

O'REILLY®

第2版



Python for Finance

Python 金融大数据分析 (影印版)

東南大學出版社

Yves Hilpisch 著

第2版

Python金融大数据分析 (影印版)

Python for Finance

Yves Hilpisch 著



Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

O'Reilly Media, Inc. 授权东南大学出版社出版

南京 东南大学出版社

图书在版编目(CIP)数据

Python 金融大数据分析:第2版:英文/(德)伊夫·希尔皮斯科(Yves Hilpisch)著. —影印本. —南京:东南大学出版社,2019.6

书名原文:Python for Finance, 2nd Edition

ISBN 978-7-5641-8372-1

I. ①P… II. ①伊… III. ①软件工具—程序设计—应用—金融—分析—英文 IV. ①F830.41-39

中国版本图书馆 CIP 数据核字(2019)第 074539 号

图字:10-2019-062 号

© 2018 by O'Reilly Media, Inc.

Reprint of the English Edition, jointly published by O'Reilly Media, Inc. and Southeast University Press, 2019. Authorized reprint of the original English edition, 2019 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版 2019。

英文影印版由东南大学出版社出版 2019。此影印版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

Python 金融大数据分析 第2版(影印版)

出版发行:东南大学出版社

地 址:南京四牌楼2号 邮编:210096

出 版 人:江建中

网 址:<http://www.seupress.com>

电子邮件:press@seupress.com

印 刷:常州市武进第三印刷有限公司

开 本:787毫米×980毫米 16开本

印 张:44.75

字 数:876千字

版 次:2019年6月第1版

印 次:2019年6月第1次印刷

书 号:ISBN 978-7-5641-8372-1

定 价:158.00元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Preface

These days, Python is undoubtedly one of the major strategic technology platforms in the financial industry. When I started writing the first edition of this book in 2013, I still had many conversations and presentations in which I argued relentlessly for Python's competitive advantages in finance over other languages and platforms. Toward the end of 2018, this is not a question anymore: financial institutions around the world now simply try to make the best use of Python and its powerful ecosystem of data analysis, visualization, and machine learning packages.

Beyond the realm of finance, Python is also often the language of choice in introductory programming courses, such as in computer science programs. Beyond its readable syntax and multiparadigm approach, a major reason for this is that Python has also become a first class citizen in the areas of artificial intelligence (AI), machine learning (ML), and deep learning (DL). Many of the most popular packages and libraries in these areas are either written directly in Python (such as `scikit-learn` for ML) or have Python wrappers available (such as `TensorFlow` for DL).

Finance itself is entering a new era, and two major forces are driving this evolution. The first is the programmatic access to basically all the financial data available—in general, this happens in real time and is what leads to *data-driven finance*. Decades ago, most trading or investment decisions were driven by what traders and portfolio managers could read in the newspaper or learn through personal conversations. Then came terminals that brought financial data in real time to the traders' and portfolio managers' desks via computers and electronic communication. Today, individuals (or teams) can no longer keep up with the vast amounts of financial data generated in even a single minute. Only machines, with their ever-increasing processing speeds and computational power, can keep up with the volume and velocity of financial data. This means, among other things, that most of today's global equities trading volume is driven by algorithms and computers rather than by human traders.

The second major force is the increasing importance of AI in finance. More and more financial institutions try to capitalize on ML and DL algorithms to improve opera-

tions and their trading and investment performances. At the beginning of 2018, the first dedicated book on “financial machine learning” was published, which underscores this trend. Without a doubt, there are more to come. This leads to what might be called *AI-first finance*, where flexible, parameterizable ML and DL algorithms replace traditional financial theory—theory that might be elegant but no longer very useful in the new era of data-driven, AI-first finance.

Python is the right programming language and ecosystem to tackle the challenges of this era of finance. Although this book covers basic ML algorithms for unsupervised and supervised learning (as well as deep neural networks, for instance), the focus is on Python’s data processing and analysis capabilities. To fully account for the importance of AI in finance—now and in the future—another book-length treatment is necessary. However, most of the AI, ML, and DL techniques require such large amounts of data that mastering data-driven finance should come first anyway.

This second edition of *Python for Finance* is more of an upgrade than an update. For example, it adds a complete part (Part IV) about algorithmic trading. This topic has recently become quite important in the financial industry, and is also quite popular with retail traders. It also adds a more introductory part (Part II) where fundamental Python programming and data analysis topics are presented before they are applied in later parts of the book. On the other hand, some chapters from the first edition have been deleted completely. For instance, the chapter on web techniques and packages (such as Flask) was dropped because there are more dedicated and focused books about such topics available today.

