

# ATP Data Analysis

19.1.2024.

Instalacija potrebnih paketa.

```
# install.packages("dplyr")
# install.packages("lubridate")
# install.packages("ggplot2")
# install.packages("caret")
# install.packages("nortest")
# install.packages("fastDummies")
# install.packages("car")
```

Učitavanje biblioteka.

```
library(dplyr)
library(lubridate)
library(ggplot2)
library(caret)
library(nortest)
library(fastDummies)
library(car)
```

Učitavanje i opis podataka

```
all_matches <- data.frame()
for (year in 1991:2023) {
  file_name <- paste0("dataset/atp_matches_", year, ".csv")
  matches_year <- read.csv(file_name, stringsAsFactors = FALSE)
  all_matches <- rbind(all_matches, matches_year)
}

dim(all_matches)
```

## [1] 104682 49

Skup podataka sadrži informacije o 104682 teniska meča održana od 1991. do 2023. godine uključivo. Svaki meč opisan je s 49 ispod navedenih značajki:

```
names(all_matches)

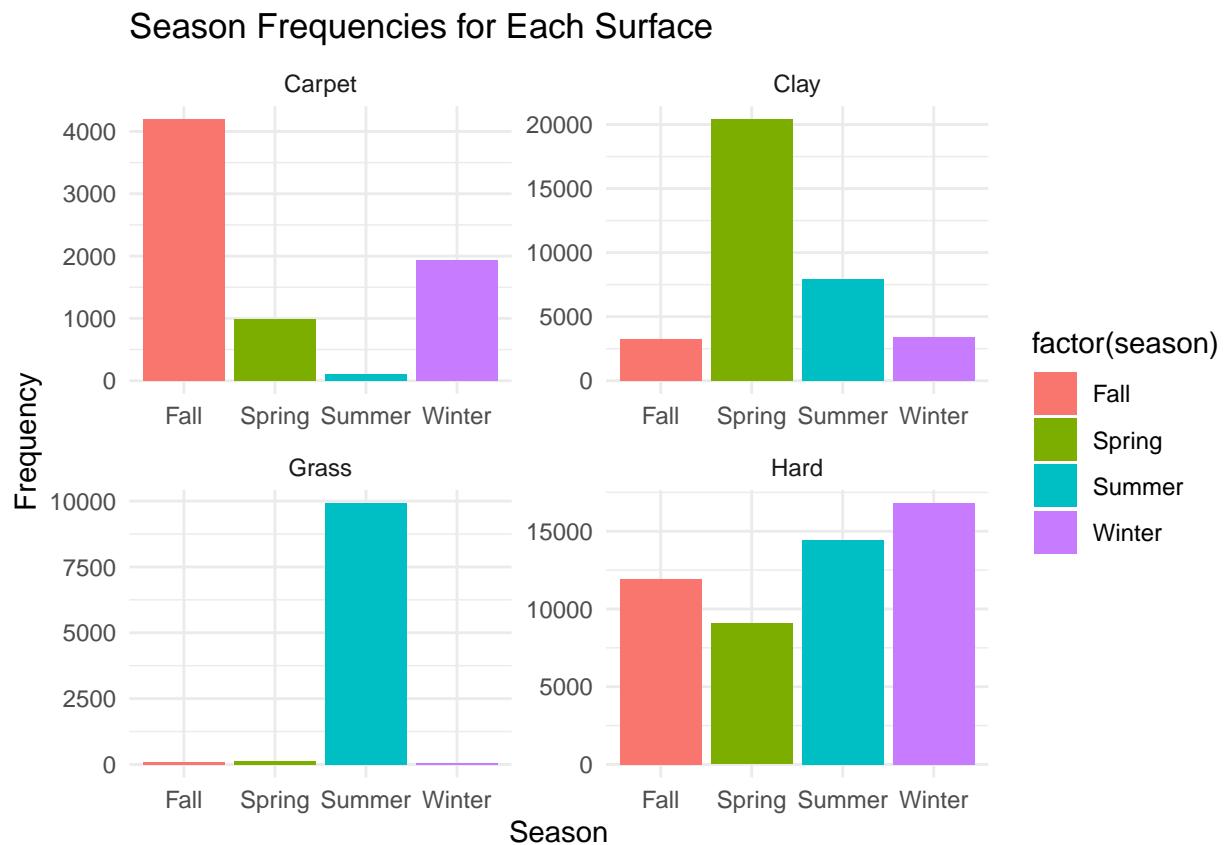
## [1] "tourney_id"          "tourney_name"        "surface"
## [4] "draw_size"           "tourney_level"       "tourney_date"
## [7] "match_num"           "winner_id"          "winner_seed"
## [10] "winner_entry"        "winner_name"         "winner_hand"
```

```

## [13] "winner_ht"
## [16] "loser_id"
## [19] "loser_name"
## [22] "loser_ioc"
## [25] "best_of"
## [28] "w_ac"
## [31] "w_1stIn"
## [34] "w_SvGms"
## [37] "l_ac"
## [40] "l_1stIn"
## [43] "l_SvGms"
## [46] "winner_rank"
## [49] "loser_rank_points"
## [13] "winner_ioc"
## [16] "loser_seed"
## [19] "loser_hand"
## [22] "loser_age"
## [25] "round"
## [28] "w_df"
## [31] "w_1stWon"
## [34] "w_bpSaved"
## [37] "l_df"
## [40] "l_1stWon"
## [43] "l_bpSaved"
## [46] "winner_rank_points"
## [49] "loser_rank"

```

Zadatak 1. Kakva je distribucija mečeva na specifičnim podlogama u različitim godišnjim dobima?



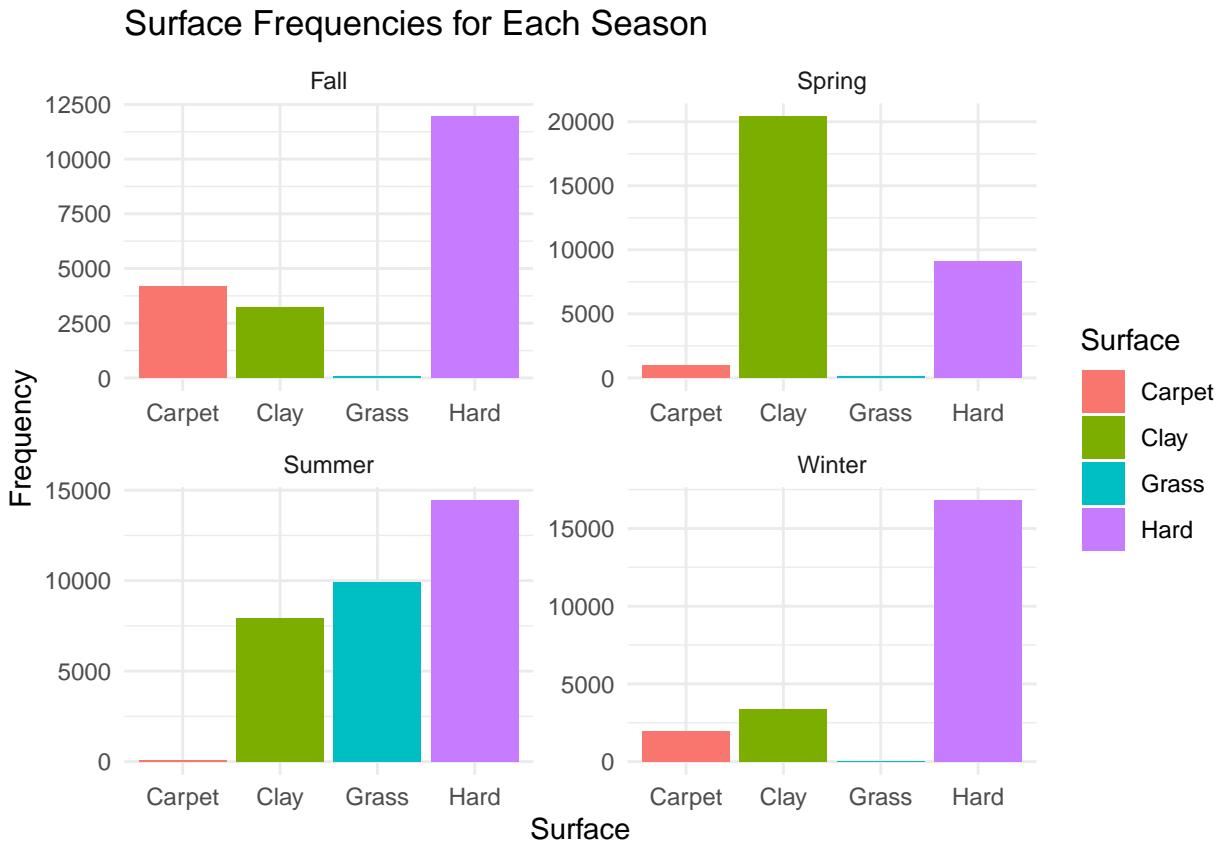
U prvom histogramu prikazana je raspodjela teniskih mečeva prema godišnjim dobima na podlozi od tepiha. Podloga od tepiha najmanje je korištena podloga za igranje mečeva. Najčešće se podloga od tepiha koristila u jesen, dosta rjeđe zimi, zatim na proljeće, a najmanje se mečeva na podlozi od tepiha igra na ljeto.

