Projekt iz kolegija: Statistička analiza podataka

Statistika NBA igrača

Disperzanti - Vlado Galić, Tin Komerički, Marino Voćanec

12/05/2019

Predgovor

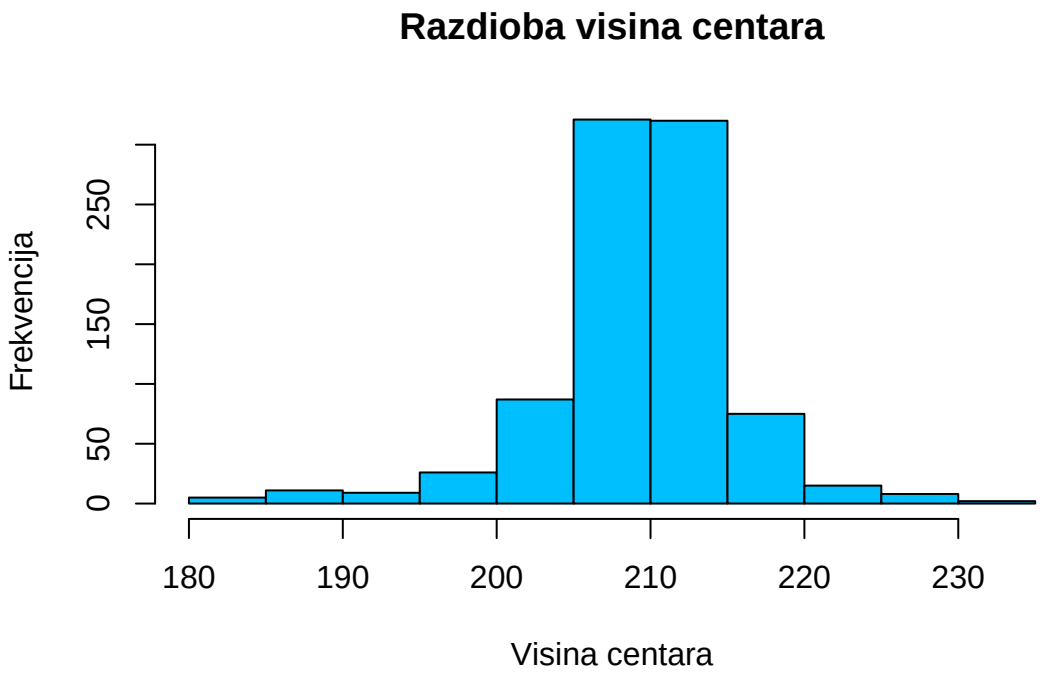
U sklopu projekta iz kolegija "Statistička analiza podataka" naš zadatak je osmisliti zanimljiva pitanja i hipoteze koje ćemo pokušati dokazati (odnosno opovrgnuti) koristeći dosad stećeno znanje na kolegiju. Konkretno naš projekt je vezan uz statistiku NBA igrača od 1950. godine. Skup podataka koji proučavamo sadrži preko 3000 igrača kroz 67 sezona. Za svakog igrača prisutne su sve važne vrijednosti koje će nam poslužiti kako bi mogli donositi zanimljive zaključke.

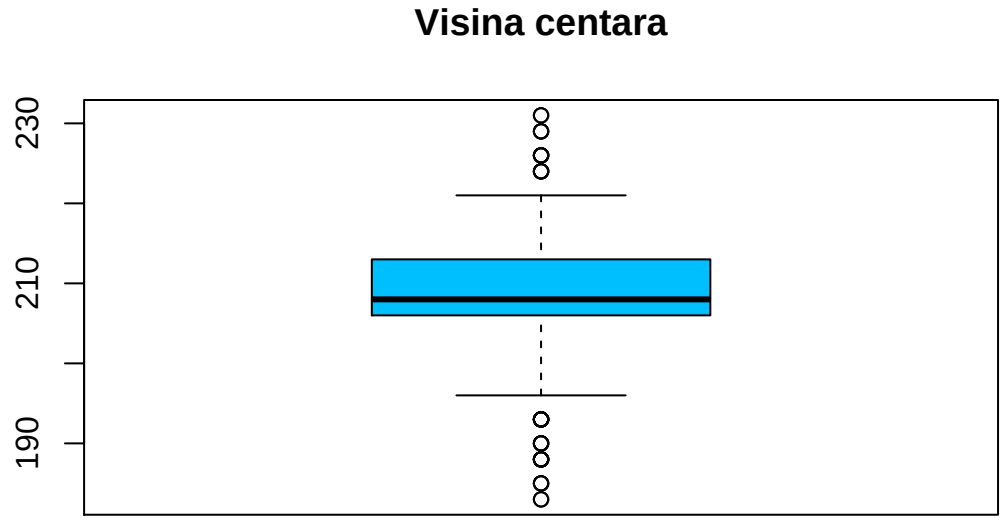
Projekt je koncipiran po cjelinama na takav način da svaka pojedina cjelina zapravo predstavlja jedno zanimljivo pitanje ili hipotezu koju onda dokazujemo. Testovi i regresije koje provodimo su popraćene brojnim vizualizacijama podataka kako bi čitatelju pojednostavili proučavanje dokumenta i dodatno približili ono što se testom ili regresijom zapravo i pokazuje.

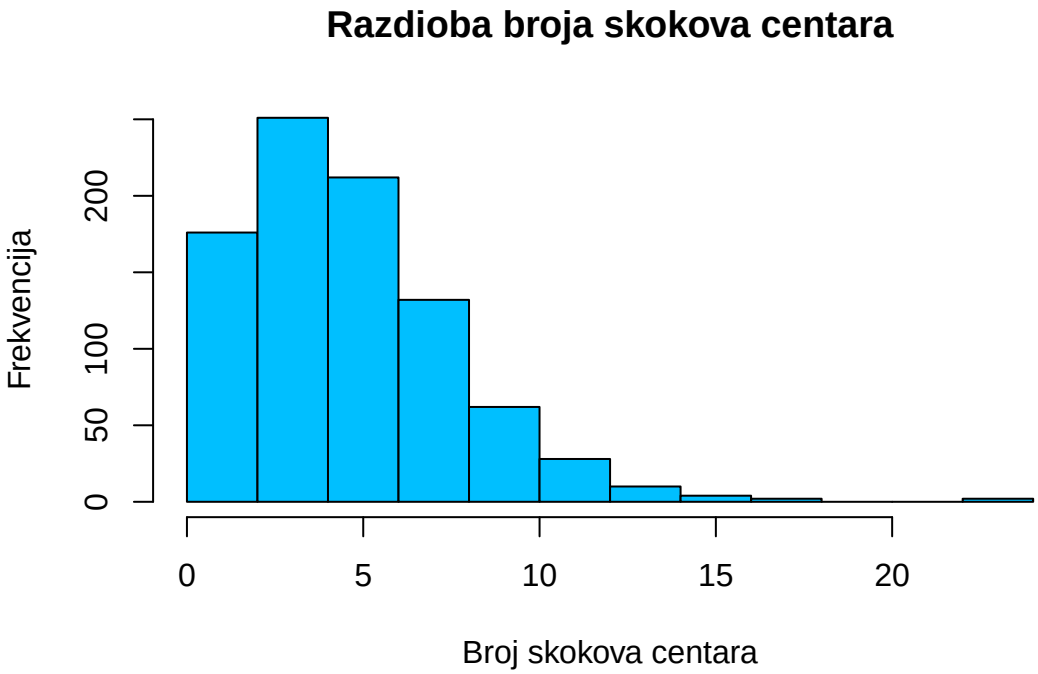
1. Jesu li su viši centri nužno i bolji skakači?

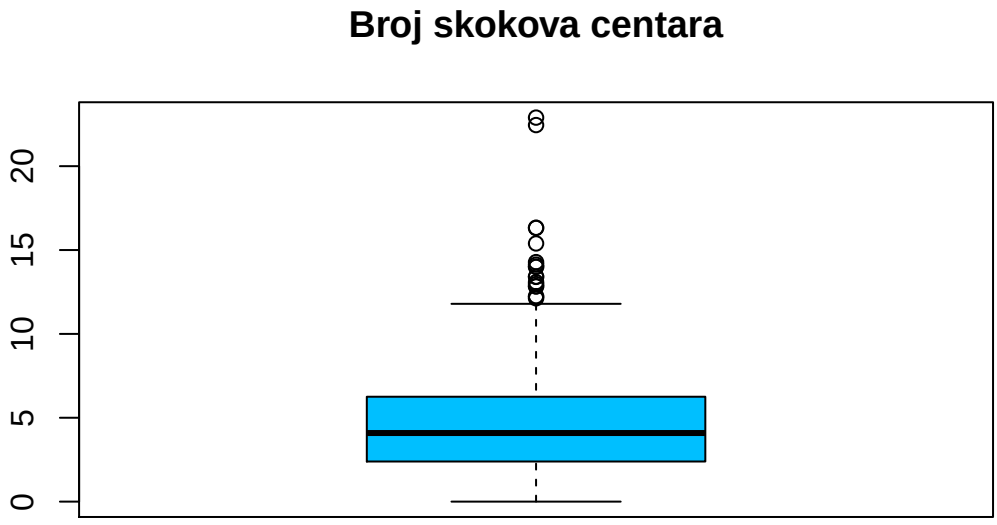
Centar je igrač koji većinu svog vremena provodi pod košem vrebajući skokove. Zbog toga je zanimljivo proučavati je li visina igrača bitan parametar za skupljanje skokova? Prirodno bi bilo očekivati da će visina igrača pozitivno utjecati na količinu skokova. No, je li to baš uvijek tako? Ljubitelji košarke bi mogli tvrditi nešto što se možda na prvi pogled ne čini tako intuitivnim.

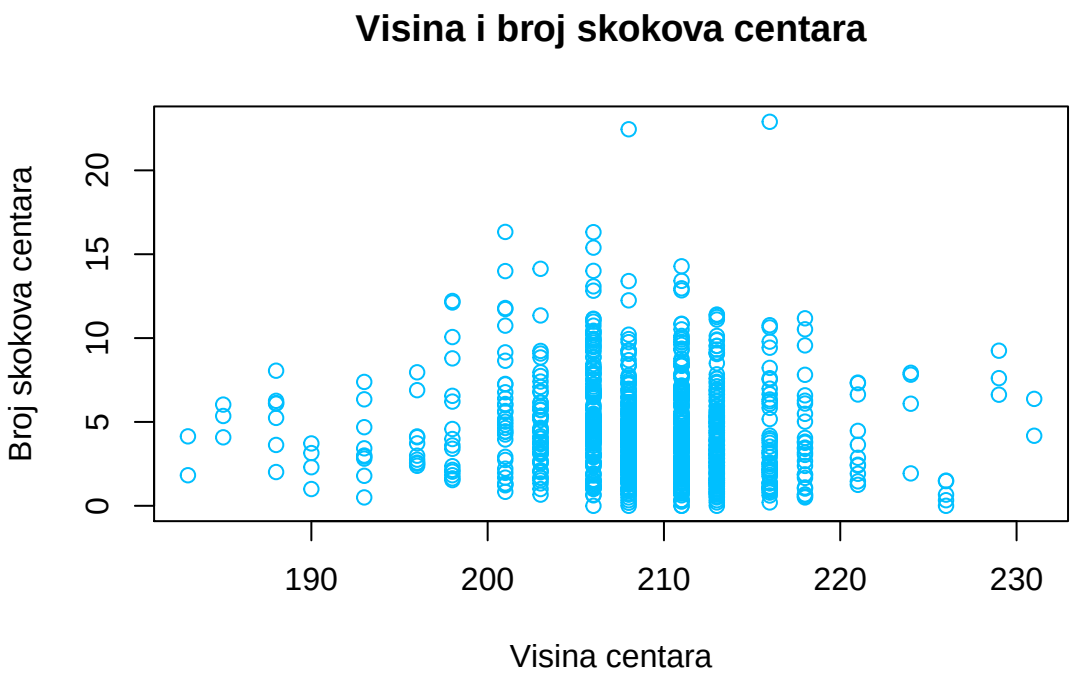
Najprije ćemo vizualizirati podatke kako bismo vidjeli postoji li bilo kakva povezanost između visine i broja skokova.






Razdioba broja skokova centara
Frekvencija
0
50
100
200
0
5
10
15
20
Broj skokova centara


Broj skokova centara
0
5
10
15
20



Sljedeće što ćemo napraviti jest razdijeliti centre u 2 skupine: više i niže centre. Niži centri su svi oni koji imaju visinu manju od srednje vrijednosti umanjene za 5% ranga visina. Na isti način se određuju i viši centri, samo što se kod viših centara izračunati postotak nadodaje na srednju vrijednost.

Nakon što smo centre podijelili na niže i više primijenti ćemo Bootstrap tehniku kako bismo utvrdili da li NIŽI igrači na poziciji centra imaju višeskokova od VIŠIH centara. Ovo možda na prvi pogled djeluje čudo, ali pogledajmo što će reći rezultati postupka.

```
## Srednja vrijednost broja skokova nizih centara: 5.1941
```

```
## Srednja vrijednost broja skokova visih centara: 4.10826
```

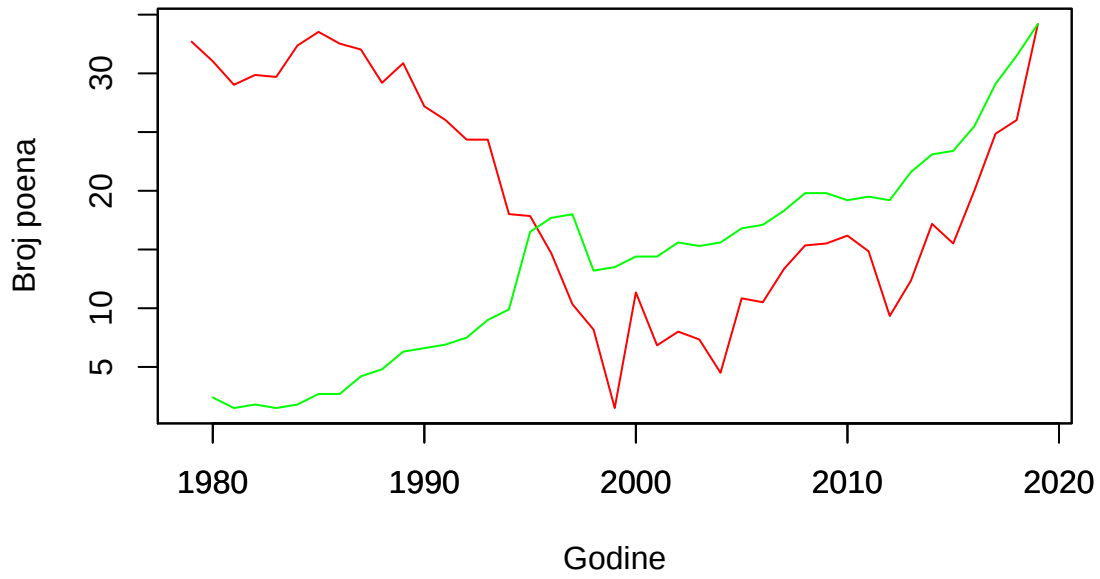
```
## Interval povjerenja jednak je: [0.45569, Inf].
```

Kako vrijednost 0 'nije upala' u 95%-tni interval povjerenja za razliku broja skokova između nižih i viših centara, možemo tvrditi na razini značajnosti od 5% kako niži centri imaju više skokova od viših. No, valja biti oprezan! Ovaj rezultat jest statistički značajan, ali je značajnost u domeni košarke upitna. Ipak, radi se o razlici između broja skokova po utakcimi ne većoj od 0.5 što možda u nekom kontekstu i nije toliko bitna razlika.

