Analiza podataka o igračima NHL lige

Tomislav Lovrenčić, Bruno Dević, Domagoj Blažanin, Dino Tognon

11.12.2020

Contents

Uvod

Kroz ovaj rad ćemo analizirati podatke o igračima u NHL-u. Većinom ćemo se baviti pitanjem predviđanja plaće i pokušavati odgovoriti na pitanje o kojim atributima igraća ovisi njegova plaća. Naš završni cilj je na temelju podataka koje imamo izgraditi linearni model pomoću kojeg ćemo moći predvidjeti plaću igrača. Uz analizu plaće igrača želimo odgovoriti i na druga zanimljiva pitanja kao što su: utječe li preferiranost ruke na poziciju igrača , utječe li težina golmana na njegovu sposobnost, itd.

Skup podataka

Koristi se skup podataka o NHL igračima iz 100. sezone NHL-a koja se igra 2016.-2017. godine.

Trideset momčadi natjecalo se u regularnoj sezoni od 82 utakmice od 12. listopada 2016. do 9. travnja 2017. Tijekom tih utakmica prikupljen je obilan broj podataka o igračima koje ćemo u ovom radu koristiti kako bi izgradili pripadne modele koji će nas zanimati te za testiranje hipoteza na koje želimo dobiti odgovor.

Podatci se sastoje od 888 zapisa o igračima, te 95 zapisa o golmanima.

Statistička analiza općenito

Pregled ulaznih podataka

Dataset Igraci

```
igraci <-
  read.csv("C:/Users/Tomislav Lovrencic/Desktop/SAP-Projekt/igraciProm.csv")

#Izbacivanje viška redova na kraju
igraci <- igraci[complete.cases(igraci[ , 5:6]),]
```

Svaki redak predstavlja određenog igraća sa 199 podataka o tom igraču. Kako ih ima jako puno izdvojit ćemo najvažnije:

- Salary - plaća igrača
- PTS - broj bodova koje je igrač osvojio(bod se dobije za gol ili asistenciju)
- GP - Broj utakmica koje je igrač odigrao u navedenoj sezoni
- Position - Pozicija na kojoj igrač igra
- Team - Klub za koji igrač igra (Troslovna kratica)
- Hand - Ruka koju igrač preferira (L - lijeva, R - Desna)
- GS - Igračev ukupni Game Score
- DftYr - Godina u kojoj je igrač draftan
- DftRd - Runda u kojoj je igrač draftan
- GS/G - Igračev prosječni Game Score
- iFF - Broj ne blokiranih udaraca
- iHA - Broj oduzetih “puckova” na silu
- TOI/GP - Broj minuta na ledu / broj odigranih utakmica
- Born - Datum rođenja
- Ovrl - Pozicija na draftu

Dataset Golmani

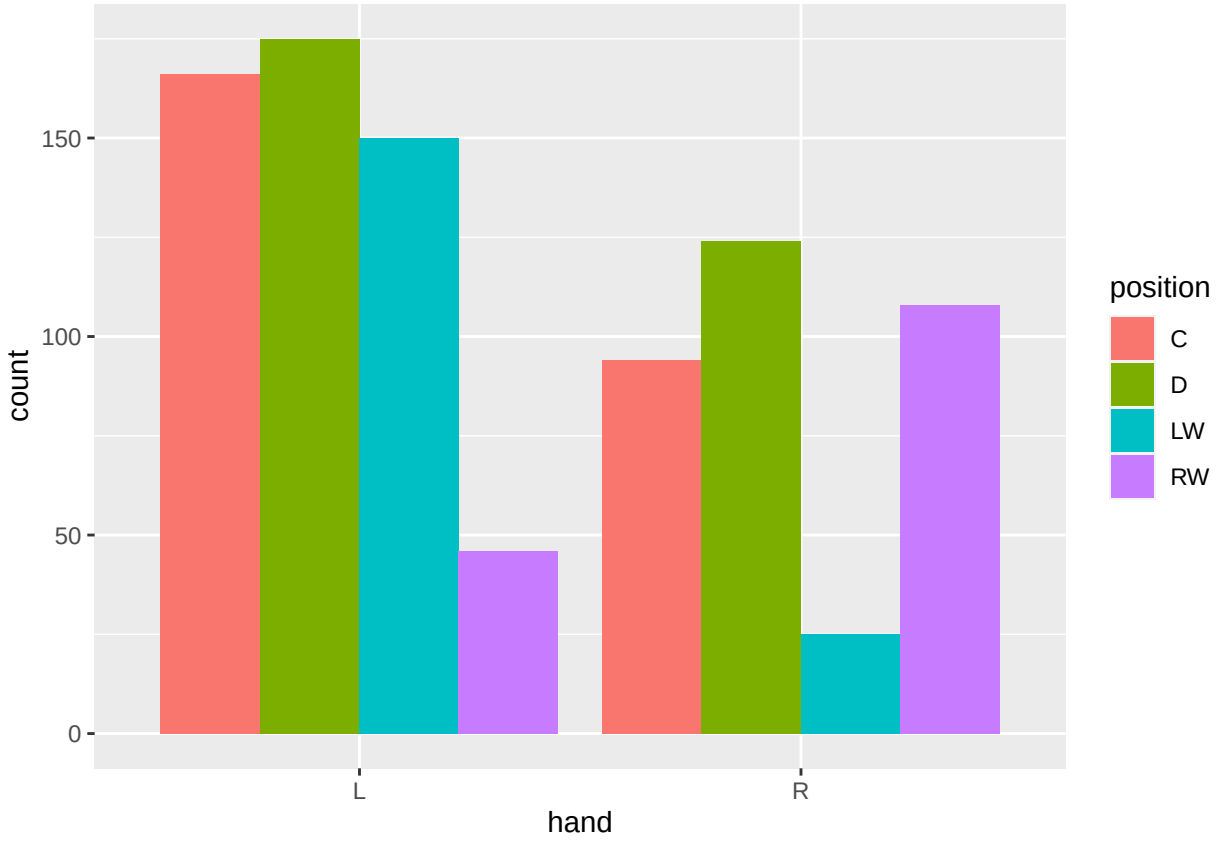
```
golmani <- read.csv("C:/Users/Tomislav Lovrencic/Desktop/SAP-Projekt/golmaniProm.csv")
```

Svaki redak predstavlja određenog golmana s 121 podataka o tom golmanu. Kako ih ima jako puno izdvojit ćemo najvažnije:

- Salary - plaća golmana u milijunima dolara
- SV% - Postotak obrana golmana
- MIN - Broj minuta koje je golman odigrao u navedenoj sezoni
- Wt - Težina golmana u funtama

Provjera zavisnosti igračeve preferirane ruke i njegove pozicije na terenu

Zanima nas postoji li zavisnost između ruke kojom igrač igra i pozicije na kojoj igra. Naravno, očekujemo da zavisnost postoji.



Iz grafa vidimo da puno više lijevaka igra na poziciji LW nego dešnjaka, te da puno više dešnjaka igra na poziciji RW nego lijevaka. Također je vidljivo da ima značajno više lijevaka nego dešnjaka što je vrlo neočekivano. Kako bi provjerili jesu li te razlike statistički značajne napravit ćemo test prilagodbe razdiobi, tj. hi-kvadrat test.

```
test <- chisq.test(table(df$hand, df$position))
#Kontingencijska tablica
test$expected
```

```
##
##            C        D       LW       RW
##   L 157.2297 180.8142 105.8277 93.12838
##   R 102.7703 118.1858  69.1723 60.87162
```

```
#Stvarna tablica
test$observed
```

```
##
##       C   D  LW  RW
##   L 166 175 150  46
##   R  94 124  25 108
```

Prvo gledamo kontingencijsku tablicu te vidimo da imamo dovoljan broj očekivanih vrijednosti u svakoj ćeliji tablice (U svakoj ćeliji više od 5). Također usporedbom dviju tablica jasno je vidljivo da postoje značajne razlike u vrijednostima te očekujemo da će test odbaciti nultu hipotezu da je uniformna razdioba lijevaka i dešnjaka u odnosu na poziciju.

```
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$hand, df$position)
## X-squared = 108.69, df = 3, p-value < 2.2e-16
```

Izvršavanjem Hi-kvadrat testa dobili smo očekivane rezultate. Pogledom na p-vrijednost dobivenu ovim testom, možemo sa sigurnošću odbaciti nul hipotezu(p-vrijednost je izrazito mala) te logički zaključiti da su preferirana ruka igrača i pozicija na kojoj igrač igra međusobno zavisne na bilo kojoj razini značajnosti.

Zanimljivost - Razdioba ljevaka i dešnjaka u hokeju u odnosu na prosjek

Tijekom prethodne analize zavisnosti igračeve preferirane ruke i pozicije na kojoj igra uočili smo jednu zanimljivost, a to je da ima značajno više ljevaka nego dešnjaka među hokejašima. Sada želimo provjeriti je li moguće da je to slučajnost ili je to specifičnost hokeja.

Kratkom pretragom interneta otkrili smo da dešnjaci čine oko 90% populacije, a ljevaci preostalih 10%.

