Statistička analiza podataka - projekt

Sapunanje

2023-01-12

Statistika nogometaša engleske Premier lige

Studenti: Karlo Boroš, Petar Novak, Vlado Perković i Mislav Rendulić

Asistent: Krunoslav Jurčić

Cilj ovoga projekta je uzeti dane podatke i iz njih probati izvući zaključke i faktore koji mogu utjecati na rezultat, broj golova i sl. Naravno, nije potrebno naglasiti važnost korištenja ispravnih testova te dobivanje rezultata koji su validni.

1. Sadržaj

2. Osnovna prilagodba podataka

Podatke je prvo potrebno učitati. Bitno je dobro ih proučiti kako ne bismo slučajno pogriješili u nekom zaključku. Nakon dobre analize možemo krenuti sa našim zadacima.

```
nogometasi <- read.csv('dataset.csv', encoding = "UTF-8", stringsAsFactors = F)
```

Nakon enkodiranja početnih podataka, postajala su odstupanja od stvarnih imena kod nekih igrača pa smo ta imena ručno ispravili.

3. Pregled sezone

Ekipe koje su se natjecale u Premier Ligi u sezoni 2021/2022

```
klubovi
```

```
##  [1] "Arsenal"                "Aston Villa"
##  [3] "Brentford"              "Brighton & Hove Albion"
##  [5] "Burnley"                "Chelsea"
##  [7] "Crystal Palace"         "Everton"
##  [9] "Leeds United"           "Leicester City"
## [11] "Liverpool"              "Manchester City"
## [13] "Manchester United"      "Newcastle United"
## [15] "Norwich City"           "Southampton"
## [17] "Tottenham Hotspur"      "Watford"
## [19] "West Ham United"        "Wolverhampton Wanderers"
```

Najbolji strijelci

```
najbolji_strijelci
```

```
##              Player              Team Gls Gls per 90 min
## 1     Mohamed Salah         Liverpool  23           0.75
## 2     Son Heung-min Tottenham Hotspur  23           0.69
## 3 Cristiano Ronaldo Manchester United  18           0.66
## 4        Harry Kane Tottenham Hotspur  17           0.47
## 5        Sadio Mané         Liverpool  16           0.51
```

Najbolji asistenti

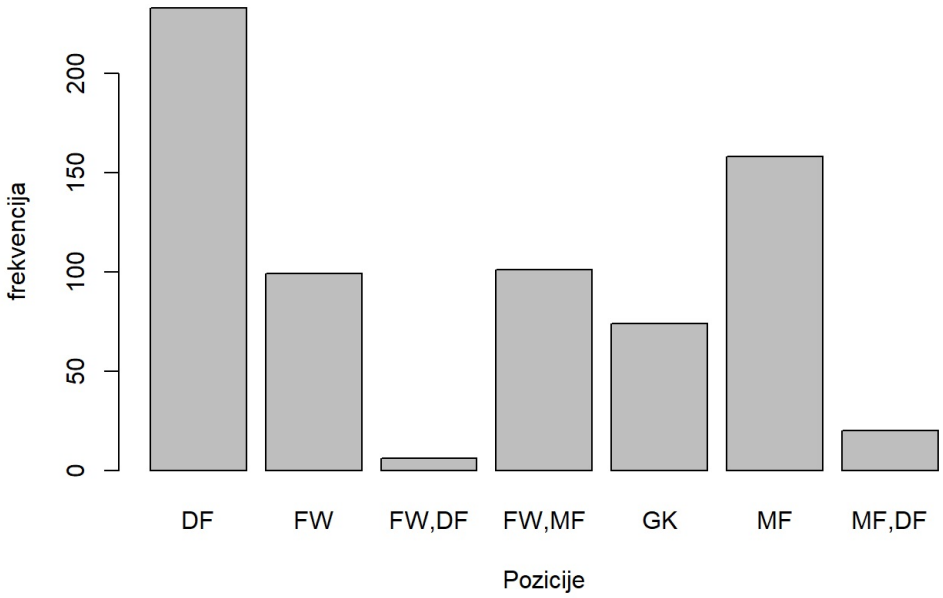
```
najbolji_asistenti
```

```
##                  Player            Team Ast Ast per 90 min
## 1          Mohamed Salah       Liverpool  13           0.42
## 2 Trent Alexander-Arnold       Liverpool  12           0.38
## 3            Mason Mount         Chelsea  10           0.38
## 4          Harvey Barnes  Leicester City  10           0.43
## 5       Andrew Robertson       Liverpool  10           0.35
## 6           Jarrod Bowen West Ham United  10           0.30
```

Pozicije igrača

Vizualizacija razdiobe igrača po pozicijama:

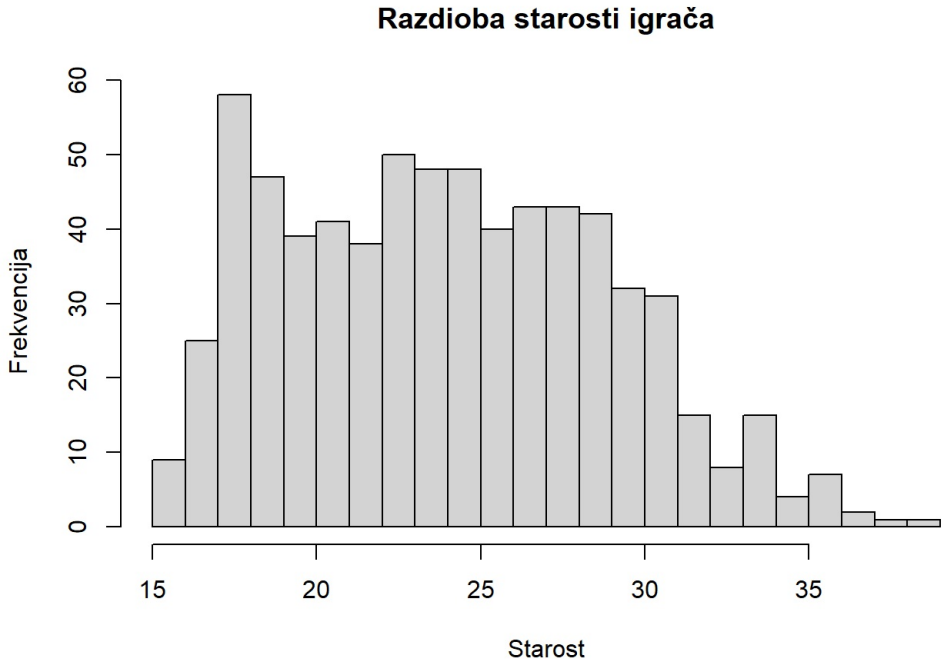
```
nogometasi %>% select(Pos) %>% summarise(uniPos = ifelse(Pos == "DF,FW", "FW,DF", ifelse(Pos == "MF,FW", "FW,MF",
ifelse(Pos == "DF,MF", "MF,DF", Pos)))) %>% arrange(uniPos) -> popravak
barplot(table(popravak), xlab = "Pozicije", ylab = "frekvencija")
```



