Contents at a Glance

Introduction 1

Part 1: Getting Started with Data Science and Python 7

CHAPTER 1: Discovering the Match between Data Science and Python 9
CHAPTER 2: Introducing Python's Capabilities and Wonders 21
CHAPTER 3: Setting Up Python for Data Science 33
CHAPTER 4: Working with Google Colab 49

Part 2: Getting Your Hands Dirty with Data 71

CHAPTER 5: Working with Jupyter Notebook 73
CHAPTER 6: Working with Real Data 83
CHAPTER 7: Processing Your Data 105
CHAPTER 8: Reshaping Data 131
CHAPTER 9: Putting What You Know into Action 143

Part 3: Visualizing Information 157

CHAPTER 10: Getting a Crash Course in Matplotlib 159
CHAPTER 11: Visualizing the Data 177

Part 4: Wrangling Data 199

CHAPTER 12: Stretching Python's Capabilities 201
CHAPTER 13: Exploring Data Analysis 223
CHAPTER 14: Reducing Dimensionality 251
CHAPTER 15: Clustering 273
CHAPTER 16: Detecting Outliers in Data 291

Part 5: Learning from Data 305

CHAPTER 17: Exploring Four Simple and Effective Algorithms 307
CHAPTER 18: Performing Cross-Validation, Selection, and Optimization 327
CHAPTER 19: Increasing Complexity with Linear and Nonlinear Tricks 351
CHAPTER 20: Understanding the Power of the Many 391

Part 6: The Part of Tens 413

CHAPTER 21: Ten Essential Data Resources 415
CHAPTER 22: Ten Data Challenges You Should Take 421

Index 431

COPYRIGHTED MATERIAL

Table of Contents

INTRODUCTION 1
About This Book 2
Foolish Assumptions 3
Icons Used in This Book 4
Beyond the Book 4
Where to Go from Here 5

PART 1: GETTING STARTED WITH DATA SCIENCE AND PYTHON 7

CHAPTER 1: Discovering the Match between Data Science and Python 9
Understanding Python as a Language 10
Viewing Python's various uses as a general-purpose language . . . 10
Interpreting Python 11
Compiling Python 12
Defining Data Science 12
Considering the emergence of data science 12
Outlining the core competencies of a data scientist 13
Linking data science, big data, and AI 14
Creating the Data Science Pipeline 15
Understanding Python's Role in Data Science 16
Considering the shifting profile of data scientists 16
Working with a multipurpose, simple, and efficient language 17
Learning to Use Python Fast 17
Loading data 18
Training a model 18
Viewing a result 19

CHAPTER 2: Introducing Python's Capabilities and Wonders . . . 21
Working with Python 22
Contributing to data science 22
Getting a taste of the language 23
Understanding the need for indentation 24
Working with Jupyter Notebook and Google Colab 24
Performing Rapid Prototyping and Experimentation 24
Considering Speed of Execution 26
Visualizing Power 27
Using the Python Ecosystem for Data Science 28
Accessing scientific tools using SciPy 29
Performing fundamental scientific computing using NumPy 29

Performing data analysis using pandas 29
Implementing machine learning using Scikit-learn 30
Going for deep learning with Keras and TensorFlow 30
Performing analysis efficiently using XGBoost 31
Plotting the data using Matplotlib 31
Creating graphs with NetworkX 31

CHAPTER 3: **Setting Up Python for Data Science** 33
Working with Anaconda 34
Using Jupyter Notebook 34
Accessing the Anaconda Prompt 35
Installing Anaconda on Windows 36
Installing Anaconda on Linux 39
Installing Anaconda on Mac OS X 40
Downloading the Datasets and Example Code 42
Using Jupyter Notebook 42
Defining the code repository 43
Understanding the datasets used in this book 47

CHAPTER 4: **Working with Google Colab** 49
Defining Google Colab 50
Understanding what Google Colab does 50
Considering the online coding difference 51
Using local runtime support 53
Working with Notebooks 53
Creating a new notebook 54
Opening existing notebooks 54
Saving notebooks 56
Downloading notebooks 59
Performing Common Tasks 59
Creating code cells 60
Creating text cells 62
Creating special cells 63
Editing cells 64
Moving cells 64
Using Hardware Acceleration 64
Executing the Code 65
Viewing Your Notebook 66
Displaying the table of contents 66
Getting notebook information 66
Checking code execution 67
Sharing Your Notebook 67
Getting Help 68

PART 2: GETTING YOUR HANDS DIRTY WITH DATA 71

CHAPTER 5: Working with Jupyter Notebook 73

Using Jupyter Notebook 74
Working with styles 74
Getting Python help 75
Using magic functions 76
Discovering objects 77
Restarting the kernel 79
Restoring a checkpoint 79
Performing Multimedia and Graphic Integration 79
Embedding plots and other images 80
Loading examples from online sites 80
Obtaining online graphics and multimedia 80

