

Jarno Tuimala

R-KIELI: PERUSTEET

R-kieli: Perusteet

R-kieli: Perusteet

Jarno Tuimala

© 2021 Jarno Tuimala

Kustantaja: BoD – Books on Demand, Helsinki, Suomi

Valmistaja: BoD – Books on Demand, Norderstedt, Saksa

ISBN: 978-952-80-6148-9

Sisällys

Lukijalle	ix
Tavoitteet	ix
Kenelle kirja on tarkoitettu?	ix
Tekniset vaatimukset	ix
Lukuohjeita	ix
Muu kirjallisuus	x
Kiitokset	x
Kirjoittajasta	xi
 OSA I	 1
R - alkeet	3
Mikä on R?	3
Mitä R:llä voi tehdä?	5
R:n ja RStudio asentaminen Windows-ympäristöön	7
R:n käynnistäminen ja käyttäminen	9
RStudio käynnistäminen ja käyttäminen	12
Laajennuspakettien asentaminen ja käyttöönotto	15
 Dokumentaatio ja tuki	 19
Käsikirjat	20
Laajennuspakettien ohjeet	20
Postilistat	23
Keskustelupalstat	24
Blogit	25

Aineistojen lukeminen R:ään	27
Taulukkomuotoisen aineiston rakenteesta	27
Tekstimuotoisen taulukon lukeminen	30
Tekstimuotoisen tiedoston lukeminen	36
Tilasto-ohjelmien datatiedostojen lukeminen	37
Merkintäkielisten tiedostojen lukeminen	42
Tietokannan kyseleminen	46
Aineiston kirjoittaminen tiedostoon	49
Tulosteiden kirjoittaminen tiedostoon	49
Tekstin kirjoittaminen tiedostoon	50
Taulukkomuotoisten aineistojen kirjoittaminen tiedostoon	51
Toimisto-ohjelmistojen tiedostojen luominen	51
Raporttien laatiminen	53
OSA II	57
R-kielen perusteet	59
Komentorivi	59
Komentojen rakenne	59
R laskimena	61
Desimaalien määrä luvuissa ja pyöristäminen	63
Laskentatarkkuus	65
Äärettömät luvut ja puuttuvat arvot	67
Puuttuvat arvot	68
Vertailuoperaattorit	69
Objektityypit	70
Objektityyppien muuntaminen toiseksi	78
Päivämäärien käsittely	81

Tekstivektoreiden käsittely	87
Joukko-opilliset operaatiot	92
Toistuvien arvojen käsittely	95
Datan muokkaaminen	101
Millainen on siisti aineisto?	101
Vektorit	103
Taulukot	120
Listat	135
Tidyverse	139
Laajennuspaketti data.table	156
Merkkijonojen käsittely	160
Lokaali	163
Numero- ja merkkijonojen järjestäminen	167
Päivämäärät	169
Ohjelmointirakenteet	171
Silmukkarakenteet	171
Ehtolauseet	178
Uusien kommentojen ohjelmoiminen	180
Ympäristöt	182
Lisäpakettien luominen	184
Ohjelmointitekniikka	189
Virheiden metsästys ja virheistä toipuminen	189
Kommentojen muuttaminen vikasietoisiksi	196
Virheisiin varautuminen	198
Lokitus	199
Yksikkötestien toteuttaminen	201

Välimuistin käyttäminen	203
Suunnittelumallit (design patterns)	205
Ohjelmointiparadigmat R:ssä	209
Imperatiivinen ohjelmointi	210
Proseduraalinen ohjelmointi	210
Funktionaalinen ohjelmointi	212
Olio-ohjelmointi	215
Suorituskyvyn optimointi	235
Käytettävät lähestymistavat	235
Laskentaoperaatioiden minimointi	235
Tunnettujen ratkaisuiden hyödyntäminen	236
Silmukat ja vektoroidut operaatiot	236
Koodin kääntäminen	238
Rinnakkaisajo	239
Muistin käyttö	240
Koodin profilointi	241
Staattinen koodianalyysi	241
Koodin muotoilu	242
Vältä kopioiden tekemistä	242
OSA III	245
Toistettava analytiikka	247
Toistettavan analytiikan peruseriaatteet	248
R:n sisuskalut	249
Miten R käynnistyy?	251
R:n ja pakettiversioiden historiointi	252

Kotoistusasetukset	253
Projektimallit	253
Notebook-teknologiat	254
Versionhallintajärjestelmät	256
Toistettava analytiikka käytännössä	257
Graafisten käyttöliittymien kehittäminen	259
Yleiskatsaus ratkaisuihin	259
Tcl/Tk	259
RGtk2+	261
gWidgets	262
shiny	262
Dash	264
fgui	265
miniGUI	266
.NET	266
R tuotantokäytössä	267
Ohjelmistotuotannon työvälineet	267
Tuotantotapoja	269
Arkkitehtuuriesimerkkejä	270
Tuotantoesimerkkejä	273
Testaus	290
Turvallisuus tuotannossa	293
Lisensointikysymyksiä	296