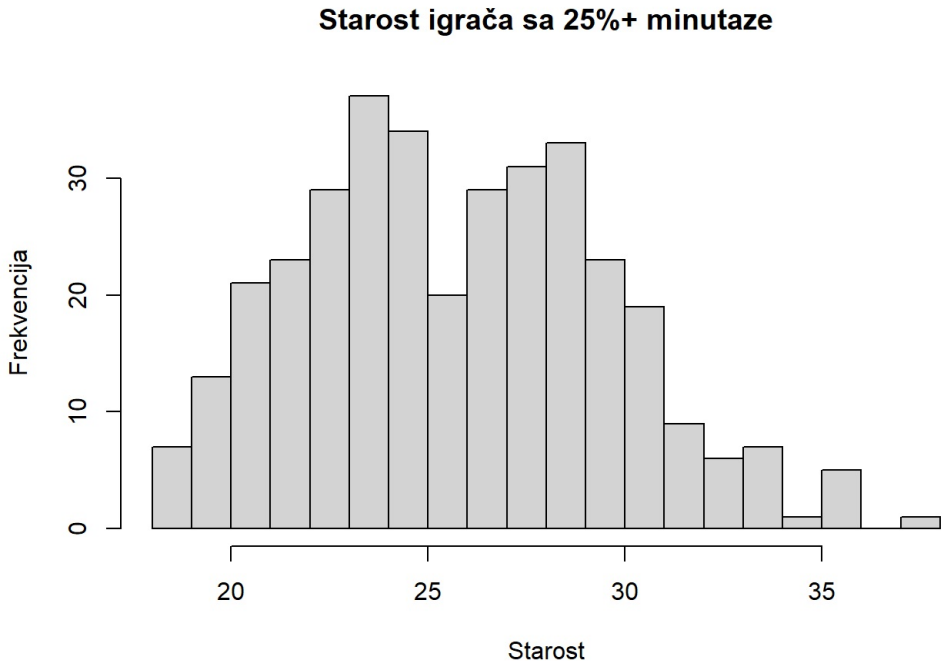
Primijetimo veliki broj obrambenih igrača što i ima smisla kada pogledamo da ekipe najčešće igraju s 4 igrača u obrani. Neki igrači su igrali pozicije beka i napadačkog krila pa spadaju u skupinu “FW,DF” koja je na prvi pogled dosta neuobičajena.

Godine igrača

```
distrStarosti <- hist(nogometasi$Age,
                      breaks = 20,
                      main="Razdioba starosti igrača",
                      xlab="Starost",
                      ylab='Frekvencija'
                      )
```




```
x <- nogometasi %>% filter(!is.na(X90s) & X90s >= 9.5)
distrStarosti <- hist(x$Age,
                      breaks = 20,
                      main="Starost igrača sa 25%+ minutaze",
                      xlab="Starost",
                      ylab='Frekvencija'
                      )
```



4. Postoji li razlika u broju odigranih minuta mladih igrača (do 25 godina) među premierligaškim ekipama?

Podijelimo igrače…

```
mladi <- nogometasi %>% filter(Age <= 25)
cat("Broj mladih igrača do 25 godina iznosi: ", nrow(mladi), "\n")
```

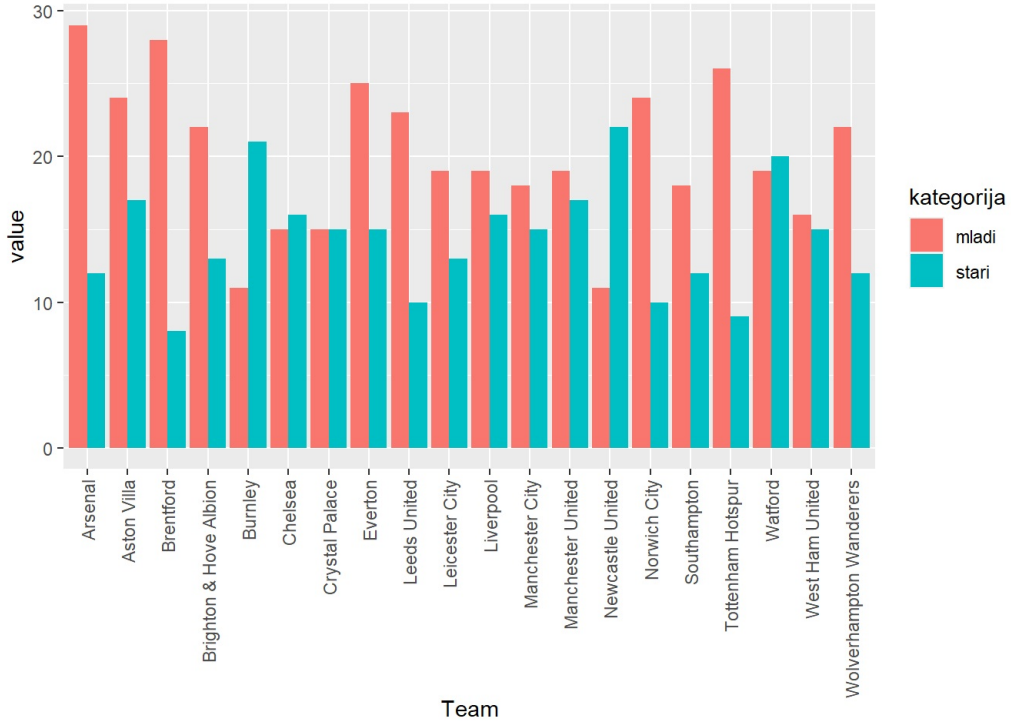
```
## Broj mladih igrača do 25 godina iznosi:  403
```

```
stari <- nogometasi %>% filter(Age > 25)
cat("Broj igrača iznad 25 godina iznosi: ", nrow(stari))
```

```
## Broj igrača iznad 25 godina iznosi:  284
```

Vizualizirajmo podjelu igrača u samim klubovima:

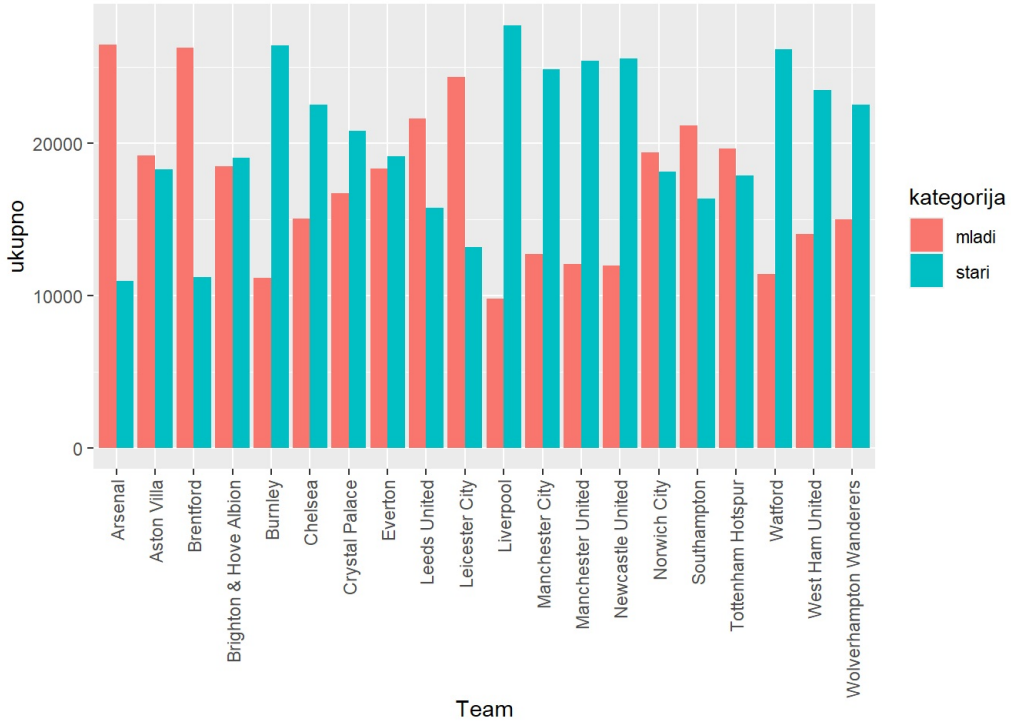
```
nogometasi_god <- nogometasi %>% summarise(mladi = ifelse(Age <= 25, 1, 0), Team) %>% group_by(Team) %>% summaris
e(mladi = sum(mladi, na.rm = T), stari = n() - mladi) %>% pivot_longer(cols = mladi:stari, names_to = "kategorija
")

ggplot(nogometasi_god, aes(x=Team, y=value, fill=kategorija)) +
    geom_bar(stat="identity", position="dodge") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



Vidimo da klubovi pretežno nastoje priključivati mlađe igrače u ekipu uz iznimke timova Burnley i Newcastle United.

Pogledajmo sada koliko te iste mlade igrače timovi zapravo i koriste…

```
nogometasi_min <- nogometasi %>% filter(!is.na(Age)) %>% summarise(Team, kategorija = ifelse(Age <= 25, "mladi",
"stari"), minutaza = X90s*90) %>% group_by(Team, kategorija) %>%  summarise(Team, kategorija, ukupno = sum(minuta
za, na.rm = T)) %>% unique()

ggplot(nogometasi_min, aes(x=Team, y=ukupno, fill=kategorija)) +
    geom_bar(stat="identity", position="dodge") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```

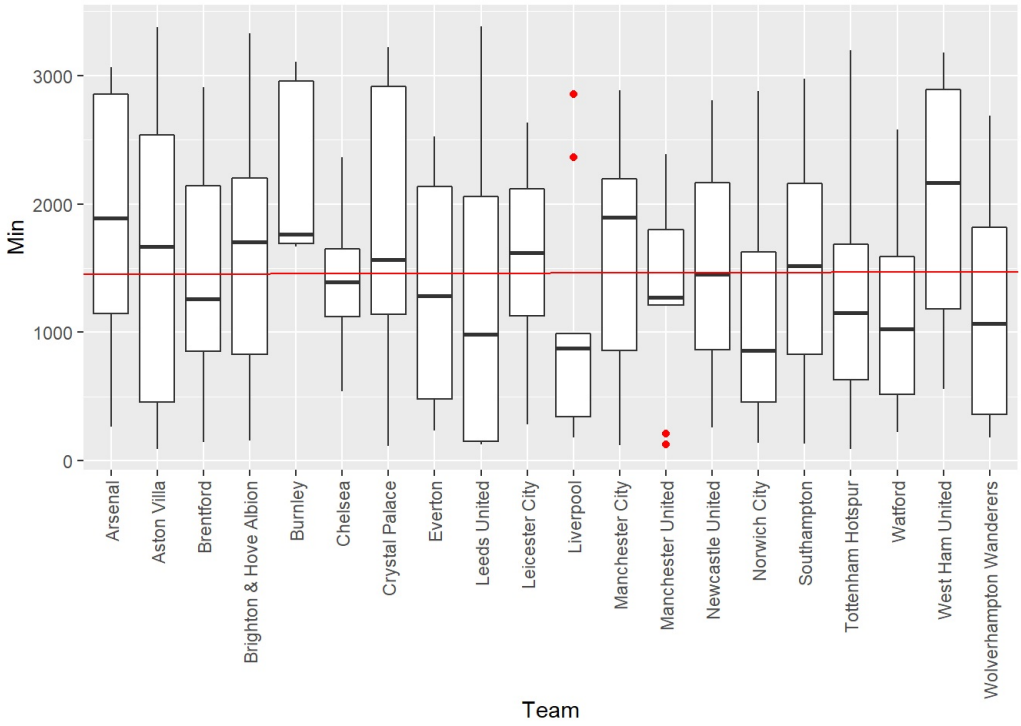


Kod analize u obzir ćemo uzeti mlade igrače koji su upisali barem 90 minuta.

```
mladi90 <- nogometasi %>% filter(Age <= 25 & Min >= 90)
```

Pogledajmo koliko su u prosjeku klubovi davali minuta svojim mladim igračima

```
ggplot(mladi90, aes(x = Team, y = Min)) +
  geom_boxplot(outlier.color = "red") +
  geom_abline(intercept = mean(mladi90$Min), col = "Red") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



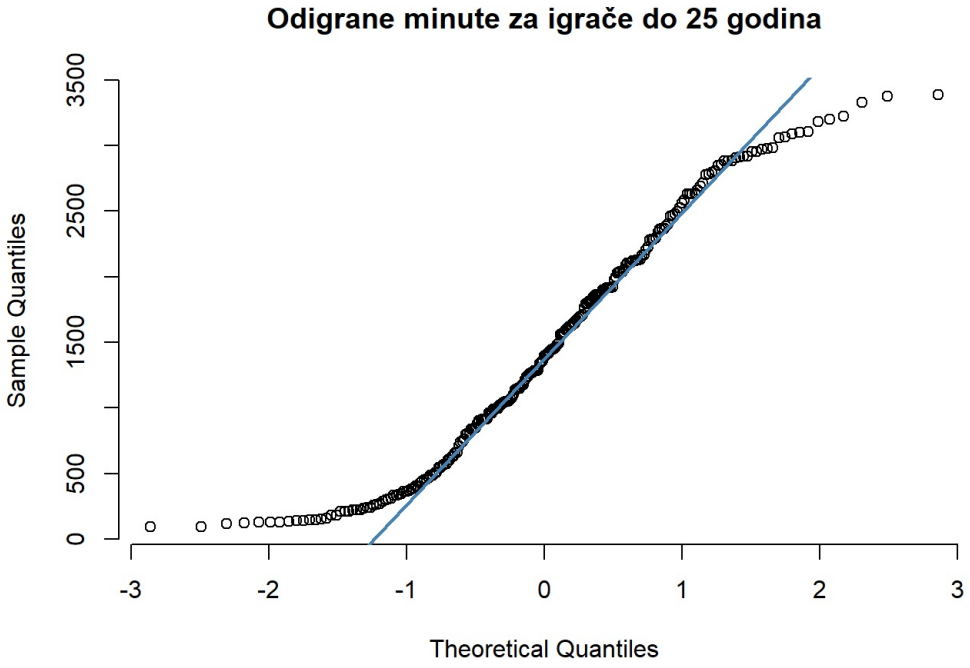
Testirat ćemo homogenost varijance raspodijele minuta mladih igraca po klubovima:

```
bartlett.test(mladi90$Min ~ mladi90$Team)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  mladi90$Min by mladi90$Team
## Bartlett's K-squared = 12.618, df = 19, p-value = 0.8575
```

Sada je potrebno testirati normalnost distribucije odigranih minuta za igrače do 25 godina ukupno i po klubovima:

```
qqnorm(mladi90$Min, pch = 1, frame = FALSE,main='Odigrane minute za igrače do 25 godina')
qqline(mladi90$Min, col = "steelblue", lwd = 2)
```




```
require(nortest)

lillie.test(mladi90$Min[mladi90$Team == "Arsenal"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Arsenal"]
## D = 0.18585, p-value = 0.2112
```

```
lillie.test(mladi90$Min[mladi90$Team == "Aston Villa"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Aston Villa"]
## D = 0.18316, p-value = 0.3232
```

```
lillie.test(mladi90$Min[mladi90$Team == "Brentford"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Brentford"]
## D = 0.12846, p-value = 0.6017
```

```
lillie.test(mladi90$Min[mladi90$Team == "Brighton & Hove Albion"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Brighton & Hove Albion"]
## D = 0.11693, p-value = 0.9435
```

```
lillie.test(mladi90$Min[mladi90$Team == "Burnley"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Burnley"]
## D = 0.34198, p-value = 0.05652
```

```
lillie.test(mladi90$Min[mladi90$Team == "Chelsea"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Chelsea"]
## D = 0.16879, p-value = 0.5134
```

```
lillie.test(mladi90$Min[mladi90$Team == "Crystal Palace"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Crystal Palace"]
## D = 0.24401, p-value = 0.127
```

```
lillie.test(mladi90$Min[mladi90$Team == "Leeds United"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Leeds United"]
## D = 0.17923, p-value = 0.1555
```

```
lillie.test(mladi90$Min[mladi90$Team == "Leicester City"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Leicester City"]
## D = 0.14925, p-value = 0.4919
```

```
lillie.test(mladi90$Min[mladi90$Team == "Liverpool"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Liverpool"]
## D = 0.31804, p-value = 0.009129
```

```
lillie.test(mladi90$Min[mladi90$Team == "Manchester City"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Manchester City"]
## D = 0.24065, p-value = 0.188
```

```
lillie.test(mladi90$Min[mladi90$Team == "Manchester United"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Manchester United"]
## D = 0.2188, p-value = 0.2451
```

```
lillie.test(mladi90$Min[mladi90$Team == "Newcastle United"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Newcastle United"]
## D = 0.13818, p-value = 0.9242
```

```
lillie.test(mladi90$Min[mladi90$Team == "Norwich City"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Norwich City"]
## D = 0.1879, p-value = 0.09276
```

```
lillie.test(mladi90$Min[mladi90$Team == "Southampton"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Southampton"]
## D = 0.13111, p-value = 0.7402
```

```
lillie.test(mladi90$Min[mladi90$Team == "Tottenham Hotspur"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Tottenham Hotspur"]
## D = 0.12067, p-value = 0.7712
```

```
lillie.test(mladi90$Min[mladi90$Team == "Watford"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Watford"]
## D = 0.13074, p-value = 0.8984
```

```
lillie.test(mladi90$Min[mladi90$Team == "West Ham United"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "West Ham United"]
## D = 0.20373, p-value = 0.5104
```

```
lillie.test(mladi90$Min[mladi90$Team == "Wolverhampton Wanderers"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mladi90$Min[mladi90$Team == "Wolverhampton Wanderers"]
## D = 0.2024, p-value = 0.1538
```

Na razini znacajnosti od 5% jedino Liverpool pravi probleme kod normalnosti. Iako varijanca i sredina ne odudaraju, uzorak ima izražene stršeće vrijednosti (Trent i Jota).

```
mladi90 %>% filter(Team == "Liverpool") %>% select(Player, Min)
```

```
##                   Player  Min
## 1 Trent Alexander-Arnold 2853
## 2              Diogo Jota 2364
## 3         Ibrahima Konaté  990
## 4               Luis Díaz  958
## 5            Curtis Jones  851
## 6         Kostas Tsimikas  877
## 7          Harvey Elliott  346
## 8               Joe Gomez  331
## 9      Caoimhín Kelleher  180
```

```
mladi90bezL <- mladi90 %>% filter(Team != "Liverpool")
```

Sada kada smo pretpostavili homogenost varijance, normalnost i nezavisnost provest ćemo ANOVA test:

H0: Raspodjela minuta igračima do 25 godina se ne razlikuje po klubovima

H1: Raspodijela minuta igračima do 25 godina razlikuje se u barem jednom klubu alpha = 0.05.

```
anova(lm(Min ~ Team, data = mladi90bezL))
```

```
## Analysis of Variance Table
##
## Response: Min
##            Df    Sum Sq Mean Sq F value Pr(>F)
## Team       18  17052162  947342  1.1298 0.3251
## Residuals 209 175244441  838490
```

Zaključci:

Ne odbacujemo nultu hipotezu da se raspodjela minuta razlikuje po klubovima.

Liverpool nismo uvrstili u test jer nismo mogli pretpostaviti normalnost, ali ni za tu ekipu ne možemo reći da značajno odstupa od prosjeka.

```
mean(mladi90$Min[mladi90$Team == "Liverpool"])
```

```
## [1] 1083.333
```

```
mean(mladi90$Min)
```

```
## [1] 1451.515
```

5. Dobivaju li u prosjeku više žutih kartona napadači ili igrači veznog reda?

Uzmimo za početak prosječne vrijednosti dobijenih žutih kartona kao motivaciju za statističko ispitivanje.

Moramo pripaziti na činjenicu da postoji podosta igrača s vrlo malo minuta odigrano, stoga ima smisla gledati igrače koji su u cijeloj sezoni sveukupno barem 50% minuta odigrali.

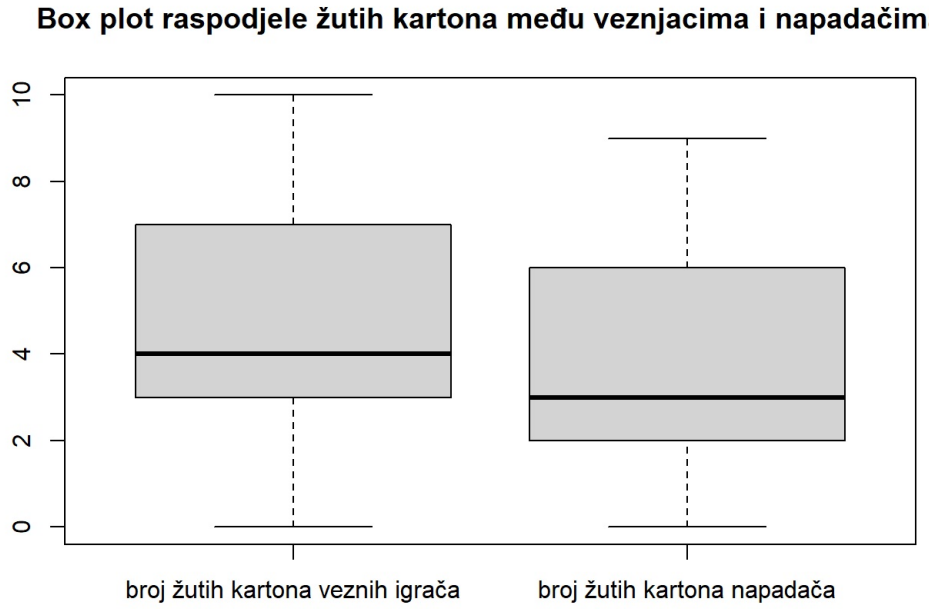
```
veznjaci <- nogometasi %>% filter(Pos == "MF" | Pos == "MF,FW" | Pos == "MF,DF") %>% filter(!is.na(X90s) & X90s >
= 18)
napadaci <- nogometasi %>% filter(Pos == "FW" | Pos == "FW,MF" | Pos == "FW,DF") %>% filter(!is.na(X90s) & X90s >
= 18)
cat("Prosječan broj žutih kartona igrača veznog reda iznosi: ", mean(veznjaci$CrdY, na.rm = T), "\n")
```