Sljedeći histogram predstavlja raspodjelu mečeva prema godišnjim dobima na zemljanoj podlozi. Mečevi na zemlji najčešće se igraju u proljetnom dijelu sezone. Dosta manje mečeva igra se na ljeto zatim otprilike podjednako na jesen i zimi.

Treći histogram opisuje distribuciju teniskih mečeva prema godišnjim dobima na travi. Teniski mečevi na

travi igraju se uglavnom ljeti, a svega nekoliko mečeva igra se u preostalim godišnjim dobima.

U posljednjem histogramu promatrana je raspodjela mečeva prema godišnjim dobima na tvrdoj podlozi. Sveukupno najviše mečeva igra se na tvrdoj podlozi te je raspodjela prema godišnjim dobima manje izražena nego kod drugih podloga. Najviše mečeva na tvrdoj podlozi održava se zimi, zatim u ljeto pa na jesen te najmanje u proljetnom dijelu sezone.



Prvi histogram prikazuje raspodjelu mečeva prema podlogama u jesen. Uvjerljivo najviše mečeva u jesen održava se na tvrdoj podlozi. Dosta manje mečeva igra se na podlozi od tepiha, a nešto malo manje na zemlji. Najmanje mečeva u jesenskom dijelu sezone igra se na travi.

Idući histogram prikazuje raspodjelu mečeva prema podlogama u proljeće. U proljetnom dijelu sezone uvjerljivo najviše teniskih mečeva igra se na podlozi od zemlje. Više od dvostruko manje mečeva održava se na tvrdoj podlozi. Jako malo mečeva održava se na podlozi od tepiha, a još manje na travi.

U trećem histogramu promatramo raspodjelu mečeva prema podlogama tijekom ljeta. Najviše mečeva održava se na tvrdoj podlozi, zatim na travi pa na podlozi od zemlje. Svega nekoliko mečeva igra se na podlozi od tepiha.

Zadnji histogram opisuje raspodjelu mečeva prema podlogama zimi. Tijekom zime prednjače mečevi na tvrdoj podlozi. Dosta manje mečeva igra se na zemlji, zatim na podlozi od tepiha te najmanje na travi.

**Zadatak 2. Postoji li značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu?**

Na početku ispisujemo statistiku o podatcima, prvo za mečeve odigrane na otvorenom pa na zatvorenom:

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 0.000 3.000 5.000 6.221 8.000 42.000
```

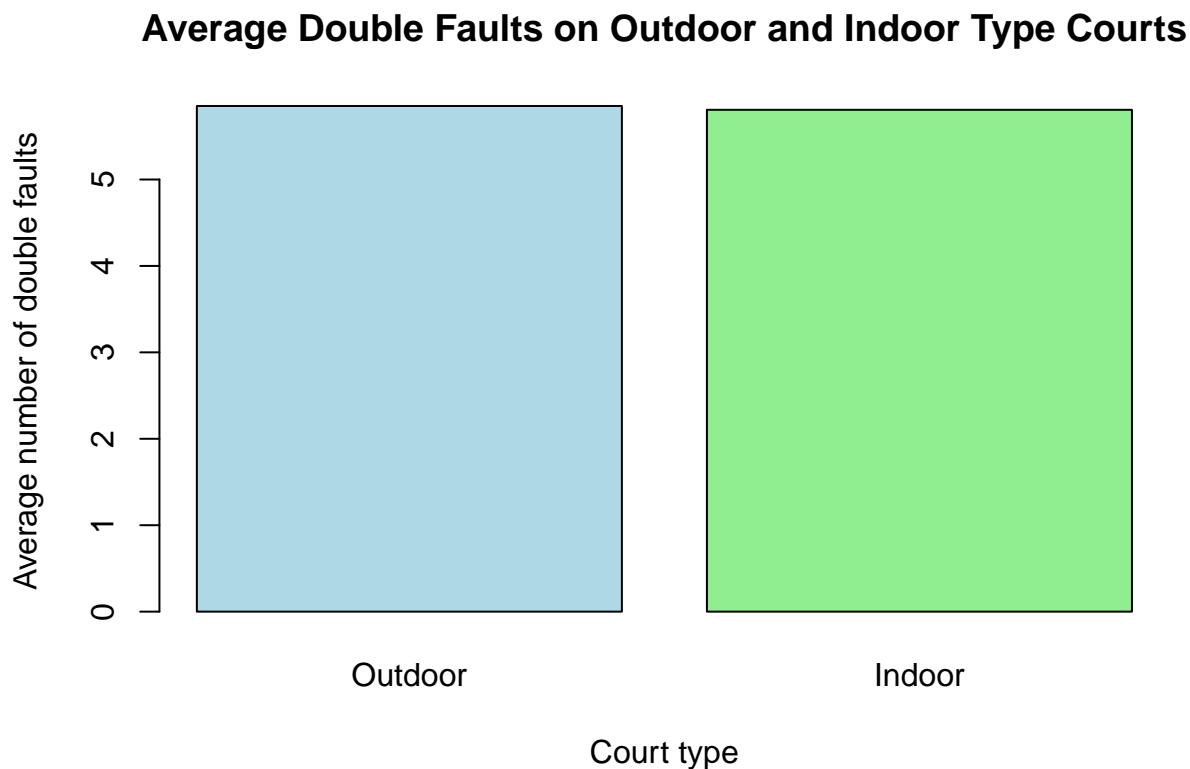
```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 0.000 4.000 6.000 6.177 8.000 30.000
```

S obzirom na veliku razliku između mean i max. vrijednosti pronađazimo outliere te ih izbacujemo iz podataka:

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 0.000 3.000 5.000 5.851 8.000 15.000
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 0.000 3.000 5.000 5.806 8.000 14.000
```

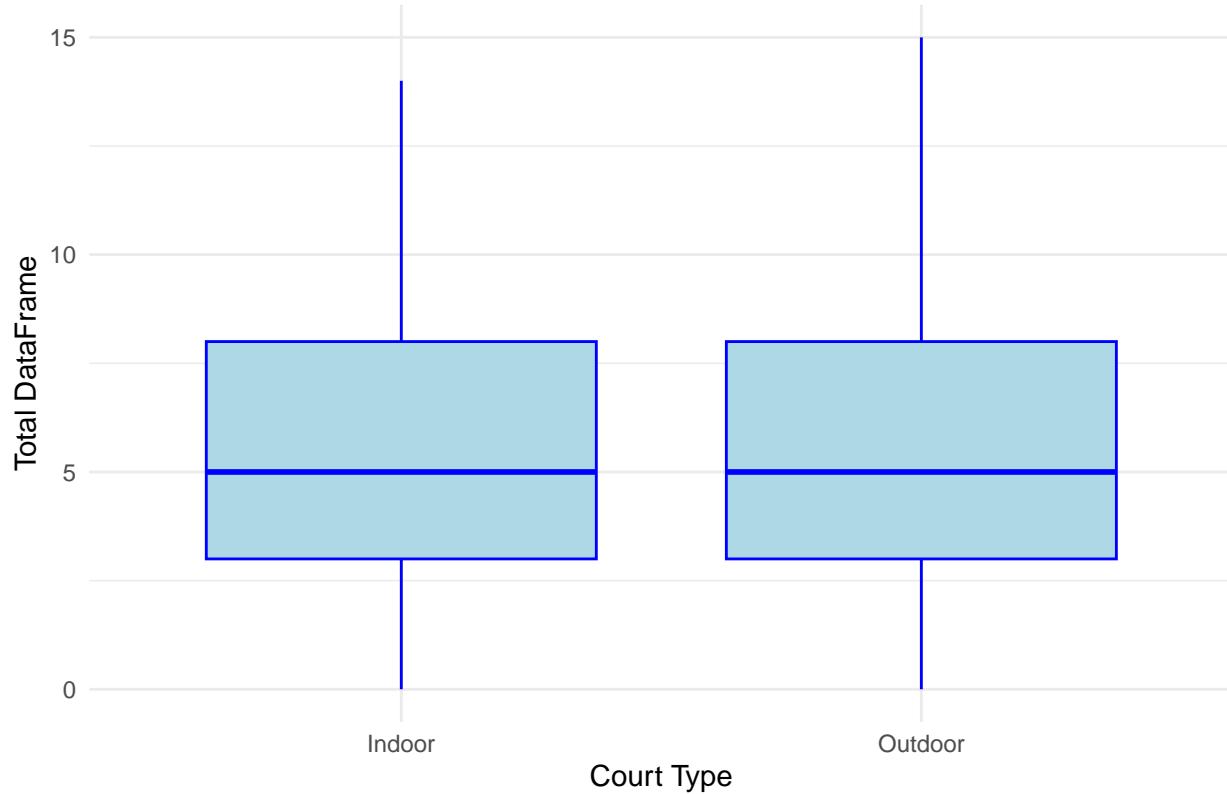
Nakon toga stvaramo bar plot za usporedbu prosječnog broja dvostrukih pogrešaka ovisno o tome jesu li mečevi odigrani na otvorenom ili zatvorenom:



Iz barplota vidljivo je da su prosjecne dvostrukе pogreške podjednake na unutarnjim i vanjskim terenima.

