Analiza podataka teniskih mečeva

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

14. siječanj 2022.

Sadržaj

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, ime i prezime, dominantna ruka, visina, nacionalnost, dob, ranking itd.

2.2 Učitavanje skupa podataka

Zadani skup podataka učitan je iz .csv datoteke *tennis_atp_matches.csv*.

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

Imena svih varijabli u skupu podataka dana su u nastavku.

```
names(tennis)
```

```
##  [1] "X"                 "tourney_id"        "tourney_name"
##  [4] "surface"           "draw_size"         "tourney_level"
##  [7] "tourney_date"      "match_num"         "winner_id"
## [10] "winner_seed"       "winner_entry"      "winner_name"
## [13] "winner_hand"       "winner_ht"         "winner_ioc"
## [16] "winner_age"        "loser_id"          "loser_seed"
## [19] "loser_entry"       "loser_name"        "loser_hand"
## [22] "loser_ht"          "loser_ioc"         "loser_age"
## [25] "score"             "best_of"           "round"
## [28] "minutes"           "w_ace"             "w_df"
## [31] "w_svpt"            "w_1stIn"           "w_1stWon"
## [34] "w_2ndWon"          "w_SvGms"           "w_bpSaved"
## [37] "w_bpFaced"         "l_ace"             "l_df"
```

```
## [40] "l_svpt"             "l_1stIn"            "l_1stWon"
## [43] "l_2ndWon"           "l_SvGms"            "l_bpSaved"
## [46] "l_bpFaced"          "winner_rank"        "winner_rank_points"
## [49] "loser_rank"         "loser_rank_points"
```

Možemo saznati neke osnovne informacije o skupu podataka, npr. njegove dimenzije, odnosno broj redaka i stupaca.

```
dim(tennis)
```

```
## [1] 96602    50
```

Prije odgovaranja na projektna pitanja, ispisat ćemo osnovne podatke o svim varijablama kako bismo se površno upoznali sa skupom podataka.

```
summary(tennis)
```

```
##        X          tourney_id        tourney_name         surface
##  Min.   :    0   Length:96602       Length:96602       Length:96602
##  1st Qu.:24150   Class :character   Class :character   Class :character
##  Median :48301   Mode  :character   Mode  :character   Mode  :character
##  Mean   :48301
##  3rd Qu.:72451
##  Max.   :96601
##
##    draw_size      tourney_level       tourney_date        match_num
##  Min.   :  4.00   Length:96602       Min.   :19901231   Min.   :   1.00
##  1st Qu.: 32.00   Class :character   1st Qu.:19970421   1st Qu.:   9.00
##  Median : 32.00   Mode  :character   Median :20040614   Median :  22.00
##  Mean   : 52.75                      Mean   :20044619   Mean   :  58.42
##  3rd Qu.: 64.00                      3rd Qu.:20120220   3rd Qu.:  48.00
##  Max.   :128.00                      Max.   :20201116   Max.   :1701.00
##
##    winner_id       winner_seed    winner_entry       winner_name
##  Min.   :100284   Min.   : 1.00   Length:96602       Length:96602
##  1st Qu.:102025   1st Qu.: 3.00   Class :character   Class :character
##  Median :103344   Median : 5.00   Mode  :character   Mode  :character
##  Mean   :104291   Mean   : 6.85
##  3rd Qu.:104571   3rd Qu.: 8.00
##  Max.   :210013   Max.   :35.00
##                   NA's   :57628
##  winner_hand          winner_ht      winner_ioc          winner_age
##  Length:96602       Min.   :160.0   Length:96602       Min.   :14.35
##  Class :character   1st Qu.:180.0   Class :character   1st Qu.:22.96
##  Mode  :character   Median :185.0   Mode  :character   Median :25.49
##                     Mean   :185.5                      Mean   :25.74
##                     3rd Qu.:190.0                      3rd Qu.:28.24
##                     Max.   :211.0                      Max.   :42.79
##                     NA's   :4070                       NA's   :58
##     loser_id        loser_seed    loser_entry         loser_name
##  Min.   :100282   Min.   : 1.0   Length:96602       Length:96602
##  1st Qu.:102035   1st Qu.: 4.0   Class :character   Class :character
##  Median :103333   Median : 6.0   Mode  :character   Mode  :character
##  Mean   :104537   Mean   : 8.2
##  3rd Qu.:104594   3rd Qu.:11.0
##  Max.   :210013   Max.   :35.0
```

```
##                   NA's   :75349
##   loser_hand          loser_ht      loser_ioc          loser_age
##  Length:96602       Min.   :160.0   Length:96602       Min.   :14.51
##  Class :character   1st Qu.:180.0   Class :character   1st Qu.:23.00
##  Mode  :character   Median :185.0   Mode  :character   Median :25.63
##                     Mean   :185.1                      Mean   :25.82
##                     3rd Qu.:190.0                      3rd Qu.:28.42
##                     Max.   :211.0                      Max.   :46.04
##                     NA's   :7671                       NA's   :127
##     score              best_of         round              minutes
##  Length:96602       Min.   :3.000   Length:96602       Min.   :   0.0
##  Class :character   1st Qu.:3.000   Class :character   1st Qu.:  74.0
##  Mode  :character   Median :3.000   Mode  :character   Median :  96.0
##                     Mean   :3.446                      Mean   : 102.8
##                     3rd Qu.:3.000                      3rd Qu.: 124.0
##                     Max.   :5.000                      Max.   :1146.0
##                                                        NA's   :12410
##      w_ace              w_df            w_svpt          w_1stIn
##  Min.   :  0.000   Min.   : 0.000   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  3.000   1st Qu.: 1.000   1st Qu.: 56.00   1st Qu.: 34.00
##  Median :  5.000   Median : 2.000   Median : 73.00   Median : 44.00
##  Mean   :  6.493   Mean   : 2.745   Mean   : 78.03   Mean   : 47.44
##  3rd Qu.:  9.000   3rd Qu.: 4.000   3rd Qu.: 94.00   3rd Qu.: 58.00
##  Max.   :113.000   Max.   :26.000   Max.   :491.00   Max.   :361.00
##  NA's   :9793      NA's   :9793     NA's   :9793     NA's   :9793
##     w_1stWon         w_2ndWon        w_SvGms        w_bpSaved
##  Min.   :  0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 26.00   1st Qu.:12.00   1st Qu.: 9.00   1st Qu.: 1.00
##  Median : 33.00   Median :16.00   Median :11.00   Median : 3.00
##  Mean   : 35.76   Mean   :16.79   Mean   :12.38   Mean   : 3.53
##  3rd Qu.: 43.00   3rd Qu.:21.00   3rd Qu.:15.00   3rd Qu.: 5.00
##  Max.   :292.00   Max.   :82.00   Max.   :90.00   Max.   :24.00
##  NA's   :9793     NA's   :9793    NA's   :9793    NA's   :9793
##    w_bpFaced          l_ace             l_df            l_svpt
##  Min.   : 0.000   Min.   :  0.000   Min.   : 0.000   Min.   :  0.00
##  1st Qu.: 2.000   1st Qu.:  2.000   1st Qu.: 2.000   1st Qu.: 59.00
##  Median : 4.000   Median :  4.000   Median : 3.000   Median : 76.00
##  Mean   : 5.174   Mean   :  4.806   Mean   : 3.502   Mean   : 80.85
##  3rd Qu.: 7.000   3rd Qu.:  7.000   3rd Qu.: 5.000   3rd Qu.: 97.00
##  Max.   :34.000   Max.   :103.000   Max.   :26.000   Max.   :489.00
##  NA's   :9793     NA's   :9793      NA's   :9793     NA's   :9793
##     l_1stIn          l_1stWon        l_2ndWon         l_SvGms
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 34.00   1st Qu.: 22.00   1st Qu.: 10.00   1st Qu.: 9.00
##  Median : 44.00   Median : 29.00   Median : 14.00   Median :11.00
##  Mean   : 47.86   Mean   : 31.78   Mean   : 15.02   Mean   :12.18
##  3rd Qu.: 58.00   3rd Qu.: 39.00   3rd Qu.: 19.00   3rd Qu.:15.00
##  Max.   :328.00   Max.   :284.00   Max.   :101.00   Max.   :91.00
##  NA's   :9793     NA's   :9793     NA's   :9793     NA's   :9793
##    l_bpSaved        l_bpFaced       winner_rank      winner_rank_points
##  Min.   :-6.000   Min.   : 0.000   Min.   :   1.00   Min.   :    1
##  1st Qu.: 2.000   1st Qu.: 6.000   1st Qu.:  18.00   1st Qu.:  517
##  Median : 4.000   Median : 8.000   Median :  46.00   Median :  860
##  Mean   : 4.813   Mean   : 8.752   Mean   :  81.35   Mean   : 1387
```

```
##  3rd Qu.: 7.000   3rd Qu.:11.000   3rd Qu.:  89.00   3rd Qu.: 1551
##  Max.   :28.000   Max.   :35.000   Max.   :2101.00   Max.   :16950
##  NA's   :9793     NA's   :9793     NA's   :1040      NA's   :2032
##    loser_rank     loser_rank_points
##  Min.   :   1.0   Min.   :    1.0
##  1st Qu.:  37.0   1st Qu.:  385.0
##  Median :  71.0   Median :  639.0
##  Mean   : 119.9   Mean   :  867.6
##  3rd Qu.: 119.0   3rd Qu.: 1015.0
##  Max.   :2159.0   Max.   :16950.0
##  NA's   :2289     NA's   :3278
```

3 Projektna pitanja

U sklopu zadatka postavljena su određena pitanja, uz mogućnost postavljanja vlastitih pitanja i pokretanja dodatne problematike vezano za dani skup podataka. Sva su pitanja vezana uz gradivo koje se obrađuje na predmetu Statistička analiza podataka na Fakultetu elektrotehnike i računarstva.

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, profiltriraju se na način da se svaki igrač pojavljuje samo jednom.To je skup podataka koji će se koristiti u analizi i predstavljati najuspješnije igrače.

S obzirom na činjenicu da se u skupu podataka na nekim mjestima pojavljuju igrači kojima nije definiran ranking, postavit 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 desetoricom 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)
```

Nakon što smo izdvojili najuspješnije i one manje uspješne igrače u skupove podataka `mostSuccesfulPlayers` i `notSoSuccessfulPlayers`, možemo početi s analizom podataka. Za početak, ispisujemo neke osnovne informacije o jednim i drugim igračima kako bi čitatelj dobio sliku. Primjeri važnijih mjera centralne tendencije jesu aritmetička sredina i medijan.