Znači da bi slučajnu varijablu koja određuje je li netko dešnjak ili ljevak mogli definirati kao binomnu slučajnu varijablu s parametrom p = 0.9 i q = 0.1. Ako ima x dešnjaka, moramo izračunati koja je vjerojatno da bude x ili manje dešnjaka u n realizacija te binomne slučajne varijable

```
brojLjevaka <- nrow(igraci[igraci$Hand == 'L',])
brojDesnjaka <- nrow(igraci[igraci$Hand == 'R',])
#ukupni broj observacija(igrača)
n <- nrow(igraci)

x <- rbinom(10, n, 0.1)
x
```

```
##  [1]  97 101  91  88  92  98  86  85  98  90
```

Vidimo da u 10 različitih realizacija najviše smo slučajno dobili 101 ljevaka a mi ih u uzorku imamo 537. Sigurno ćemo odbaciti hipotezu da je stvarni udio ljevaka 10%.

```
x <- pbinom(brojDesnjaka, n, 0.9)
x
```

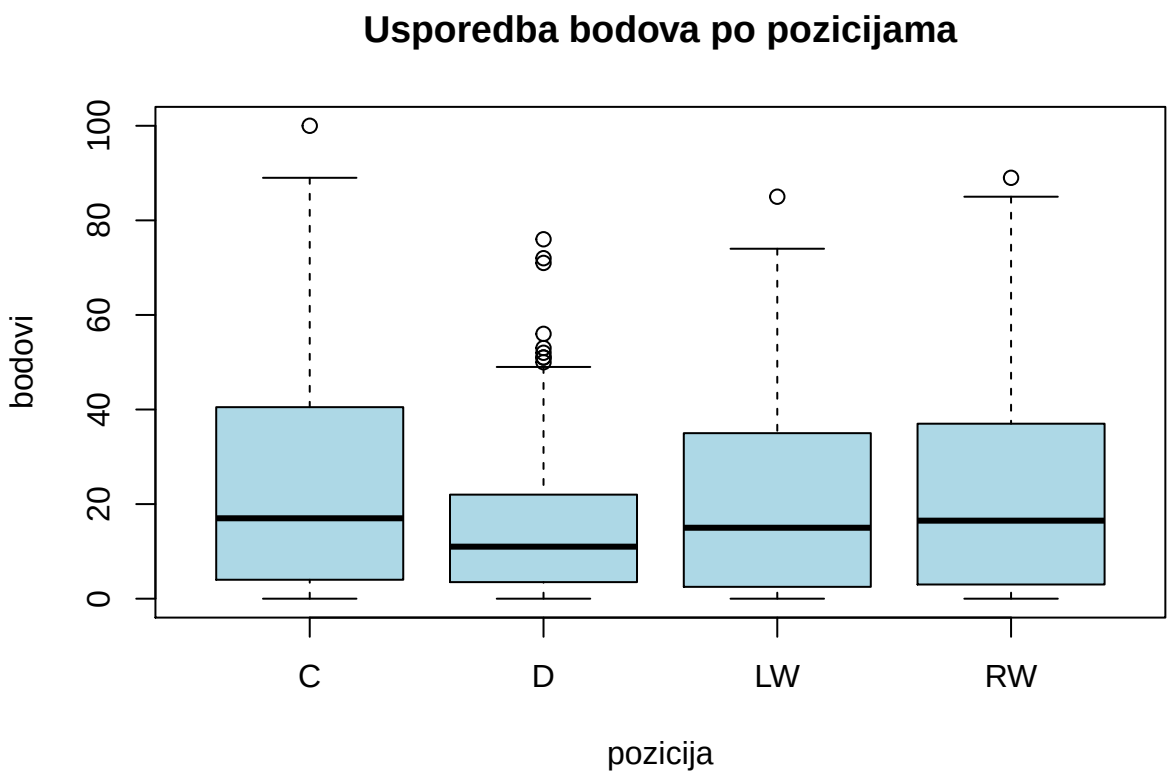
```
## [1] 1.588037e-296
```

Vidimo da kao što je i očekivano p vrijednost je ekstremno malena te ne možemo na smislenom nivou značajnosti tvrditi da je stvarni udio ljevaka među hokejašima 10%.

Nakon još malo istraživanja otkrili smo da se djeca dok su još malena uče igrati hokej tako da jačom rukom drže vrh štapa, što je razlog zašto puno djece kad odraste nastavi igrati tako te koristiti štap za ljevake, čime ulaze u statistiku kao ljevaci.

Provjera zavisnosti pozicije igrača i broja osvojenih bodova

Zanima nas osvajaju li igrači na svim pozicijama jednako bodova ili ipak neke pozicije osvajaju više bodova od ostalih. Očekujemo da će igrači na napadačkim pozicijama(LW, C, RW) osvajati više bodova od igrača u obrani(D).



Iz ovog vizualnog prikaza na prvu možemo vidjeti da sve pozicije otprilike imaju jednaki medijan bodova uz manje razlike te neznatno različite raspršenosti. No u sljedećem testu ćemo upravo provjeriti jednakost bodova i provjeriti dali su sve pozicije jednake po bodovima ili postoje odstupanja, te logički zaključiti jesu li pozicija i broj bodova međusobno zavisne varijable.

Prije nego provedemo ANOVA test moramo provjeriti pretpostavke: - nezavisnost pojedinih podataka u uzorcima, - normalna razdioba podataka, - homogenost varijanci među populacijama

Podatci su nezavisni jer svaki redak predstavlja drugog igrača. Možda postoji sitna zavisnost budući da igrači igraju zajedno u timovima a svaki tim najčešće sličan broj na svakoj poziciji(npr. 4 obrambena igrača) što bi značilo da odabir zadnjeg igrača za tim nije u potpunosti nezavisan, no smatramo da je to vrlo malena količina zavisnosti na poprilično veliki skup podataka te da ne stvara problem.

Iz grafa je vidljivo da su grupe okvirno podjednake veličine, što je uvijek poželjno. Provesti ćemo Lillieforsovu inačicu Kolmogorov-Smirnovljev testa kako bi provjerili pretpostavku o normalnosti.

```
## # A tibble: 4 x 5
##   position variable     n  mean    sd
##   <fct>    <chr>    <dbl> <dbl> <dbl>
## 1 C        PTS        260  23.8  22.5
## 2 D        PTS        299  15.3  15.0
## 3 LW       PTS        175  21.7  21.0
## 4 RW       PTS        154  22.0  21.6
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df2$PTS[df2$position == "RW"])
## D = 0.15468, p-value = 1.212e-09
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df2$PTS[df2$position == "LW"])
## D = 0.1514, p-value = 1.613e-10
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df2$PTS[df2$position == "C"])
## D = 0.14427, p-value = 2.619e-14
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df2$PTS[df2$position == "D"])
## D = 0.15392, p-value < 2.2e-16
```

Test je odbacio pretpostavku o normalnosti podataka no smatramo da nije toliko loša situacija na temelju histograma pa ćemo pogledati je li varijanca među populacijama homogena koristeći Bartlettov test te ako je ćemo nastaviti s ANOVOM.

```
bartlett.test(df2$PTS ~ df2$position)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df2$PTS by df2$position
## Bartlett's K-squared = 51.002, df = 3, p-value = 4.887e-11
```

```r
var((df2$PTS[df2$position=='LW']))
```

```
## [1] 441.4812
```

```r
var((df2$PTS[df2$position=='RW']))
```

```
## [1] 466.9728
```

```r
var((df2$PTS[df2$position=='C']))
```

```
## [1] 504.1489
```

```r
var((df2$PTS[df2$position=='D']))
```

```
## [1] 225.1306
```

Iako Bartlettov test odbacuje hipotezu o jednakosti varijanci. Budući da su varijance vrlo slične osim za obrambene igraće, a ni kod njih nije jako loša situacija, provesti ćemo ANOVU do kraja.

```r
#ANOVA test
a = aov(df2$PTS ~ df2$position)
summary(a)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## df2$position   3  11453    3818   9.756 2.45e-06 ***
## Residuals    884 345928     391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Iz gornje ANOVA tablice vidljivo je da postoje značajne razlike između skupina (p = 2,45e-06) te možemo odbaciti nultu hipotezu tj. odbacujemo hipotezu da su srednje vrijednosti bodova svake pozicije jednake. Logički zaključujemo da vrlo vjerojatno postoji zavisnost između pozicije na kojoj igrač igra i osvojenih bodova tog igrača.

Provjera zavisnosti težine golmana s postotkom obrana

Očekujemo da su težina golmana i postotak obrana nezavisne varijable.

```
## [1] "Podjela težine golmana u kategorije"
```

```
## [1] "do 80kg - lagan(l)"
```

```
## [1] "80-90kg - srednje tezak(st)"
```

```
## [1] "90-100kg - tezak(t)"
```

```
## [1] "preko 100kg - jako tezak(jt)"
```

```
##  golmani.tezKat1  golmani.SV.
##  l : 6           Min.   :0.7330
##  st:39           1st Qu.:0.9000
##  t :39           Median :0.9130
##  jt:10           Mean   :0.9077
##                  3rd Qu.:0.9227
##                  Max.   :1.0000
```

Prvo provjeravamo imamo li dovoljno opservacija u svakoj kategoriji. Vidimo da je samo 6 opservacija u kategoriji lagan te samo 10 opservacija u kategoriji jako težak. Kako bi malo izjednačili broj opservacija po kategorijama pomaknuti ćemo kategoriju lagan do 83kg i kategoriju jako težak od 98kg.

```
## [1] "Preraspodjela golmana u kategorije"
```

```
## [1] "do 83kg - lagan(l)"
```

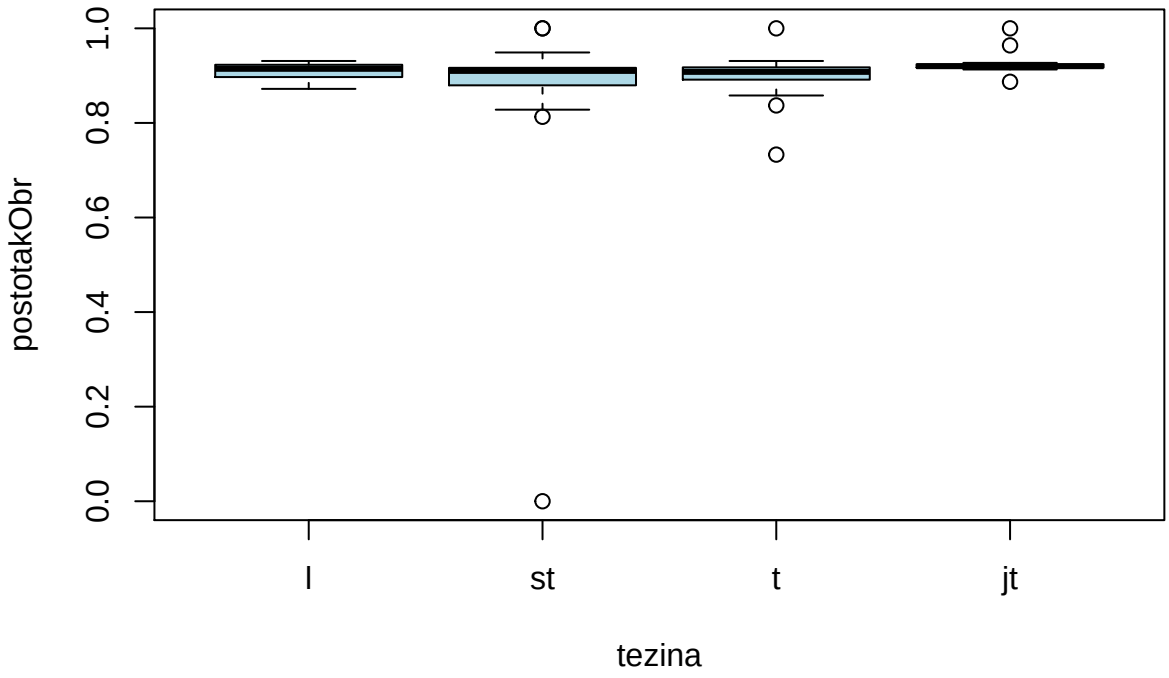
```
## [1] "#83-90kg - srednje tezak(st)"
```