Zatim provjeravamo ukazuju li boxplot na moguću značajnu razliku.

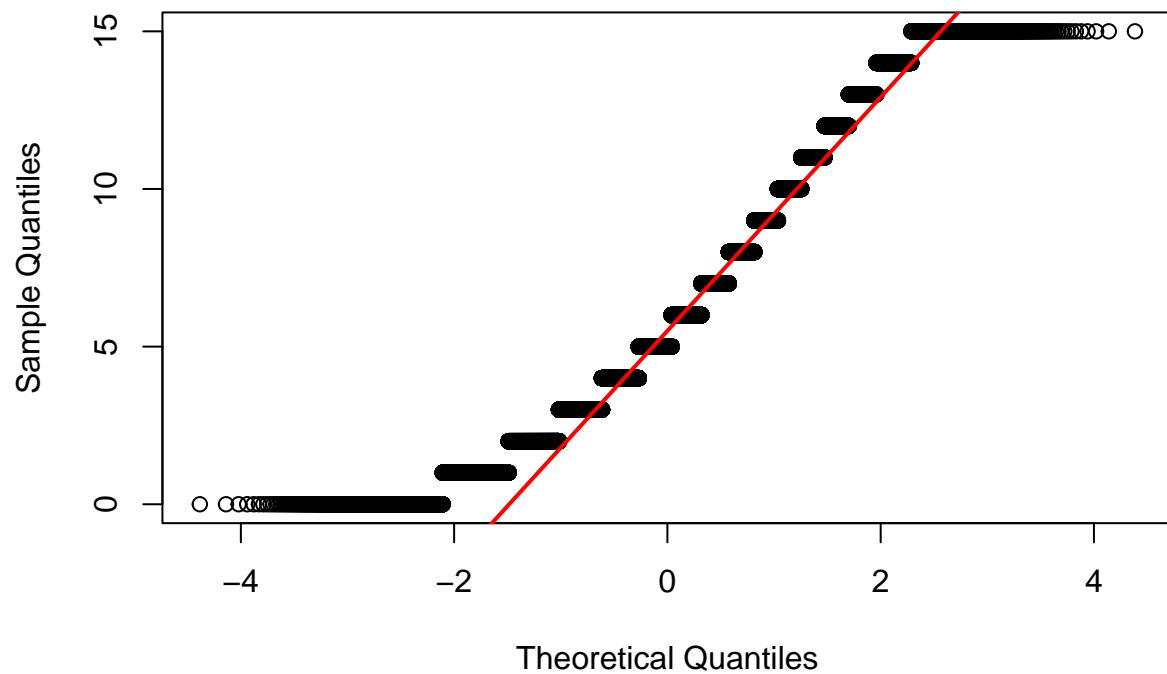
## Comparison of Outdoor and Indoor Court Types



Boxplot ukazuje na to da nema razlike između prosječnog broja dvostrukih pogrešaka između mečeva održanih na otvorenom i zatvorenom. Kako bismo provjerili možemo li prihvati nullu hipotezu koja pretostavlja da nema razlike, provedećemo t-test. Najprije moramo provjeriti pretpostavke o normalnoj distribuciji i homogenosti varijanci. Normalnu distribuciju prvo provjeravamo pomoću qq-plota, a zatim i Lilliefors testom.

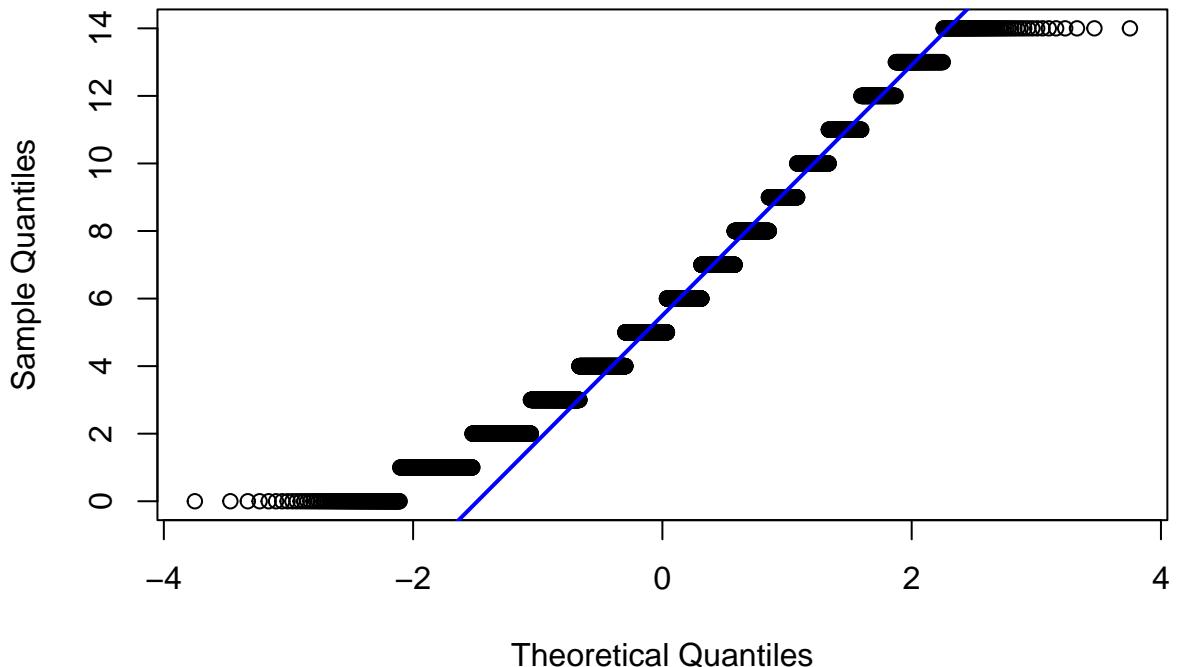
```
qqnorm(open_surface_data_no_outliers, main = "Normal Q-Q Plot for outdoor courts")
qqline(open_surface_data_no_outliers, col = "red", lwd = 2)
```

### Normal Q-Q Plot for outdoor courts



```
qqnorm(closed_surface_data_no_outliers, main = "Normal Q-Q Plot for indoor courts")
qqline(closed_surface_data_no_outliers, col = "blue", lwd = 2)
```

## Normal Q-Q Plot for indoor courts



Iz qq-plota vidljivo je da su distribucije približno normalne uz minimalna odstupanja pa zbog velike dimenzije uzorka možemo reći da je pretpostavka o normalnosti zadovoljena i za mečeve na otvorenom i za mečeve u zatvorenom. Zatim provodimo Lilliefors test:

```
lillie_test_outdoor <- lillie.test(open_surface_data_no_outliers)
lillie_test_indoor <- lillie.test(closed_surface_data_no_outliers)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: open_surface_data_no_outliers
## D = 0.11612, p-value < 2.2e-16

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: closed_surface_data_no_outliers
## D = 0.11589, p-value < 2.2e-16
```

Prema Lilliefors testu odbacili bi pretpostavku o normalnosti distribucije, ali zbog značajno velike količine podataka i centralnog graničnog teorema zaključujemo da je pretpostavka ipak zadovoljena.

Provodimo sada F-test za provjeru homogenosti varijanci:

```

var_test_result <- var.test(open_surface_data_no_outliers, closed_surface_data_no_outliers)
print(var_test_result)

##
## F test to compare two variances
##
## data: open_surface_data_no_outliers and closed_surface_data_no_outliers
## F = 1.1235, num df = 86050, denom df = 5681, p-value = 3.576e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.081253 1.166568
## sample estimates:
## ratio of variances
## 1.12346

```

F-test za usporedbu varijanci pokazuje da postoji razlika u varijancama između otvorenog terena i zatvorenog terena (p-vrijednost < 0.05).

Provjerimo jesu li veličine uzoraka podjednake.

```
print(length(open_surface_data_no_outliers))
```

```
## [1] 86051
```

```
print(length(closed_surface_data_no_outliers))
```

```
## [1] 5682
```

Kako veličine uzoraka nisu približno jednake ne možemo pretpostaviti homogenost varijanci na temelju toga što imamo mnogo podataka pa koristimo neparametarsku alternativu t-testu - Wilcoxon rank-sum test.

```

#t.test(open_surface_data, closed_surface_data, var.equal = TRUE)
print(wilcox.test(open_surface_data_no_outliers, closed_surface_data_no_outliers, alternative = "two.sided"))

##
## Wilcoxon rank sum test with continuity correction
##
## data: open_surface_data_no_outliers and closed_surface_data_no_outliers
## W = 243522192, p-value = 0.6221
## alternative hypothesis: true location shift is not equal to 0

```

Wilcoxon rank-sum test ne pokazuje značajnu razliku u srednjim vrijednostima (medijanama) između otvorenog i zatvorenog terena (p-vrijednost > 0.05, dakle ne možemo odbaciti nullu hipotezu).

Na temelju ovih rezultata, možemo zaključiti da nema značajne razlike u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom terenu i mečeva odigranih na zatvorenom terenu.