```
summary(mostSuccessfulPlayers$player_ht)
```

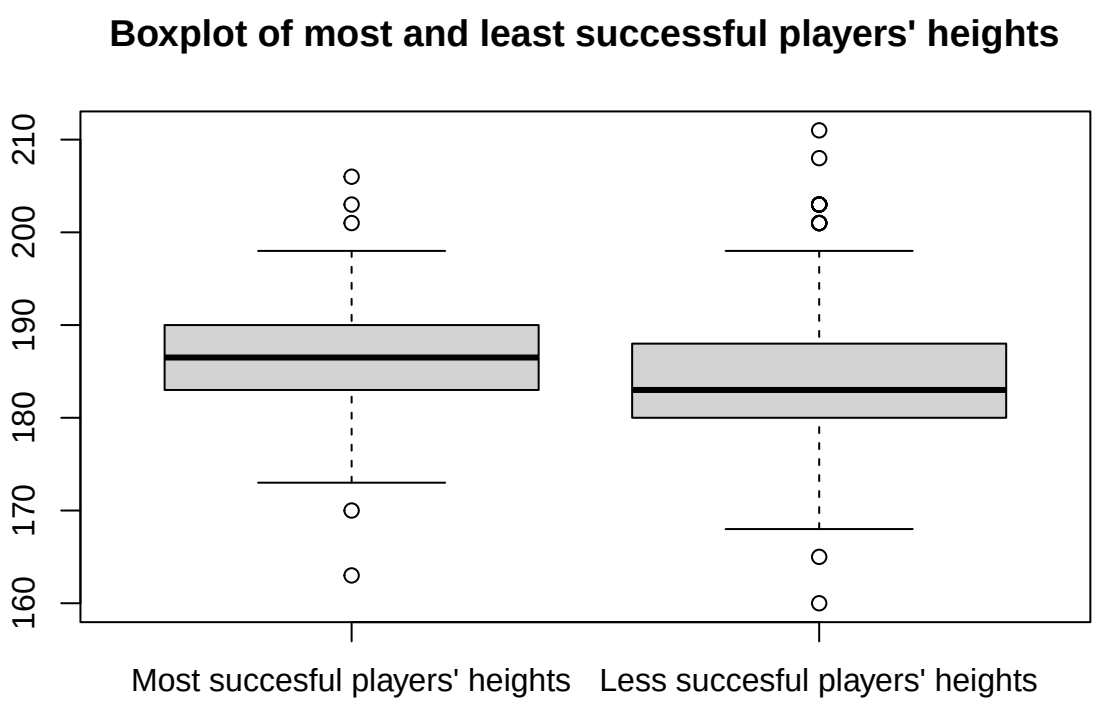
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   163.0   183.0   186.5   186.6   190.0   206.0
```

```
notSoSuccessfulPlayers$player_ht <- as.numeric(as.character(notSoSuccessfulPlayers$player_ht))
summary(notSoSuccessfulPlayers$player_ht)
```

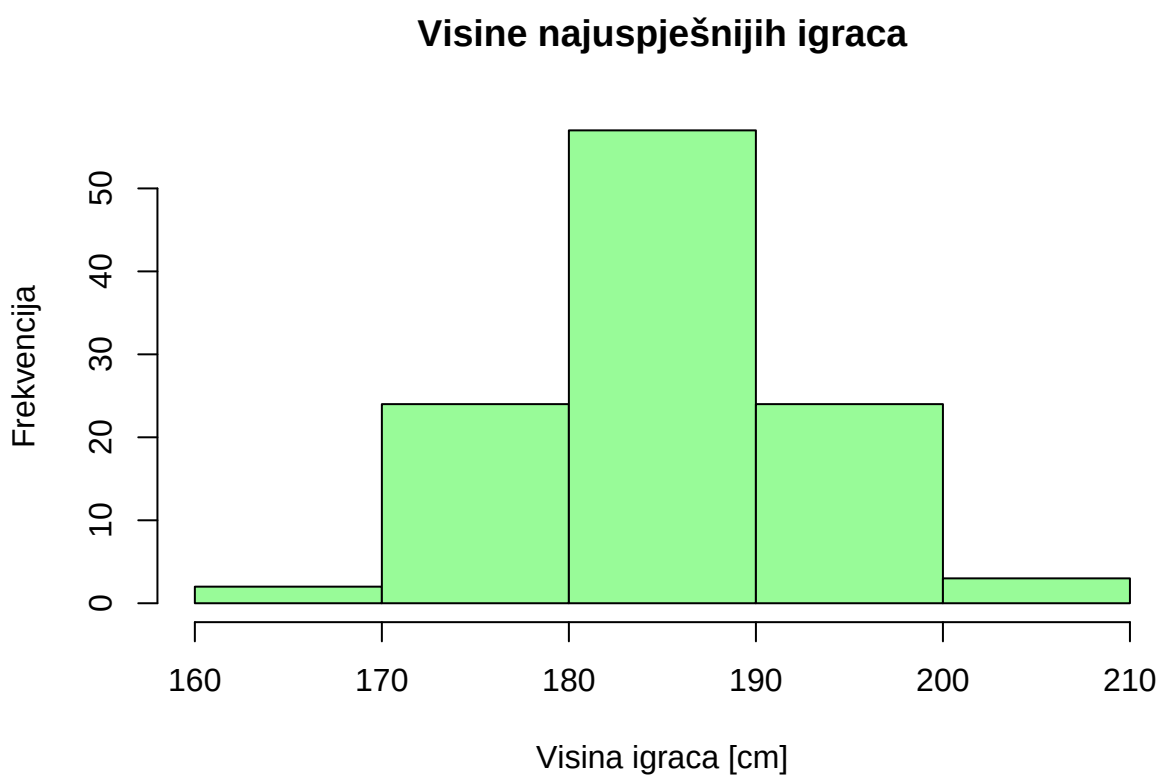
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     160     180     183     184     188     211
```

Vizualiziramo podatke, prvo za najuspješnije igrače, zatim za one manje uspješne.

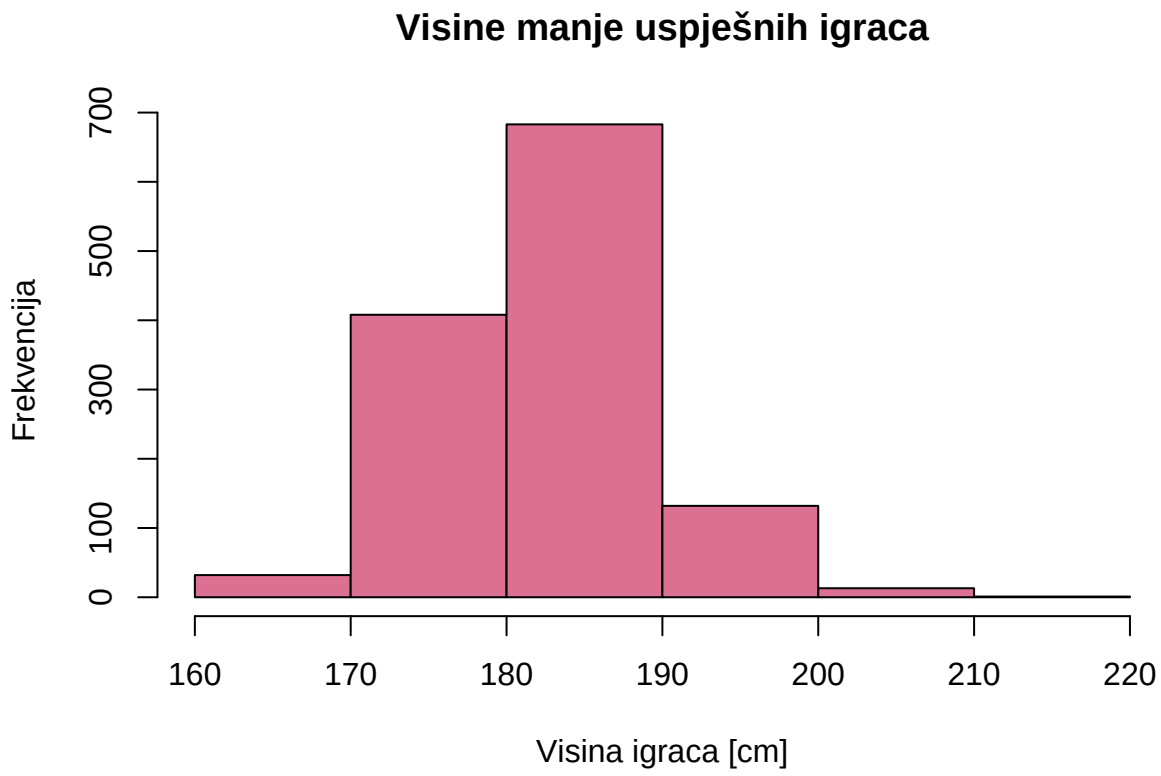
```
boxplot(mostSuccessfulPlayers$player_ht, notSoSuccessfulPlayers$player_ht,
        names = c('Most succesful players\' heights','Less succesful players\' heights'),
        main='Boxplot of most and least successful players\' heights')
```