```
## Prosječan broj žutih kartona igrača veznog reda iznosi:  4.723077
```

```
cat("Prosječan broj žutih kartona napadača iznosi: ", mean(napadaci$CrdY, na.rm = T))
```

```
## Prosječan broj žutih kartona napadača iznosi:  3.765957
```

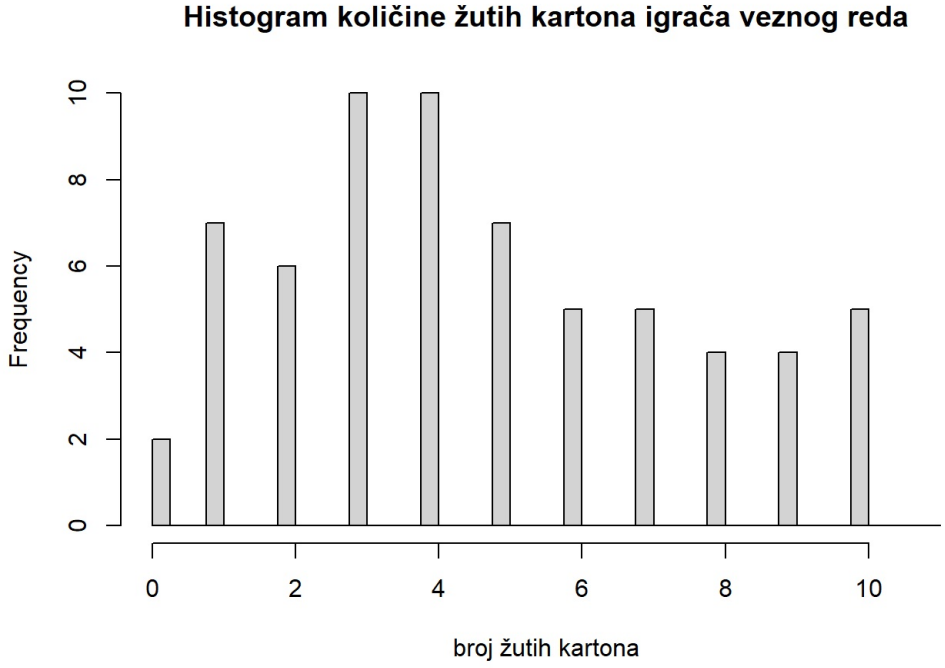
```
boxplot(veznjaci$CrdY, napadaci$CrdY,
        names = c('broj žutih kartona veznih igrača','broj žutih kartona napadača'),
        main='Box plot raspodjele žutih kartona među veznjacima i napadačima')
```



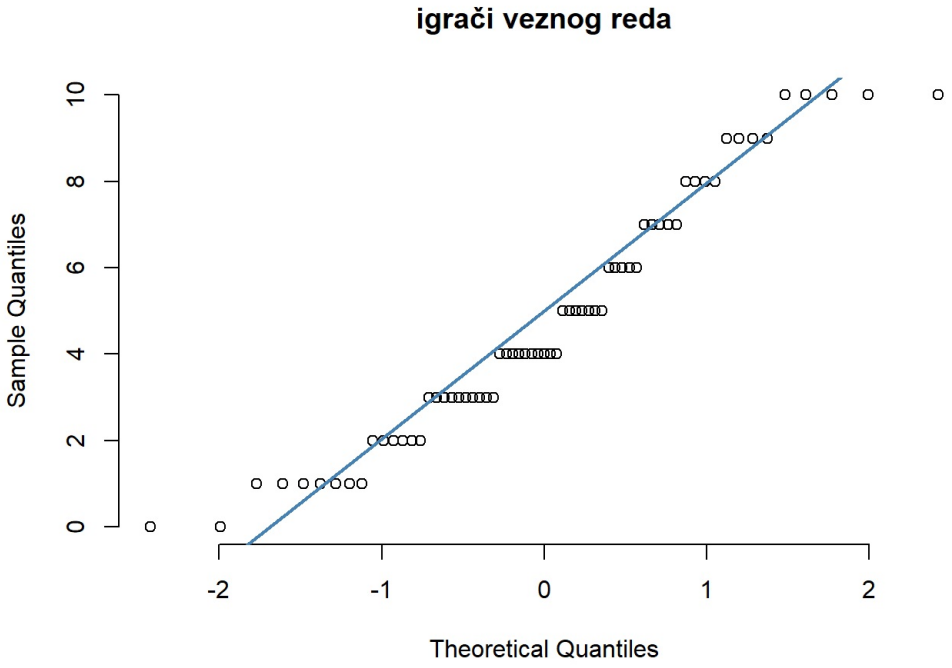
Vizualiziramo li podatke pomoću box plota dobijemo bolju sliku stvarne raspodjele žutih kartona u kojoj vidimo neke indikacije da bi mogla postojati razlika u broju žutih kartona. Ovakvo ispitivanje bismo mogli provesti klasičnim t-testom, no prvo se moramo uvjeriti da raspodjele kartona dolaze iz približno normalne razdiobe.

Normalnost ćemo provjeriti histogramom i qq plotom.

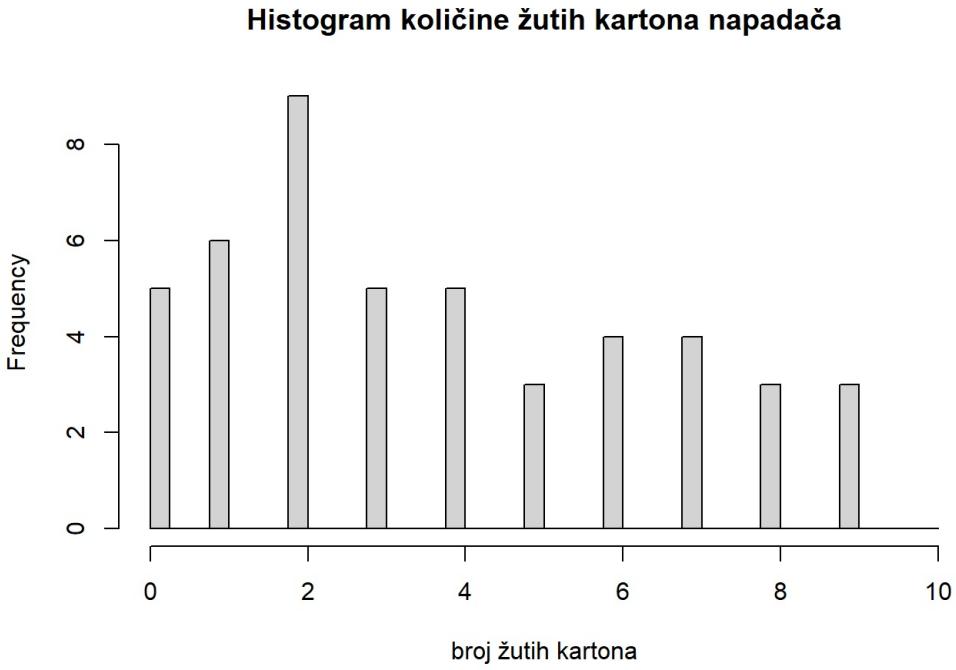
```
hist(veznjaci$CrdY,
     breaks=seq(min(veznjaci$CrdY, na.rm = T),max(veznjaci$CrdY, na.rm = T)+1,0.25),
     main='Histogram količine žutih kartona igrača veznog reda',
     xlab='broj žutih kartona')
```



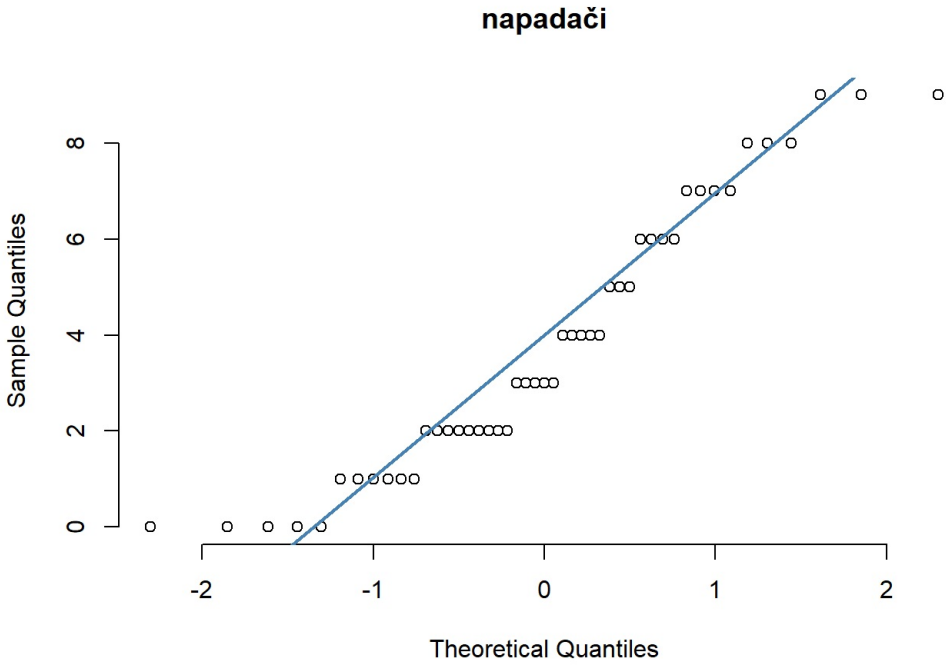