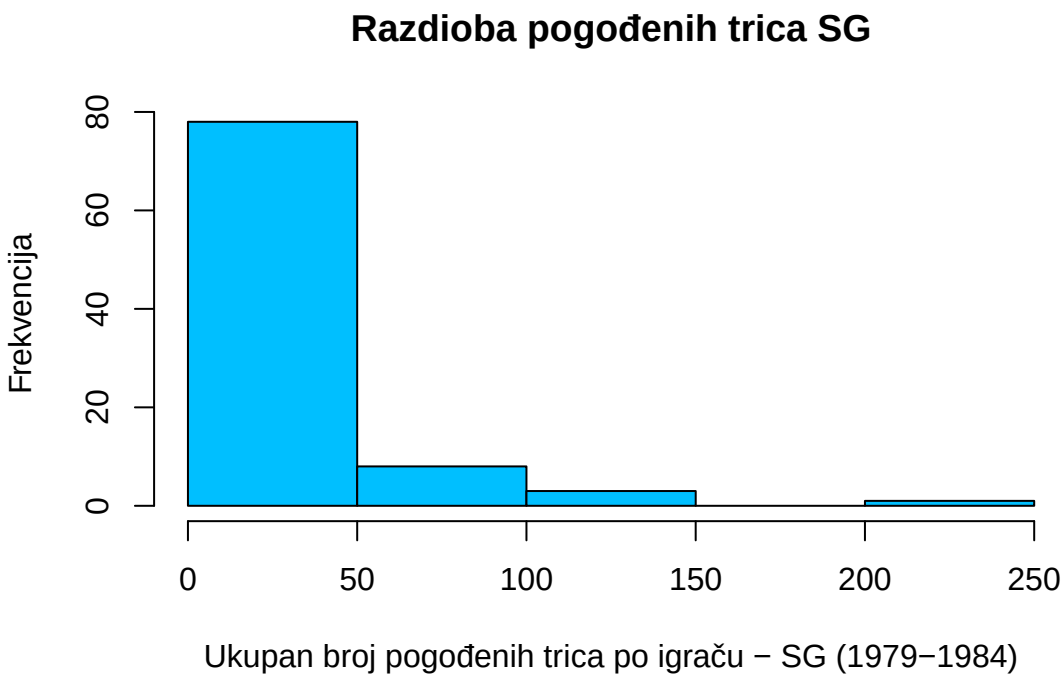
2. Evolucija šuta za tri poena

Čak i netko tko površno prati NBA ligu može uočiti jedan zanimljiv trend a to je da se u današnjoj košarci šuta puno više trica. Teško je naći baš razlog zbog čega se to događa ali očito je da su igrači sve više vještiji u šutiranju trica pa čak i oni koji igraju poziciju centra (inače loši šuteri za 3). Na grafu prikazujemo ZELENOM bojom broj poena po timu po sezoni, a CRVENOM prosječan broj zabijenih trica u toj sezoni.

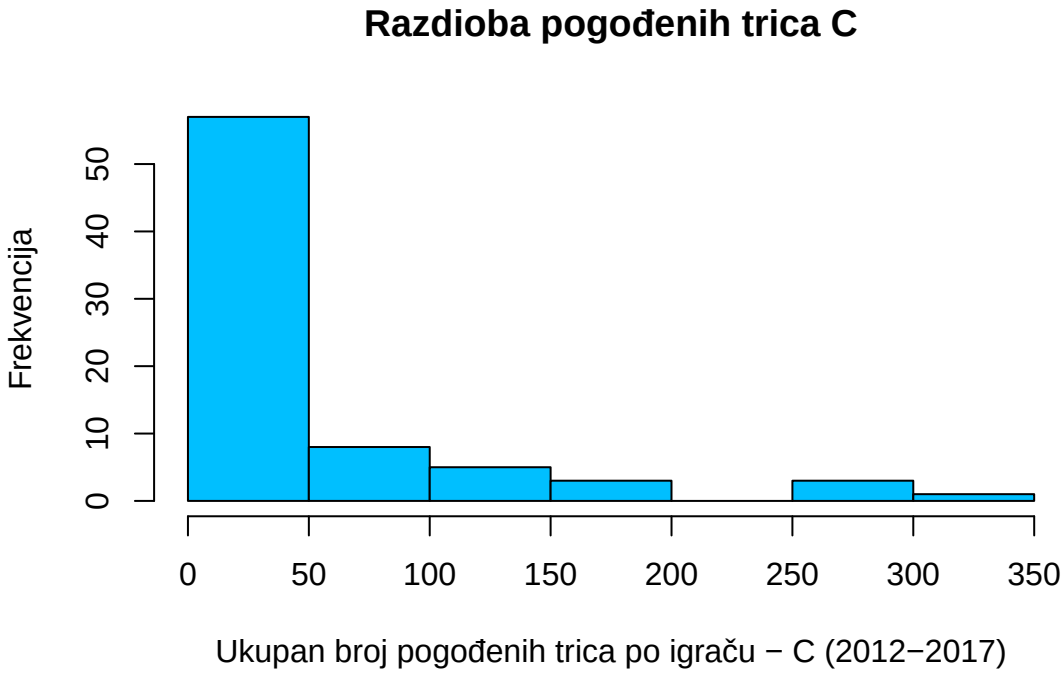
Broj koševa po NBA sezoni:



Također možemo zapaziti da današnji igrači u prosjeku više zabiju trica nego tadašnji bek šuteri. To je također dokaz evolucije igre i uvježbavanja vještine kao što je šut za tri poena. Nadalje, zanimljivo je uočiti kako je od 1979. godine (odnosno godine kada je uvedena trica) postepeno rastao broj poena po timu po sezoni što može biti dokaz evolucije u tom segmentu igre - igrači postaju sve vještiji u zabijanju trica.




```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.00    3.00   12.00   24.81   29.00  247.00
```

