Statistika za NBA

Učitavanje podataka i biblioteka.

```
seasons.stats = read.csv(file = "../data/Seasons_Stats.csv", header = TRUE)
player.data = read.csv("../data/player_data.csv")
players = read.csv('../data/Players.csv')

library(tidyverse) # sluzi za grupiranja, joinove, filtriranja
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0       v purrr   0.3.1
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

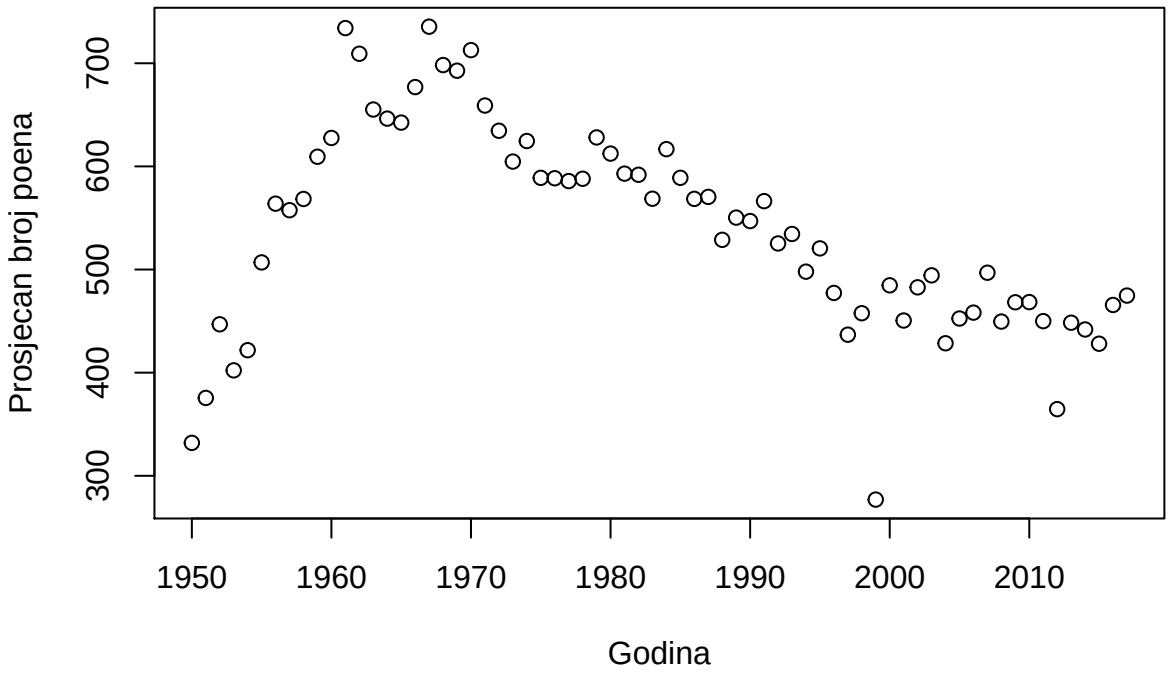
```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Deskriptivna statistika za NBA

Ovisnost prosječnog 3PA (pokušaji trica) i 2PA (pokušaji dvica) o godinama te jedan o drugom

Filozofija igre je 60-ih i 70-ih godina bila takva da je tempo igre je bio puno brzi, odnosno bilo je ukupno vise posjeda lopte. Valja napomenuti zanimljivu činjenicu da je linija za tricu uvedena tek 80-ih zbog čega su podaci za iste navedeni tek od 80-ih pa nadalje. Skok u pokušajima trica je vidljiv 1996. iz razloga što je te godine linija za tricu pomaknuta bliže košu za skoro 60cm, ali je 1998. godine crta vraćena na staru poziciju. Razlog zbog kojeg je crta za tricu bila pomaknuta bliže košu je taj što su htjeli povećati broj bodova na utakmici.

```
# prosječni broj poena (PTS) po godini
seasons.stats %>% group_by(Year) %>% summarize(avg.PTS = mean(PTS)) -> average.points
plot(average.points, xlab = "Godina", ylab = "Prosjecan broj poena")
```



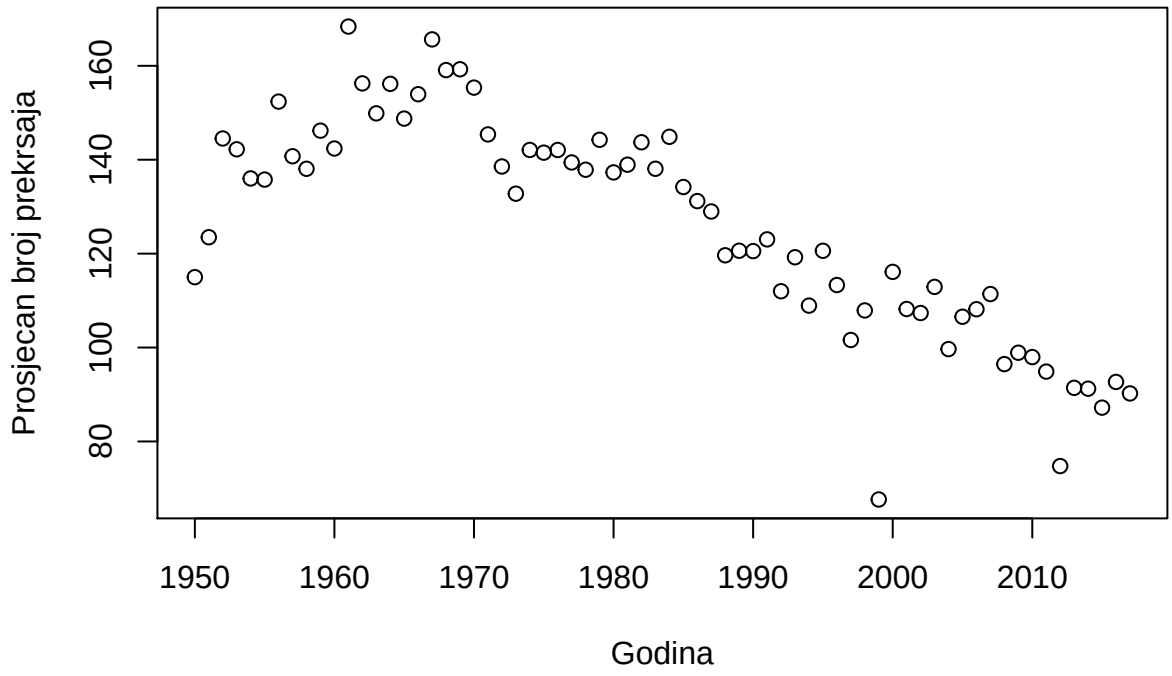

```
# projekcija tablice na godinu, 3PA, 2PA i G (broj utakmica)
season.stats.points = seasons.stats[c('Year', 'X3PA', 'X2PA', 'G')]

# prosječan broj 3PA i 2PA po igri
season.stats.points$X3PAPerGame = season.stats.points$X3PA / season.stats.points$G
season.stats.points$X2PAPerGame = season.stats.points$X2PA / season.stats.points$G

# brisanje nepotpunih redaka
season.stats.points.complete = season.stats.points[complete.cases(season.stats.points),]

# prosječan broj 3PA i 2PA po godini po utakmici
season.stats.points.complete %>% group_by(Year) %>% summarize(ThreePointAttempts = mean(X3PAPerGame)) ->
season.stats.points.complete %>% group_by(Year) %>% summarize(TwoPointAttempts = mean(X2PAPerGame)) -> t

# grafički prikaz 3PA po godinama
plot(
  three.point.attempts$Year,
  three.point.attempts$ThreePointAttempts,
  xlab = "Godina",
  ylab = "3PA po utakmici",
  col = 'blue'
)
```


```
# grafički prikaz 2PA po godinama
plot(
  two.point.attempts$Year,
  two.point.attempts$TwoPointAttempts,
  xlab = 'Godina',
  ylab = '2PA po utakmici',
  col = 'red'
)
```

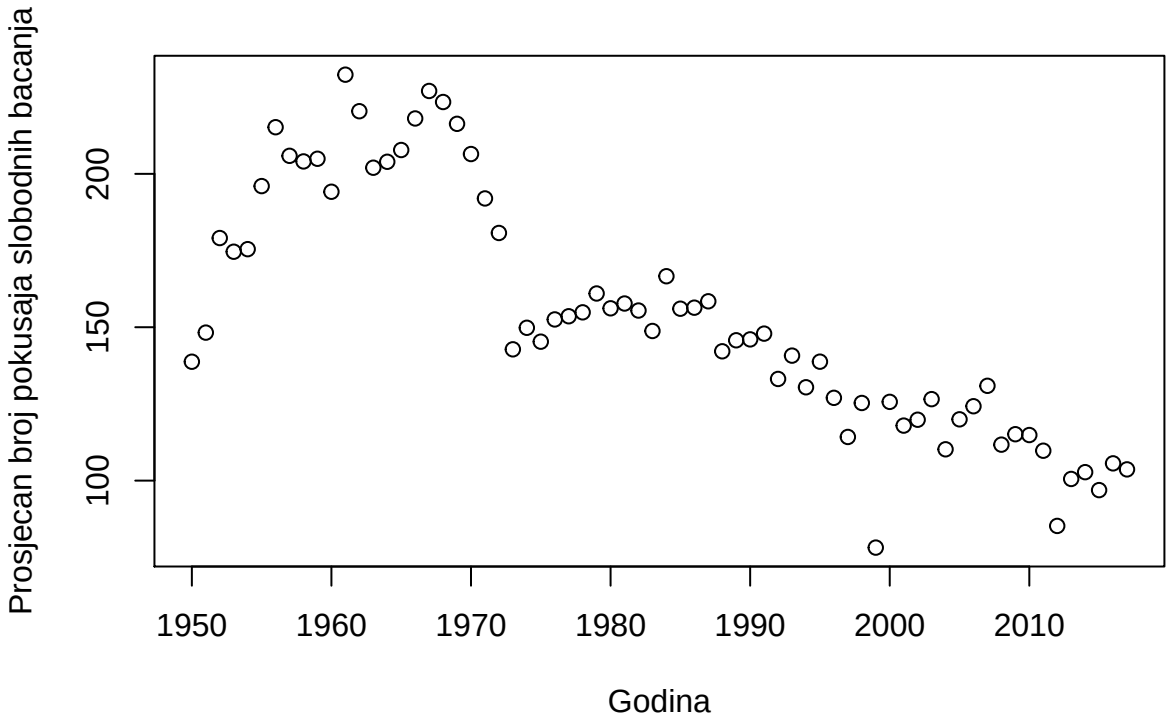
Ovisnost prosječnog broja prekršaja i pokušaja slobodnih bacanja po godinama

NBA je entertainment industrija i gledanost im je jako bitan aspekt s obzirom da gledanost nosi profit. Foulovi usporavaju igru i u korist tome govori novo uvedeno pravilo - da se prekršaji u kontri kažnjavaju nesportskim greškama (2 bacanja i posjed lopte), što obeshrabruje obrambene igrače da rade foulove u kontri. Namjera tog pravila je da se kontre maksimalno dopuštaju s obzirom da najčešće rezultiraju atraktivnom završnicom što pogoduje gledanosti.

```
# prosjecan broj prekrsaja po godini
seasons.stats %>% group_by(Year) %>% summarize(PFAvg = mean(PF)) -> average.fouls
# graficki prikaz prosjecnog broja prekrsaja po godini
plot(
  average.fouls,
  xlab = "Godina",
  ylab = 'Prosjecan broj prekrsaja'
)
```



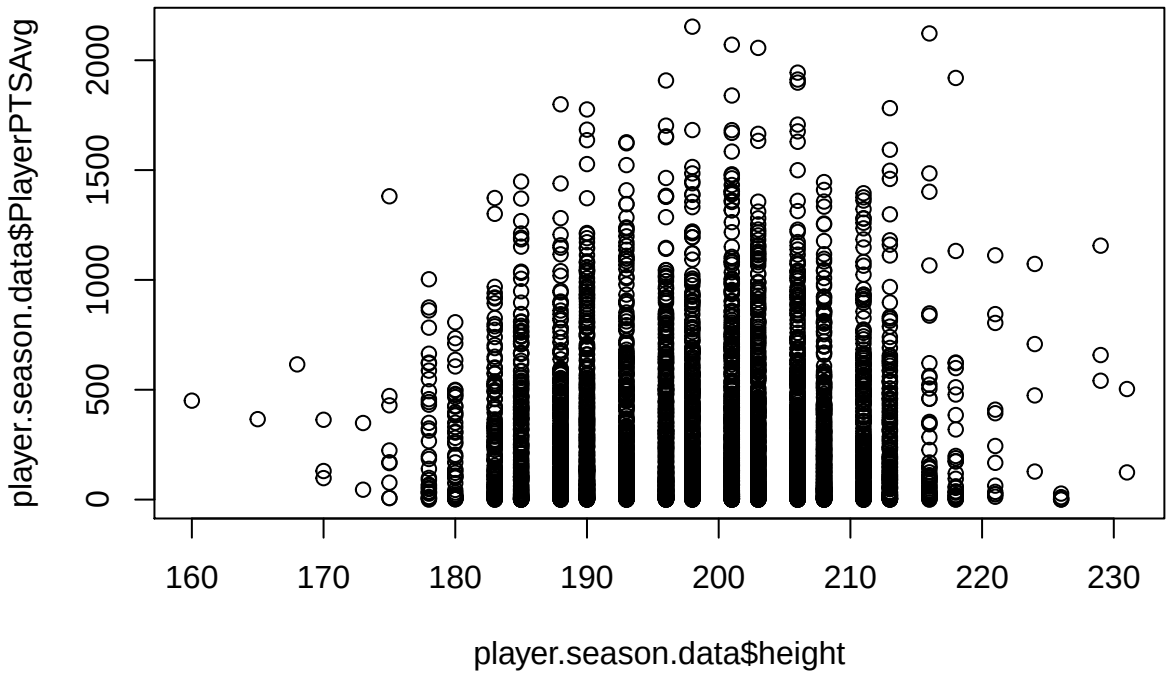

```
# prosjecan broj FTA (pokusaja slobodnih udaraca)
seasons.stats %>% group_by(Year) %>% summarize(FTAAvg = mean(FTA)) -> average.free.throw.attempts
# graficki prikaz prosjecnog broja slobodnih bacanja po godini
plot(
  average.free.throw.attempts,
  xlab = "Godina",
  ylab = 'Prosjecan broj pokusaja slobodnih bacanja'
)
```