```
## [1] "#90-98kg - tezak(t)"
```

```
## [1] "#preko 98kg - jako tezak(jt)"
```

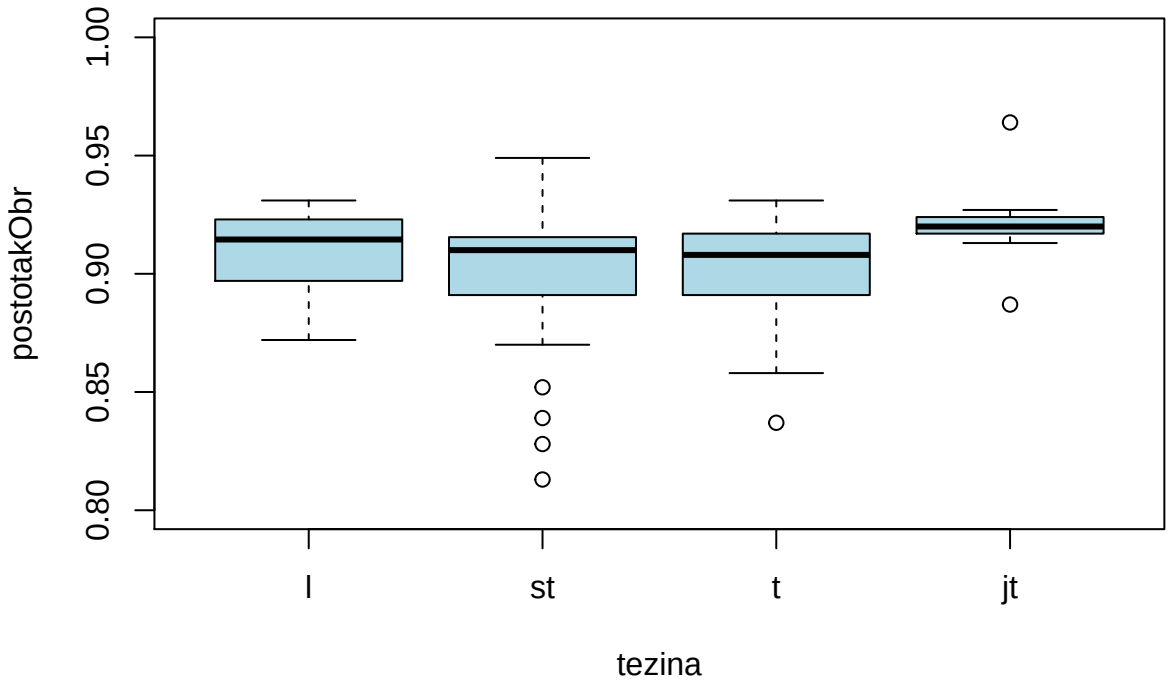
```
##  tezina   postotakObr
##  l :14   Min.   :0.0000
##  st:32   1st Qu.:0.8990
##  t :32   Median :0.9130
##  jt:17   Mean   :0.8981
##          3rd Qu.:0.9225
##          Max.   :1.0000
```

Sada kada su kategorije bolje raspodijeljene vizualizirat ćemo podatke pomoću box plota.



Vidimo da imamo više outliera koji dosta odskaču i nemaju previše smisla. Recimo imamo bar 3 golmana koji imaju 100% obrana i imamo jednog golmana s 0% obrana što se vjerojatno desilo jer su igrali jako malo minuta ili jednostavno nisu ispravno uneseni podatci. Zato ćemo sada izbaciti sve golmane koji su igrali manje od 30 minuta tijekom sezone.

Zanimljivost: Malo dubljim istraživanjem saznali smo da je golman kojeg vidimo u grafu s 0% obrana Jorge Alves, upravitelj opreme za Carolina Hurricanese, koji drži rekord za najkraću karijeru u povijesti NHL-a. Igrao je svega 7,6 sekundi kada je ušao da zamijeni golmana koji se ozlijedio.

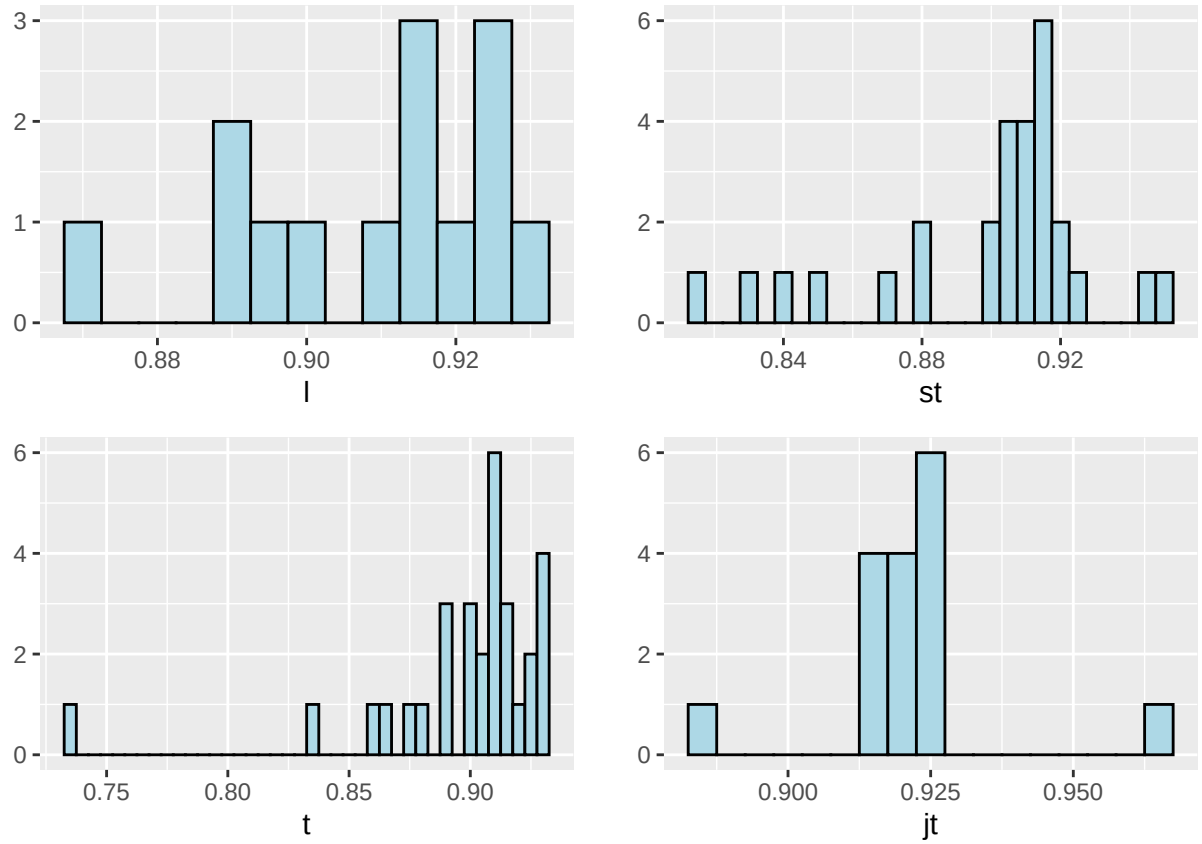


Vidimo da smo izbacivanjem outliera dobili puno pregledniji i smisleniji graf. Vidimo da su medijani vrlo blizu za sve težinske kategorije te očekujemo da će ANOVA test pokazati da ne možemo odbaciti hipotezu o jednakosti srednjih vrijednosti.

Prije provođenja ANOVE kao i ranije moramo provjeriti pretpostavke. Nezavisnost pojedinih podatak je ispunjena budući da je svaki podatak druga osoba, razdiobu podataka ćemo provjeriti pomoću Lillieforsove inačice Kolmogorov-Smirnovljev testa.

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(tezObr$postotakObr[tezObr$tezina == "l"])
## D = 0.19123, p-value = 0.1786

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(tezObr$postotakObr[tezObr$tezina == "st"])
## D = 0.28563, p-value = 2.93e-06

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(tezObr$postotakObr[tezObr$tezina == "t"])
## D = 0.20906, p-value = 0.001748

##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(tezObr$postotakObr[tezObr$tezina == "jt"])
## D = 0.27737, p-value = 0.001828
```



Iako Lilliefors odbacuje normalnost osim za laku kategoriju, smatramo da su rezultati dovoljno dobri te idemo dalje provjeriti homogenost varijance.

```
bartlett.test(tezObr$postotakObr ~ tezObr$tezina)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  tezObr$postotakObr by tezObr$tezina
## Bartlett's K-squared = 19.678, df = 3, p-value = 0.0001979
```

```
var((tezObr$postotakObr[tezObr$tezina=='l']))
```

```
## [1] 0.0002923352
```

```
var((tezObr$postotakObr[tezObr$tezina=='st']))
```

```
## [1] 0.001046247
```

```r
var((tezObr$postotakObr[tezObr$tezina=='t']))
```

```
## [1] 0.001455637
```

```r
var((tezObr$postotakObr[tezObr$tezina=='jt']))
```

```
## [1] 0.0002156625
```

