SAP - Projekt - Analiza tesniskih mečeva

Bilić Ante, Paradžik Mario, Kaštelan Niko, Benjak Petar

Motivacija i opis problema

Statistika i predvidanje sportskih rezultata mogu pomoći menadžerima, trenerima, procjeniteljima kvota i drugima u donošenju odluka. U tenisu je statistika kao alat dobila dodatnu popularnost zahvaljujući bivšem treneru Craigu O'Shaughnessyu, 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.

Učitavanje potrebnih paketa

```
library(dplyr)
library(tidyverse)
```

Učitavanje podataka

```
tennisMatches = read.csv('tennis_atp_matches.csv')
```

Dimenzija podataka

```
dim(tennisMatches)
```

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

Nazivi varijabli

```
names(tennisMatches)
```

```
##  [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"
```

Prikaz podataka

```
View(tennisMatches)
```

Tipovi varijabli u skupu podataka

```
sapply(tennisMatches, class)
```

```
##                  X         tourney_id       tourney_name            surface
##          "integer"        "character"        "character"        "character"
##          draw_size      tourney_level       tourney_date          match_num
##          "integer"        "character"          "integer"          "integer"
##          winner_id        winner_seed       winner_entry        winner_name
##          "integer"          "numeric"        "character"        "character"
##        winner_hand          winner_ht         winner_ioc         winner_age
##        "character"          "numeric"        "character"          "numeric"
##           loser_id         loser_seed        loser_entry         loser_name
##          "integer"          "numeric"        "character"        "character"
##         loser_hand          loser_ht          loser_ioc          loser_age
##        "character"          "numeric"        "character"          "numeric"
##              score           best_of              round            minutes
##        "character"          "integer"        "character"          "numeric"
##              w_ace              w_df             w_svpt            w_1stIn
##          "numeric"          "numeric"          "numeric"          "numeric"
##           w_1stWon           w_2ndWon            w_SvGms          w_bpSaved
##          "numeric"          "numeric"          "numeric"          "numeric"
##          w_bpFaced              l_ace              l_df             l_svpt
##          "numeric"          "numeric"          "numeric"          "numeric"
##            l_1stIn           l_1stWon           l_2ndWon            l_SvGms
##          "numeric"          "numeric"          "numeric"          "numeric"
##          l_bpSaved          l_bpFaced        winner_rank winner_rank_points
##          "numeric"          "numeric"          "numeric"          "numeric"
##         loser_rank  loser_rank_points
##          "numeric"          "numeric"
```

Kod učitavanja podataka može doći do situacije gdje se tipovi podatak pogrešno prepoznaju

pa ih je potrebno ručno izmijeniti. U ovom se slučaju krivo prepoznaju tipovi varijabli: tourney_date, winner_id te loser_id.

```
tennisMatches = tennisMatches %>% mutate(tourney_date = as.Date(as.character(tourney_date), "%Y%m%d"),
                  winner_id = as.factor(winner_id),
                  loser_id = as.factor(loser_id))
```

```
summary(tennisMatches)
```

```
##        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.   :1990-12-31   Min.   :   1.00
##  1st Qu.: 32.00   Class :character   1st Qu.:1997-04-21   1st Qu.:   9.00
##  Median : 32.00   Mode  :character   Median :2004-06-14   Median :  22.00
##  Mean   : 52.75                      Mean   :2004-10-30   Mean   :  58.42
##  3rd Qu.: 64.00                      3rd Qu.:2012-02-20   3rd Qu.:  48.00
##  Max.   :128.00                      Max.   :2020-11-16   Max.   :1701.00
##
##    winner_id      winner_seed    winner_entry       winner_name
##  103819 : 1250   Min.   : 1.00   Length:96602       Length:96602
##  104745 : 1012   1st Qu.: 3.00   Class :character   Class :character
##  104925 :  946   Median : 5.00   Mode  :character   Mode  :character
##  103970 :  740   Mean   : 6.85
##  101736 :  692   3rd Qu.: 8.00
##  101948 :  687   Max.   :35.00
##  (Other):91275   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
##  103852 :  457   Min.   : 1.0    Length:96602       Length:96602
##  104269 :  424   1st Qu.: 4.0    Class :character   Class :character
##  104022 :  423   Median : 6.0    Mode  :character   Mode  :character
##  102148 :  422   Mean   : 8.2
##  104312 :  404   3rd Qu.:11.0
##  104259 :  388   Max.   :35.0
##  (Other):94084   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
```