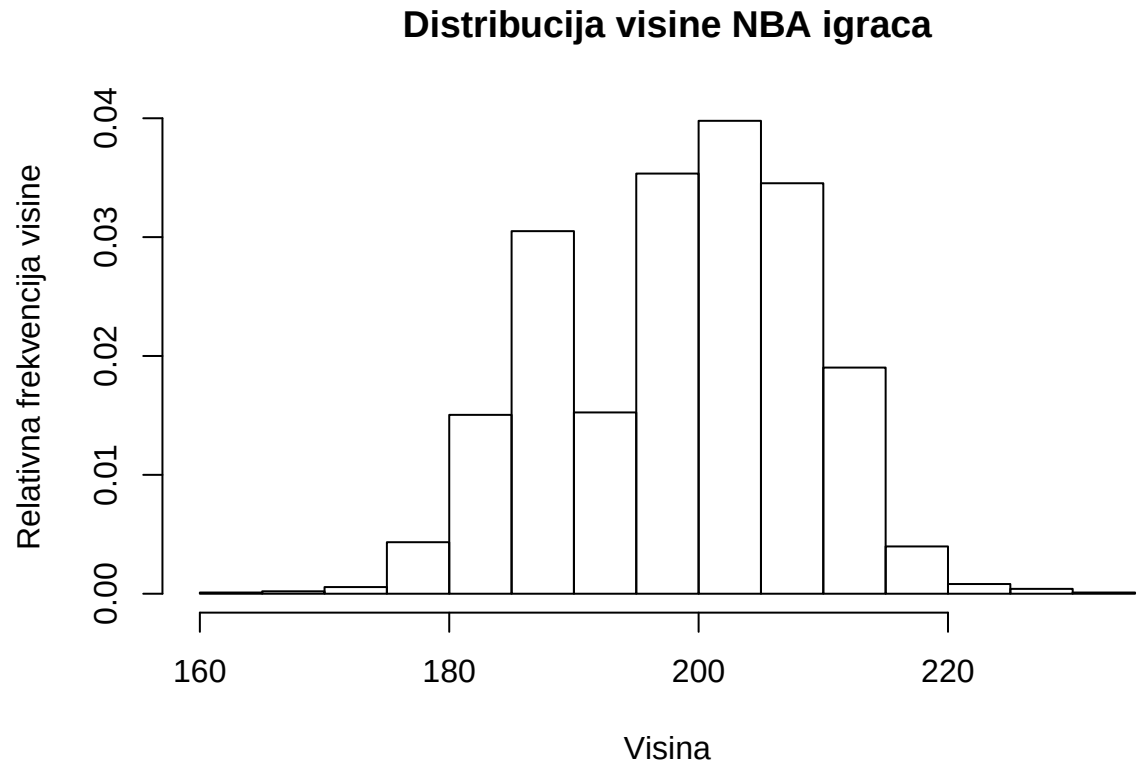
Visine i mase igrača u NBA ligi

Htjeli smo pokazati da visina utječe na prosječan broj zabijenih poena i to je prikazano na grafu u nastavku. Nažalost iz grafa se ne vidi nikakva korelacija između visine i zabijenih poena.

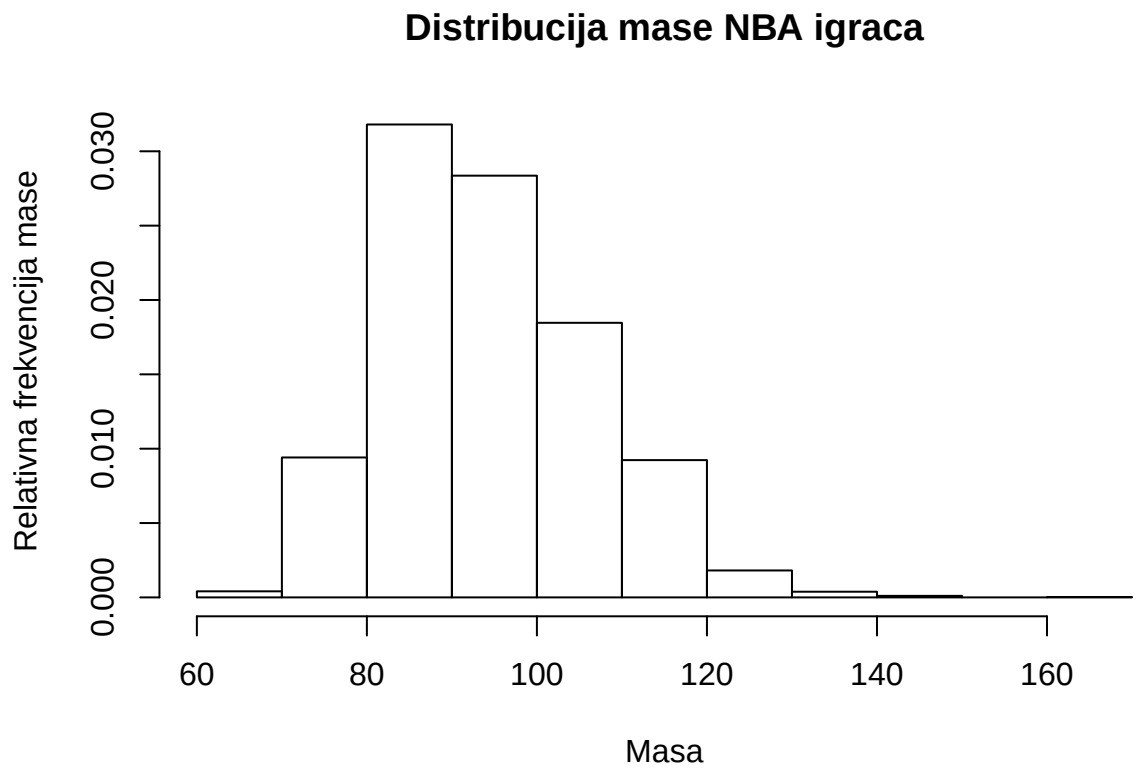
```
# prosjecan broj poena po igracu
seasons.stats %>% group_by(Player) %>% summarize(PlayerPTSAvg = mean(PTS)) -> playerPTSAverage

playerPTSAverage$Player = as.character(playerPTSAverage$Player)
players$Player = as.character(players$Player)

# spajanje prosjecnog broja poena po igracu s informacijama o igracu
player.season.data = inner_join(playerPTSAverage, players, by = c("Player" = "Player"))

# graficki prikaz ovisnosti prosjecnog broja zabijenih poena o visini
plot(player.season.data$height, player.season.data$PlayerPTSAvg)
```




```r
# graficki prikazi distribucije visine i mase
hist(
  player.season.data$height,
  main = "Distribucija visine NBA igraca",
  xlab = "Visina",
  ylab = "Relativna frekvencija visine",
  probability = TRUE
)
```



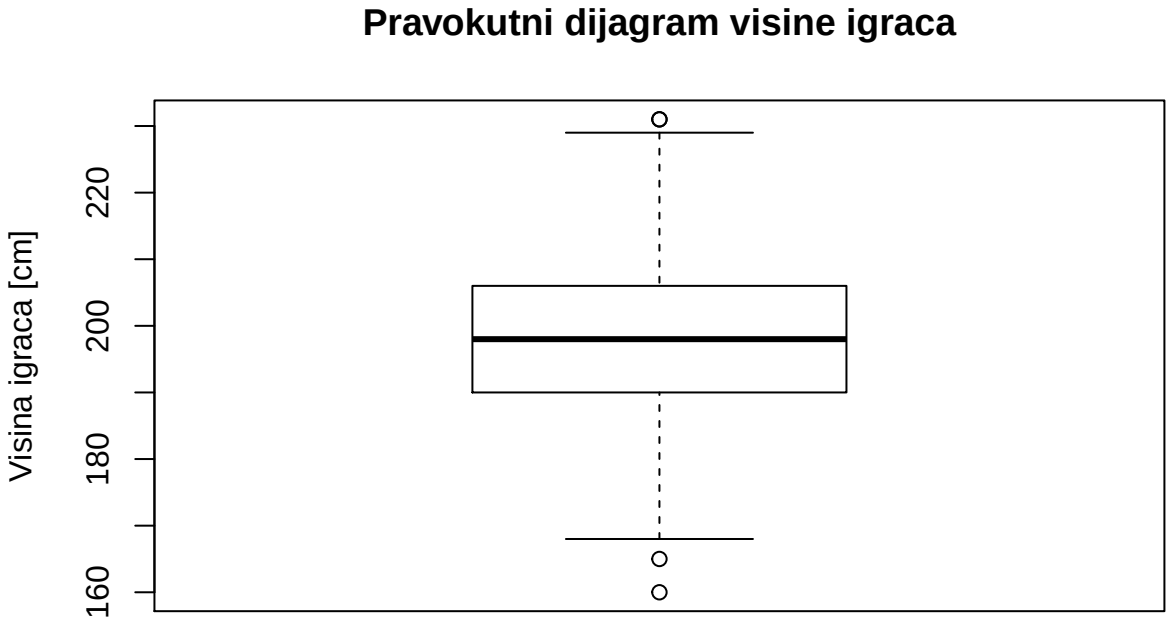
```r
hist(
  player.season.data$weight,
  main = "Distribucija mase NBA igraca",
  xlab = "Masa",
  ylab = "Relativna frekvencija mase",
  probability = TRUE
)
```




```r
player.heights.complete = player.season.data$height[complete.cases(player.season.data$height)]

summary(player.heights.complete)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   160.0   190.0   198.0   198.7   206.0   231.0
```

```r
boxplot(
  player.heights.complete,
  main = 'Pravokutni dijagram visine igraca',
  ylab = 'Visina igraca [cm]'
)
```



Visine igrača po pozicijama

Htjeli smo pokazati da se visine igrača po pozicijama znatno razlikuju. To smo učinili ANOVA-om koja je dala izuzetno malu p-vrijednost što znači da možemo odbaciti hipotezu H0, tj. ne možemo tvrditi da se visine igrača po pozicijama ne razlikuju.

Dobiveni su podaci kao i ocekivani, pozicije G (Guards) ne zahtijevaju da su igraci visoki, dok za pozicije F (Forward) i C (Center) se zahtijeva visina zbog uloge koje imaju u igri.

Braniče (Guards - G) tradicionalno čine razigravač (poznatiji pod imenom playmaker) i bek šuter. Neke od njihovih osobina su brzina, odlična kontrola lopte, spretnost i mehanička potkovanost koja najčešće jako korelira s dobrim šuterskim sposobnostima. S obzirom na spomenute osobine braniči su najčešće ispodprosječne visine.

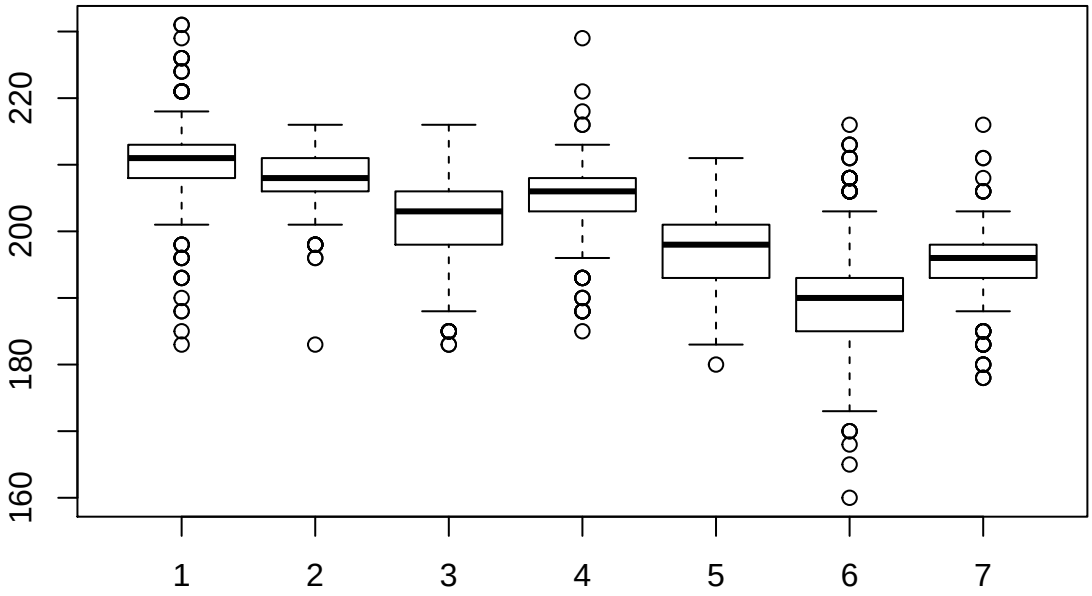
Krila (Forwards - F) tradicionalno čine nisko i visoko krilo. Od niskog krila se očekuje svestranost i u napadu i u obrani, što podrazumijeva relativno nadprosječnu visinu s obzirom da su im obrambene zadaće izrazito šarolike (očekuje se switchableness kroz većinu pozicija u obrani od blokova), a napadački moraju moći završiti ulaze pored takozvanih rim protectora (braniča obruča). Visoka krila uz centre najčešće sačinjavaju "tornjeve u reketu" koji odrađuju funkcije visokih igrača (postavljanje blokova, blokiranje ulaza i sl.), no u modernijoj košarci se od njih podrazumijeva da mogu širiti teren, odnosno očekuje se nadprosječan postotak ubačaja za tri poena. Često se i koriste kao "small-ball petice" kako bi se ubrzala igra i maksimizirala obrambena fleksibilnost dok se pritom relativno zadržala visina spomenute pozicije.

Centri (Centers - C) su zaduženi za "prljavi posao" u reketu (obrambeni i napadački skok, rim protection, postavljanje blokova i sl.) u čemu im visina maksimalno pogoduje, no povlači i tromost koja nije toliki problem s obzirom da su većinom usidreni u reketu, no to ih čini dosta ranjivima prilikom preuzimanja u obrani.