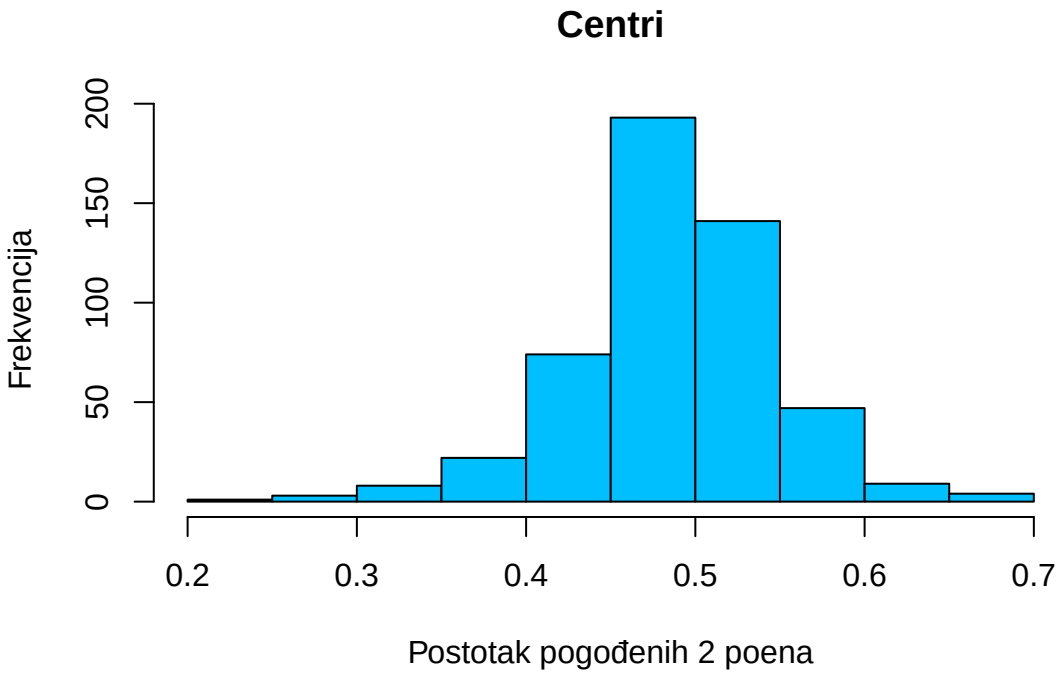


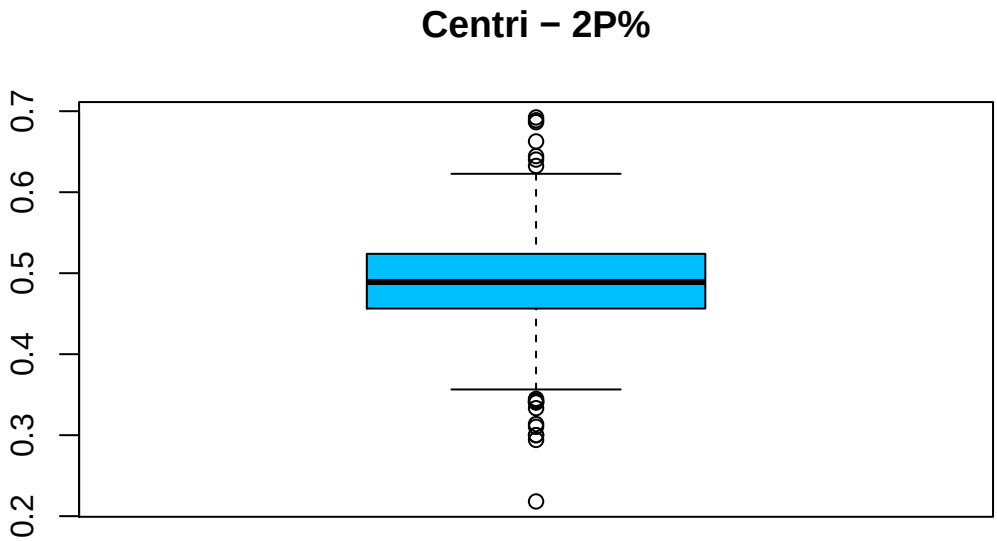

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.00    2.00    8.00   44.29   55.00  338.00
```

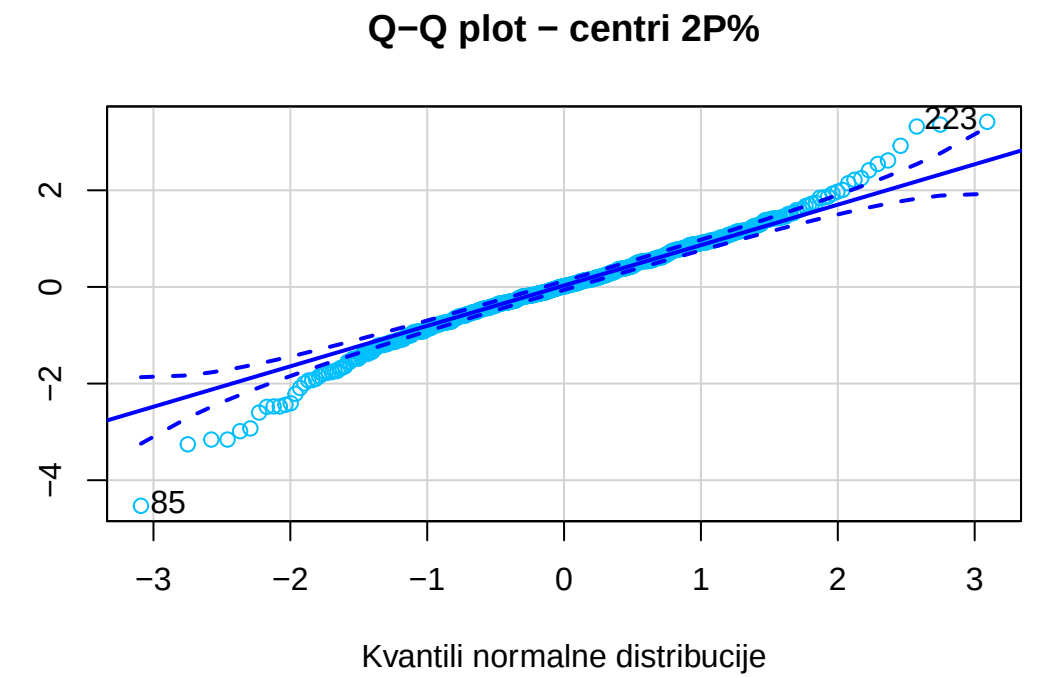
3. Tko je bolji šuter za tri i dva poena: 'Center' ili 'Shooting Guard'?

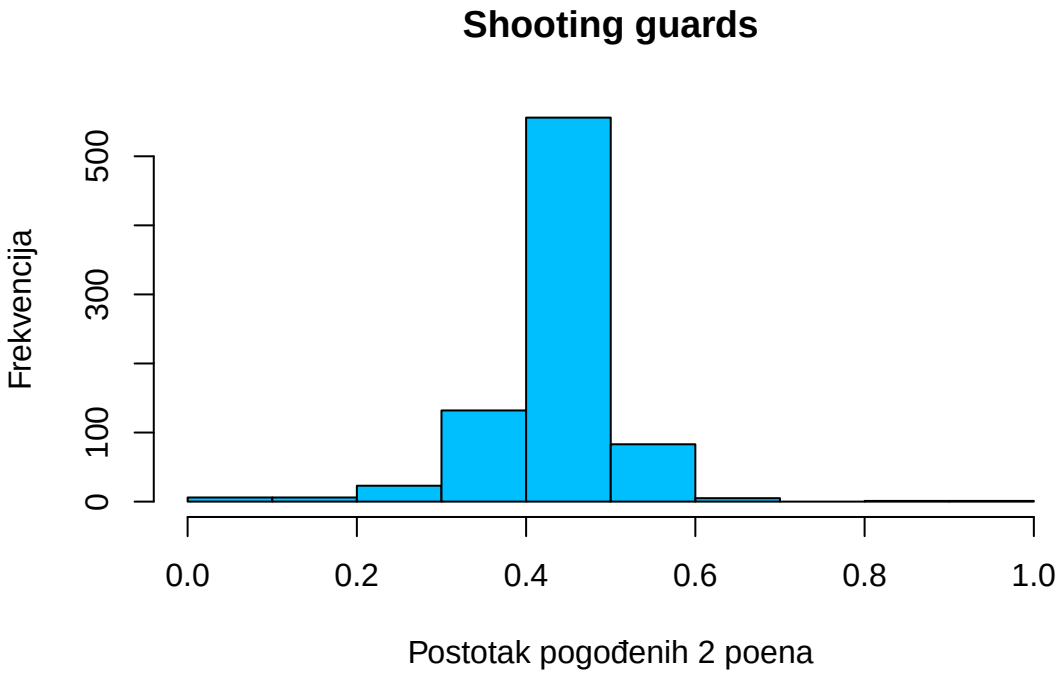
Općenito zanima nas koja je od navedenih pozicija bolji šuter. No, osim toga zanimljivo bi bilo vidjeti je li ta pozicija nužno bolja i za dva i za tri poena. Možda je jedna pozicija uspješnija od druge samo u pogledu trica, ali je zato ova druga bolja za pogotke koji vrijede 2 poena. Pogledajmo najprije kako izgledaju podatci i kojim tehnikama možemo pribjeći.

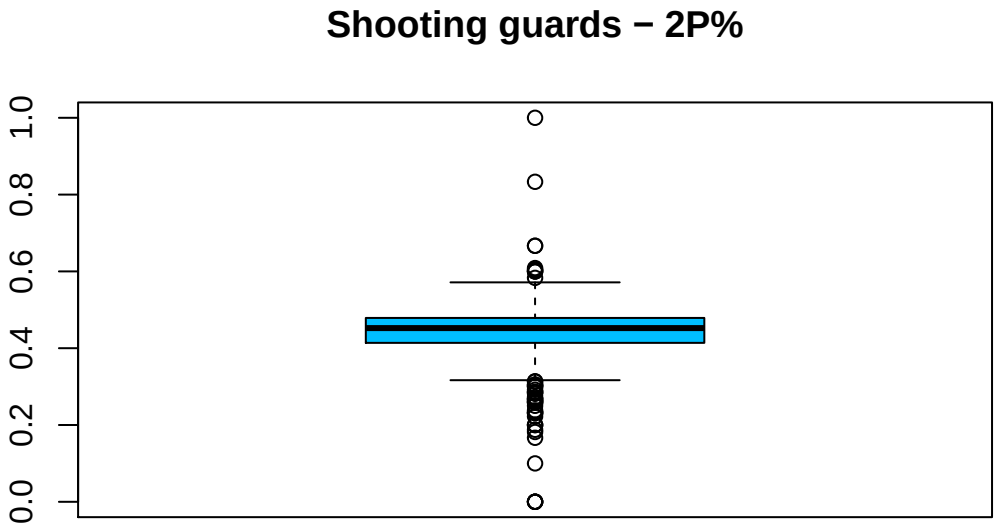
$Q-Q$ plot i histogram za postotak pogođenih 2 point šuta za 'Center' i 'Shooting Guard':

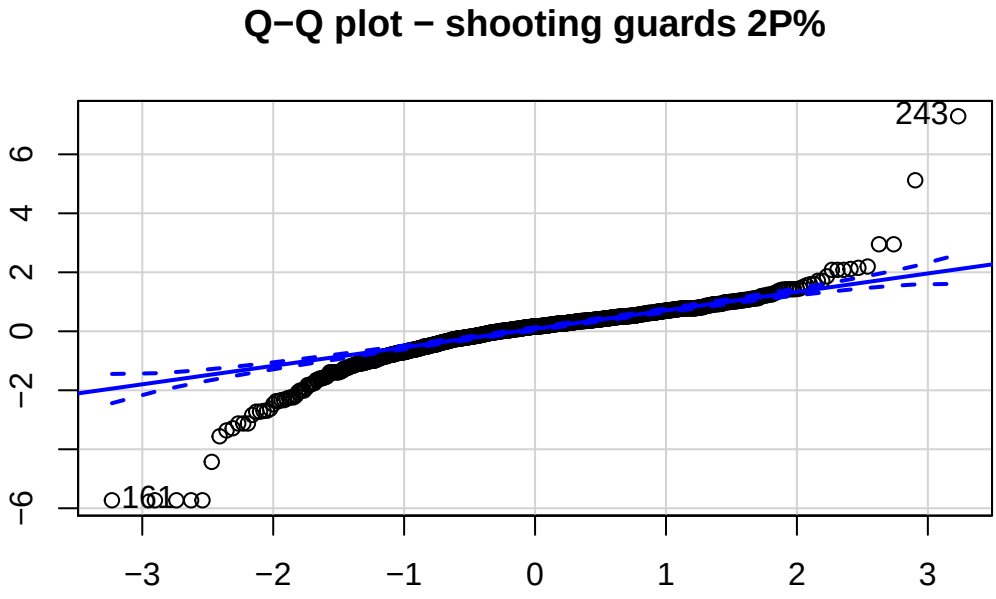


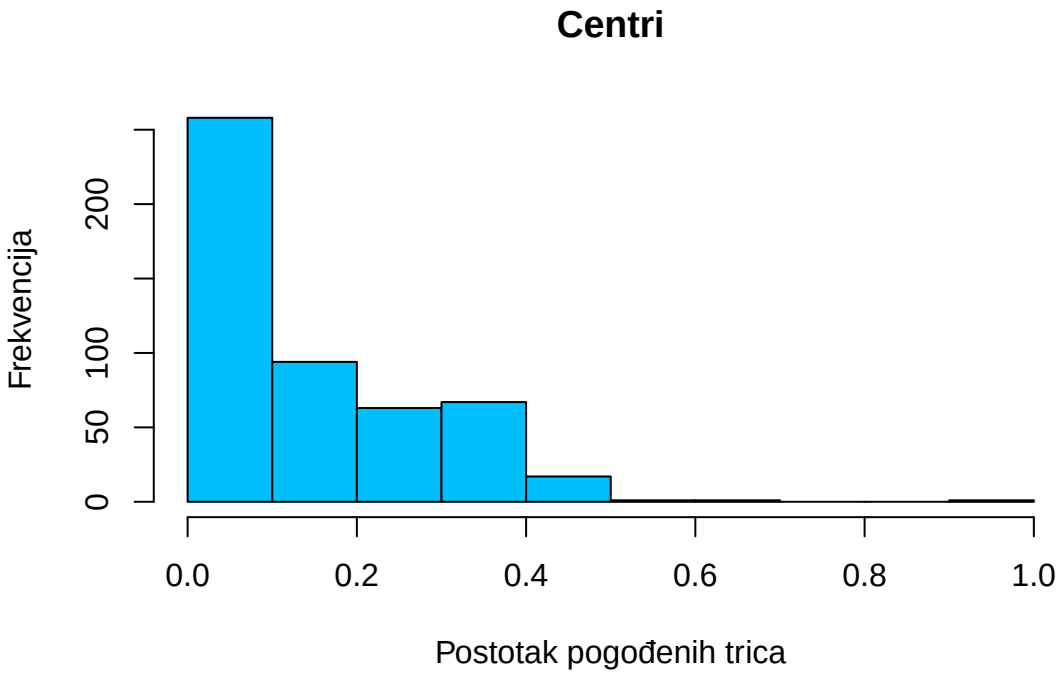


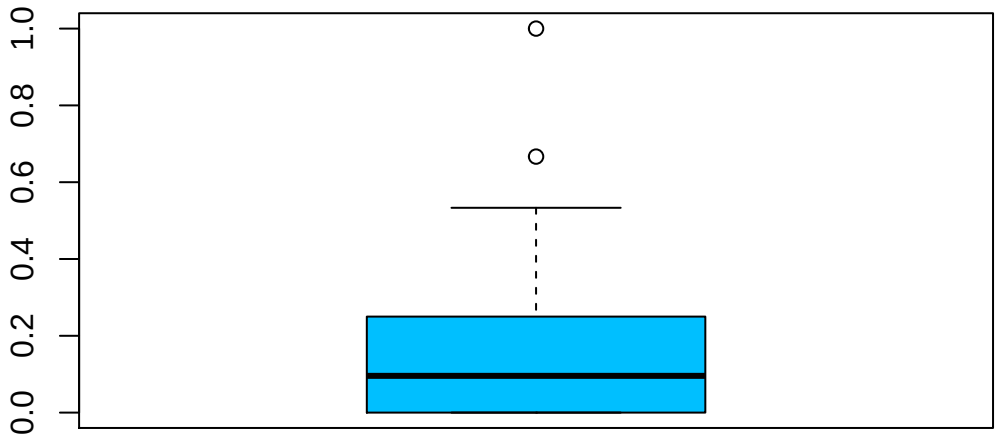

Q−Q plot − centri 2P%
223
85
2
0
−2
−4
−3
−2
−1
0
1
2
3
Kvantili normalne distribucije

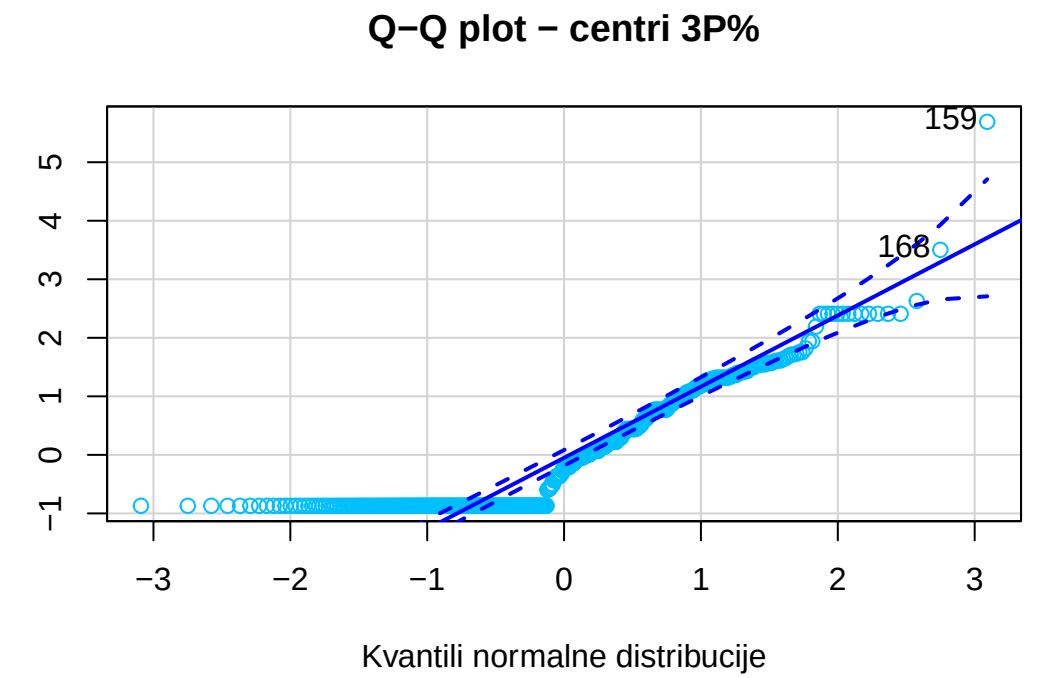

Shooting guards
Frekvencija
0
100
300
500
0.0
0.2
0.4
0.6
0.8
1.0
Postotak pogođenih 2 poena

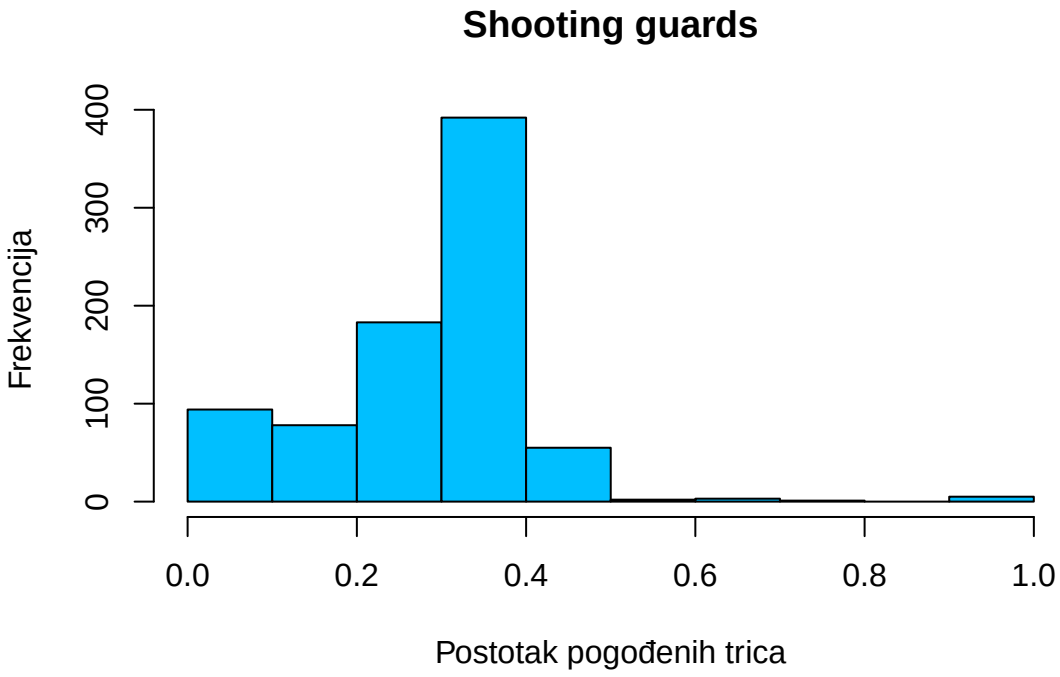

Shooting guards − 2P%
0.0
0.2
0.4
0.6
0.8
1.0

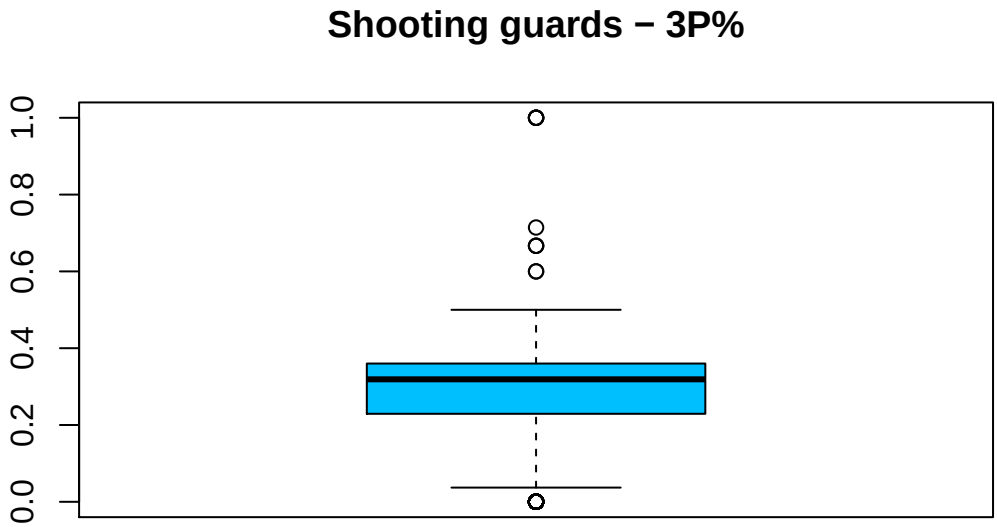

Q−Q plot − shooting guards 2P%
243
161
−6
−2
0
2
4
6
−3
−2
−1
0
1
2
3
Kvantili normalne distribucije

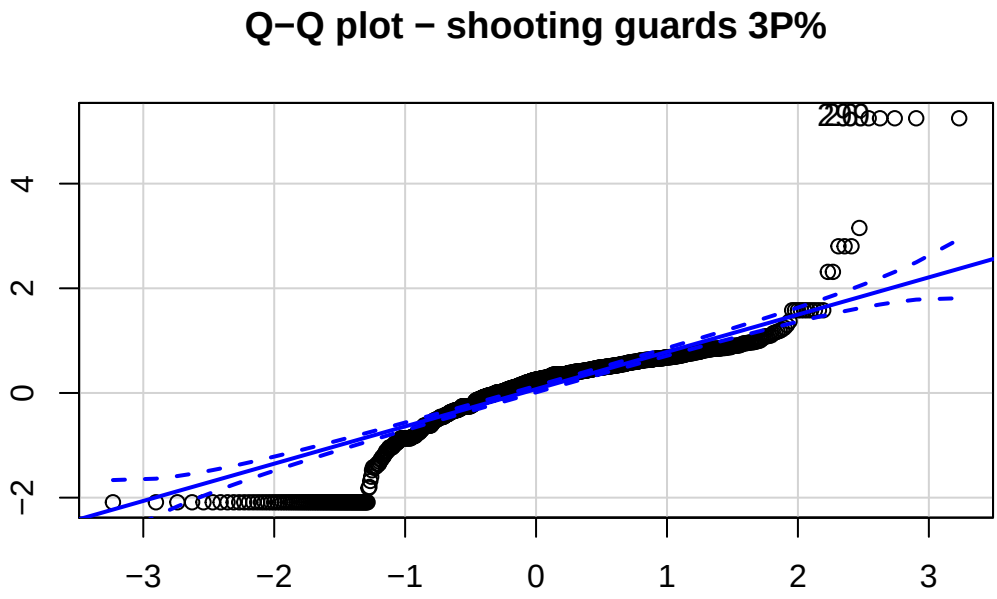

Centri
Frekvencija
0
50
100
200
0.0
0.2
0.4
0.6
0.8
1.0
Postotak pogođenih trica


Centri − 3P%
0.0
0.2
0.4
0.6
0.8
1.0


Q−Q plot − centri 3P%
159
168
−1
0
1
2
3
4
5
−3
−2
−1
0
1
2
3
Kvantili normalne distribucije


Shooting guards
Frekvencija
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
Postotak pogođenih trica


Shooting guards − 3P%
0.0
0.2
0.4
0.6
0.8
1.0


Q−Q plot − shooting guards 3P%
−2
0
2
4
−3
−2
−1
0
1
2
3
Kvantili normalne distribucije

Iz prikazanih grafova možemo zaključiti da parametarski t-test možemo koristiti samo za postotak realizacije 2 point šuta. Stoga, pogledajmo najprije uz pomoć f-testa možemo li pretpostaviti jednakost varijanci za postotke realizacije 2 point šuta pozicija 'Center' i 'Shooting Guard'. Na koncu, provedimo sam t-test.

```
##
##  Welch Two Sample t-test
##
## data:  centers$X2PPG and s.guards$X2PPG
## t = 12.758, df = 1246, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.04208705        Inf
## sample estimates:
## mean of x mean of y
## 0.4884182 0.4400964
```

Zaključujemo kako na razini značajnosti od 5% možemo tvrditi da je postotak realizacije 2 point šuta veći kod pozicije 'Center' u odnosu na poziciju 'Shooting Guard'.

U prethodnom koraku rekli smo kako ne možemo koristiti parametarski t-test za postotak realizacije 3 point šuta jer ne vrijedi pretpostavka normalnosti podataka. No, možemo si pomoći koristeći neparametarski test Bootstrap.

