

Analiza podataka teniskih mečeva

Antonio Babić, Iva Maria Ivanković, Gabrijel Jambrošić, Antun Jurelinac

14. siječanj 2022.

Sadržaj

1	Motivacija i opis problema	2
2	Opis i učitavanje skupa podataka	2
2.1	Opis skupa podataka	2
2.2	Učitavanje skupa podataka	2
3	Projektna pitanja	2
3.1	Zadana pitanja	3
3.1.1	Distribucija visine igrača	3
3.1.2	Odnos ljevaka i dešnjaka	4
3.1.3	Pobjeda prvog seta	4
3.1.4	Predviđanje pobjednika meča	4
3.2	Vlastita pitanja	4

1 Motivacija i opis problema

Statistička analiza podataka oduvijek je prisutna u sportu. Njome se služe komentatori koji prije neke važne utakmice trebaju naučiti što više činjenica o igraču ili timu, što spada pod deskriptivnu statistiku. Investitori na temelju statistike kluba raspoređuju svoja ulaganja, što za posljedicu može imati napredak kluba ili njegovu potpunu propast. Plaće igrača i njihove cijene na tržištu transfera izravno ovise o njihovoj statistici u prethodnoj sezoni, a kladionice provode iscrpne analize podataka kako bi postavile kvote.

U tenisu je statistika kao alat dobila dodatnu popularnost zahvaljujući bivšem treneru Craigu O'Shaughnessyju, strategu s uporištem u statistici čija je analiza bila ključna u rezultatima Novaka Đokovića protiv njegovih najvećih rivala. Svojim zaključcima izvedenim iz povijesnih podataka mečeva tenisačima je moguće prilagoditi kondicijske pripreme, teniske treninge i strategiju protiv pojedinih protivnika, što rezultira boljom i konzistentnijom igrom.

U nastavku teksta analizirat će se skup podataka o teniskim mečevima i tenisačima te će se iz podataka pokušati izvesti zaključci i pomoću njih odgovoriti na projektna pitanja. Analiza podataka bit će provedena u programskom jeziku *R*, a odabrano okruženje je *RStudio*.

2 Opis i učitavanje skupa podataka

2.1 Opis skupa podataka

Podaci se sastoje od svih ATP mečeva odigranih između 1991. i 2020. godine. Svakom igraču pridodano je više značajki kao što su visina, starost, ruka kojom igra, igra li jednoručni ili dvoručni backhand itd. Dodatno je svaki meč opisan s više značajki poput rankinga pobjednika, rankinga gubitnika, trajanja meča, broja winnera pobjednika, broja neprisiljenih grešaka gubitnika, broja spašenih break prilika i sl.

Jedan redak u tablici skupa podataka sadrži podatke raspoređene u sljedeće stupce (ne moraju sve vrijednosti biti definirane): redni broj podatka, identifikacijska oznaka turnira, naziv turnira, vrsta podloge, broj natjecatelja na turniru, razina turnira, datum održavanja turnira, redni broj meča, rezultat meča, broj setova (*best of x*), razina meča (npr. kvalifikacijski, četvrtfinale, finale) i trajanje meča. Informacije o pobjedniku i gubitniku sadržane su u sljedećim stupcima, za svakog od dvojice igrača zasebno: identifikacijska oznaka, jakosna skupina u ždrijebu, **entry**, ime i prezime, dominantna ruka, visina, nacionalnost, dob, **ace**, **df**, **svpt**, **1stIn**, **1stWon**, **2ndWon**, **svGms**, **bpSaved**, **bpFaced**, ranking te ranking bodovi.

2.2 Učitavanje skupa podataka

Zadani skup podataka učitani su iz .csv datoteke *tennis_atp_matches.csv*.

```
tennis <- read.csv("tennis_atp_matches.csv")
```

Odmah na početku jasno je da su neki od podataka suvišni za analizu koja će biti provedena u sklopu projekta pa će odmah na početku odgovarajući stupci biti eliminirani iz tablice radi bolje preglednosti podataka. To su podaci poput naziva i datuma održavanja turnira te rednog broja meča. **Ovdje ćemo kasnije dodati sve stupce koje nismo koristili uz objašnjenje da se tih stupaca naša pitanja ne tiču. Slobodno izbacite neke stupce koji vam ipak trebaju. Zakomentirala sam jer mi inače javlja error koji nisam još skužila zašto se pojavljuje.**

```
#tennis <- subset(tennis, select = -c(tourney_name, tourney_date, match_num))
```

Možda ovdje dodati neki summary podataka ako budemo imali vremena i volje.

3 Projektna pitanja

U sklopu zadatka postavljena su četiri pitanja, uz mogućnost postavljanja vlastitih pitanja i pokretanja dodatne problematike vezano za dani skup podataka. **U slučaju da ipak nećemo postavljati vlastita pitanja,**

ovaj dio teksta trebalo bi malo izmijeniti.

