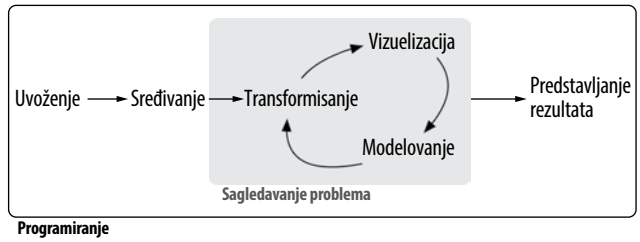
# Predgovor

Nauka koja se bavi statističkom obradom i analizom podataka (engl. *data science*) uzbudljiva je disciplina koja vam omogućava da svoje sirove podatke pretvorite u razumevanje, uvid i znanje. Cilj knjige *R za statističku obradu podataka* jeste da vam pomogne u savladavanju najvažnijih alatki R-a i tako budete u stanju da analizirate podatke. Nakon čitanja ove knjige, moći ćete da se suočite s najrazličitijim izazovima obrade podataka, koristeći najbolje delove R-a.

## Šta ćete naučiti

Nauka o statističkoj obradi i analizi podataka ogromno je polje, i svakako je ne možete savladati čitajući samo jednu knjigu. Ova knjiga treba da vam pruži solidnu osnovu kad je reč o najvažnijim alatkama i tehnikama. Naš model alatki potrebnih za uobičajen projekat obrade podataka izgleda otprilike ovako:




Prvo morate *uvesti* (engl. *import*) podatke u R. To obično znači da uzmete podatke iz datoteke, baze podataka ili veb API-ja i učitate ih u okvir za podatke (engl. *data frame*) u R-u. Ako ne možete da prenesete podatke u R, ne možete ih ni obrađivati u tom programskom okruženju!

Nakon što uvezete podatke, preporučljivo je da ih *pročistite*, tj. *sredite* (engl. *tidy*). Sređivanje podataka znači njihovo čuvanje u konzistentnom obliku koji odgovara semantici (značenju) datog skupa podataka. Ukratko, kada su vam podaci sređeni (uredni), svaka kolona je jedna promenljiva a svaki red je jedna opservacija. Urednost podataka je važna jer vam njihova konzistentna struktura omogućava da se bavite pitanjima o tim podacima, a ne time kako da ih dovedete u odgovarajući oblik za potrebe različitih funkcija.

Kada imate uredne podatke, uobičajen prvi korak je da ih *transformišete*. Transformacija uključuje sužavanje skupa podataka na opservacije koje vas zanimaju (na primer, svi ljudi iz jednog grada ili svi podaci iz prošle godine), pravljenje novih promenljivih koje su funkcije postojećih promenljivih (recimo, izračunavanje ubrzanja na osnovu brzine i vremena), i izračunavanje skupa zbirnih statističkih pokazatelja (na primer, ukupnog broja opservacija ili srednjih vrednosti). Uvoženje, sređivanje i transformisanje zajedno se zovu *priprema podataka za analizu* ili „*borba s podacima*" (engl. *wrangling*), pošto dovođenje podataka u oblik s kojim je prirodno raditi često podseća na tuču!

Kada imate uredne podatke s promenljivama koje vam trebaju, na raspolaganju su vam dva glavna načina za stizanje do znanja: vizuelizacija i modelovanje. Obe metode imaju svoje jake i slabe strane, pa ćete u svakoj realističnoj analizi prelaziti s jedne na drugu mnogo puta.

*Vizuelizacija* je – u osnovi – ljudska aktivnost. Dobra vizuelizacija će vam ukazati na neočekivane stvari ili otvoriti nova pitanja o podacima. Dobra vizuelizacija vam može i nagovestiti da postavljate pogrešno pitanje ili da treba da prikupite drugačije podatke. Vizuelizacije vas mogu iznenaditi, ali se ne mogu tako lako skalirati (prilagođavati različitim obimima ispitivanja) jer je neophodno da ih protumači čovek.

*Modelovanje* je tehnika komplementarna vizuelizaciji. Nakon što dovoljno precizno postavite pitanja, možete koristiti model da biste dobili odgovore na njih. Modeli su – u osnovi – matematička, računska alatka, pa se najčešće lako skaliraju. A čak i kada to nije moguće, obično je jeftinije kupiti još računara nego još mozgova! Međutim, pri svakom modelovanju prave se pretpostavke, a priroda modela je takva da ne može dovoditi u pitanje sopstvene pretpostavke. To znači da vas model ne može mnogo iznenaditi.

Poslednji korak u statističkoj obradi podataka jeste *predstavljanje rezultata* (engl. *communication*) – apsolutno najvažniji deo svakog projekta analize podataka. Bez obzira na to koliko dobro ste istražili i razumeli podatke zahvaljujući modelima i vizuelizaciji, to neće mnogo vredeti ukoliko ne uspete da predstavite svoje rezultate drugima.