Valja napomenuti kako se u zadnje vrijeme sve više gubi tradicionalna podjela po pet pozicija, a sve se više gura tzv. "positionless basketball" gdje svaki trener kroji petorke na način koji mu odgovara protiv određenog protivnika te za željeni stil igre. Jedan od prijedloga modernih pozicija čine "ball-handleri" (koji najviše odgovaraju braničima), krila (koji najviše odgovaraju niskim krilima) i visoki (koji najviše odgovaraju visokim krilima i centrima).

```
players$Player = as.character(players$Player)
player.data$name = as.character(player.data$name)
players.joined = inner_join(player.data, players, by=c("name" = "Player"))
```

```r
# usporedni pravokutni dijagrami za distribuciju visine po pozicijama
boxplot(
  players.joined[players.joined$position == 'C',]$height.y,
  players.joined[players.joined$position == 'C-F',]$height.y,
  players.joined[players.joined$position == 'F',]$height.y,
  players.joined[players.joined$position == 'F-C',]$height.y,
  players.joined[players.joined$position == 'F-G',]$height.y,
  players.joined[players.joined$position == 'G',]$height.y,
  players.joined[players.joined$position == 'G-F',]$height.y
)
```




```r
# ANOVA - razlika izmedu visina po pozicijama
players.joined %>% group_by(position) %>% summarize(averageHeight = mean(height.y)) -> averageHeightFor

model = lm(players.joined$height.y ~ players.joined$position)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: players.joined$height.y
##                            Df Sum Sq Mean Sq F value    Pr(>F)
## players.joined$position     7 197300 28185.7  846.54 < 2.2e-16 ***
## Residuals                3806 126721    33.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Usporedba igrača s Duke University-ja u odnosu na druge fakultete u SAD-u

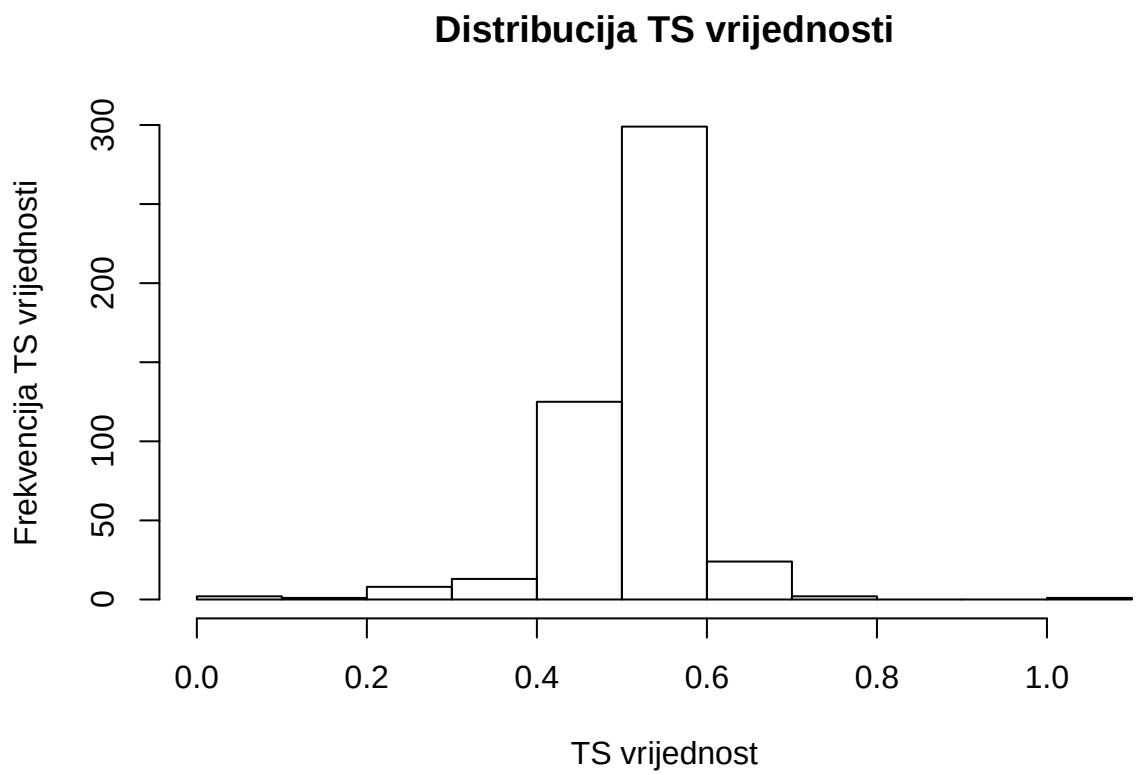
Uspoređivali smo studente koji su draftani s Duke Universityja u odnosu na ostale, zato što se u današnjici Duke slovi kao sveučilište s najboljim programom za košarku. Htjeli smo pokazati da TS vrijednost za studente s Duke Universityja veći u odnosu ostale.

```r
seasons.stats$Player = as.character(seasons.stats$Player)
player.data$name = as.character(player.data$name)
season.player.nice.data = inner_join(seasons.stats, player.data, by=c("Player" = "name"))
```

```r
# igraci koji su draftani s Duke Uneverityja
season.player.nice.data %>% filter(college == "Duke University") -> duke.players

summary(duke.players$TS.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.4870  0.5250  0.5168  0.5610  1.0420       1
```

```r
hist(
  main = "Distribucija TS vrijednosti",
  xlab = "TS vrijednost",
  ylab = "Frekvencija TS vrijednosti",
  duke.players$TS.
)
```

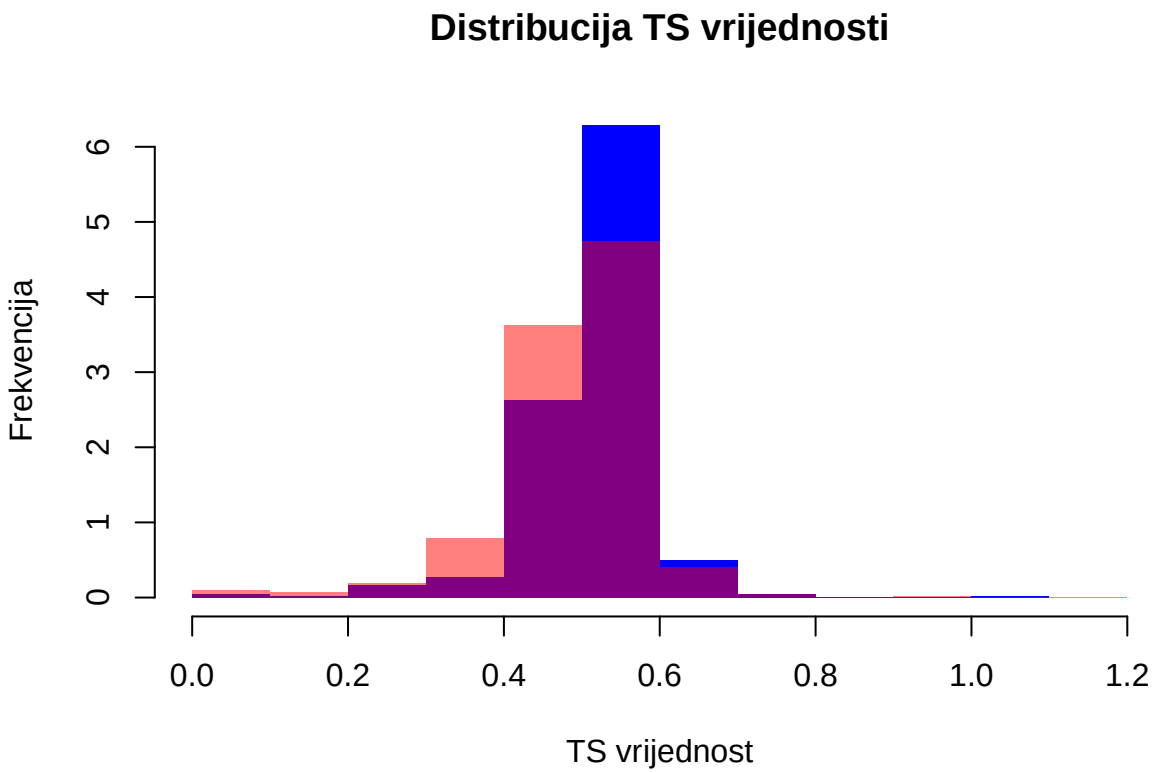



```r
# igraci koji su draftani s drugih sveucilista
season.player.nice.data %>% filter(college != "Duke University") -> not.duke.players
summary(not.duke.players$TS.)
```

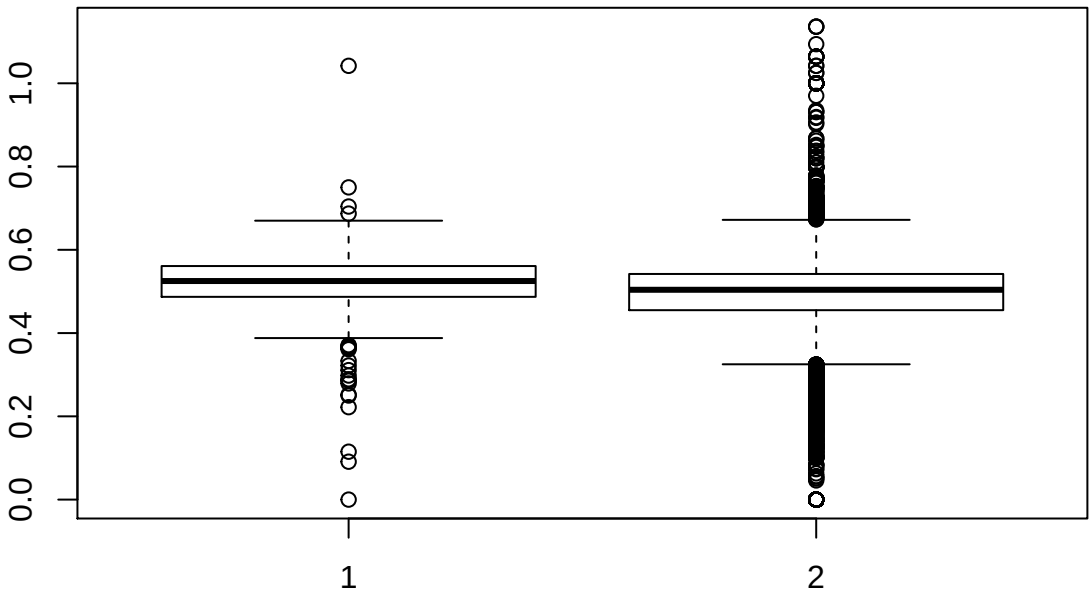
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.4550  0.5040  0.4907  0.5420  1.1360      86
```

```r
# donja granica
dg = 0
# gornja granica
gg = 1.2

# graficki prikaz frekvencije TS vrijednosti za igrace sa Duke universityja i ostalih
# TS vrijednosti za igrace s Duke-a su oznaceni plavom bojom, dok su za ostale oznacene crvenom
hist(
```

```
  duke.players$TS.,
  xlim = c(dg,gg),
  prob = TRUE,
  col = 'blue',
  border = F,
  xlab = "TS vrijednost",
  ylab = "Frekvencija",
  main = "Distribucija TS vrijednosti"
)
hist(
  not.duke.players$TS.,
  add = T,
  xlim = c(dg,gg),
  prob = TRUE,
  col = scales::alpha('red',.5),
  border = F,
  xlab = "TS vrijednost",
  ylab = "Frekvencija",
  main = "Distribucija TS vrijednosti"
)
```




```
# usporedni pravokutni dijagrami za TS vrijednosti igraca
boxplot(duke.players$TS., not.duke.players$TS.)
```




```
# t.test(duke.players$TS., not.duke.players$TS., alternative = "greater", var.equal = TRUE)
# t.test(duke.players$TS., not.duke.players$TS., alternative = "greater", var.equal = FALSE)
```

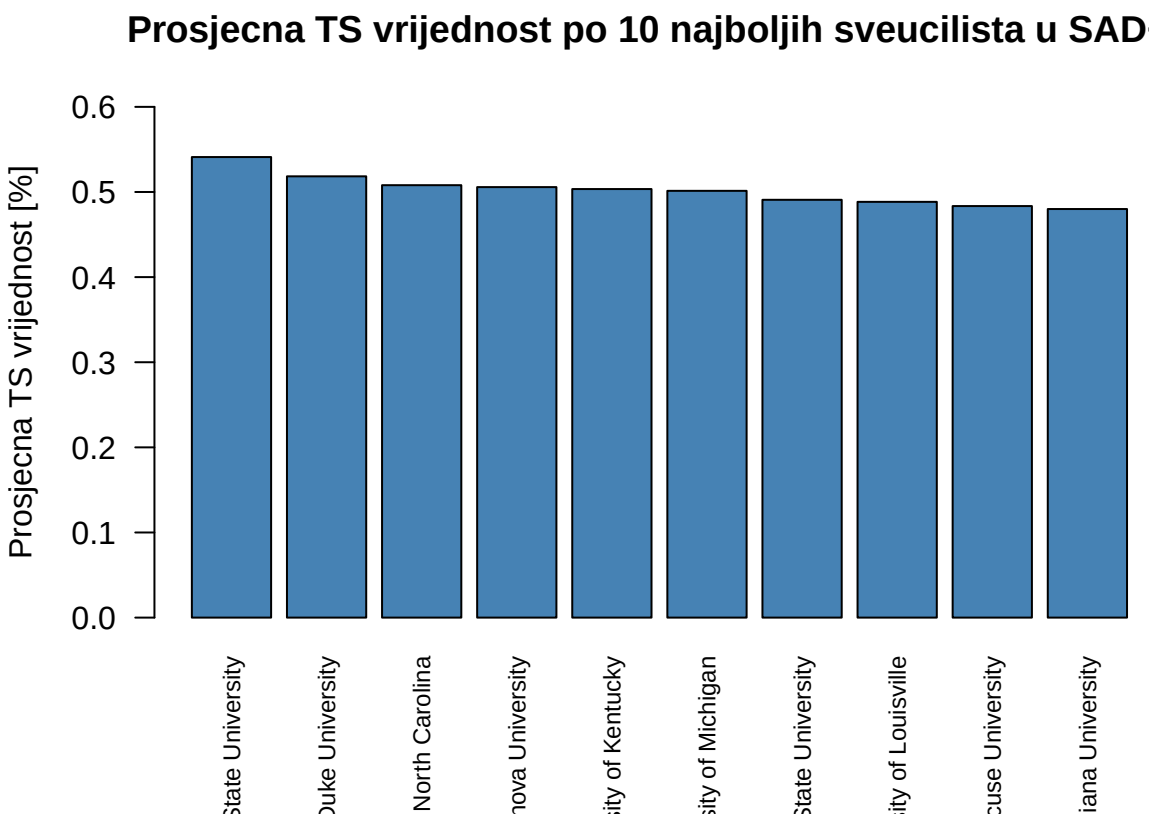
Usporedba TS vrijednosti među sveučilištima s najboljim košarkaškim programom nakon 1990. godine