3.1 Zadana pitanja

3.1.1 Distribucija visine igrača

Postavljeno pitanje bilo je: Možemo li nešto zaključiti iz distribucije visine najboljih deset igrača u posljednjih 30 godina u odnosu na distribuciju visine igrača koji nisu bili tako uspješni?

Kako je u skupu podataka svakom igraču pridružen njegov ranking, taj će se podatak koristiti pri određivanju najuspješnijih igrača. Svake godine igrač dobije novi ranking te se za svaku godinu može odrediti popis deset igrača s najboljim rankingom. Nakon što se prikupe podaci svih trideset godina, profiliraju se na način da se svaki igrač pojavljuje samo jednom. To je skup podataka koji će se koristiti u analizi i predstavljati najuspješnije igrače. **Malo nespretna način izdvajanja odmah na početku. Što ako je visina NA i pretvori se u 1000? Nije se dogodilo, ali treba pripaziti na općeniti slučaj.**

S obzirom na činjenicu da se u skupu podataka na nekim mjestima pojavljuju igrači kojima nije definiran ranking, postaviti ćemo im ranking na 1000 kako ne bi ušli u selekciju igrača s najboljim rankingom. Broj 1000 odabran je donekle proizvoljno - mogao je biti i 11, bitno je da je veći od 10.

```
rankingRelevantData <- tennis[c("winner_id", "winner_name", "winner_rank", "winner_ht")]
rankingRelevantData[is.na(rankingRelevantData)] = 1000
```

Iz podataka se zatim izvuče popis svih pobjednika mečeva za koje je u bilo kojem meču zabilježen ranking ≤ 10 . Razlog zašto se gledaju samo pobjednici jasan je ako se malo promisli o samom sustavu rangiranja - niti jedan igrač koji je u nekom trenutku bio među najboljom desetericom nije se mogao ne pojaviti u barem jednom meču kao pobjednik.

```
#svi igrači koji su u nekom trenutku imali ranking <= 10
winnersBestRanking <- rankingRelevantData[rankingRelevantData$winner_rank <= 10,]

#izdvajanje relevantnih stupaca
bestRanking <- winnersBestRanking[c("winner_id", "winner_name", "winner_ht")]

#brisanje duplikata
mostSuccessfulPlayers <- unique(bestRanking)
colnames(mostSuccessfulPlayers) <- c("player_id", "player_name", "player_ht")
```

Skup igrača koji nisu bili tako uspješni ustvari je skup svih ostalih igrača.

Taj popis dobijemo tako što iz tablice s popisom svih igrača izuzmemo one retke koji se nalaze u tablici s popisom najuspješnijih igrača.

```
#popis svih igrača koji imaju barem jednu zabilježenu pobjedu
winners <- tennis[c("winner_id", "winner_name", "winner_ht")]
groupedWinners <- subset(as.data.frame(table(winners)), Freq != 0)
groupedWinners[4] = NULL
colnames(groupedWinners) <- c("player_id", "player_name", "player_ht")

#popis svih igrača koji imaju barem jedan zabilježen gubitak
losers <- tennis[c("loser_id", "loser_name", "loser_ht")]
groupedLosers <- subset(as.data.frame(table(losers)), Freq != 0)
groupedLosers[4] = NULL
colnames(groupedLosers) <- c("player_id", "player_name", "player_ht")

#full outer join pobjednika i gubitnika
allPlayers <- merge(groupedWinners, groupedLosers, all = TRUE)
```

```
#izuzimamo igrače koji su među najuspješnijima
notSoSuccessfulPlayers <- subset(allPlayers, !player_id %in% mostSuccessfulPlayers$player_id)
```

Ovdje krećemo s analizom podataka.

3.1.2 Odnos ljevaka i dešnjaka

3.1.3 Pobjeda prvog seta

3.1.4 Predviđanje pobjednika meča

3.2 Vlastita pitanja

- 1) Možemo li nešto zaključiti iz distribucije visine najboljih deset igrača u posljednjih 30 godina u odnosu na distribuciju visine igrača koji nisu bili tako uspješni?