For the second edition, I tried to cover even more finance-related topics and to focus on Python techniques that are particularly useful for financial data science, algorithmic trading, and computational finance. As in the first edition, the approach is a practical one, in that implementation and illustration come before theoretical details and I generally focus on the big picture rather than the most arcane parameterization options of a certain class, method, or function.

Having described the basic approach for the second edition, it is worth emphasizing that this book is neither an introduction to Python programming nor to finance in general. A vast number of excellent resources are available for both. This book is located at the intersection of these two exciting fields, and assumes that the reader has some background in programming (not necessarily Python) as well as in finance. Such readers learn how to apply Python and its ecosystem to the financial domain.

The Jupyter Notebooks and codes accompanying this book can be accessed and executed via our Quant Platform. You can sign up for free at <http://py4fi.pqp.io>.

My company (The Python Quants) and myself provide many more resources to master Python for financial data science, artificial intelligence, algorithmic trading, and computational finance. You can start by visiting the following sites:

- Our company website (<http://tpq.io>)
- My private website (<http://hilpisch.com>)
- Our Python books website (<http://books.tpq.io>)
- Our online training website (<http://training.tpq.io>)
- The Certificate Program website (<http://certificate.tpq.io>)

From all the offerings that we have created over the last few years, I am most proud of our *Certificate Program in Python for Algorithmic Trading*. It provides over 150 hours of live and recorded instruction, over 1,200 pages of documentation, over 5,000 lines of Python code, and over 50 Jupyter Notebooks. The program is offered multiple times per year and we update and improve it with every cohort. The online program is the first of its kind, in that successful delegates obtain an official university certificate in cooperation with htw saar University of Applied Sciences (<http://htwsaar.de>).

In addition, I recently started The AI Machine (<http://aimachine.io>), a new project and company to standardize the deployment of automated, algorithmic trading strategies. With this project, we want to implement in a systematic and scalable fashion what we have been teaching over the years in the field, in order to capitalize on the many opportunities in the algorithmic trading field. Thanks to Python—and data-driven and AI-first finance—this project is possible these days even for a smaller team like ours.

I closed the preface for the first edition with the following words:

I am really excited that Python has established itself as an important technology in the financial industry. I am also sure that it will play an even more important role there in the future, in fields like derivatives and risk analytics or high performance computing. My hope is that this book will help professionals, researchers, and students alike make the most of Python when facing the challenges of this fascinating field.

When I wrote these lines in 2014, I couldn't have predicted how important Python would become in finance. In 2018, I am even happier that my expectations and hopes have been so greatly surpassed. Maybe the first edition of the book played a small part in this. In any case, a big thank you is in order to all the relentless open source developers out there, without whom the success story of Python couldn't have been written.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic

Indicates new terms, URLs, and email addresses.

Constant width

Used for program listings, as well as within paragraphs to refer to software packages, programming languages, file extensions, filenames, program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

Constant width italic

Shows text that should be replaced with user-supplied values or by values determined by context.



This element signifies a tip or suggestion.



This element signifies a general note.



This element indicates a warning or caution.

Using Code Examples

Supplemental material (in particular, Jupyter Notebooks and Python scripts/modules) is available for usage and download at <http://py4fi.pqp.io>.

This book is here to help you get your job done. In general, if example code is offered with this book, you may use it in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: “*Python for Finance*, 2nd Edition, by Yves Hilpisch (O'Reilly). Copyright 2019 Yves Hilpisch, 978-1-492-02433-0.”

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at permissions@oreilly.com.

O'Reilly Safari



Safari (formerly Safari Books Online) is a membership-based training and reference platform for enterprise, government, educators, and individuals.

Members have access to thousands of books, training videos, Learning Paths, interactive tutorials, and curated playlists from over 250 publishers, including O'Reilly Media, Harvard Business Review, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Adobe, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, and Course Technology, among others.

For more information, please visit <http://oreilly.com/safari>.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at <http://bit.ly/python-finance-2e>.

To comment or ask technical questions about this book, send email to bookquestions@oreilly.com.

For more information about our books, courses, conferences, and news, see our website at <http://www.oreilly.com>.

Find us on Facebook: <http://facebook.com/oreilly>

Follow us on Twitter: <http://twitter.com/oreillymedia>

Watch us on YouTube: <http://www.youtube.com/oreillymedia>

Acknowledgments

I want to thank all those who helped to make this book a reality—in particular, the team at O’Reilly, who really improved my manuscript in many ways. I would like to thank the tech reviewers, Hugh Brown and Jake VanderPlas. The book benefited from their valuable feedback and their many suggestions. Any remaining errors, of course, are mine.