Sve pomenute tehnike okružuje *programiranje*. Programiranje se koristi u svakom delu projekta. Ne morate biti ekspert za programiranje da biste se bavili statističkom obradom i analizom podataka, ali se učenje programiranja isplati jer će vam to što ćete postati bolji programer omogućiti da automatizujete uobičajene zadatke i lakše rešavate nove probleme.

Pomenute alatke koristićete u svakom projektu statističke obrade podataka, ali za većinu projekata one nisu dovoljne. Grubo govoreći, važi pravilo 80-20; oko 80% svakog projekta možete završiti pomoću alatki opisanih u ovoj knjizi, ali će vam trebati i druge alatke da završite preostalih 20%. Kroz celu knjigu ukazivaćemo vam na izvore dodatnog znanja.

## Organizacija knjige

Prethodni opis alatki i tehnika za statističku obradu podataka naveden je otprilike po redosledu kojim ćete ih i koristiti u analizi (mada ćete – naravno – mnogo puta prelaziti s jedne na drugu). Međutim, prema našem iskustvu, to nije najbolji redosled za učenje:

- Započinjanje učenja sa uvoženjem i sređivanjem podataka nije optimalan pristup, zato što 80% tog vremena otpada na rutinski, dosadan posao, a preostalih 20% na čudne i frustrirajuće stvari. Stoga to nikako nije dobar način za upoznavanje s novom temom! Umesto toga, počećemo s vizuelizacijom i transformisanjem podataka koji su već uvezeni i sređeni. Zahvaljujući tome, kada budete učitavali i sređivali sopstvene podatke, bićete snažno motivisani jer ćete znati da će vam se trud isplatiti.
- Neke teme je najbolje objasniti pomoću drugih tema. Na primer, verujemo da ćete lakše shvatiti kako funkcionišu modeli ako već znate šta su vizuelizacija, uredni podaci i programiranje.
- Programerske alatke ne moraju biti zanimljive same po sebi, ali vam omogućavaju da rešavate znatno teže probleme. Negde u sredini knjige opisaćemo izabrane programerske alatke, pa ćete videti da se one mogu kombinovati sa alatkama za statističku obradu podataka kako bi se rešili zanimljivi problemi modelovanja.

Pokušali smo da se u svakom poglavlju držimo sličnog šablona: počinjemo s nekim inspirativnim primerima da biste sagledali širu sliku, a zatim se bavimo detaljima. Uz svaki odeljak knjige date su vežbe koje će vam pomoći da isprobate ono što ste naučili. Iako ćete možda pasti u iskušenje da preskočite te vežbe, ne postoji bolji način učenja od rešavanja stvarnih problema.

## Šta nećete naučiti

Postoje važne teme koje nisu obrađene u ovoj knjizi. Verujemo da je bitno ostati strogo fokusiran na osnove, tako da se što brže osposobite za započinjanje rada. To znači da ova knjiga ne može da pokrije baš svaku važnu temu.

### Obimni podaci

U ovoj knjizi namerno radimo s malim skupovima podataka, smeštenim u memoriji računara. To je pravi pristup početku rada jer se ne možete baviti obimnim skupovima podataka (engl. *big data*) ukoliko nemate iskustva sa onim malim. Alatke o kojima ćete učiti u ovoj knjizi lako će izaći na kraj sa stotinama megabajta podataka, a – uz malo pažnje – obično ih možete koristiti i sa 1-2 Gb podataka. Ako rutinski radite s mnogo podataka (recimo, 10-100 Gb), trebalo bi da naučite više o jednom od hiljada paketa za programski jezik R – paketu data.table (*http://bit.ly/Rdatatable*). Ova knjiga ne obrađuje data.table zato što taj paket ima veoma sveden interfejs i teže ga je savladati pošto nudi

mnogo manje jezičkih sugestija. Ali ako radite s velikim skupovima podataka, isplatiće vam se da ovladate njime jer ćete postići bolje performanse.

Ukoliko radite s još većim skupovima podataka, pažljivo razmotrite da li je vaš problem s mnogo podataka možda prikriven problem malog skupa podataka. Bez obzira na to što su kompletni podaci obimni, odgovor na određeno pitanje o njima često se može dobiti pomoću malog skupa podataka (engl. *small data*). Možda možete naći podskup (engl. *subset*), manji uzorak (engl. *subsample*) ili sažetak podataka (engl. *summary*) koji staje u memoriju a ipak vam omogućava da dobijete odgovor na pitanje koje vas zanima. Tu je izazov pronaći pravi mali skup podataka, za šta je često potrebno obaviti mnogo iteracija.