```
qqnorm(veznjaci$CrdY, pch = 1, frame = FALSE,main='igrači veznog reda')
qqline(veznjaci$CrdY, col = "steelblue", lwd = 2)
```




```
hist(napadaci$CrdY,
     breaks=seq(min(napadaci$CrdY, na.rm = T),max(napadaci$CrdY, na.rm = T)+1,0.25),
     main='Histogram količine žutih kartona napadača',
     xlab='broj žutih kartona')
```




```
qqnorm(napadaci$CrdY, pch = 1, frame = FALSE,main='napadači')
qqline(napadaci$CrdY, col = "steelblue", lwd = 2)
```



Budući da znamo da je t-test poprilično robustan, dajemo si za pravo koristiti ga iako gore prikazane razdiobe nisu distribuirane normalnom razdiobom, no nisu ni predaleko od iste.

Provjeravamo jesu li varijance uzoraka značajno različite:

```
cat("Varijanca broja žutih kartona kod veznjaka iznosi: ", var(veznjaci$CrdY), "\n")
```

```
## Varijanca broja žutih kartona kod veznjaka iznosi:  7.984615
```

```
cat("Varijanca broja žutih kartona kod napadača iznosi: ", var(napadaci$CrdY))
```

```
## Varijanca broja žutih kartona kod napadača iznosi:  7.617946
```

ispitajmo…

```
var.test(veznjaci$CrdY, napadaci$CrdY)
```

```
##
##  F test to compare two variances
##
## data:  veznjaci$CrdY and napadaci$CrdY
## F = 1.0481, num df = 64, denom df = 46, p-value = 0.876
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6024446 1.7798549
## sample estimates:
## ratio of variances
##           1.048132
```

Ne odbacujemo H0 koja kaže da su varijance jednake. Dakle koristit ćemo **t-test za dva uzorka s pretpostavkom jednakih varijanci**.

H0: broj žutih kartona između veznjaka i napadača je jednak.

H1: broj žutih kartona kod veznjaka veći je od onog kod napadača.

Odabir H1 motiviran je saznanjem da očekujemo da veznjaci imaju više žutih kartona.