```
h_mostSuccesfulPlayers = hist(mostSuccessfulPlayers$player_ht,
                              main = "Visine najuspješnijih igrača",
                              xlab = "Visina igrača [cm]",
                              ylab = "Frekvencija",
                              breaks = 5,
                              col = "palegreen")
```



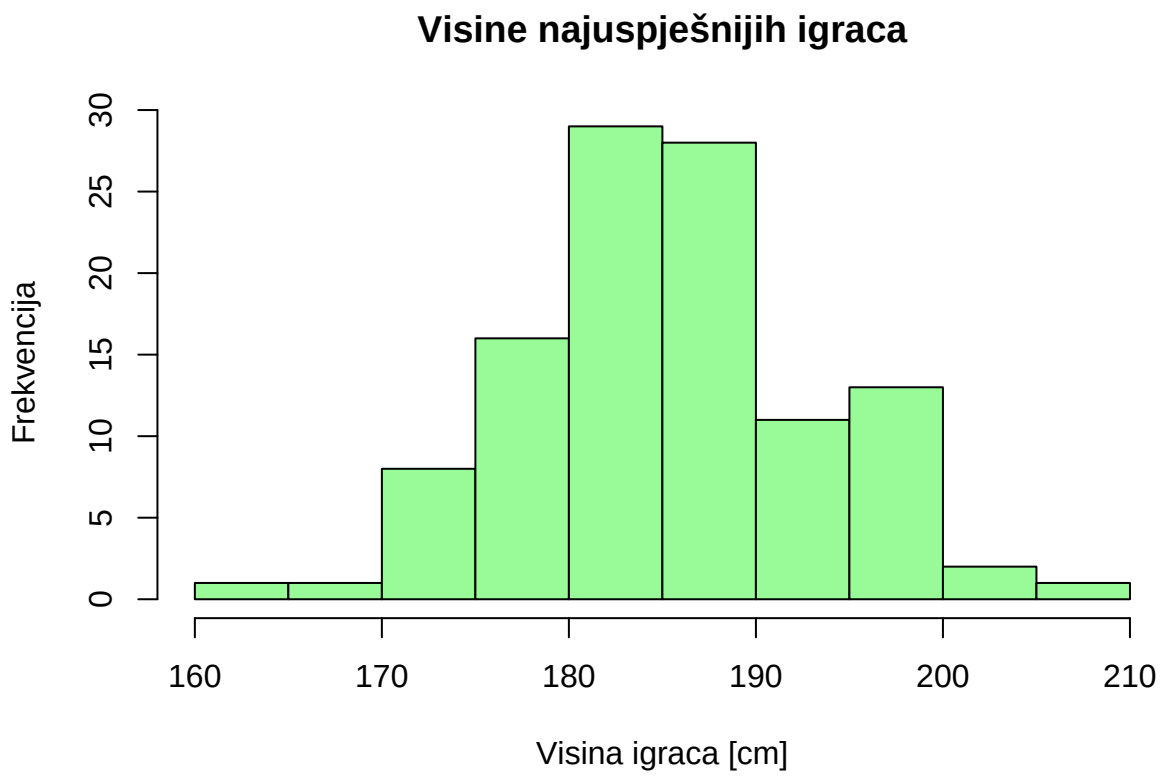

```
h_notSoSuccesfulPlayers = hist(notSoSuccessfulPlayers$player_ht,
                               main = "Visine manje uspješnih igrača",
                               xlab = "Visina igrača [cm]",
                               ylab = "Frekvencija",
                               breaks = 5,
                               col = "palevioletred")
```



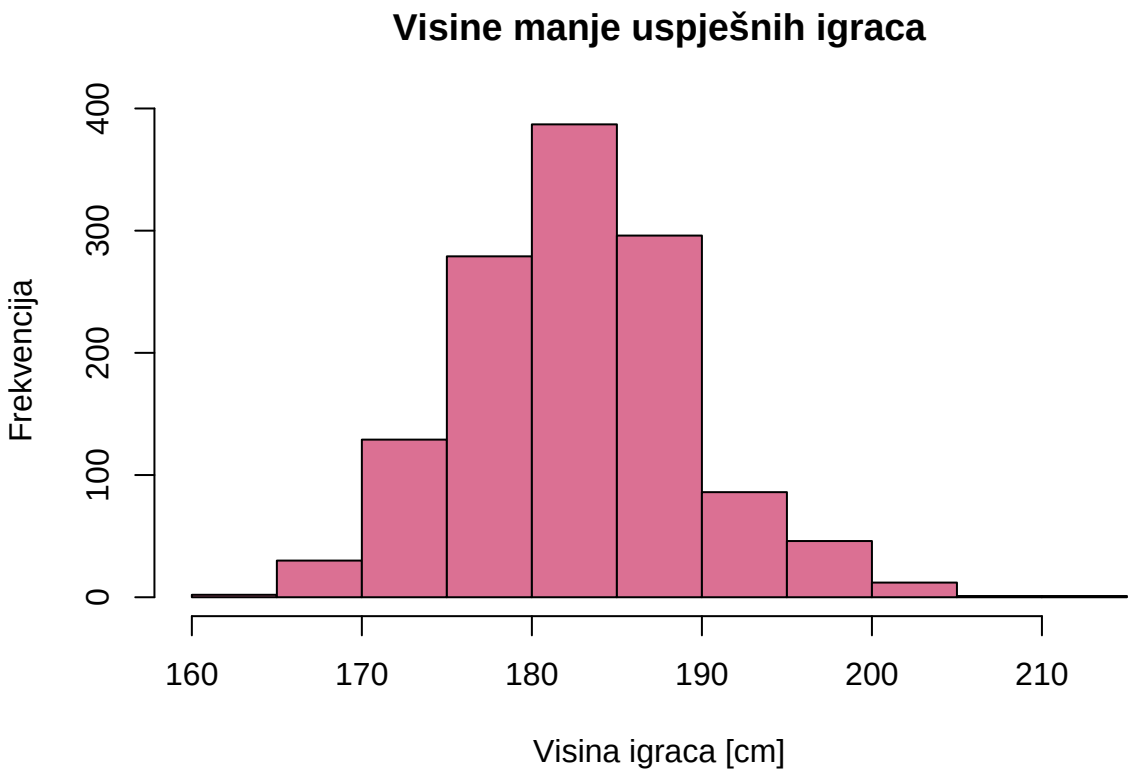
Ovaj je prikaz dosta grub, ali iz njega i dalje možemo izvući neke zaključke. Naime, usporedbom *boxplota* uviđamo da je srednja vrijednost visine nešto viša za najuspješnije igrače. Usporedbom histograma uviđamo da, iako je visina većine igrača i jedne i druge skupine između 180 i 190 cm, kod onih manje uspješnih igrača broj onih čija je visina manja od 180 cm znatno je veći od onih čija je visina veća od 190 cm, dok to kod najuspješnijih igrača nije slučaj.

Histogram vrijednosti visina najuspješnijih igrača ima zvonolik oblik, a pretpostavka je da bi i histogram vrijednosti visina manje uspješnih igrača imao sličan oblik ako se broj razreda poveća. Da bismo se u to uvjerili, možemo podatke prikazati histogramom s većim brojem razreda:

```
h2_mostSuccesfulPlayers = hist(mostSuccessfulPlayers$player_ht,
                               main = "Visine najuspješnijih igrača",
                               xlab = "Visina igrača [cm]",
                               ylab = "Frekvencija",
                               breaks = 10,
                               col = "palegreen")
```

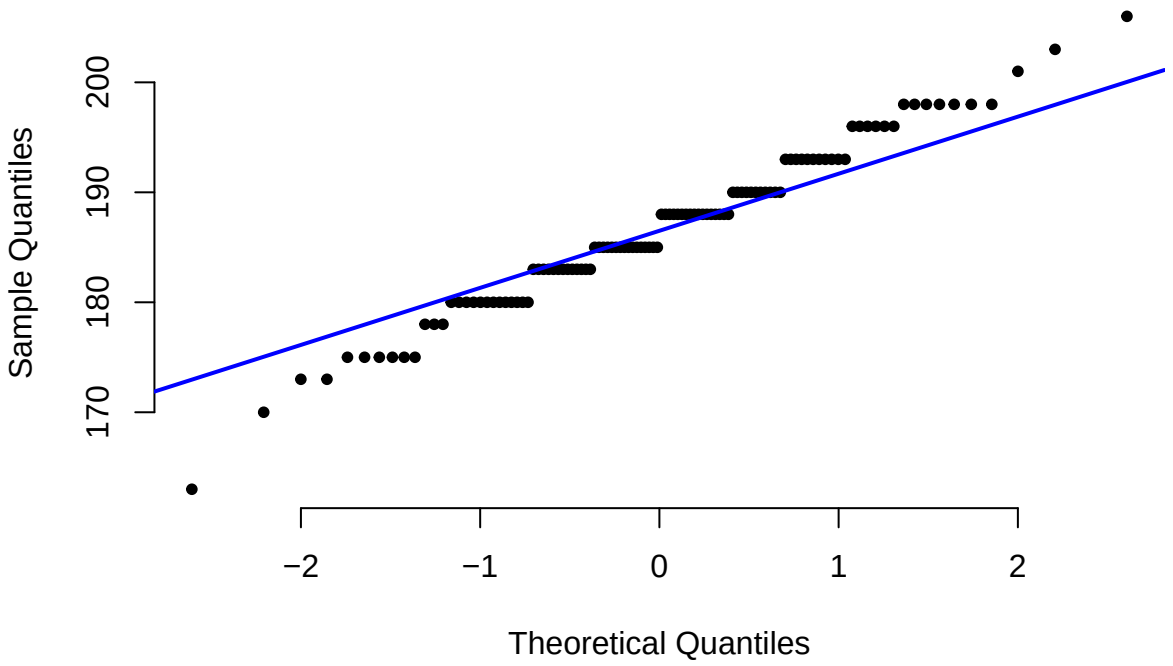