```
## Interval povjerenja jednak je: [0.13832, Inf].
```

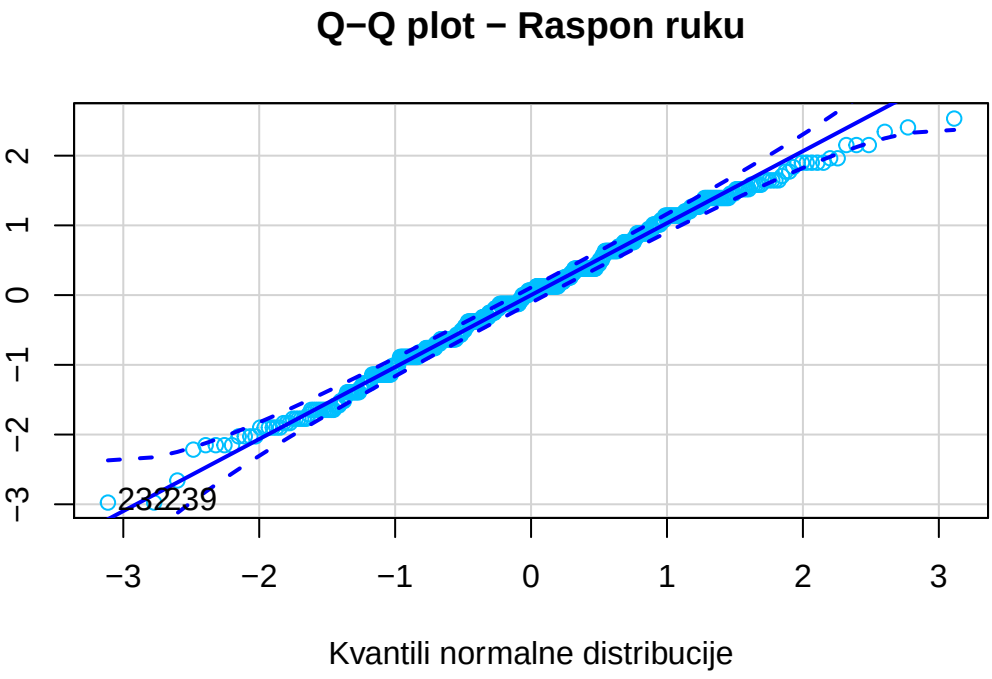
```
## [1] 0.2846649
```

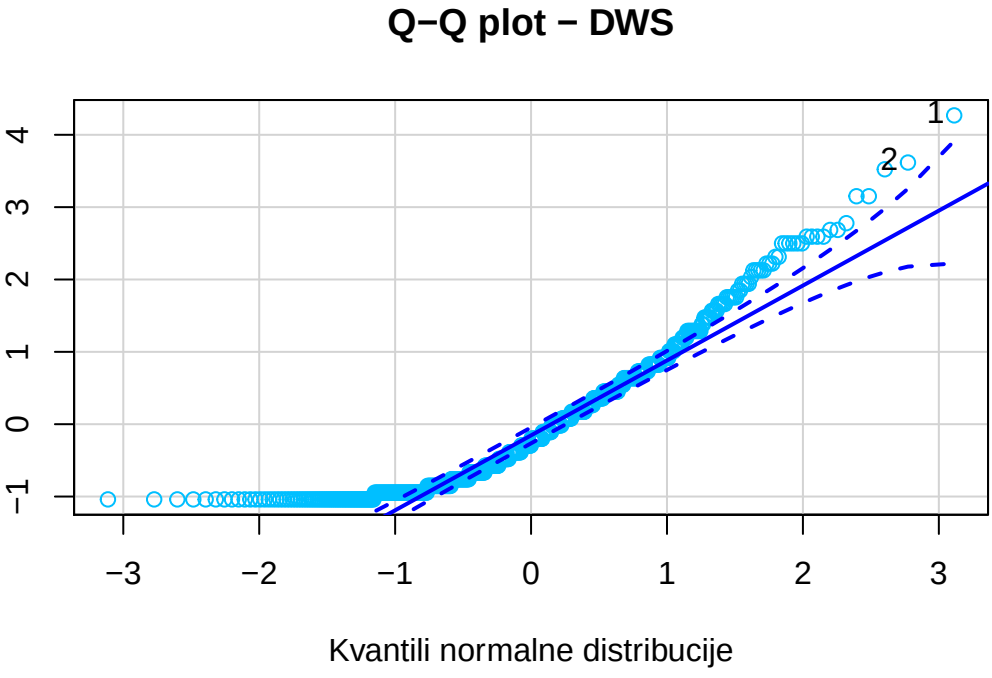
```
## [1] 0.1326266
```

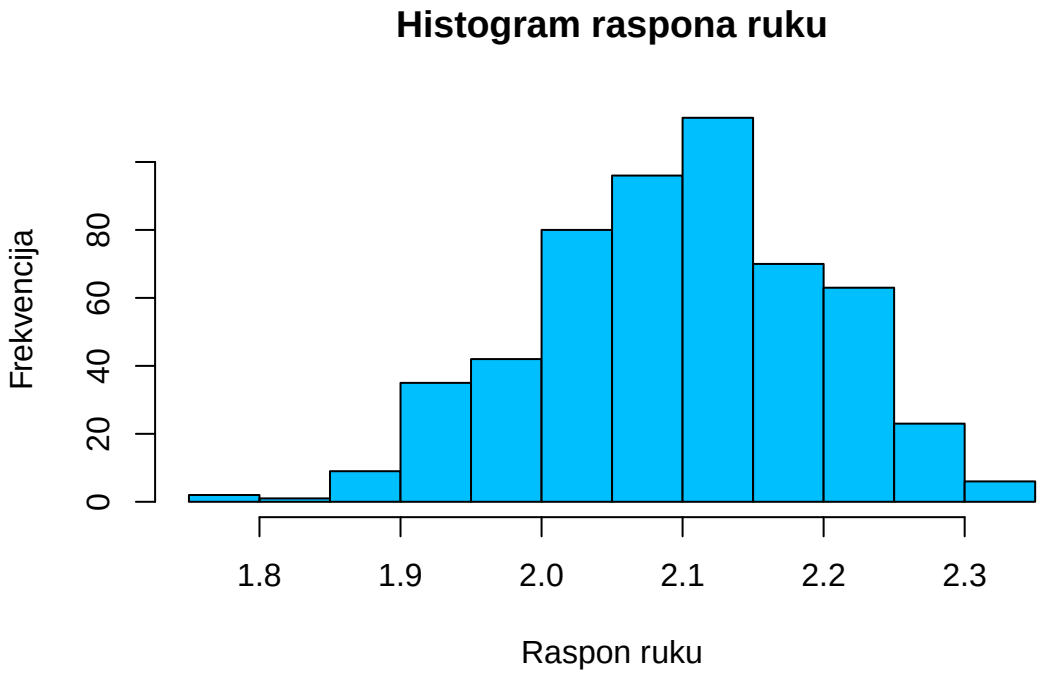
Nakon provedenog testa vidimo da 0 'nije upala' u interval povjerenja pa možemo uz razinu značajnosti od 5% tvrditi kako pozicija 'Shooting guard' zaista ima veći postotak realizacije 3 point šuta u odnosu na poziciju 'Center'.

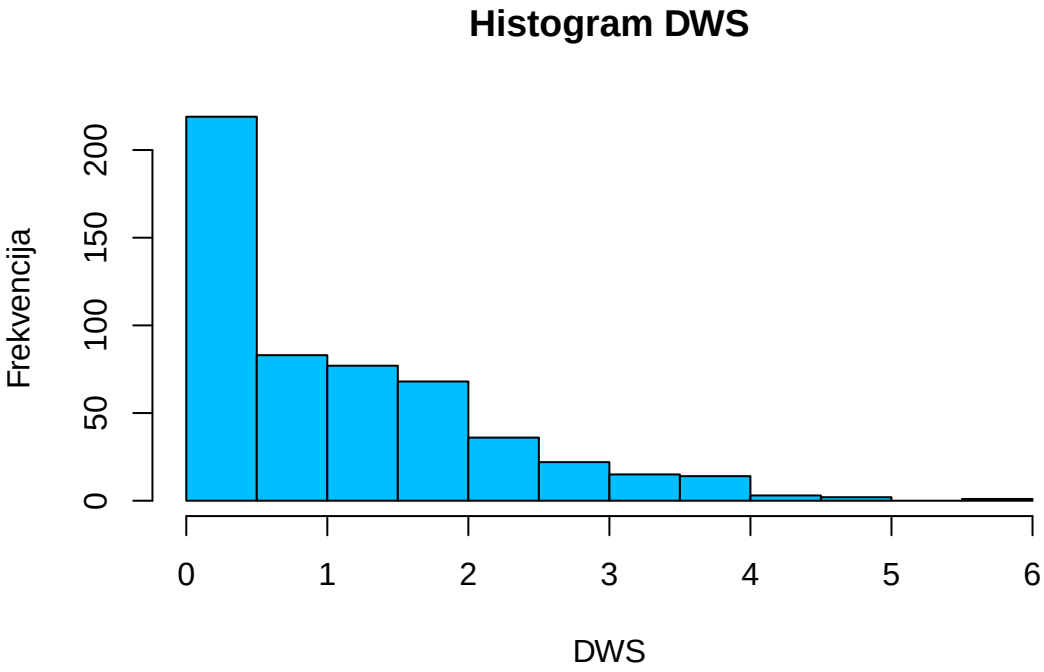
4. Utječe li širina raspona ruku na efikasnost u obrani?

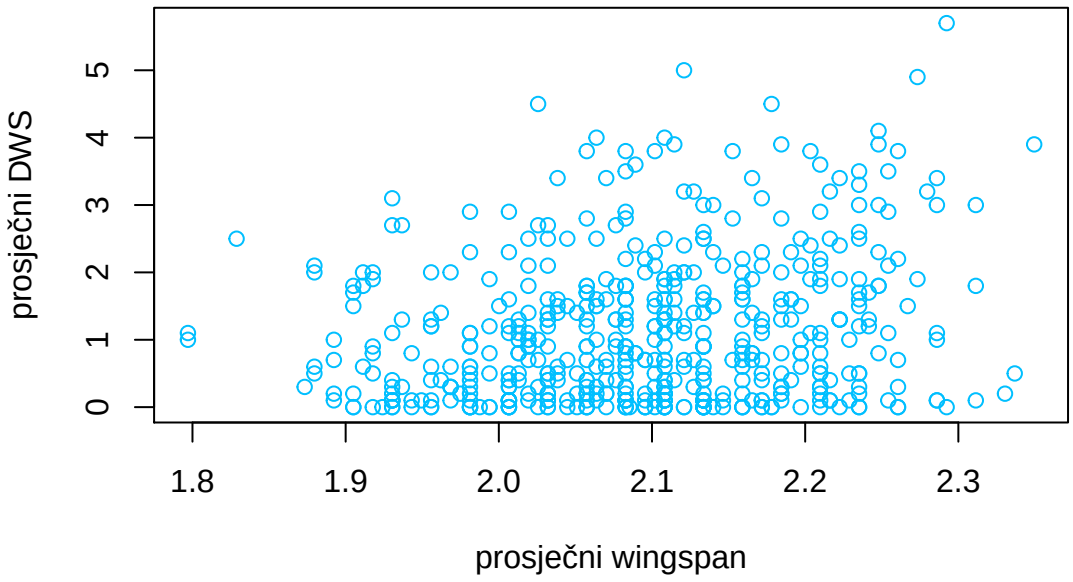
Kako bismo odgovorili na ovo pitanje moramo odrediti kako ćemo definirati efikasnost obrane. NBA statistika nam ovdje značajno pomaže sa podatkom DWS (Defensive Win Shares). To je složeni podatak koji uzima u obzir nekoliko ostalih podataka kao što su broj pogodaka, posjed lopte te neke timske podatke te na osnovu njih računa vlastitu vrijednost. Pokušat ćemo linearnom regresijom provjeriti jesu li ova dva podatka (wingspan i DWS) linearno povezani.






Histogram raspona ruku
Frekvencija
0
20
40
60
80
1.8
1.9
2.0
2.1
2.2
2.3
Raspon ruku


Histogram DWS
Frekvencija
0
50
100
150
200
0
1
2
3
4
5
6
DWS



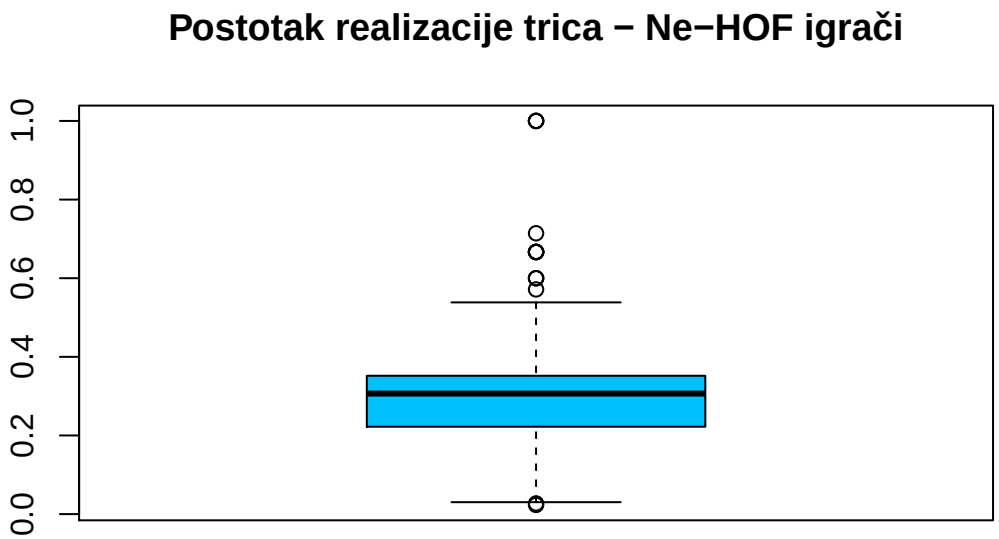