Valja napomenuti kako su navedena sveučilišta najbolja na temelju trenutne ljestvice. Pošto svako sveučilište ima različitu politiku igranja, na temelju TS vrijednosti htjeli smo vidjeti kako se politike odražavaju na iste.

```
# najboljih 10 sveucilista s kosarkaskim programom
top.10.colleges = factor(c(
  'Duke University',
  'University of North Carolina',
  'University of Kentucky',
  'Kansas State University',
  'Michigan State University',
  'Villanova University',
  'Indiana University',
  'University of Michigan',
  'University of Louisville',
  'Syracuse University'
))

# filtritanje igraca koji su igrali nakon 1990, a dosli su s nekog od najboljih 10 sveucilista
season.player.nice.data %>% filter(Year > 1990 & college %in% top.10.colleges) -> seasons.stats.after.19
# projekcija na TS vrijednosti i svucilista
seasons.stats.after.1990 = seasons.stats.after.1990[c("TS.", "college")]
seasons.stats.after.1990 = seasons.stats.after.1990[complete.cases(seasons.stats.after.1990),]

# prosjecna TS vrijednost po najboljih 10 sveucilista
seasons.stats.after.1990 %>% group_by(college) %>% summarize(ts.college.average = mean(TS.)) -> avg.ts.t

# padajuci poredak po prosjecnoj TS vrijednosti
avg.ts.top.10.colleges <- avg.ts.top.10.colleges[order(-avg.ts.top.10.colleges$ts.college.average),]
avg.ts.top.10.colleges
```

```
## # A tibble: 10 x 2
##    college                       ts.college.average
##    <fct>                                      <dbl>
##  1 Kansas State University                    0.541
##  2 Duke University                            0.518
##  3 University of North Carolina               0.508
##  4 Villanova University                       0.506
##  5 University of Kentucky                     0.503
##  6 University of Michigan                     0.501
##  7 Michigan State University                  0.491
##  8 University of Louisville                   0.488
##  9 Syracuse University                        0.483
## 10 Indiana University                         0.480
```

```r
# graficki prikaz TS vrijednosti po 10 najboljih sveucilista
barplot(
  avg.ts.top.10.colleges$ts.college.average,
  ylim = c(0, 0.6),
  main = "Prosjecna TS vrijednost po 10 najboljih sveucilista u SAD-u",
  ylab = "Prosjecna TS vrijednost [%]",
  xlab = "",
  names.arg = avg.ts.top.10.colleges$college,
  col = "steelblue",
  las = 2,
  cex.names = 0.75
)
```

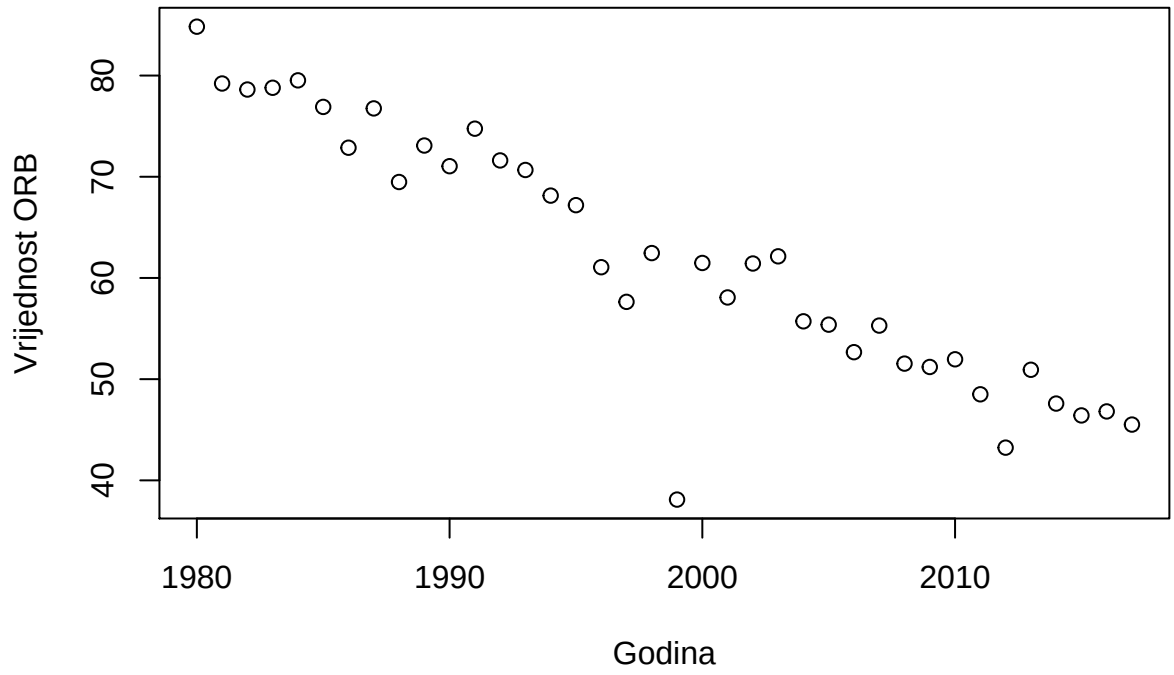


Ovisnost 3PA o skoku u napadu

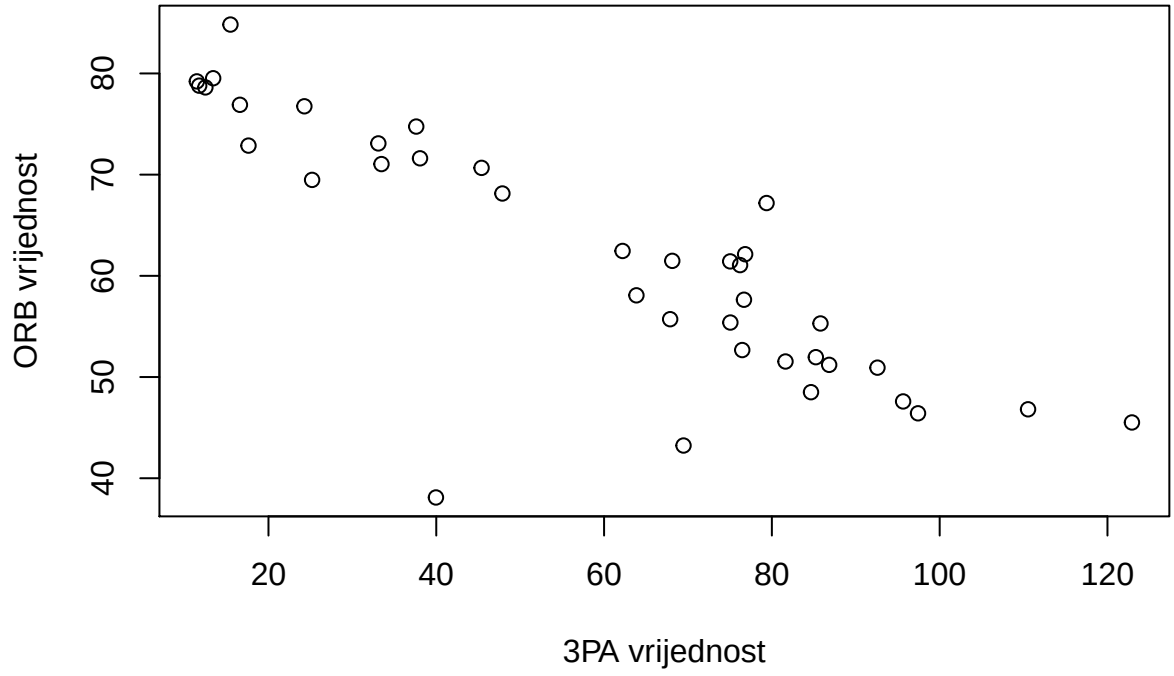
Htjeli smo pokazati kako porastom pokušaja bacanja trica, raste i broj skokova u napadu, točnije rađa se više šansi da napadački igrač uhvati loptu koja se odbije nakon neuspjele trice. Rezultati su neugodno iznenadili, jer su pokazali upravo suprotno od očekivanog, odnosno da broj skokova u napadu pada s porastom pokušaja trica. Jedna točka na grafu ovisnosti ORB o 3PA označuje prosječnu godišnju ORB vrijednost za pridruženu 3PA vrijednost.

```
# projekcija na godinu, 3PA, ORB (skok u napadu), ORB. (postotak skokova u napadus)
help.table = seasons.stats[c("Year", "X3PA", "ORB", "ORB.")]
help.table = help.table[complete.cases(help.table),]

# prosjecan broj 3PA po godini
help.table %>% group_by(Year) %>% summarize(three.point.attempts = mean(X3PA)) -> average.three.point.a
# prosjecan broj ORB po godini
help.table %>% group_by(Year) %>% summarize(offensive.rebounds = mean(ORB)) -> average.orb.by.year
# prosjecan broj ORB% po godini
help.table %>% group_by(Year) %>% summarize(offensive.rebounds.percentage = mean(ORB)) -> average.orb.p

plot(
  average.three.point.attempts.by.year,
  xlab = "Godina",
  ylab = "Vrijednost 3PA"
)
```


```
plot(
  average.orb.by.year,
  xlab = "Godina",
  ylab = "Vrijednost ORB"
)
```




```
plot(
  average.three.point.attempts.by.year$three.point.attempts,
  average.orb.by.year$offensive.rebounds,
  xlab = "3PA vrijednost",
  ylab = "ORB vrijednost"
)
```

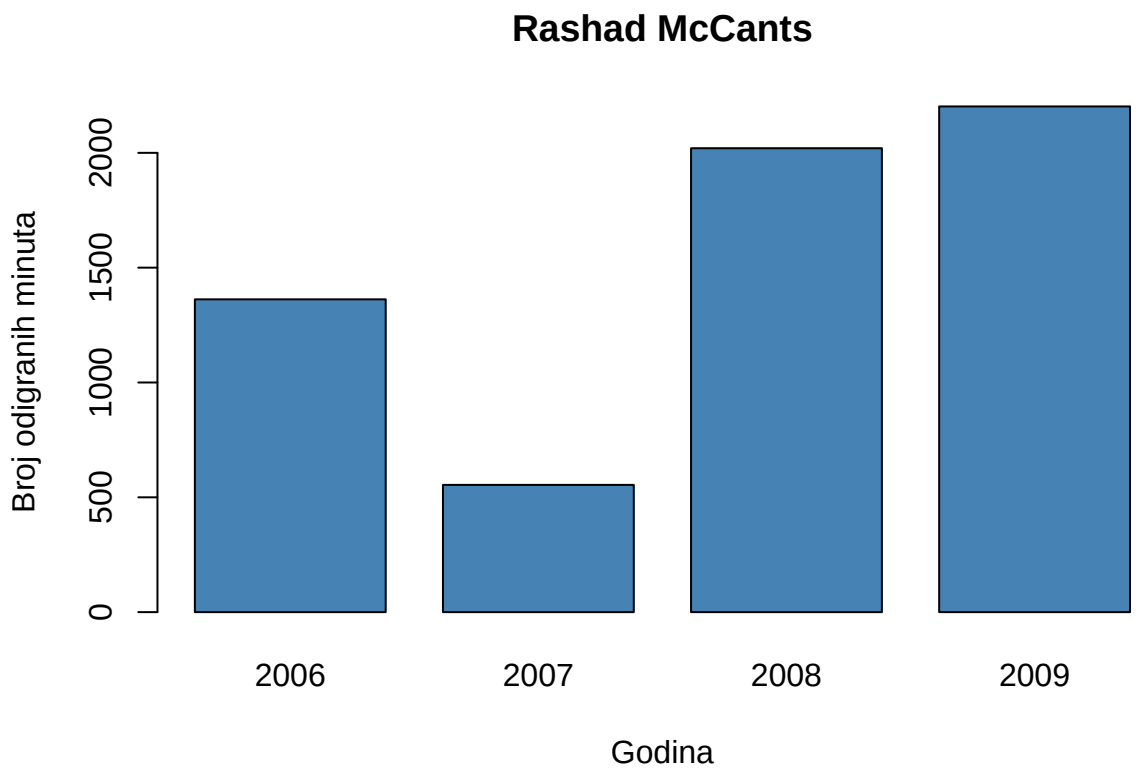


Kardashian Curse

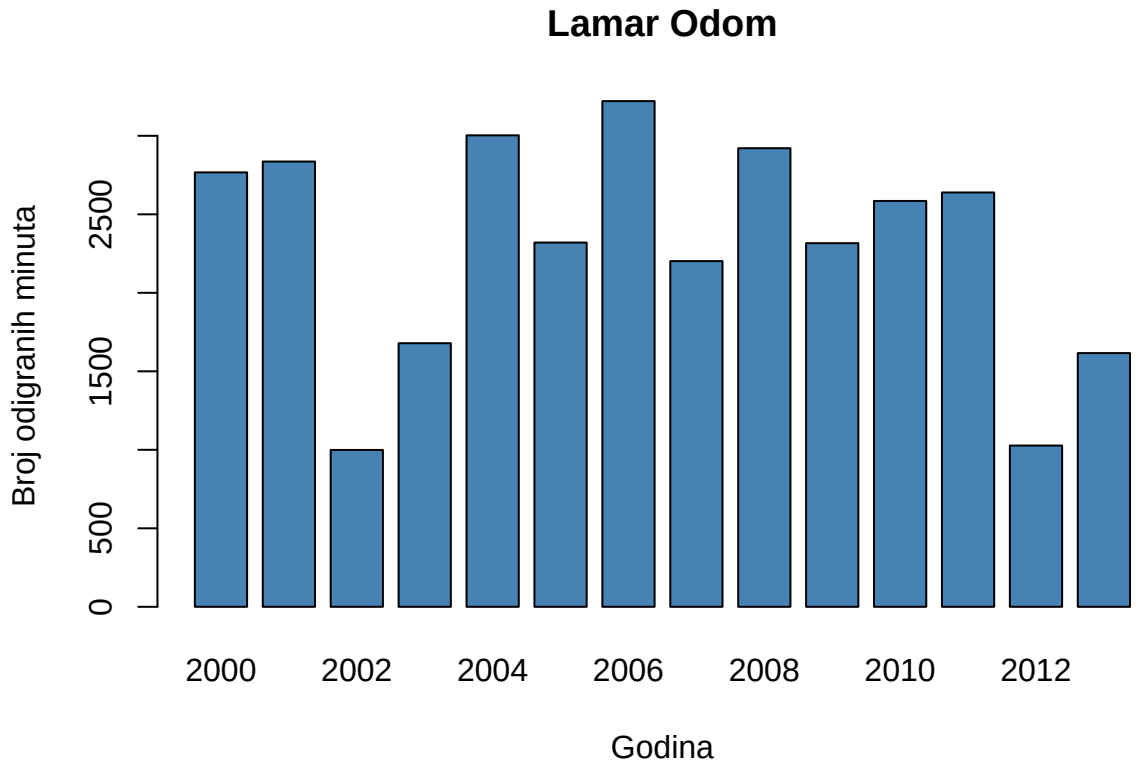
Postoji teorija da igrači koji su bili u vezi s Kardashiankama da je od tada putanja njihove karijere išla samo prema dolje. Treba napomenuti da za nedavne partnere nemamo podatke na temelju kojih bismo to mogli

dokazati, ali činjenica je da je nekima karijera nakon toga završila što ide u korist toj teoriji.

