



Univerzitet u Istočnom Sarajevu
Filozofski fakultet



PROGRAMSKI JEZICI R I PYTHON

Jovana Forcan

Univerzitet u Istočnom Sarajevu - Filozofski fakultet

Akademska misao

PROGRAMSKI JEZICI R I PYTHON

Autor:

Jovana Forcan

Recenzenti:

Dr Dragoljub Krneta, vanredni profesor

Univerzitet u Istočnom Sarajevu

Filozofski fakultet

Dr Miloš Ljubojević, vanredni profesor

Univerzitet u Banjoj Luci

Elektrotehnički fakultet

Dizajn korica:

Boris Popović

Izdavač:

Univerzitet u Istočnom Sarajevu, Filozofski fakultet

Akademski misao, Beograd

Štampa:

Akademski misao, Beograd

Tiraž:

200 primjeraka

ISBN:

978-86-7466-999-0

Mjesto i godina izdanja:

Beograd, Istočno Sarajevo, 2024.

Odlukom Senata Univerziteta u Istočnom Sarajevu broj 01-C-75-LXIII/24 od 29.2.2024. godine, rukopis „Programski jezici R i Python” autora Jovane Forcan, odobren je za objavljivanje kao univerzitetski udžbenik.

*S ljubavlju i zahvalnošću
mojim dragim roditeljima, koji su mi uvijek bili oslonac*

Predgovor

R i Pajton (*Python*) su dva istaknuta programska jezika u domenu analize podataka i statističkih istraživanja. Oba jezika se odlikuju širokim spektrom naprednih biblioteka i alata, čime omogućavaju jednostavno i efikasno rješavanje složenih analitičkih izazova u svijetu analize podataka. Ovi izazovi uključuju statističko modelovanje, mašinsko učenje i vizualizaciju podataka.

Ovaj udžbenik kreiran je sa ciljem da pruži sveobuhvatan pregled programskih jezika R i Pajton. Namijenjen je kako početnicima, tako i iskusnim korisnicima. Udžbenik je strukturiran u poglavlja koja pokrivaju različite aspekte analize podataka, prateći korake od osnovnih koncepata R i Pajton jezika prema naprednijim temama. Svako poglavlje sadrži brojne primjere kako bi koncepti i tehnike bili ilustrovani i omogućili čitaocima bolje razumijevanje.

Sadržaj udžbenika je organizovan u dva dijela.

Prvi dio udžbenika detaljno istražuje karakteristike programskog jezika R. Ovdje se obrađuju teme kao što su strukture podataka, manipulacija podacima, vizualizacija i statističko modelovanje. Čitalac će se takođe upoznati sa nekim od najpopularnijih R paketa i biblioteka, uključujući `dplyr`, `ggplot2` i druge, uz praktične smjernice o njihovoj primjeni za različite zadatke analize podataka. Prvi dio udžbenika sastoji se od 6 poglavlja.

- Uvod u R: ovo poglavlje pruža osnovni uvod u programski jezik R, njegovo integrirano razvojno okruženje, proces preuzimanja i instalacije R-a i RStudio-a, te uvodi različite programske paradigme.
- Osnove programiranja u R-u: u ovom poglavlju razmatra se sintaksa R-a, osnovne operacije, rad sa stringovima, upotreba paketa i različite strukture podataka. Takođe se obrađuje kontrola toka programa, rad sa funkcijama, te koncepti objektno orijentisanog programiranja.
- Obrada podataka: ovo poglavlje fokusira se na upotrebu paketa `dplyr` i `data.table` za obradu podataka, kao i na različite pakete za povezivanje R-a sa bazom podataka.

- Grafika u R-u: poglavlje razmatra kreiranje različitih vrsta grafikona pomoću paketa graphics i ggplot2.
- Uvod u statistiku: u ovom poglavlju detaljno se istražuju koncepti deskriptivne i inferencijalne statistike, sa posebnim fokusom na njihovu primjenu unutar programskog jezika R.
- Regresiona analiza: poglavlje razmatra linearnu i nelinearnu regresiju, istražujući postupke i tehnike za izradu efikasnih regresionih modela unutar R-a.

Drugi dio udžbenika fokusira se na osnovne karakteristike jezika Pajton i obuhvata najpopularnije Pajton biblioteke i module, uključujući NumPy, Pandas, Matplotlib i Scikit-Learn, pružajući korisne smjernice za njihovu primjenu u različitim analitičkim zadacima. Drugi dio udžbenika sastoji se od 5 poglavlja.

