

Claudiu VINȚE

Titus Felix FURTUNĂ

Python pentru analiza datelor

Colecția
Informatică

**Editura ASE
București
2020**



Academia de Studii Economice din București

Copyright © 2020, Editura ASE

Toate drepturile asupra acestei ediții sunt rezervate editurii.

Editura ASE

Piața Romană nr. 6, sector 1, București, România

cod 010374

www.ase.ro

www.editura.ase.ro

editura@ase.ro

Descrierea CIP a Bibliotecii Naționale a României

VINȚE, CLAUDIU

Python pentru analiza datelor / Claudiu Vințe, Titus Felix Furtună. -

București : Editura ASE, 2020

Conține bibliografie

ISBN 978-606-34-0361-3

I. Furtună, Titus Felix

004

Editura ASE

Redactor: Luiza Constantinescu

Coperta: Violeta Rogoian

Autorii își asumă întreaga responsabilitate pentru: ideile exprimate, corectitudinea științifică, originalitatea materialului și sursele bibliografice menționate.

Copierea și utilizarea codului Python prezentat în această lucrare este permisă doar însoțită de menționarea sursei..

Celor pentru care aparentul nu este suficient.

Autorii

Cuprins

1	Scurtă introducere în mediul de programare Python	13
1.1	Instalare și configurare	13
1.2	Configurare PATH	13
1.3	Variabile de mediu Python	14
1.4	Moduri de rulare a interpretorului Python	15
1.4.1	Interpretor interactiv	15
1.4.2	Lansare script din linia de comandă	16
1.4.3	Utilizarea unui mediu de dezvoltare integrat	16
2	Elemente de bază ale limbajului Python	17
2.1	Sintaxa generală a limbajului Python	18
2.1.1	Identificatori Python	18
2.1.2	Cuvinte rezervate	19
2.1.3	Linii de cod și indentare	19
2.1.4	Declararea unei linii logice pe mai multe linii fizice	20
2.1.5	Utilizarea citărilor în Python	20
2.1.6	Comentarii în Python	21
2.1.7	Utilizarea liniilor libere	21
2.1.8	Declarații multiple pe o singură linie	22
2.1.9	Gruparea declarațiilor în blocuri	22
2.2	Tipuri de variabile	22
2.2.1	Atribuirea de valori variabilelor	22
2.2.2	Atribuire multiplă	23
2.2.3	Tipuri de date standard	23
2.2.4	Testarea valorii de adevăr	24
2.2.5	Comparații	24
2.2.6	Tipuri de date numerice	25
2.2.7	Șiruri de caractere	27
2.2.7.1	Accesarea valorilor în șiruri	27
2.2.7.2	Caractere de control al tipăririi	28
2.2.7.3	Citarea triplă	29
2.2.7.4	Modificarea șirurilor de caractere	30
2.2.7.5	Operatori speciali pentru șiruri de caractere	30
2.2.7.6	Operatori de formatare a șirurilor de caractere	31
2.2.7.7	Șiruri Unicode	32
2.3	Operatori fundamentali	32
2.3.1	Operatori aritmetici	32
2.3.2	Operatori de comparație	33
2.3.3	Operatori de atribuire	33
2.3.4	Operatori logici pe biți (bitwise)	34
2.4	Structuri de control	35
2.4.1	Structura alternativă	35
2.4.2	Structuri repetitive	36
2.4.2.1	Structura while	36
2.4.2.2	Structura for	36
2.5	Tratarea excepțiilor	36