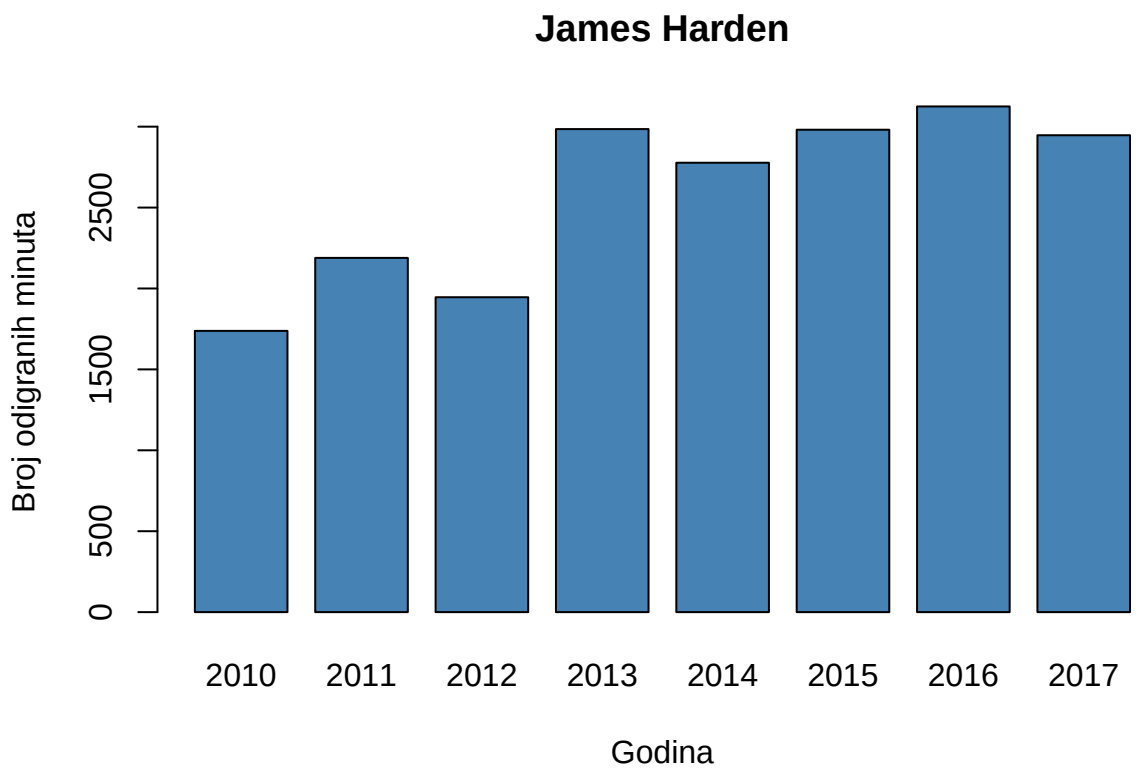
```
# kosarkasi koji su bili u vezi s Khloe Kardashian
khloe.kardashian.boyfriends = factor(c(
  "Rashad McCants", # 2009
  "Lamar Odom", # 2009 - 2013
  "James Harden", # 2015 (trajala je kratko)
  "Tristan Thompson" # 2016 - 2019
))

# kosarkasi koji su bili u vezi s Kim Kardashian
kim.kardashian.boyfriends = c(
  "Kris Humphries" # 2010 - 2011
)

bar_plot_boyfriends <- function(boyfriends) {
  for (boyfriend in boyfriends) {
    seasons.stats %>% filter(Player == boyfriend) %>% group_by(Year) %>% summarize(mp.sum = sum(MP)) ->
    barplot(
      boyfriend.data$mp.sum,
      main = boyfriend,
      names.arg = boyfriend.data$Year,
      ylab = "Broj odigranih minuta",
      xlab = "Godina",
      col = "steelblue",
      space = 0.3
    )
    print("")
  }
}

bar_plot_boyfriends(khloe.kardashian.boyfriends)
```


Rashad McCants
Broj odigranih minuta
0
500
1000
1500
2000
2006
2007
2008
2009
Godina

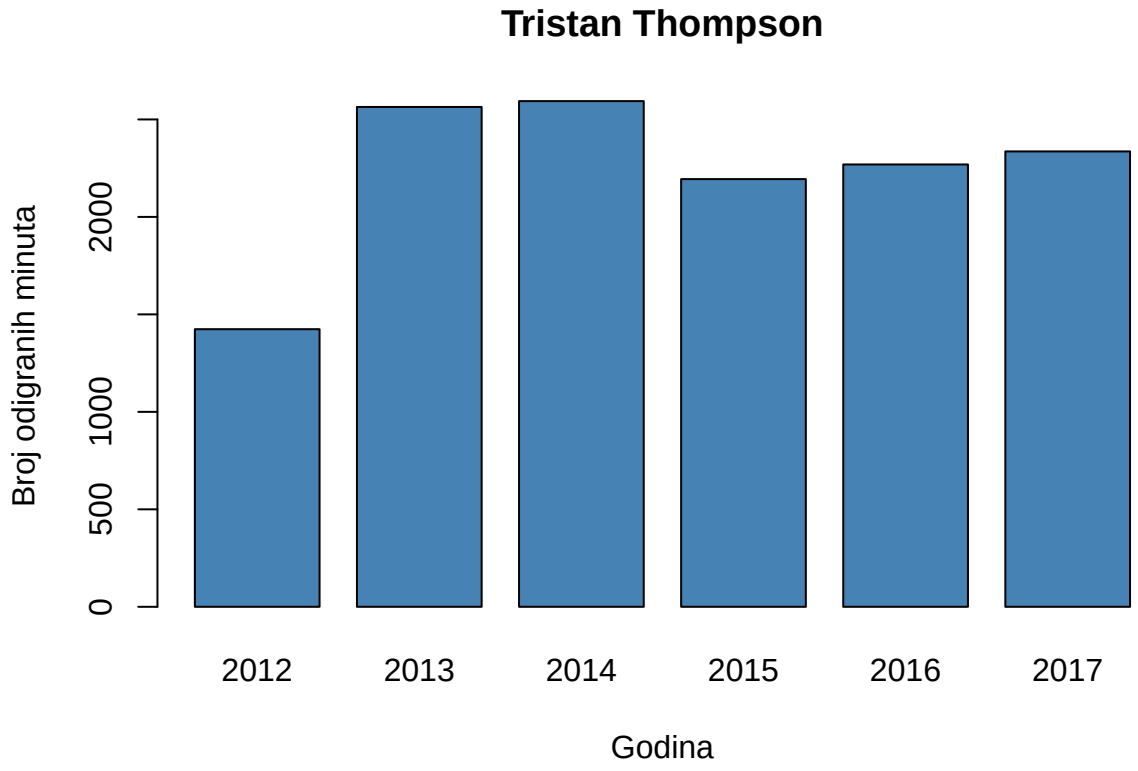
```
## [1] ""
```


Lamar Odom
Broj odigranih minuta
0
500
1500
2500
2000
2002
2004
2006
2008
2010
2012
Godina

```
## [1] ""
```

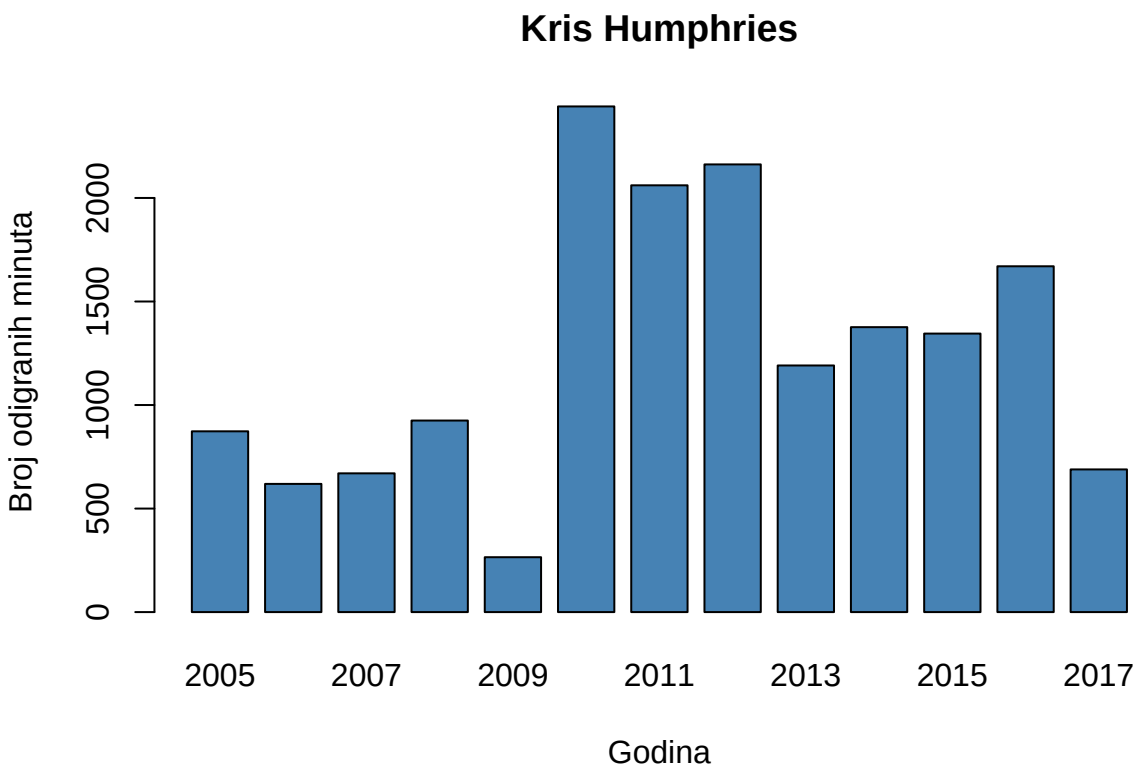



```
## [1] ""
```




```
## [1] ""
```

```
bar_plot_boyfriends(kim.kardashian.boyfriends)
```




```
## [1] ""
```

Udio stranih igrača u NBA ligi u odnosu na igrače iz SAD-a po desetljećima

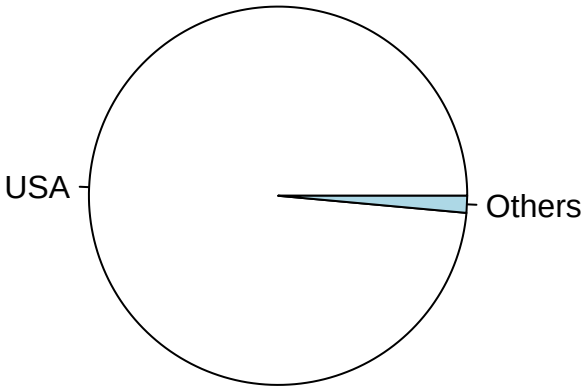
Za očekivati je da će tijekom vremena zbog globalizacije i ostalih uzroka udio stranih igrača rasti. Uspjeli smo pokazati da je kroz desetljeća zbilja bio u porastu. Iznimka su 1960.-e godine u kojima se bilježi pad u odnosu na prethodno desetljeće, za što unatoč trudu, nismo uspijeli naći potencijalan uzrok.

```
usa.states = factor(c(
  "Alabama",
  "Alaska",
  "Arizona",
  "Arkansas",
  "California",
  "Colorado",
  "Connecticut",
  "Delaware",
  "Florida",
  "Georgia",
  "Hawaii",
  "Idaho",
  "Illinois",
  "Indiana",
  "Iowa",
  "Kansas",
  "Kentucky",
  "Louisiana",
  "Maine",
  "Maryland",
  "Massachusetts",
```