- Uvod u Pajton: ovo poglavlje pruža uvod u Pajton, njegovu instalaciju i predstavlja platforme Anaconda i Google Colab, koje olakšavaju rad sa Pajtonom.
- Kurs programiranja u Pajtonu: ovo poglavlje obuhvata različite tipove podataka u Pajtonu, kontrolu toka programa (if-else, petlje), module, funkcije, rad sa izuzecima, objektno orijentisanu paradigmu, rad sa datotekama i povezivanje sa bazom podataka.
- Moduli NumPy, Pandas i SciPy: poglavlje obrađuje tri ključna modula u Pajtonu koji se široko koriste u naučnim i analitičkim aplikacijama.
- Grafika u Pajtonu: poglavlje se bavi bibliotekama Matplotlib, Seaborn i drugim alatima za vizualizaciju podataka.
- Regresiona analiza i osnove klasifikacije u Pajtonu: u ovom poglavlju istražuju se ključne tehnike analize podataka koje se koriste za modelovanje i predviđanje.

Glavni cilj ovog udžbenika je osigurati čvrsto razumijevanje jezika R i Pajton i omogućiti čitaocima efikasno korišćenje ovih jezika za raznovrsne analitičke zadatke povezane sa podacima. Knjiga je namijenjena studentima prirodnih, računarskih i tehničkih nauka, kao i studentima ekonomije i društvenih nauka. Takođe je idealna za one koji pohađaju interdisciplinarne programe koji kombinuju elemente matematike, informatike, statistike, nauke o podacima i drugih disciplina. Ova knjiga pruža uvide i vještine koje su primjenljive u različitim oblastima, čineći

je idealnim izvorom za raznovrsnu akademsku zajednicu, posebno za one koji teže da integrišu napredne analitičke i programerske vještine u svoje akademsko i profesionalno usavršavanje.

Zahvaljujem se profesoru dr Dragoljubu Krneti i profesoru dr Milošu Ljubojeviću na angažovanju i stručnosti koje su unijeli u recenziju mog djela, čime su značajno doprinijeli njegovom kvalitetu.

Sadržaj

I	R	1
1	Uvod u R	3
1.1	Šta predstavlja R?	3
1.2	Šta je integrisano razvojno okruženje (IRO) R-a?	4
1.3	Preuzimanje i instalacija R-a	5
1.4	Preuzimanje i instalacija RStudio-a	6
1.4.1	Desktop verzija	6
1.4.2	RStudio Cloud	8
1.5	Programske paradigme	8
2	Osnove programiranja u R-u	13
2.1	Sintaksa programskog jezika R	13
2.1.1	Osnovne operacije u R-u	16
2.1.2	String	20
2.2	Paketi	22
2.3	Strukture podataka	23
2.3.1	Vektori	23
2.3.2	Liste	26
2.3.3	Matrice	29
2.3.4	Višedimenzionalni nizovi	32
2.3.5	Faktori	33
2.3.6	Okvir podataka – Data frame	34
2.3.7	Datum i vrijeme - Datetime	44
2.4	Kontrola toka	46
2.4.1	If-Then-Else	47
2.4.2	Naredbe <code>break</code> i <code>next</code>	48
2.4.3	Funkcija <code>ifelse()</code>	49
2.4.4	Petlje	49
2.5	Funkcije	51
2.5.1	Porodica funkcija <code>apply</code>	53
2.5.2	Rad sa izuzecima u R-u	54

2.6	OO paradigma programskog jezika R	56
3	Obrada podataka	65
3.1	Paket dplyr	65
3.1.1	Funkcija <code>select</code>	66
3.1.2	Funkcija <code>filter</code>	67
3.1.3	Funkcija <code>slice</code>	67
3.1.4	Funkcija <code>arrange</code>	68
3.1.5	Funkcija <code>mutate</code>	68
3.1.6	Funkcija <code>reallocate</code>	69
3.1.7	Funkcije <code>summarize</code> i <code>group_by</code>	69
3.2	Spajanje tabela - <code>join</code>	70
3.3	Povezivanje sa podacima iz baza podataka	73
3.3.1	Povezivanje sa bazom podataka pomoću paketa RODB	74
3.3.2	Povezivanje sa bazom podataka pomoću paketa DBI	78
3.3.3	Povezivanje sa bazom podataka pomoću paketa RJDBC	82
4	Grafika u R-u	85
4.1	Paket <code>graphics</code>	85
4.1.1	Funkcija <code>plot</code>	85
4.1.2	Pita grafikoni u paketu <code>graphics</code>	88
4.1.3	Stubičasti grafikoni u paketu <code>graphics</code>	89
4.1.4	Histogrami u paketu <code>graphics</code>	90
4.2	Paket <code>ggplot2</code>	91
4.2.1	Dijagrami raspršenosti u paketu <code>ggplot2</code>	94
4.2.2	Dijagrami kutija: <code>graphics</code> i <code>ggplot2</code>	97
4.2.3	Histogrami: <code>graphics</code> i <code>ggplot2</code>	100
4.2.4	Stubičasti grafikoni: <code>graphics</code> i <code>ggplot2</code>	102
4.2.5	Pita grafikoni u paketu <code>ggplot2</code>	105
4.2.6	Linijski grafikoni u paketu <code>ggplot2</code>	107
4.2.7	Krive gustine raspodjele: <code>graphics</code> i <code>ggplot2</code>	109
4.2.8	Toplotne mape u paketu <code>ggplot2</code>	112
4.2.9	Dodavanje zaglađivača	115
4.3	Izvoz grafikona	120
5	Uvod u statistiku	121
5.1	Deskriptivna statistika	122
5.2	Inferencijalna statistika	129
5.2.1	Statistički testovi	130