```
h_notSoSuccesfulPlayers = hist(notSoSuccessfulPlayers$player_ht,
                               main = "Visine manje uspješnih igrača",
                               xlab = "Visina igrača [cm]",
                               ylab = "Frekvencija",
                               breaks = 10,
                               col = "palevioletred")
```

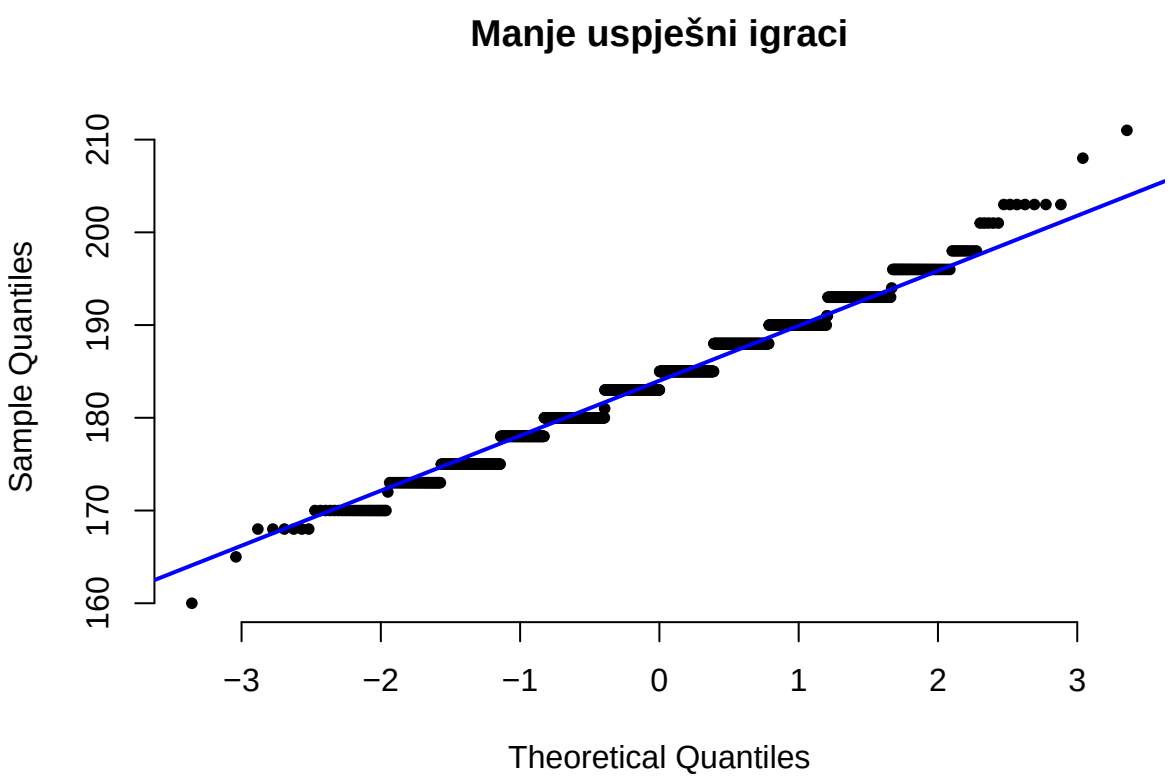


Oblik histograma upućuje na to da se podaci ravnaju po normalnoj razdiobi. Da bismo to sa sigurnošću mogli tvrditi, potrebno je provesti test normalnosti varijabli.

```
qqnorm(mostSuccessfulPlayers$player_ht, pch = 20, frame = FALSE,
       main='Najuspješniji igrači')
qqline(mostSuccessfulPlayers$player_ht, col = "blue", lwd = 2)
```




```
qqnorm(notSoSuccessfulPlayers$player_ht, pch = 20, frame = FALSE,
       main='Manje uspješni igrači')
qqline(notSoSuccessfulPlayers$player_ht, col = "blue", lwd = 2)
```



Čini se da podaci prate ravnu liniju pa možemo pretpostaviti da je razdioba normalna.

Pod tom pretpostavkom prvo ćemo se pozabaviti jednakošću varijanci, odnosno dokazivanjem iste, a zatim ćemo provesti testove i postaviti hipoteze o jednakosti prosječnih vrijednosti visina uspješnih i onih manje uspješnih igrača.

Imamo li na raspolaganju dva nezavisna slučajna uzorka $X_1^1, X_1^2, \dots X_1^{n_1}$ i $X_2^1, X_2^2, \dots, X_2^{n_2}$, pod pretpostavkom da oni dolaze iz populacija s normalnom razdiobom i varijancama σ_1^2 i σ_2^2, tada slučajna varijabla

$$F = \frac{S_{X_1}^2/\sigma_1^2}{S_{X_2}^2/\sigma_2^2}$$

ima Fisherovu razdiobu s $(n_1 - 1, n_2 - 1)$ stupnjeva slobode pri čemu vrijedi:

$$S_{X_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_1^i - \bar{X}_1)^2, \quad S_{X_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_2^i - \bar{X}_2)^2.$$

Hipoteze testa jednakosti varijanci glase:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2 \quad , \quad \sigma_1^2 > \sigma_2^2 \quad , \quad \sigma_1^2 \neq \sigma_2^2$$

Ispitujemo jednakost varijanci prikupljenih uzoraka:

```
var.test(mostSuccessfulPlayers$player_ht, notSoSuccessfulPlayers$player_ht)
```

```
##
##  F test to compare two variances
##
```

```
## data:  mostSuccessfulPlayers$player_ht and notSoSuccessfulPlayers$player_ht
## F = 1.3039, num df = 109, denom df = 1268, p-value = 0.04659
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.003946 1.750071
## sample estimates:
## ratio of variances
##           1.303941
```

p-vrijednost od 0.04659 govori nam da ne možemo odbaciti hipotezu o jednakosti varijanci, tj. da varijance možemo smatrati jednakima.

Sada se možemo pozabaviti srednjim vrijednostima visine.

Imamo li na raspolaganju dva nezavisna slučajna uzorka $X_1^1, X_1^2, \ldots, X_1^{n_1}$ i $X_2^1, X_2^2, \ldots, X_2^{n_2}$, pod pretpostavkom da oni dolaze iz populacija s normalnom razdiobom s očekivanjima μ_1 i μ_2 te s nepoznatim, ali jednakim varijancama σ. Zajednička disperzija uzorka računa se kao težinska sredina disperzija S_{X_1} i S_{X_2}:

$$S_X^2 = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2].$$

Slučajna varijabla

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ima jediničnu normalnu razdiobu, a slučajna varijabla

$$W^2 = \frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{\sigma^2}$$

ima χ^2 razdiobu s $n_1 + n_2 - 2$ stupnja slobode. Iz tog razloga možemo reći da slučajna varijabla

$$T = \frac{Z\sqrt{n_1 + n_2 - 2}}{W} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_X\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ima egzaktnu t razdiobu s $n_1 + n_2 - 2$ stupnja slobode.

Hipoteze testa jednakosti srednjih vrijednosti glase:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2 \quad , \quad \mu_1 > \mu_2 \quad , \quad \mu_1 \neq \mu_2$$

Test možemo provesti samo pod pretpostavkom da uzorak dolazi iz populacije koja prati normalnu razdiobu, što imamo. Također, uzorci moraju biti nezavisni, što je i slučaj kod igrača koji dolaze iz dvije različite skupine.

Provedimo sada t-test uz pretpostavku jednakosti varijanci:

```
t.test(mostSuccessfulPlayers$player_ht, notSoSuccessfulPlayers$player_ht,
       alt = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  mostSuccessfulPlayers$player_ht and notSoSuccessfulPlayers$player_ht
## t = 3.9518, df = 1377, p-value = 4.075e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  1.495637      Inf
## sample estimates:
## mean of x mean of y
##  186.5727  184.0095
```

p-vrijednost od 0.00004075 govori nam da trebamo odbaciti nultu hipotezu, odnosno da smijemo zaključiti da su najuspješniji tenisači u prosjeku znatno viši od onih manje uspješnih tenisača.

3.1.2 Odnos ljevaka i dešnjaka

Postavljeno pitanje bilo je: Jesu li ljevaci nezgodniji protivnici dešnjacima koji igraju jednoručni backhand?

Pitanje je interpretirano na malo drugačiji način - provjeravat će se jesu li dešnjaci zapravo nezgodniji protivnici od ljevaka.

Postavljene su hipoteze:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 < p_2$$

Nulta hipoteza glasi da je dešnjacima s jednoručnim backhandom jednako nezgodno igrati protiv dešnjaka i ljevaka, dok je alternativna hipoteza da su dešnjaci su nezgodniji protivnici dešnjacima koji igraju s jednoručnim backhandom.

Jesu li protivnici zahtjevni, lagani, nezgodniji i slično promatrat će se po broju pobjeda naspram broja odigranih mečeva. Tu se pojavljuju dva uzorka i oba uzorka će biti ispunjena isključivo sa dva podatka, pobjeda ili poraz. Našu hipotezu i takve podatke odlučili smo testirati Testom o dvije proporcije: dva uzorka

Potreban nam je popis svih igrača koji igraju jedoručni backhand.

```
library(rvest)
stranica <- read_html("http://www.tennisdrawchallenge.com/data/list/one-handed-backhand")
tables <- stranica %>% html_table(fill = TRUE)
jedno_back <- tables[[1]]
igraci_1HBH <- jedno_back["Name"]
lista_1HBH = igraci_1HBH[["Name"]]
```

Pomoću dobivenog popisa, svi mečevi bit će izdvojeni u dvije tablice. U jednoj će biti mečevi između igrača koji su dešnjaci i preferiraju jednoručni backhand i igrača koji su ljevaci, a u drugoj tablici će biti biti mečevi između igrača koji su dešnjaci i preferiraju jednoručni backhand i igrača koji su dešnjaci, ali ne igraju jednoručnim backhandom.

```
jedno_back_mecevi_protiv_L = subset(tennis, ((as.character(winner_name) %in% lista_1HBH &
                                        !(as.character(loser_name) %in% lista_1HBH) &
                                        as.character(winner_hand)=="R" & as.character(loser_hand)=="L")
                                        (as.character(loser_name) %in% lista_1HBH &
                                          as.character(loser_hand)=="R" & as.character(winner_hand)=="L
                                          !(as.character(winner_name) %in% lista_1HBH) )))
jedno_back_mecevi_protiv_R = subset(tennis, ((as.character(winner_name) %in% lista_1HBH &
                                        as.character(winner_hand)=="R" & as.character(loser_hand)=="R")
                                        (as.character(loser_name) %in% lista_1HBH &
                                          as.character(loser_hand)=="R" & as.character(winner_hand)=="R
```

Sada je potrebno proći kroz sve pojedine mečeve i provjeriti jesu li igrači s jednoručnim backhandom pobjedili(1) ili izgubili (-1).

```
rezultati_L <- data.frame(value = numeric())
rezultati_R <-data.frame(value = numeric())
k1=0
k2=0
```

```
for (i in 1:nrow(jedno_back_mecevi_protiv_L)){
  if (as.character(jedno_back_mecevi_protiv_L[i,"winner_name"] %in% lista_1HBH)){
    rezultati_L[i,"value"] <- 1
    k1=k1+1
  } else {
    rezultati_L[i,"value"] <- 0
  }
}
for (i in 1:nrow(jedno_back_mecevi_protiv_R)){
  if (as.character(jedno_back_mecevi_protiv_R[i,"winner_name"] %in% lista_1HBH)){
    rezultati_R[i,"value"] <- 1
    k2=k2+1
  } else {
    rezultati_R[i,"value"] <- 0
  }
}
n1=nrow(rezultati_L)
n2=nrow(rezultati_R)
```

Nad tim podacima provodi se Test o dvije proporcije (k_1, k_2 predstavljaju broj pobjeda, a n_1, n_2 broj mečeva). Korištena je Z-statistika:

$$Z = \frac{\frac{k1}{n1} - \frac{k2}{n2}}{\sqrt{p(1-p)(\frac{1}{n1} + \frac{1}{n2})}} \sim N(0,1)$$

Slijedi računanje z_vrijednosti i p-vrijednosti.