OSA IV	299
Pikaohje	301
R-ympäristö	301
Aineiston lataaminen	304
Aineiston muokkaaminen	306
Tapausesimerkki – rairuoho	309
Mahdollisia virhetilanteita ja niiden ratkaisuja	314
Yleisimmät virheilmoitukset	317
Virheilmoitusten yleinen rakenne	317
Sulkeiden puuttuminen	317
Operaattorien puuttuminen	318
Lainausmerkkien käyttö	318
Pakettien asentaminen	319
Non-numeric argument to binary operator	320
Object of type ‘closure’ is not subsettable	320
There is no package called...	321
Could not find function...	321
+	322
Rivinvaihdot	322
Object ‘x’ not found	324
Cannot open the connection	325
Error in if	325
Number of items to replace is not a multiple of replacement length	326
Python	327
Pythonin analytiikkapaketit	327
Pythonin ja R:n yhteiskäyttö	328
Ohjelmistoversiot	331

Lukijalle

Tavoitteet

R on tilastolliseen työskentelyyn tarkoitettu avoimen lähdekoodin ohjelmisto ja ohjelmointiympäristö. Tämä kirja esittelee R-kielen perusteet.

Kenelle kirja on tarkoitettu?

Kirja on tarkoitettu kaikille R-kielestä kiinnostuneille.

Tekniset vaatimukset

R-ohjelmisto toimii Windows-, macOS- ja Linux/UNIX-ympäristöissä, mutta tämä kirja on kirjoitettu Windows-käyttäjän näkökulmasta. Valtaosa R-kielen toiminnallisuuksista toimii kaikissa ympäristöissä samalla tavalla.

Lukuohjeita

Kirjassa on paljon R-koodia. R-kieliset osuudet tunnistaa erilaisesta ladonnasta:

```
# Luo vektorin v luvuista 1-10
v <- c(1:10)
```

```
# Tulostaa vektorin ruudulle
print(v)
```

tai:

```
# Luo vektorin v luvuista 1-10
v <- c(1:10)

# Tulostaa vektorin ruudulle
print(v)
```

```
## [1] 1 2 3 4 5 6 7 8 9 10
```

Merkillä # alkavat rivit ovat kommentteja, joilla selvennetään R-koodia. R ei suorita #-merkillä alkavia rivejä tai mitään mikä tulee samalla rivillä #-merkin jälkeen.

Kirjassa esitetyissä R:n antamissa tulosteissa esiintyy rivejä, jotka alkavat kahdella kommenttimerkillä, ja ne ovat R-komentojen ruudulle tulostamia viestejä tai tuloksia.

Toisinaan koodissa voi esiintyä kahden kaksoispisteen merkintätapa:

```
maps::map()
```

Tätä ei normaalisti tarvita, mutta sitä on jouduttu tässä kirjassa paikoin käyttämään, jos joissakin laajennuspaketeissa on ollut saman nimisiä komentoja. Merkintätavalla voidaan esittää yksiselitteisesti, että komentoa kutsutaan nimenomaan tietysti laajennuspaketista. Yllä olevassa esimerkissä kutsutaan komentoa `map()` laajennuspaketista `maps`.

Kirjan koodiesimerkkejä kannattaa testata itse R:ssä, sillä se edistää oppimista. Samalla kannattaa myös lukea ainakin kaikkien uusien komentojen ohjesivut, jotta niiden käyttö tulee tutuksi.

Muu kirjallisuus

Kirjan pyrkimyksenä ei ole esitellä edes suurta osaa R:n toiminnallisuuksista, sillä R-ekosysteemi on nykyisin niin laaja, ettei se olisi edes mahdollista. Niinpä lukijan on tarvittaessa tutustuttava muihin syventäviin lähteisiin. R:stä on muun muassa saatavilla runsaasti kirjallisuutta niin painetussa kuin elektronisessa muodossakin. Kirjallisuutta on merkittävässä määrin saatavilla myös ilmaiseksi.

Kiitokset

Tämä kirja on kirjoitettu RStudio-editorilla ja R:n laajennuspaketilla `bookdown`. PDF-muotoinen versio on käännetty LaTeXin TinyTeX-versiolla, joka on asennettu saman

nimisellä R-laajennuspaketilla. Kiitän sekä mainittujen ohjelmistojen että muun R-ekosysteemin kehittämiseen osallistuneita kehittäjiä heidän panoksestaan avoimien työvälineiden kehittämisessä.

Kirjoittajasta

Kirjoittaja on käyttänyt R-kieltä yli 20 vuoden ajan niin tutkimuksessa kuin ohjelmistokehityksessäkin.

OSA I

R on tilastolliseen työskentelyyn tarkoitettu avoimen lähdekoodin ohjelmisto ja ohjelmointiympäristö. Tämä kirja esittelee R-kielen perusteet.

BoD



9 789528 061489