Kreiramo novu tablicu u koju rangiramo igrače po broju pobjeda, za početak. Raspravljali smo o tome da ih rangiramo po postotku pobjeda, ali to nema smisla zbog toga što netko npr. može imati 2/2 pobjede, a netko 19/20. (Postaviti pitanje asistentu.)

```
pobjednici <- tennis[c("winner_id", "winner_name", "winner_ht")]
grupiraniPobjednici <- subset(as.data.frame(table(pobjednici)), Freq != 0)
colnames(grupiraniPobjednici) <- c("player_id", "player_name", "player_ht", "no_of_wins")
sortiraniPobjednici <- grupiraniPobjednici[order(grupiraniPobjednici$no_of_wins, decreasing = TRUE),]
desetNajboljih <- head(sortiraniPobjednici, 10)
```

Problem se pojavio kod traženja igrača koji nisu bili tako uspješni. Naime, nema smisla rangirati ih po najmanjem broju pobjeda, jer će biti puno igrača s 0 ili 1 pobjedom. Isto tako, rangiranje po najmanjem postotku pobjeda ne bi bilo baš sretno rješenje. Iz tog razloga, rangiramo ih po najvećem broju gubitaka i nadamo se da će podaci imati smisla.

Time smo dobili tablicu u kojoj se pojavljuje Andy Murray, što nikako nema smisla. Ovaj kod ispod ne treba gledati!

```
gubitnici <- tennis[c("loser_id", "loser_name", "loser_ht")]
grupiraniGubitnici <- subset(as.data.frame(table(gubitnici)), Freq != 0)
colnames(grupiraniGubitnici) <- c("player_id", "player_name", "player_ht", "no_of_losses")
sortiraniGubitnici <- grupiraniGubitnici[order(grupiraniGubitnici$no_of_losses, decreasing = TRUE),]
desetNajlosijih <- head(sortiraniGubitnici, 10)
```

Sljedeće rješenje, koje smatramo najboljim, jest da ‘odsiječemo’ igrače s najmanjim brojem mečeva i onda za ostale gledamo najmanji postotak pobjeda.

```
#full outer join pobjednika i gubitnika
pobjedeIPorazi <- merge(grupiraniPobjednici, grupiraniGubitnici, all = TRUE)

#mijenja NA s 0 gdje god se pojavljuje da se može zbrajati
pobjedeIPorazi[is.na(pobjedeIPorazi)] = 0

#dodajemo novi stupac u kojem piše ukupan broj odigranih mečeva
for (i in 1:nrow(pobjedeIPorazi)) {
  pobjedeIPorazi$total[i] <- pobjedeIPorazi$no_of_wins[i] + pobjedeIPorazi$no_of_losses[i]
}

#sortiramo tenisace po ukupnom broju mečeva
pobjedeIPoraziSortirano <- pobjedeIPorazi[order(pobjedeIPorazi$total, decreasing = TRUE),]
```

```

#uzimamo samo one koji imaju vise od 100 meceva
pobjedeIPoraziSortiranoBezNajlosijih <- subset(pobjedeIPoraziSortirano, pobjedeIPoraziSortirano$total > 100)

#dodajemo novi stupac u kojem pise postotak pobjeda u ukupnom broju meceva
for(i in 1:nrow(pobjedeIPoraziSortiranoBezNajlosijih)) {
  pobjedeIPoraziSortiranoBezNajlosijih$win_percentage[i] <- pobjedeIPoraziSortiranoBezNajlosijih$no_of_wins[i] / pobjedeIPoraziSortiranoBezNajlosijih$total[i]
}

#sortiramo tenisace po postotku pobjeda
pobjedeIPoraziSortiranoBezNajlosijihSortiranoPoPostotku <- pobjedeIPoraziSortiranoBezNajlosijih[order(pobjedeIPoraziSortiranoBezNajlosijih$win_percentage),]

#uzmemo one koji imaju <33% pobjede
najlosiji <- subset(pobjedeIPoraziSortiranoBezNajlosijihSortiranoPoPostotku, pobjedeIPoraziSortiranoBezNajlosijih$win_percentage < 0.33)
desetNajlosijih <- tail(najlosiji, 10)

```

Potrebno je sada spojiti igrače s njihovom visinom i odrediti distribuciju tih podataka.

```

prosjecnaVisinaNajboljih <- mean(as.numeric(as.character(desetNajboljih$player_ht)))
varijancaVisineNajboljih <- var(as.numeric(as.character(desetNajboljih$player_ht)))
standardnaDevijacijaVisineNajboljih <- sd(as.numeric(as.character(desetNajboljih$player_ht)))

prosjecnaVisinaNajlosijih <- mean(as.numeric(as.character(desetNajlosijih$player_ht)))
varijancaVisineNajlosijih <- var(as.numeric(as.character(desetNajlosijih$player_ht)))
standardnaDevijacijaVisineNajlosijih <- sd(as.numeric(as.character(desetNajlosijih$player_ht)))

```

Izrađujemo dijagram raspršenja za visinu igrača i postotak pobjeda. Iz nekog razloga, pojavljuju se mali box plotovi unutar scatter plot (pitati asistenta zašto bi to moglo biti).

```

#x <- pobjedeIPoraziSortiranoBezNajlosijih$player_ht
#y <- pobjedeIPoraziSortiranoBezNajlosijih$win_percentage
#plot(x, y, xlab = 'Visina igrača [cm]', ylab = 'Postotak pobjeda')

