Analiza rezultata teniskih mečeva

Tipseri: Alan Đurđević, Matej Magat, Tin Šaban, Ivan Zeba

Zagreb, siječnja 2024.

Contents

Uvod

Za pravilno donošenje odluka o treningu tenisača i organizaciji teniskih mečeva, ključna je detaljna obrada postojećih podataka o prijašnjim mečevima.

U ovom projektu obrađujemo prikupljene podatke o preko 100,000 odigranih teniskih mečeva ATP turnira za svaku sezonu od 1991. do 2023. godine. Neke od izmjerenih varijabla su: vrsta podloge, broj asova, broj dvostrukih pogrešaka, postotak uspješnosti prvog servisa, ukupan broj osvojenih poena, itd.

Projekt navodi mnoge analize deskriptivne statistike:

- Udio različitih podloga u ATP mečevima
- Starost pobjednika i svih sudionika ATP-a
- Visina pobjednika i svih sudionika ATP-a
- Udio različitih nacionalnosti igrača koji su sudjelovali na ATP-u
- Udio dominantne ruke među svim sudionicima ATP-a

Također obrađujemo sljedeća istraživačka pitanja:

- Kakva je distribucija mečeva na specifičnim podlogama u različitim godišnjim dobima?
- Postoji li značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu?
- Ima li razlike u broju serviranih asova na različitim podlogama?
- Kakva je veza između vrste terena i vjerojatnosti da će mečevi otići u peti set?
- Možemo li procijeniti broj asova koje će igrač odservirati u tekućoj (zadnjoj dostupnoj sezoni) na temelju njegovih rezultata iz prethodnih sezona?

1 Deskriptivna statistika

1.0.1 Učitavanje potrebnih paketa

```
packages <- c("dplyr", "readr", "lubridate", "stringr", "nortest", "fastDummies")
lapply(packages, library, character.only=TRUE)
```

1.0.2 Spajanje podataka u jedinstvenu tablicu

```
matches <- list.files(path=paste(getwd(), "/ATP-Matches", sep = ""), pattern =
↪  "*.csv" ,full.names = TRUE) %>% lapply(read_csv) %>% bind_rows
```

1.1 Svojstva skupa podataka

1.1.1 Mjereni podatci o teniskim mečevima

```
names(matches)
```

```
 [1] "tourney_id"         "tourney_name"       "surface"
 [4] "draw_size"          "tourney_level"      "tourney_date"
 [7] "match_num"          "winner_id"          "winner_seed"
[10] "winner_entry"       "winner_name"        "winner_hand"
[13] "winner_ht"          "winner_ioc"         "winner_age"
[16] "loser_id"           "loser_seed"         "loser_entry"
[19] "loser_name"         "loser_hand"         "loser_ht"
[22] "loser_ioc"          "loser_age"          "score"
[25] "best_of"            "round"              "minutes"
[28] "w_ace"              "w_df"               "w_svpt"
[31] "w_1stIn"            "w_1stWon"           "w_2ndWon"
[34] "w_SvGms"            "w_bpSaved"          "w_bpFaced"
[37] "l_ace"              "l_df"               "l_svpt"
[40] "l_1stIn"            "l_1stWon"           "l_2ndWon"
[43] "l_SvGms"            "l_bpSaved"          "l_bpFaced"
```

```
[46] "winner_rank"        "winner_rank_points" "loser_rank"
[49] "loser_rank_points"
```

1.1.2 Dimenzije podataka

```
cat("Dimenzije podataka: ", dim(matches))
```

```
Dimenzije podataka:  104682 49
```

1.1.3 Postotci nedostajućih vrijednosti

```
Postotak nedostajućih vrijednosti za varijablu:
```

```
winner_seed :    59.5 %   winner_entry :     87.76 %   winner_hand :     0.01 %
winner_ht :  2.34 %   winner_age :   0 %   loser_seed :  77.74 %
loser_entry :    79.86 %   loser_hand :  0.04 %   loser_ht :     4.64 %
loser_age :  0.02 %   minutes :  12.45 %   w_ace :   9.75 %
w_df :   9.75 %   w_svpt :   9.75 %   w_1stIn :  9.75 %
w_1stWon :   9.75 %   w_2ndWon :     9.75 %   w_SvGms :  9.75 %
w_bpSaved :  9.75 %   w_bpFaced :    9.75 %   l_ace :    9.75 %
l_df :   9.75 %   l_svpt :   9.75 %   l_1stIn :  9.75 %
l_1stWon :   9.75 %   l_2ndWon :     9.75 %   l_SvGms :  9.75 %
l_bpSaved :  9.75 %   l_bpFaced :    9.75 %   winner_rank :  1.14 %
winner_rank_points :     2.08 %   loser_rank :   2.42 %   loser_rank_points :    3.36 %
```

1.1.4 Čišćenje podataka

Podatke starije od 1991. nismo uključili u testni skup podataka.

Obrisali smo meč 11. 1. 2016. između Gillesa Mullera i Jeremyja Chardyja koji je prema podatku trajao 1146 minuta i meč 1. 5. 2017. između Hyeona Chunga i Martina Kližana koji je trajao 987 minuta. Mečeve smo izbacili jer je najduži teniski meč službeno zabilježen trajao 665 minuta između Johna Isnera i Nicolasa Mahuta na Wimbledonu 2010.

https://olympics.com/en/news/longest-tennis-match-history-grand-slam-record

Na internetu postoji podatak da je meč između Hyeona Chunga i Martina Kližana uistinu trajao 987 minuta, ipak smo odlučili podatak izostaviti iz skupa podataka.

https://www.tennisabstract.com/cgi-bin/player-classic.cgi?p=MartinKlizan&f=ACareerqqDMunichqq

Zbog velikog postotka nedostajućih vrijednosti za stupce winner_seed, winner_entry, loser_seed i loser_entry, odlučili smo izbaciti te stupce.

```
matches <- matches[matches$minutes <= 665 | is.na(matches$minutes), ]
matches <- select(matches, -winner_seed, -winner_entry, -loser_seed, -loser_entry)
```

1.1.5 Dimenzije očišćenih podataka

```
cat("Dimenzije očišćenih podataka: ", dim(matches))
```

```
Dimenzije očišćenih podataka:  104680 45
```

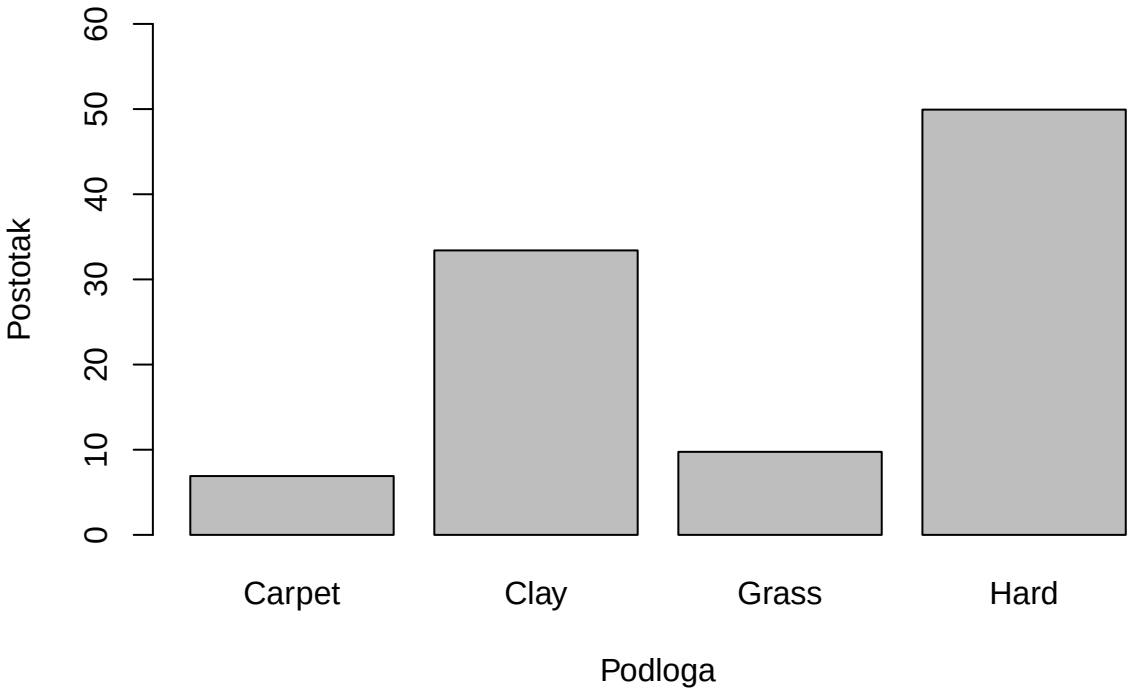
1.1.6 Prikaz podataka

```
View(matches)
```

1.2 Histogram podloga

Histogram podloga nam pokazuje postotak teniskih mečeva odigranih na pojedinoj podlozi. Histogram nam pokazuje da se najviše teniskih mečeva odigrava na tvrdoj podlozi (oko 50 %) dok je na 2. mjestu zemljana podloga (oko 35 %), a trava i tepih imaju podjednak udio (oko 10 %).

```
barplot(table(matches$surface) / nrow(matches) * 100, ylab = "Postotak", xlab =
↪  "Podloga", ylim = range(pretty(c(0, table(matches$surface) / nrow(matches) * 100
↪  * 1.1))))
```



1.3 Usporedba starosti tenisača

```
winner_cols <- matches[, c("winner_id", "winner_age")]
loser_cols <- matches[, c("loser_id", "loser_age")]
colnames(loser_cols) <- colnames(winner_cols)
combined_matches <- rbind(winner_cols, loser_cols)

winner_tour <- matches[matches$round == 'F', ]
cat('Prosječna godina svih tenisača: ', mean(combined_matches$winner_age, na.rm =
↪  TRUE), "\n")
```

```
Prosječna godina svih tenisača:  25.82735
```