Traženje nedostajećih vrijednosti

Dani skup podataka nerijetko sadrži nedostajuće podatke. Rad nad takvim podacima može dovesti do pogrešaka u testiranju hipoteza i zaključivanju. Varijable s velikim udjelom nedostajećih vrijednosti odbacujemo iz skupa podataka.

```
for (col_name in names(tennisMatches)){
  if (sum(is.na(tennisMatches[,col_name])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu ',col_name, ': ', sum(is.na(tennisMatches[,col_nam
```

```
  }
}
```

```
## Ukupno nedostajucih vrijednosti za varijablu  winner_seed :  57628
## Ukupno nedostajucih vrijednosti za varijablu  winner_ht :  4070
## Ukupno nedostajucih vrijednosti za varijablu  winner_age :  58
## Ukupno nedostajucih vrijednosti za varijablu  loser_seed :  75349
## Ukupno nedostajucih vrijednosti za varijablu  loser_ht :  7671
## Ukupno nedostajucih vrijednosti za varijablu  loser_age :  127
## Ukupno nedostajucih vrijednosti za varijablu  minutes :  12410
## Ukupno nedostajucih vrijednosti za varijablu  w_ace :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_df :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_svpt :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_1stIn :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_1stWon :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_2ndWon :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_SvGms :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_bpSaved :  9793
## Ukupno nedostajucih vrijednosti za varijablu  w_bpFaced :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_ace :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_df :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_svpt :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_1stIn :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_1stWon :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_2ndWon :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_SvGms :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_bpSaved :  9793
## Ukupno nedostajucih vrijednosti za varijablu  l_bpFaced :  9793
## Ukupno nedostajucih vrijednosti za varijablu  winner_rank :  1040
## Ukupno nedostajucih vrijednosti za varijablu  winner_rank_points :  2032
## Ukupno nedostajucih vrijednosti za varijablu  loser_rank :  2289
## Ukupno nedostajucih vrijednosti za varijablu  loser_rank_points :  3278
```

Varijabla winner_seed ima 59% nedostajećih vrijednosti, a varijabla loser_seed ima 78% nedostajećih vrijednosti što znači da nam ne daju puno informacija o teniskim mečevima pa se takve varijable uklanjanju iz skupa podataka.

```
tennisMatches = select(tennisMatches, -c("winner_seed", "loser_seed"))
dim(tennisMatches)
```

```
## [1] 96602    48
```

Problem 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?

Potrebno je izdvojiti visine tenisača u dva različita skupa podataka. Prvi skup podataka sadrži sve visine igrača koji su u trenutku odigravanje meča bili u top deset najboljih tenisača. Drugi skup podataka sadrži sve visine igrača koji u trenutku odigravanje meča nisu bili u top deset najboljih tenisača.