Budući da su razlike varijanci malene(iako je Bartlettov test odbacio hipotezu o jednakosti varijanci) provesti ćemo ANOVU.

```r
a = aov(tezObr$postotakObr ~ tezObr$tezina)
summary(a)
```

```
##               Df Sum Sq   Mean Sq F value Pr(>F)
## tezObr$tezina  3 0.0070 0.0023318   2.527 0.0629 .
## Residuals     84 0.0775 0.0009226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Iako smo očekivali da će p vrijednosti biti veća, p vrijednost je ispala tek 6.3%. Što znači da na razini značajnosti od 5% ne odbacujemo hipotezu o različitosti srednjih vrijednosti postotka obrane golmana, no već na razini značajnosti od 7% bi mogli odbaciti tu hipotezu. Dakle nije lako zaključiti utječe li težina golmana na njegovu sposobnost branjenja. Budući da je uobičajeno koristiti razinu značajnosti od 5%, ne ćemo odbacit hipotezu H0 i zaključiti da su srednje vrijednosti jednake.

Statistička analiza plaće

Obrada i Analiza plaće

```
## [1] "Prikaz formata plaće:"
```

```
## [1] 0.5750 5.5000 0.8425 0.8925 0.6250 0.9250
```

```
## [1] "Saženak plaća igrača:"
```

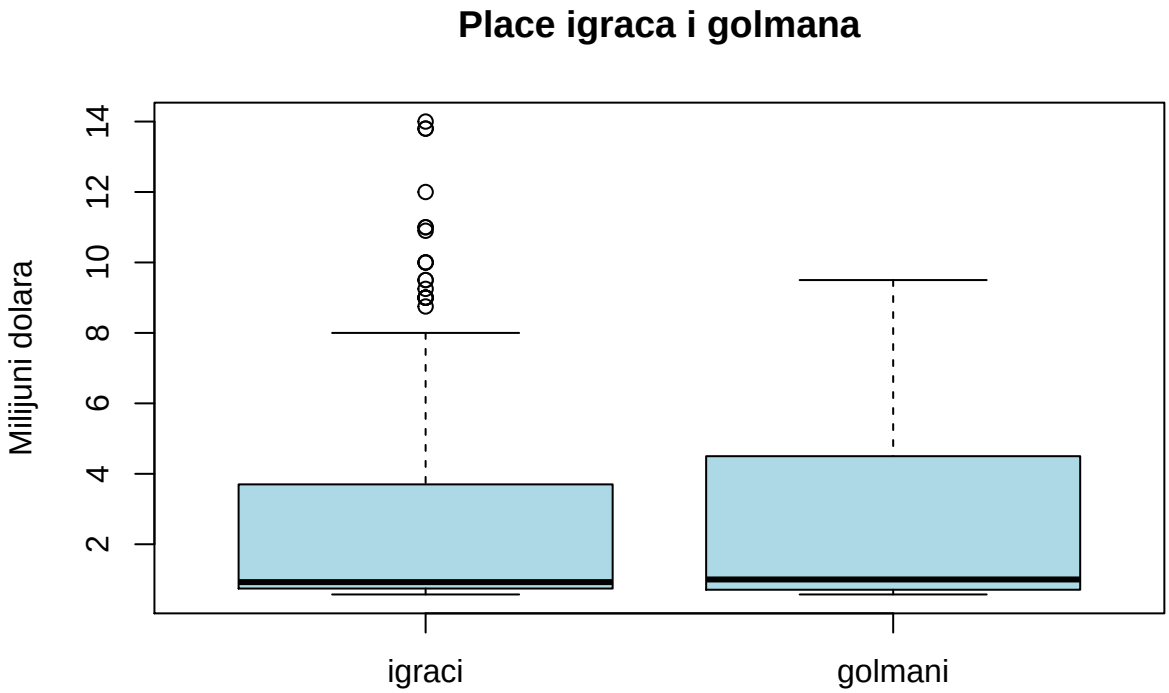
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.5750  0.7425  0.9250  2.3253  3.7000 14.0000      14
```

```
## [1] "Saženak plaća golmana:"
```

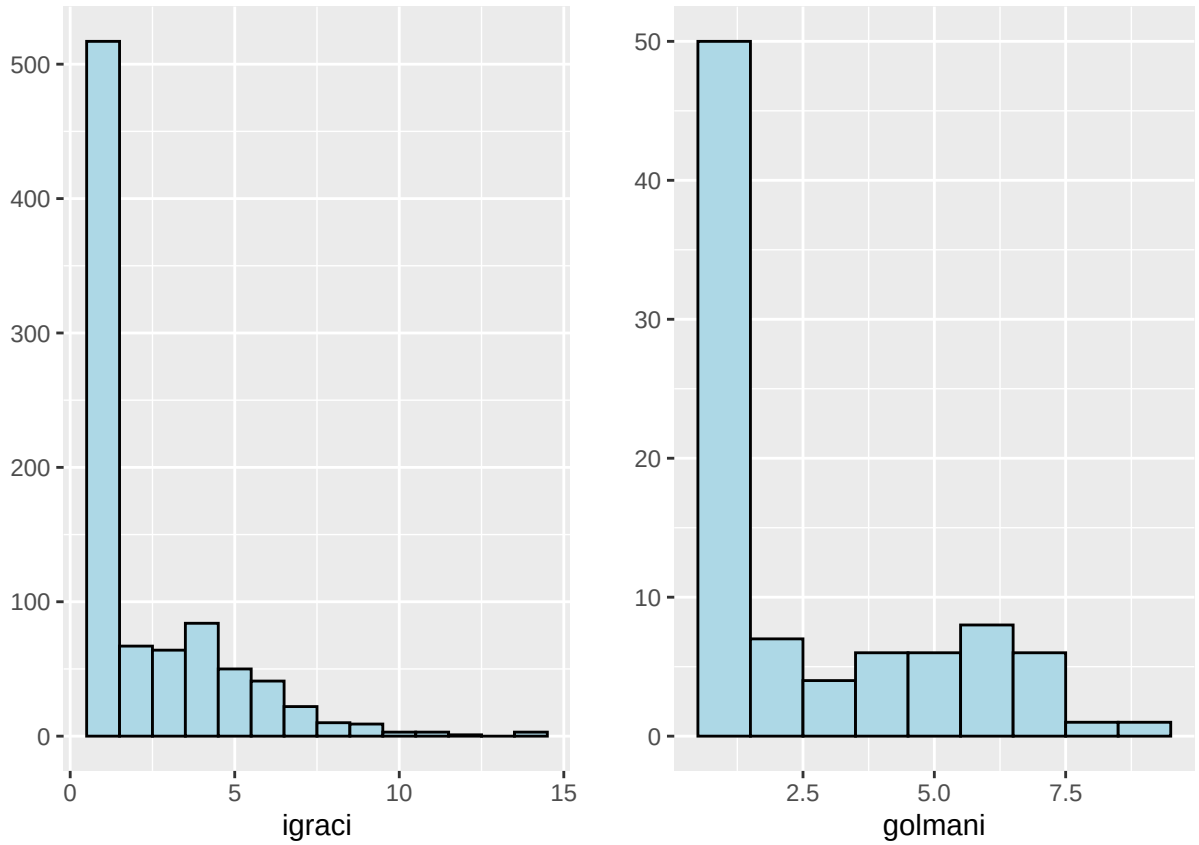
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.5750  0.7087  1.0000  2.5823  4.5000  9.5000       6
```

Vidimo da za 142 igrača i 6 golmana nema podataka o plaći. Srednja vrijednost plaće igrača je 2.32 milijuna dolara a medijan tek 925 tisuća dolara. Srednja vrijednost plaće golmana je 2.58 milijuna dolara a medijan milijun dolara, dakle nešto više nego za igrače. Iz velike razlike srednje vrijednosti i medijana možemo iščitati

veliku nagnutost distribucije ulijevo. Kako bi grafički prikazali ove podatke najbolje da pogledamo boxplot dijagrame.



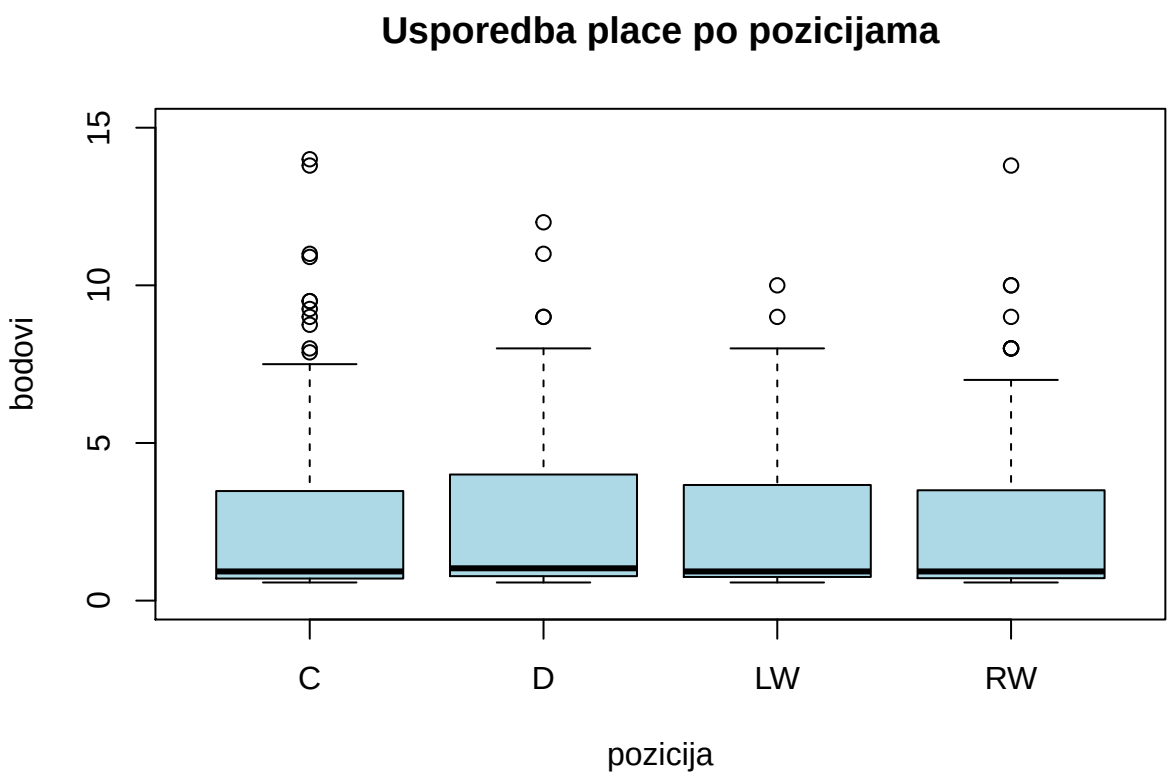
Iz boxplot dijagrama plaće igrača vidimo da 75% posto igrača ima plaću manju od 4 milijuna dolara. Budući da je medijan jako blizu prvog kvartila(Q1) možemo zaključiti da je distribucija jako nagnuta ulijevo što možemo provjeriti pomoću histograma. Iz boxplot dijagrama plaće golmana vidimo da je treći kvartil nešto viši (oko 4.5 milijuna dolara) te da je medijan neznatno udaljeniji od prvog kvartila što nam govori kako je distribucija plaće golmana možda malo manje nagnuta od distribucije plače igrača što ćemo najbolje provjeriti pomoću histograma. Također vidimo da je nešto veća raspršenost plaće golmana.



Histogrami su vrlo slični, gotovo pa identični. Vidljivo je da distribucija plaća nije simetrična te se velika većina vrijednosti nalazi u intervalu do milijun dolara. Zaključujemo da se plaća ne ravna po normalnoj distribuciji.

Provjera zavisnosti pozicije igrača i njegove plaće

Zanima nas jesu li igrači na svim pozicijama plaćeni jednako, ili su neke pozicije više plaćene.



Iz grafa vidimo da ne postoje gotovo nikakve razlike između plaća igrača na različitim pozicijama.