```
z_vrijednost = (k1/n1 - k2/n2)/sqrt(((k1+k2)/(n1+n2))*(1-(k1+k2)/(n1+n2))*(1/k1+1/k2))
z_vrijednost
```

```
## [1] -4.991745
```

```
pnorm(z_vrijednost)
```

```
## [1] 2.991818e-07
```

p-vrijednost manja od 0.01 upućuje na odbacivanje nulte hipoteze u korist prve hipoteze na razini značajnosti $\alpha = 0.01$. Zaključujemo da su igračima s desnim jednoručnim backhandom nezgodniji protivnici dešnjaci koji ne preferiraju jednoručni backhand.

```
ruke = tennis[c("winner_hand", "loser_hand")]
razl_ruke = subset(ruke,(as.character(winner_hand)!=as.character(loser_hand) &
                           as.character(loser_hand)!= "U" &
                           as.character(winner_hand)!= "U" &
                           as.character(loser_hand)!= "" &
                           as.character(winner_hand)!= ""))
grupiraneruke = subset(as.data.frame(table(razl_ruke)), Freq != 0)
grupiraneruke$Freq[1]/nrow(razl_ruke)
```

```
## [1] 0.5094129
```

```
grupiraneruke$Freq[2]/nrow(razl_ruke)
```

```
## [1] 0.4905871
```

3.1.3 Pobjeda prvog seta

Postavljeno pitanje bilo je: Možemo li na temelju dobitnika prvog seta predvidjeti dobitnika cijelog meča? Ono je također preoblikovano te se u ovom dijelu zapravo ispituje možemo li reći da je dobitnik prvog seta bolji igrač.

Zadana je nulta hipoteza koja glasi da su igrači jednako dobri odnosno imaju jednaku vjerojatnost dobitka pojedinog seta. Ovdje pretpostavljamo da svi igrači jednako igraju u svim setovima, iako će se neki igrači npr. brže umoriti. Sada je alternativna hipoteza da je igrač koji je dobio prvi set bolji igrač.

U nastavku promatramo samo *best of 3* mečeve. Vjerojatnost da dobitnik prvog seta pobijedi uz uvjet da je H0 istinita je $0,5 + 0,5^2 = 0,75$. To znači da od n mečeva očekujemo da će dobitnik prvog seta pobijediti u njih $0,75 * n$. Broj takvih mečeva je varijabla podvrgnuta binomnoj razdiobi s parametrima n i $p = 0,75$. Budući da nam je na raspolaganju mnoštvo podataka, možemo binomnu razdiobu aproksimirati normalnom s parametrima np i npq.

Sada možemo hipoteze postaviti na sljedeći način:

$$H_0 : p = 0,75$$
$$H_1 : p > 0,75$$

gdje je p vjerojatnost pobjede prvog igrača odnosno onoga koji je odnio pobjedu u prvom setu.

Izračunajmo sada postotak mečeva u kojem je dobitnik prvog seta dobio meč te p-vrijednost za distribuciju $N(np, npq)$, gdje je $p = 0,75$, a $q = 0,25$.

```
full_sets <- tennis[!grepl("[A-Za-z]", tennis$score),]
bo3 = full_sets[full_sets$best_of == 3,]
podatkovna_vjerojatnost = sum(substr(bo3$score, 1, 1) > substr(bo3$score, 3, 3)) / nrow(bo3)
n = nrow(bo3)
p = 0.75
o = sum(substr(bo3$score, 1, 1) > substr(bo3$score, 3, 3))
print(podatkovna_vjerojatnost)
```

```
## [1] 0.8151183
```

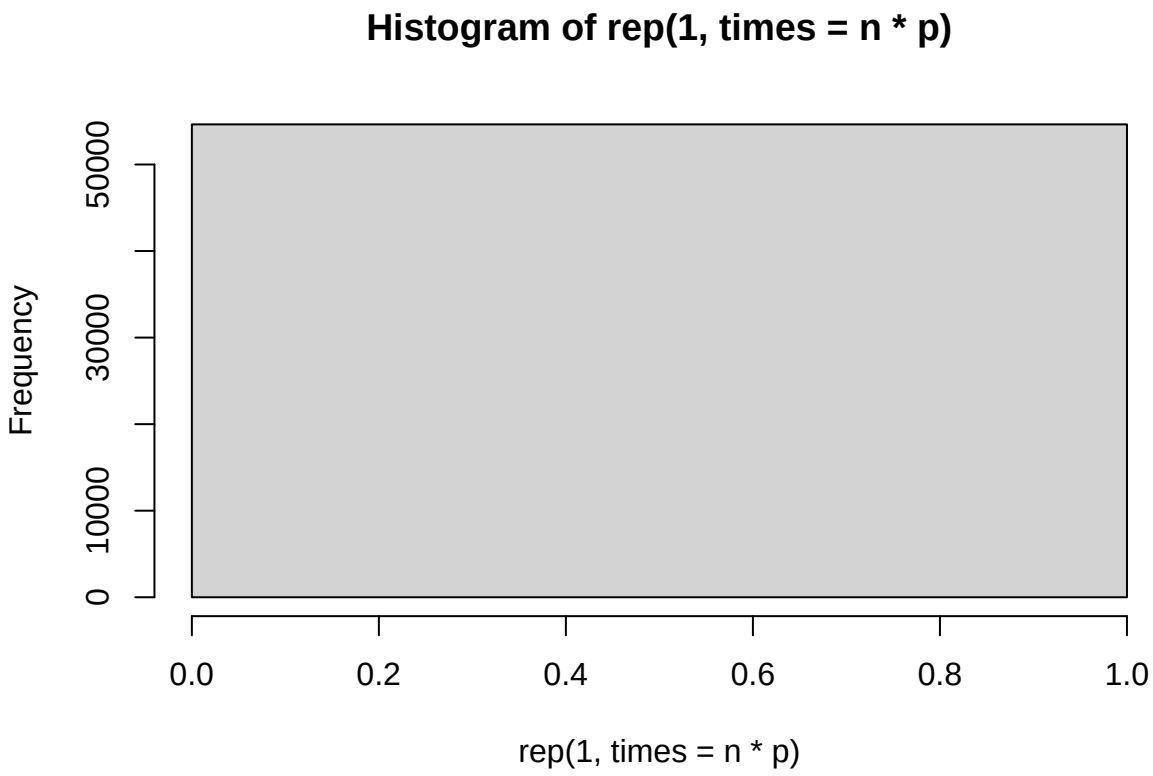
```
pnorm(o, mean = n * p, sd = sqrt(n * p * (1 - p)), lower.tail = FALSE)
```

```
## [1] 0
```

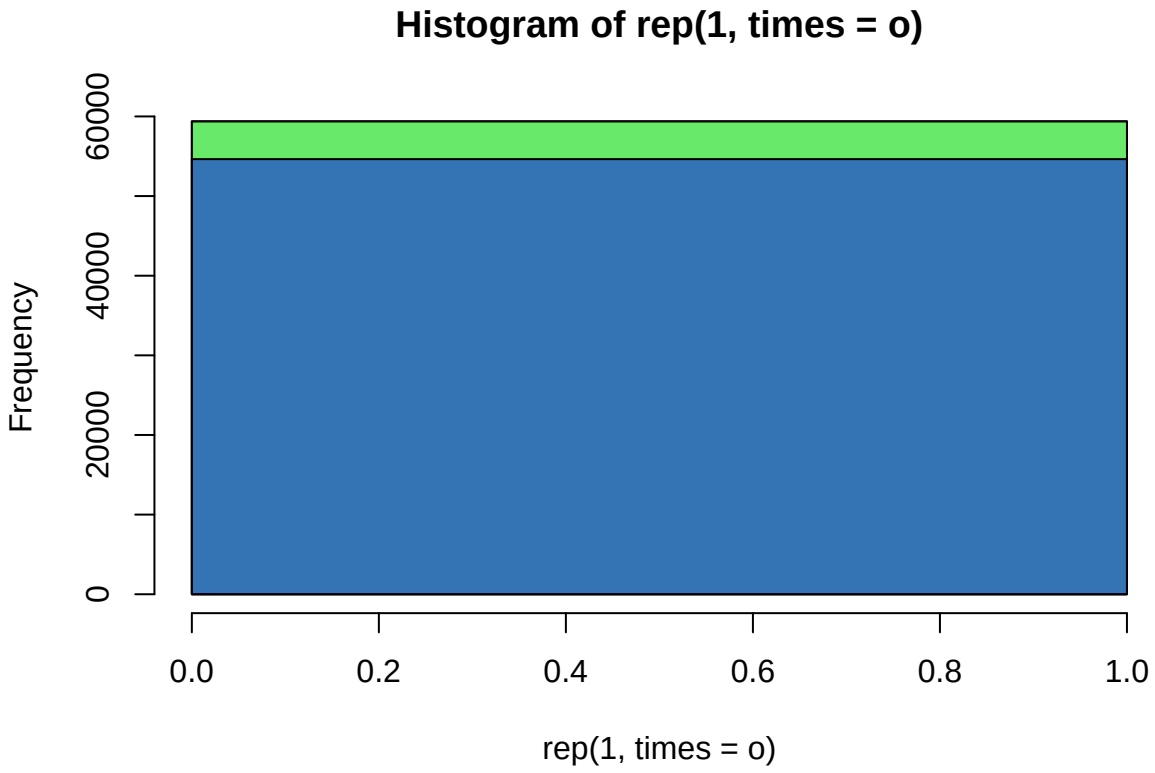
Vidimo da postotak takvih mečeva iznosi otprilike 82%; p-vrijednost je praktički 0. Odbacujemo nultu hipotezu na jako velikoj razini značajnosti i zaključujemo da su dobitnici prvog seta u prosjeku bolji igrači (imaju veću vjerojatnost dobitka pojedinog seta).

Rezultate ovog testa možemo prikazati i grafički:

```
p1 <- hist(rep(1, times = n * p))
```




```
p2 <- hist(rep(1, times = o))
plot(p2, col=rgb(0,1,0,1/2), add=T)
plot(p1, col=rgb(0,0,1,1/2), add=T)
```



Zelenom bojom je prikazan višak opaženih u odnosu na očekivane mečeve. Razlika je manja od 5000 mečeva, ali je s danim parametrima distribucije i više nego dovoljna da opovrgne H_0.

Promotrimo vezu između razlike u rangu i trajanja meča - pretpostavka je da igrači sličnijeg ranga igraju dulje mečeve.

Izdvajanje podataka i grafiranje:

```
library(dplyr)
```

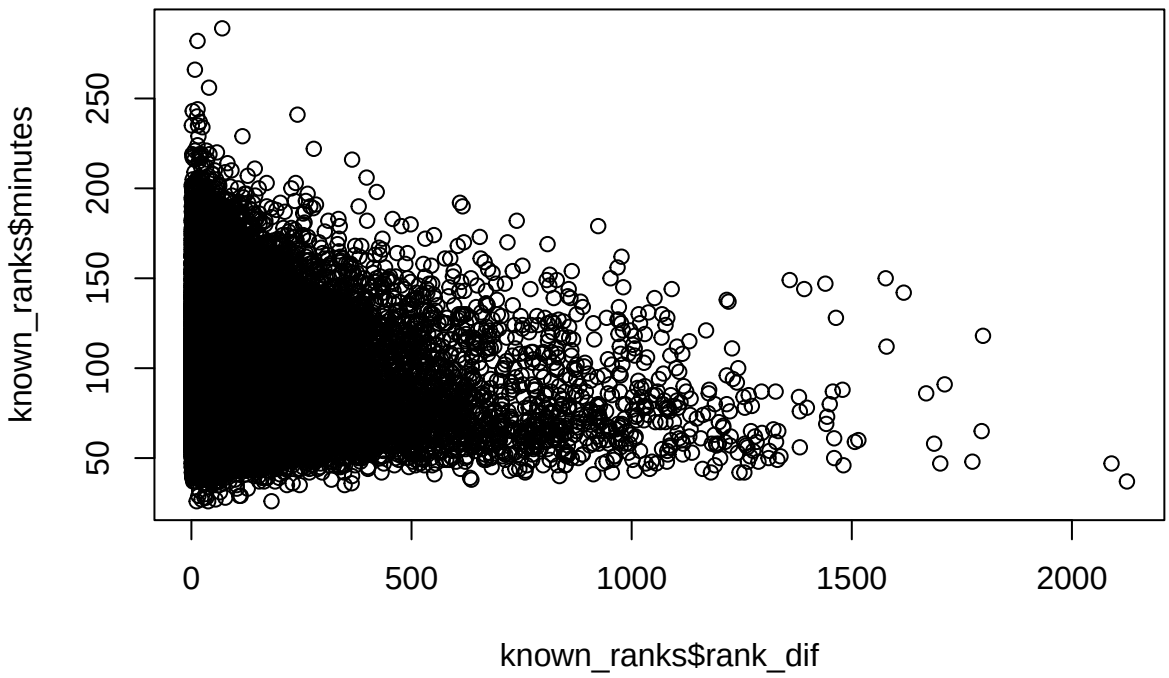
```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

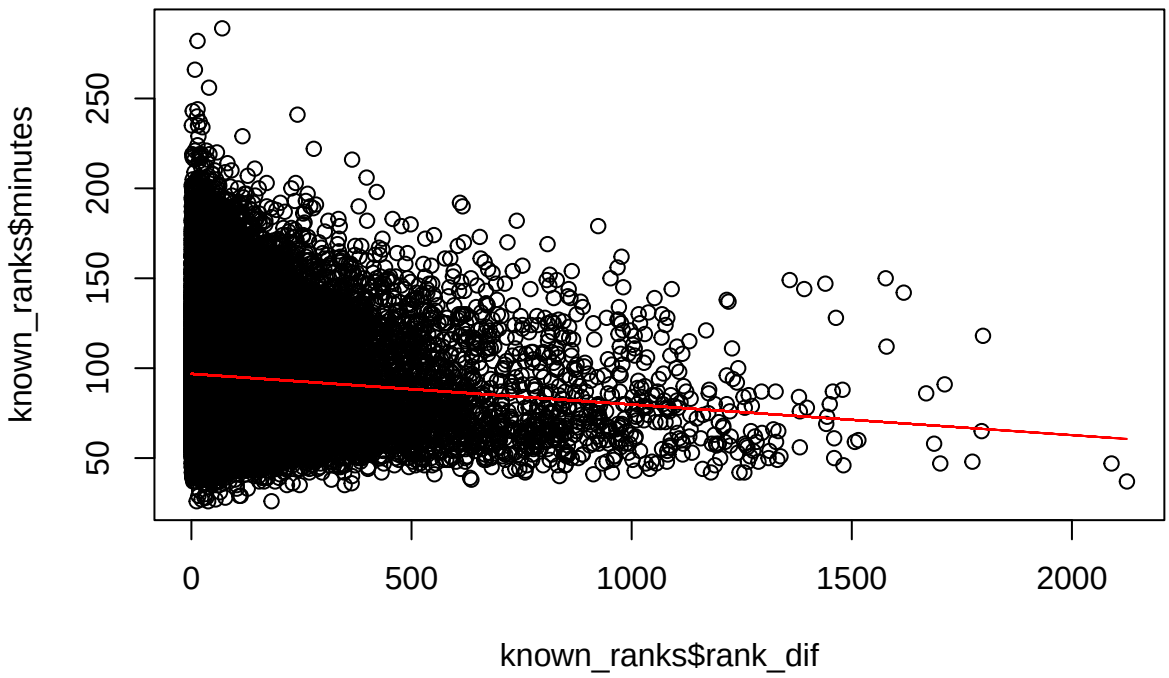
```
known_ranks = na.omit(bo3[,c("minutes", "winner_rank", "loser_rank")])
known_ranks = arrange(known_ranks[known_ranks$minutes < 500,], desc(minutes)) # postoje par nerealnih m
known_ranks$rank_dif = abs(known_ranks$winner_rank - known_ranks$loser_rank)
plot(known_ranks$rank_dif, known_ranks$minutes)
```



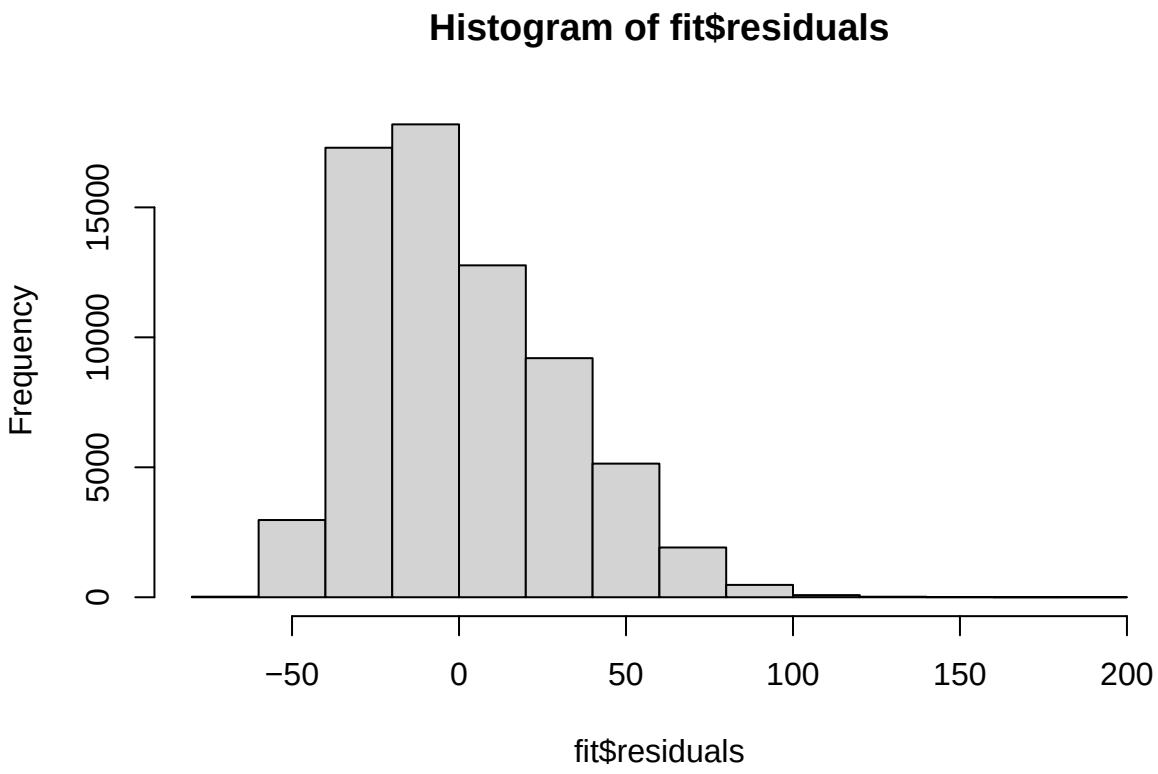
Vidimo da podaci imaju padajući trend, ali također vidimo i kratke mečeve slično rangiranih igrača. To možemo objasniti količinom podatka, pogotovo onih gdje su igrači sličnog ranga. Zato ćemo nastaviti s provođenjem regresijskog testa.