```
##
## Call:
## lm(formula = wingspan$DWS ~ wingspan$Wingspan.m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5751 -0.8272 -0.2513  0.5551  4.1249
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.7707     0.9451  -3.990 7.53e-05 ***
## wingspan$Wingspan.m   2.3320     0.4505   5.177 3.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.049 on 538 degrees of freedom
## Multiple R-squared:  0.04745,    Adjusted R-squared:  0.04568
## F-statistic:  26.8 on 1 and 538 DF,  p-value: 3.195e-07
```

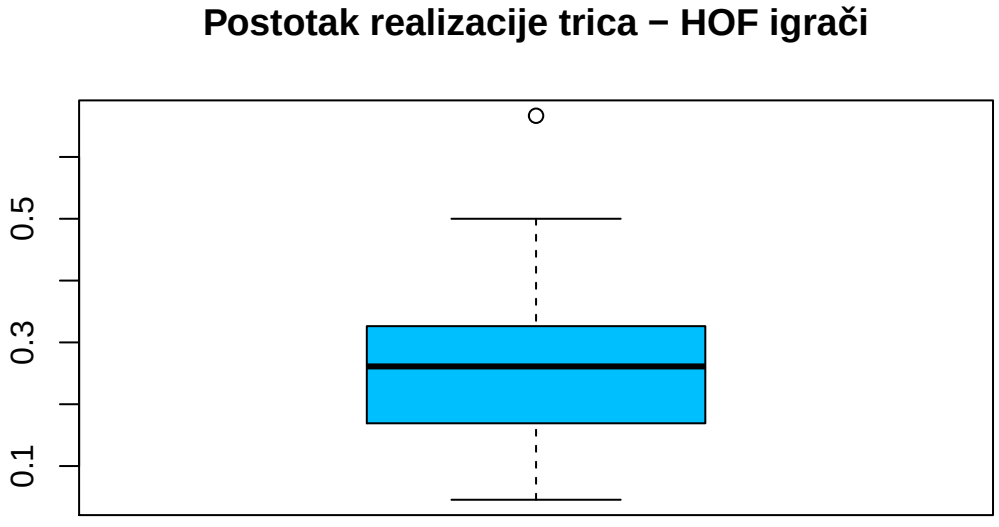
Na osnovu dobivenih rezultata zaključujemo da navedene podatke ne možemo nikako povezati. Kada malo bolje razmislimo, veći raspon ruku omogućuje veću pokrivenost terena oko sebe, međutim, on također i unosi potencijalne negativne učinke kao što su tromost, manju ravnotežu i sl. Zbog toga ova statistika i nije toliko začuđujuća.

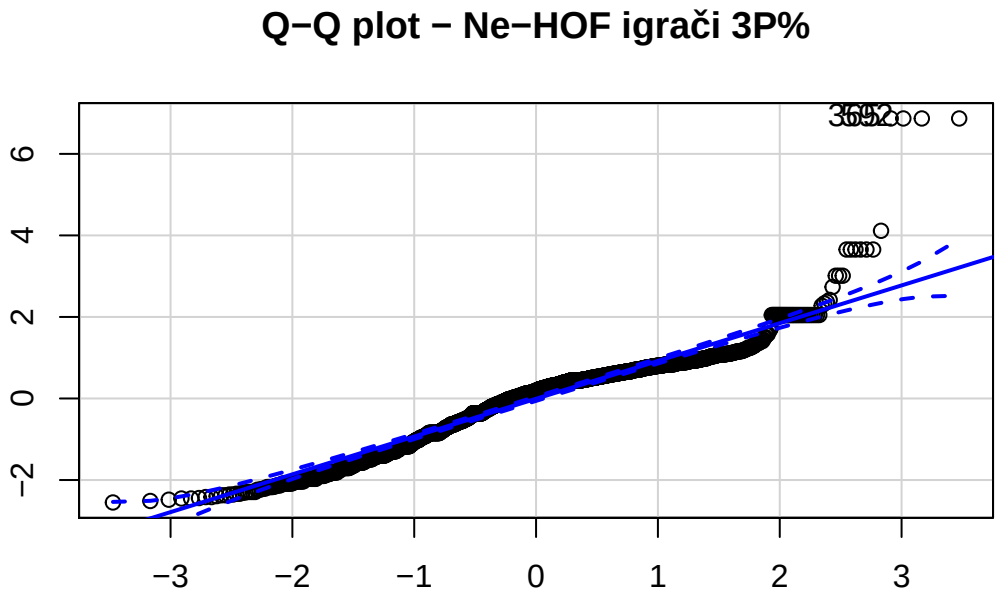
5. Hall of Fame igrači - najbolji baš u svemu?

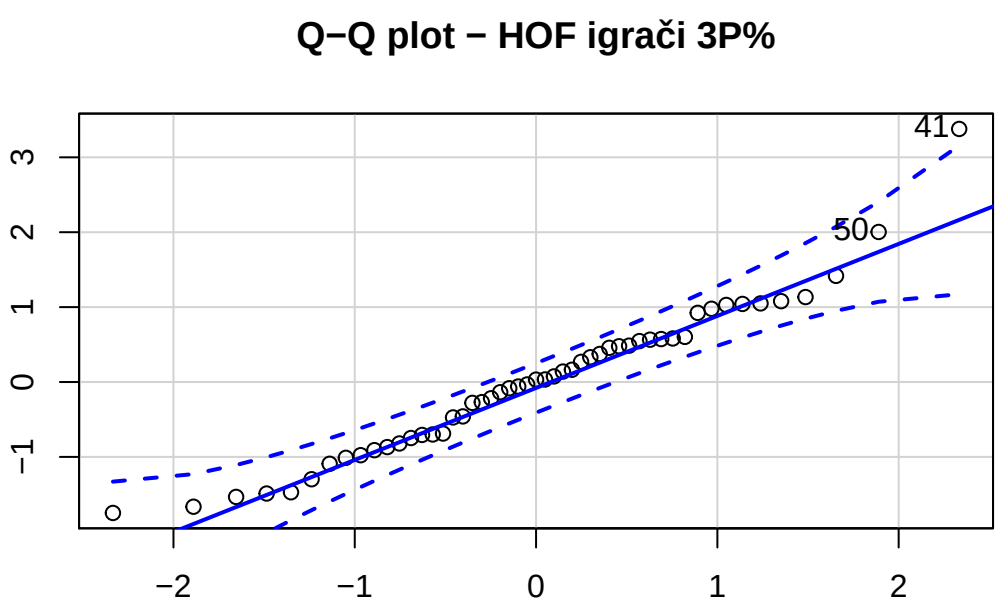
Jedno od zanimljivh pitanja koje se postavlja jest jesu li igrači koji su dospjeli u Košarkašku Kuću slavnih bolji od ostalih igrača. Konkretno ovdje ćemo se osvrnui na postotak realizacija trica. Na prvi pogled, osoba koja nije previše upućena u svijet košarke bi mogla tvrditi kako su slavni igrači zasigurno bolji. Ovim testom pokušati ćemo pokazati baš suprotno. Razdvojit ćemo igrače koji su u Košarkaškoj Kući od onih koji nisu. Najprije ćemo provesti f-test kako bi znali možemo li pretpostaviti jednakost varijanci, zatim crtamo $Q-Q$ plot kojim provjeravamo jesu li zadovoljene pretpostavke o normalnosti populacija te konačno provodimo t-test.

Boxplot te $Q-Q$ plot za igrače koji nisu dospjeli u Košarkašku Kuću slavnih i one koji jesu:






Q−Q plot − Ne−HOF igrači 3P%
Kvantili normalne distribucije


Q−Q plot − HOF igrači 3P%
41
50
Kvantili normalne distribucije

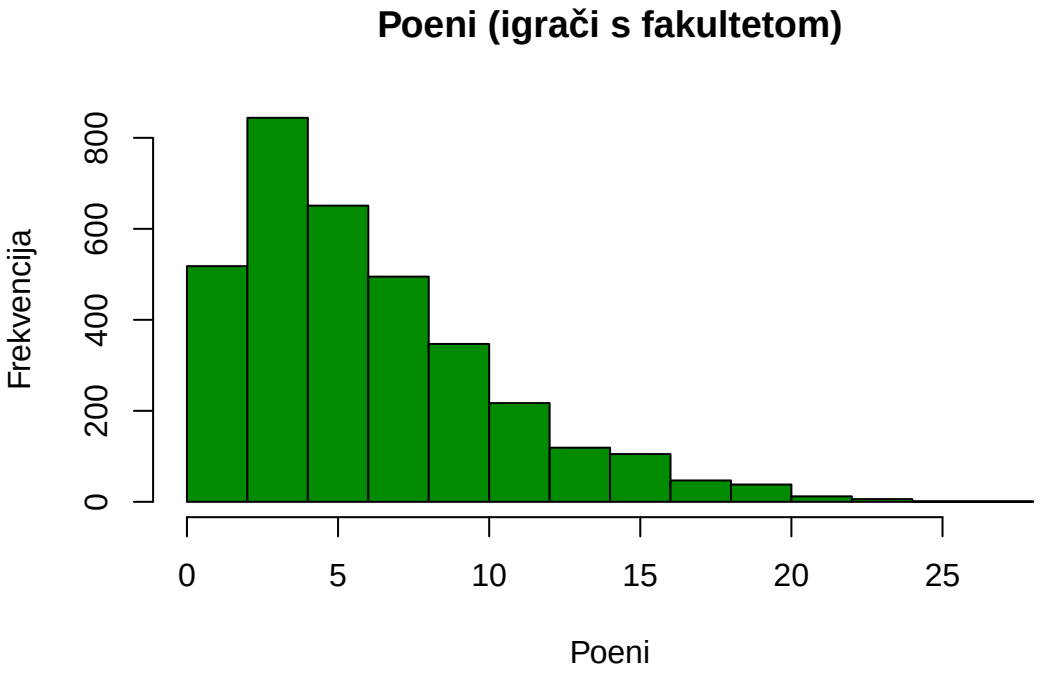
Rezultati t-testa:

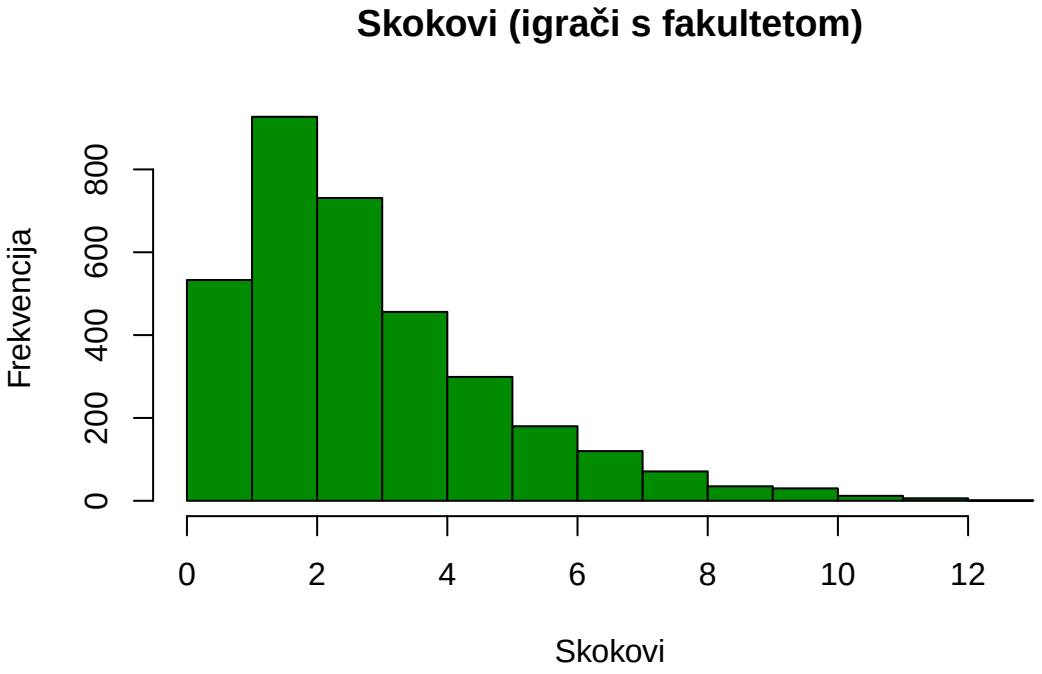
```
##
##  Two Sample t-test
##
## data:  NOR.players$three.p and HOF.players$three.p
## t = 2.0637, df = 1991, p-value = 0.01959
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.006177714         Inf
## sample estimates:
## mean of x mean of y
## 0.2877831 0.2572887
```

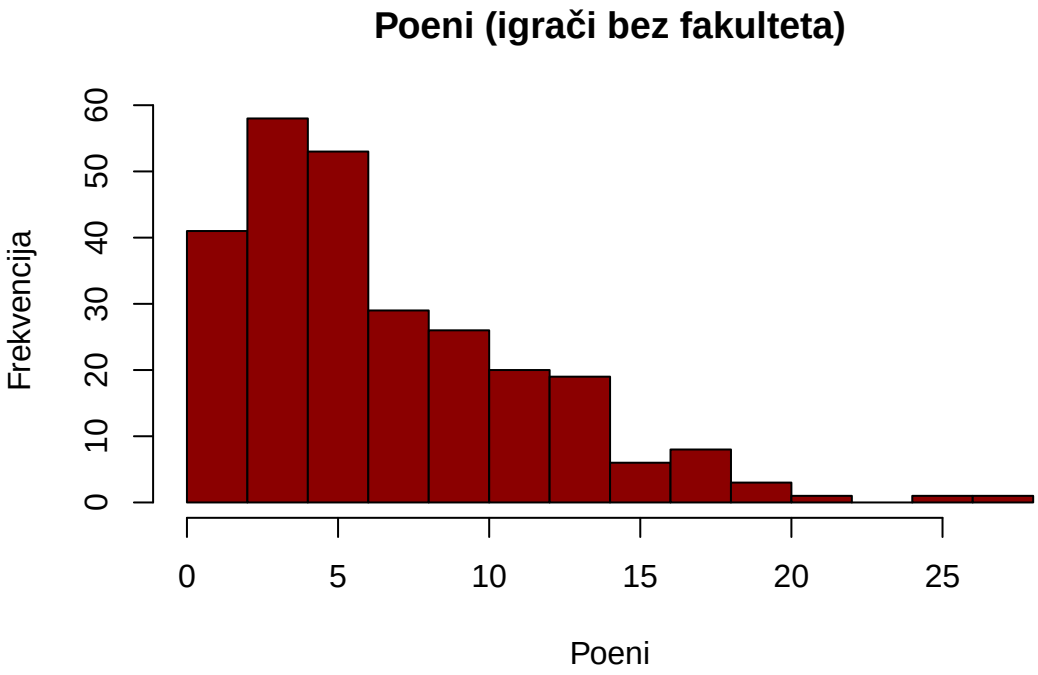
Nakon provedenog t-testa možemo zaključiti na razini značajnosti od 5% kako igrači koji su dospjeli u Košarkašku Kuću slavnih imaju manji postotak pogođenih 3 point šuta u odonsu na igrače koji nisu u Košarkaškoj Kući slavnih.

6. Fakultet prije NBA karijere? Da ili ne?

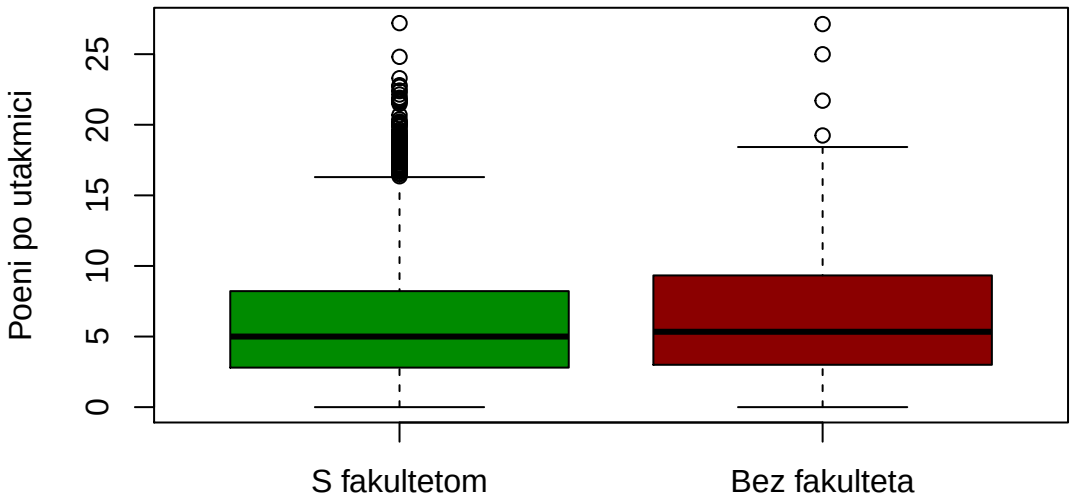
Poznato nam je da mnogi najbolji NBA igrači nisu završili fakultet. Neki od njih su Kobe Brynat, Kevin Garnett, LeBron James, Tracy McGrady te mnogi drugi, stoga bi bilo zanimljivo istražiti jesu li općenito ti igrači bolji od onih koji su se školovali na fakultetima. Uglavnom, isplati li se budućim NBA igračima završiti fakultet ili odmah započeti NBA karijeru.Najprije ćemo vizualizirati podatke kako bi dobili ideju što možemo učiniti po pitanje provjere jesu li doista igrači bez završenog fakulteta bolji od onih koji su ga završili.

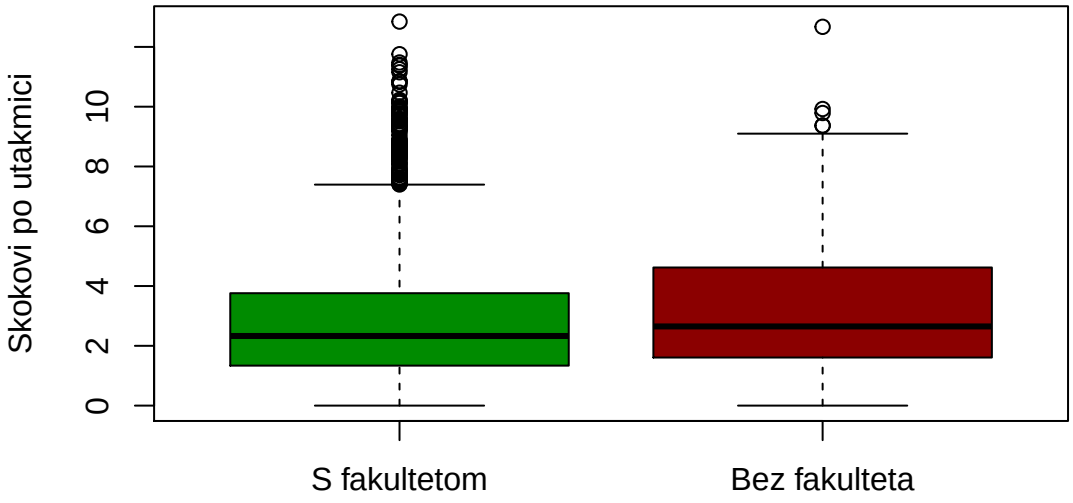


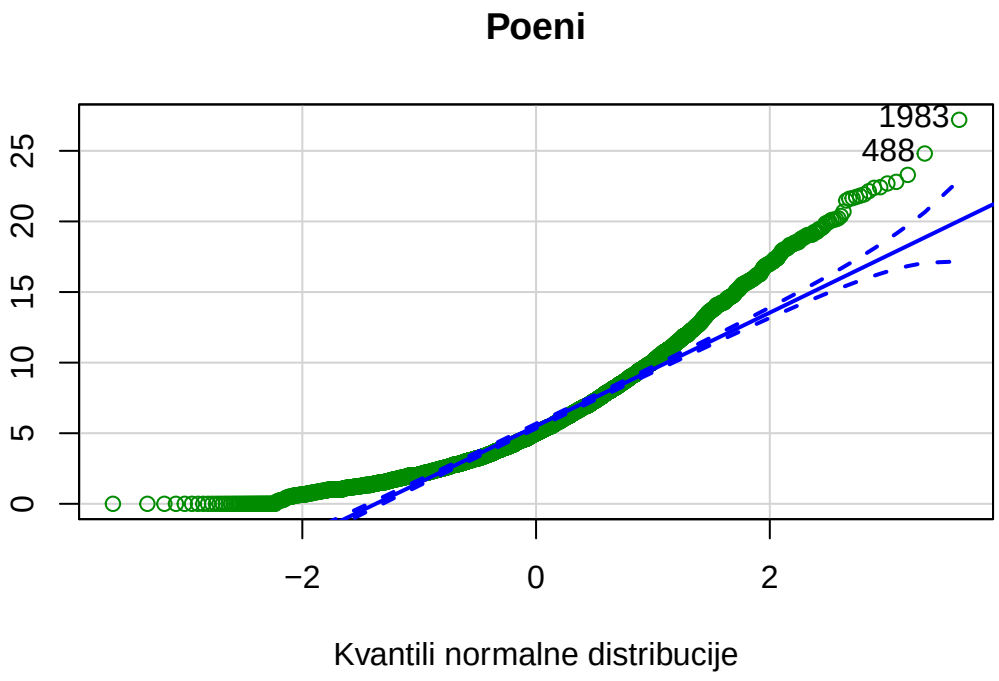


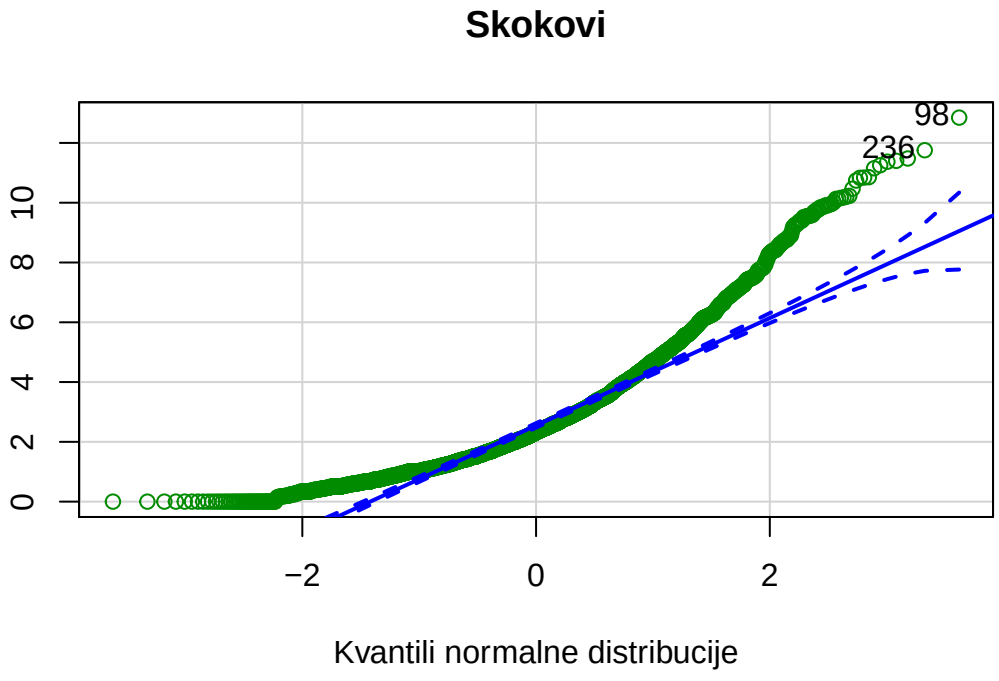

Poeni (igrači bez fakulteta)
Frekvencija
0
10
20
30
40
50
60
0
5
10
15
20
25
Poeni


Skokovi (igrači bez fakulteta)
Frekvencija
0
10
20
30
40
50
0
2
4
6
8
10
12
Skokovi

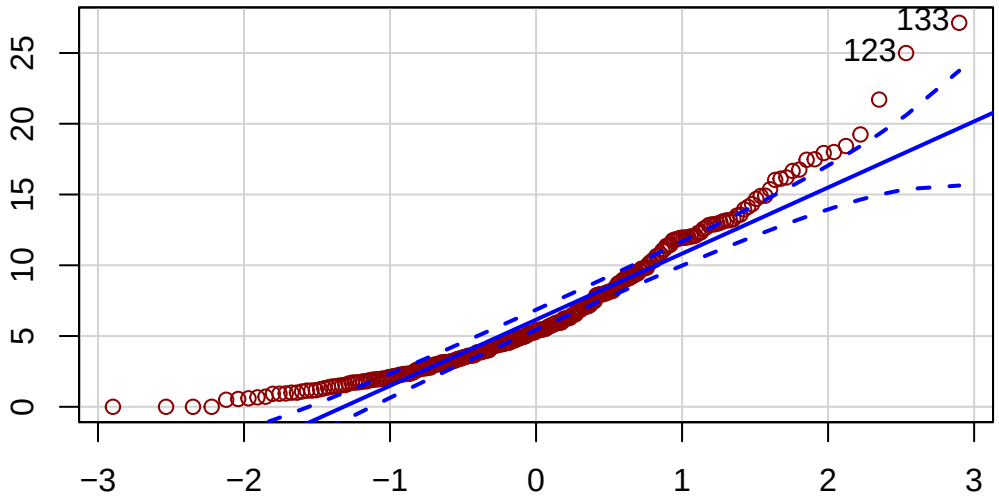

Poeni po utakmici
0
5
10
15
20
25
S fakultetom
Bez fakulteta


Skokovi po utakmici
0
2
4
6
8
10
S fakultetom
Bez fakulteta

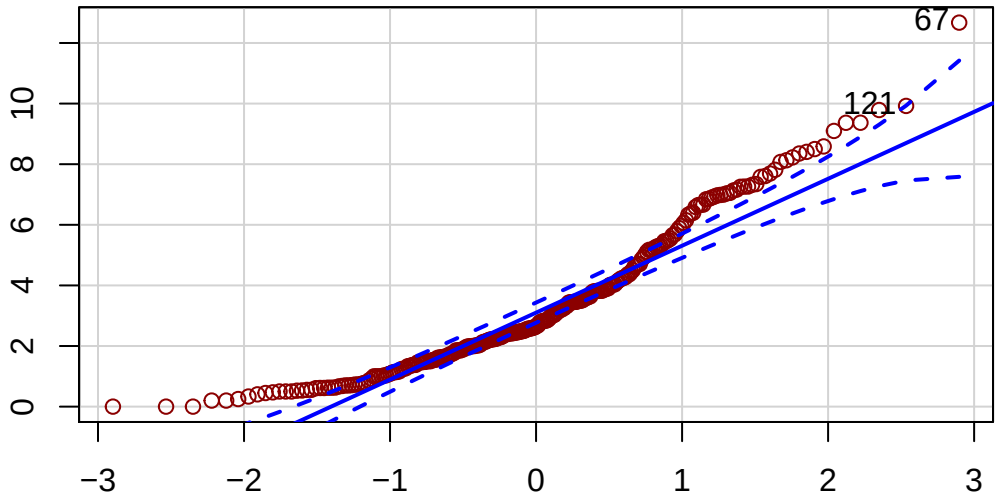

Poeni
1983
488
0
5
10
15
20
25
−2
0
2
Kvantili normalne distribucije


Skokovi
98
236
0
2
4
6
8
10
−2
0
2
Kvantili normalne distribucije

Poeni



Skokovi



Vidimo da su igrači bez završenog fakulteta ipak malo bolji i to i u broju poena i broju skokova po utakmici. Također, vidimo da podatci nisu normalno distribuirani pa ćemo provesti Bootstrap metodu testiranja s intervalom pouzdanosti. Konkretno, testirati ćemo hipotezu da su obje populacije igrača jednako dobre (i u broju poena i u skokovima) nasuprot alternative da su igrači bez završenog fakulteta bolji. Zbog zakrivljenosti distribucija umjesto srednje vrijednosti koristit ćemo medijan jer je on u tom slučaju bolji pokazatelj sredine distribucije.