```
model2 = lm(df4$Salary~df4$Position, data = df4)

summary(model2)
```

```
##
## Call:
## lm(formula = df4$Salary ~ df4$Position, data = df4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.819 -1.585 -1.349  1.376 11.683
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.31690    0.14329  16.169   <2e-16 ***
## df4$PositionD   0.07697    0.19619   0.392    0.695
## df4$PositionLW -0.06251    0.22777  -0.274    0.784
## df4$PositionRW -0.03167    0.23534  -0.135    0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.302 on 870 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.000542,   Adjusted R-squared:  -0.002904
## F-statistic: 0.1573 on 3 and 870 DF,  p-value: 0.925
```

```
df4 <- df4[complete.cases(df4),]
```

U ovom slučaju naš linearan model u kojem gledamo dali je pozicija igrača statistički značajna varijabla ne daje dobre rezultate te pogledom na p-vrijednosti pozicija možemo zaključiti da pozicija nije statistički značajna. Mogli bi logički zaključiti da su igrači na svim pozicijama podjednako plaćeni.

Testirat ćemo jednakost srednjih vrijednosti plaće po pozicijama pomoću ANOVE. Da bi to mogli potrebno je zadovoljiti kriterije ANOVE kao i u prethodnom primjeru. Pretpostavka nezavisnosti je zadovoljena budući da svaki redak predstavlja jednog igrača. Sada ćemo provjeriti pretpostavku normalnosti pomoću Lillieforseove inačice KS testa kao i ranije.

```
## # A tibble: 4 x 5
##   Position variable     n  mean    sd
##   <fct>    <chr>    <dbl> <dbl> <dbl>
## 1 C        Salary     258  2.32  2.50
## 2 D        Salary     295  2.39  2.15
## 3 LW       Salary     169  2.25  2.14
## 4 RW       Salary     152  2.28  2.40
```

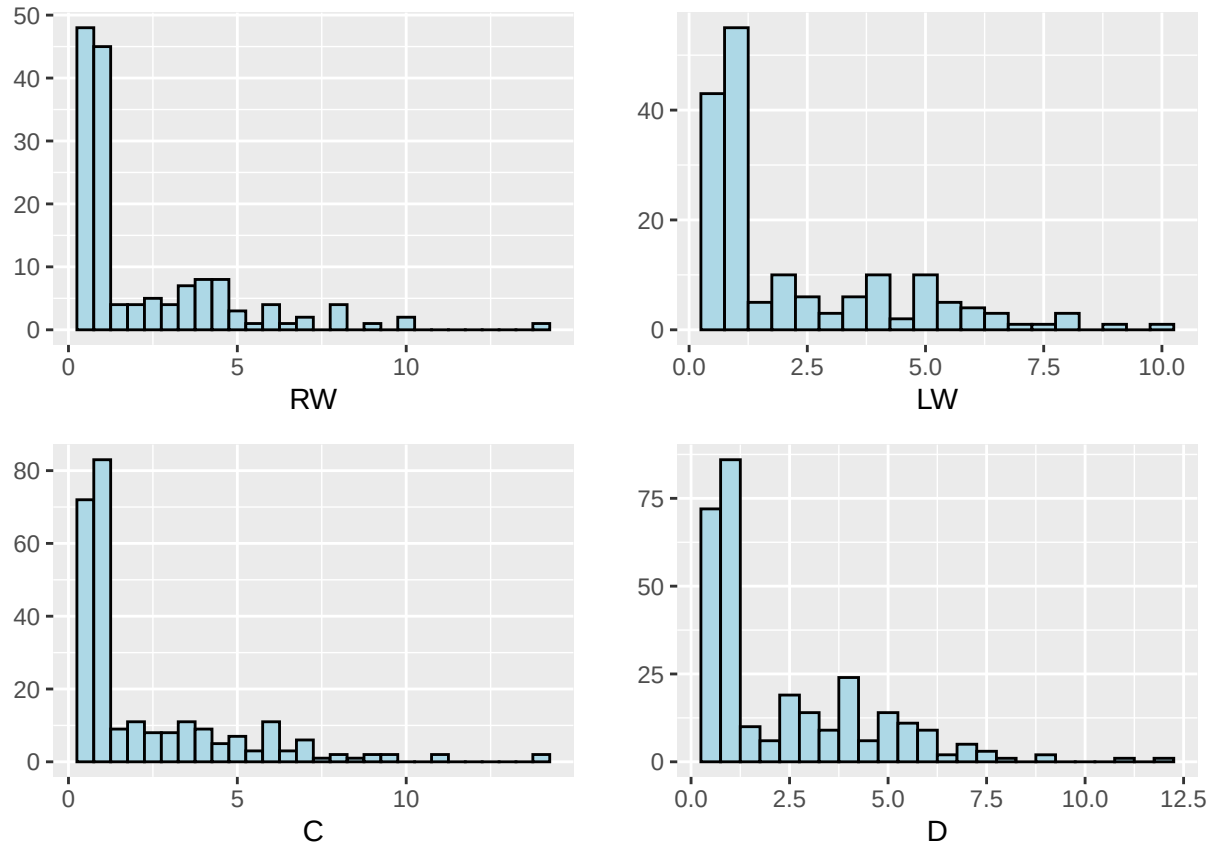
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5750  0.7425  0.9250  2.3253  3.7000 14.0000
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df4$Salary[df4$Position == "RW"])
## D = 0.27887, p-value < 2.2e-16
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df4$Salary[df4$Position == "LW"])
## D = 0.27905, p-value < 2.2e-16
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df4$Salary[df4$Position == "C"])
## D = 0.27959, p-value < 2.2e-16
```