Izgradimo sada regresijski model:

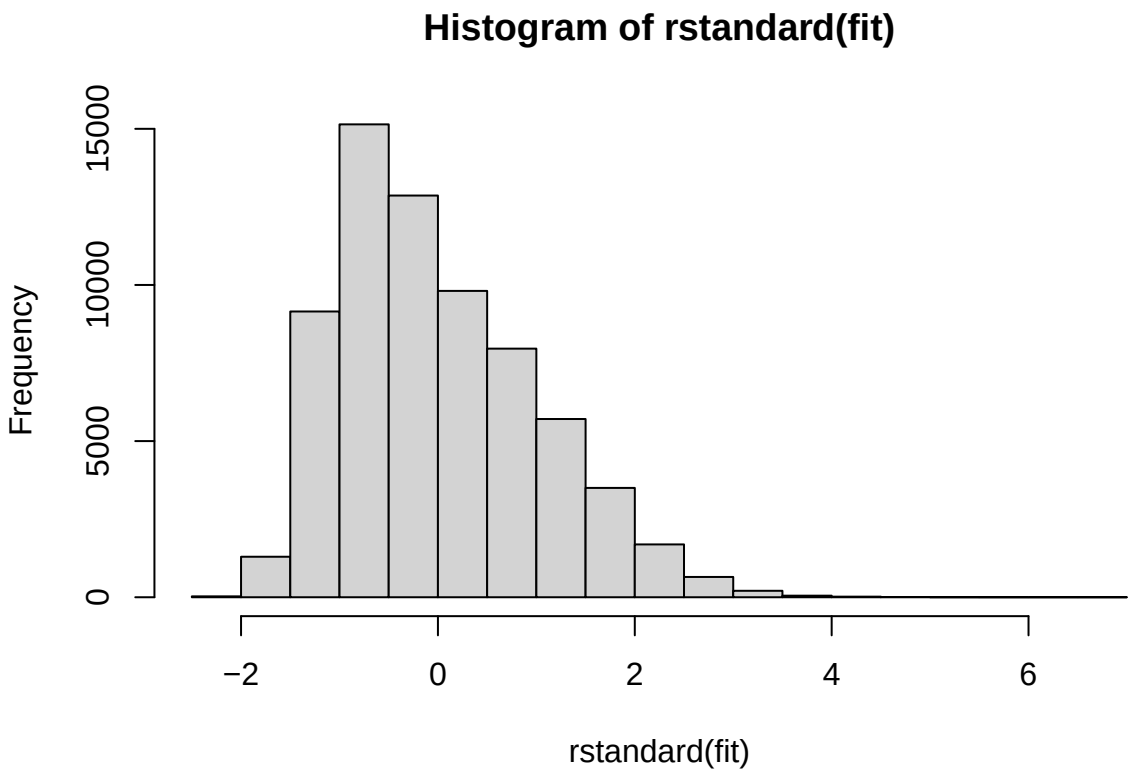
```
fit = lm(minutes~rank_dif, data = known_ranks)
plot(known_ranks$rank_dif, known_ranks$minutes)
lines(known_ranks$rank_dif, fit$fitted.values, col = 'red')
```




```
hist(fit$residuals)
```




```
hist(rstandard(fit))
```



Pogledamo li histograme reziduala, vidimo da su nakošeni, ali poznato je da je t-distribucija robusna na nenormalnosti pa ćemo svejedno provesti test do kraja.

```
summary(fit)
```

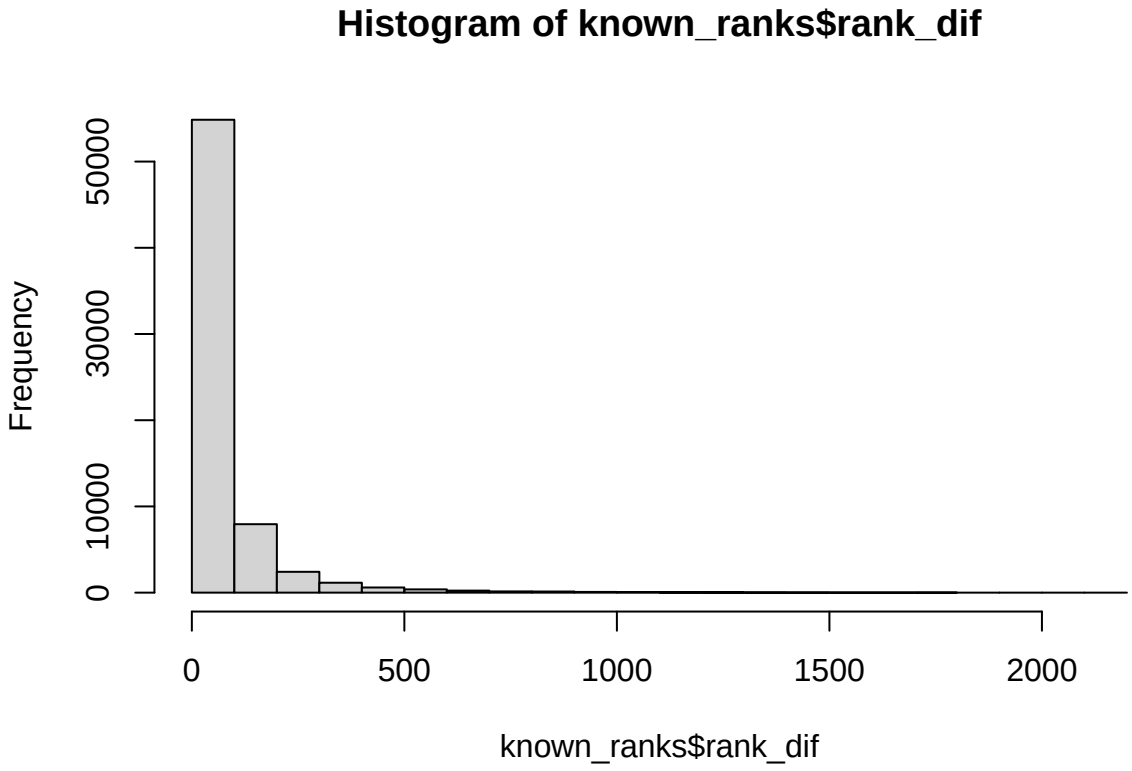
```
##
## Call:
## lm(formula = minutes ~ rank_dif, data = known_ranks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.628 -23.100  -5.577  19.610 193.360
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 96.8327249  0.1346822  718.97   <2e-16 ***
## rank_dif    -0.0170334  0.0009418  -18.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.63 on 68092 degrees of freedom
## Multiple R-squared:  0.004781,   Adjusted R-squared:  0.004766
## F-statistic: 327.1 on 1 and 68092 DF,  p-value: < 2.2e-16
```

Kao što vidimo, dobivamo vrlo mali R^2 te ne možemo zaključiti da su varijable linearno zavisne. Preduvjeti nisu bili dobri, a prividan padajući trend možemo objasniti i brojem odigranih mečeva u kojima su igrači imali sličan rang. Naime, veći ekstremi se prirodno postižu na mjestima brojnijeg uzorkovanja. Na grafu

možemo uočiti gornju envelopu koja uistinu ima padajući trend, ali također postoji i donja envelopa koja nema takav trend. Mečevi blizu gornje envelope su rijetki i ne utječu toliko na regresiju koliko utječu oni drugi, čisto zbog svoje brojnosti.

Da bismo se uvjerili da je ovaj prividan padajući trend uistinu rezultat neravnomjerne raspoređenosti podataka po horizontalnoj osi, razmotrimo sljedeći graf:

```
hist(known_ranks$rank_dif)
```



koji potvrđuje tu teoriju.

3.2 Vlastita pitanja

3.2.1 Vrsta podloge

Postavljeno pitanje bilo je: Je li odustajanje od meča nezavisno od vrste podloge na kojoj se on igra?

Postavljena je nulta hipoteza koja glasi da su podloga teniskog terena i odustajanje od meča nezavisne varijable, dok je alternativa da su to zavisne varijable.

```
podloge <- subset(tennis, as.character(surface)!="")
podloge <- droplevels(podloge)
podloge$score <- ifelse(grepl("RET", podloge$score), "Retired", "Not retired")
```

Pogledajmo kontingencijsku tablicu varijabli podloge i odustajanja od meča:

```
tbl = table(podloge$surface, podloge$score)
tbl
```

```
##
```

```
##          Not retired Retired
##   Carpet        7089     126
##   Clay         31528     867
##   Grass         9013     237
##   Hard         46222    1322
```

Kontingencijskoj tablici dodajemo sume redaka i stupaca:

```
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##          Not retired Retired   Sum
##   Carpet        7089     126  7215
##   Clay         31528     867 32395
##   Grass         9013     237  9250
##   Hard         46222    1322 47544
##   Sum          93852    2552 96404
```

Test nezavisnosti χ^2 test u programskom paketu R implementiran je u funkciji `chisq.test()` koja kao ulaz prima kontingencijsku tablicu podataka koje testiramo na nezavisnost. Ispitat ćemo nezavisnost podloge teniskog terena i odustajanja od meča.

Pretpostavka testa je da očekivana frekvencija pojedinog razreda mora biti veća ili jednaka 5 (`chisq.test()` pretpostavlja da je ovaj uvjet zadovoljen stoga je prije provođenja testa potrebno to provjeriti).

```
for (col_names in colnames(added_margins_tbl)){
  for (row_names in rownames(added_margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,': ',(added_margins_tbl[row_names,
    }
  }
}
```

```
## Ocekivane frekvencije za razred  Not retired - Carpet :  7024.005
## Ocekivane frekvencije za razred  Not retired - Clay :  31537.44
## Ocekivane frekvencije za razred  Not retired - Grass :  9005.135
## Ocekivane frekvencije za razred  Not retired - Hard :  46285.42
## Ocekivane frekvencije za razred  Retired - Carpet :  190.995
## Ocekivane frekvencije za razred  Retired - Clay :  857.5582
## Ocekivane frekvencije za razred  Retired - Grass :  244.8654
## Ocekivane frekvencije za razred  Retired - Hard :  1258.581
```

Očekivane frekvencije veće su od 5, što znači da možemo nastaviti sa χ^2 testom.

```
chisq.test(tbl,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 26.368, df = 3, p-value = 7.987e-06
```