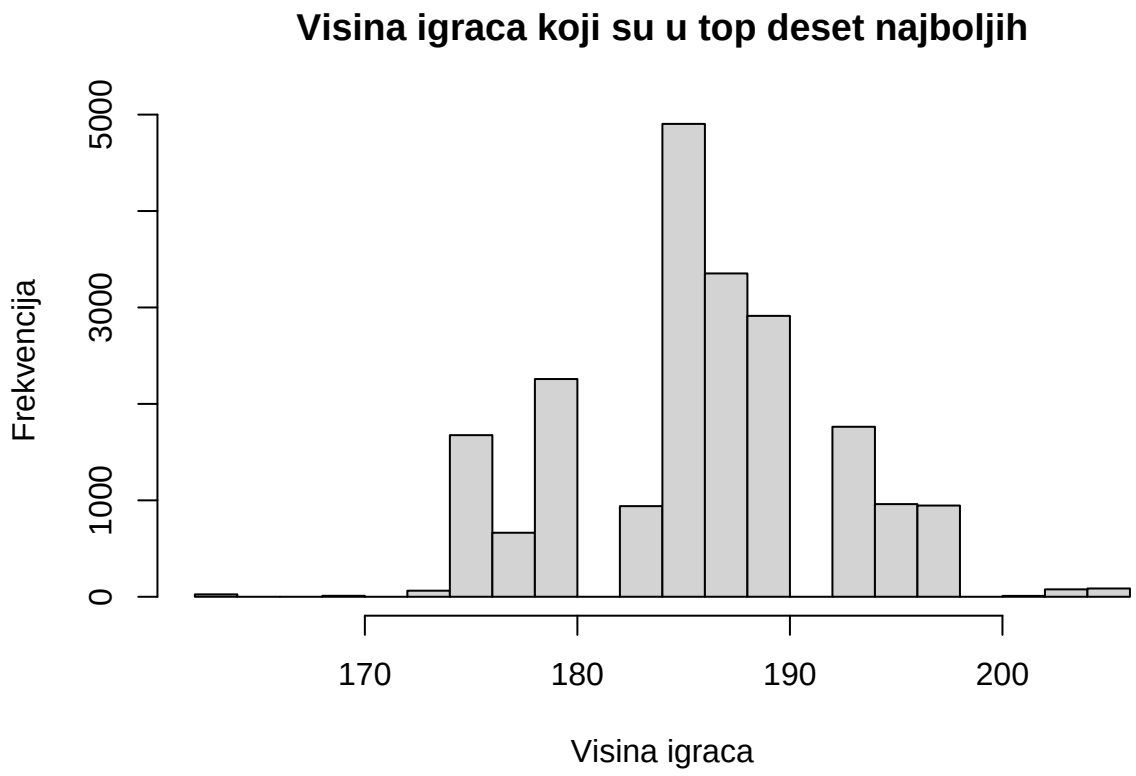
```
# Spremanje visina igrača koji su u trenutku igranja meča u top deset najboljih
topTenWinnerHeight = tennisMatches[tennisMatches$winner_rank <= 10, ]$winner_ht
topTenLosserHeight = tennisMatches[tennisMatches$loser_rank <= 10, ]$loser_ht

topTenHeight = append(topTenWinnerHeight, topTenLosserHeight)
```

```r
# Odbacivanje nedostajećih vrijednosti
topTenHeight = topTenHeight[complete.cases(topTenHeight)]
summary(topTenHeight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   163.0   183.0   185.0   186.4   190.0   206.0
```

```r
hist(topTenHeight,  xlab="Visina igrača", ylab="Frekvencija", main="Visina igrača koji su u top deset na
```