### Zadatak 3. Ima li razlike u broju serviranih asova na različitim podlogama?

Analizirajmo deskriptivnu statistiku za svaku od podloga:

```
## Summary for Hard Surface:  
  
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  
##      0.00   7.00 11.00   12.58   16.00 108.00  
  
##  
## Summary for Clay Surface:  
  
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  
##      0.00   4.00  6.00    7.62   10.00  74.00  
  
##  
## Summary for Carpet Surface:  
  
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  
##      0.00   8.00 12.00   13.36   18.00  70.00  
  
##  
## Summary for Grass Surface:  
  
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  
##      0.00   9.00 14.00   16.16   21.00 216.00
```

Izbacit ćemo stršeće podatke čija se prisutnost jasno vidi u maksimumu na svakoj od podloga te opet ispisati statistiku.

```
remove_outliers <- function(data) {  
  q <- quantile(data, c(0.25, 0.75), na.rm = TRUE)  
  iqr <- q[2] - q[1]  
  lower_bound <- q[1] - 1.5 * iqr  
  upper_bound <- q[2] + 1.5 * iqr  
  
  data_filtered <- data[data >= lower_bound & data <= upper_bound]  
  
  return(data_filtered)  
}  
  
hard_aces = remove_outliers(t3$aces[t3$surface == 'Hard'])  
carpet_aces = remove_outliers(t3$aces[t3$surface == 'Carpet'])  
clay_aces = remove_outliers(t3$aces[t3$surface == 'Clay'])  
grass_aces = remove_outliers(t3$aces[t3$surface == 'Grass'])
```

```
## Summary for Hard Surface:  
  
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  
##      0.00   7.00 11.00   11.59   16.00  29.00
```

```

##  

## Summary for Clay Surface:  
  

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  

##      0.000   4.000  6.000   6.911 10.000 19.000  
  

##  

## Summary for Carpet Surface:  
  

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  

##      0.00   8.00  11.00  12.68 17.00 33.00  
  

##  

## Summary for Grass Surface:  
  

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  

##      0.00   9.00  14.00  15.08 20.00 39.00

```

Sada možemo i vizualizirati prosječni broj asova serviranih na svakoj od podloga kako bi uočili potencijalnu razliku.

```

combined_data <- data.frame(  

  surface = rep(c("Hard", "Carpet", "Clay", "Grass"),  

                c(length(hard_aces), length(carpet_aces), length(clay_aces), length(grass_aces))),  

  aces = c(hard_aces, carpet_aces, clay_aces, grass_aces)  

)  
  

ggplot(combined_data, aes(x = surface, y = aces, fill = surface)) +  

  stat_summary(fun = "mean", geom = "bar") +  

  labs(title = "Average Served Aces by Surface",  

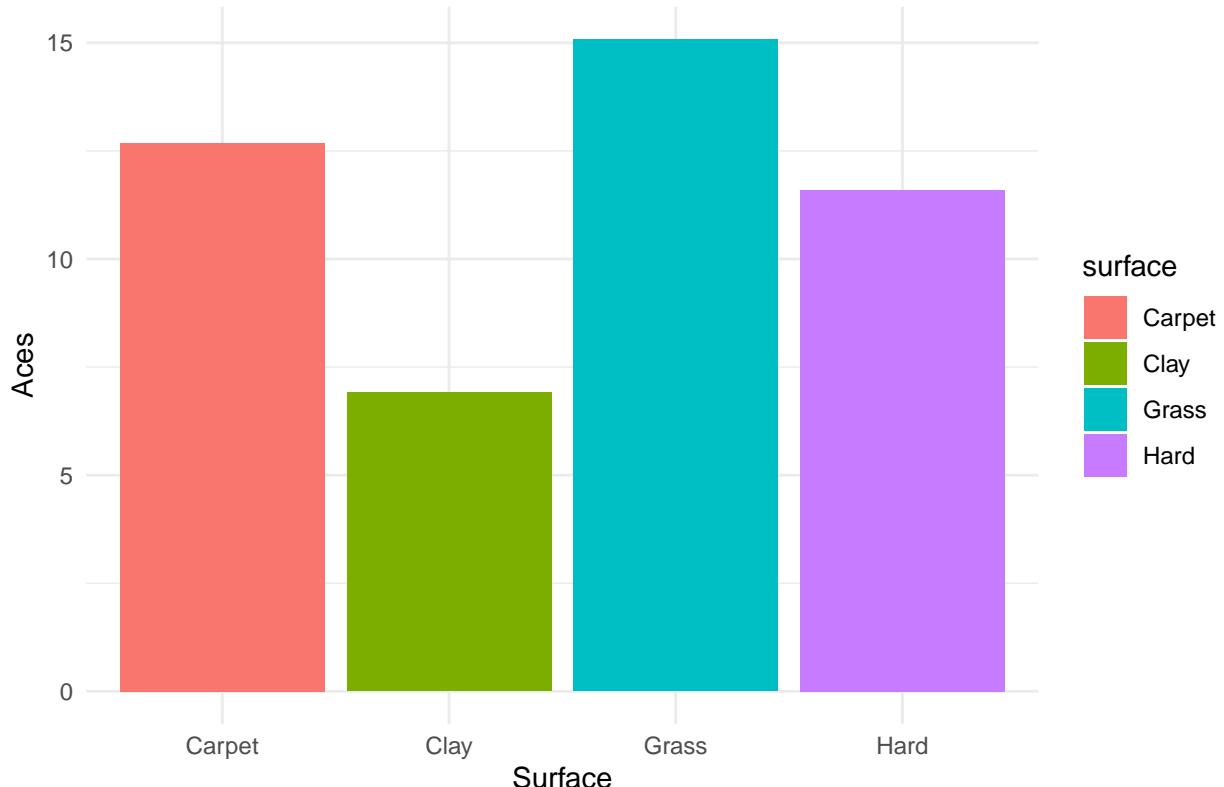
       x = "Surface",  

       y = "Aces") +  

  theme_minimal()

```

## Average Served Aces by Surface

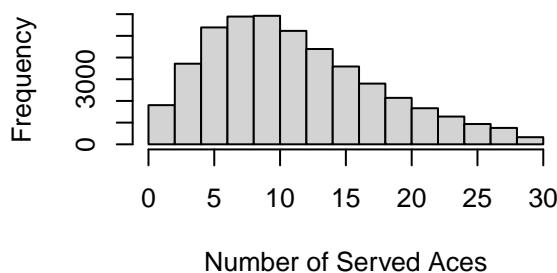
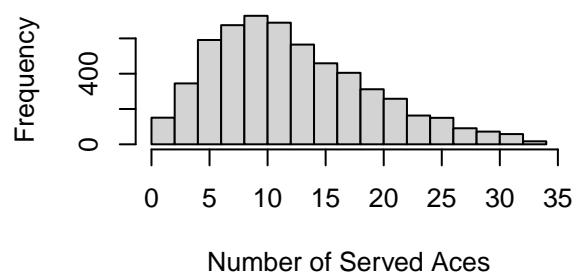
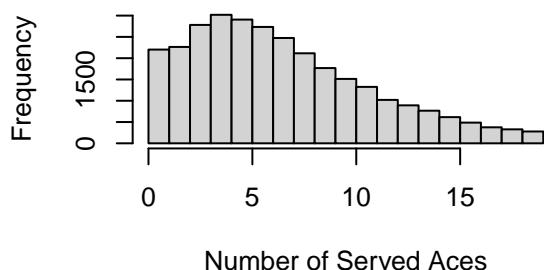
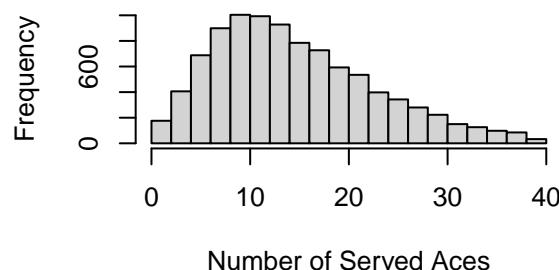


Iz barplota može se naslutiti da se zaista servira više asova na nekim površinama. Provjerimo je li to zaista tako koristeći ANOVA test. Nulta hipoteza jest da nema razlika u broju serviranih asova na različitim površinama. Kako bismo mogli provesti ANOVA statistički test moramo provjeriti jesu li zadovoljene pretpostavke o normalnosti distribucije i homogenosti varijanci za svaku od podloga.

1. provjera normalne distribucije

Prvo ćemo iscrtati histograme serviranih asova za svaku od površina:

```
par(mfrow = c(2, 2))
for (surface in unique(combined_data$surface)){
  hist(combined_data$aces[combined_data$surface==surface],
       main = paste("Histogram of served aces on" , surface),
       xlab = "Number of Served Aces",
       ylab = "Frequency")
}
```

**Histogram of served aces on Hard****Histogram of served aces on Carpet****Histogram of served aces on Clay****Histogram of served aces on Grass**

Iz histograma distribucije na svim podlogama čine se približno normalne osim na zemlji(Clay) koja ima nešto veće odstupanje. Pretpostaviti ćemo da je distribucija normalna s obzirom na to da su uzorci prilično veliki za svaku od grupe.