Michael Schwed, with whom I have been working closely for more than ten years, deserves a special thank you. Over the years, I have benefited in innumerable ways from his work, support, and Python know-how.

I also want to thank Jason Ramchandani and Jorge Santos of Refinitiv (formerly Thomson Reuters) for their continued support not only of my work but also of the open source community in general.

As with the first edition, the second edition of this book has tremendously benefited from the dozens of “Python for finance” talks I have given over the years, as well as the hundreds of hours of “Python for finance” trainings. In many cases the feedback from participants helped to improve my training materials, which often ended up as chapters or sections in this book.

Writing the first edition took me about a year. Overall, writing and upgrading the second edition also took about a year, which was quite a bit longer than I expected. This is mainly because the topic itself keeps me very busy travel- and business-wise, which I am very grateful for.

Writing books requires many hours in solitude and such hours cannot be spent with the family. Therefore, thank you to Sandra, Lilli, Henry, Adolf, Petra, and Heinz for all your understanding and support—not only with regard to writing this book.

I dedicate the second edition of this book, as the first one, to my lovely, strong, and compassionate wife Sandra. She has given new meaning over the years to what family is really about. Thank you.

— Yves
Saarland, November 2018

Table of Contents

Preface.....	xiii
--------------	------

Part I. Python and Finance

1. Why Python for Finance.....	3
The Python Programming Language	3
A Brief History of Python	5
The Python Ecosystem	6
The Python User Spectrum	7
The Scientific Stack	8
Technology in Finance	9
Technology Spending	9
Technology as Enabler	10
Technology and Talent as Barriers to Entry	11
Ever-Increasing Speeds, Frequencies, and Data Volumes	11
The Rise of Real-Time Analytics	13
Python for Finance	14
Finance and Python Syntax	14
Efficiency and Productivity Through Python	18
From Prototyping to Production	23
Data-Driven and AI-First Finance	24
Data-Driven Finance	24
AI-First Finance	28
Conclusion	31
Further Resources	31

2. Python Infrastructure.....	33
conda as a Package Manager	35
Installing Miniconda	35
Basic Operations with conda	37
conda as a Virtual Environment Manager	41
Using Docker Containers	45
Docker Images and Containers	45
Building an Ubuntu and Python Docker Image	46
Using Cloud Instances	50
RSA Public and Private Keys	51
Jupyter Notebook Configuration File	52
Installation Script for Python and Jupyter Notebook	53
Script to Orchestrate the Droplet Setup	55
Conclusion	56
Further Resources	57

Part II. Mastering the Basics

3. Data Types and Structures.....	61
Basic Data Types	62
Integers	62
Floats	63
Booleans	66
Strings	69
Excursion: Printing and String Replacements	71
Excursion: Regular Expressions	74
Basic Data Structures	75
Tuples	75
Lists	76
Excursion: Control Structures	78
Excursion: Functional Programming	80
Dicts	81
Sets	82
Conclusion	84
Further Resources	84
4. Numerical Computing with NumPy.....	85
Arrays of Data	86
Arrays with Python Lists	86
The Python array Class	88
Regular NumPy Arrays	90

The Basics	90
Multiple Dimensions	94
Metainformation	97
Reshaping and Resizing	98
Boolean Arrays	101
Speed Comparison	103
Structured NumPy Arrays	105
Vectorization of Code	106
Basic Vectorization	107
Memory Layout	110
Conclusion	112
Further Resources	112
5. Data Analysis with pandas.....	113
The DataFrame Class	114
First Steps with the DataFrame Class	114
Second Steps with the DataFrame Class	119
Basic Analytics	123
Basic Visualization	126
The Series Class	128
GroupBy Operations	130
Complex Selection	132
Concatenation, Joining, and Merging	135
Concatenation	136
Joining	137
Merging	139
Performance Aspects	141
Conclusion	143
Further Reading	143
6. Object-Oriented Programming.....	145
A Look at Python Objects	149
int	149
list	150
ndarray	151
DataFrame	152
Basics of Python Classes	154
Python Data Model	159
The Vector Class	163
Conclusion	164
Further Resources	164