6	Regresiona analiza	141
6.1	Linearna regresija	141
6.1.1	Jednostavna linearna regresija u R-u	144
6.1.2	Unakrsna validacija (Cross-validation)	147
6.1.3	Višestruka linearna regresija	150
6.2	Pretpostavke linearnog modela	155
6.3	Linearna regresija sa kategoričkim promjenljivim	160
6.4	Nelinearna regresija	165
6.4.1	Polinomijalna regresija	167
6.4.2	Logaritamska transformacija	169
6.4.3	Segmentirana (splajn) regresija	170
6.4.4	Generalizovani aditivni model	172
II	Pajton	175
7	Uvod u Pajton	177
7.1	Instalacija Pajtona	178
7.1.1	Anakonda	179
7.1.2	Google Colab	181
7.2	PIP	182
8	Kurs programiranja u Pajtonu	183
8.1	Tipovi podataka	184
8.1.1	Stringovi	186
8.1.2	Liste	189
8.1.3	Torke	191
8.1.4	Skupovi	192
8.1.5	Rječnici	193
8.2	Operatori	195
8.3	Kontrola toka programa	198
8.3.1	If-Else	198
8.3.2	Petlje	198
8.4	Moduli	200
8.4.1	Modul <code>datetime</code>	201
8.5	Izuzeci	203
8.6	Funkcije	205
8.6.1	Lambda funkcija	206
8.6.2	Funkcija <code>input()</code>	207
8.7	OO paradigma Pajtona	208
8.8	Rad sa datotekama	212
8.8.1	Rad sa binarnim datotekama	214

8.8.2	Rad sa .csv datotekama	215
8.8.3	Rad sa JSON datotekama	216
8.9	Povezivanje sa bazama podataka u Pajtonu	216
9	Moduli NumPy, Pandas i SciPy	223
9.1	NumPy	223
9.1.1	ndarray	223
9.1.2	Generisanje slučajnih brojeva	234
9.1.3	NumPy ufuncs	237
9.2	Pandas	238
9.2.1	Rad sa datotekama	243
9.2.2	Čišćenje podataka	244
9.2.3	Korelacija podataka	251
9.3	SciPy	253
9.3.1	Interpolacija	253
9.3.2	Inferencijalna statistika	258
10	Grafika u Pajtonu	265
10.1	Matplotlib	265
10.1.1	Seaborn	281
11	Regresiona analiza i osnove klasifikacije u Pajtonu	299
11.1	Linearna regresija	300
11.1.1	Višestruka linearna regresija	306
11.1.2	Linearna regresija sa kategoričkim promjenljivim	308
11.2	Logistička regresija	313
11.3	Polinomijalna regresija	321
11.4	Metoda k-najbližih susjeda	324
	Literatura	335

Dio I

R

Poglavlje 1

Uvod u R

R je moćan i svestran programski jezik sa širokom primjenom u analizi podataka, statističkom modelovanju i vizualizaciji podataka. Njegova funkcionalnost obuhvata obilje ugrađenih funkcija, opsežan broj paketa i aktivnu zajednicu korisnika i programera koji neprekidno doprinose njegovom razvoju i održavanju. To ga čini popularnim izborom među stručnjacima u oblasti analize podataka i statističarima. Pored toga, R se ističe svojom paradigmom funkcionalnog programiranja, obiljem grafičkih mogućnosti i širokim spektrom primjena u različitim oblastima, što ga čini neizostavnim alatom za rješavanje različitih izazova.

1.1 Šta predstavlja R?

R je programski jezik, interpreter i platforma.

R, kao programski jezik omogućava pisanje programa za različite svrhe.

R, kao interpretirani jezik omogućava izvršavanje R programa liniju po liniju iz konzole ili pokretanje skripata iz .R datoteka pomoću ugrađenog interpretera.

R, kao platforma otvorenog koda pruža bogat skup alata za manipulaciju podacima, izvođenje statističkih proračuna i vizualizaciju rezultata.

R je kompatibilan s operativnim sistemima poput Windows-a, Mac OS-a i raznih UNIX platformi, uključujući Linux. Većina platformi distribuira R u obliku binarnih datoteka kako bi se olakšala instalacija. Ovaj softverski projekat prvobitno su pokrenuli Ros Ihak (*Ross Ihak*) i Robert Džentlmen (*Robert Gentleman*). R se smatra implementacijom jezika S, kojeg su razvili Rik Beker (*Rick Becker*), Džon Čejmbers (*John Chambers*) i Alan Vilks (*Allan Wilks*).