## 2. provjera homogenosti varijanci

Homogenost varijanci provjerava se Bartlettovim testom:

```
bartlett.test(combined_data$aces ~ combined_data$surface)
```

```
##  
##  Bartlett test of homogeneity of variances  
##  
## data: combined_data$aces by combined_data$surface  
## Bartlett's K-squared = 7740.8, df = 3, p-value < 2.2e-16
```

Bartlettov test ukazuje na to da varijance nisu homogene no zbog velike količine podataka provjerimo koliko se razlikuju veličine uzoraka za svaku od podloga i njihove varijance.

```
length(hard_aces)
```

```
## [1] 45839
```

```

length(carpet_aces)

## [1] 5729

length(clay_aces)

## [1] 29866

length(grass_aces)

## [1] 9470

var((combined_data$aces[combined_data$surface=='Carpet']))

## [1] 46.13074

var((combined_data$aces[combined_data$surface=='Hard']))

## [1] 40.87476

var((combined_data$aces[combined_data$surface=='Clay']))

## [1] 18.9474

var((combined_data$aces[combined_data$surface=='Grass']))

## [1] 66.02822

```

Svaka grupa ima preko 5000 podataka i postoje odstupanja u varijancama no ANOVA je robustna na manja odstupanja homogenosti uz dovoljno velik sample size stoga ćemo ipak provesti i taj test radi usporedbe. Nulta hipoteza jest da nema razlike u broju serviranih asova na različitim podlogama.

```

res <- aov(aces~surface, data=combined_data)
summary(res)

##                               Df  Sum Sq Mean Sq F value Pr(>F)
## surface                  3  665174   221725     6054 <2e-16 ***
## Residuals    90900 3328940        37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

P-vrijednost manja je od 0.05 stoga možemo odbaciti nullu hipotezu, odnosno postoji razlika u serviranim asovima na različitim podlogama.

Provedimo i Kruskal-Wallis koji je neparametarska alternativa ANOVA testu i koristi se kad nisu zadovoljene pretpostavke o normalnosti i homogenosti varijanci.

```

kruskal.test(aces~surface, data=combined_data)

##
##  Kruskal-Wallis rank sum test
##
## data:  aces by surface
## Kruskal-Wallis chi-squared = 15060, df = 3, p-value < 2.2e-16

```

Nakon provedenog testa također dobivamo p-vrijednost manju od 0.05 što znači da postoji razlika u broj serviranih asova u odnosu na podlogu. Intuitivno ovaj rezultat ima smisla jer loptica ne odskače jednako od svih podloga, npr. na zemljanoj podlozi loptica se sporije odbija i obično zadržava niže dok se na travi odbija brzo.

#### Zadatak 4. Kakva je veza između vrste terena i vjerojatnosti da će mečevi otici u peti set?

Postavljamo nultu hipotezu kako ne postoji statistički značajne veze između vrste terena i vjerojatnosti da će mečevi otici u peti set, a alternativna hipoteza sugerira prisutnost takve veze. Kako bismo testirali ovu hipotezu, koristit ćemo  $\chi^2$  test.

Najprije, provjeravamo pretpostavke kako bismo osigurali ispravnu primjenu testa.

Omogućena je nezavisnost podataka jer rezultat jednog teniskog meča ne utječe na rezultat drugog meča.

Također, osiguravamo da su nam podaci kategorički, klasifikacijom vrsta terena i ishoda mečeva u diskretne kategorije. Stvaramo kontingencijsku tablicu:

```

t4 <- all_matches[all_matches$best_of == 5, ]
t4$sets_played <- sapply(strsplit(as.character(t4$score), ""), function(x) sum(x == "-"))

contingency_table <- table(t4$surface, t4$sets_played == 5)
print(contingency_table)

```

```

##
##          FALSE  TRUE
## Carpet    700 179
## Clay     5550 1240
## Grass    3471  819
## Hard     9090 2054

```

Kontingencijskoj tablici dodajemo sume redaka i stupaca:

```

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

```

Još jedna pretpostavka testa je da očekivana frekvencija pojedinog razreda mora biti veća ili jednaka 5, stoga i to provjeravamo:

```

## Očekivane frekvencije za razred FALSE - Carpet : 715.7022
## Očekivane frekvencije za razred FALSE - Clay : 5528.576
## Očekivane frekvencije za razred FALSE - Grass : 3493.018
## Očekivane frekvencije za razred FALSE - Hard : 9073.704
## Očekivane frekvencije za razred TRUE - Carpet : 163.2978
## Očekivane frekvencije za razred TRUE - Clay : 1261.424
## Očekivane frekvencije za razred TRUE - Grass : 796.9822
## Očekivane frekvencije za razred TRUE - Hard : 2070.296

```

Sve očekivane pretpostavke su zadovoljene, nastavljamo sa  $\chi^2$  testom.