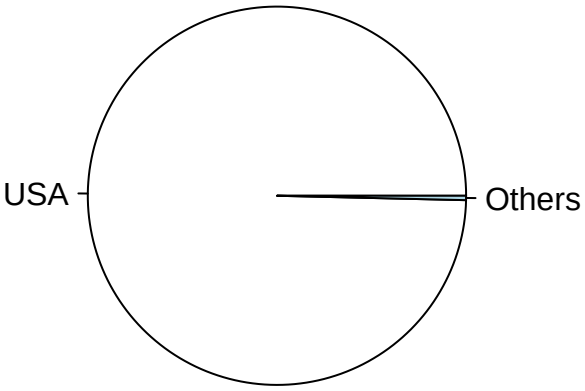
```
  "Michigan",
  "Minnesota",
  "Mississippi",
  "Missouri",
  "Montana",
  "Nebraska",
  "Nevada",
  "New Hampshire",
  "New Jersey",
  "New Mexico",
  "New York",
  "North Carolina",
  "North Dakota",
  "Ohio",
  "Oklahoma",
  "Oregon",
  "Pennsylvania",
  "Rhode Island",
  "South Carolina",
  "South Dakota",
  "Tennessee",
  "Texas",
  "Utah",
  "Vermont",
  "Virginia",
  "Washington",
  "West Virginia",
  "Wisconsin",
  "Wyoming",
  "District of Columbia"
))

years = seq(1950, 2019, by=10)

count = 0
for (year in years) {
  count = count + 1
  test = seasons.stats$Year >= year & seasons.stats$Year < (year + 10)
  combined = seasons.stats[test,]
  combined
  combined$Player = as.character(combined$Player)
  players$Player = as.character(players$Player)

  combined.1 = inner_join(combined, players, by=c('Player' = 'Player'))
  combined.1
  combined.names.countries = combined.1[c('Player','birth_state')]
  combined.names.countries = combined.names.countries[complete.cases(combined.names.countries),]

  combined.names.countries
  combined.names.countries$birth_state = as.character(combined.names.countries$birth_state)

  combined.names.countries = combined.names.countries[combined.names.countries$birth_state != '',]
  combined.names.countries
```

```
  combined.names.countries %>% filter(!(birth_state %in% usa.states)) %>% distinct(Player, birth_state)
  combined.names.countries %>% filter(birth_state %in% usa.states) %>% distinct(Player, birth_state) ->

  slices <- c(nrow(usa.countries), nrow(other.countries))
  lbls <- c("USA", "Others")
  title = paste(as.character(year), '-', as.character(year + 9))
  pie(slices, labels = lbls, main = title)

  if (count %% 2 == 0){
    print("================================================================================")
  }
}
```

1950 – 1959

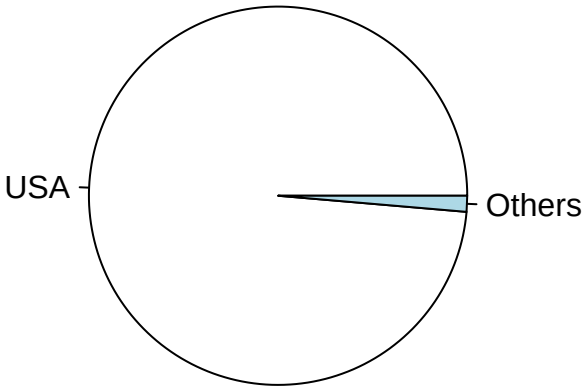


1960 – 1969

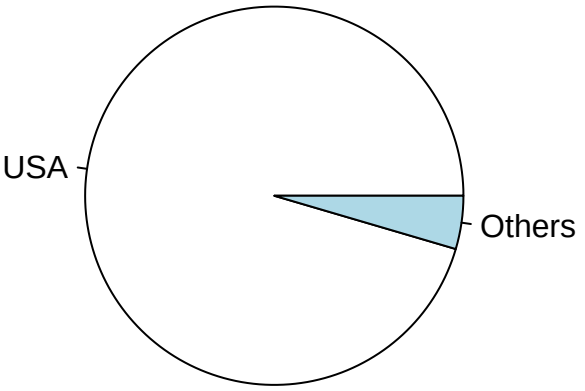



```
## [1] "================================================================================"
```

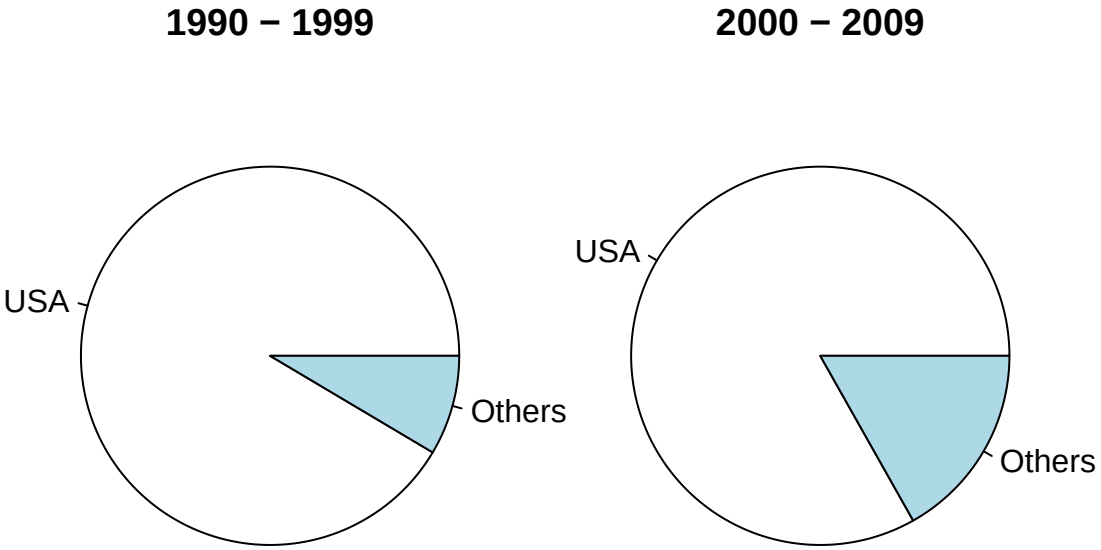
1970 – 1979



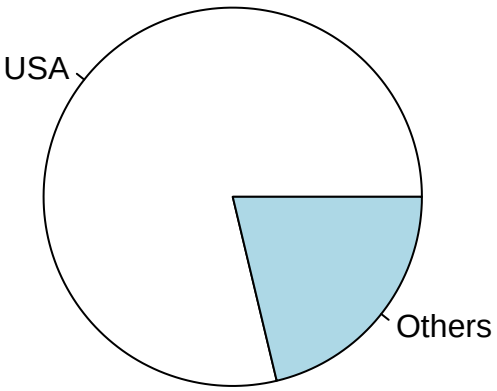
1980 – 1989




```
## [1] "================================================================================"
```




```
## [1] "================================================================================"
```



Testovi za NBA statistiku

Usporedba 3P bacanja za pozicije centra i shooting garda

Pokušali smo otkloniti zavisnost među podacima tako što smo uzeli prosjek trica po svakom igraču, za svaku poziciju. Zatim smo izvršili t-test na dobivenim podacima.

```
# selekcija shooting garda igrača koji su igrali nakon 2010. godine
sg.players = seasons.stats[seasons.stats$Pos == "SG" & seasons.stats$Year > 2010,][c('Player', 'X3P.')]
# selekcija centara koji su igrali nakon 2010. godine
c.players = seasons.stats[seasons.stats$Pos == "C" & seasons.stats$Year > 2010,][c('Player', 'X3P.')]

sg.players.complete = sg.players[complete.cases(sg.players),]
c.players.complete = c.players[complete.cases(c.players),]

# grupiranje sg.players tablice po igračima kako bismo otklonili ovisnosti među podacima
sg.players.complete %>% group_by(Player) %>% summarize(X3P.Average = mean(X3P.)) -> sg.3p.shoots
```

```
c.players.complete %>% group_by(Player) %>% summarize(X3P.Average = mean(X3P.)) -> c.3p.shoots

summary(sg.players)
```

```
##     Player              X3P.
##  Length:859         Min.   :0.0000
##  Class :character   1st Qu.:0.2960
##  Mode  :character   Median :0.3485
##                     Mean   :0.3280
##                     3rd Qu.:0.3852
##                     Max.   :1.0000
##                     NA's   :15
```

```
summary(c.players)
```

```
##     Player              X3P.
##  Length:795         Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.0000
##                     Mean   :0.1722
##                     3rd Qu.:0.3330
##                     Max.   :1.0000
##                     NA's   :365
```

```
# H0: prosjecan 3P% SG = prosjecan 3P% C
# H1: prosjecan 3P% SG > prosjecan 3P% C
t.test(sg.players$X3P., c.players$X3P., alt="greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  sg.players$X3P. and c.players$X3P.
## t = 13.396, df = 543.86, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1365877       Inf
## sample estimates:
## mean of x mean of y
## 0.3279739 0.1722302
```

Usporedba PER (Player efficiency rating) statistike za igrače koji su u karijeri igrali u 1 timu i za igrače koji su igrali u više timova (nakon 1980.)

Na temelju prikupljenih podataka iz deskriptivne statistike htjeli smo testom pokazati da je osjetna razlika između igrača koji su igrali u 1 timu u svojoj NBA karijeri u odnosu na one koji su igrali u više. U obzir smo uzeli statistiku PER (efikasnost igrača) te smo došli do zaključka da igrači koji su igrali u više timova imaju očito veći PER.

```
seasons.stats$Player = as.character(seasons.stats$Player)
seasons.stats$Tm = as.character(seasons.stats$Tm)

# filtriranje igraca koji su igrali nakon 1980. godine
seasons.stats %>% filter(Year > 1980) -> filtered
player.teams = filtered %>% distinct(filtered$Player, filtered$Tm)

names(player.teams) = c("Player", "Tm")
```

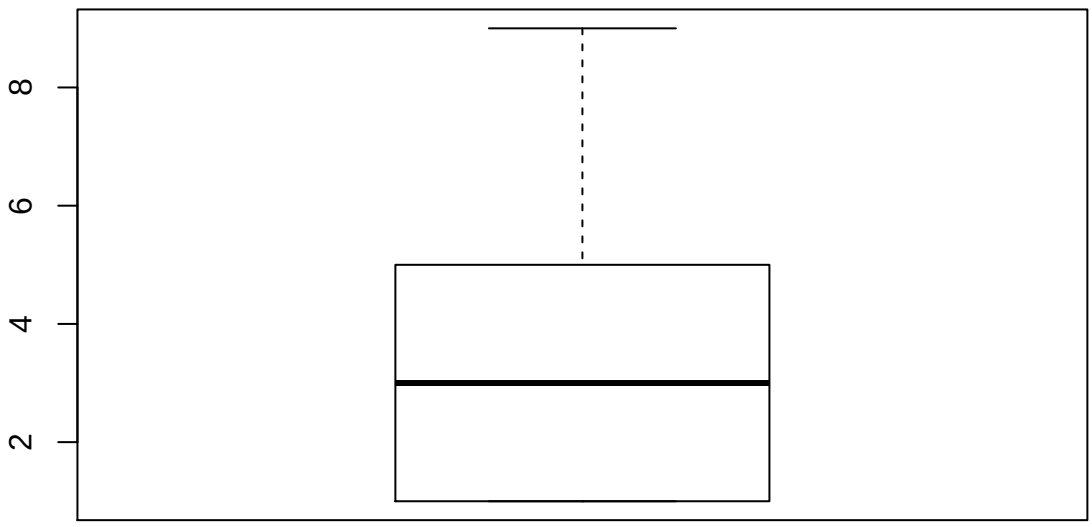
```r
player.teams = player.teams[complete.cases(player.teams),]

# određivanje broja timova za svakog igraca
player.teams %>% group_by(Player) %>% summarize(team.count = length(Tm)) -> player.number.of.teams

player.number.of.teams = player.number.of.teams[-c(1),]

player.pers = seasons.stats[c("Player", "PER")]
player.pers = player.pers[complete.cases(player.pers),]
player.pers %>% group_by(Player) %>% summarize(avg.per = mean(PER)) -> player.per

player.teams.per = inner_join(player.number.of.teams, player.per, by=c("Player" = "Player"))
player.teams.per
```

```
## # A tibble: 2,770 x 3
##    Player         team.count avg.per
##    <chr>               <int>   <dbl>
##  1 A.J. Bramlett           1   -0.4
##  2 A.J. English            1   11.6
##  3 A.J. Guyton             2    4.37
##  4 A.J. Hammons            1    8.4
##  5 A.J. Price              6   10.4
##  6 A.J. Wynder             1    7.6
##  7 Aaron Brooks            7   12.3
##  8 Aaron Gordon            1   14.3
##  9 Aaron Gray              5   10.9
## 10 Aaron Harrison          1    1.05
## # ... with 2,760 more rows
```