Postoji i mogućnost da je vaš problem sa obimnim podacima – u stvari – veliki broj problema s malim skupovima podataka. Svaki pojedinačni problem možda staje u memoriju, ali imate ih milione. Na primer, možda želite da napravite model za svaku osobu u datom skupu podataka. To bi bilo trivijalno kada biste imali samo 10 ili 100 ljudi, ali imate ih milion. Srećom, svaki problem je nezavisan od drugih (postavka koja se ponekad naziva „neprijatni paralelizam", engl. *embarrassingly parallel*), pa vam samo treba sistem (kao što je Hadoop ili Spark) koji vam omogućava da različite skupove šaljete različitim računarima na obradu. Kad shvatite kako da odgovorite na dato pitanje za samo jedan podskup koristeći alatke opisane u ovoj knjizi, ovladaćete i novim alatkama – kao što su sparklyr, rhipe i ddr – da biste rešili isto pitanje i za ceo skup podataka.

## Python, Julia i prijatelji

U ovoj knjizi nećete učiti o jezicima Python, Julia i drugim programskim jezicima korisnim za statističku obradu podataka. To nije zato što smatramo da su ti jezici loši. Naprotiv, veoma su dobri! Štaviše, u praksi, većina timova koji se bave obradom i analizom podataka koristi razne jezike – često su to baš R i Python.

Ipak, čvrsto verujemo da je najbolje ovladavati jednom po jednom alatkom. Brže ćete napredovati ako se zaista udubite u jednu temu nego ukoliko se rasplinete i površno bavite mnogima. To ne znači da treba da poznajete samo jednu stvar, već samo da ćete brže učiti ako se u jednom trenutku usredsredite samo na jedno. Treba da težite učenju novih stvari tokom cele svoje karijere, ali je važno da dobro shvatite određenu temu pre nego što pređete na naredni predmet interesovanja.

Smatramo da je R odlično polazište na putu kroz statističku obradu podataka zato što je to okruženje od početka projektovano kao podrška radu s podacima. R nije samo programski jezik već i interaktivno okruženje za obradu i analizu podataka. Da bi podržao interakcije, R je mnogo fleksibilniji jezik od mnogih svojih srodnika. Uz takvu fleksibilnost idu i neke mane, ali je velika prednost lakoća iznalaženja odgovarajućih „gramatičkih" rešenja za određene delove procesa obrade podataka. Ti mini-jezici vam

pomažu da razmišljate o problemima kao neko ko se bavi naučnom analizom podataka, podržavajući pritom glatku interakciju između vašeg mozga i računara.

### Nepravougaoni podaci

U ovoj knjizi govori se isključivo o pravougaonim podacima (engl. *rectangular data*): skupovima vrednosti od kojih je svaka povezana s jednom promenljivom i jednom opservacijom. Mnogi skupovi podataka – uključujući slike, zvukove, stabla i tekst – ne zadovoljavaju ovu pretpostavku. Ipak, pravougaoni skupovi podataka sasvim su uobičajeni u nauci i industriji, pa verujemo da su odlični za započinjanje putovanja kroz statističku obradu podataka.

### Potvrda hipoteze

Analizu podataka moguće je podeliti na dva segmenta: postavljanje hipoteze i potvrđivanje hipoteze (koja se ponekad zove i potvrđujuća analiza, engl. *confirmatory analysis*). U ovoj knjizi se fokusiramo na postavljanje hipoteze, tj. istraživanje podataka (engl. *data exploration*). Znači, dubinski ispitujete podatke i – kombinujući ta saznanja sa svojim poznavanjem datog predmeta istraživanja – postavljate mnoge zanimljive hipoteze koje treba da vam pomognu da objasnite zašto se podaci ponašaju tako kako se ponašaju. Neformalno procenjujete hipoteze, skeptično analizirajući podatke na razne načine.

Potvrđivanje hipoteze (engl. *hypothesis confirmation*) komplementarno je njenom postavljanju, a teško je iz dva razloga:

- Treba vam precizan matematički model da biste napravili predikcije za koje se može dokazati da su lažne. Za to je često potrebno veliko statističko umeće.
- Jednu opservaciju možete koristiti samo jednom da biste potvrdili određenu hipotezu. Čim je upotrebite više od jedanput, vraćate se na istraživačku analizu. To znači sledeće: da biste potvrdili hipotezu, morate unapred napisati svoj plan analize, i ne odstupati od njega čak i nakon što vidite podatke. O nekim strategijama koje možete koristiti da biste olakšali ovaj postupak, govorićemo u delu IV.

Uobičajeno je posmatrati modelovanje kao alatku za potvrđivanje hipoteze a vizuelizaciju kao alatku za postavljanje hipoteze. Ipak, to je lažna dihotomija: modeli se često koriste za istraživanje, a – uz malo pažnje – vizuelizacija vam može poslužiti za potvrđivanje. Ključna razlika je u tome koliko često koristite svaku opservaciju: ako je koristite samo jednom, to je potvrđivanje; ukoliko je koristite više puta, reč je o istraživanju.

