

# Table of Contents

**About the Author .....xvii**

**About the Technical Reviewer .....xix**

**Preface .....xxi**

**■ Chapter 1: An Introduction to Data Analysis ..... 1**

    Data Analysis..... 1

    Knowledge Domains of the Data Analyst ..... 2

        Computer Science ..... 2

        Mathematics and Statistics ..... 3

        Machine Learning and Artificial Intelligence ..... 3

        Professional Fields of Application..... 3

    Understanding the Nature of the Data..... 4

        When the Data Become Information..... 4

        When the Information Becomes Knowledge..... 4

        Types of Data ..... 4

    The Data Analysis Process ..... 4

        Problem Definition ..... 5

        Data Extraction ..... 6

        Data Preparation..... 6

        Data Exploration/Visualization ..... 7

        Predictive Modeling ..... 7

        Model Validation ..... 8

        Deployment..... 8



■ TABLE OF CONTENTS

Quantitative and Qualitative Data Analysis.....	9
Open Data.....	9
Python and Data Analysis.....	12
Conclusions.....	13
■ <b>Chapter 2: Introduction to the Python World</b> .....	<b>15</b>
Python—The Programming Language.....	15
The Interpreter and the Execution Phases of the Code.....	16
Installing Python.....	18
Python Distributions.....	19
Using Python.....	23
Writing Python Code.....	26
IPython.....	30
PyPI—The Python Package Index.....	36
The IDEs for Python.....	37
SciPy.....	42
NumPy.....	42
Pandas.....	43
matplotlib.....	43
Conclusions.....	43
■ <b>Chapter 3: The NumPy Library</b> .....	<b>45</b>
NumPy: A Little History.....	45
The NumPy Installation.....	46
ndarray: The Heart of the Library.....	47
Create an Array.....	48
Types of Data.....	49
The dtype Option.....	50
Intrinsic Creation of an Array.....	50
Basic Operations.....	51
Arithmetic Operators.....	52
The Matrix Product.....	53



Increment and Decrement Operators .....	54
Universal Functions (ufunc) .....	54
Aggregate Functions.....	55
Indexing, Slicing, and Iterating .....	55
Indexing .....	55
Slicing.....	57
Iterating an Array .....	59
Conditions and Boolean Arrays.....	60
Shape Manipulation .....	61
Array Manipulation .....	62
Joining Arrays.....	62
Splitting Arrays .....	63
General Concepts .....	64
Copies or Views of Objects .....	64
Vectorization.....	65
Broadcasting .....	66
Structured Arrays .....	68
Reading and Writing Array Data on Files .....	70
Loading and Saving Data in Binary Files .....	70
Reading Files with Tabular Data .....	70
Conclusions .....	72
■ Chapter 4: The pandas Library—An Introduction.....	73
pandas: The Python Data Analysis Library .....	73
Installation of pandas .....	74
Installation from Anaconda.....	74
Installation from PyPI.....	78
Getting Started with pandas.....	78
Introduction to pandas Data Structures .....	79



The Series.....	80
The Dataframe .....	87
The Index Objects .....	94
Other Functionalities on Indexes .....	96
Reindexing.....	96
Dropping .....	98
Arithmetic and Data Alignment.....	99
Operations Between Data Structures .....	100
Flexible Arithmetic Methods .....	100
Operations Between Dataframes and Series.....	101
Function Application and Mapping .....	102
Functions by Element .....	102
Functions by Row or Column.....	102
Statistics Functions .....	103
Sorting and Ranking.....	104
Correlation and Covariance .....	107
“Not a Number” Data .....	108
Assigning a NaN Value.....	108
Filtering Out NaN Values.....	109
Filling in NaN Occurrences .....	110
Hierarchical Indexing and Leveling .....	110
Reordering and Sorting Levels .....	112
Summary Statistics with groupby Instead of with Level .....	113
Conclusions .....	114
■ Chapter 5: pandas: Reading and Writing Data.....	115
I/O API Tools.....	115
CSV and Textual Files .....	116
Reading Data in CSV or Text Files.....	116
Using Regexp to Parse TXT Files .....	119
Reading TXT Files Into Parts.....	121
Writing Data in CSV.....	121



