Drawing Logistic Curves.	110
Fitting the Logistic Function to Our Data	112
Applications of Binary Classification	114
Summary	115

6 SUPERVISED LEARNING 117

Predicting Website Traffic	118
Reading and Plotting News Article Data	119
Using Linear Regression as a Prediction Method	121
Understanding Supervised Learning	122
k-Nearest Neighbors	124
Implementing k-NN	126
Performing k-NN with Python’s sklearn	127
Using Other Supervised Learning Algorithms	128
Decision Trees	130
Random Forests	131
Neural Networks	132
Measuring Prediction Accuracy	134
Working with Multivariate Models	136
Using Classification Instead of Regression	137
Summary	139

7 UNSUPERVISED LEARNING 141

Unsupervised Learning vs. Supervised Learning	142
Generating and Exploring Data	143
Rolling the Dice	143
Using Another Kind of Die	147
The Origin of Observations with Clustering	149
Clustering in Business Applications.	152
Analyzing Multiple Dimensions	154
E-M Clustering	155
The Guessing Step	157
The Expectation Step	158
The Maximization Step	160
The Convergence Step.	162
Other Clustering Methods	164
Other Unsupervised Learning Methods	167
Summary	168

8 WEB SCRAPING 169

Understanding How Websites Work	170
Creating Your First Web Scraper	171
Parsing HTML Code	173
Scraping an Email Address	173
Searching for Addresses Directly.	175

Performing Searches with Regular Expressions	176
Using Metacharacters for Flexible Searches	178
Fine-Tuning Searches with Escape Sequences	178
Combining Metacharacters for Advanced Searches	180
Using Regular Expressions to Search for Email Addresses	181
Converting Results to Usable Data	182
Using BeautifulSoup	184
Parsing HTML Label Elements	185
Scraping and Parsing HTML Tables	186
Advanced Scraping	188
Summary	189

9

RECOMMENDATION SYSTEMS 191

Popularity-Based Recommendations	192
Item-Based Collaborative Filtering	194
Measuring Vector Similarity	195
Calculating Cosine Similarity	198
Implementing Item-Based Collaborative Filtering	199
User-Based Collaborative Filtering	201
Case Study: Music Recommendations	204
Generating Recommendations with Advanced Systems	206
Summary	207

10

NATURAL LANGUAGE PROCESSING 209

Using NLP to Detect Plagiarism	210
Understanding the word2vec NLP Model	211
Quantifying Similarities Between Words	211
Creating a System of Equations	213
Analyzing Numeric Vectors in word2vec	218
Manipulating Vectors with Mathematical Calculations	221
Detecting Plagiarism with word2vec	222
Using Skip-Thoughts	223
Topic Modeling	225
Other Applications of NLP	228
Summary	229

11

DATA SCIENCE IN OTHER LANGUAGES 231

Winning Soccer Games with SQL	232
Reading and Analyzing Data	232
Getting Familiar with SQL	234
Setting Up a SQL Database	234
Running SQL Queries	235
Combining Data by Joining Tables	238

Winning Soccer Games with R	241
Getting Familiar with R	241
Applying Linear Regression in R	243
Using R to Plot Data	244
Gaining Other Valuable Skills	246
Summary	248

INDEX **249**