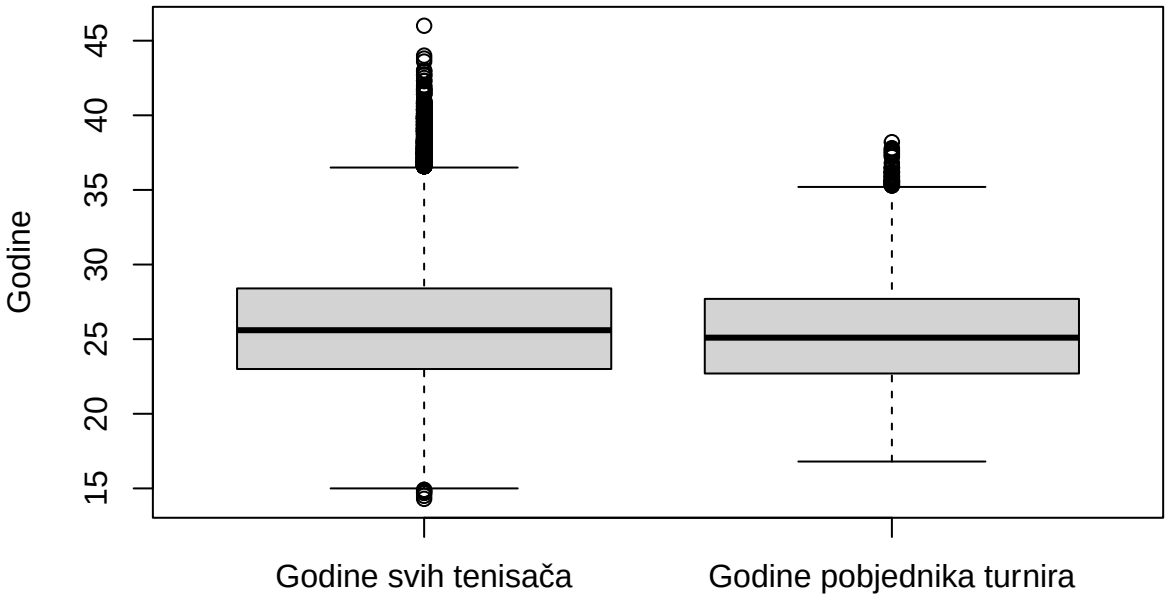
```
cat('Prosječna godina pobjednika turnira: ', mean(winner_tour$winner_age, na.rm =
↪  TRUE), "\n")
```

```
Prosječna godina pobjednika turnira:  25.47182
```

1.3.1 Pravokutni dijagram starosti

Pravokutni dijagram starosti tenisača nam pokazuje da je medijalna vrijednost oko 25 godina te da su također pobjednici turnira u toj vrijednosti. Pedeset posto igrača se nalazi između 20 i 30 godina dok najstariji igrač ima preko 45 godina, a najmlađi ispod 15.

```
boxplot(combined_matches$winner_age, winner_tour$winner_age, names = c("Godine svih
↪  tenisača", "Godine pobjednika turnira"), ylab="Godine")
```



1.4 Usporedba visine tenisača

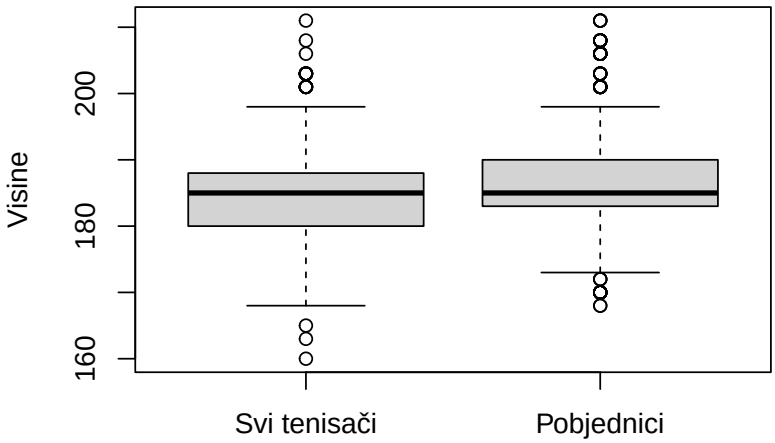
```
winner_cols <- matches[, c("winner_id", "winner_ht")]
loser_cols <- matches[, c("loser_id", "loser_ht")]
colnames(loser_cols) <- colnames(winner_cols)
combined_matches <- rbind(winner_cols, loser_cols)
unique_matches <- combined_matches %>% distinct(winner_id, .keep_all = TRUE)
winner_tour <- matches[matches$round == 'F', ]
cat('Prosječna visina svih tenisača: ', mean(unique_matches$winner_ht, na.rm = TRUE),
↪  "\n")
```

```
Prosječna visina svih tenisača:  184.3402
```

```
cat('Prosječna visina pobjednika turnira: ', mean(winner_tour$winner_ht, na.rm =
↪  TRUE), "\n")
```

```
Prosječna visina pobjednika turnira:  186.3783
```

1.4.1 Pravokutni dijagram visina tenisača



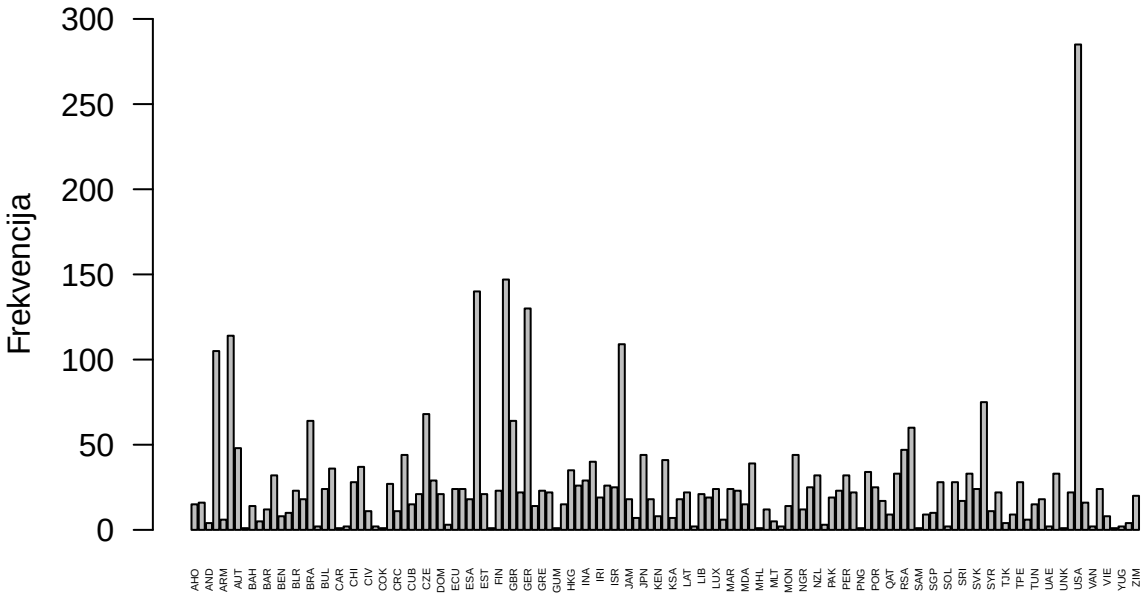
Pravokutni dijagram visina tenisača nam pokazuje da su tenisači većinom veći od 180 cm, te također nam pokazuje da je prosječna visina pobjednika turnira viša nego li je to prosječna visina svih tenisača.

1.5 Nacionalnosti tenisača

1.5.1 Histogram država

Histogram država prikazuje koliko su pojedine države imale različitih tenisača. Prvo smo izdvojili stupce koji su nam potrebni (id i ioc) za pobjednike i gubitnike, zatim smo spojili podatke (rbind) te smo očistili podatke od višestrukog pojavljivanja istog id-a. Histogram je malo nepregledan zbog velikog broja država, no možemo iščitati da najveći broj tenisača dolazi iz Sjedinjenih Američkih Država, zatim slijedi Australija. Još se izdvajaju Španjolska, Francuska, Velika Britanija, Njemačka i Italija po broju tenisača, dok su ostale države manje uočljive.

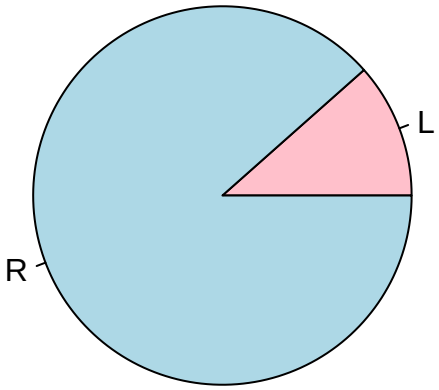
```
winner_cols <- matches[, c("winner_id", "winner_ioc")]
loser_cols <- matches[, c("loser_id", "loser_ioc")]
colnames(loser_cols) <- colnames(winner_cols)
combined_matches <- rbind(winner_cols, loser_cols)
unique_matches <- combined_matches %>% distinct(winner_id, .keep_all = TRUE)
barplot(table(unique_matches$winner_ioc), las=2,cex.names=.3, ylab = "Frekvencija",
↪  ylim = range(pretty(c(0, table(unique_matches$winner_ioc)))), xlim = c(0,155))
```



1.6 Graf udjela dominantne ruke

Za prikaz udjela dominantne ruke koristili smo kružni graf. Iz grafa možemo vidjeti znatno veći udio dešnjaka od ljevaka. Oko 10 % ljudske populacije je ljevoruko pa nas ovaj podatak ne iznenađuje.

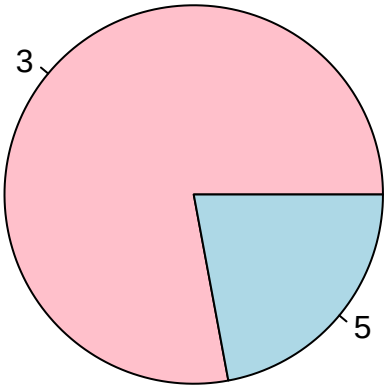
```
winner_cols <- matches[, c("winner_id", "winner_hand")]
loser_cols <- matches[, c("loser_id", "loser_hand")]

colnames(loser_cols) <- colnames(winner_cols)

combined_matches <- rbind(winner_cols, loser_cols)

combined_matches <- filter(combined_matches, winner_hand %in% c("L", "R"))

unique_matches <- combined_matches %>% distinct(winner_id, .keep_all = TRUE)
# table(unique_matches$winner_hand)
pie(table(unique_matches$winner_hand), col = c("pink", "lightblue"))
```



1.7 Graf udjela mečeva po broju “best of” setova

Strukturnim krugom prikazano je koliki udio mečeva se igra “best of 3” (na dva dobivena seta), a koliki “best of 5” (na tri dobivena seta). Vidljivo je da puno više mečeva (preko 80%) igra “best of 3”.