## Preduslovi

Pošli smo od nekoliko pretpostavki o tome šta bi već trebalo da znate kako biste izvukli maksimum iz ove knjige. Trebalo bi da poznajete osnove rada s brojevima, a dobro bi

bilo i da imate barem malo iskustva u programiranju. Ako nikada niste programirali, preporučujemo priručnik *Hands on Programming with R* koji je napisao Garrett i koji će vam biti korisna dopuna ove knjige.

Da biste izvršavali kôd iz ove knjige, trebaju vam četiri stvari: R, RStudio, skup paketa za R pod imenom *tidyverse* i nekoliko drugih paketa. Paketi su fundamentalne jedinice ponovljivog R koda. Oni sadrže funkcije za višekratno korišćenje, dokumentaciju koja opisuje njihovu upotrebu i primere podataka.

## R

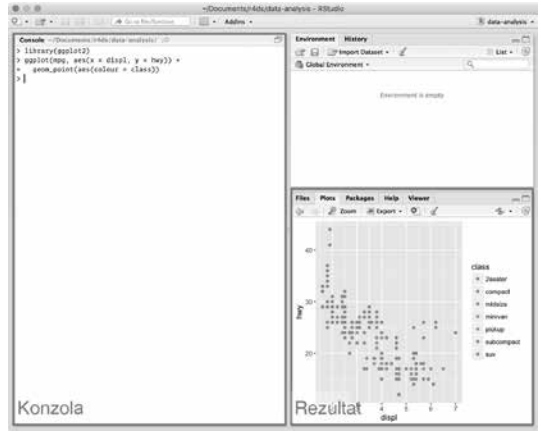
Da biste preuzeli R sa veba, idite na mrežu CRAN (*comprehensive R archive network*). CRAN se sastoji od skupa identičnih (preslikanih) servera raspoređenih širom sveta i koristi se za distribuiranje R-a i paketa za R. Nemojte pokušavati da sami izaberete server koji vam je u blizini već koristite server u oblaku, *https://cloud.r-project.org*, koji će vam automatski pronaći odgovarajući server.

Nova glavna verzija R-a objavljuje se jednom godišnje, a svake godine pojavljuju se i 2–3 izdanja s manjim izmenama. Ažuriranje može biti pomalo komplikovano – posebno za glavne verzije, za koje je neophodno da ponovo instalirate sve svoje pakete – ali je odlaganje prelaska na novu verziju još gore.

## RStudio

RStudio je integrisano razvojno okruženje, tj. IDE, za programiranje na jeziku R. Preuzmite ga s veb lokacije *http://www.rstudio.com/download* i instalirajte. RStudio se ažurira nekoliko puta godišnje. Kada se pojavi nova verzija, RStudio će vas obavestiti. Preporučuje se da ga redovno ažurirate kako biste iskoristili njegove najnovije i najbolje mogućnosti. Da biste pratili ovu knjigu, trebalo bi da imate RStudio 1.0.0.

Kada pokrenete RStudio, na interfejsu ćete videti dve ključne oblasti:



Zasad samo znajte da R kôd treba da unosite u okno konzole i da treba da pritisnete Enter kako biste ga izvršili. Više ćete naučiti kako budete napredovali kroz knjigu!

## Paket tidyverse

Uz R morate instalirati i neke pakete za njega. *Paket za R* (engl. *R package*) predstavlja skup funkcija, podataka i dokumentacije koji proširuje mogućnosti osnovnog R-a (engl. *base R*). Upotreba paketa je ključna za uspešnu primenu R-a. Većina paketa opisanih u ovoj knjizi deo je skupa pod imenom tidyverse. Za sve pakete iz skupa zajednička je filozofija rada s podacima i programiranja na R-u, i svi su napravljeni tako da prirodno rade zajedno.

Kompletan tidyverse možete instalirati unošenjem jednog reda koda:

```
install.packages("tidyverse")
```

Unesite taj red na konzolu svog računara, a zatim pritisnite Enter da biste ga izvršili. R će preuzeti pakete s mreže CRAN i instalirati ih na vaš računar. Ako imate problema sa instaliranjem, proverite da li ste povezani na internet i da li su – možda – vaša zaštitna barijera (engl. *firewall*) ili zastupnički server (engl. *proxy*) blokirali lokaciju *https://cloud.r-project.org/*.

Funkcije, objekte i pomoćne datoteke iz paketa nećete moći da koristite dok ga ne učitate pomoću funkcije `library()`. Nakon što instalirate paket, možete ga učitati pomoću funkcije `library()`:

```
library(tidyverse)
#> Loading tidyverse: ggplot2
#> Loading tidyverse: tibble
#> Loading tidyverse: tidyr
#> Loading tidyverse: readr
#> Loading tidyverse: purrr
#> Loading tidyverse: dplyr
#> Conflicts with tidy packages ----------------------------------------------
#> filter(): dplyr, stats
#> lag():    dplyr, stats
```

Poruke koje vidite govore vam da tidyverse učitava pakete **ggplot2**, **tibble**, **tidyr**, **readr**, **purrr** i **dplyr**. Oni čine *jezgro* (engl. *core*) skupa paketa tidyverse jer ćete ih koristiti u gotovo svakoj analizi.