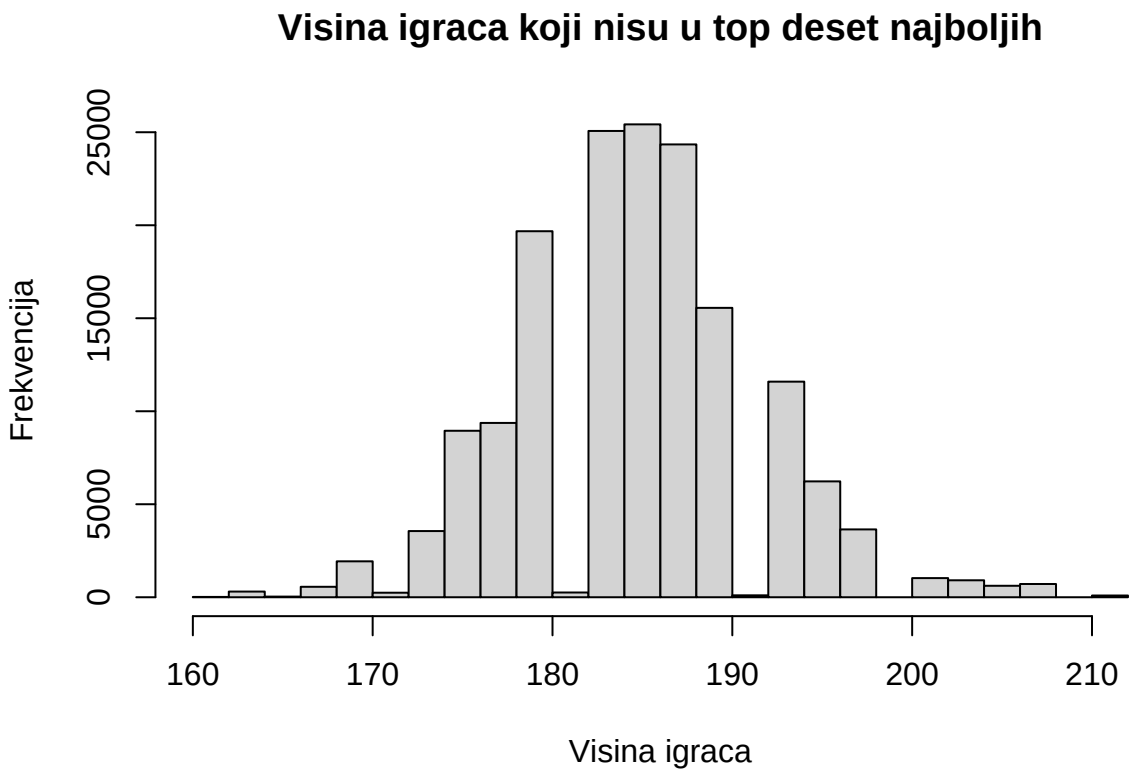

```r
# Spremanje visina igrača koji u trenutku igranja meča nisu u top deset najboljih
notTopTenWinnerHeight = tennisMatches[tennisMatches$winner_rank > 10, ]$winner_ht
noTopTenLosserHeight = tennisMatches[tennisMatches$loser_rank > 10, ]$loser_ht

notTopTenHeight = append(notTopTenWinnerHeight, noTopTenLosserHeight)

# Odbacivanje nedostajećih vrijednosti
notTopTenHeight = notTopTenHeight[complete.cases(notTopTenHeight)]
summary(notTopTenHeight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   160.0   180.0   185.0   185.2   190.0   211.0
```

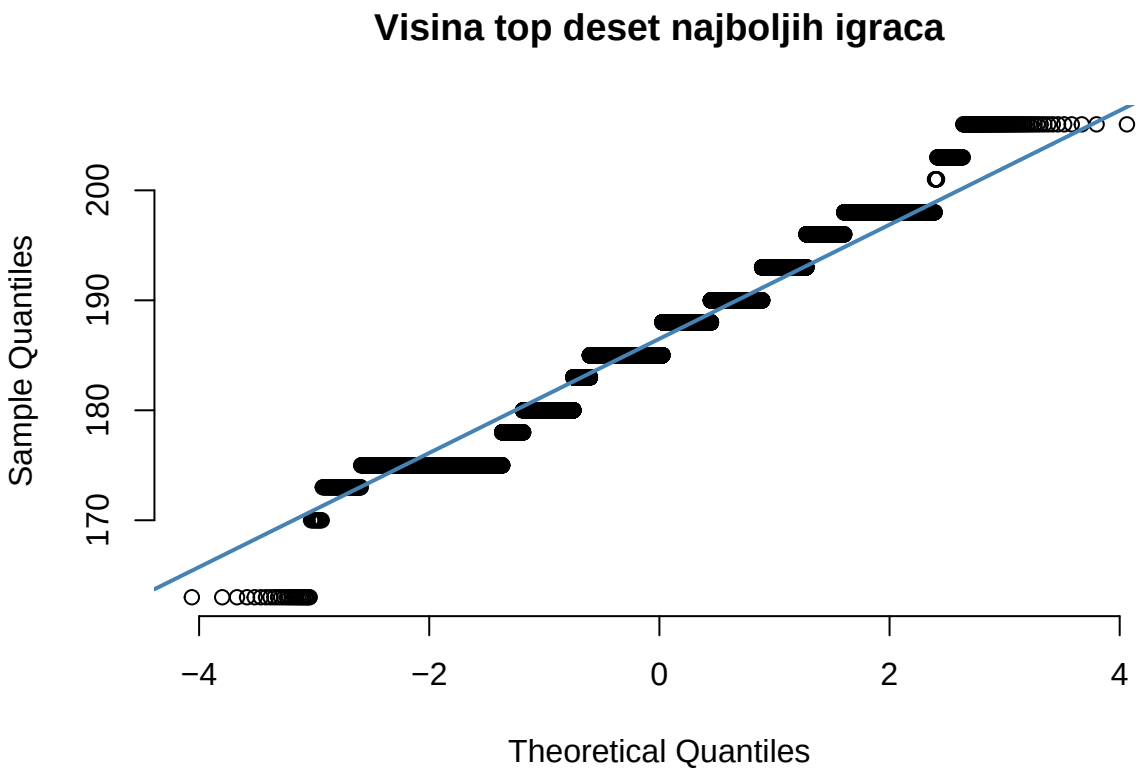
```r
hist(notTopTenHeight, xlab="Visina igrača", ylab="Frekvencija", main="Visina igrača koji nisu u top dese
```