```

chi_square_result <- chisq.test(contingency_table)
print(chi_square_result)

```

```

##
## Pearson's Chi-squared test
##
## data: contingency_table
## X-squared = 3.2059, df = 3, p-value = 0.361

```

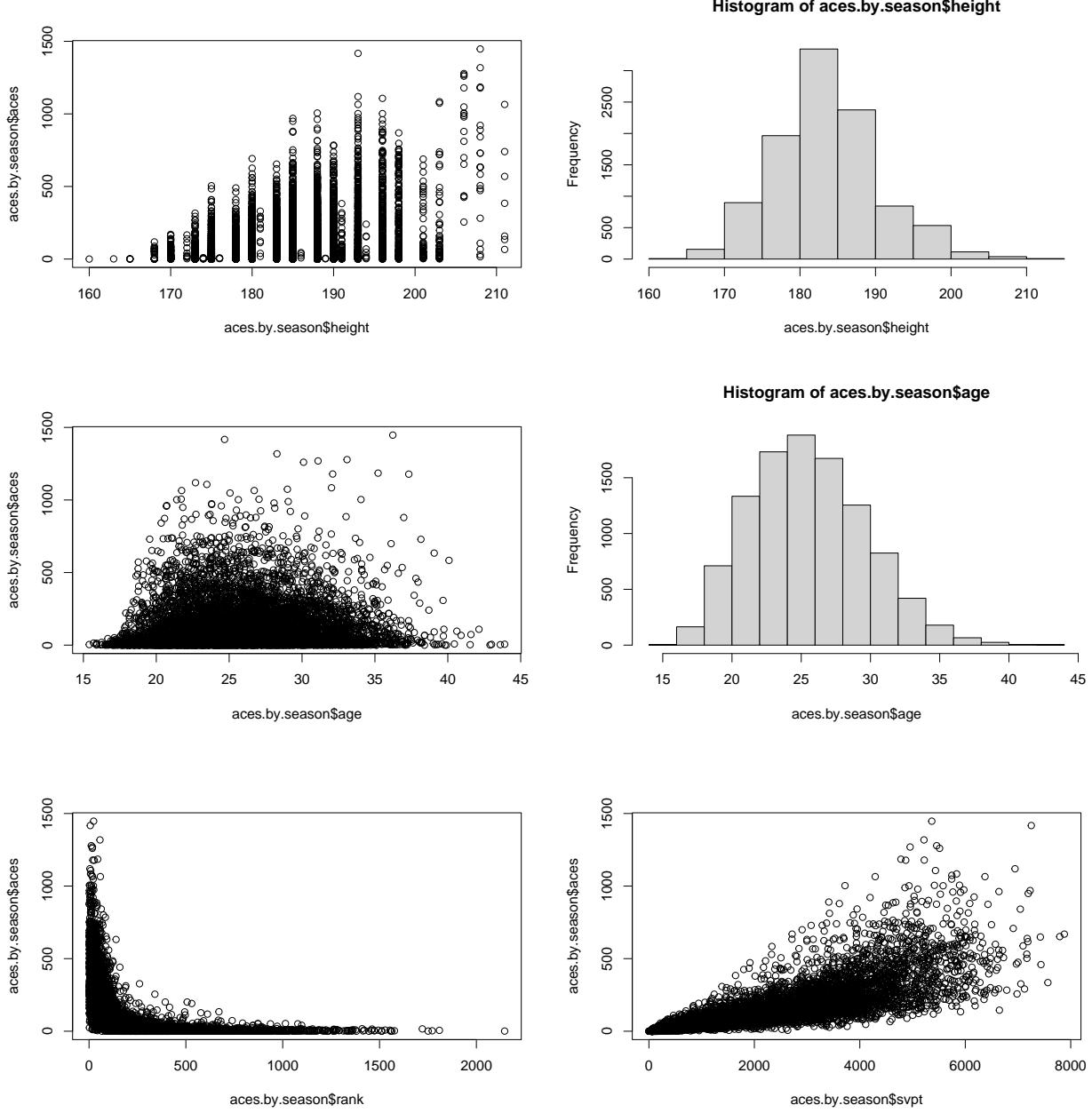
Rezultati  $\chi^2$  testa ukazuju na to da ne postoji statistički značajna veza između vrste terena na kojem se održavaju teniski mečevi i vjerojatnosti da će mečevi otići u peti set (p-vrijednost = 0.361). S obzirom na p-vrijednost veću od 0.05, ne odbacujemo nultu hipotezu.

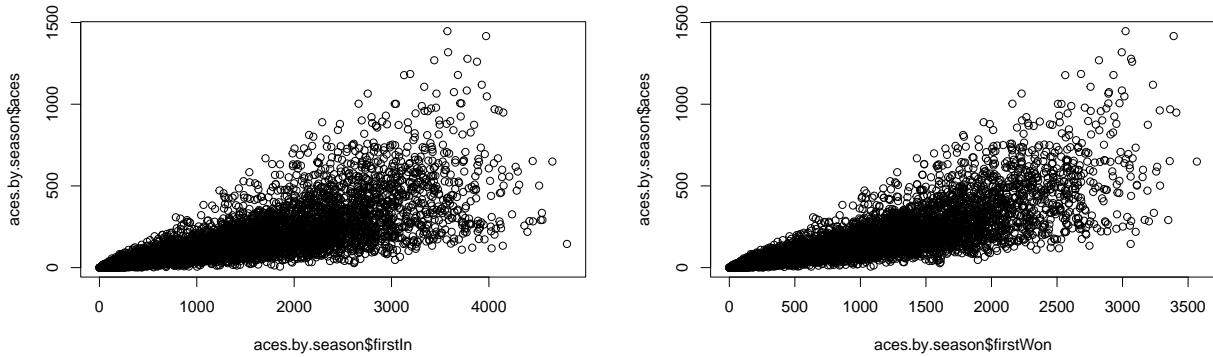
**Zadatak 5.** Možemo li procijeniti broj asova koje će igrač odservirati u tekućoj godini (zadnjoj dostupnoj sezoni) na temelju njegovih rezultata iz prethodnih sezona?

Kako bismo predviđeli broj asova za nekog igrača prvo moramo sagledati koje bi sve značajke mogle utjecati na broj odserviranih asova:

- Visina igrača
- Starost igrača
- Rank igrača
- Broj servi u sezoni
- Broj uspješnih prvih servi
- Broj osvojenih prvih servi

S obzirom da se podatci odnose pojedinačne mečeve, potrebno ih je agregirati za svakog igrača na razini sezone. Prilikom aggregacije uzet ćemo srednju vrijednost za starost i rank igrača dok ćemo ostale statistike o servi sumirati. Kako bi dobili dojam o odnosu ovih varijabli i broja asova prikazat ćemo ih pojedinačno grafički pomoću scatter plota.





Starost i visina igrača naizgled imaju nelinearni odnos prema broju aseva. Međutim, ako ispišemo histograme starosti i visina vidimo da je taj odnos proizašao iz frekvencije pojavljivanja tih značajki. Ovo ne vrijedi u potpunosti za visine igrača, ali pokazalo se da visina nije znatno utjecala na točnost modela. Zbog toga te dvije značajke nećemo dalje uzimati u obzir.

Rank igrača ima snažan obrnuto proporcionalan odnos s brojem asova što intuitivno ima smisla. Bolji igrači će imati veći broj asova. Važno je uočiti da ovaj odnos nije linearan i te ćemo morati uzeti inverz ranga prilikom regresije.

Ukupan broj servi, prvih uspješnih i prvih osvojenih servi ima linearan odnos prema broju aseva. Također uočavamo da su ove tri značajke naizgled jako korelirane. To ćemo morati ispitati i po potrebi uzeti samo jedno od ovih značajki za našu regresiju.

Valja spomenuti da su u obzir još bili uzeti podatci o servama, ali u postotcima. Oni su imali sličan problem kao i starost i visina. Scatter plot u odnosu na broj asova je imao oblik normalne distribucije. Dominantna ruka također nije uzeta u obzir jer prednost koju ona donosi ovisi o dominantnoj ruci suparnika, a pošto su podatci agregirani prednosti i hendikepi dominantne ruke će se poništiti.

Provjerit ćemo naše pretpostavke koristeći model jednostavne regresije, po jedan za svaku značajku. Broj asova će biti zavisna varijabla

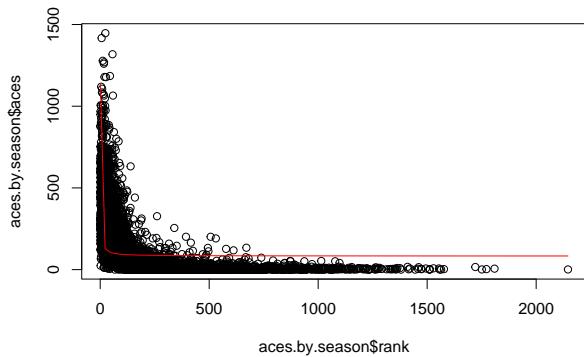
```
fit.invRank = lm(aces ~ invRank, data=aces.by.season)

fit.svpt = lm(aces ~ svpt, data=aces.by.season)

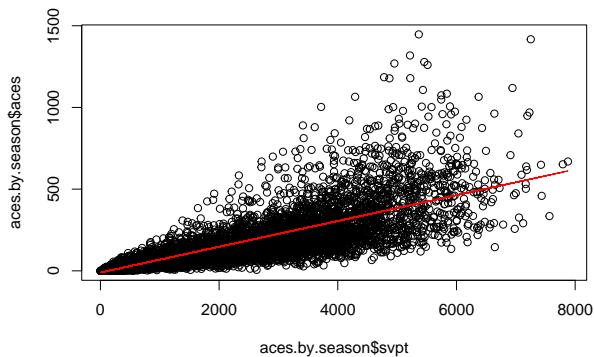
fit.firstIn = lm(aces ~ firstIn, data=aces.by.season)

fit.firstWon = lm(aces ~ firstWon, data=aces.by.season)

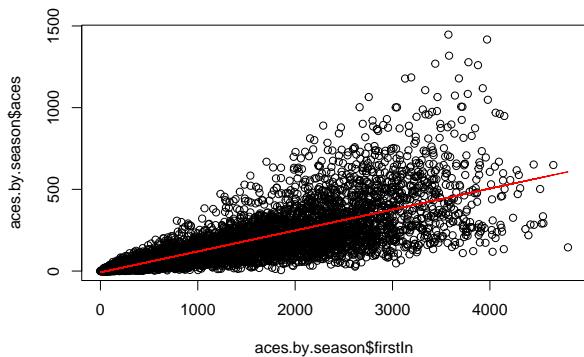
f = function(x, coeffs)
  return(coeffs[[1]] + coeffs[[2]] * (1 / x))
plot(aces.by.season$rank, aces.by.season$aces)
curve(f(x, fit.invRank$coefficients), add = TRUE, col = "red")
```



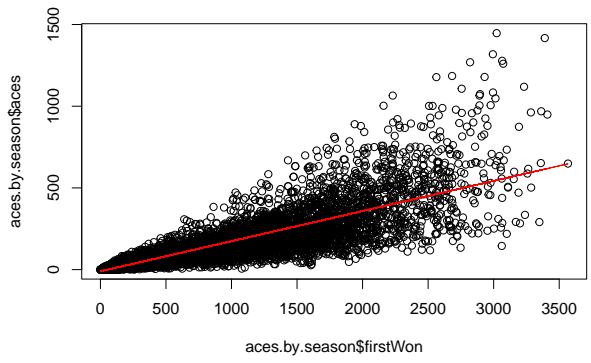
```
plot(aces.by.season$svpt, aces.by.season$aces)
lines(aces.by.season$svpt , fit.svpt$fitted.values, col='red')
```