Paketi u skupu tidyverse menjaju se prilično često. Ako izvršite funkciju `tidyverse_update()`, možete saznati postoje li ažurirane komponente i instalirati ih – ukoliko želite.

## Drugi paketi

Postoje i mnogi drugi odlični paketi koji nisu deo skupa tidyverse, zato što rešavaju probleme iz drugih oblasti ili se zasnivaju na drugačijem skupu principa. To ih ne čini boljim niti gorim – samo različitim. Drugim rečima, komplement paketa *tidyverse* nije

*messyverse*, već brojni drugi univerzumi međusobno povezanih paketa. Kako se budete upuštali u sve više i više projekata obrade i analize podataka pomoću R-a, upoznavaćete nove pakete i nove načine razmišljanja o podacima.

U ovoj knjizi koristićemo tri paketa izvan skupa tidyverse:

```
install.packages(c("nycflights13", "gapminder", "Lahman"))
```

Ti paketi sadrže podatke o avionskim letovima, razvoju zemalja sveta i bejzbolu, a koristićemo ih da bismo iustrovali ključne ideje obrade i analize podataka.

## Izvršavanje R koda

U prethodnom odeljku navedeno je nekoliko primera izvršavanja R koda. U ovoj knjizi, kôd je prikazan ovako:

```
1 + 2
#> [1] 3
```

Ako izvršite isti kôd na svojoj lokalnoj konzoli, izgledaće ovako:

```
> 1 + 2
[1] 3
```

Postoje dve glavne razlike. Na svojoj konzoli, komande unosite iza znaka >, koji se zove *odzivnik* (engl. *prompt*); u knjizi ga ne prikazujemo. Osim toga, u knjizi su izlazni rezultati pretvoreni u komentare i označeni sa #>; na vašoj konzoli se pojavljuju odmah iza koda. Znači, ako koristite elektronsku verziju knjige (na engleskom), možete lako kopirati kôd iz knjige na konzolu.

U celoj knjizi koristimo dosledan skup pravila (konvencija) za predstavljanje koda:

- Funkcije su napisane fontom za kôd i slede im zagrade – na primer, `sum()` ili `mean()`.
- Ostali R objekti (recimo, podaci ili argumenti funkcije) napisani su fontom za kôd, ali ne sadrže zagrade – na primer `flights` ili `x`.
- Ako želimo jasno da naglasimo iz kog paketa potiče neki objekat, navodimo ime paketa praćeno znakom dve tačke – na primer, `dplyr::mutate()` ili `nycflights13::flights`. To je takođe ispravan R kôd.

## Dobijanje pomoći i izvori dodatnog znanja

Ova knjiga nije ostrvo; ne možete ovladati R-om koristeći samo jedan resurs. Kada počnete da primenjujete tehnike opisane u ovoj knjizi na sopstvene podatke, brzo ćete se susresti s pitanjima na koje ova knjiga ne daje odgovor. U ovom odeljku navedeno je nekoliko saveta o dobijanju pomoći i nastavljanju učenja.

Ako se zaglavite, počnite od Googlea. Obično je dodavanje slova „R“ upitu dovoljno da ograniči rezultate pretrage na one relevantne: ukoliko je takva pretraga beskorisna, to često znači da nema dostupnih rezultata specifičnih za R. Google je posebno koristan za poruke o greškama. Ako dobijete poruku o grešci i nemate pojma šta ona znači, pokušajte da je unesete u Google! Vrlo je verovatno da je ista poruka već zbunjivala nekoga i da ćete naći pomoć negde na vebu. (Ukoliko poruka o grešci nije na engleskom jeziku, zadajte komandu `Sys.setenv(LANGUAGE = "en")` i ponovo izvršite kôd; lakše se nalazi pomoć za poruke o greškama na engleskom.)

U slučaju da Google ne pomogne, pokušajte na lokaciji stackoverflow (*http://stackoverflow.com*). Počnite tako što ćete provesti malo vremena tražeći postojeći odgovor; ako u pojam za pretragu uključite `[R]`, ograničićete je na pitanja i odgovore koji se odnose na R. Ukoliko ne nađete ništa korisno, pripremite najmanji primer koji se može reprodukovati – tzv. **reprex**. Uz dobar reprex, drugi korisnici će vam mnogo lakše pomoći, a često ćete i sami rešiti problem dok budete pravili reprex.

Tri su stvari koje morate uvrstiti u svoj primer da bi se on mogao reprodukovati: potrebni paketi, podaci i kôd.