2.6	Utilizarea funcțiilor în Python	37
2.6.1	Definirea unei funcții.....	37
2.6.2	Pasarea parametrilor în funcții	38
2.6.2.1	Argumente necesare	40
2.6.2.2	Argumente cu cuvinte-cheie.....	40
2.6.2.3	Argumente implicite	41
2.6.2.4	Argumente de lungime variabilă	41
2.6.2.5	Funcțiile anonime (<code>lambda</code>).....	42
2.6.2.6	Instrucțiunea <code>return</code>	43
2.6.3	Decoratori	43
2.7	Tipuri de date secvențiale.....	44
2.7.1	Conceptul de listă în Python	44
2.7.1.1	Accesarea valorilor unei liste	45
2.7.1.2	Actualizarea listelor	45
2.7.1.3	Ștergerea de elemente din listă.....	46
2.7.1.4	Operații de bază cu liste	46
2.7.1.5	Indexare și partiționare	46
2.7.1.6	Funcții și metode implicite pentru liste	47
2.7.2	Conceptul de tuplu.....	48
2.7.2.1	Modificarea tuplurilor	49
2.7.2.2	Ștergerea elementelor unui tuplu	49
2.7.2.3	Funcții implicite pentru tuplu	50
2.7.3	Tipul interval (<code>range</code>).....	51
2.8	Python ca limbaj de programare orientat obiect	52
2.8.1	Crearea obiectelor	53
2.8.2	Metode speciale.....	53
2.8.3	Crearea atributelor	54
2.8.4	Proprietăți	54
2.9	Structuri de date secvențiale modelate cu tipurile implicite Python	55
2.9.1	Vector (<code>array</code>)	55
2.9.1.1	Reprezentare vector (<code>array</code>).....	56
2.9.1.2	Operații de bază cu vectori.....	56
2.9.1.3	Parcurgerea elementelor unui vector	57
2.9.1.4	Accesarea elementelor unui vector.....	57
2.9.1.5	Operațiunea de inserare în vector.....	58
2.9.1.6	Operațiunea de ștergere	58
2.9.1.7	Operațiunea de căutare.....	59
2.9.1.8	Operațiunea de actualizare	59
2.9.2	Stivă (<code>stack</code>)	60
2.9.2.1	Adăugarea sau punerea unui element în stivă – PUSH într-o stivă	60
2.9.2.2	Extragerea sau scoaterea unui element din stivă – POP dintr-o stivă	61
2.9.3	Coadă (<code>queue</code>).....	61
2.9.3.1	Adăugarea de elemente la o coadă – PUT într-o stivă	62
2.9.3.2	Scoaterea elementelor dintr-o coadă.....	62
2.10	Conceptul de dicționar (tipul <code>map</code>).....	63
2.10.1	Accesarea valorilor în dicționar.....	63
2.10.2	Actualizarea dicționarului	64
2.10.3	Ștergerea elementelor unui dicționar.....	64
2.10.4	Proprietățile cheilor de dicționar	65
2.10.5	Funcții și metode implicite pentru tipul dicționar.....	65

2.11	Lucrul cu fișiere	66
2.11.1	Funcții de bază pentru citirea datelor din fișiere	66
2.11.2	Citirea datelor utilizând pachetul pandas	68
3	Lucru cu masive multidimensionale din pachetul numpy	69
3.1	Masive multidimensionale ndarray	69
3.1.1	Funcții de creare	70
3.1.2	Atribute ale masivelor ndarray	72
3.1.3	Indexarea masivelor numpy.ndarray	73
3.1.4	Specificarea dimensiunilor	74
3.2	Operații aritmetice și logice	75
3.3	Funcții universale	76
3.4	Funcții statistice	77
3.5	Funcții pentru inversare, sortare, căutare, selecție	79
3.6	Funcții matematice și de algebră liniară	80
4	Tipuri de date ale pachetului pandas	81
4.1	Serii	81
4.2	Tabele de date (DataFrame)	82
4.3	Modificare, salvare, sortare, agregare și joncțiune în DataFrame	84
4.3.1	Modificarea datelor din DataFrame	84
4.3.2	Salvarea în fișier CSV	85
4.3.3	Sortarea	85
4.3.4	Agregarea datelor	85
4.3.5	Joncțiuni	87
4.4	Date de tip categorial	88
4.5	Clase utilitare	89
5	Elemente de reprezentare grafică în Python	91
5.1	Adăugarea de subgrafice	92
5.2	Tipuri de grafice în matplotlib	92
5.2.1	Grafice linie – matplotlib.plot	92
5.2.2	Grafice puncte – matplotlib.scatter	94
5.2.3	Grafice puncte seaborn.scatterplot	95
5.2.4	Grafice corelogramă – seaborn.heatmap	96
5.2.5	Histograme matplotlib	97
5.2.6	Histograme pandas	98
5.2.7	Reprezentări grafice utilizate frecvent în analiza datelor	99
5.2.7.1	Cercul corelațiilor	100
5.2.7.2	Graficul varianței explicate de componentele principale	103
5.2.7.3	Graficul corelogramă	104
5.2.7.4	Graficul histogramă cu bare multiple	106
6	Implementarea unor metode de analiză a datelor în Python	108
6.1	Analiza în componente principale (ACP)	108
6.1.1	Datele de prelucrat și etapele analizei	108
6.1.2	ACP – studiu de caz	110
6.1.2.1	Setul de date	110
6.1.2.2	Abordarea ACP în Python	112
6.1.2.3	Interpretarea rezultatelor	117
6.2	Analiza exploratorie a factorilor (AEF)	126
6.2.1	Datele de prelucrat	126
6.2.2	Ipoteze modelului	127

