

**Stelian Stancu**

**DATA SCIENCE  
în mediul R  
Teorie și aplicații**

Colecția  
Cibernetică

**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](http://www.ase.ro)

[www.editura.ase.ro](http://www.editura.ase.ro)

[editura@ase.ro](mailto:editura@ase.ro)

**Descrierea CIP a Bibliotecii Naționale a României**

**STANCU, STELIAN**

**Data science în mediul R : teorie și aplicații / Stelian Stancu. –**

București : Editura ASE, 2020

Conține bibliografie

ISBN 978-606-34-0331-6

311

**Editura ASE**

**Redactor, tehnoredactor și copertă:** Claudia-Marinela Dumitru

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

*Febra rațiunii o regăsim în perioada de tinerețe,  
pentru ca la maturitate să avem suficientă putere  
pentru a ne urma întru totul adevărata rațiune.*

*În viață e mai comod să nu primești decât să dai,  
pentru că mai lesne te hotărăști a nu primi ceva  
decât a ceda ceva ce deja îți aparține.  
Să fie oare așa și în relația cunoaștere-experiență?*

# Cuprins

---

<b>Despre autor .....</b>	<b>13</b>
<b>Cuvânt-înainte .....</b>	<b>15</b>
<b>Capitolul 1. Introducere în Data Science .....</b>	<b>17</b>
1.1. Statistica.....	17
1.1.1. Definirea Statisticii .....	17
1.1.2. Caracterizarea Statisticii .....	17
1.2. Știința Datelor (Data Science).....	18
1.2.1. Definirea științei datelor (Data Science) .....	18
1.2.2. Caracterizarea științei datelor (Data Science) .....	19
1.3. Analiza Datelor (Data Mining).....	20
1.3.1. Definirea analizei datelor (Data Mining) .....	20
1.3.2. Părțile constitutive principale ale CRISP-DM.....	22
1.4. Analiza de text (Text Mining).....	23
1.4.1. Definirea analizei de text (Text Mining).....	23
1.4.2. Părțile constitutive principale ale analizei de text (Text Mining).....	25
1.5. Prelucrarea limbajului natural (NLP – Natural Language Processing).....	26
1.6. Tipuri de analiză (Analytics).....	26
1.7. Big Data .....	27
1.7.1. Caracteristici generale.....	27
1.7.2. Varietatea datelor.....	28
<b>Capitolul 2. Introducere în mediul R .....</b>	<b>29</b>
2.1. Instalarea pachetelor din mediul R.....	30
2.1.1. Instalarea pachetului R.....	30
2.1.2. Pachete R necesare.....	32
2.2. Citirea/Importul datelor în mediul R.....	32
2.2.1. Citirea datelor elementare în R .....	32
2.2.2. Utilizarea <code>scan()</code> pentru a citi/importa în R date stocate într-un fișier.....	33
2.2.3. Utilizarea <code>read.table()</code> pentru a citi/importa în R date stocate într-un fișier.....	34
2.2.4. Citirea/Importul fișierelor Excel în R.....	39
2.2.4.1. Copierea datelor din Excel și citirea/importul acestora în R .....	39
2.2.4.2. Citirea/Importul fișierelor Excel în R, utilizând pachetul <code>readxl</code> .....	40

2.2.4.3. Citirea/Importul fișierelor Excel în R, utilizând pachetul <code>xlsx</code> .....	41
2.2.5. Utilizarea <code>read.csv()</code> și <code>read.csv2()</code> pentru a citi/importa date în R .....	42
2.3. Lucrul cu <code>dataframe</code> în R .....	44
2.3.1. Selecția de părți din <code>dataframe</code> .....	44
2.3.2. Sortarea datelor în R .....	45
2.3.3. Rezumarea ( <code>summary</code> , <code>with</code> , <code>aggregate</code> ) conținutului fișierelor de date .....	46
2.4. Cunoașterea datelor în R .....	48
2.5. Operații cu numere și vectori în R .....	49
2.5.1. Principalele comenzi pentru accesarea datelor, în lucrul cu un vector ..	49
2.5.2. Operații bazate pe funcții R .....	50
2.5.3. Convertirea unui vector în <code>dataframe</code> .....	53
2.5.4. Formatarea rezultatelor într-un tabel <code>dataframe</code> , utilizând pachetul <code>stargazer</code> .....	54
2.6. Operații cu matrice în R .....	56
2.6.1. Principalele comenzi pentru accesarea datelor, în lucrul cu matrice ....	56
2.6.2. Operații bazate pe funcții R .....	58
2.7. Căutarea în R a unui răspuns primit anterior .....	59
2.8. Actualizarea R la o versiune mai nouă .....	59