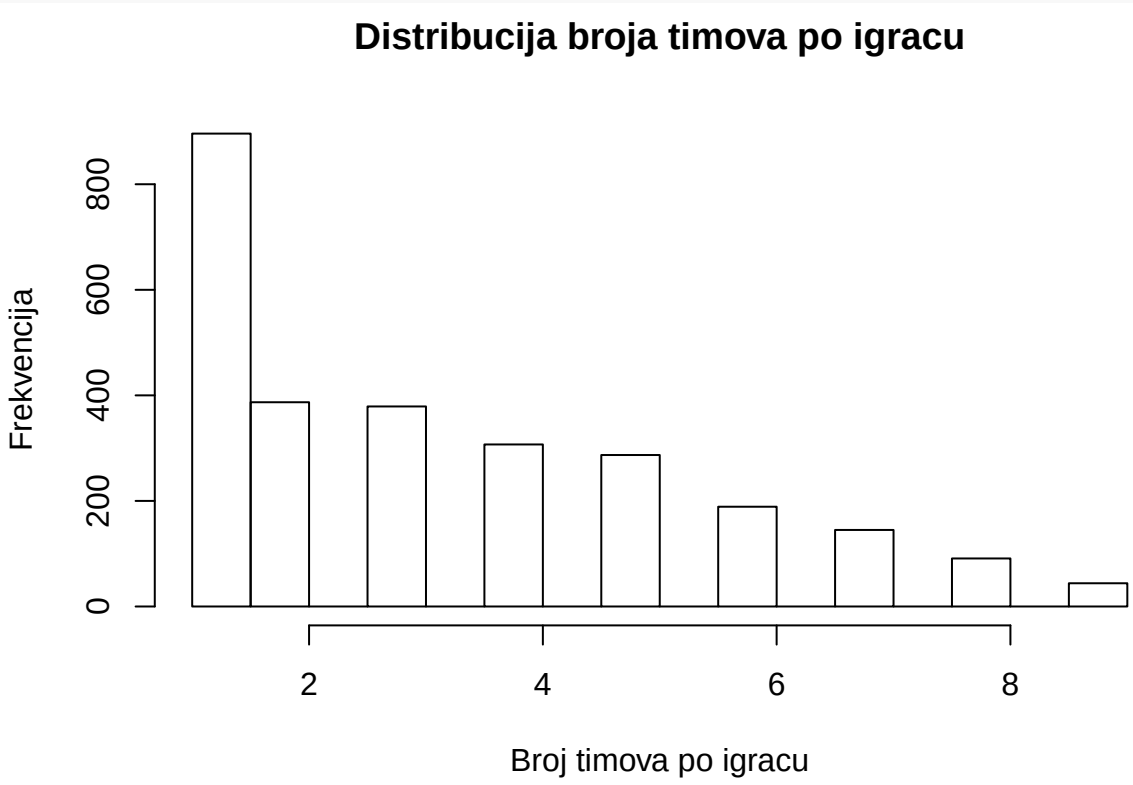
```r
summary(player.teams.per$team.count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    3.00    3.33    5.00   14.00
```

```r
player.teams.per = player.teams.per[player.teams.per$team.count < 10, ]

# pravokutni dijagram za prikaz broja timova po igracu
boxplot(
  player.teams.per$team.count,
  main = "Broj timova po igracu"
)
```




```r
# distribucija broja timova po igracu
hist(
  player.teams.per$team.count,
  xlab = "Broj timova po igracu",
  ylab = "Frekvencija",
  main = "Distribucija broja timova po igracu"
)
```




```r
players.one.team = player.teams.per[player.teams.per$team.count == 1,]
players.multiple.teams = player.teams.per[player.teams.per$team.count > 1,]
```

```r
summary(players.one.team$avg.per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -48.600   6.675  10.300   9.819  13.500  88.300
```

```r
summary(players.multiple.teams$avg.per)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -31.567   9.125  11.700  11.399  13.957  38.233
```

```r
# H0: prosjecan PER igraca s 1 timom = prosjecan PER igraca s vise timova
# H1: prosjecan PER igraca s 1 timom < prosjecan PER igraca s vise timova
t.test(players.one.team$avg.per, players.multiple.teams$avg.per, alt = "less", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  players.one.team$avg.per and players.multiple.teams$avg.per
## t = -4.766, df = 1104.5, p-value = 1.065e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -1.03388
## sample estimates:
## mean of x mean of y
##  9.819352 11.398783
```

Velika trojka: Michael Jordan vs. Kobe Bryant vs. LeBron James

Htjeli smo svesti njihove karijere u par statistika: prosječni PER, prosječni TS, prosječni turnovers, prosječna minutaža, prosječni BPM, prosječan broj prekršaja.

```r
the.great.three = factor(c("Michael Jordan*", "Kobe Bryant", "LeBron James"))

seasons.stats %>% filter(Player %in% the.great.three) %>% group_by(Player) %>% summarize(per.avg = mean

michael = the.great.three.stats[the.great.three.stats$Player == "Michael Jordan*",]
kobe = the.great.three.stats[the.great.three.stats$Player == "Kobe Bryant",]
lebron = the.great.three.stats[the.great.three.stats$Player == "LeBron James",]

michael
```

```
## # A tibble: 1 x 7
##   Player          per.avg ts.avg turnovers.avg mp.avg bpm.avg fouls.avg
##   <chr>             <dbl>  <dbl>         <dbl>  <dbl>   <dbl>     <dbl>
## 1 Michael Jordan*    27.4  0.559          195.  2734.    7.46      186.
```

```r
kobe
```

```
## # A tibble: 1 x 7
##   Player      per.avg ts.avg turnovers.avg mp.avg bpm.avg fouls.avg
##   <chr>         <dbl>  <dbl>         <dbl>  <dbl>   <dbl>     <dbl>
## 1 Kobe Bryant    21.6  0.543          200.  2432.    2.90      168.
```

```r
lebron
```

```
## # A tibble: 1 x 7
##   Player       per.avg ts.avg turnovers.avg mp.avg bpm.avg fouls.avg
##   <chr>          <dbl>  <dbl>         <dbl>  <dbl>   <dbl>     <dbl>
```

```
## 1 LeBron James    27.7  0.586          258.   2948    9.19      141.
```

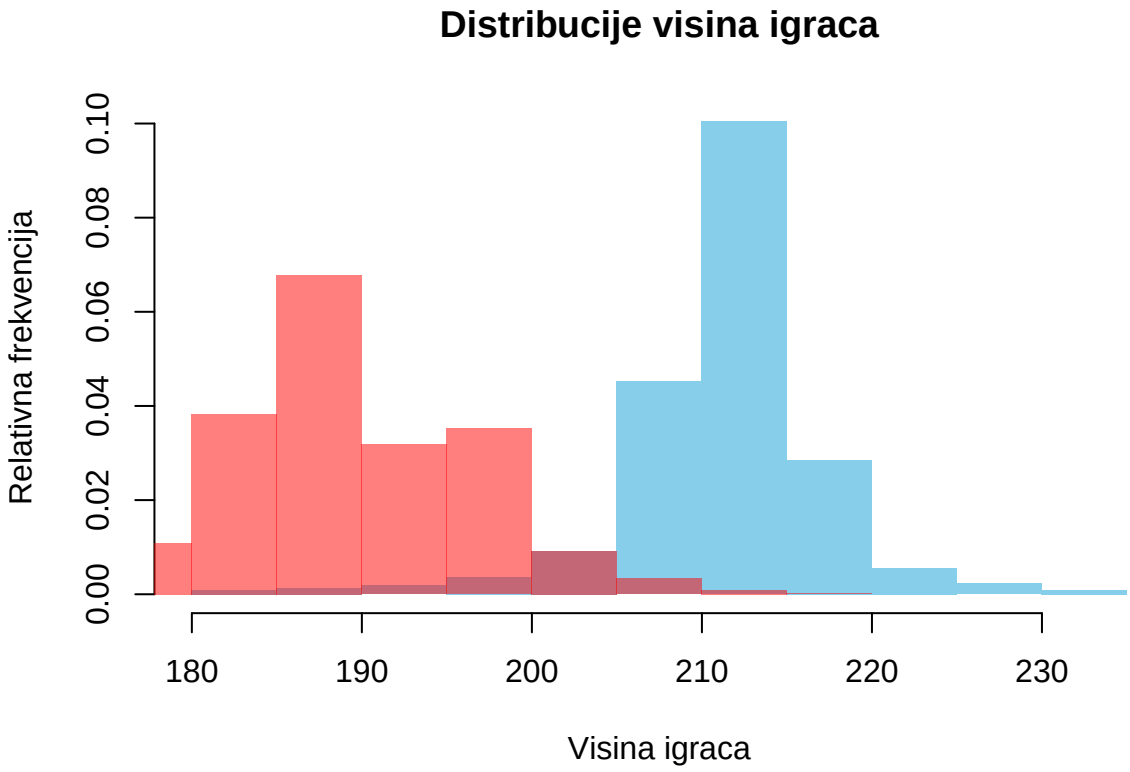
Usporedba visine igrača koji su igrali centra i igrača koji su igrali guard poziciju

S obzirom da smo deskriptivnom statistikom zaključili da visina igrača jako utječe na njihovu ulogu u igri. Testnom statistikom htjeli smo to i pokazati.

```
combined.player.data = inner_join(players, player.data, by=c("Player" = "name"))

combined.player.data %>% filter(position == "C") -> centers.data
combined.player.data %>% filter(position == "G") -> guards.data

# H0: visina centara = visina guardova
# H1: visina centara > visina guardova
t.test(centers.data$height.x, guards.data$height.x, alternative = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  centers.data$height.x and guards.data$height.x
## t = 59.563, df = 1754, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  20.08712      Inf
## sample estimates:
## mean of x mean of y
##  211.0369  190.3790
```

```
hist(
  centers.data$height.x,
  prob = TRUE,
  col = "skyblue",
  border = F,
  xlab = "Visina igraca",
  ylab = "Relativna frekvencija",
  main = "Distribucije visina igraca"
)
hist(
  guards.data$height.x,
  add = T,
  prob = TRUE,
  col = scales::alpha('red', .5),
  border = F,
  xlab = "Visina igraca",
  ylab = "Relativna frekvencija",
  main = "Distribucije visina igraca"
)
```



Izgubljena lopta: stranci vs. Amerikanci

Budući da postoji tvrdnja da je lakše zabiti koš i igrati u NBA ligi, nego u primjerice, Europskoj ligi zbog razlike u pravilima, htjeli smo pokazati da je prosječan broj izgubljenih lopti stranaca u NBA-u statistički manji od Amerikanaca u NBA-u. Unatoč prethodno navedenoj teoriji, nismo uspjeli odbaciti nultu hipotezu da su prosječni brojevi izgubljenih lopti stranaca i Amerikanaca jednaki.

```
seasons.stats$Player = as.character(seasons.stats$Player)
players$Player = as.character(players$Player)

joined.seasons.stats.players = inner_join(seasons.stats,players, by=c("Player"="Player"))

joined.seasons.stats.players %>% group_by(Player, birth_state) %>% summarize(avg.turnover = mean(TOV)) -

players.turnovers %>% filter(!(birth_state %in% usa.states)) -> non.usa.players
players.turnovers %>% filter(birth_state %in% usa.states) -> usa.players

usa.players = usa.players[complete.cases(usa.players),]
non.usa.players = non.usa.players[complete.cases(non.usa.players),]

usa.players = usa.players[sample(nrow(usa.players), 400),]
non.usa.players = non.usa.players[sample(nrow(non.usa.players), 400),]

summary(usa.players$avg.turnover)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.00   34.08   46.95   67.25  273.29
```

```
summary(non.usa.players$avg.turnover)
```

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

```
##    0.00   10.00   33.82   48.51   72.24  209.80
```

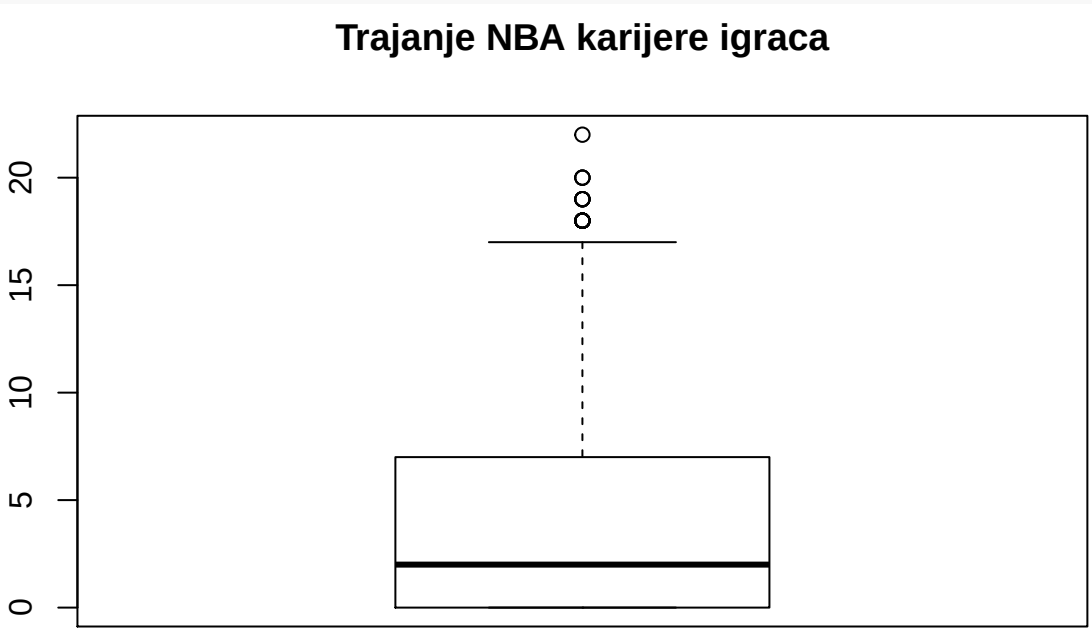
```
# H0: broj izgubljenih lopti Amerikanaca = broj izgubljenih lopti stranaca
# H1: broj izgubljenih lopti Amerikanaca > broj izgubljenih lopti stranaca
t.test(usa.players$avg.turnover, non.usa.players$avg.turnover, alternative = "greater", var.equal = FALS
```

```
##
##  Welch Two Sample t-test
##
## data:  usa.players$avg.turnover and non.usa.players$avg.turnover
## t = -0.48541, df = 796.29, p-value = 0.6862
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -6.816314       Inf
## sample estimates:
## mean of x mean of y
##  46.95346  48.50524
```