```
t.test(veznjaci$CrdY, napadaci$CrdY, alt = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  veznjaci$CrdY and napadaci$CrdY
## t = 1.7863, df = 110, p-value = 0.03841
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.06828648        Inf
## sample estimates:
## mean of x mean of y
##  4.723077  3.765957
```

Budući da je p-value značajno malen, možemo odbaciti H0 u korist H1. Čak i ako bi koristili dvostrani test, svejedno bi odbacili našu hipotezu H0 u koristi H1.

Zaključci:

Na razini pouzdanosti od 95% odbacujemo H0 u korist H1, odnosno zaključujemo da je broj žutih kartona kod vezniaka veći od onog kod napadača.

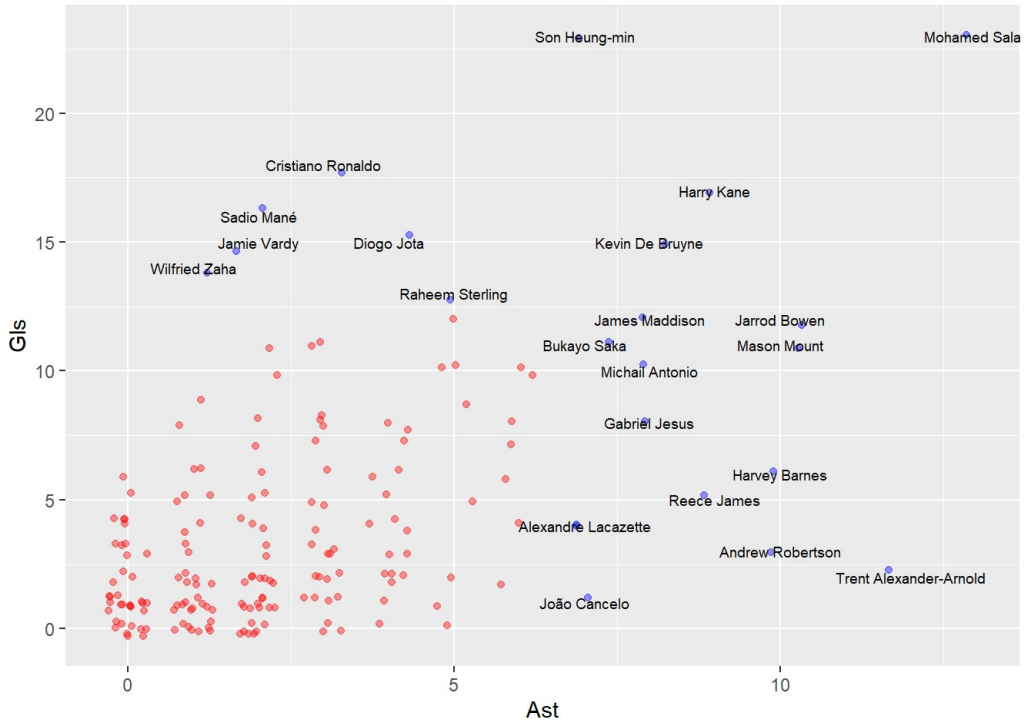
6. Možete li na temelju zadanih parametara odrediti uspješnost pojedinog igrača?

Što je zapravo uspješnost igrača? To je pitanje kojim smo se prvotno morali baviti i secirati što čini dobrog igrača ovisno o pozicijama.

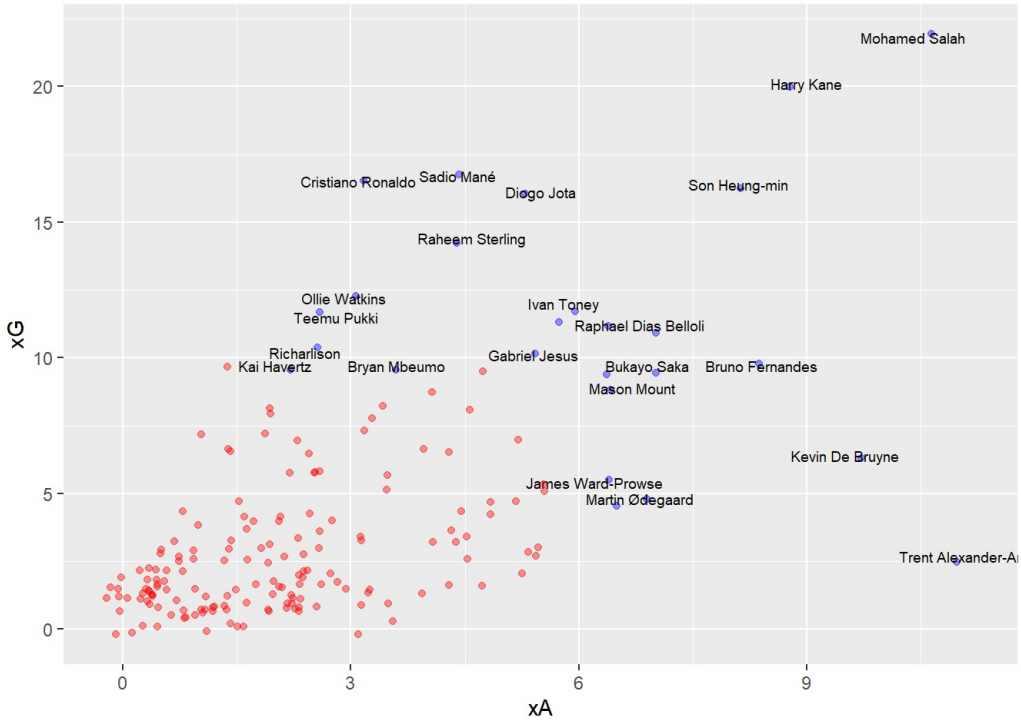
```
\newpage
```

Kao mjere uspješnosti igrača na raspolaganju imamo broj golova i broj asistencija. Naravno, nije objektivno uspoređivati obrambene, vezne i napadače prema broju golova tako da za neke pozicije sljedeća analiza nije najpogodnija.

```
nog <- nogometasi %>% filter(Pos != "GK") %>% filter(X90s >= 19)
dobri <- nog %>% filter(Ast > 6 | Gls > 12)
losi <- nog %>% filter(Ast <= 6 & Gls <= 12)
ggplot(dobri, aes(x = Ast, y = Gls)) +
  geom_jitter(width = 0.4, height = 0.4, alpha = 0.4, color="blue") +
  geom_text(aes(label = Player), check_overlap = T, size = 2.5) +
  geom_jitter(data = losi, aes(x = Ast, y = Gls), color="red", width = 0.3, height = 0.3, alpha = 0.4)
```




```
dobrixG <- nog %>% filter(xG > 9.5 | xA > 6)
losixG <- nog %>% filter(xG <= 9.5 & xA <= 6)
ggplot(dobrixG, aes(x = xA, y = xG)) + geom_jitter(color="blue",width = 0.3, height = 0.3, alpha = 0.4) + geom_te
xt(aes(label = Player), check_overlap = T, size = 2.5) + geom_jitter(data = losixG,aes(x = xA, y = xG), color="re
d", width = 0.3, height = 0.3, alpha = 0.4 )
```



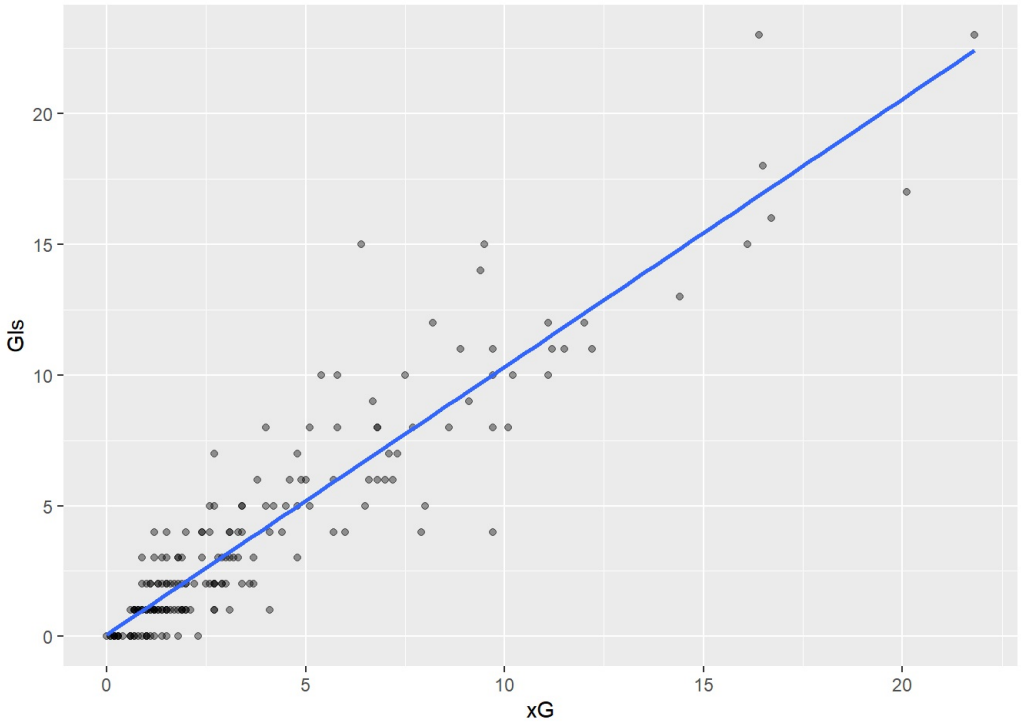
napomena: u gornja dva grafa dodan je *jitter* efekt kako bi se stekao bolji dojam količine točaka jer se koriste diskretni podaci

Određivanje uspješnosti po broju golova preko mjere očekivanih golova.

Osobi koja ne prati nogomet pojam očekivanih golova (xG) je možda nepoznat pa ćemo napomenuti da se radi o mjeri koja pokazuje procjenu vjerojatnosti u kojima neka prilika završi zgoditkom.

Gls/xG