Prije određivanja testa potrebno je provjeriti normalnost podataka.

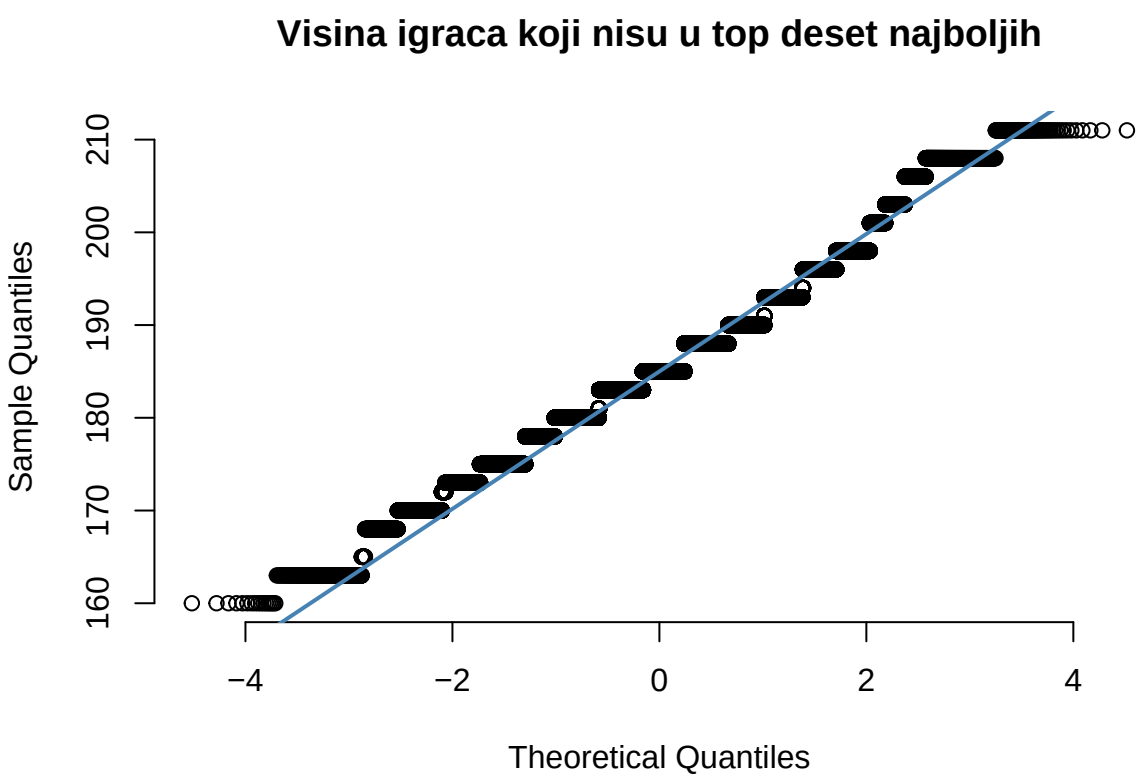
Histogram te QQ-dijagram upućuju na normalnost podataka visine top deset najboljih igrača.

```
# QQ-dijagram
qqnorm(topTenHeight, pch = 1, frame=FALSE, main="Visina top deset najboljih igrača")
qqline(topTenHeight, col="steelblue", lwd=2)
```



