

Table of Contents

Preface	1
Section 1: Foundation for Data Analysis	
Chapter 1: Getting Started with Python Libraries	9
Understanding data analysis	10
The standard process of data analysis	11
The KDD process	12
SEMMA	13
CRISP-DM	14
Comparing data analysis and data science	16
The roles of data analysts and data scientists	16
The skillsets of data analysts and data scientists	17
Installing Python 3	19
Python installation and setup on Windows	19
Python installation and setup on Linux	20
Python installation and setup on Mac OS X with a GUI installer	20
Python installation and setup on Mac OS X with brew	20
Software used in this book	21
Using IPython as a shell	22
Reading manual pages	25
Where to find help and references to Python data analysis libraries	26
Using JupyterLab	26
Using Jupyter Notebooks	28
Advanced features of Jupyter Notebooks	29
Keyboard shortcuts	30
Installing other kernels	31
Running shell commands	32
Extensions for Notebook	32
Summary	38
Chapter 2: NumPy and pandas	39
Technical requirements	40
Understanding NumPy arrays	40
Array features	43
Selecting array elements	44
NumPy array numerical data types	45
dtype objects	47
Data type character codes	48

dtype constructors	49
dtype attributes	49
Manipulating array shapes	50
The stacking of NumPy arrays	52
Partitioning NumPy arrays	55
Changing the data type of NumPy arrays	57
Creating NumPy views and copies	58
Slicing NumPy arrays	60
Boolean and fancy indexing	62
Broadcasting arrays	63
Creating pandas DataFrames	65
Understanding pandas Series	67
Reading and querying the Quandl data	70
Describing pandas DataFrames	74
Grouping and joining pandas DataFrame	77
Working with missing values	81
Creating pivot tables	83
Dealing with dates	85
Summary	87
References	87
Chapter 3: Statistics	89
Technical requirements	90
Understanding attributes and their types	90
Types of attributes	90
Discrete and continuous attributes	92
Measuring central tendency	92
Mean	92
Mode	93
Median	94
Measuring dispersion	94
Skewness and kurtosis	98
Understanding relationships using covariance and correlation coefficients	99
Pearson's correlation coefficient	100
Spearman's rank correlation coefficient	101
Kendall's rank correlation coefficient	101
Central limit theorem	101
Collecting samples	103
Performing parametric tests	104
Performing non-parametric tests	110
Summary	116
Chapter 4: Linear Algebra	117

Technical requirements	118
Fitting to polynomials with NumPy	118
Determinant	120
Finding the rank of a matrix	120
Matrix inverse using NumPy	121
Solving linear equations using NumPy	122
Decomposing a matrix using SVD	123
Eigenvectors and Eigenvalues using NumPy	125
Generating random numbers	126
Binomial distribution	127
Normal distribution	129
Testing normality of data using SciPy	130
Creating a masked array using the numpy.ma subpackage	134
Summary	136
 Section 2: Exploratory Data Analysis and Data Cleaning	
 Chapter 5: Data Visualization	139
Technical requirements	139
Visualization using Matplotlib	140
Accessories for charts	141
Scatter plot	143
Line plot	144
Pie plot	146
Bar plot	147
Histogram plot	148
Bubble plot	150
pandas plotting	152
Advanced visualization using the Seaborn package	154
lm plots	155
Bar plots	158
Distribution plots	159
Box plots	160
KDE plots	161
Violin plots	162
Count plots	163
Joint plots	165
Heatmaps	166
Pair plots	168
Interactive visualization with Bokeh	170
Plotting a simple graph	170
Glyphs	172
Layouts	173

Nested layout using row and column layouts	177
Multiple plots	179
Interactions	181
Hide click policy	181
Mute click policy	183
Annotations	184
Hover tool	187
Widgets	188
Tab panel	189
Slider	190
Summary	193
Chapter 6: Retrieving, Processing, and Storing Data	195
Technical requirements	196
Reading and writing CSV files with NumPy	196
Reading and writing CSV files with pandas	197
Reading and writing data from Excel	199
Reading and writing data from JSON	200
Reading and writing data from HDF5	201
Reading and writing data from HTML tables	202
Reading and writing data from Parquet	203
Reading and writing data from a pickle pandas object	204
Lightweight access with sqllite3	205
Reading and writing data from MySQL	206
Inserting a whole DataFrame into the database	209
Reading and writing data from MongoDB	210
Reading and writing data from Cassandra	211
Reading and writing data from Redis	212
PonyORM	213
Summary	214
Chapter 7: Cleaning Messy Data	215
Technical requirements	216
Exploring data	216
Filtering data to weed out the noise	219
Column-wise filtration	220
Row-wise filtration	222
Handling missing values	225
Dropping missing values	226
Filling in a missing value	226
Handling outliers	228
Feature encoding techniques	231
One-hot encoding	231
Label encoding	233
Ordinal encoder	234