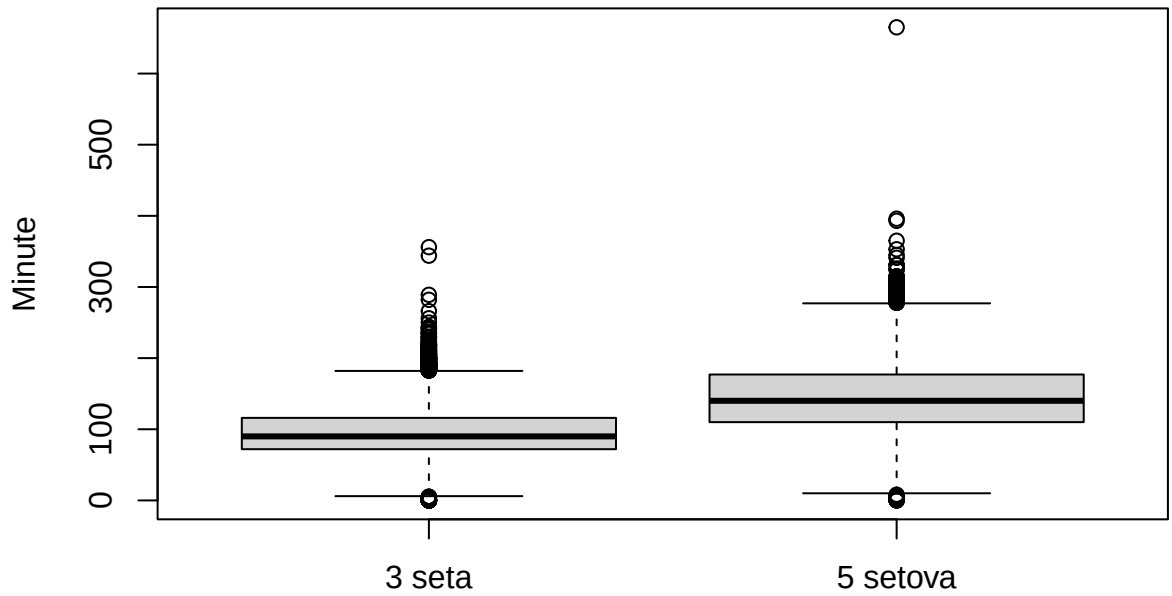
```
winner_tour <- matches[matches$round == 'F', ]
set_vrijeme <- matches[, c("best_of", "minutes")]
# table(set_vrijeme$best_of)
pie(table(set_vrijeme$best_of), col = c("pink", "lightblue"))
```



1.8 Usporedba trajanja mečeva

S pomoću pravokutnog dijagrama vizualizirali smo i usporedili trajanje mečeva s 3 odnosno 5 setova. Očekivano, medijan mečeva s 3 seta je manji od medijana mečeva s 5 setova. Također se vidi da je treći kvartil mečeva s 3 seta veći od prvog kvartila mečeva s 5 setova, kao i što je gornji izdanak mečeva s 3 seta veći od trećeg kvartila mečeva s 5 setova.

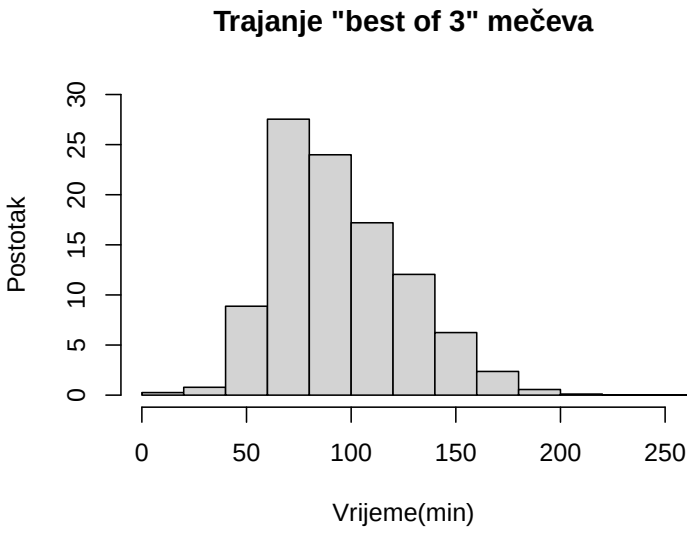
```
s3=set_vrijeme[set_vrijeme$best_of == 3 & (!is.na(set_vrijeme$minutes)),]
s5=set_vrijeme[set_vrijeme$best_of == 5 & (!is.na(set_vrijeme$minutes)),]

boxplot(s3$minutes, s5$minutes, names = c ("3 seta", "5 setova"), ylab="Minute")
```

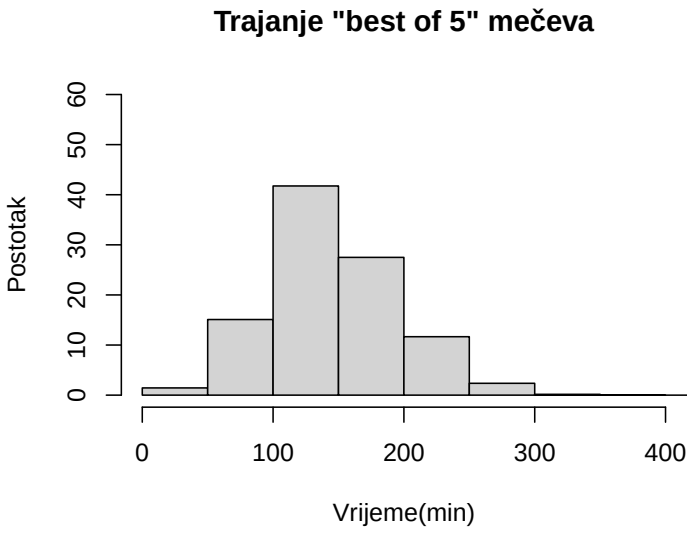


```
filter(set_vrijeme, !is.na(minutes)) %>% mutate(., katTrajanje = ntile(minutes, 20))
 ↪  -> set_vrijeme

d1 <- table(filter(set_vrijeme, best_of == 5)$katTrajanje)/nrow(filter(set_vrijeme,
 ↪  best_of == 5))
d2 <- table(filter(set_vrijeme, best_of == 3)$katTrajanje)/nrow(filter(set_vrijeme,
 ↪  best_of == 3))
data <- t(cbind(d1, d2))
```

```
h = hist(s3$minutes,plot=FALSE)
h$density = h$counts/sum(h$counts)*100
```

```
h2 = hist(s5$minutes,plot=FALSE)
h2$density = h2$counts/sum(h2$counts)*100
```





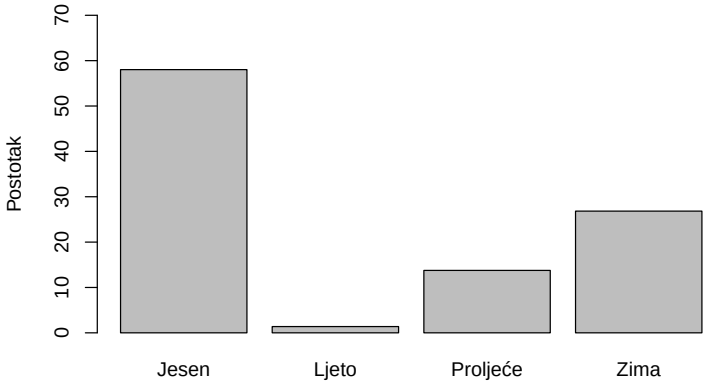
2 Istraživačka pitanja

2.1 Distribucija mečeva ovisno o podlozi i godišnjem dobu

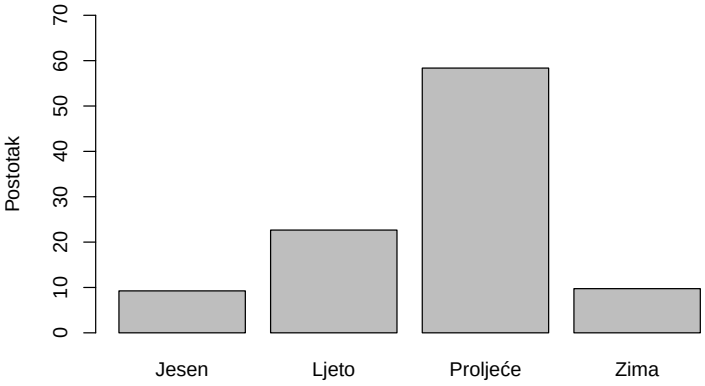
Prvo istraživačko pitanje:
Kakva je distribucija mečeva na specifičnim podlogama u različitim godišnjim dobima?

```
podloga_datum <- matches[, c("surface", "tourney_date")]
podloga_datum$tourney_date <- ymd(podloga_datum$tourney_date)
# funkcija koja mapira mjesec iz datuma na godišnje doba
get_season <- function(date) {
  month <- month(date)
  if (month %in% c(3, 4, 5)) {
    return("Proljeće")
  } else if (month %in% c(6, 7, 8)) {
    return("Ljeto")
  } else if (month %in% c(9, 10, 11)) {
    return("Jesen")
  } else {
    return("Zima")
  }
}

podloga_datum$god_doba <- sapply(podloga_datum$tourney_date, get_season)

tepih <- podloga_datum[podloga_datum$surface == "Carpet",]
zemlja <- podloga_datum[podloga_datum$surface == "Clay",]
trava <- podloga_datum[podloga_datum$surface == "Grass",]
tvrda <- podloga_datum[podloga_datum$surface == "Hard",]
```

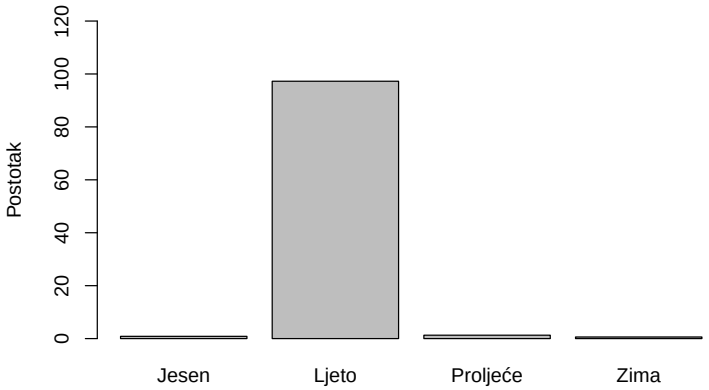
Stvaranje barplotova iz vektora stvorenih gore (kod izbačen zbog sažetosti)