Histogram te QQ-dijagram upućuju na normalnost podataka visine igrača koji nisu u top deset najboljih.

```
# QQ-dijagram
qqnorm(notTopTenHeight, pch = 1, frame=FALSE, main="Visina igrača koji nisu u top deset najboljih")
qqline(notTopTenHeight, col="steelblue", lwd=2)
```



Pod pretpostavkom da visine igrača prate normalnu distribuciju nastavlja testiranje povezanosti visine i uspješnosti igrača, koristi se t-test.

Testira se postoji li značajna razlika u očekivanju visine igrača koji su u top deset najboljih te onih koji nisu.

Provedba t-testa

Provjera jednakosti varijanci

```
varTopTen = var(topTenHeight)
varNotTopTen = var(notTopTenHeight)
cat('Varijanca top deset najboljih igrača: ', varTopTen, '\n')
```

```
## Varijanca top deset najboljih igraca:  38.77018
```

```
cat('Varijanca igrača izvan top deset najboljih: ', varNotTopTen)
```

```
## Varijanca igraca izvan top deset najboljih:  45.34102
```

Test o jednakosti varijanci

Ako imamo dva nezavisna slučajna uzorka $X_1^1, X_1^2, \dots X_1^{n_1}$ i $X_2^1, X_2^2, \dots, X_2^{n_2}$ koji dolaze iz normalnih distribucija s 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 distribuciju s $(n_1 - 1, n_2 - 1)$ stupnjeva slobode, pri čemu vrijedi:

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

Hipoteze testa jednakosti varijanci glase:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

```
var.test(topTenHeight, notTopTenHeight)
```

```
##
##  F test to compare two variances
##
## data:  topTenHeight and notTopTenHeight
## F = 0.85508, num df = 20647, denom df = 160241, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8378014 0.8728576
## sample estimates:
## ratio of variances
##          0.8550796
```

Zaključak

Odbacujemo hipotezu H_0 o jednakosti varijanci pri razini značajnosti od 0.05 s obzirom da je p-vrijednost 2.2e-16.

Provedba t-testa uz pretpostavku o nejednakosti varijanci.

```
t.test(topTenHeight, notTopTenHeight, alt = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  topTenHeight and notTopTenHeight
## t = 26.343, df = 27259, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.148039      Inf
## sample estimates:
## mean of x mean of y
##  186.3996  185.1751
```

Zaključak

Na razini značajnosti od 0.05 odbacujemo hipotezu H_0 premda je p-vrijednost 2.2e-16, postoji značajna razlika visina igrača u top deset najboljih u posljednjih 30 godina i onih izvan top deset najboljih. Visina igrača u top deset najboljih je veća od visine igrača koji nisu u top deset najboljih.

Iz provedenog testa ne možemo zaključiti da visina izravno utječe na uspješnost igrača, već samo korelaciju između visine i uspješnosti igrača.

Problem 2

Jesu li ljevaci nezgodniji protivnici dešnjacima koji igraju jednoručni backhand?

Problem 3

Predvida li pobjeda prvog seta pobjedu cijelog meča?

Problem 4

Možemo li temeljem danih varijabli predvidjeti pobjednika teniskog meča?

##Problem 5 Utječe li broj aseva na ishod pobjednika?