- *Paketi* treba da se učitaju na početku skripta, tako da se lako vidi koji su potrebni za dati primer. To je dobar trenutak da proverite koristite li najnoviju verziju svakog paketa; moguće je da ste naišli na grešku koja je popravljena otkako ste instalirali dati paket. Pakete iz skupa tidyverse najlakše ćete proveriti komandom `tidyverse_update()`.
- Najlakši način da u pitanje uvrstite *podatke* jeste da upotrebite `dput()` kako biste generisali R kôd koji ponovo pravi dati skup podataka. Na primer, da biste ponovo napravili skup podataka `mtcars` u R-u, pratite sledeće korake:
  1. Izvršite komandu `dput(mtcars)` u R-u.
  2. Kopirajte dobijeni rezultat.
  3. U primeru mog skripta koji se može reprodukovati, upišite `mtcars <-` pa prenesite kopirani rezultat komandom Paste.

  Pokušajte da nađete najmanji podskup svojih podataka u kome se i dalje ispoljava dati problem.
- Odvojte malo vremena kako biste osigurali da drugi mogu lako čitati vaš *kôd*:
  - Proverite da li ste u kodu koristili razmake i da li su imena vaših promenljivih koncizna, ali informativna.
  - Koristite komentare da biste ukazali na to gde se javlja problem.
  - Potrudite se da iz koda uklonite sve što se ne odnosi na dati problem. Što je kôd kraći, lakše ga je i razumeti i popraviti.

  Završite tako što ćete proveriti da li ste zaista napravili primer koji se može reprodukovati, tako što ćete započeti novu sesiju u R-u, pa iskopirati i preneti svoj skript u nju.

Trebalo bi, takođe, da provedete i neko vreme pripremajući se za rešavanje problema pre nego što se pojave. Ako svakog dana budete odvajali malo vremena za učenje R-a, to će vam se lepo isplatiti na duže staze. Jedna od mogućnosti za to je da pratite šta Hadley, Garrett i svi drugi saradnici na projektu RStudio rade na blogu posvećenom ovom integrisanom razvojnom okruženju (*https://blog.rstudio.org*). Tu postavljamo obaveštenja o novim paketima i novim mogućnostima, kao i odgovarajuće kurseve. Da biste bili u toku s novim mogućnostima IDE-a, možete pratiti i autore ove knjige na Twitteru: Hadley (@hadleywickham (*https://twitter.com/hadleywickham*)), Garrett (@statgarrett (*https://twitter.com/statgarrett*)), ili pratiti @rstudiotips (*https://twitter.com/rstudiotips*).

Ako želite da se povežete sa širom zajednicom korisnika R-a, preporučujemo da čitate tekstove na lokaciji *http://www.r-bloggers.com*: ona obuhvata preko 500 blogova o R-u koji potiču iz celog sveta. Ukoliko aktivno koristite Twitter, pratite oznaku #rstats. Twitter je jedna od ključnih alatki koje Hadley koristi da bi bio u toku s novinama u razvoju R-a.

## Zahvalnice

Za ovu knjigu nisu zaslužni samo Hadley i Garrett – ona je rezultat mnogih razgovora (ličnih i preko mreže) koje smo vodili s mnogim ljudima u zajednici korisnika R-a. Želimo posebno da se zahvalimo onima koji su proveli silne sate odgovarajući na naša glupa pitanja i pomažući nam da bolje razmišljamo o nauci posvećenoj statističkoj obradi i analizi podataka:

- Jenny Bryan i Lionel Henry – za mnoge korisne diskusije o radu sa listama i kolonama koje sadrže liste.
- Za tri poglavlja o radnom toku adaptiran je (s dozvolom) članak „R basics, workspace and working directory, RStudio projects“ (*http://bit.ly/Rbasicsworkflow*), čiji je autor Jenny Bryan.
- Genevera Allen – za razgovore o modelima, modelovanju, perspektivi učenja statistike i razlikama između postavljanja i potvrđivanja hipoteze.
- Yihui Xie – za njegov rad na paketu **bookdown** (*https://github.com/rstudio/bookdown*), i za neumorno odgovaranje na naše zahteve.
- Bill Behrman – za pažljivo čitanje cele knjige i njenu proveru s polaznicima njegovog kursa o statističkoj obradi podataka na Stanfordu.
- Twitter zajednica #rstats koja je pregledala nacrte svih poglavlja i dala tone korisnih sugestija.
- Tal Galili – za proširivanje njegovog paketa **dendextend** kako bi podržao odeljak o klasterovanju koji ipak nije uvršćen u finalnu verziju knjige.