### **Capitolul 3. Statistici descriptive în mediul R ..... 60**

3.1. Termeni statistici de bază .....	60
3.2. Tipuri de date .....	60
3.3. Indicatori ai tendinței centrale (medie, mediană și mod) .....	60
3.3.1. Media .....	61
3.3.2. Mediana .....	62
3.3.3. Modul .....	62
3.4. Indicatori ai variației datelor .....	63
3.4.1. Amplitudinea absolută a variației ( <code>range</code> ) .....	64
3.4.2. Abaterea intercuantilică .....	65
3.4.3. Varianța (dispersia) .....	65
3.4.4. Abaterea standard (abaterea medie pătratică) .....	66
3.5. Distribuția normală și distribuția binomială .....	68
3.5.1. Distribuția normală .....	68
3.5.1.1. Prezentarea distribuției normale .....	68
3.5.1.2. Testarea normalității unei distribuții .....	70
3.5.1.3. Funcția de distribuție cumulativă (CDF – Cumulative Distribution Function) .....	71
3.5.1.4. Modul unei distribuții .....	72
3.5.1.5. Aprecierea formei unei distribuții – asimetria ( <code>skewness()</code> ) și boltirea (indicele de aplatizare) ( <code>kurtosis()</code> ) .....	73

3.5.2. Distribuția binomială (Bernoulli) .....	76
3.5.2.1. Prezentarea distribuției binomiale .....	76
3.5.2.2. Funcția de distribuție cumulativă (CDF – Cumulative Distribution Function) .....	77
3.6. Rezumarea datelor. Funcțiile <code>summary()</code> și <code>str()</code> .....	78

## Capitolul 4. Vizualizarea datelor în R

### Reprezentare grafică a datelor în mediul R .....79

4.1. Vizualizarea datelor utilizând softul încorporat în mediul R .....	79
4.1.1. Diagrama cu bare/batoane .....	79
4.1.2. Histograma .....	81
4.1.3. Curba densității de probabilitate/repartiție (diagrama tendinței) .....	83
4.1.4. Diagrama de structură (cercul de structură) .....	85
4.1.5. Diagrama prin puncte (scatterplot-ul, norul de puncte sau diagrama împrăștierii) .....	86
4.1.6. Matricea scatterplot (matricea împrăștierii datelor) .....	87
4.1.7. Boxplot-ul .....	88
4.1.8. Bazele reprezentării grafice în cazul analizei rețelelor sociale .....	91
4.2. Vizualizarea datelor utilizând pachetul <code>ggplot2</code> .....	93
4.2.1. Introducere în <code>ggplot2</code> .....	93
4.2.2. Gramatica grafică în <code>ggplot2</code> .....	94
4.2.3. Setarea pentru <code>ggplot2</code> .....	94
4.2.4. Cartografierea (maparea) estetică în <code>ggplot2</code> .....	95
4.2.5. Geometrie în <code>ggplot2</code> .....	95
4.2.6. Etichetarea în <code>ggplot2</code> .....	97
4.2.7. Teme în <code>ggplot2</code> .....	97
4.2.8. Reprezentări grafice în <code>ggplot2</code> .....	99
4.2.8.1. Diagrama cu bare în <code>ggplot2</code> .....	99
4.2.8.2. Histograma în <code>ggplot2</code> .....	100
4.2.8.3. Funcția densitate de probabilitate în <code>ggplot2</code> .....	101
4.2.8.4. Diagrama tendinței .....	101
4.2.8.5. Scatterplot-ul (norul de puncte – diagrama împrăștierii) în <code>ggplot2</code> .....	102
4.2.8.6. Boxplot-ul în <code>ggplot2</code> .....	103
4.3. Diagrame interactive cu <code>plotly</code> și <code>ggplot2</code> .....	105

## Capitolul 5. Inferența statistică și analiza de regresie .....107

5.1. Inferența statistică – aspecte generale .....	107
5.2. Funcțiile <code>apply()</code> , <code>lapply()</code> , <code>sapply()</code> .....	108
5.3. Eșantionarea la nivelul unei populații .....	109
5.3.1. Eșantionarea aleatoare simplă .....	109
5.3.2. Eșantionarea stratificată .....	110