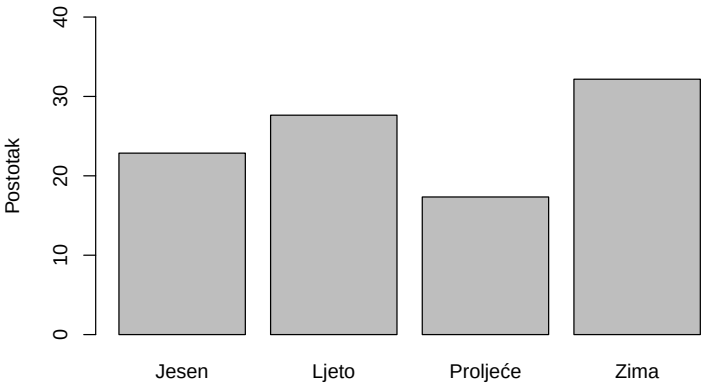
(a) Tepih



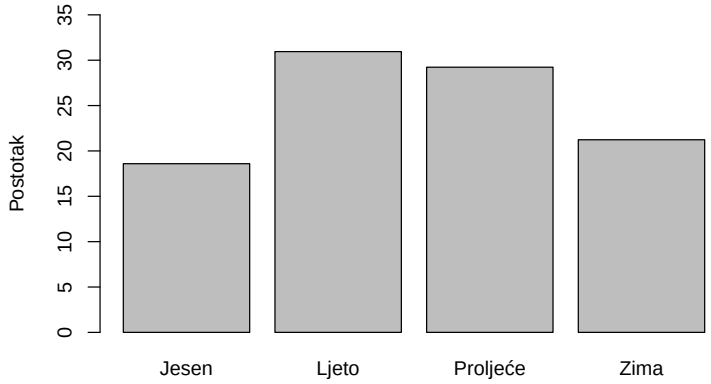
(b) Zemlja



(c) Trava



(d) Tvrda podloga



(e) Sve podloge

Graf pod a) prikazuje da se na tepihu igra najviše u jesen, a nakon toga u zimi. To ima smisla s obzirom na to da je teren s tepihom često u zatvorenom prostoru.

Tereni na otvorenom najčešće koriste zemlju ili travu te nam to i grafovi pod b) i c) potvrđuju.

Graf pod b) prikazuje da se zemljani teren najčešće koristi u proljeće i ljeto, dok graf pod c) prikazuje da se travnati teren koristi gotovo isključivo ljeti.

Graf pod d) prikazuje da je tvrda podloga skoro podjednako distribuirana, s najvećim korištenjem zimi.

2.2 Dvostruke pogreške na zatvorenom i otvorenom terenu

Drugo istraživačko pitanje:
Postoji li značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu?

```
surface_df <- matches[, c("surface", "w_df", "l_df")]
# funkcija koja mapira površinu na vanjski i unutarnji teren
get_outdoor <- function(surface) {
  if (surface %in% c("Grass", "Clay")) {
    return("T")
  } else if (surface %in% c("Carpet", "Hard")) {
    return("F")
  } else {
    return("")
  }
}

surface_df$is_outdoor <- sapply(surface_df$surface, get_outdoor)
surface_df <- transmute(surface_df, df = w_df+l_df, is_outdoor)

outdoor_df <- filter(surface_df, is_outdoor == "T")
indoor_df <- filter(surface_df, is_outdoor == "F")
```

Prvo Lillieforsovim testom provjeravamo dolaze li uzorci iz normalnih distribucija.

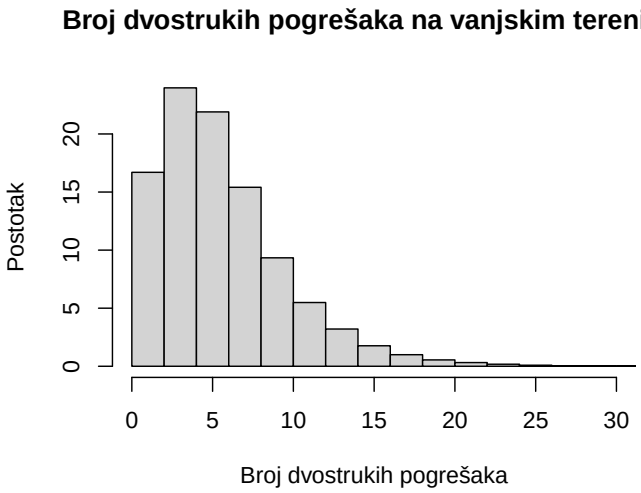
```
lillie.test(outdoor_df$df)
```

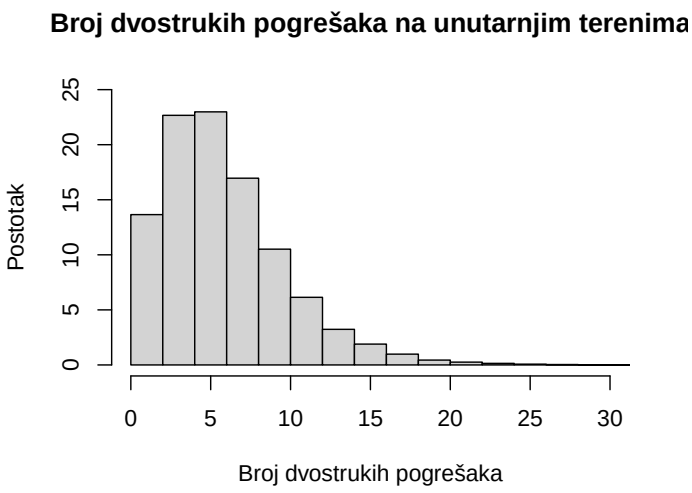
```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  outdoor_df$df
D = 0.13351, p-value < 2.2e-16
```

```
lillie.test(indoor_df$df)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  indoor_df$df
D = 0.12633, p-value < 2.2e-16
```

Razina značajnosti je 0.05 te zbog p-vrijednosti manje od 0.05 odbacujemo pretpostavku normalnosti uzoraka. Histogrami uzoraka nam dodatno osnažuju ovu tvrdnju.





2.2.1 Mann-Whitney-Wilcoxonov test

Pošto ne možemo pretpostaviti normalnost populacije, morat ćemo provesti Mann-Whitney-Wilcoxonov test - neparametarska alternativa t-testu za nezavisne uzorke.

H_0 : Razlika u sredinama dvostrukih pogrešaka na vanjskom i unutarnjem terenu jednaka je 0

H_1 : Razlika u sredinama dvostrukih pogrešaka na vanjskom i unutarnjem terenu nije jednaka 0

$$u_1 = w_1 - \frac{n_1(n_1+1)}{2}$$

$$u_2 = w_2 - \frac{n_2(n_2+1)}{2}$$

$$u = min(u_1, u_2)$$

```
wilcox.test(outdoor_df$df, indoor_df$df, alt = "two.sided")
```

```
    Wilcoxon rank sum test with continuity correction

data:  outdoor_df$df and indoor_df$df
W = 1039343474, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Razina značajnosti je 0.05 te zbog p-vrijednosti manje od 0.05 odbacujemo H_0 i zaključujemo da postoji značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu.

2.2.1 T-test

Provest ćemo i t-test iako uzorci ne proizlaze iz normalne populacije kako bismo vidjeli poklapa li se s Wilcoxon–Mann–Whitney testom.

Da bismo mogli uopće mogli provesti t-test, moramo vidjeti jesu li varijance dvije populacije jednake ili različite.

Za slučaj iz zadatka pretpostavli bismo da varijance nisu jednake, s obzirom na to da su na vanjskom terenu prisutni dodatni čimbenici. Sunce, vjetar, temperatura i vlažnost su samo od nekih njih.

Provodimo f-test za provjeru jednakosti varijanci dvije populacije.

H_0 : Omjer varijanci dvostrukih pogrešaka na vanjskom i unutarnjem terenu jednak je 1

H_1 : Omjer varijanci dvostrukih pogrešaka na vanjskom i unutarnjem terenu nije jednak 1

$$F = \frac{S_{X_1}^2}{S_{X_2}^2} \sim f(n-1, m-1)$$

```
var.test(outdoor_df$df, indoor_df$df, ratio = 1, alternative = "two.sided",
↪  conf.level = 0.95, na.action("na.exclude"))
```

```
    F test to compare two variances

data:  outdoor_df$df and indoor_df$df
F = 1.077, num df = 40866, denom df = 53605, p-value = 1.332e-15
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.057556 1.096766
sample estimates:
ratio of variances
           1.07697
```

Zbog p-vrijednosti manje od 0.05 možemo odbaciti tvrdnju da su varijance dvostrukih pogrešaka na vanjskom i zatvorenom terenu jednake.

Sada možemo provesti t-test uz pretpostavku nejednakih varijanca.

H_0 : Razlika u sredinama dvostrukih pogrešaka na vanjskom i unutarnjem terenu jednaka je 0

H_1 : Razlika u sredinama dvostrukih pogrešaka na vanjskom i unutarnjem terenu nije jednaka 0

$$T = \frac{\bar{x_1} - \bar{x_2} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t^2(\lfloor v \rfloor)$$

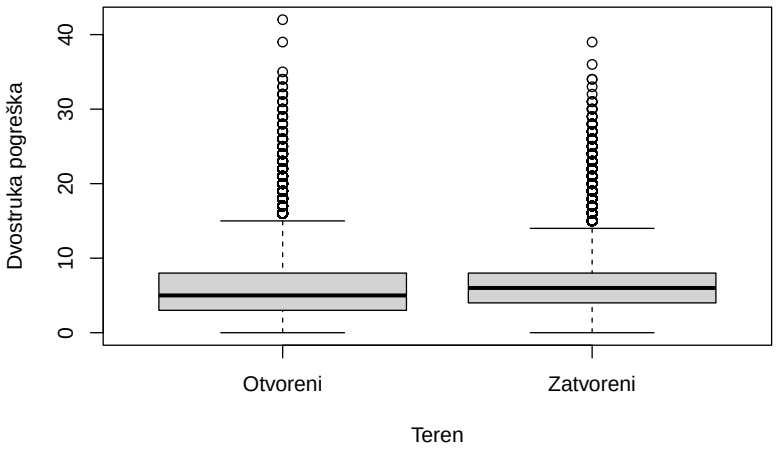
$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

```
t.test(outdoor_df$df, indoor_df$df, alt = "two.sided", var.equal = FALSE)
```

```
    Welch Two Sample t-test
```

```
data:  outdoor_df$df and indoor_df$df
t = -9.4295, df = 86251, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2965178 -0.1944635
sample estimates:
mean of x mean of y
 6.078988  6.324479
```

Zbog p-vrijednosti manje od 0.05, odbacujemo nultu hipotezu i prihvaćamo alternativu. Zaključujemo da postoji značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu.



Pravokutni dijagrami nam pokazuju da na zatvorenom terenu ima u prosjeku više dvostrukih pogrešaka, ali vanjski tereni pokazuju veću varijancu

2.3 Servirani asovi ovisno o podlozi

Treće istraživačko pitanje:
Ima li razlike u broju serviranih asova na različitim podlogama?

Da bismo provjerili ovo tvrdnju, provest ćemo ANOVA-u, metodu kojom testiramo sredine više populacija.

Pretpostavke ANOVA-e su:

- nezavisnost pojedinih podataka u uzorcima,
- normalna razdioba podataka,
- homogenost varijanci među populacijama.