Pisanje ove knjige odvijalo se javno, tako da su mnogi ljudi doprineli svojim predlozima za rešavanje manjih problema. Posebno smo zahvalni svima koji su dali svoj doprinos preko lokacije GitHub (navodimo ih abecednim redosledom): adi pradhan, Ahmed El-Gabbas, Ajay Deonarine, @Alex, Andrew Landgraf, bahadir cankardes, @batpigandme, @behrman, Ben Marwick, Bill Behrman, Brandon Greenwell, Brett Klamer, Christian G. Warden, Christian Mongeau, Colin Gillespie, Cooper Morris, Curtis Alexander, Daniel Gromer, David Clark, Derwin McGeary, Devin Pastoor, Dylan Cashman, Earl Brown, Eric Watt, Etienne B. Racine, Flemming Villalona, Gregory Jefferis, @harrismcgehee, Hengni Cai, Ian Lyttle, Ian Sealy, Jakub Nowosad, Jennifer (Jenny) Bryan, @jennybc, Jeroen Janssens, Jim Hester, @jjchern, Joanne Jang, John Sears, Jon Calder, Jonathan Page, @jonathanflint, Jose Roberto Ayala Solares, Julia Stewart Lowndes, Julian During, Justinas Petuchovas, Kara Woo, @kdpsingh, Kenny Darrell, Kirill Sevastyanenko, @koalabearski, @KyleHumphrey, Lawrence Wu, Matthew Sedaghatfar, Mine Cetinkaya-Rundel, @MJMarshall, Mustafa Ascha, @nate-d-olson, Nelson Areal, Nick Clark, @nickelas, Nirmal Patel, @nwaff, @OaCantona, Patrick Kennedy, @Paul, Peter Hurford, Rademeyer Vermaak, Radu Grosu, @rlzijdeman, Robert Schuessler, @robinlovelace, @robinsones, S'busiso Mkhondwane, @seamus-mckinsey, @seanpwilliams, Shannon Ellis, @shoili, @sibusiso16, @spirgel, Steve Mortimer, @svenski, Terence Teo, Thomas Klebel, TJ Mahr, Tom Prior, Will Beasley, @yahwes, Yihui Xie, @zeal626.

## On-lajn verzija

On-lajn verzija ove knjige na engleskom jeziku dostupna je na adresi *http://r4ds.had.co.nz*. Ona će nastaviti da se ažurira i u periodu između štampanja papirnih izdanja knjige. Izvorni kôd knjige nalazi se na adresi *https://github.com/hadley/r4ds*. Knjigu pokreće paket **bookdown** (*https://bookdown.org*), pa je datoteke napravljene pomoću jezika za označavanje R Markdown lako prevesti u format HTML, PDF i EPUB.

Za izradu ove knjige korišćeno je:

```
devtools::session_info(c("tidyverse"))
#> Session info -------------------------------------------------------------
#>  setting  value
#>  version  R version 3.4.0 (2017-04-21)
#>  system   x86_64, linux-gnu
#>  ui       X11
#>  language (EN)
#>  collate  en_US.UTF-8
#>  tz       Etc/UTC
#>  date     2017-05-04
#> Packages -----------------------------------------------------------------
#>  package      * version   date       source
#>  assertthat     0.2.0     2017-04-11 cran (@0.2.0)
#>  BH             1.62.0-1  2016-11-19 cran (@1.62.0-)
#>  broom          0.4.2     2017-02-13 cran (@0.4.2)
```

```
#>  cellranger     1.1.0     2016-07-27 cran (@1.1.0)
#>  colorspace     1.3-2     2016-12-14 cran (@1.3-2)
#>  curl           2.6       2017-04-27 CRAN (R 3.4.0)
#>  DBI            0.6-1     2017-04-01 cran (@0.6-1)
#>  dichromat      2.0-0     2013-01-24 cran (@2.0-0)
#>  digest         0.6.12    2017-01-27 CRAN (R 3.4.0)
#>  dplyr        * 0.5.0     2016-06-24 cran (@0.5.0)
#>  forcats        0.2.0     2017-01-23 cran (@0.2.0)
#>  foreign        0.8-67    2016-09-13 CRAN (R 3.4.0)
#>  ggplot2      * 2.2.1     2016-12-30 cran (@2.2.1)
#>  gtable         0.2.0     2016-02-26 cran (@0.2.0)
#>  haven          1.0.0     2016-09-23 cran (@1.0.0)
#>  hms            0.3       2016-11-22 cran (@0.3)
#>  httr           1.2.1     2016-07-03 CRAN (R 3.4.0)
#>  jsonlite       1.4       2017-04-08 CRAN (R 3.4.0)
#>  labeling       0.3       2014-08-23 cran (@0.3)
#>  lattice        0.20-35   2017-03-25 CRAN (R 3.4.0)
#>  lazyeval       0.2.0     2016-06-12 cran (@0.2.0)
#>  lubridate      1.6.0     2016-09-13 cran (@1.6.0)
#>  magrittr       1.5       2014-11-22 cran (@1.5)
#>  MASS           7.3-47    2017-02-26 CRAN (R 3.4.0)
#>  mime           0.5       2016-07-07 CRAN (R 3.4.0)
#>  mnormt         1.5-5     2016-10-15 cran (@1.5-5)
#>  modelr         0.1.0     2016-08-31 cran (@0.1.0)
#>  munsell        0.4.3     2016-02-13 cran (@0.4.3)
#>  nlme           3.1-131   2017-02-06 CRAN (R 3.4.0)
#>  openssl        0.9.6     2016-12-31 CRAN (R 3.4.0)
#>  plyr           1.8.4     2016-06-08 cran (@1.8.4)
#>  psych          1.7.5     2017-05-03 cran (@1.7.5)
#>  purrr        * 0.2.2     2016-06-18 cran (@0.2.2)
#>  R6             2.2.0     2016-10-05 CRAN (R 3.4.0)
#>  RColorBrewer   1.1-2     2014-12-07 cran (@1.1-2)
#>  Rcpp           0.12.10.2 2017-05-03 Github (RcppCore/Rcpp@c57b754)
#>  readr        * 1.1.0     2017-03-22 cran (@1.1.0)
#>  readxl         1.0.0     2017-04-18 cran (@1.0.0)
#>  rematch        1.0.1     2016-04-21 cran (@1.0.1)
#>  reshape2       1.4.2     2016-10-22 cran (@1.4.2)
#>  rvest          0.3.2     2016-06-17 cran (@0.3.2)
#>  scales         0.4.1     2016-11-09 cran (@0.4.1)
#>  selectr        0.3-1     2016-12-19 cran (@0.3-1)
#>  stringi        1.1.5     2017-04-07 cran (@1.1.5)
#>  stringr        1.2.0     2017-02-18 cran (@1.2.0)
#>  tibble       * 1.3.0     2017-04-01 cran (@1.3.0)
#>  tidyr        * 0.6.1     2017-01-10 cran (@0.6.1)
#>  tidyverse    * 1.1.1     2017-01-27 cran (@1.1.1)
#>  xml2           1.1.1     2017-01-24 cran (@1.1.1)
```