CHAPTER 6: Working with Real Data 83

Uploading, Streaming, and Sampling Data 84
Uploading small amounts of data into memory 84
Streaming large amounts of data into memory 86
Generating variations on image data 86
Sampling data in different ways 88
Accessing Data in Structured Flat-File Form 89
Reading from a text file 90
Reading CSV delimited format 90
Reading Excel and other Microsoft Office files 93
Sending Data in Unstructured File Form 94
Managing Data from Relational Databases 97
Interacting with Data from NoSQL Databases 99
Accessing Data from the Web 100

CHAPTER 7: Processing Your Data 105

Juggling between NumPy and pandas 106
Knowing when to use NumPy 106
Knowing when to use pandas 107
Validating Your Data 108
Figuring out what's in your data 108
Removing duplicates 110
Creating a data map and data plan 110
Manipulating Categorical Variables 113
Creating categorical variables 114
Renaming levels 115
Combining levels 116
Dealing with Dates in Your Data 117
Formatting date and time values 118
Using the right time transformation 118

Dealing with Missing Data 119
Finding the missing data 120
Encoding missingness 120
Imputing missing data 121
Slicing and Dicing: Filtering and Selecting Data 123
Slicing rows 123
Slicing columns 124
Dicing 124
Concatenating and Transforming 125
Adding new cases and variables 126
Removing data 127
Sorting and shuffling 128
Aggregating Data at Any Level 129

CHAPTER 8: **Reshaping Data** 131
Using the Bag of Words Model to Tokenize Data 132
Understanding the bag of words model 132
Sequencing text items with n-grams 134
Implementing TF-IDF transformations 136
Working with Graph Data 138
Understanding the adjacency matrix 138
Using NetworkX basics 139

CHAPTER 9: **Putting What You Know into Action** 143
Contextualizing Problems and Data 144
Evaluating a data science problem 145
Researching solutions 147
Formulating a hypothesis 148
Preparing your data 149
Considering the Art of Feature Creation 149
Defining feature creation 149
Combining variables 150
Understanding binning and discretization 151
Using indicator variables 151
Transforming distributions 152
Performing Operations on Arrays 153
Using vectorization 153
Performing simple arithmetic on vectors and matrices 154
Performing matrix vector multiplication 154
Performing matrix multiplication 155

PART 3: VISUALIZING INFORMATION 157

CHAPTER 10: Getting a Crash Course in Matplotlib 159

Starting with a Graph 160
Defining the plot 160
Drawing multiple lines and plots 161
Saving your work to disk 162
Setting the Axis, Ticks, and Grids 163
Getting the axes 163
Formatting the axes 164
Adding grids 165
Defining the Line Appearance 167
Working with line styles 167
Using colors 168
Adding markers 169
Using Labels, Annotations, and Legends 171
Adding labels 172
Annotating the chart 173
Creating a legend 174

CHAPTER 11: Visualizing the Data 177

Choosing the Right Graph 178
Creating comparisons with bar charts 178
Showing distributions using histograms 179
Depicting groups using boxplots 181
Seeing data patterns using scatterplots 182
Creating Advanced Scatterplots 183
Depicting groups 184
Showing correlations 184
Plotting Time Series 186
Representing time on axes 186
Plotting trends over time 188
Plotting Geographical Data 190
Using an environment in Notebook 191
Using Cartopy to plot geographic data 192
Avoiding outdated libraries: The Basemap Toolkit 194
Visualizing Graphs 195
Developing undirected graphs 195
Developing directed graphs 197

PART 4: WRANGLING DATA 199

CHAPTER 12: Stretching Python's Capabilities 201
Playing with Scikit-learn 202
Understanding classes in Scikit-learn 202
Defining applications for data science 203
Using Transformative Functions 207
Chaining estimators 208
Transforming targets 209
Composing features 209
Handling heterogeneous data 210
Considering Timing and Performance 212
Benchmarking with timeit 213
Working with the memory profiler 217
Running in Parallel on Multiple Cores 220
Performing multicore parallelism 220
Demonstrating multiprocessing 221

CHAPTER 13: Exploring Data Analysis 223
The EDA Approach 224
Defining Descriptive Statistics for Numeric Data 225
Measuring central tendency 227
Measuring variance and range 228
Working with percentiles 228
Defining measures of normality 229
Counting for Categorical Data 231
Understanding frequencies 232
Creating contingency tables 233
Creating Applied Visualization for EDA 234
Inspecting boxplots 234
Performing t-tests after boxplots 236
Observing parallel coordinates 238
Graphing distributions 238
Plotting scatterplots 240
Understanding Correlation 241
Using covariance and correlation 242
Using nonparametric correlation 244
Considering chi-square for tables 245
Working with Cramér's V 246
Modifying Data Distributions 246
Using different statistical distributions 247
Creating a Z-score standardization 247
Transforming other notable distributions 248