```
tepihAsovi <- matches[matches$surface == "Carpet" ,]
zemljaAsovi <- matches[matches$surface == "Clay",]
travaAsovi <- matches[matches$surface == "Grass",]
tvrdaAsovi <- matches[matches$surface == "Hard",]

tepihAsovi <- na.omit(transmute(tepihAsovi, ukupnoAsovi = w_ace + l_ace))
zemljaAsovi <- na.omit(transmute(zemljaAsovi, ukupnoAsovi = w_ace + l_ace))
travaAsovi <- na.omit(transmute(travaAsovi, ukupnoAsovi = w_ace + l_ace))
tvrdaAsovi <- na.omit(transmute(tvrdaAsovi, ukupnoAsovi = w_ace + l_ace))
```

```
podlogeAsovi <- na.omit(transmute(matches, podloga = surface, ukupnoAsovi = w_ace +
↪  l_ace))

podlogeAsovi$podloga <- as.factor(podlogeAsovi$podloga)
```

Lillieforsovim testom provjeravamo pretpostavku normalnosti

```
lillie.test(podlogeAsovi$ukupnoAsovi)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  podlogeAsovi$ukupnoAsovi
D = 0.12254, p-value < 2.2e-16
```

```
lillie.test(tepihAsovi$ukupnoAsovi)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  tepihAsovi$ukupnoAsovi
D = 0.10864, p-value < 2.2e-16
```

```
lillie.test(zemljaAsovi$ukupnoAsovi)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  zemljaAsovi$ukupnoAsovi
D = 0.13506, p-value < 2.2e-16
```

```
lillie.test(travaAsovi$ukupnoAsovi)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  travaAsovi$ukupnoAsovi
D = 0.10802, p-value < 2.2e-16
```

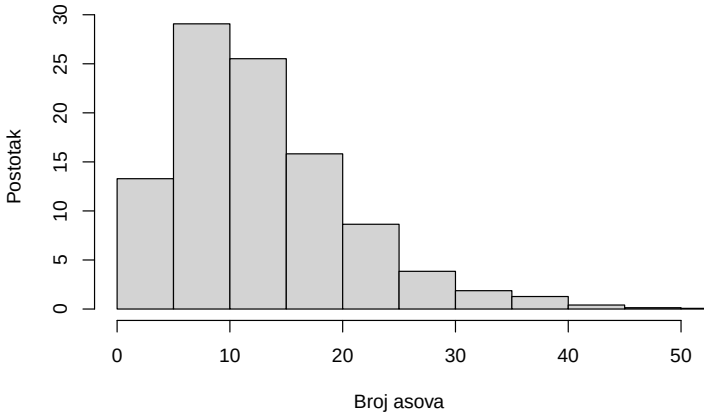
```
lillie.test(tvrdaAsovi$ukupnoAsovi)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test
```

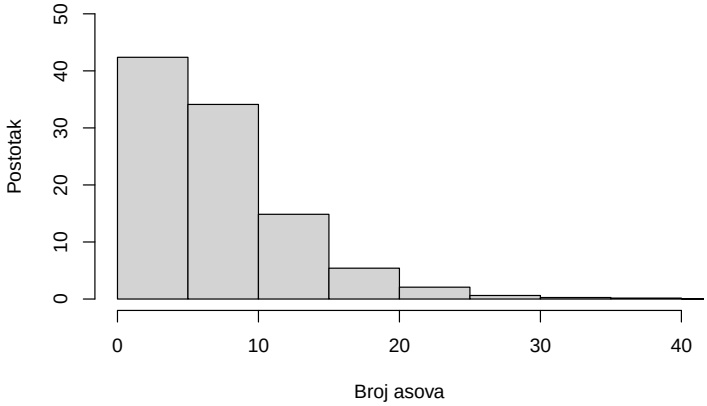
```
data:  tvrdaAsovi$ukupnoAsovi
D = 0.11437, p-value < 2.2e-16
```

Razina značajnosti je 0.05 te zbog p-vrijednosti manje od 0.05 odbacujemo tvrdnju da uzorci proizlaze iz populacije s normalnom distribucijom.

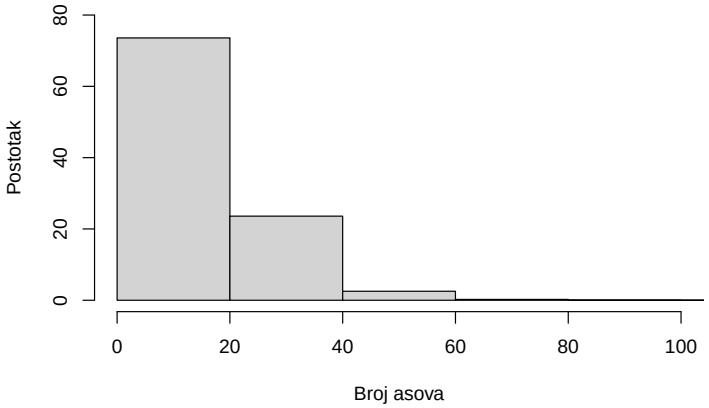
Također nam to i potvrđuju histogrami:



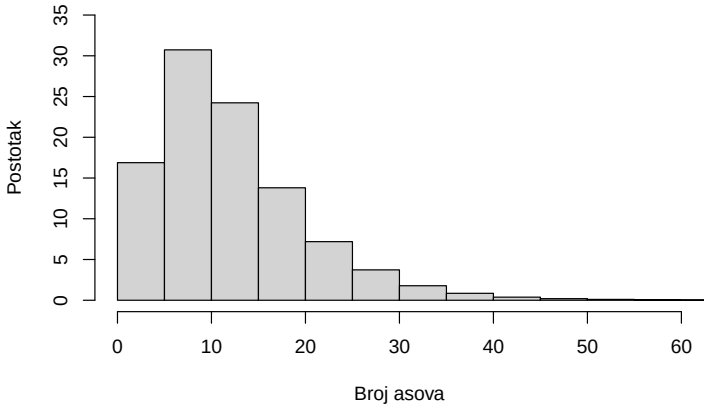
(a) Tepih



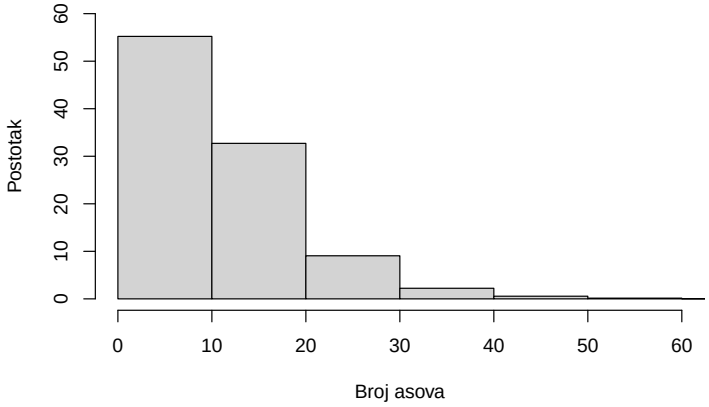
(b) Zemlja



(c) Trava



(d) Tvrda podloga



(e) Sve podloge

Pošto smo utvrdili da uzorci ne proizlaze iz normalno distribuirane populacije, ne provodimo Bartlettov test homogenosti varijanci. Ne provodimo ga jer je test osjetljiv na odstupanja od normalne razdiobe. Odnosno, ako uzorci dolaze iz nenormalnih distribucija, onda Bartlettov test samo provjerava nenormalnost.

Također smo utvrdili da ne možemo provesti ANOVA-u, budući da nismo zadovoljili sve zahtjeve ANOVA-e. Zato provodimo Kruskal-Wallisov test koji ne pretpostavlja normalnu distribuciju i uspoređuje srednje rangove koristeći varijancu rangova.

Pretpostavke za Kruskal-Wallisov test:

- Pretpostavlja se da podaci nisu normalni
- Test se najčešće koristi u analizi tri ili više skupina
- Pretpostavlja se da će podaci imati sličnu distribuciju po skupinama.
- Podaci bi trebali biti nasumično odabrani neovisni uzorci, tako da skupine ne bi trebale biti međusobno povezane.
- Svaki grupni uzorak trebao bi imati najmanje 5 promatranja za dovoljnu veličinu uzorka.

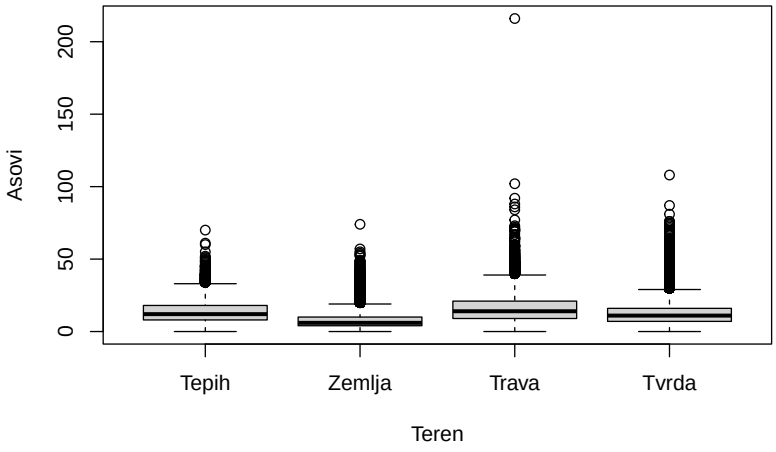
H_0 : Medijani distribucija svih uzoraka su jednaki.

H_1 : Barem dva medijana svih uzoraka nisu jednaka.

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \sim \chi^2(k-1)$$