Reading and Writing HTML Files .....	123
Writing Data in HTML.....	124
Reading Data from an HTML File .....	126
Reading Data from XML .....	127
Reading and Writing Data on Microsoft Excel Files.....	129
JSON Data .....	131
The HDF5 Format .....	135
Pickle—Python Object Serialization.....	136
Serialize a Python Object with cPickle .....	136
Pickling with pandas .....	137
Interacting with Databases .....	137
Loading and Writing Data with SQLite3 .....	138
Loading and Writing Data with PostgreSQL in a Docker Container .....	140
Reading and Writing Data with a NoSQL Database: MongoDB.....	146
Conclusions .....	148
■ <b>Chapter 6: pandas in Depth: Data Manipulation</b> .....	<b>149</b>
Data Preparation .....	149
Merging .....	150
Concatenating .....	154
Combining .....	156
Pivoting.....	157
Removing.....	160
Data Transformation .....	161
Removing Duplicates.....	161
Mapping.....	162
Discretization and Binning .....	166
Detecting and Filtering Outliers.....	168
Permutation.....	169
Random Sampling .....	170



<b>String Manipulation .....</b>	<b>170</b>
Built-in Methods for String Manipulation .....	170
Regular Expressions .....	172
<b>Data Aggregation.....</b>	<b>173</b>
GroupBy .....	174
A Practical Example.....	175
Hierarchical Grouping.....	176
<b>Group Iteration.....</b>	<b>176</b>
Chain of Transformations.....	177
Functions on Groups.....	178
<b>Advanced Data Aggregation .....</b>	<b>179</b>
<b>Conclusions .....</b>	<b>181</b>
<b>■ Chapter 7: Data Visualization with matplotlib and Seaborn.....</b>	<b>183</b>
The matplotlib Library .....	183
Installation.....	184
The matplotlib Architecture .....	185
Backend Layer .....	186
Artist Layer .....	186
Scripting Layer (pyplot) .....	188
pylab and pyplot .....	188
pyplot .....	189
The Plotting Window .....	189
<b>Data Visualization with Jupyter Notebook.....</b>	<b>191</b>
Set the Properties of the Plot.....	192
matplotlib and NumPy .....	194
<b>Using kwargs.....</b>	<b>196</b>
Working with Multiple Figures and Axes .....	196
<b>Adding Elements to the Chart.....</b>	<b>198</b>
Adding Text .....	198
Adding a Grid .....	202
Adding a Legend.....	203



<b>Saving Your Charts .....</b>	<b>206</b>
Saving the Code.....	206
Saving Your Notebook as an HTML File or as Other File Formats.....	207
Saving Your Chart Directly as an Image.....	208
<b>Handling Date Values .....</b>	<b>208</b>
<b>Chart Typology .....</b>	<b>211</b>
<b>Line Charts .....</b>	<b>211</b>
Line Charts with pandas.....	217
<b>Histograms .....</b>	<b>218</b>
<b>Bar Charts .....</b>	<b>219</b>
Horizontal Bar Charts.....	222
Multiserial Bar Charts.....	223
Multiseries Bar Charts with a pandas Dataframe.....	225
Multiseries Stacked Bar Charts .....	227
Stacked Bar Charts with a pandas Dataframe.....	229
Other Bar Chart Representations.....	230
<b>Pie Charts .....</b>	<b>231</b>
Pie Charts with a pandas Dataframe .....	234
<b>Advanced Charts .....</b>	<b>235</b>
Contour Plots .....	235
Polar Charts .....	236
<b>The mplot3d Toolkit .....</b>	<b>237</b>
3D Surfaces .....	238
Scatter Plots in 3D .....	239
Bar Charts in 3D .....	240
<b>Multipanel Plots.....</b>	<b>241</b>
Display Subplots Within Other Subplots .....	241
Grids of Subplots .....	243
<b>The Seaborn Library .....</b>	<b>245</b>
<b>Conclusions .....</b>	<b>257</b>



■ <b>Chapter 8: Machine Learning with scikit-learn</b> .....	<b>259</b>
The scikit-learn Library .....	259
Machine Learning .....	259
Supervised and Unsupervised Learning .....	259
Training Set and Testing Set .....	260
Supervised Learning with scikit-learn .....	260
The Iris Flower Dataset .....	261
The PCA Decomposition .....	264
K-Nearest Neighbors Classifier .....	267
Diabetes Dataset .....	271
Linear Regression: The Least Square Regression .....	272
Support Vector Machines (SVMs) .....	276
Support Vector Classification (SVC) .....	277
Nonlinear SVC .....	281
Plotting Different SVM Classifiers Using the Iris Dataset .....	283
Support Vector Regression (SVR) .....	285
Conclusions .....	287
■ <b>Chapter 9: Deep Learning with TensorFlow</b> .....	<b>289</b>
Artificial Intelligence, Machine Learning, and Deep Learning .....	289
Artificial Intelligence .....	289
Machine Learning Is a Branch of Artificial Intelligence .....	290
Deep Learning Is a Branch of Machine Learning .....	290
The Relationship Between Artificial Intelligence, Machine Learning, and Deep Learning .....	290
Deep Learning .....	291
Neural Networks and GPUs .....	291
Data Availability: Open Data Source, Internet of Things, and Big Data .....	292
Python .....	292
Deep Learning Python Frameworks .....	292
Artificial Neural Networks .....	293