Part III. Financial Data Science

7. Data Visualization.....	167
Static 2D Plotting	168
One-Dimensional Data Sets	169
Two-Dimensional Data Sets	176
Other Plot Styles	183
Static 3D Plotting	191
Interactive 2D Plotting	195
Basic Plots	195
Financial Plots	199
Conclusion	203
Further Resources	204
8. Financial Time Series.....	205
Financial Data	206
Data Import	206
Summary Statistics	210
Changes over Time	212
Resampling	215
Rolling Statistics	217
An Overview	218
A Technical Analysis Example	220
Correlation Analysis	222
The Data	222
Logarithmic Returns	224
OLS Regression	226
Correlation	227
High-Frequency Data	228
Conclusion	230
Further Resources	230
9. Input/Output Operations.....	231
Basic I/O with Python	232
Writing Objects to Disk	232
Reading and Writing Text Files	236
Working with SQL Databases	239
Writing and Reading NumPy Arrays	242
I/O with pandas	245
Working with SQL Databases	246
From SQL to pandas	247
Working with CSV Files	250

Working with Excel Files	252
I/O with PyTables	253
Working with Tables	254
Working with Compressed Tables	261
Working with Arrays	263
Out-of-Memory Computations	265
I/O with TsTables	268
Sample Data	268
Data Storage	270
Data Retrieval	271
Conclusion	273
Further Resources	274
10. Performance Python.....	275
Loops	276
Python	277
NumPy	278
Numba	279
Cython	280
Algorithms	282
Prime Numbers	282
Fibonacci Numbers	286
The Number Pi	290
Binomial Trees	294
Python	294
NumPy	296
Numba	297
Cython	298
Monte Carlo Simulation	299
Python	300
NumPy	302
Numba	303
Cython	303
Multiprocessing	304
Recursive pandas Algorithm	305
Python	306
Numba	308
Cython	308
Conclusion	309
Further Resources	310

11. Mathematical Tools.....	311
Approximation	312
Regression	313
Interpolation	324
Convex Optimization	328
Global Optimization	329
Local Optimization	331
Constrained Optimization	332
Integration	334
Numerical Integration	336
Integration by Simulation	337
Symbolic Computation	337
Basics	338
Equations	340
Integration and Differentiation	340
Differentiation	341
Conclusion	343
Further Resources	343
12. Stochastics.....	345
Random Numbers	346
Simulation	352
Random Variables	353
Stochastic Processes	356
Variance Reduction	372
Valuation	375
European Options	376
American Options	380
Risk Measures	383
Value-at-Risk	383
Credit Valuation Adjustments	388
Python Script	392
Conclusion	394
Further Resources	395
13. Statistics.....	397
Normality Tests	398
Benchmark Case	399
Real-World Data	409
Portfolio Optimization	415
The Data	416
The Basic Theory	417

Optimal Portfolios	421
Efficient Frontier	424
Capital Market Line	425
Bayesian Statistics	429
Bayes' Formula	429
Bayesian Regression	430
Two Financial Instruments	435
Updating Estimates over Time	439
Machine Learning	444
Unsupervised Learning	444
Supervised Learning	448
Conclusion	462
Further Resources	463

Part IV. Algorithmic Trading

14. The FXCM Trading Platform.....	467
Getting Started	469
Retrieving Data	469
Retrieving Tick Data	470
Retrieving Candles Data	472
Working with the API	474
Retrieving Historical Data	475
Retrieving Streaming Data	477
Placing Orders	478
Account Information	480
Conclusion	480
Further Resources	481
15. Trading Strategies.....	483
Simple Moving Averages	484
Data Import	485
Trading Strategy	485
Vectorized Backtesting	487
Optimization	489
Random Walk Hypothesis	491
Linear OLS Regression	494
The Data	495
Regression	497
Clustering	499
Frequency Approach	501

Classification	504
Two Binary Features	504
Five Binary Features	506
Five Digitized Features	508
Sequential Train-Test Split	509
Randomized Train-Test Split	511
Deep Neural Networks	512
DNNs with scikit-learn	512
DNNs with TensorFlow	515
Conclusion	519
Further Resources	519
16. Automated Trading.....	521
Capital Management	522
The Kelly Criterion in a Binomial Setting	522
The Kelly Criterion for Stocks and Indices	527
ML-Based Trading Strategy	532
Vectorized Backtesting	533
Optimal Leverage	537
Risk Analysis	539
Persisting the Model Object	543
Online Algorithm	544
Infrastructure and Deployment	546
Logging and Monitoring	547
Conclusion	550
Python Scripts	550
Automated Trading Strategy	550
Strategy Monitoring	553
Further Resources	554

Part V. Derivatives Analytics

17. Valuation Framework.....	557
Fundamental Theorem of Asset Pricing	558
A Simple Example	558
The General Results	559
Risk-Neutral Discounting	560
Modeling and Handling Dates	561
Constant Short Rate	563
Market Environments	565
Conclusion	568