6.2.3	Estimarea existenței factorilor	128
6.2.4	Estimarea parametrilor modelului. Extragerea factorilor	129
6.2.5	Metoda probabilității maxime	130
6.2.6	Metoda analizei componentelor principale	130
6.2.7	Estimarea numărului de factori. Testul Bartlett	131
6.2.8	AEF – studiu de caz	133
6.3	Analiza corelațiilor canonice (ACC)	141
6.3.1	Datele de prelucrat	142
6.3.2	Etapele analizei	142
6.3.3	Factorii canonici	145
6.3.4	Legăturile dintre factori	146
6.3.5	Varianța explicată și redundanța informațională	147
6.3.6	Standardizarea factorilor canonici	148
6.3.7	Relevanța rădăcinilor canonice. Testul de relevanță Bartlett χ^2	148
6.3.8	ACC – studiu de caz	149
6.4	Analiza discriminantă (AD)	158
6.4.1	Organizarea datelor și notații	158
6.4.2	Indicatori de variabilitate și împrăștiere	159
6.4.3	Semnificația modelului. Teste statistice	161
6.4.4	Analiza discriminantă liniară (Linear Discriminant Analysis) – funcțiile Fisher	163
6.4.5	Analiza discriminantă: caz particular al analizei canonice	167
6.4.6	Funcții de clasificare	168
6.4.7	Discriminarea bayesiană	170
6.4.7.1	Probabilitate totală și probabilitate bayesiană	170
6.4.7.2	Clasificatorul bayesian	171
6.4.7.3	Estimarea neparametrică a probabilităților condiționate. Metoda histogramei	172
6.4.7.4	Metode parametrice	173
6.4.8	Analiza discriminantă – studiu de caz	174
6.4.8.1	Setul de date utilizat	174
6.4.8.2	Investigarea eșantionului de bază – construirea modelului	175
6.4.8.3	Implementarea Python a analizei discriminate	175
6.4.8.4	Prezentarea și interpretarea rezultatelor	178
6.4.8.5	Clasificarea instanțelor din eșantionul de test – aplicarea modelului	182
6.5	Analiza de cluster (AC)	183
6.5.1	Algoritmi ierarhici	184
6.5.2	Metode de grupare ierarhică	186
6.5.3	Distanțe utilizate pentru calculul proximității	187
6.5.4	Analiza de cluster – studiu de caz	188
6.5.4.1	Implementarea analizei de cluster în Python	189
6.5.4.2	Rezultatele obținute și interpretarea acestora	191
6.5.4.3	Clasificarea variabilelor	194

Bibliografie	201
---------------------------	------------

Anexe

Anexa 1	205
Anexa 2	213