Feature scaling	235
Methods for feature scaling	236
Feature transformation	239
Feature splitting	240
Summary	241
Chapter 8: Signal Processing and Time Series	243
Technical requirements	244
The statsmodels modules	244
Moving averages	245
Window functions	248
Defining cointegration	250
STL decomposition	252
Autocorrelation	254
Autoregressive models	256
ARMA models	260
Generating periodic signals	263
Fourier analysis	265
Spectral analysis filtering	268
Summary	270
Section 3: Deep Dive into Machine Learning	
Chapter 9: Supervised Learning - Regression Analysis	273
Technical requirements	274
Linear regression	274
Multiple linear regression	276
Understanding multicollinearity	276
Removing multicollinearity	277
Dummy variables	279
Developing a linear regression model	281
Evaluating regression model performance	283
R-squared	283
MSE	284
MAE	284
RMSE	285
Fitting polynomial regression	286
Regression models for classification	289
Logistic regression	289
Characteristics of the logistic regression model	291
Types of logistic regression algorithms	292
Advantages and disadvantages of logistic regression	292
Implementing logistic regression using scikit-learn	293
Summary	295

Chapter 10: Supervised Learning - Classification Techniques	297
Technical requirements	298
Classification	298
Naive Bayes classification	300
Decision tree classification	304
KNN classification	307
SVM classification	310
Terminology	310
Splitting training and testing sets	313
Holdout	313
K-fold cross-validation	314
Bootstrap method	314
Evaluating the classification model performance	315
Confusion matrix	315
Accuracy	318
Precision	319
Recall	319
F-measure	319
ROC curve and AUC	320
Summary	323
Chapter 11: Unsupervised Learning - PCA and Clustering	325
Technical requirements	326
Unsupervised learning	326
Reducing the dimensionality of data	327
PCA	328
Performing PCA	329
Clustering	332
Finding the number of clusters	333
The elbow method	334
The silhouette method	336
Partitioning data using k-means clustering	338
Hierarchical clustering	341
DBSCAN clustering	345
Spectral clustering	347
Evaluating clustering performance	350
Internal performance evaluation	351
The Davies-Bouldin index	351
The silhouette coefficient	351
External performance evaluation	352
The Rand score	352
The Jaccard score	352
F-Measure or F1-score	353
The Fowlkes-Mallows score	353
Summary	356

Section 4: NLP, Image Analytics, and Parallel Computing

Chapter 12: Analyzing Textual Data	359
Technical requirements	360
Installing NLTK and SpaCy	360
Text normalization	361
Tokenization	362
Removing stopwords	366
Stemming and lemmatization	368
POS tagging	370
Recognizing entities	371
Dependency parsing	372
Creating a word cloud	373
Bag of Words	375
TF-IDF	376
Sentiment analysis using text classification	377
Classification using BoW	378
Classification using TF-IDF	383
Text similarity	386
Jaccard similarity	387
Cosine similarity	388
Summary	389
Chapter 13: Analyzing Image Data	391
Technical requirements	392
Installing OpenCV	392
Understanding image data	393
Binary images	393
Grayscale images	394
Color images	394
Color models	395
Drawing on images	398
Writing on images	403
Resizing images	404
Flipping images	406
Changing the brightness	409
Blurring an image	410
Face detection	414
Summary	418
Chapter 14: Parallel Computing Using Dask	419
Parallel computing using Dask	420
Dask data types	421

Dask Arrays	422
Dask DataFrames	423
DataFrame Indexing	424
Filter data	427
Groupby	428
Converting a pandas DataFrame into a Dask DataFrame	429
Converting a Dask DataFrame into a pandas DataFrame	429
Dask Bags	430
Creating a Dask Bag using Python iterable items	430
Creating a Dask Bag using a text file	431
Storing a Dask Bag in a text file	432
Storing a Dask Bag in a DataFrame	432
Dask Delayed	433
Preprocessing data at scale	435
Feature scaling in Dask	435
Feature encoding in Dask	437
Machine learning at scale	439
Parallel computing using scikit-learn	440
Reimplementing ML algorithms for Dask	442
Logistic regression	442
Clustering	444
Summary	446
Other Books You May Enjoy	449
Index	453