5.3.3. Eșantionarea de tip cluster .....	111
5.4. Covarianța dintre variabile (matricea de covarianță) .....	113
5.5. Corelația dintre variabile (coeficientul de corelație – matricea de corelație) ..	114
5.6. Testarea ipotezelor statistice, în R .....	115
5.6.1. Teoria testelor de semnificație .....	115
5.6.2. Testul $t$ (Student) .....	117
5.6.3. Testul hi-pătrat .....	122
5.7. Analiza de regresie .....	125
5.7.1. Introducere în conceptul de regresie .....	125
5.7.2. Regresia liniară simplă .....	125
5.7.3. Regresia liniară multiplă .....	127

## **Capitolul 6. Analiza în componente principale (ACP), utilizând mediul R.....130**

6.1. Analiza în componente principale – aspecte generale .....	130
6.1.1. Introducere .....	130
6.1.2. Încărcarea pachetelor de bază, necesare suplimentar în mediul R.....	131
6.1.3. Pregătirea datelor .....	131
6.2. Metoda ACP/PCA propriu-zisă .....	133
6.2.1. Valori proprii / vectori proprii / varianțe .....	133
6.2.2. Prezentarea metodei ACP/PCA .....	134
6.2.3. Vizualizarea și interpretarea datelor și rezultatelor obținute .....	138
6.2.4. Funcții ACP/PCA încorporate .....	138
6.2.5. Proporția de variație explicată la nivel de ACP/PCA .....	140
6.2.6. Determinarea numărului optim de componente principale, la nivelul unui model .....	142

## **Capitolul 7. Analiza cluster și algoritmi de clustering, în mediul R ....144**

7.1. Reprezentarea datelor multidimensionale primare .....	144
7.1.1. Matricea observațiilor (obiectelor, indivizilor) .....	144
7.1.2. Matricea sau tabelul de contingență .....	145
7.1.3. Matricea sau tabelul de proximitate .....	146
7.2. Analiza cluster – aspecte generale .....	147
7.2.1. Introducere .....	147
7.2.2. Scopul analizei cluster .....	147
7.2.3. Încărcarea pachetelor de bază în mediul R .....	148
7.3. Metoda k-mean clustering sau algoritmul centrului mobil .....	149
7.3.1. Aspecte generale .....	149
7.3.2. Pregătirea datelor .....	149
7.3.3. Metoda propriu-zisă .....	152
7.3.4. Măsurarea calității unei partiții k-mean .....	156
7.4. Metode de grupare ierarhică (Hierarchical Clustering Algorithms) .....	157
7.4.1. Aspecte generale .....	157

---

7.3.2. Pregătirea datelor .....	158
7.4.3. Metoda propriu-zisă .....	160
7.4.4. Lucrul cu dendrograme .....	168
7.5. Determinarea numărului optim de clustere în cazul metodei <i>k</i> -mean clustering .....	171
7.5.1. Metoda cotului .....	171
7.5.2. Metoda siluetei medii .....	174
7.5.3. Metoda statisticii gap .....	176
7.5.4. Extragerea rezultatelor .....	179
7.6. Determinarea numărului optim de clustere în cazul metodelor de grupare ierarhică (Hierarchical Clustering Algorithms).....	182
7.6.1. Metoda cotului .....	183
7.6.2. Metoda siluetei medii .....	183
7.6.3. Metoda statisticii gap .....	184
7.7. Alegerea celui mai bun dintre algoritmi de clustering .....	185
<b>Capitolul 8. Suport vectori mașină (SVM), utilizând mediul R.....</b>	<b>188</b>
8.1. SVM – aspecte generale .....	188
8.1.1. Introducere .....	188
8.1.2. Încărcarea pachetelor suplimentare, necesare pe lângă pachetul de bază din mediul R .....	188
8.2. Metoda SVM folosită în clasificarea simplă (booleană).....	189
8.2.1. SVM – clasificator linear. Construirea modelului și prezentarea clasificatorului de marjă maximă .....	189
8.2.2. SVM – clasificator neliniar .....	192
8.3. Metoda SVM folosită în clasificarea multiplă.....	196
8.4. Metoda SVM aplicată setului de date <i>productia</i> .....	199
8.4.1. Pregătirea datelor pentru aplicarea SVM .....	199
8.4.2. Prelucrarea propriu-zisă a datelor utilizând SVM .....	201
<b>Bibliografie.....</b>	<b>207</b>