## Konvencije korišćene u knjizi

U ovoj knjizi korišćene su sledeće tipografske konvencije:

*Kurziv*

Koristi se za nove termine, reči na engleskom, URL adrese, adrese e-pošte, imena datoteka i oznake tipova datoteka.

**Podebljano**

Koristi se za imena paketa za R.

`Konstantne širine`

Koristi se za listinge programa i – unutar pasusa – za označavanje programskih elemenata kao što su imena promenljivih ili funkcija, baza podataka, tipova podataka, promenljivih okruženja, naredbi i rezervisanih reči.

**`Konstantne širine i podebljano`**

Koristi se za komande ili drugi tekst koji korisnik treba da unese doslovno onako kako je napisano.

*`Konstantne širine i kurziv`*

Koristi se za tekst koji treba zameniti vrednostima koje unosi korisnik ili određuje kontekst.

Ovaj element označava savet ili predlog.

## Korišćenje primera koda

Izvorni kôd se može preuzeti s veb lokacije *https://github.com/hadley/r4ds*.

Svrha ove knjige je da vam pomogne da obavite posao. U opštem slučaju, ako je u knjizi naveden određen primer koda, možete ga koristiti u svojim programima i dokumentaciji. Ne morate stupati u kontakt sa izdavačem izvornog izdanja ove knjige radi dobijanja ovlašćenja za upotrebu, osim ako reprodukujete značajan deo koda. Na primer, za pisanje programa koji sadrži nekoliko blokova koda iz ove knjige ne treba vam ovlašćenje. Za prodaju ili distribuiranje CD-ROM-a s primerima iz knjiga čiji je izdavač O'Reilly, potrebno je ovlašćenje izdavača. Za odgovaranje na pitanja citiranjem ove knjige i navođenjem primera koda iz knjige, nije potrebno ovlašćenje. Za umetanje značajne količine koda iz ove knjige u dokumentaciju vašeg proizvoda, treba vam ovlašćenje.

Bićemo vam zahvalni ako nas navedete kao izvor, ali to nije obavezno. Navođenje izvora obično podrazumeva naslov, autora, izdavača i ISBN broj.

Ako smatrate da vaš slučaj upotrebe primera koda iz ove knjige nije obuhvaćen prethodno opisanim ovlašćenjima, obratite se izdavaču izvornog izdanja ove knjige na adresu *permissions@oreilly.com*.

Spisak grešaka, primere iz ove knjige i dodatne informacije o njoj potražite na veb lokaciji *http://bit.ly/r-for-data-science*.

Komentare ili tehnička pitanja o knjizi šaljite na adresu *bookquestions@oreilly.com*.

Više informacija o knjigama, kursevima, konferencijama i novostima kompanije O'Reilly, naći ćete na veb lokaciji *http://www.oreilly.com*.

Više informacija o izdanjima Mikro knjige potražite na adresi *www.mikroknjiga.rs*.