```
## Bootstrap test za poene...

## Interval povjerenja jednak je: [-0.21237, Inf].

##
## Bootstrap test za skokove...

## Interval povjerenja jednak je: [0.12138, Inf].
```

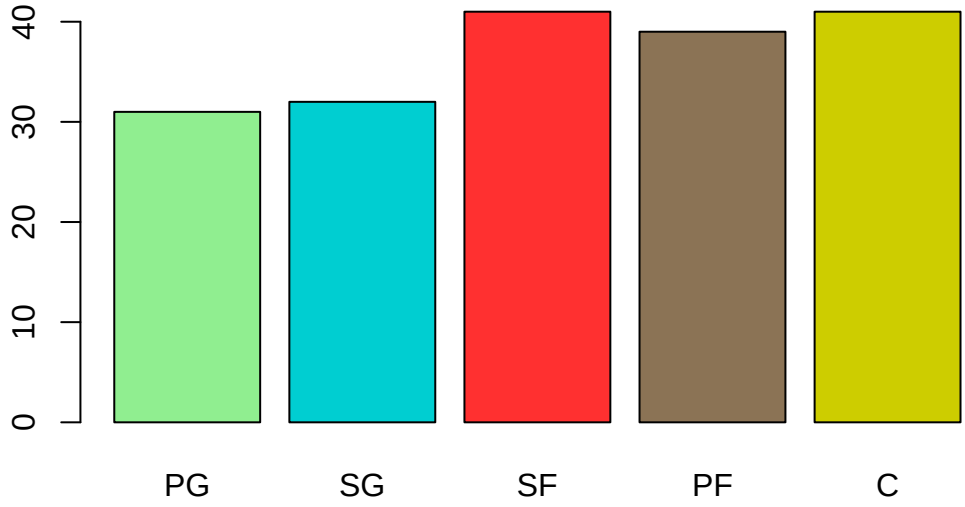
Uz pomoć Bootstrap metode možemo na razini značajnosti $\alpha = 0.05$ tvrditi kako su igrači bez završenog fakulteta bolji samo u broju skokova po utakmici. Ipak, valja primijetiti kako je razlika izrazito mala i kako možda nije dovoljno značajna da bi donosili nekakve čvrste, kategoričke odluke. Uprkos tome, test ipak ima značaj jer bi imalo smisla tvrditi da su igrači koji su se dodatno obrazovali na fakultetu bolji zbog dodatnog stečenog znanja, a mi smo s testiranjem pokazali ne samo da su jednaki kao oni bez završenog fakulteta nego čak i lošiji.

7. Jesu li pozicije igrača podjednako raspodijeljene u Kući slavnih?

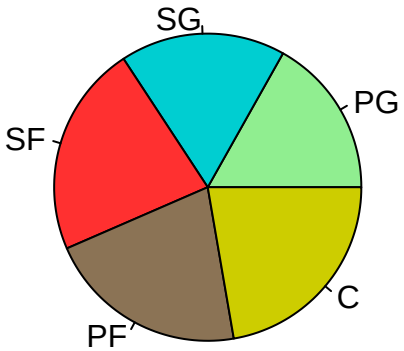
Jesu li se pozicije igrača koji su dospjeli u Košarkašku Kuću slavnih ravnomjerno raspodijeljene? Točnije, zanima nas jesu li određene pozicije popularnije u smislu da igrači koji igraju na tim pozicijama ostvaruju takve rezultate i uspjehe da to u konačnici rezultirati ulaskom u Košarkašku Kuću slavnih. Koristit ćemo χ^2-test odnosno test prilagodbe modela podatcima.

Rezultati χ^2-testa te raspodjela Hall of Fame igrača po pozicijama:

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 1.3484, df = 4, p-value = 0.8531
```



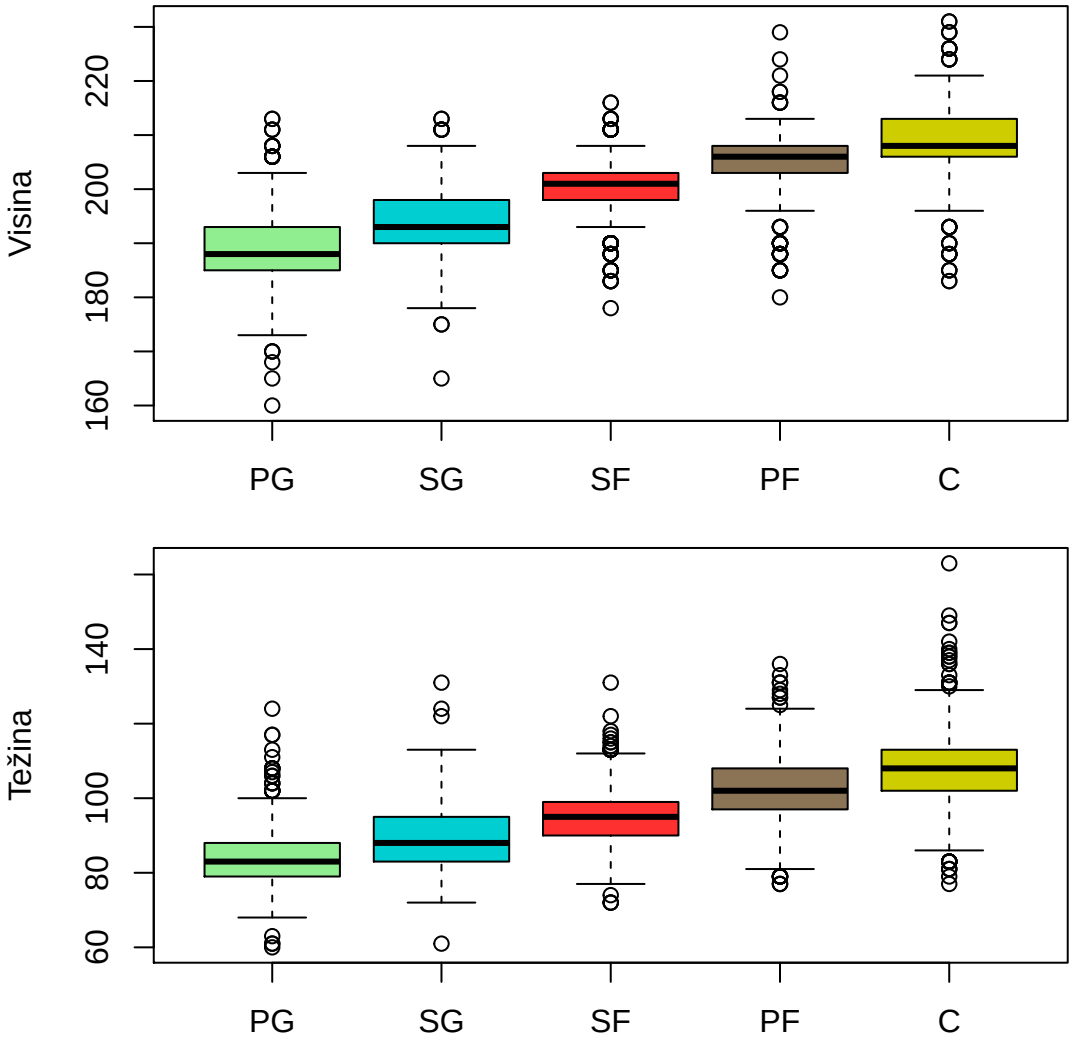
Raspodjela Hall of Fame igrača po pozicijama



8. Vizualizacija fizičkih karakteristika igrača po pozicijama

U prošlom dijelu smo testirali je li raspodjela pozicija igrača koji su dospjeli u Košarkašku kuću slavnih ravnomjerna. Sada ćemo promatrati fizičke karakteristike igrača kod pojedinih pozicija. Preciznije, zanima nas kakva je distribucija visina i težina za pojedine pozicije. To ćemo ilustrirati na 2 dijagrama (jedan za visinu, drugi za težinu) gdje će istovremeno biti 5 pravokutnih dijagrama s izdancima - za svaku poziciju po jedan.