```
plot(aces.by.season$firstIn, aces.by.season$aces)
lines(aces.by.season$firstIn, fit.firstIn$fitted.values, col='red')
```

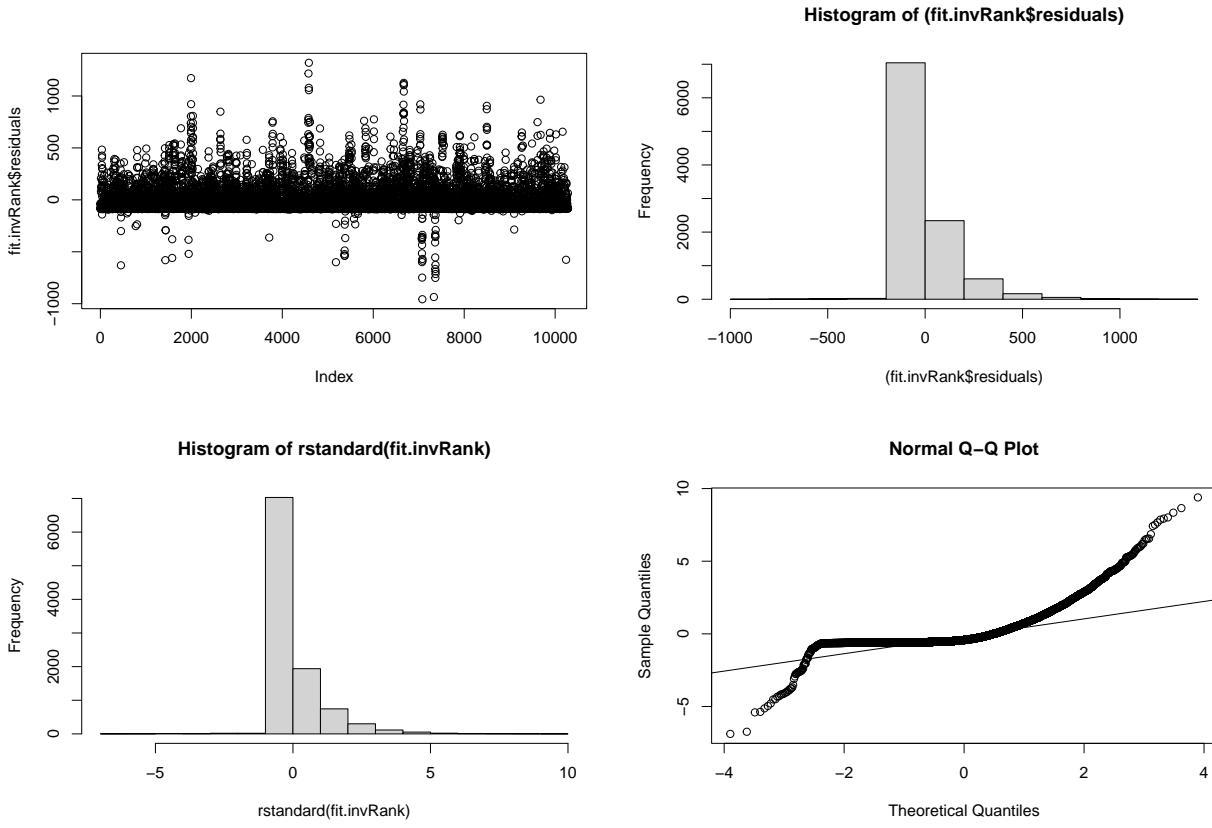


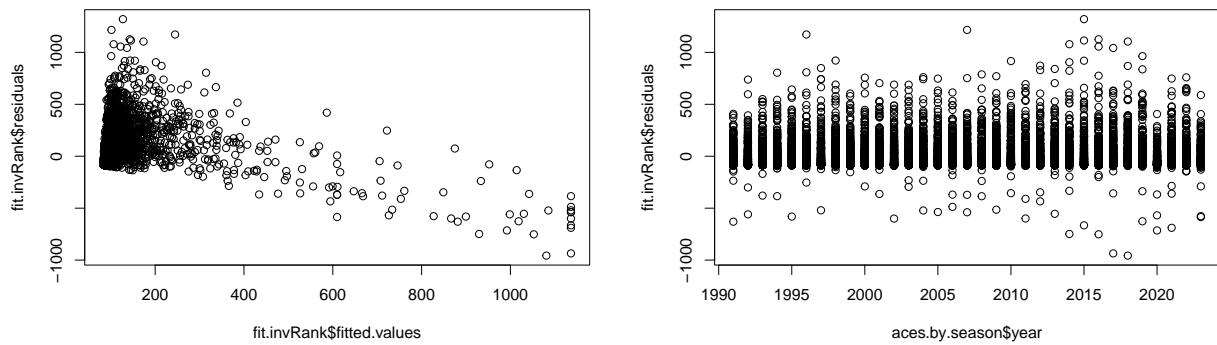
```
plot(aces.by.season$firstWon, aces.by.season$aces)
lines(aces.by.season$firstWon, fit.firstWon$fitted.values, col='red')
```



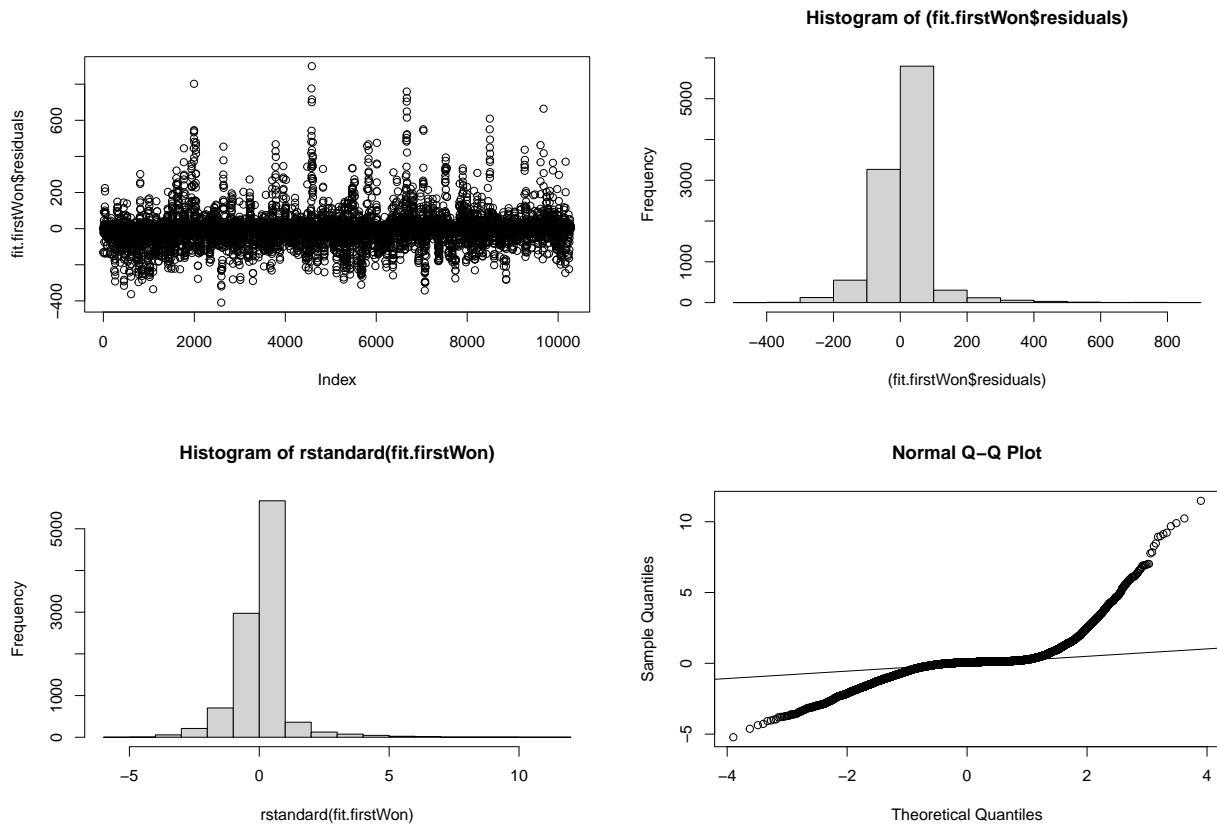
### Normalnost reziduala i homogenost varijance

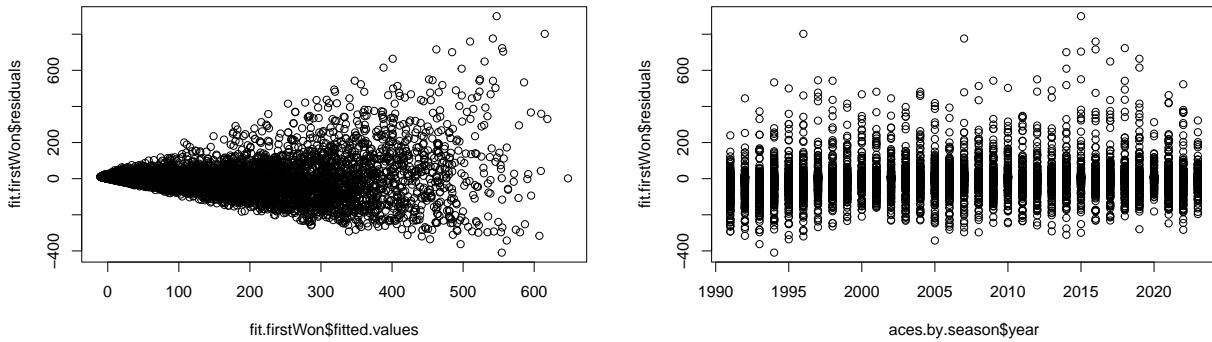
Normalnost reziduala provjerit ćemo grafički pomoću histograma i qq plota te statistički pomoću Lillieforsovog testa.





```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: rstandard(fit.invRank)  
## D = 0.25099, p-value < 2.2e-16
```





```
## 
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.firstWon)
## D = 0.23869, p-value < 2.2e-16
```

Iz histograma i qq plota možemo zaključiti da reziduali ukupan broj servi, uspješnih prvih servi i osvojenih prvih servi značajno odstupaju od normalne distribucije dok za inverz ranka reziduali više nalikuju eksponencijalnoj distribuciji nego normalnoj.

Također treba prijetiti da za ukupan broj servi, uspješnih prvih servi i osvojenih prvih servi u ovisnosti o predviđanjima reziduali pokazuju heterogenost varijance, šire se povećanjem  $\hat{y}$ . Međutim u ovisnosti o godinama reziduali pokazuju homogenu varijancu.

### Korelacija značajki

Unatoč neobećavajućim rezultatima ispitivanja normalnosti i homogenosti varijance reziduala pokušat ćemo napraviti regresiju nad skupom relevantnih značajki. Kako bi to napravili prvo moramo odrediti koje su značajke korelirane.