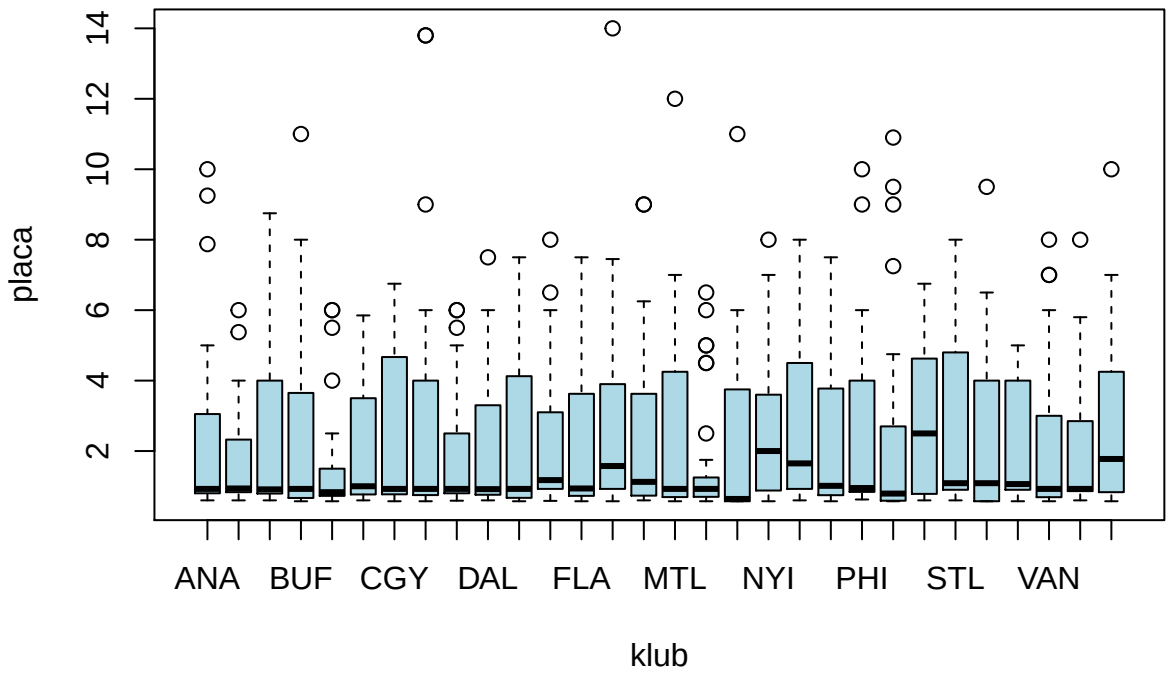
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df4$Salary[df4$Position == "D"])
## D = 0.23995, p-value < 2.2e-16
```



Test odbacuje hipotezu o normalnosti podatak i iz histogram je vidljivo da je pretpostavka normalnosti ozbiljno narušena te ne možemo provesti ANOVU u ovom slučaju.

Provjera zavisnosti tima(kluba) igrača i njegove plaće

Zanima nas jesu li igrači nekih klubova plaćeni više od igrača drugih klubova, tj. jesu li u prosjeku igrači svih klubova jednako plaćeni. Očekujemo da ne plaćaju svi klubovi isto svoje igrače već da bogatiji klubovi više plaćaju svoje igrače.



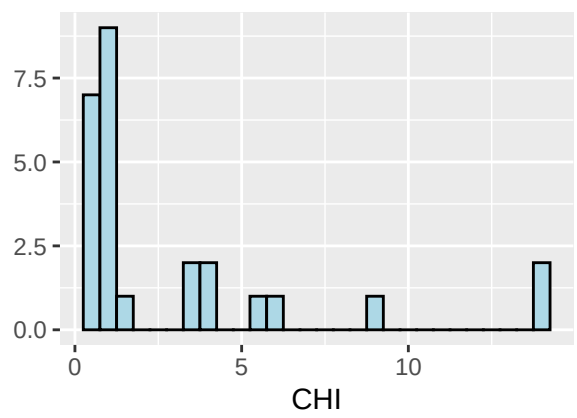
Iz histograma je vidljivo da postoje neke razlike između prosječne plaće u klubovima no ne vrlo značajne, pretpostavljamo da s testom nećemo moći odbaciti hipotezu da su srednje vrijednosti plaća jednake za svaki tim.

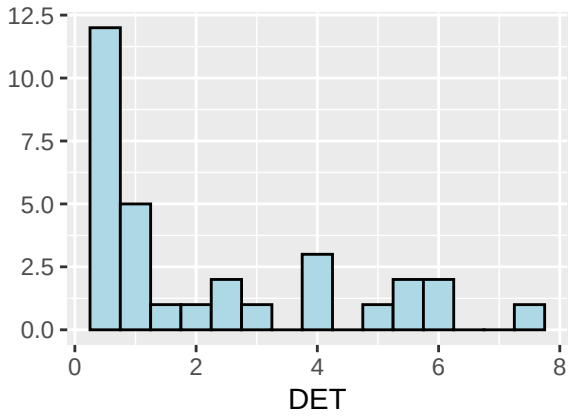
Prije provođenja ANOVE kao i ranije moramo provjeriti pretpostavke. Nezavisnost pojedinih podatak je ispunjena, razdiobu podatak ćemo provjeriti pomoću Lillieforsove inačice Kolmogorov-Smirnovljev testa. Provjeravamo samo za nekoliko timova budući da će rezultati za sve timove biti podjednaki.

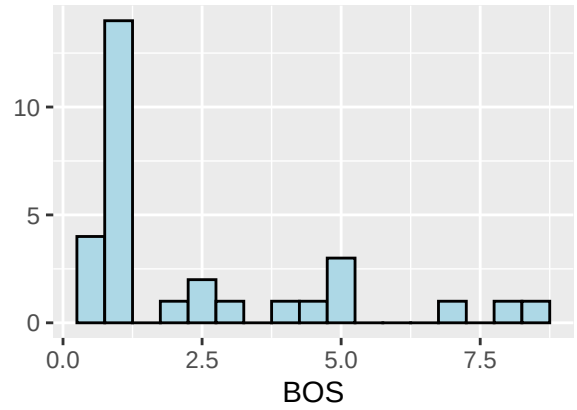
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df5$Salary[df5$Team == "CHI"])
## D = 0.31721, p-value = 2.834e-07
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df5$Salary[df5$Team == "DET"])
## D = 0.27647, p-value = 1.878e-06
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  scale(df5$Salary[df5$Team == "BOS"])
## D = 0.30639, p-value = 9.103e-08
```







Iz rezultata testa vidimo da podatci nisu normalni, te iz histograma vidimo da su odstupanja značajna te zaključujemo da ne smijemo koristiti ANOVU u ovom slučaju.

Linearni modeli za plaće

Prije nego što prihvatimo linearan model, prvo moramo biti sigurni da su ispunjene četiri pretpostavke:

1. Linearna veza: Postoji linearna veza između ovisne varijable x i nezavisnih varijabli y,z...
2. Nezavisnost.
3. Homoskedastičnost: Reziduali imaju konstantnu varijancu na svakoj razini x-a.
4. Normalnost: Reziduali modela imaju normalnu distribuciju.

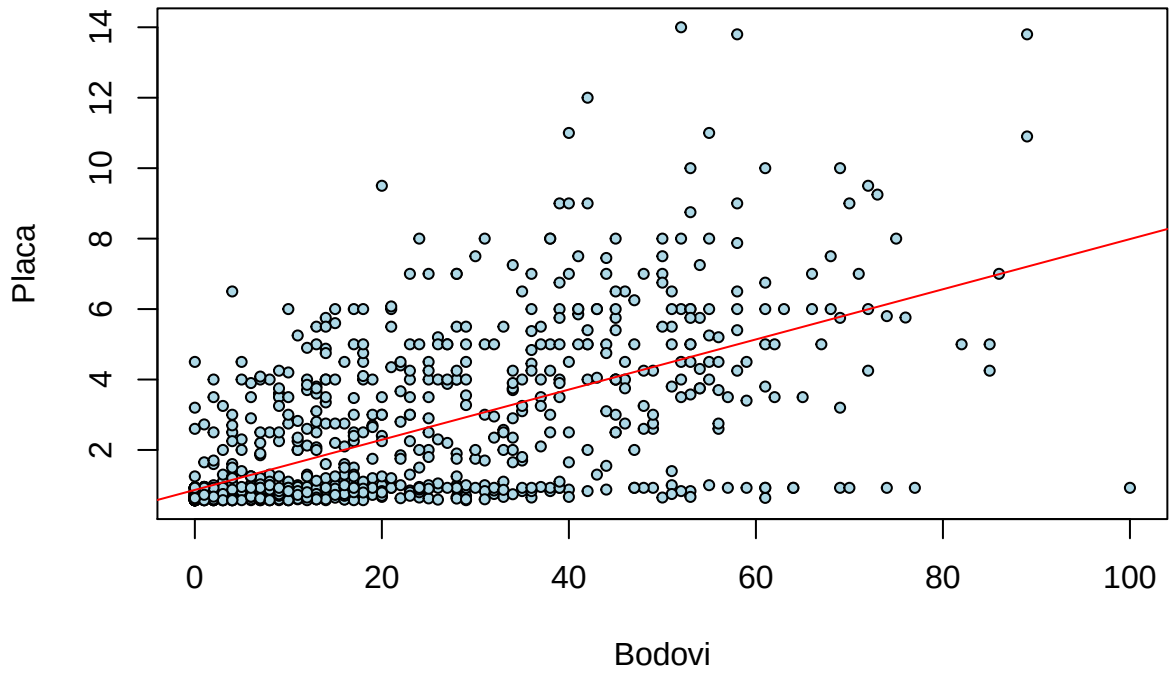
Ako se prekrši jedna ili više ovih pretpostavki, tada rezultati naše linearne regresije mogu biti nepouzdani. Za svaki model pozivom plot() metode možemo na temelju grafova ili pomoću pripadnih testova odrediti sve 4 pretpostavke i na taj način se uvjeriti da je model značajan.

U nastavku možemo vidjeti postupak dolaska i razvijanja linearnih regresijskih modela od jednostavnijih pa sve do završnih modela koji su se koristili pri predviđanju plaća. Završni modeli su proizvod stupnjevitog ažuriranja jednostavnih modela s novim varijablama koje su dodatno poboljšavale linearne regresijske modele.Pošto je prikazan postupak dolaska do završnog modela pretpostavke koje trebaju vrijediti za svaki od njih smo odlučili prikazati samo za završne modele iz razloga da ne dolazi do redundancije koda i teksta , no testovi su napravljeni te za svaki od jednostavnijih modela vrijedi uspostavljen linearni odnos , neovisnost , normalnost te homoskedastičnost.

Linearni model za predikciju plaće na temelju broja bodova

Očekujemo da će igrači koji osvajaju veći broj bodova imati veće plaće.

```
Salary <- data.frame(igraci$Salary)

PTS <- data.frame(igraci$PTS)

df3 <- data.frame(Salary,PTS)

plot(df3$igraci.PTS,df3$igraci.Salary,
     ylab = "Plaća",
     xlab = "Bodovi",
     col= "black",
     bg = "lightblue",
     cex = 0.7,
     pch = 21)

abline(lm(df3$igraci.Salary~df3$igraci.PTS), col="red")
```




```
model1 = lm(df3$igraci.Salary~df3$igraci.PTS, data = df3)


summary(model1)
```

```
##
```

```
## Call:
## lm(formula = df3$igraci.Salary ~ df3$igraci.PTS, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0621 -0.9229 -0.2654  0.6465  9.4313
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.86544    0.08697   9.951   <2e-16 ***
## df3$igraci.PTS  0.07122    0.00303  23.500   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.799 on 872 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.3878, Adjusted R-squared:  0.3871
## F-statistic: 552.3 on 1 and 872 DF,  p-value: < 2.2e-16
```

Budući da je p-vrijednost jako niska (p <0,001), možemo zaključiti da PTS ima statistički značajan učinak na plaću igrača. Iz t-testa koji ima jako veliku vrijednost možemo zaključiti da su varijable plaća i PTS zavisne. Ako pogledamo R-kvadrat , on nam mjeri udio varijacije u našoj ovisnoj varijabli (Plaća) objašnjenu neovisnom varijablom (Bodovi) a on je u našem slučaju 0.3878 , iz čega možemo zaključiti da bi ovaj linearni model mogao objasniti ~ 38% odstupanja od dobivenih vrijednosti.