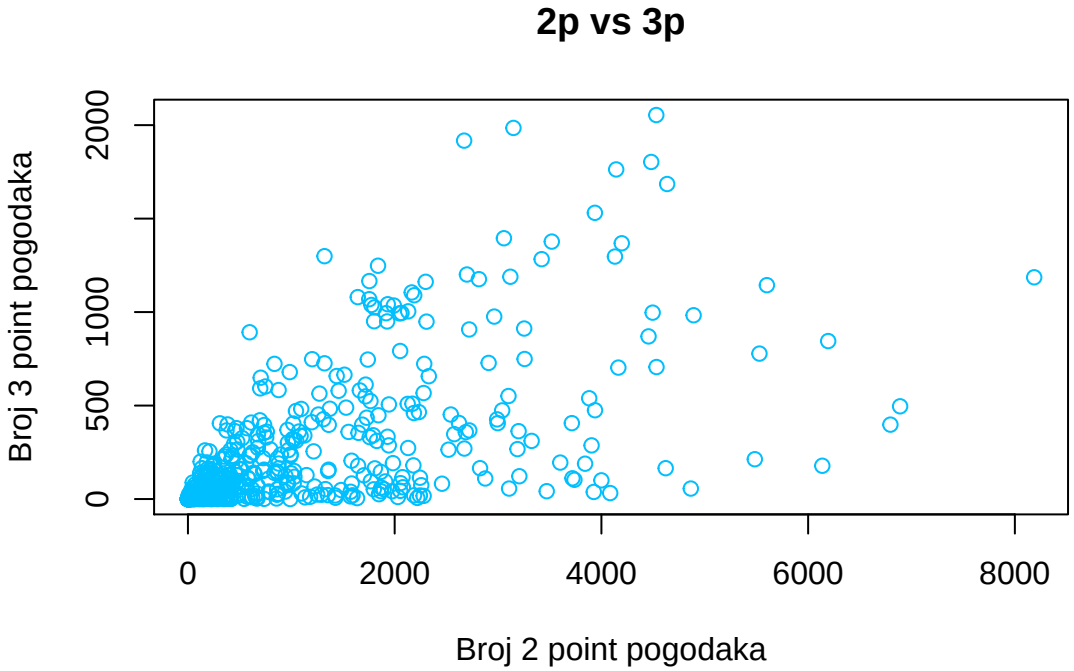
```
## Joining, by = "Player"
```



9. Postoji li povezanost između broja realiziranih šuteva koji nose 2 i 3 poena?

Konkretno, zanima nas postoji li čvrsta LINEARNA povezanost u smislu da oni igrači koji imaju više pogodaka za 2 poena nužno imaju i više pogođenih trica.

Najprije ćemo vizualizirati podatke kako bi vidjeli postoji li mogućnost povezivanja varijabli linearnim modelom.



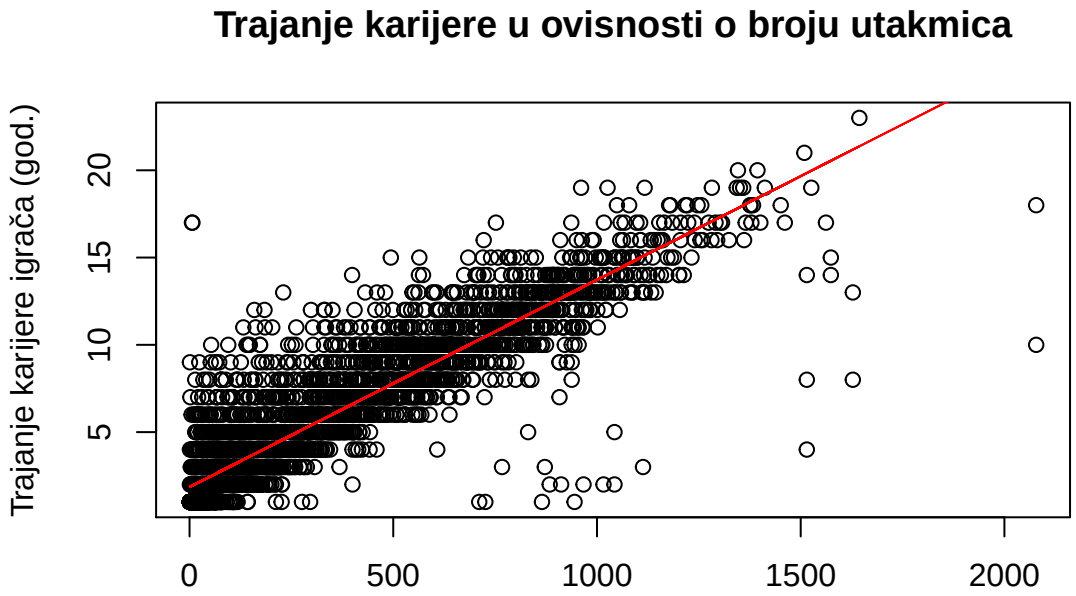
Kao što vidimo podatci su prilično "rasprešni" i možemo reći da ne postoji pretjerana (linearna) povezanost između broja 2 point pogodaka i trica. No, u to se možemo i dodatno uvjeriti pomoću linearne regresije. Rezultati regresije na sljedećoj stranici:

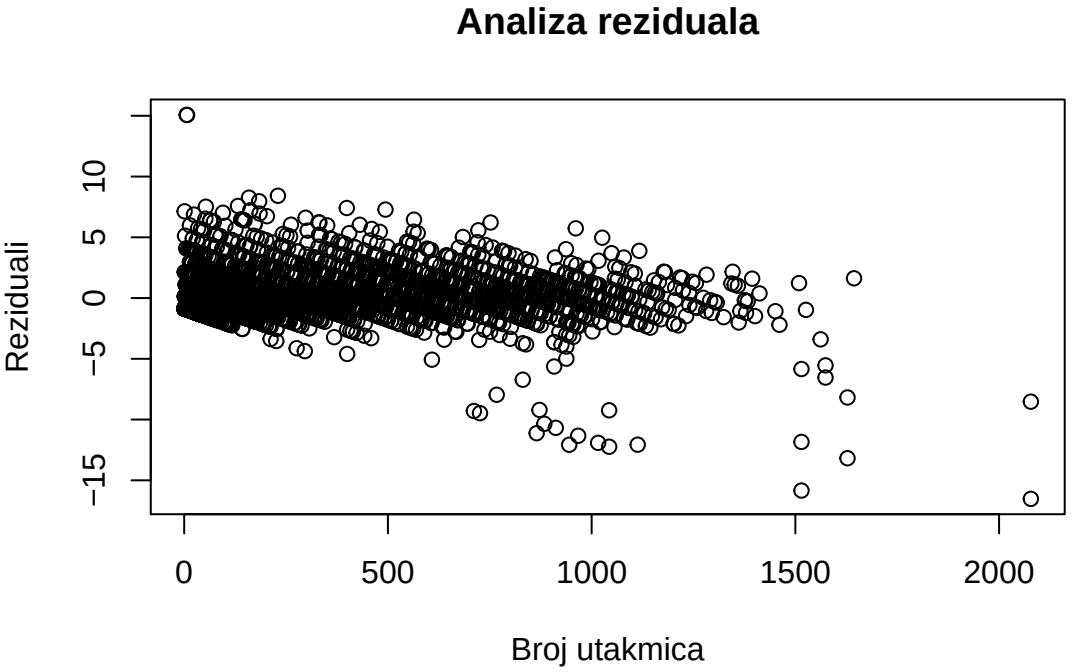
```
##
## Call:
## lm(formula = players$X3sum ~ players$X2sum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -922.74  -63.79  -42.20   23.79 1413.46
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   43.023123  12.244647   3.514 0.000473 ***
## players$X2sum  0.172350   0.008097  21.287  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 253.8 on 635 degrees of freedom
## Multiple R-squared:  0.4164, Adjusted R-squared:  0.4155
## F-statistic: 453.1 on 1 and 635 DF,  p-value: < 2.2e-16
```

Linearna korelacija je premala kako bi mogli govoriti o čvrstoj povezanosti između šuta za 2 i 3 poena što nam dodatno potvđuje ono što smo i nasulutili iz grafa.

10. Povezanost duljine karijere i broja utakmica igrača.

Cilj ovog zadatka je konstruirati dovoljno dobar model koji će na osnovu broja utakmica igrača predvidjeti koliko dugo je trajala karijera igrača. Ideja je sljedeća: najprije ćemo podijeliti početni skup igrača na onaj za učenje i ispitavanje (provjere točnosti) modela. Zatim ćemo pristupu konstrukcije modela pristupiti na 2 načina: linearnom regresijom te strojnim učenjem odnosno pomoću umjetnih neuronskih mreža.





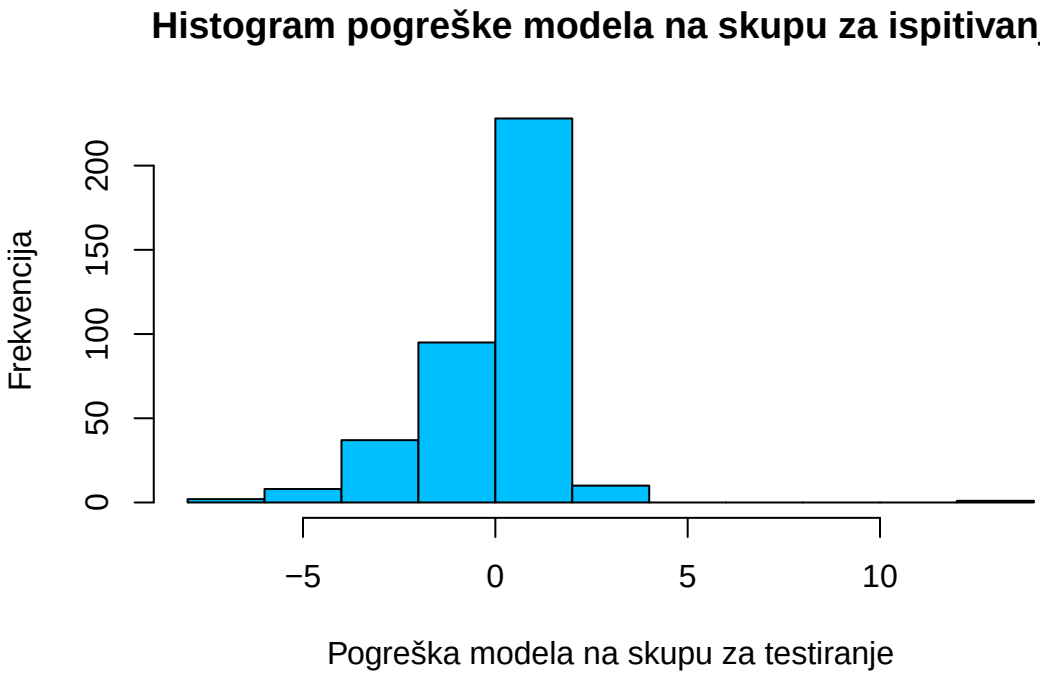
Već pri analizi samih grafova možemo zaključiti kako postoji dosta čvrsta veza između broja utakmica i trajanja karijere igrača. Početnu hipotezu potvrđuje i relativno visok Pearsonov koeficijent korelacije (81.13%) što je relativno zadovoljavajuće s obzirom na stvarne podatke.

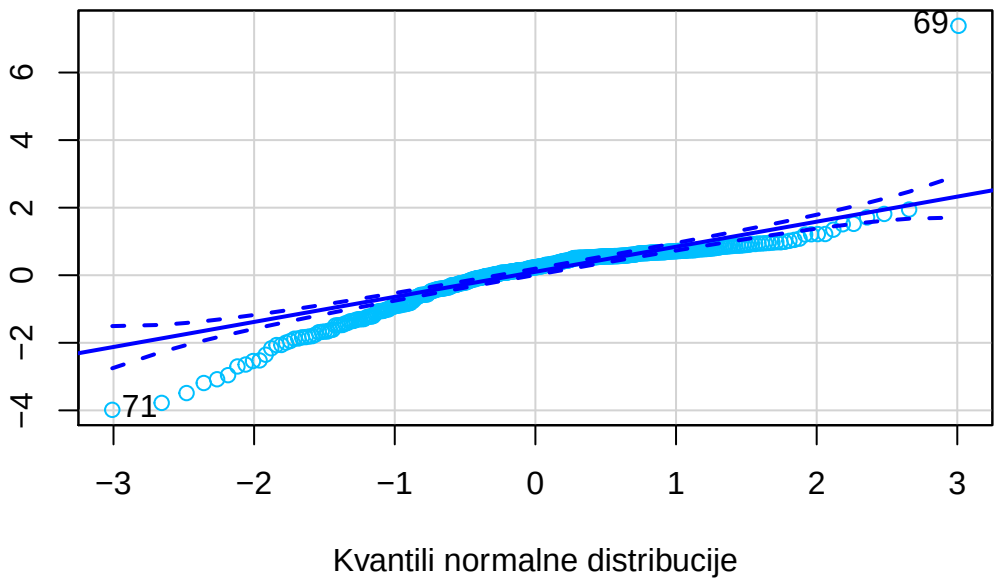
Rezultati regresije:

```
##
## Call:
## lm(formula = all.players.train$duration ~ all.players.train$Games)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5263  -1.0290  -0.4446   0.8022  15.0779
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.851e+00  4.429e-02   41.79   <2e-16 ***
## all.players.train$Games 1.187e-02  9.776e-05  121.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.918 on 3430 degrees of freedom
## Multiple R-squared:  0.8114, Adjusted R-squared:  0.8113
## F-statistic: 1.476e+04 on 1 and 3430 DF,  p-value: < 2.2e-16
```

Nakon provedenih t-testova za odječak na osi y - β_0 te koeficijenta smjera - β_1 možemo zaključiti kako su oba koeficijenta različita od 0 ($p - vrijednost < 10^{-10}$). $\beta_0 = 1.8508$ $\beta_1 = 0.0119$

Točnost modela provjeravamo na skupu za ispitivanje. Kao mjeru pogreške uzet ćemo MSE (engl. Mean Square Error) odnosno srednju kvadratnu pogrešku. Ispada da je navedena mjera jednaka 2.73 godine što je 14.4% od ukupnog raspona trajanja karijera igrača u skupu za ispitivanje.