```
kruskal.test(ukupnoAsovi ~ podloga, data = podlogeAsovi)
```

```
    Kruskal-Wallis rank sum test

data:  ukupnoAsovi by podloga
Kruskal-Wallis chi-squared = 13658, df = 3, p-value < 2.2e-16
```



Zbog razine značajnosti 0.05 i zbog p-vrijednosti manje od 0.05, odbacujemo nul-hipotezu i zaključujemo da uzorci ne proizlaze iz iste populacije. To znači da postoji razlika u broju serviranih asova na različitim podlogama.

2.4 Veza između vrste terena i ulaska u peti set

Četvrto istraživačko pitanje:
Kakva je veza između vrste terena i vjerojatnosti da će mečevi otići u peti set?

Stvaramo kontingencijsku tablicu:

```
pet_setova <- filter(matches, best_of == 5)
pet_setova <- select(pet_setova, score, surface)
pet_setova$score <- if_else(str_count(pet_setova$score, "-") == 5, T, F)
tab = table(pet_setova)
tab_old = tab
tab = addmargins(tab)
tab
```

```
       surface
score   Carpet  Clay Grass  Hard   Sum
  FALSE    700  5550  3471  9090 18811
  TRUE     179  1240   819  2054  4292
  Sum      879  6790  4290 11144 23103
```

Da bismo mogli provesti Hi-kvadrat test nezavisnosti moramo zadovoljiti ove kriterije:

- Broj stupnjeva slobode

$$v = (brojStupaca - 1) * (brojRedaka - 1) > 1$$

- Očekivana vrijednost ćelije trebala bi biti veća ili jednaka od 5.

```
dof <- (nrow(tab_old) - 1) * (ncol(tab_old) - 1)
cat("Degrees of freedom: ", dof, "\n")
```

```
Degrees of freedom:  3
```

```
for (col_names in colnames(tab)){
  for (row_names in rownames(tab)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,':
      ↪  ',(tab[row_names,'Sum'] * tab['Sum',col_names]) / tab['Sum','Sum'],'\n')
    }
  }
}
```

```
Očekivane frekvencije za razred  Carpet - FALSE :  715.7022
Očekivane frekvencije za razred  Carpet - TRUE :  163.2978
```

```
Očekivane frekvencije za razred  Clay - FALSE :  5528.576
Očekivane frekvencije za razred  Clay - TRUE :  1261.424
Očekivane frekvencije za razred  Grass - FALSE :  3493.018
Očekivane frekvencije za razred  Grass - TRUE :  796.9822
Očekivane frekvencije za razred  Hard - FALSE :  9073.704
Očekivane frekvencije za razred  Hard - TRUE :  2070.296
```

Pošto su zahtjevi ispunjeni, provodimo Hi-kvadrat test nezavisnosti nad tablicom.

```
       surface
score   Carpet Clay Grass Hard
  FALSE    700 5550  3471 9090
  TRUE     179 1240   819 2054
```

H_0 : Odlazak meča u peti set ne ovisi o vrsti podloge.

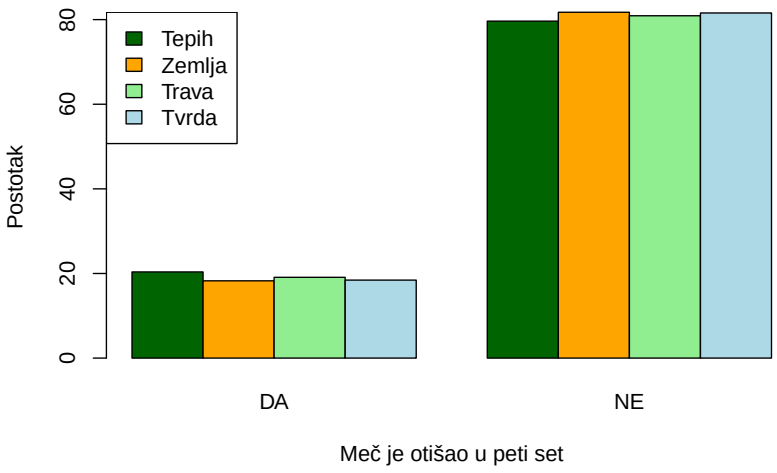
H_1 : Odlazak meča u peti set ovisi o vrsti podloge.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

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

```
    Pearson's Chi-squared test

data:  tab_old
X-squared = 3.2059, df = 3, p-value = 0.361
```

Nulta hipoteza se ne odbacuje na razini značajnosti 0.05, p-vrijednost nije manja od 0,05. Zaključujemo da odlazak meča u peti ne set ovisi o vrsti podloge.



Iz grafa vidimo da su vjerojatnosti odlaska i neodlaska u peti set podjednake za svaku vrstu podloge.

2.5 Predviđanje broja asova

Peto istraživačko pitanje:
Možemo li procijeniti broj asova koje će igrač odservirati u tekućoj (zadnjoj dostupnoj sezoni) na temelju njegovih rezultata iz prethodnih sezona?

Da bismo mogli procijeniti broj asova koje će igrač odservirati u tekućoj sezoni temeljem njegovih prethodnih rezultata, sastavit ćemo model linearne regresije.

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

Modelirat ćemo regresiju tako da odservirani asovi u tekućoj godini ovise on njegovoj visini i broju odigranih mečeva prethodnih sezona. Također ćemo s pomoću "dummy" variable testirati ovise li odservirani asovi o tome je li igrač ljevak ili dešnjak.

```
winners <- matches[,c("winner_name", "tourney_date", "winner_hand", "winner_ht",
↪  "w_ace")]
winners$tourney_date <- year(ymd(winners$tourney_date))
winners <- na.omit(winners)
winners <- winners %>% group_by(winner_name, tourney_date, winner_ht, winner_hand)
↪  %>% summarise(broj_meceva = n(), broj_asova = sum(w_ace))
losers <- matches[,c("loser_name", "tourney_date", "loser_hand", "loser_ht",
↪  "l_ace")]
losers$tourney_date <- year(ymd(losers$tourney_date))
losers <- na.omit(losers)
losers <- losers %>% group_by(loser_name, tourney_date, loser_ht, loser_hand) %>%
↪  summarise(broj_meceva = n(), broj_asova = sum(l_ace))

colnames(losers) <- colnames(winners)
svi_igraci_tab <- rbind(winners, losers)
svi_igraci_tab <- svi_igraci_tab %>% group_by(winner_name, tourney_date, winner_ht,
↪  winner_hand) %>% summarise(broj_meceva = sum(broj_meceva), broj_asova =
↪  sum(broj_asova))
svi_igraci_tab_2023 <- svi_igraci_tab[svi_igraci_tab$tourney_date == 2023,]
svi_igraci_tab <- svi_igraci_tab[svi_igraci_tab$tourney_date < 2023,]
```

Lillieforsovim testom provjeravamo dolaze li varijable visina igrača i broj mečeva iz normalnih populacija.

```
lillie.test(svi_igraci_tab$winner_ht)
```

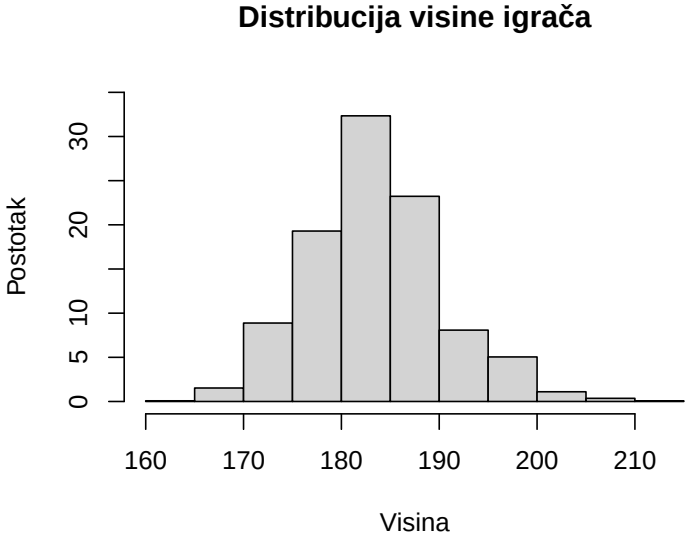
```
    Lilliefors (Kolmogorov-Smirnov) normality test
```

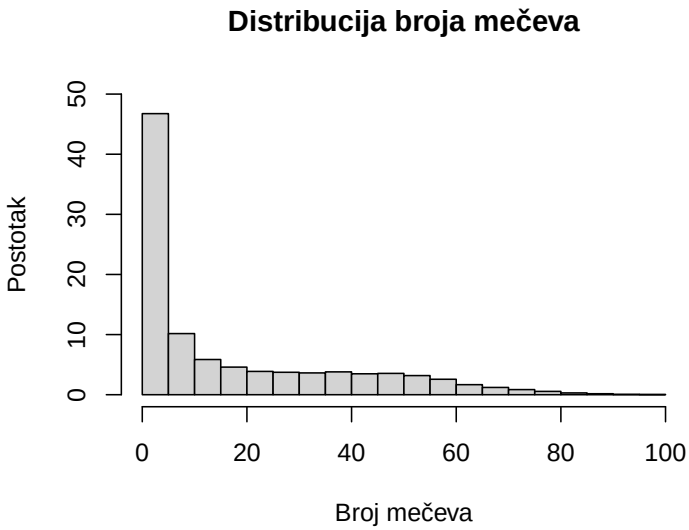
```
data:  svi_igraci_tab$winner_ht
D = 0.10544, p-value < 2.2e-16
```

```
lillie.test(svi_igraci_tab$broj_meceva)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  svi_igraci_tab$broj_meceva
D = 0.21884, p-value < 2.2e-16
```

Razina značajnosti je 0.05 te zbog p-vrijednosti manje od 0.05 odbacujemo tvrdnju da varijable proizlaze iz populacije s normalnom distribucijom. Dodatno ćemo provjeriti ovu tvrdnju crtajući grafove.





Graf distribucije visine igrača ukazuje na normalnu distribuciju iako nas Lillieforsov test navodi da zaključimo suprotno. Takav nije slučaj s distribucijom broja mečeva jer nam i oblik grafa pokazuje da se ne radi o normalnoj distribuciji.

Pošto o pretpostavci normalnosti ovisi hoćemo li koristiti Pearsonov ili Spearmanov korelacijski koeficijent, izračunat ćemo oba i vidjeti što nam govore.