CHAPTER 14: **Reducing Dimensionality** 251
Understanding SVD 252
Looking for dimensionality reduction 253
Using SVD to measure the invisible 255
Performing Factor Analysis and PCA 256
Considering the psychometric model 257
Looking for hidden factors 257
Using components, not factors 258
Achieving dimensionality reduction 259
Squeezing information with t-SNE 259
Understanding Some Applications 261
Recognizing faces with PCA 262
Extracting topics with NMF 266
Recommending movies 268

CHAPTER 15: **Clustering** 273
Clustering with K-means 274
Understanding centroid-based algorithms 275
Creating an example with image data 277
Looking for optimal solutions 278
Clustering big data 281
Performing Hierarchical Clustering 283
Using a hierarchical cluster solution 284
Visualizing aggregative clustering solutions 285
Discovering New Groups with DBScan 287

CHAPTER 16: **Detecting Outliers in Data** 291
Considering Outlier Detection 292
Finding more things that can go wrong 293
Understanding anomalies and novel data 294
Examining a Simple Univariate Method 295
Leveraging on the Gaussian distribution 297
Remediating outliers 298
Developing a Multivariate Approach 300
Using principal component analysis 300
Using cluster analysis for spotting outliers 302
Automating detection with Isolation Forests 303

PART 5: LEARNING FROM DATA 305

CHAPTER 17: Exploring Four Simple and Effective Algorithms 307
Guessing the Number: Linear Regression 307
Defining the family of linear models 308
Using more variables 309
Understanding limitations and problems 312
Moving to Logistic Regression 313
Applying logistic regression 313
Considering the case when there are more classes 315
Making Things as Simple as Naïve Bayes 317
Finding out that Naïve Bayes isn't so naïve 318
Predicting text classifications 320
Learning Lazily with Nearest Neighbors 322
Predicting after observing neighbors 323
Choosing your k parameter wisely 324

CHAPTER 18: Performing Cross-Validation, Selection, and Optimization 327
Pondering the Problem of Fitting a Model 328
Understanding bias and variance 329
Defining a strategy for picking models 330
Dividing between training and test sets 333
Cross-Validating 336
Using cross-validation on k folds 337
Sampling stratifications for complex data 338
Selecting Variables Like a Pro 341
Selecting by univariate measures 341
Employing forward and backward selection 343
Pumping Up Your Hyperparameters 344
Implementing a grid search 345
Trying a randomized search 348

CHAPTER 19: Increasing Complexity with Linear and Nonlinear Tricks 351
Using Nonlinear Transformations 352
Doing variable transformations 353
Creating interactions between variables 354
Regularizing Linear Models 357
Relying on Ridge regression (L2) 358
Using the Lasso (L1) 359
Leveraging regularization 360
Combining L1 & L2: Elasticnet 361

Fighting with Big Data Chunk by Chunk 361
Determining when there is too much data 362
Implementing Stochastic Gradient Descent 362
Understanding Support Vector Machines 365
Relying on a computational method 366
Fixing many new parameters 368
Classifying with SVC 371
Going nonlinear is easy 376
Performing regression with SVR 378
Creating a stochastic solution with SVM 380
Playing with Neural Networks 385
Understanding neural networks 385
Classifying and regressing with neurons 386

CHAPTER 20: **Understanding the Power of the Many** 391
Starting with a Plain Decision Tree 392
Understanding a decision tree 392
Creating classification trees 395
Creating regression trees 397
Getting Lost in a Random Forest 399
Making machine learning accessible 399
Working with a Random Forest classifier 402
Working with a Random Forest regressor 404
Optimizing a Random Forest 404
Boosting Predictions 406
Knowing that many weak predictors win 406
Setting a gradient boosting classifier 407
Running a gradient boosting regressor 408
Using GBM hyperparameters 409
Using XGBoost 410

PART 6: THE PART OF TENS 413

CHAPTER 21: **Ten Essential Data Resources** 415
Discovering the News with Reddit 416
Getting a Good Start with KDnuggets 416
Locating Free Learning Resources with Quora 416
Gaining Insights with Oracle's AI & Data Science Blog 417
Accessing the Huge List of Resources on Data Science Central 417
Discovering New Beginner Data Science Methodologies at Data Science 101 418
Obtaining the Most Authoritative Sources at Udacity 418
Receiving Help with Advanced Topics at Conductrics 419
Obtaining the Facts of Open Source Data Science from Springboard 420
Zeroing In on Developer Resources with Jonathan Bower 420

CHAPTER 22: **Ten Data Challenges You Should Take** 421
Removing Personally Identifiable Information 422
Creating a Secure Data Environment 423
Working with a Multiple-Data-Source Problem. 423
Honing Your Overfit Strategies 424
Trudging Through the MovieLens Dataset 425
Locating the Correct Data Source 425
Working with Handwritten Information. 426
Working with Pictures 427
Indentifying Data Lineage 428
Interacting with a Huge Graph 429

INDEX 431