Histogram pogreške modela na skupu za ispitivanje
Frekvencija
0
50
100
150
200
−5
0
5
10
Pogreška modela na skupu za testiranje


Q−Q plot − distribucija pogrešaka na skupu za testiranje
69
71
−4
−2
0
2
4
6
−3
−2
−1
0
1
2
3
Kvantili normalne distribucije

Konačno, kako bi pokušali nači optimalnu funkciju preslikavanja između broja utakmica i trajanja karijere, problem smo pokušali riješiti koristeći umjetne neuronske mreže. Konkretno, prilagođavanje težina u neuronskim mrežama izvedeno je genetskim populacijskim algoritmom sa sljedećim parametrima:
- veličina populacije: 30
- broj elitnih jedinki: 5
- vjerojatnost mutacije: 5%
- broj iteracija: 10 000

Što se tiče same arhitekture mreže, sastoji se od 3 skrivena sloja:
1. ulazni sloj (1 neuron - broj utakcima)
2. potpuno povezan sloj (7 neurona)
3. aktivacijska funkcija (ReLU)
4. potpuno povezan sloj (5 neurona)
5. potpuno povezan sloj (3 neurona)
6. izlazni sloj (1 neuron - trajanje karijere u godinama)

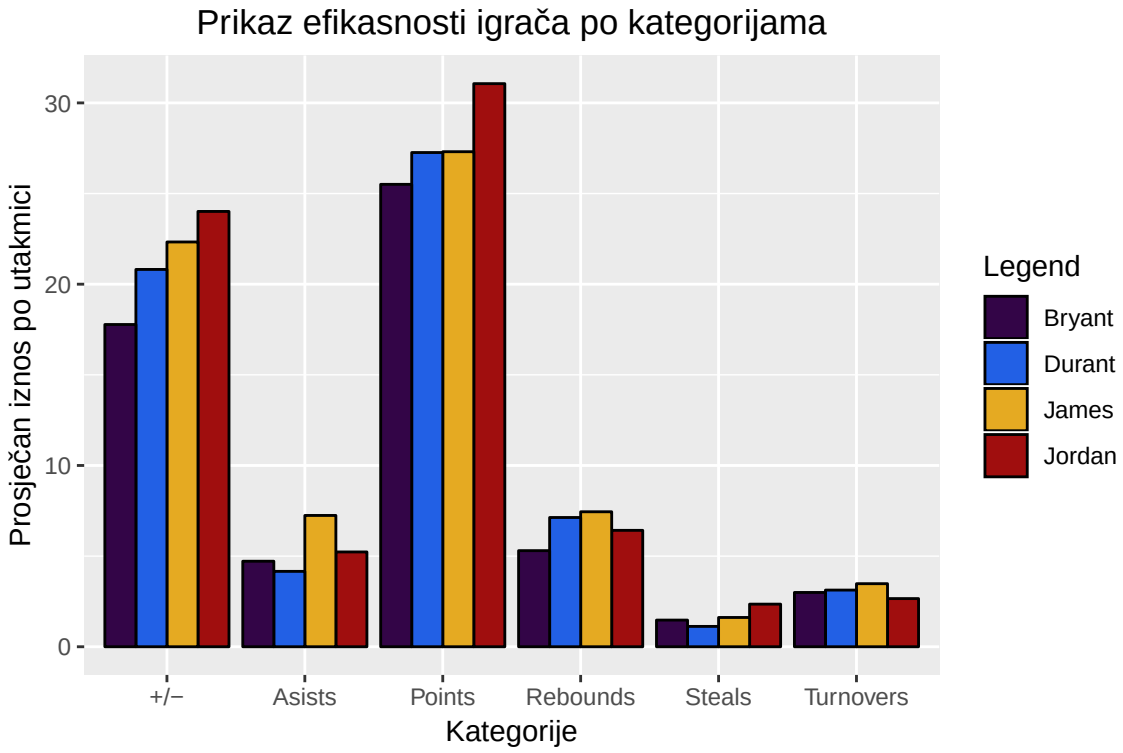
Zanimljivo je da je neuronska mreža bila nešto bolja u predikciji rezultata, konkretno dobivena je pogreška od 2.5 godine na skupu za ispitivanje.

11. Usporedba najboljih igrača u povijesti NBA lige (G.O.A.T. debate)

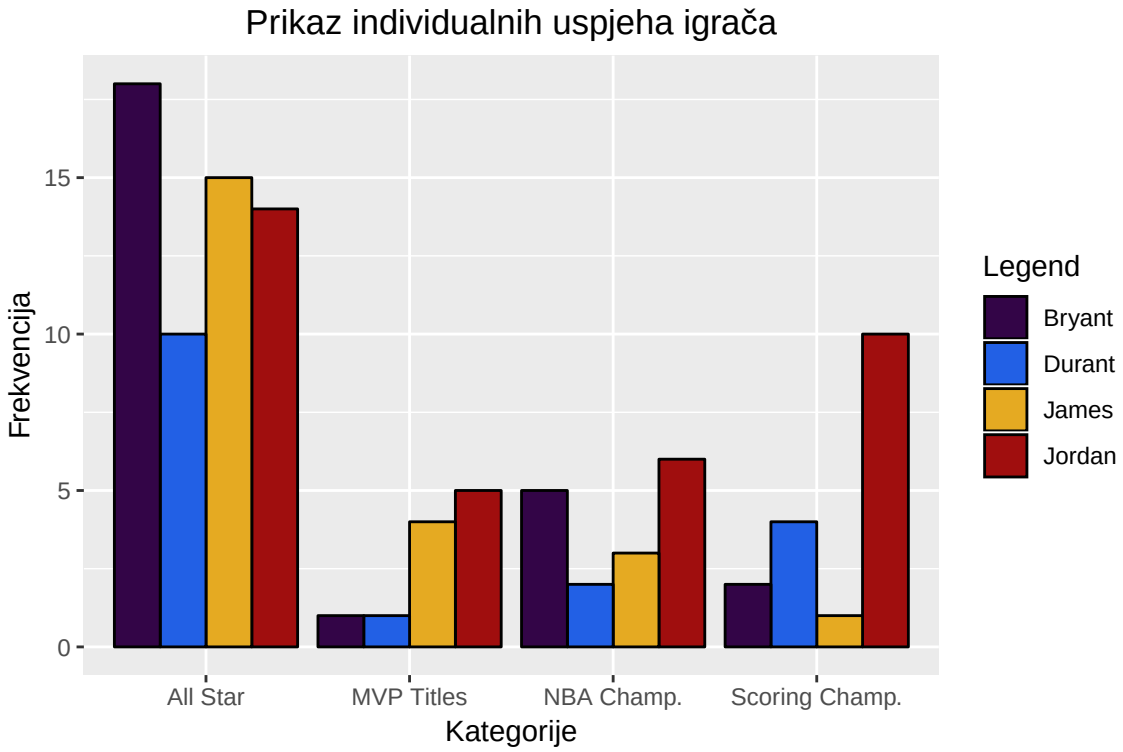
I tako, došli smo do kraja gdje ćemo se dotaknuti najškakljivije teme u NBA ligi a to je tko je najbolji igrač ikada? Navijači su podijeljeni otprilike ravnoporavno i svaka grupa je naoružana validnim argumentima zašto baš njihov miljenik zaslužuje titulu najboljeg igrača u povijesti NBA lige. U ovom dijelu nećemo dati odgovor na spomenuto pitanje jer bi to bilo relativno amaterski da na par statisčki zaključaka donesemo konačni sud. Umjesto toga prezentirat ćemo podatke i ostaviti čitatelju da sam procijeni tko je najbolji.

Prikaz kategorijskih podataka

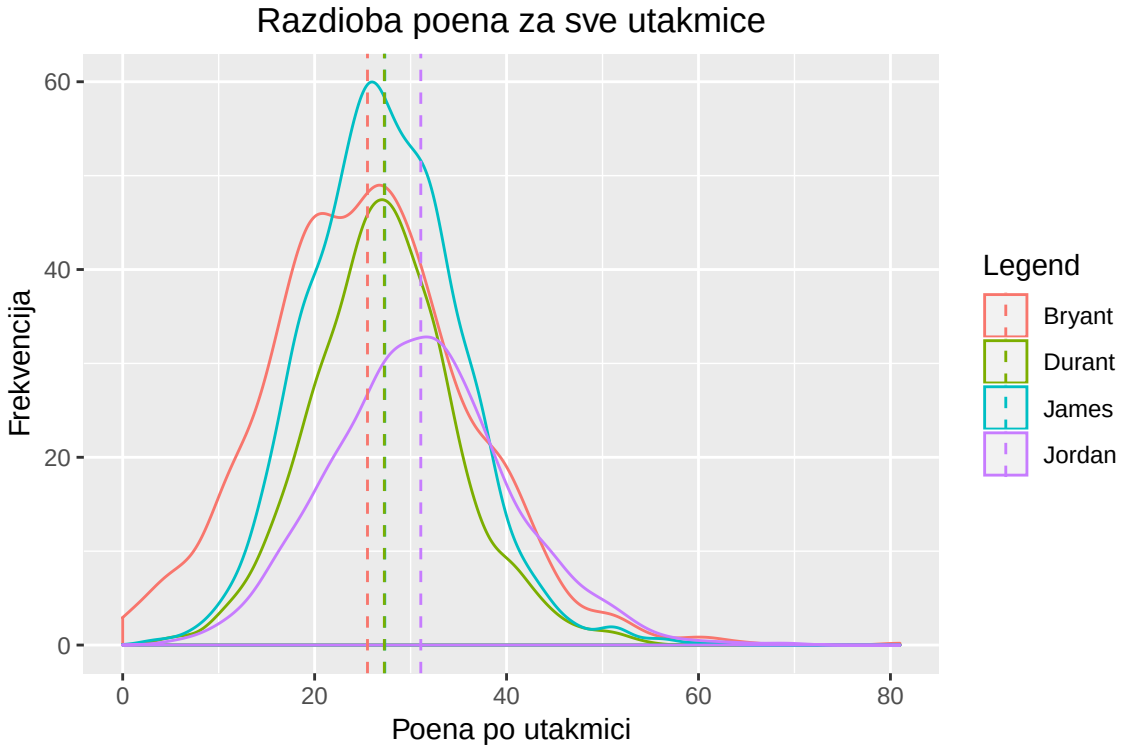
Prikaz najbitnijh karakterstika po kategorijama. Ovaj prikaz je površni te dosta grubo predočava situaciju.



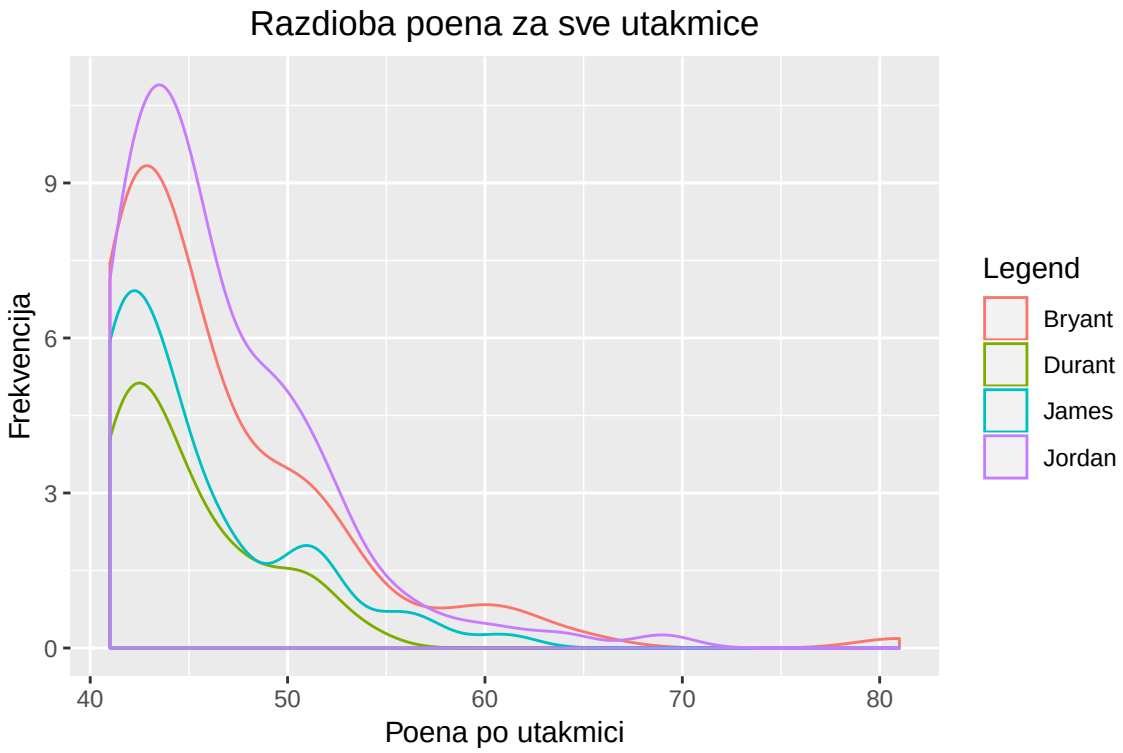
Individualni i timski uspjesi



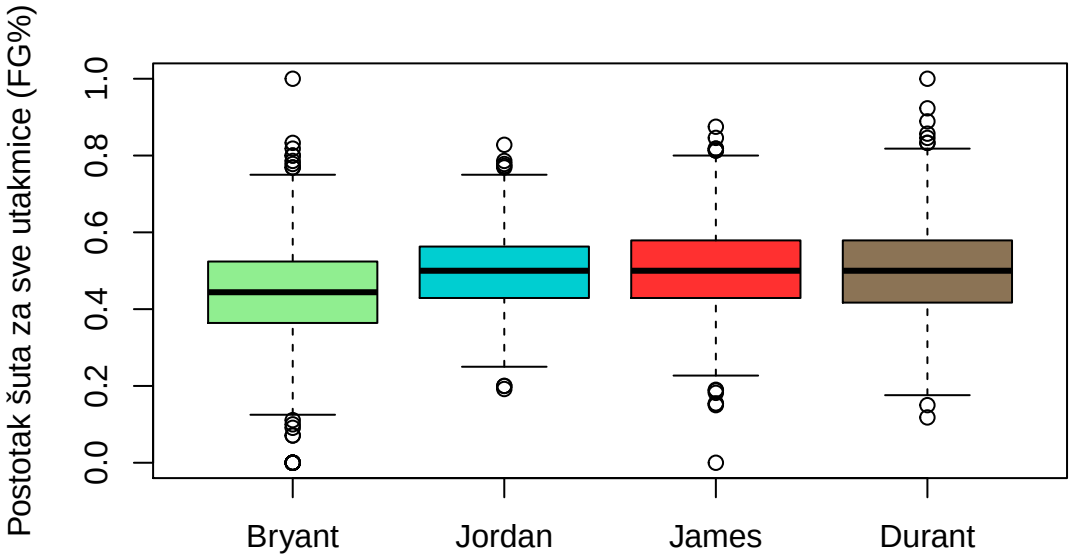
Šuterske karakterstike igrača

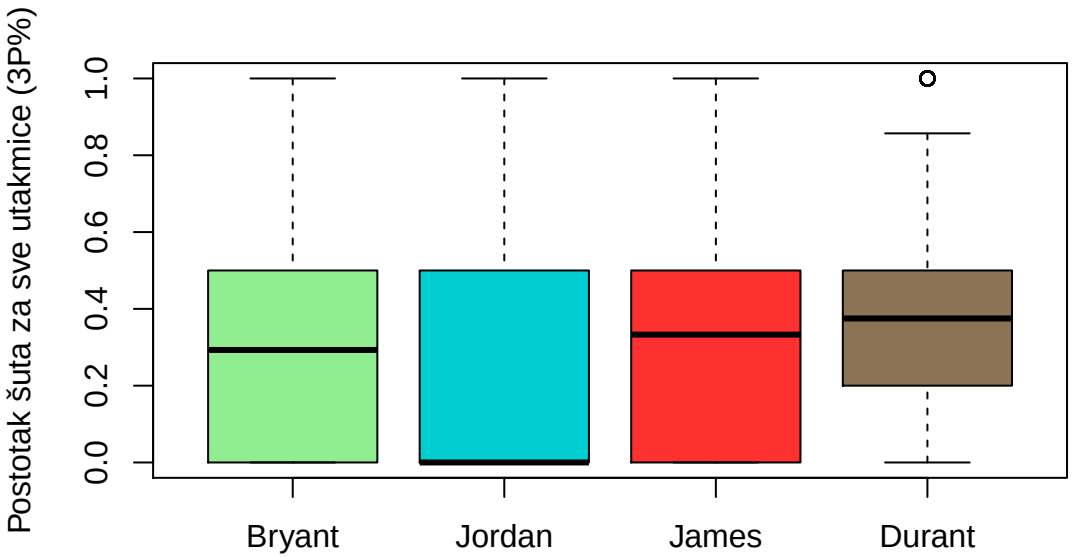


Iz prethodne slike ne možemo jasno vidjeti što se događa u repu distribucije koji je također bitan. Zbog toga u sljedećem prikazu povećavamo prikaz repova. Iz slike vidimo da zapravo Kobe Bryant i Michael Jordan imaju najviše utakmica s najviše zabijenih koševa.



No isto tako bitan je postotak pogođenih šuteva za dva i tri poena.




Postotak šuta za sve utakmice (3P%)
0.0
0.2
0.4
0.6
0.8
1.0
Bryant
Jordan
James
Durant