Linearni regresijski model za predviđanje plaće koristeći godine igrača i Ovrl

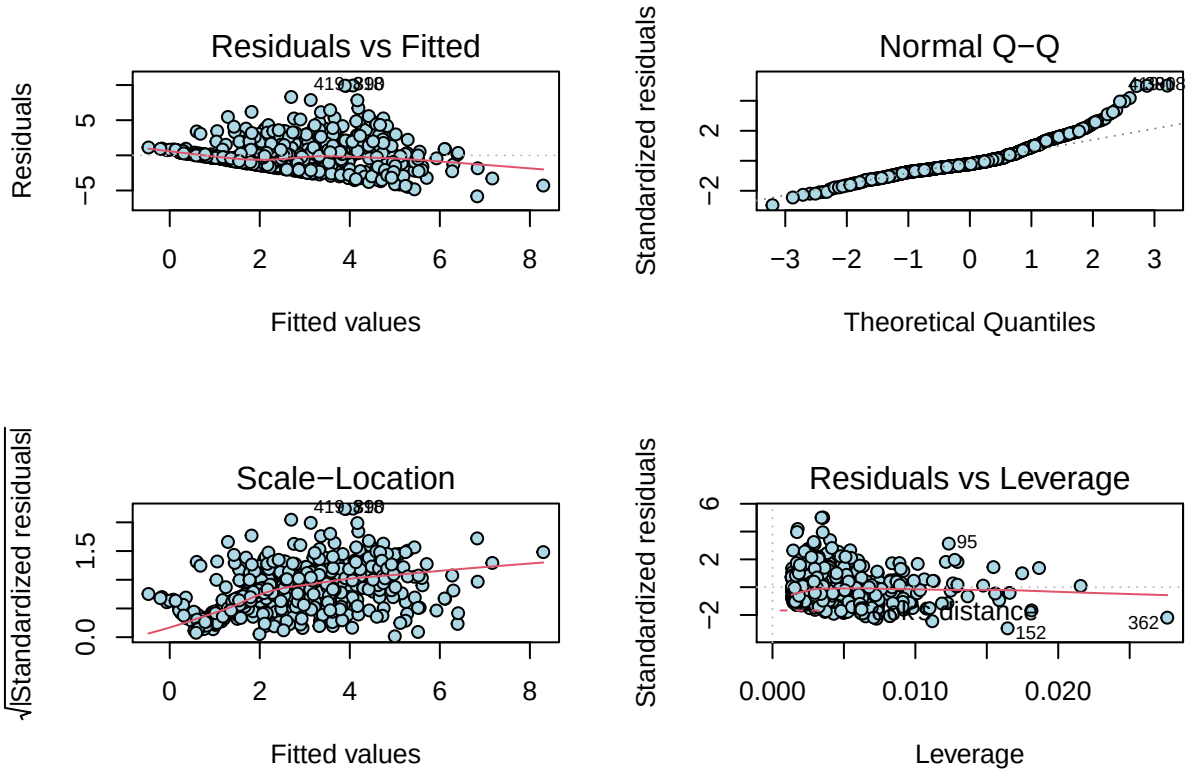
```
born <- ymd(igraci$Born)
playerAge <- 2017 - year(born)

model4 = lm(Salary ~ playerAge + draftOverall)

summary(model4)
```

```
##
## Call:
## lm(formula = Salary ~ playerAge + draftOverall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8295 -1.1436 -0.4421  0.8338  9.9298
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.108127   0.438372  -9.371   <2e-16 ***
## playerAge     0.277008   0.016441  16.848   <2e-16 ***
## draftOverall -0.011992   0.001173 -10.226   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.986 on 746 degrees of freedom
```

```
##   (139 observations deleted due to missingness)
## Multiple R-squared:  0.3097, Adjusted R-squared:  0.3079
## F-statistic: 167.4 on 2 and 746 DF,  p-value: < 2.2e-16
```

```
car::vif(model4)
```

```
##    playerAge draftOverall
##     1.038025     1.038025
```



Prvi pokušaj u stvaranju regresijskog modela napravljen je s varijablama playerAge i draftOverall. Iz sumiranog modela možemo vidjeti da su obje varijable uz malu p-vrijednost statistički značajne. Proveli smo usporedbu za međuovisnost između korištenih varijabli te se u ovom slučaju pokazalo da su doista nezavisne što možemo saznati naredbom vif(model4) gdje su vrijednosti jako malene(skoro jednake 1). Možemo zaključiti da je ovo dobar pokušaj u stvaranju modela za predviđanje plaće igrača.

Linearni regresijski model za predviđanje plaće koristeći godine igrača, Ovrl i TOI/GP

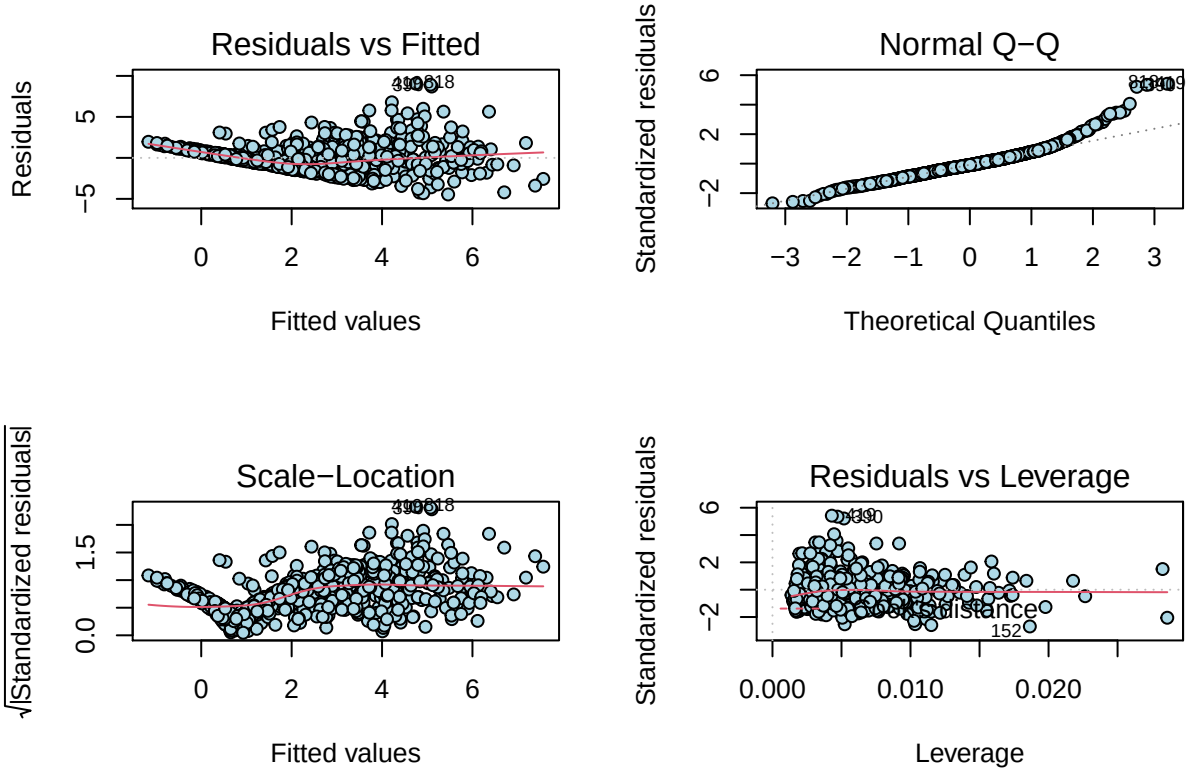
Ovaj model je nadogradnja na prethodni model s varijablom TOI/GP.

```
model5 = lm(Salary ~ playerAge + draftOverall + toiGp)

summary(model5)
```

```
##
## Call:
## lm(formula = Salary ~ playerAge + draftOverall + toiGp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4623 -1.0440 -0.1485  0.7912  9.0511
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.943367   0.403662 -17.201  < 2e-16 ***
## playerAge     0.218435   0.014260  15.318  < 2e-16 ***
## draftOverall -0.007634   0.001020  -7.487 1.99e-13 ***
## toiGp         0.267027   0.015296  17.457  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.674 on 745 degrees of freedom
##   (139 observations deleted due to missingness)
## Multiple R-squared:  0.5101, Adjusted R-squared:  0.5082
## F-statistic: 258.6 on 3 and 745 DF,  p-value: < 2.2e-16
```

```
car::vif(model5)
```

```
##    playerAge draftOverall        toiGp
##     1.098860     1.104217     1.102735
```



Iz ovog modela vidimo da su zbog malenih p-vrijednosti sve varijable statistički značajne te je dodavanjem varijable toiGp poboljšalo model. Iz naredbe vif(model5) možemo vidjeti da su varijable korištene u modelu međusobno nezavisne.

```
anova(model4,model5)
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ playerAge + draftOverall
## Model 2: Salary ~ playerAge + draftOverall + toiGp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    746 2942.5
## 2    745 2088.3  1    854.23 304.75 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Iz rezultata testa ANOVE možemo zaključiti da je model5 bolji. To možemo iščitati iz tablice gdje vidimo da je p-vrijednost jako mala , što znači da je dodavanje varijabli toiGp značajno poboljšalo model.

Odabir završnog modela

Postoje različiti atributi koji određuju plaću te se oni razlikuju s obzirom na kojoj poziciji igrač igra. Razdvajanjem modela po pozicijama smo uspjeli objasniti veći postotak odstupanja a samim time nam je model postao bolji za previđanje plaća.

Model za igrače napada

Za završni model koji ćemo koristiti u previđanju plaće igračima napada nadogradili smo prethodni model dodajući varijablu GS.G(Prosječni rezultat igrača)

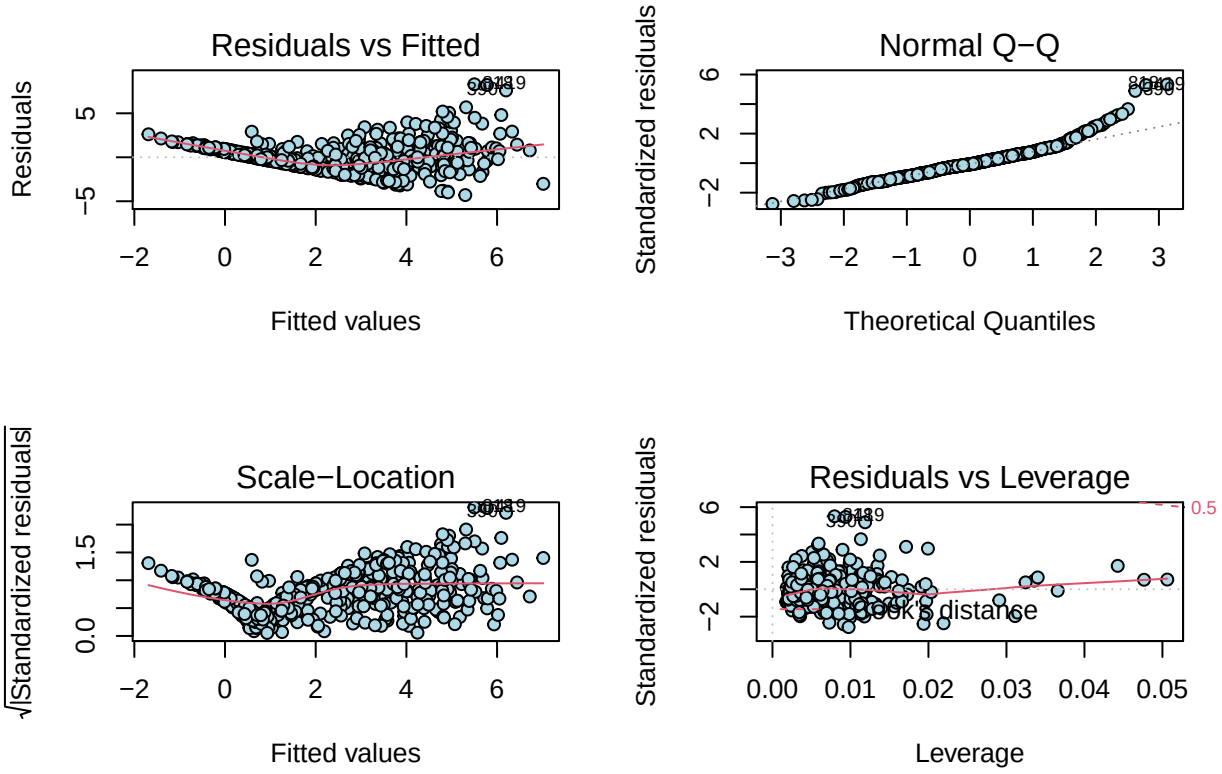
```
model7 = lm(Salary ~ playerAge  + toiGp+ GS.G,data = foward)