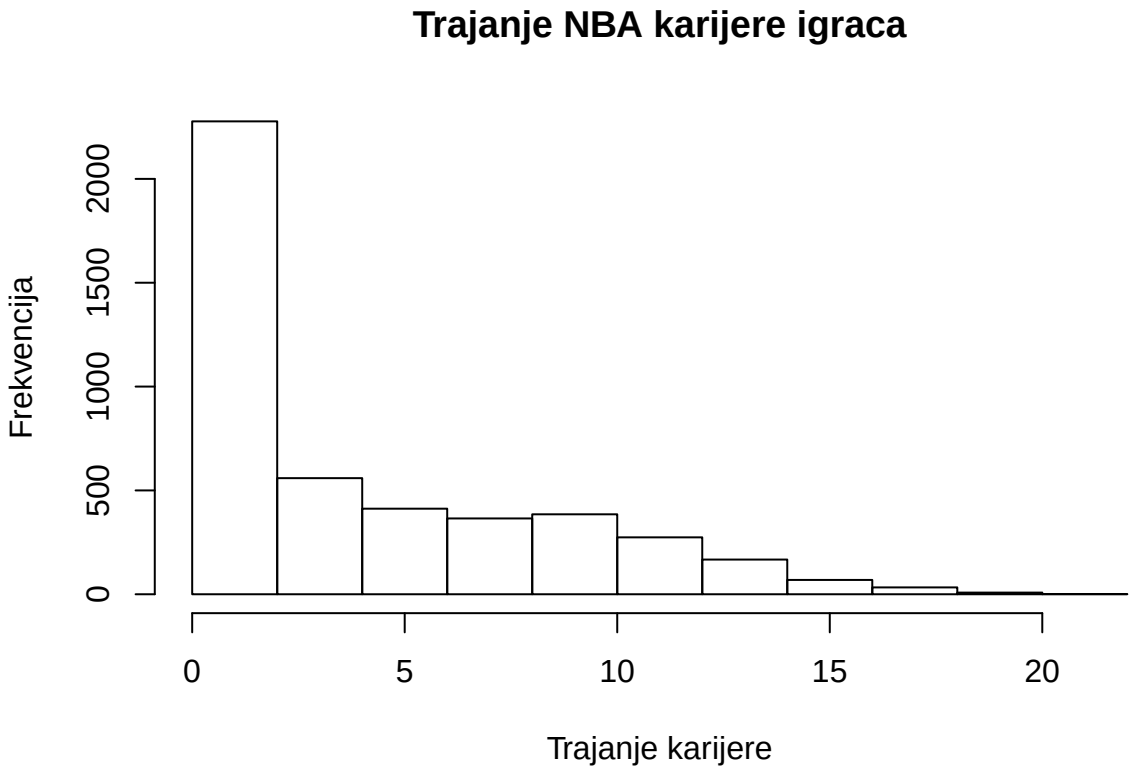
Prosječno trajanje NBA karijere igrača

Deskriptivnom statistikom smo htjeli pokazati distribuciju trajanja NBA karijere.

```
player.data$career.duration = player.data$year_end - player.data$year_start

boxplot(
  player.data$career.duration,
  main = "Trajanje NBA karijere igraca"
)
```




```
hist(
  player.data$career.duration,
  xlab = "Trajanje karijere",
  ylab = "Frekvencija",
  main = "Trajanje NBA karijere igraca"
)
```




```
summary(player.data$career.duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   4.196   7.000  22.000
```

Drazen Petrovic vs Michael Jordan

Dražen Petrović i Michael Jordan igrali su na istoj poziciji (SG) što predstavlja odličan temelj za provođenje statistike. Kako je (nažalost) karijera Dražena Petrovića trajala puno kraće nego karijera Michaela Jordana odlučili smo iskoristiti Bootstrap metodu uzorkovanja i stvoriti podatke za Dražena Petrovića koji zapravo ne postoje i usporediti statistiku ove dvojice igrača kako bi potvrdili teoriju da je Dražen Petrović u svojoj kraćoj karijeri postigao puno bolje rezultate nego Michael Jordan.

```
require(bootstrap)
```

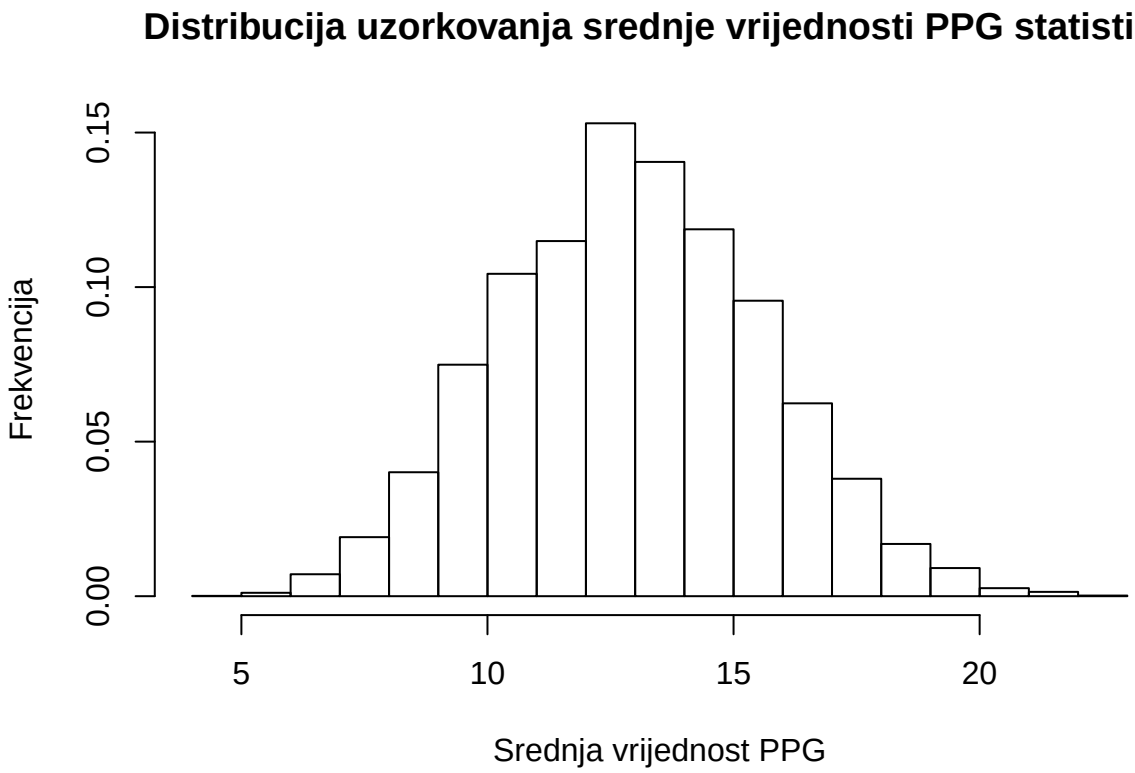
```
## Loading required package: bootstrap
```

```
# bootstrap(podaci, N, statistika)
drazen.petrovic = seasons.stats[seasons.stats$Player == "Drazen Petrovic*",]

drazen.petrovic$PPG = drazen.petrovic$PTS / drazen.petrovic$G

thetastar = bootstrap(drazen.petrovic$PPG, 10000, mean)$thetastar
summary(thetastar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.966  11.067  12.970  12.947  14.762  22.056
```

```
# graficki prikaz distribucije uzorkovanja srednje vrijednosti PPG statistike
hist(
  thetastar,
  main = "Distribucija uzorkovanja srednje vrijednosti PPG statistike",
```

```
  xlab = "Srednja vrijednost PPG",
  ylab = "Frekvencija",
  prob = TRUE
)
```



Linearna regresija

Provjera ovisnosti 3PA o godini linearnom regresijom

Samim pogledom na graf ovisnosti 3PA po utakmici o godini vidljivo je da postoji određena korelacija između te dvije vrijednosti. To nas je inspiriralo da napravimo linearnu regresiju da vidimo u kojoj mjeri zaista vrijedi ta ovisnost.

```
library(tidyverse) # sluzi za grupiranja, joinove, filtriranja

# plotting average three and two point attempts per year
season.stats.points = seasons.stats[c('Year', 'X3PA', 'X2PA', 'G')]

season.stats.points$X3PAPerGame = season.stats.points$X3PA / season.stats.points$G
season.stats.points.complete = season.stats.points[complete.cases(season.stats.points),]

season.stats.points.complete %>% group_by(Year) %>% summarize(ThreePointAttempts = mean(X3PAPerGame)) -

test.data = three.point.attempts[three.point.attempts$Year > 2010,]
three.point.attempts = three.point.attempts[three.point.attempts$Year <= 2010,]

fit = lm(ThreePointAttempts~Year, data=three.point.attempts)
summary(fit)
```

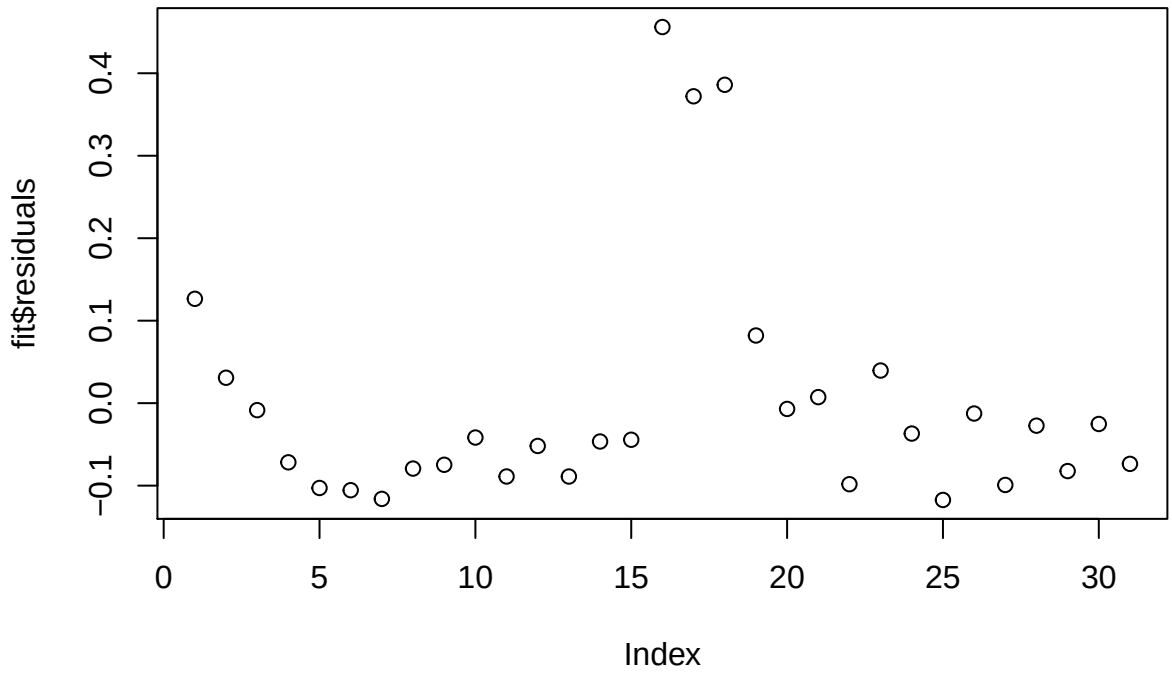
```
##
## Call:
## lm(formula = ThreePointAttempts ~ Year, data = three.point.attempts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11734 -0.08561 -0.04436  0.00017  0.45601
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -96.587516   5.977700  -16.16 4.86e-16 ***
## Year          0.048839   0.002996   16.30 3.86e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1492 on 29 degrees of freedom
## Multiple R-squared:  0.9016, Adjusted R-squared:  0.8982
## F-statistic: 265.7 on 1 and 29 DF,  p-value: 3.864e-16
```

```
fit$coefficients
```

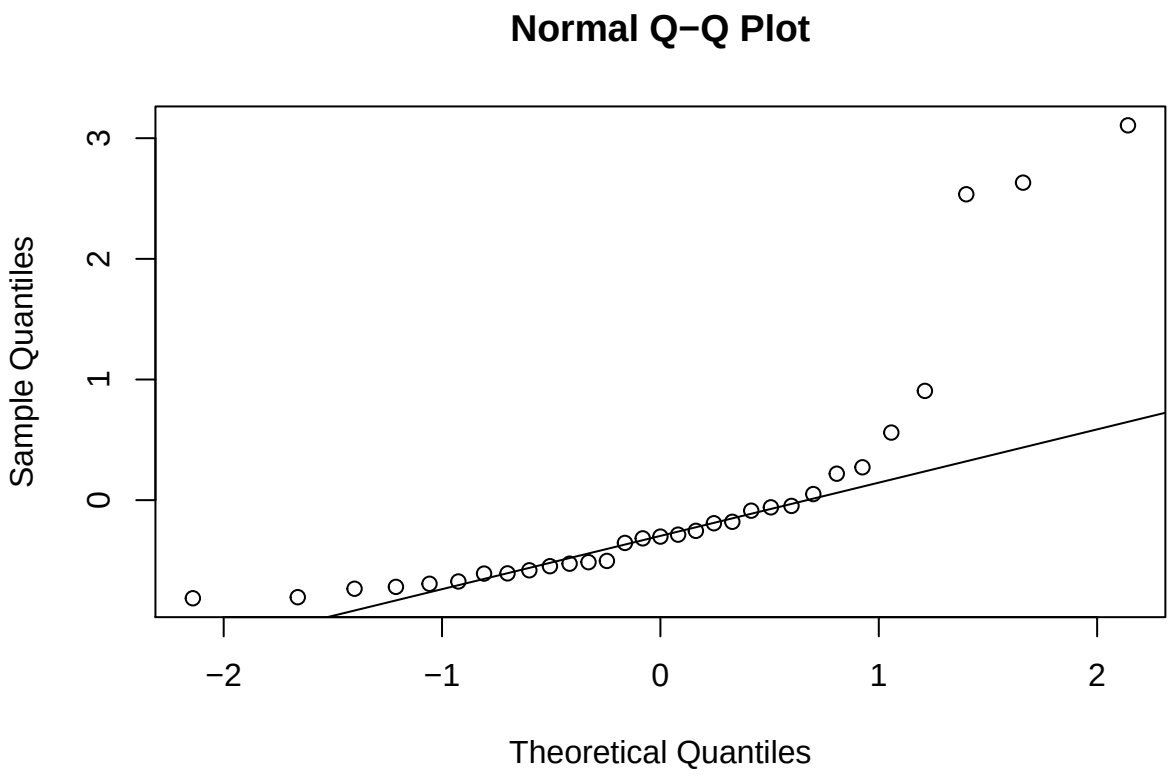
```
##  (Intercept)         Year
## -96.58751620   0.04883948
```

```
plot(
  three.point.attempts$Year,
  three.point.attempts$ThreePointAttempts,
  xlab = "Godina",
  ylab = "3PA po utakmici"
)
lines(
  three.point.attempts$Year,
  fit$fitted.values,
  col = 'skyblue'
)
```


```
# dijagram reziduala
plot(fit$residuals)
```




```
# qq plot
qqnorm(rstandard(fit))
qqline(rstandard(fit))
```




```
ks.test(rstandard(fit),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit)
## D = 0.26099, p-value = 0.02371
## alternative hypothesis: two-sided
```

```
prediction = predict(fit,data=test.data, interval = "confidence")

actual <- data.frame(cbind(actuals=three.point.attempts$ThreePointAttempts, predicteds=prediction))
```