```

```

pobjednici = tennis[c("winner_id", "winner_name")]
grupiraniPobjednici = subset(as.data.frame(table(pobjednici)), Freq != 0)
sortiraniPobjednici = grupiraniPobjednici[order(grupiraniPobjednici$Freq, decreasing = TRUE),]
desetNajboljih = head(sortiraniPobjednici, 10)

```

```

ruke = tennis[c("winner_hand", "loser_hand")]
razl_ruke = subset(ruke, (as.character(winner_hand) != as.character(loser_hand) & as.character(loser_hand) != "0"))
grupiraneruke = subset(as.data.frame(table(razl_ruke)), Freq != 0)

```

Možemo li na temelju dobitnika prvog seta predvidjeti dobitnika cijelog meča?

```

full_sets <- tennis[!grepl("[A-Za-z]", tennis$score),]
sum(substr(full_sets$score, 1, 1) > substr(full_sets$score, 3, 3)) / nrow(full_sets)

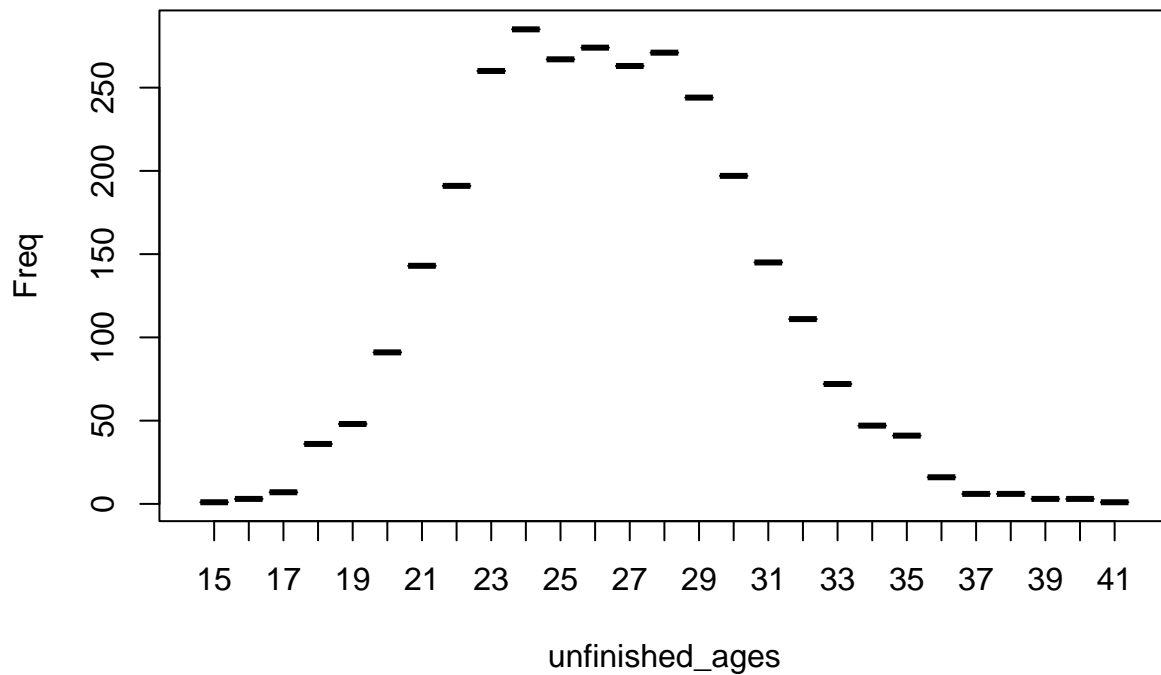
```

```
## [1] 0.8084622
```

```

unfinished_sets <- tennis[grepl("[A-Za-z]", tennis$score),]
unfinished_ages <- round(unfinished_sets[c("loser_age")])
grupirani_unf <- subset(as.data.frame(table(unfinished_ages)), Freq != 0)
plot(grupirani_unf)

```



Promatranje napretka najboljih igrača

```
tri_najbolja <- sort(table(tennis$winner_id), decreasing=TRUE)[1:3]
godine <- sort(unique(substr(tennis$tourney_date, 1, 4)))
for(player_id in dimnames(tri_najbolja)[[1]]) {
  pobjede_po_godinama <- rep()
  for(godina in godine) {
    pobjede_po_godinama <- append(pobjede_po_godinama, nrow(tennis[which(tennis$winner_id == player_id &
      / nrow(tennis[which((tennis$winner_id == player_id | tennis$loser_id == player_id) & substr(tennis$tourney_date, 1, 4) == godina])
    ]))
  }
  plot(godine, pobjede_po_godinama, type="b")
}
```