summary(model7)
```

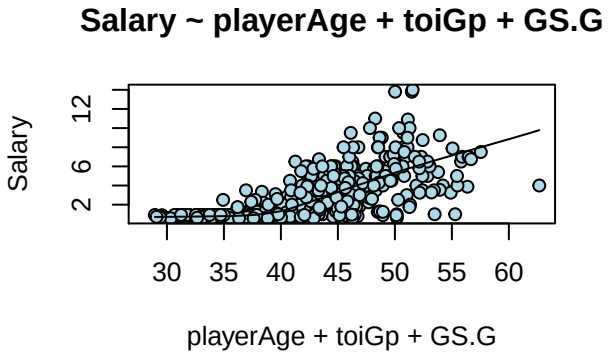
```
##
## Call:
## lm(formula = Salary ~ playerAge + toiGp + GS.G, data = foward)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2984 -0.9857 -0.1309  0.7730  8.2993
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.37210    0.47058 -15.666  < 2e-16 ***
## playerAge    0.17598    0.01472  11.952  < 2e-16 ***
## toiGp        0.32962    0.02903  11.354  < 2e-16 ***
## GS.G         1.29596    0.32512   3.986 7.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.563 on 574 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.5681, Adjusted R-squared:  0.5658
## F-statistic: 251.7 on 3 and 574 DF,  p-value: < 2.2e-16
```

```
car::vif(model7)
```

```
## playerAge     toiGp      GS.G
##  1.024429  2.492066  2.455021
```





Model za igrače obrane

Za završni model koji ćemo koristiti u previđanju plaće igračima obrane nadogradili smo prethodni model s dvije nove varijable , iFF(Pokušaji odblokiranog udarca) i iHA(sudari u tijelo od strane drugog igrača kako bi oduzeo "puck")

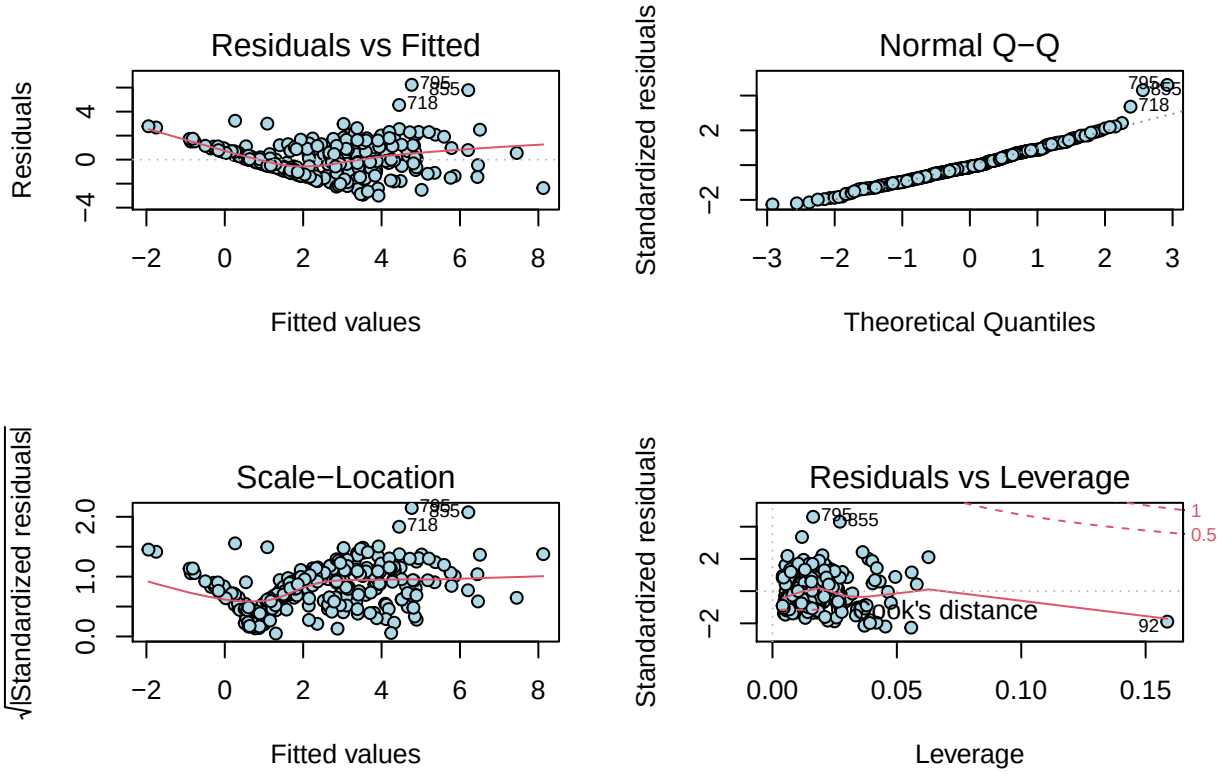
```
model8 = lm(Salary ~ playerAge  +toiGp + iFF +iHA
            ,data = defence)

summary(model8)
```

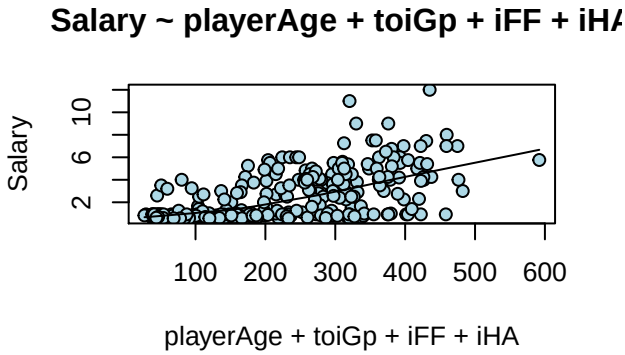
```
##
## Call:
## lm(formula = Salary ~ playerAge + toiGp + iFF + iHA, data = defence)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9966 -0.9121 -0.1268  0.8842  6.2305
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.081315   0.634646 -11.158  < 2e-16 ***
## playerAge    0.183928   0.019523   9.421  < 2e-16 ***
## toiGp        0.228764   0.032055   7.137 8.11e-12 ***
```

```
## iFF          0.008262   0.001636   5.049 7.97e-07 ***
## iHA         -0.007166   0.002208  -3.246  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.362 on 282 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.6071, Adjusted R-squared:  0.6015
## F-statistic: 108.9 on 4 and 282 DF,  p-value: < 2.2e-16
```

```
car::vif(model8)
```

```
## playerAge     toiGp       iFF       iHA
##  1.076596  2.653117  2.957304  1.753917
```





Svojstvo linearne veze možemo najlakše iščitati iz scatter plota gdje vizualno vidimo linearnu vezu između promatranih varijabla. Za svojstvo normalnosti možemo pogledati Q-Q graf , ako vrijednosti podataka u grafu padaju približno duž pravca, tada se podaci normalno distribuiraju.Budući da svaki podatak predstavlja drugog igrača svojstvo nezavisnosti je sigurno zadovoljeno. Provjeru za homoskedastičnost možemo provjeriti na grafu "residuals vs fitted" gdje gledamo crvenu liniju koja je otprilike poravnata s vrijednosti 0 na y osi , te je ključno da se reziduali ne povećavaju kako se povećavaju izračunate vrijednosti.Iz našeg primjera možemo zaključiti da heteroskedastičnost ne postoji te možemo prihvatiti ovaj regresijski model.

Predviđanje plaće igračima

Prikaz rezultata predviđanja plaće igračima napada

```
## [1] "3.5 mil (stvarna vrijednost)"
```

```
##        1
## 4.476994
```

```
## [1] "5.2 mil (stvarna vrijednost)"
```

```
##        1
## 2.316479
```

```
## [1] "0.715 mil (stvarna vrijednost)"
```

```
##        1
## 1.080765
```

Prikaz rezultata predviđanja plaće igračima obrane

```
## [1] "0.635 mil (stvarna vrijednost)"
```

```
##       1
## 1.03081
```

```
## [1] "7 mil(stvarna vrijednost)"
```

```
##        1
## 4.891231
```

```
## [1] "5.6 mil (stvarna vrijednost)"
```

```
##        1
## 4.621003
```