```
cor(cbind(aces.by.season$aces, aces.by.season$invRank, aces.by.season$svpt, aces.by.season$firstIn, ace
```

```
##          [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.4164105 0.8363657 0.8230129 0.8617121
## [2,] 0.4164105 1.0000000 0.4393532 0.4467202 0.4660817
## [3,] 0.8363657 0.4393532 1.0000000 0.9948097 0.9942303
## [4,] 0.8230129 0.4467202 0.9948097 1.0000000 0.9965806
## [5,] 0.8617121 0.4660817 0.9942303 0.9965806 1.0000000
```

Kao što smo i pretpostavili ranije, ukupan broj servi, uspješnih prvih servi i osvojenih prvih servi pokazuju snažnu pozitivnu korelaciju. Stoga ćemo od te tri značajke uzeti samo ukupan broj osvojenih prvih servi jer ima najveću korelaciju s brojem asova od te tri.

Inverzni rank također pokazuje značajnu korelaciju s ostatkom značajki.

```
fit.multi = lm(aces ~ invRank + firstWon, data=aces.by.season)
summary(fit.multi)
```

```

## 
## Call:
## lm(formula = aces ~ invRank + firstWon, data = aces.by.season)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -406.35   -15.93    5.03   11.35  903.20 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -10.05495   1.01610 -9.896 < 2e-16 ***
## invRank      47.76983  14.30098  3.340  0.00084 ***
## firstWon     0.18255   0.00121 150.863 < 2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 78.41 on 10279 degrees of freedom
## Multiple R-squared:  0.7428, Adjusted R-squared:  0.7428 
## F-statistic: 1.485e+04 on 2 and 10279 DF,  p-value: < 2.2e-16

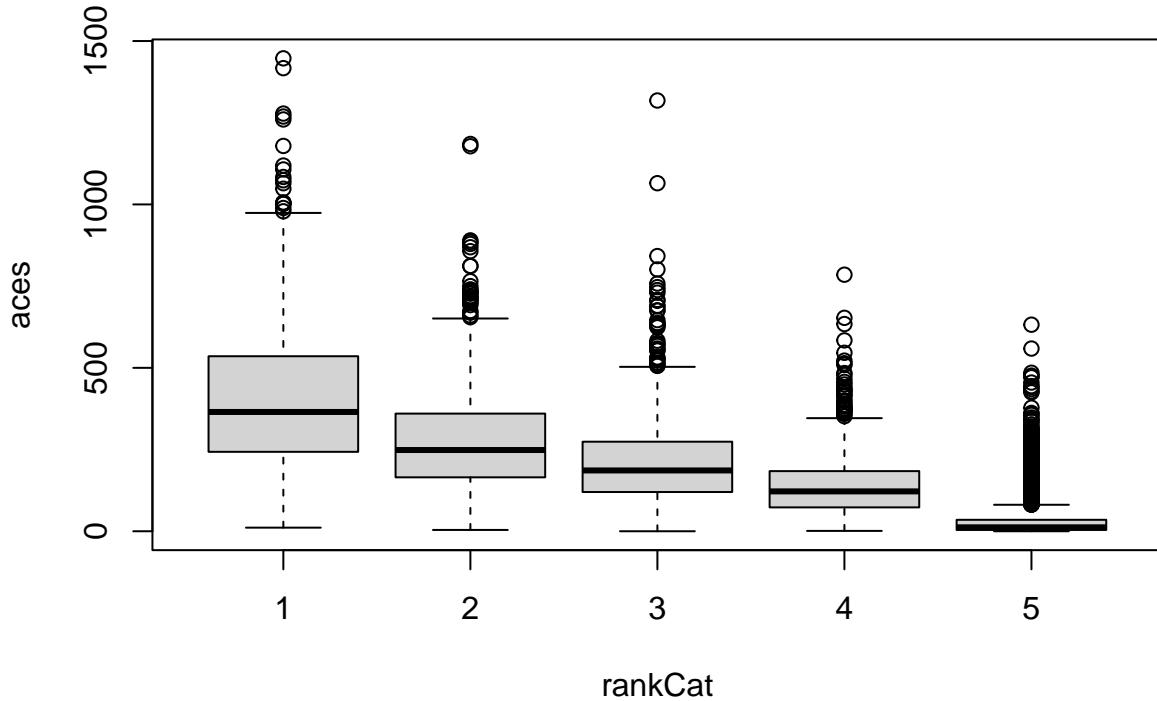
```

Iz modela možemo vidjeti da rank igrača jako slabo utječe na objašnjavanje varijance podataka. Stoga ćemo pokušati pretvoriti rank igrača u kategoričku varijablu na način da svrtamo igrače u sljedeće kategorije: - Rank 1 - 20 - Rank 21 - 50 - Rank 50 - 100 - Rank 100+

```

aces.by.season <- aces.by.season %>%
  mutate(rankCat = case_when(
    rank < 25 ~ 1,
    rank < 50 ~ 2,
    rank < 75 ~ 3,
    rank < 100 ~ 4,
    TRUE ~ 5
  ))
  boxplot(aces~rankCat, data=aces.by.season)

```



```

aces.by.season.d = dummy_cols(aces.by.season, select_columns = 'rankCat')

fit.multi.d = lm(aces ~ firstWon + rankCat_1 + rankCat_2 + rankCat_3 + rankCat_4, data = aces.by.season)
summary(fit.multi.d)

##
## Call:
## lm(formula = aces ~ firstWon + rankCat_1 + rankCat_2 + rankCat_3 +
##     rankCat_4, data = aces.by.season.d)
##
## Residuals:
##      Min      1Q      Median      3Q      Max 
## -446.70   -15.48     2.97     9.77   862.75 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -7.875695  1.035250 -7.608 3.04e-14 ***
## firstWon     0.191089  0.002302  83.018 < 2e-16 ***
## rankCat_1    14.459791  5.356223   2.700  0.00695 **  
## rankCat_2   -26.230447  4.436337  -5.913 3.47e-09 ***
## rankCat_3   -31.537385  3.784755  -8.333 < 2e-16 *** 
## rankCat_4   -29.510775  3.162099  -9.333 < 2e-16 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

```

## Residual standard error: 77.37 on 10276 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7495
## F-statistic:  6154 on 5 and 10276 DF,  p-value: < 2.2e-16

```

Ova metoda nažalost također nije dala nikakvo značajno poboljšanje. Međutim postoji još jedan način da poboljšamo rezultate a to je uvođenje nove zakašnjele varijable. Naime pretpostavljamo da će broj aseva za nekog igrača biti jako povezan s brojem aseva koje je ostavio u prošloj sezoni. Zato ćemo uvesti novu značajku prevAces u naš model.

```

aces.by.season.d$prevAces = c(NA, aces.by.season.d$aces[1:length(aces.by.season.d$aces)-1])

fit.multi.timelag = lm(aces ~ firstWon + rankCat_1 + rankCat_2 + rankCat_3 + rankCat_4 + prevAces, data = aces.by.season.d)
summary(fit.multi.timelag)

```

```

##
## Call:
## lm(formula = aces ~ firstWon + rankCat_1 + rankCat_2 + rankCat_3 +
##      rankCat_4 + prevAces, data = aces.by.season.d)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -378.24 -23.90   7.86  18.12  749.22 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.077e+01  8.288e-01 -25.06  <2e-16 ***
## firstWon     1.674e-01  1.832e-03   91.41  <2e-16 ***
## rankCat_1   -1.097e+02  4.484e+00  -24.47  <2e-16 ***
## rankCat_2   -9.625e+01  3.592e+00  -26.80  <2e-16 ***
## rankCat_3   -6.649e+01  3.004e+00  -22.13  <2e-16 ***
## rankCat_4   -4.641e+01  2.492e+00  -18.62  <2e-16 ***
## prevAces     4.498e-01  5.626e-03   79.94  <2e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.76 on 10274 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.8457, Adjusted R-squared:  0.8456
## F-statistic:  9382 on 6 and 10274 DF,  p-value: < 2.2e-16

```

Ovime smo dobili osjetno bolji model. Za kraj ćemo još primjeniti model za nekoliko igrača da vidimo kako radi.

```

player.samples <- aces.by.season.d[sample(nrow(aces.by.season.d), 10),]

name <- player.samples$name
year <- player.samples$year
target_aces <- player.samples$aces
prediction_aces <- predict(fit.multi.timelag, newdata = player.samples)

print(data.frame(name, year, target_aces, prediction_aces))

##          name year target_aces prediction_aces

```

## 1	Jim Courier	1991	379	305.1266393
## 2	Tomas Barrios Vera	2021	12	2.6588568
## 3	Feliciano Lopez	2001	39	18.2418452
## 4	Juan Antonio Marin	2004	4	-10.2423839
## 5	Felipe Rivera	1993	8	-7.2288229
## 6	Arnaud Clement	2012	36	52.0653560
## 7	Kamil Majchrzak	2018	15	-0.7093369
## 8	Dominic Thiem	2020	244	277.3161710
## 9	Thomas Muster	2011	6	-11.9888850
## 10	Frederico Ferreira Silva	2016	2	-17.6910632

Model na žalost ne radi sjajno, ali to je i za očekivati jer su narušeni uvjeti normalnosti i homogenosti varijanci reziduala. Za ovaj problem potreban je složeniji model.