How Artificial Neural Networks Are Structured.....	293
Single Layer Perceptron (SLP) .....	294
Multilayer Perceptron (MLP) .....	296
Correspondence Between Artificial and Biological Neural Networks .....	297
<b>TensorFlow .....</b>	<b>298</b>
TensorFlow: Google's Framework.....	298
TensorFlow: Data Flow Graph .....	298
<b>Start Programming with TensorFlow .....</b>	<b>299</b>
TensorFlow 2.x vs TensorFlow 1.x .....	299
Installing TensorFlow .....	300
Programming with the Jupyter Notebook.....	300
Tensors .....	300
Loading Data Into a Tensor from a pandas Dataframe .....	303
Loading Data in a Tensor from a CSV File.....	304
Operation on Tensors.....	306
<b>Developing a Deep Learning Model with TensorFlow.....</b>	<b>307</b>
Model Building .....	307
Model Compiling .....	308
Model Training and Testing.....	309
Prediction Making .....	309
<b>Practical Examples with TensorFlow 2.x .....</b>	<b>310</b>
Single Layer Perceptron with TensorFlow .....	310
Multilayer Perceptron (with One Hidden Layer) with TensorFlow .....	317
Multilayer Perceptron (with Two Hidden Layers) with TensorFlow .....	319
<b>Conclusions .....</b>	<b>321</b>
<b>■Chapter 10: An Example—Meteorological Data .....</b>	<b>323</b>
A Hypothesis to Be Tested: The Influence of the Proximity of the Sea .....	323
The System in the Study: The Adriatic Sea and the Po Valley.....	323
Finding the Data Source .....	327
Data Analysis on Jupyter Notebook.....	328



Analysis of Processed Meteorological Data .....	332
The RoseWind .....	343
Calculating the Mean Distribution of the Wind Speed .....	347
Conclusions .....	348
■ <b>Chapter 11: Embedding the JavaScript D3 Library in the IPython Notebook ....</b>	<b>349</b>
The Open Data Source for Demographics .....	349
The JavaScript D3 Library .....	352
Drawing a Clustered Bar Chart.....	355
The Choropleth Maps .....	358
The Choropleth Map of the U.S. Population in 2022 .....	362
Conclusions .....	366
■ <b>Chapter 12: Recognizing Handwritten Digits.....</b>	<b>367</b>
Handwriting Recognition .....	367
Recognizing Handwritten Digits with scikit-learn .....	367
The Digits Dataset .....	368
Learning and Predicting .....	370
Recognizing Handwritten Digits with TensorFlow .....	372
Learning and Predicting with an SLP .....	376
Learning and Predicting with an MLP .....	381
Conclusions .....	384
■ <b>Chapter 13: Textual Data Analysis with NLTK.....</b>	<b>385</b>
Text Analysis Techniques .....	385
The Natural Language Toolkit (NLTK) .....	386
Import the NLTK Library and the NLTK Downloader Tool.....	386
Search for a Word with NLTK .....	389
Analyze the Frequency of Words .....	390
Select Words from Text.....	392
Bigrams and Collocations.....	393
Preprocessing Steps.....	394



Use Text on the Network.....	397
Extract the Text from the HTML Pages.....	398
Sentiment Analysis .....	399
Conclusions .....	401
■ <b>Chapter 14: Image Analysis and Computer Vision with OpenCV .....</b>	<b>403</b>
Image Analysis and Computer Vision .....	403
OpenCV and Python .....	404
OpenCV and Deep Learning.....	404
Installing OpenCV .....	404
First Approaches to Image Processing and Analysis.....	404
Before Starting .....	404
Load and Display an Image .....	405
Work with Images.....	406
Save the New Image.....	407
Elementary Operations on Images.....	407
Image Blending.....	411
Image Analysis .....	412
Edge Detection and Image Gradient Analysis.....	413
Edge Detection .....	413
The Image Gradient Theory.....	413
A Practical Example of Edge Detection with the Image Gradient Analysis .....	415
A Deep Learning Example: Face Detection.....	420
Conclusions .....	422
■ <b>Appendix A: Writing Mathematical Expressions with LaTeX .....</b>	<b>423</b>
■ <b>Appendix B: Open Data Sources .....</b>	<b>435</b>
<b>Index.....</b>	<b>437</b>