```
cor.test(svi_igraci_tab$winner_ht, svi_igraci_tab$broj_meceva)
```

```
    Pearson's product-moment correlation

data:  svi_igraci_tab$winner_ht and svi_igraci_tab$broj_meceva
t = 10.12, df = 10347, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07988135 0.11803732
sample estimates:
       cor
```

```
0.09899572
```

```
cor.test(svi_igraci_tab$winner_ht, svi_igraci_tab$broj_meceva, method = "spearman")
```

```
    Spearman's rank correlation rho

data:  svi_igraci_tab$winner_ht and svi_igraci_tab$broj_meceva
S = 1.6825e+11, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.08923651
```

```
cor.test(svi_igraci_tab$winner_ht, svi_igraci_tab$broj_asova, method = "spearman")
```

```
    Spearman's rank correlation rho

data:  svi_igraci_tab$winner_ht and svi_igraci_tab$broj_asova
S = 1.3982e+11, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2431139
```

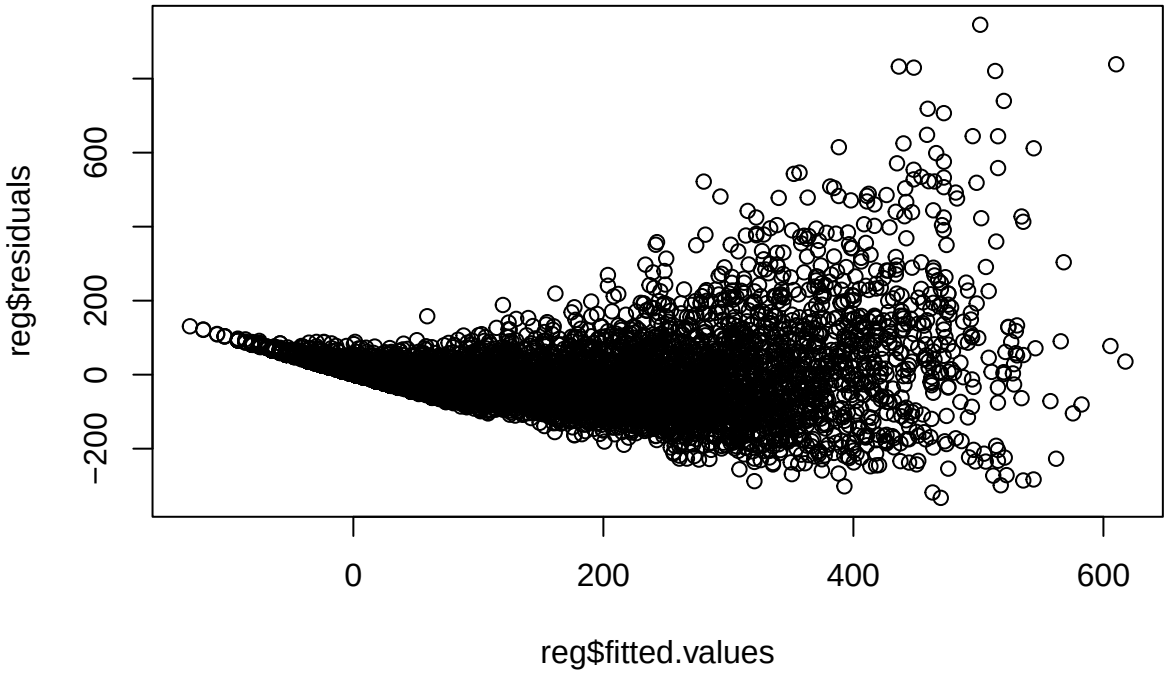
```
cor.test(svi_igraci_tab$broj_meceva, svi_igraci_tab$broj_asova, method = "spearman")
```

```
    Spearman's rank correlation rho

data:  svi_igraci_tab$broj_meceva and svi_igraci_tab$broj_asova
S = 1.2916e+10, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9300823
```

Pearsonov i Spearmanov korelacijski koeficijent visine igrača i broja mečeva nam oba ukazuju na vrlo slabu korelaciju s s obzirom na to da je rezultat u intervalu [0, 0.19]. Zaključujemo da je korelacija više nego dovoljno slaba da bismo te dvije varijable koristili pri modeliranju regresijskog modela.

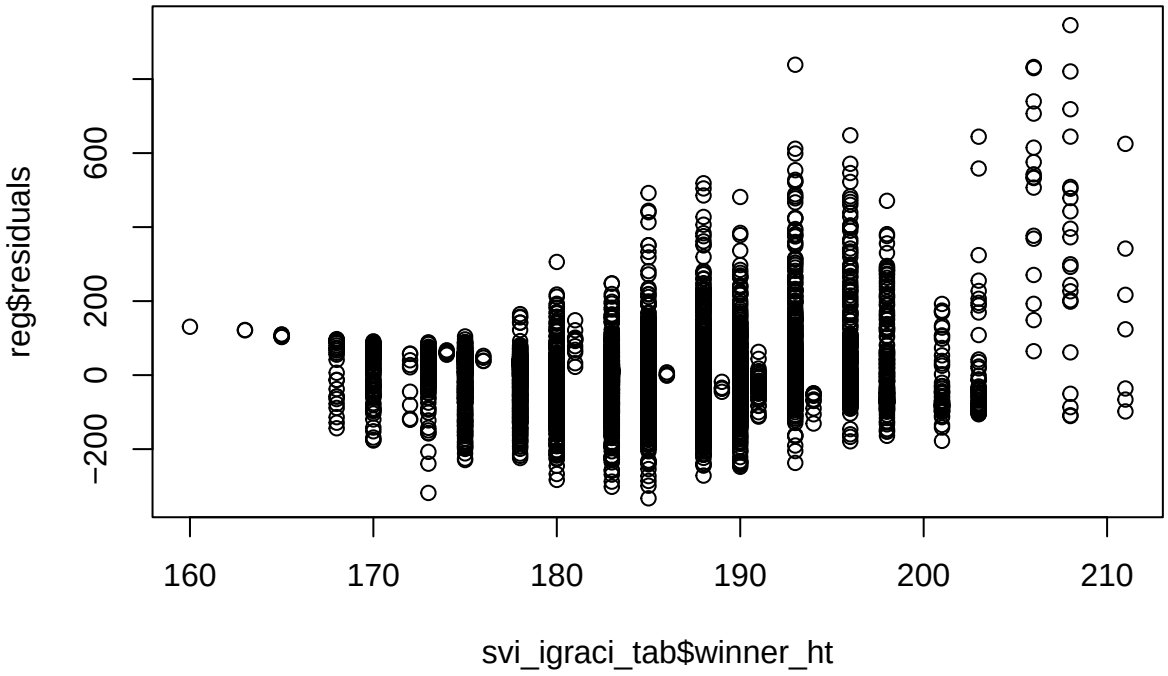
Stvaramo regresijski model i crtamo graf ovisnosti reziduala o vrijednostima prema kojima je model istreniran.

```
reg = lm(broj_asova ~ winner_ht + broj_meceva, svi_igraci_tab)
plot(reg$fitted.values, reg$residuals)
```



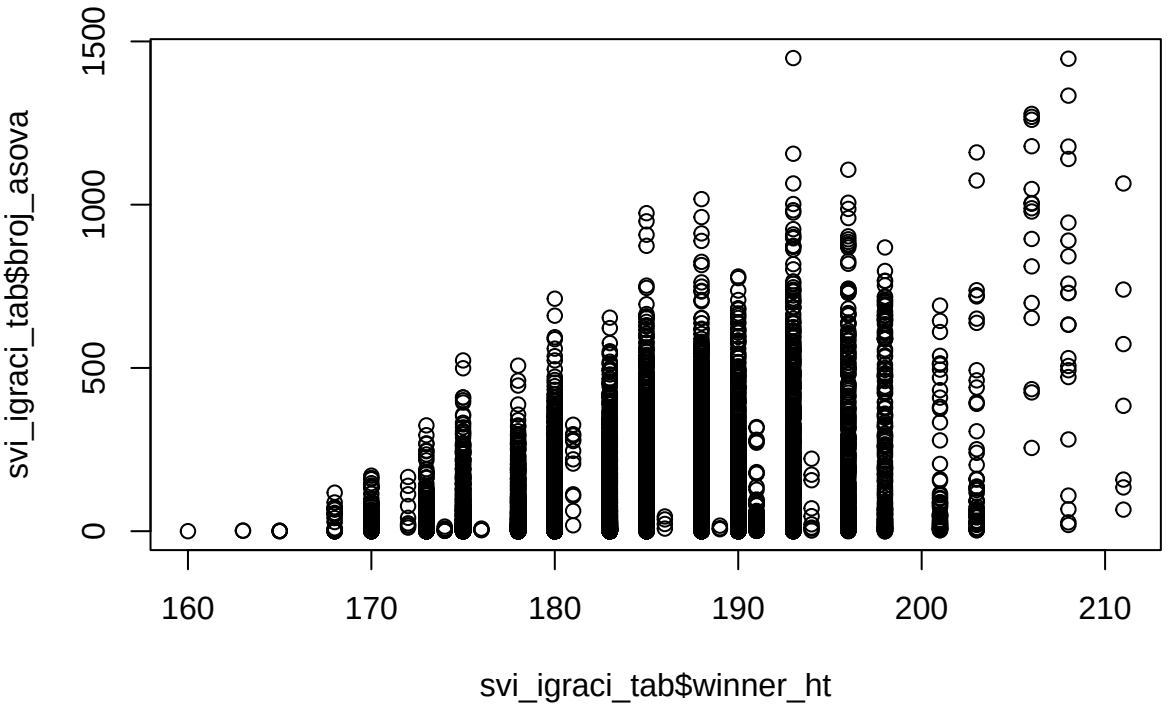
Ovisnost reziduala o visini igrača

```
plot(svi_igraci_tab$winner_ht, reg$residuals)
```



Ovisnost broja asova o visini igrača

```
plot(svi_igraci_tab$winner_ht, svi_igraci_tab$broj_asova)
```




```
summary(reg)
```

```
Call:
lm(formula = broj_asova ~ winner_ht + broj_meceva, data = svi_igraci_tab)

Residuals:
    Min      1Q  Median      3Q     Max
-332.89  -33.29   -1.64   27.16  945.56