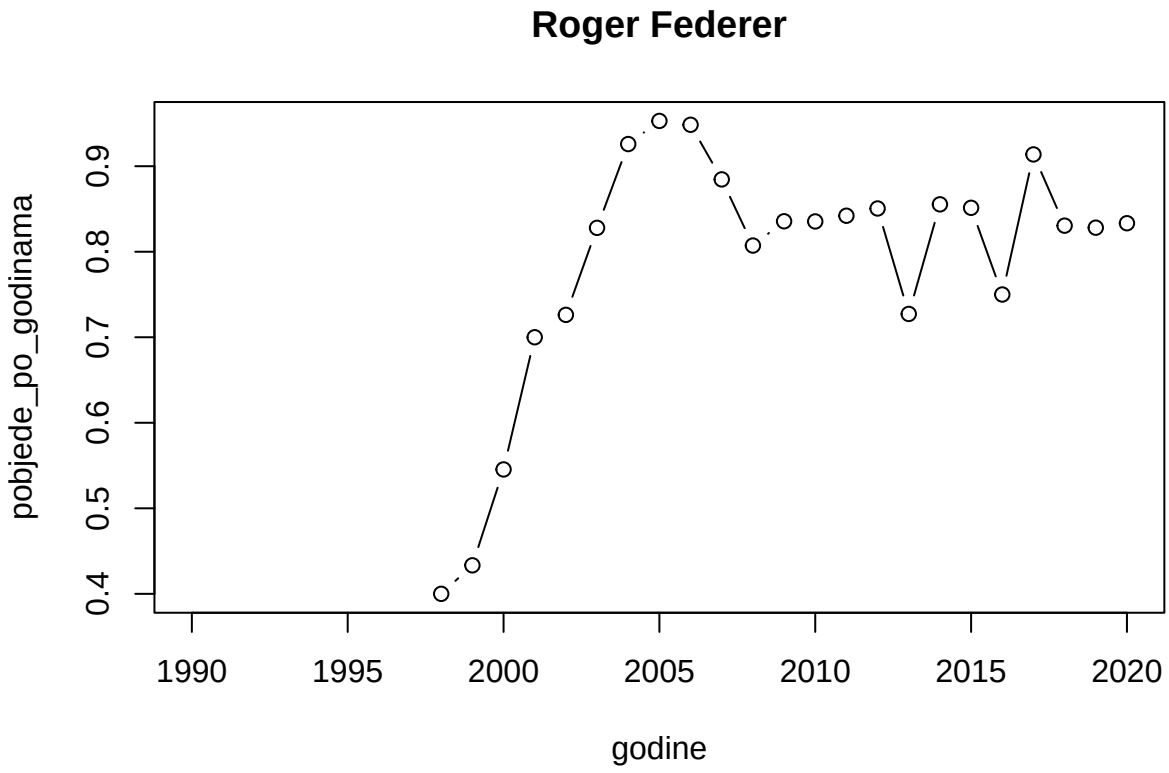
Odbacujemo H_0 u korist H_1 i zaključujemo da su podloga teniskog terena i odustajanje igrača od meča međusobno zavisni.

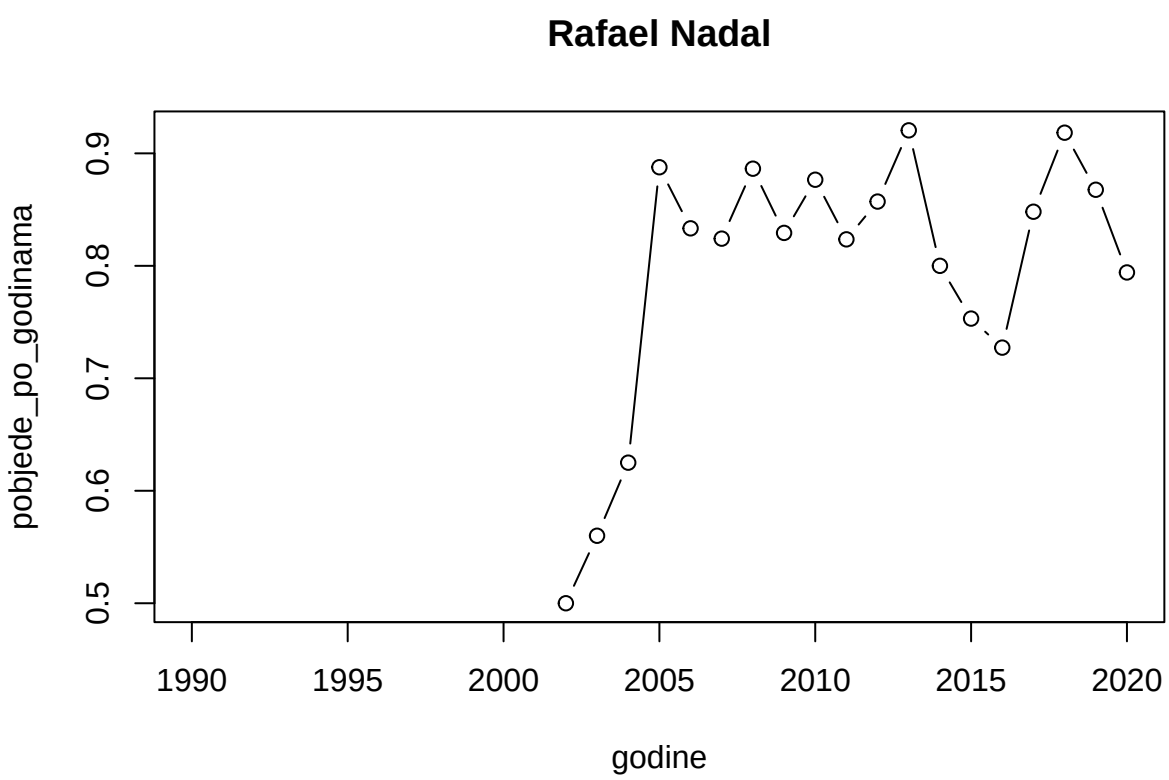
3.3 Dodatak

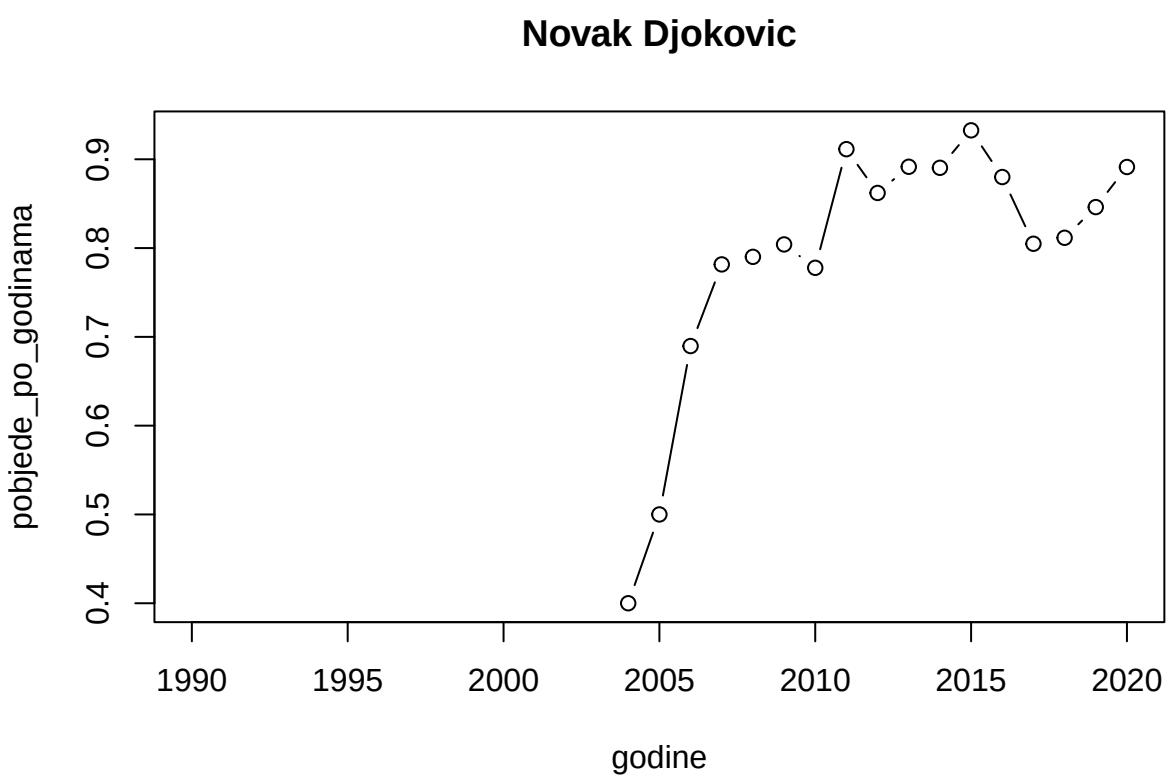
3.3.1 Promatranje napretka najboljih igrača

U raspravi o setu podataka i eventualnim dodatnim pitanjima koja bi se mogla nametnuti, došlo se na ideju praćenja napretka najboljih tenisača u zadnjih 30 godina. Ovdje su rezultati analize:

```
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$loser_id == player_id) & substr(tennis$
  }
  plot(godine, pobjede_po_godinama, type="b", main = (tennis[tennis$winner_id == player_id,][1,]$winner_
}
```




Rafael Nadal
pobjede_po_godinama
0.5
0.6
0.7
0.8
0.9
1990
1995
2000
2005
2010
2015
2020
godine



Možemo vidjeti da najbolji igrači imaju rast prvih 5 do 10 godina, a onda manje više stagniraju.

3.3.2 Marin Čilić

```
cilic = subset(tennis,(as.character(winner_name)=="Marin Cilic" | as.character(loser_name) == "Marin Ci
cilic_bof_5 = subset(cilic, as.character(best_of)=="5")
cilic_loss = subset(cilic_bof_5, as.character(loser_name) == "Marin Cilic")
cilic_superloss = subset(cilic_loss, substr(cilic_loss$score,1,1) < substr(cilic_loss$score,3,3))
cilic_superduperLoss = 4
```

4 Zaključak

Analiza podataka teniskih mečeva iznjedrila je neke zanimljive zaključke. S obzirom na veličinu skupa podataka i broj varijabli koje se u njemu pojavljuju, bio je velik izazov uopće i razumjeti što sve te varijable znače, a kamoli analizirati svaku sa svojim eventualnim ovisnostima. Ipak, neka od projektnih pitanja dobila su pripadne odgovore, od kojih su neki bili i iznenađujući.

U nastavku rada na projektu moglo bi se analizirati pojedine tenisače i njihove karijere. Primjerice, pojedini mečevi hrvatskog tenisača Marina Čilića mogli bi zadati pravu glavobolju ne samo onima koji ih gledaju, već i onima koji bi na temelju rezultata pokušali izvesti bilo kakve zaključke ili predviđanja. To je, pak, posao za nekog vrsnog poznavatelja tenisa, koji bi se tako mogao i dobro zabaviti.