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.030e+03  2.246e+01  -45.86   <2e-16 ***
winner_ht    5.545e+00  1.218e-01   45.53   <2e-16 ***
broj_meceva  5.999e+00  3.813e-02  157.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.88 on 10346 degrees of freedom
Multiple R-squared:  0.7339,    Adjusted R-squared:  0.7338
F-statistic: 1.426e+04 on 2 and 10346 DF,  p-value: < 2.2e-16
```

Ovaj sažetak nam pokazuje da visina i broj mečeva značajno utječu na broj asova. Spearmanov koeficijent nam je ukazivao da visina igrača slabo korelira s brojem asova pa možemo zaista zaključiti da je riječ o Simpsonovom paradoksu.

Koeficijent determinacije je čak 73.39 %, a prilagođeni koeficijent determinacije 73.38 %. Koeficijent je

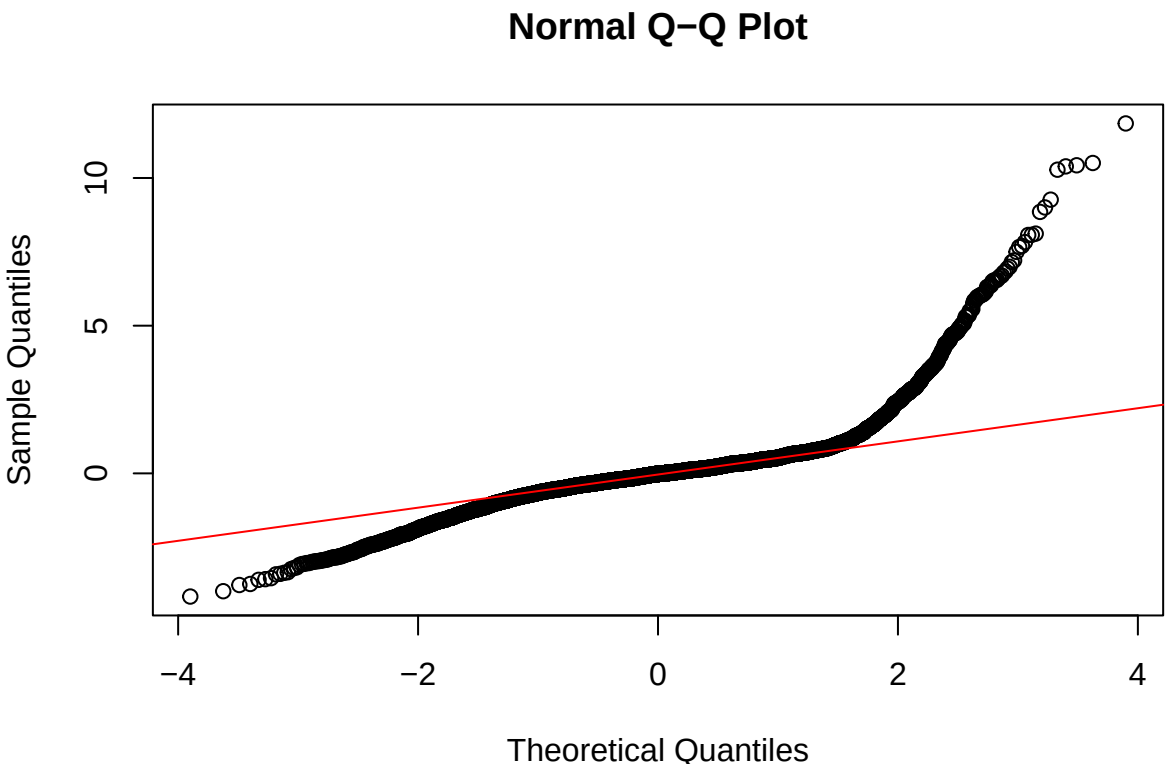
vrlo visok, što znači da model odlično predviđa broj asova ovisno o visini igrači i broju mečeva u sezoni (73.39% varijance u podacima je objašnjeno linearnim modelom).

```
svi_igraci_tab.d = dummy_cols(svi_igraci_tab, select_columns = "winner_hand")
reg.d = lm(broj_asova ~ winner_ht + broj_meceva + winner_hand_L, svi_igraci_tab.d)
summary(reg.d)
```

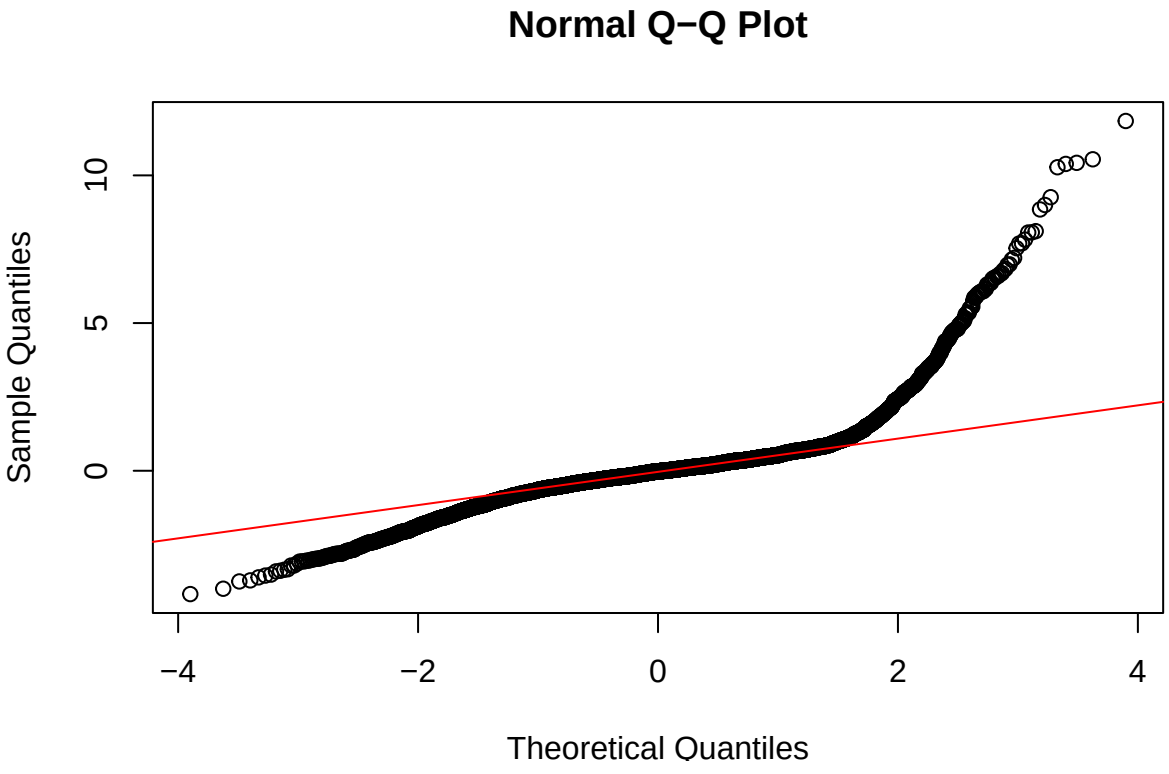
```
Call:
lm(formula = broj_asova ~ winner_ht + broj_meceva + winner_hand_L,
    data = svi_igraci_tab.d)

Residuals:
    Min      1Q  Median      3Q     Max
-333.36  -33.27   -1.37   27.37  945.06

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.030e+03  2.246e+01 -45.856   <2e-16 ***
winner_ht      5.547e+00  1.218e-01  45.542   <2e-16 ***
broj_meceva    6.000e+00  3.813e-02 157.355   <2e-16 ***
winner_hand_L -3.128e+00  2.251e+00  -1.389    0.165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.88 on 10345 degrees of freedom
Multiple R-squared:  0.7339,    Adjusted R-squared:  0.7338
F-statistic:  9510 on 3 and 10345 DF,  p-value: < 2.2e-16
```

Velika p vrijednost kod lijeve ruke na nam ukazuje na to da ruka ne utječe na broj odserviranih asova.

```
qqnorm(rstandard(reg))
qqline(rstandard(reg), col = "red")
```




```
qqnorm(rstandard(reg.d))
qqline(rstandard(reg.d), col = "red")
```



Predviđamo broj asova za nekoliko tenisača

```
predvidi <- function(ime) {
  cat(ime, "(", svi_igraci_tab_2023[svi_igraci_tab_2023$winner_name ==
  ↪  ime,c("winner_ht", "broj_meceva")]$winner_ht, ",",
      svi_igraci_tab_2023[svi_igraci_tab_2023$winner_name == ime,c("winner_ht",
↪  "broj_meceva")]$broj_meceva, ")\n")
```

```
  broj_asova_stv <- svi_igraci_tab_2023[svi_igraci_tab_2023$winner_name ==
↪  ime,c("broj_asova")]
  p <- predict(reg, svi_igraci_tab_2023[svi_igraci_tab_2023$winner_name ==
↪  ime,c("winner_ht", "broj_meceva")])
 cat("predviđen broj asova: ", p , ", stvaran: ",  round(broj_asova_stv$broj_asova,
 ↪  2), ", rezidual: ", round( broj_asova_stv$broj_asova - p, 2), "\n\n")
}
cat("Tenisač ( visina , broj mečeva )\n\n")
```

```
Tenisač ( visina , broj mečeva )
```

```
predvidi("Novak Djokovic")
```

```
Novak Djokovic ( 188 , 50 )
predviđen broj asova:  312.539 , stvaran:  333 , rezidual:  20.46
```

```
predvidi("Carlos Alcaraz")
```

```
Carlos Alcaraz ( 185 , 65 )
predviđen broj asova:  385.8949 , stvaran:  261 , rezidual:  -124.89
```

```
predvidi("Daniil Medvedev")
```

```
Daniil Medvedev ( 198 , 67 )
predviđen broj asova:  469.9831 , stvaran:  469 , rezidual:  -0.98
```

```
predvidi("Borna Coric")
```

```
Borna Coric ( 188 , 39 )
predviđen broj asova:  246.5449 , stvaran:  190 , rezidual:  -56.54
```

```
predvidi("Dominic Thiem")
```

```
Dominic Thiem ( 185 , 31 )
predviđen broj asova:  181.9132 , stvaran:  160 , rezidual:  -21.91
```

```
predvidi("Borna Gojo")
```

```
Borna Gojo ( 196 , 13 )
predviđen broj asova:  134.9214 , stvaran:  170 , rezidual:  35.08
```