```
ggplot(nog, aes(x = xG, y = Gls)) + geom_point(alpha = 0.4) + stat_smooth(method = lm, formula = y~x, se = F)
```



Prema grafu se da naslutiti da postoji jasna linearna veza između golova i očekivanih golova što daje motivaciju za daljne istraživanje.

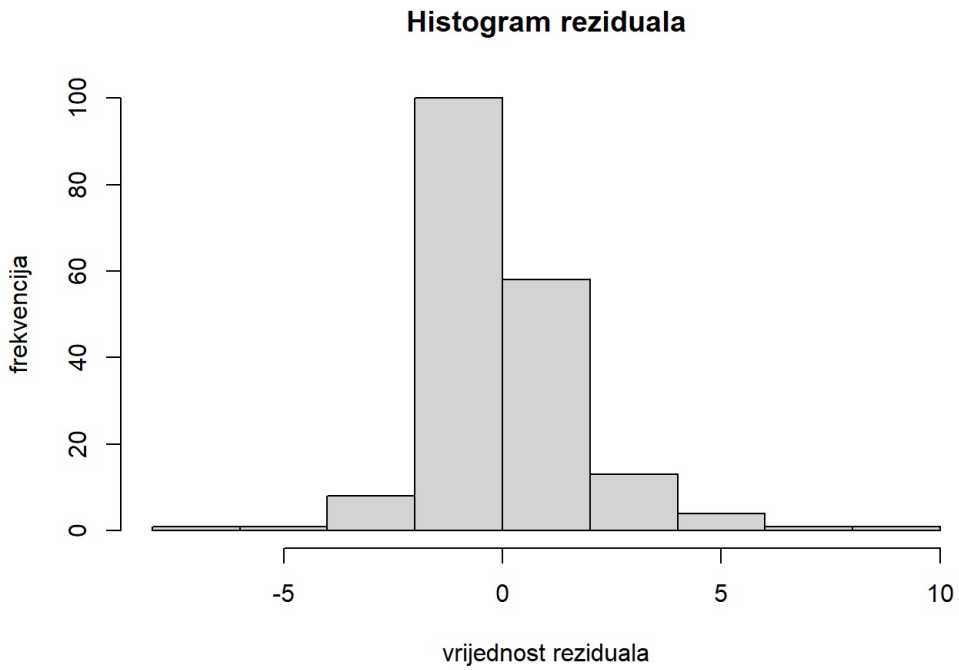
```
fit.gls = lm(Gls~xG,data=nog)
```

Potrebno je provjeriti jesu li narušene osnovne pretpostavke o rezidualima prije nego nastavimo dalje. Pretpostavke reziduala su normalnost i homogenost varijance.

Normalnost

Normalnost možemo provjeriti grafički pomoću histograma.

```
hist(fit.gls$residuals,
     main = "Histogram reziduala",
     xlab = "vrijednost reziduala",
     ylab = "frekvencija")
```



Statistički ju možemo provjeriti pomoću Kolmogorov-Smirnovljevog testa.

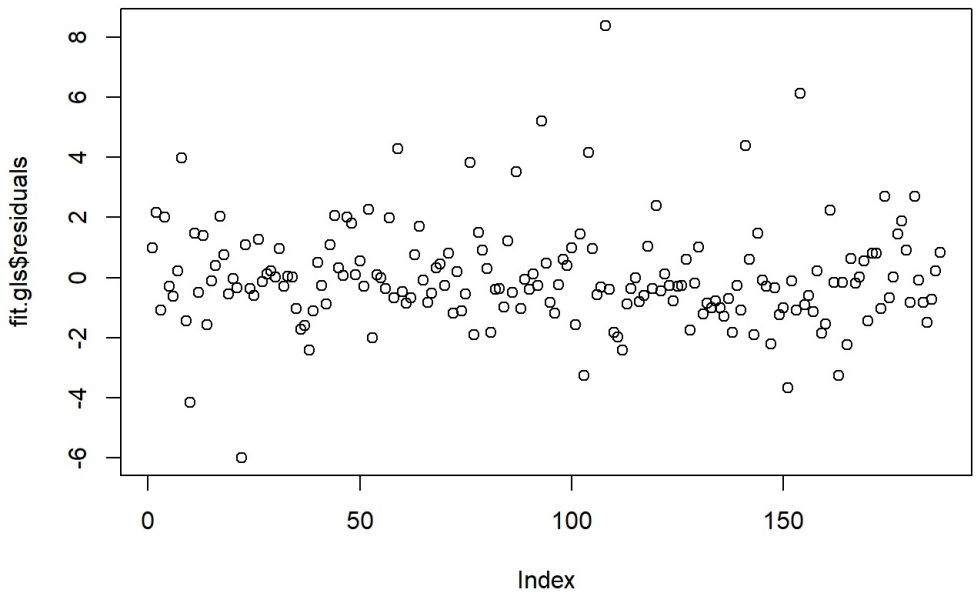
```
require(nortest)
lillie.test(fit.gls$residuals)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fit.gls$residuals
## D = 0.12357, p-value = 2.428e-07
```

Budući da reziduali ne pokazuju preveliko odstupanje od normalnosti (u smislu zakrivljenosti ili drugih razlika u distribuciji) te je poznato da je t-test robustan na (ne)normalnost - u analizi podataka se u ovakvim slučajevima i dalje mogu donositi statistički zaključci iz regresijskih modela.

Homogenost varijance provjerit ćemo grafički prikazom reziduala. Bitno nam je da se reziduali ne šire povećanjem y.

```
plot(fit.gls$residuals)
```



Pogledajmo rezultat analize…

```
summary(fit.gls)
```

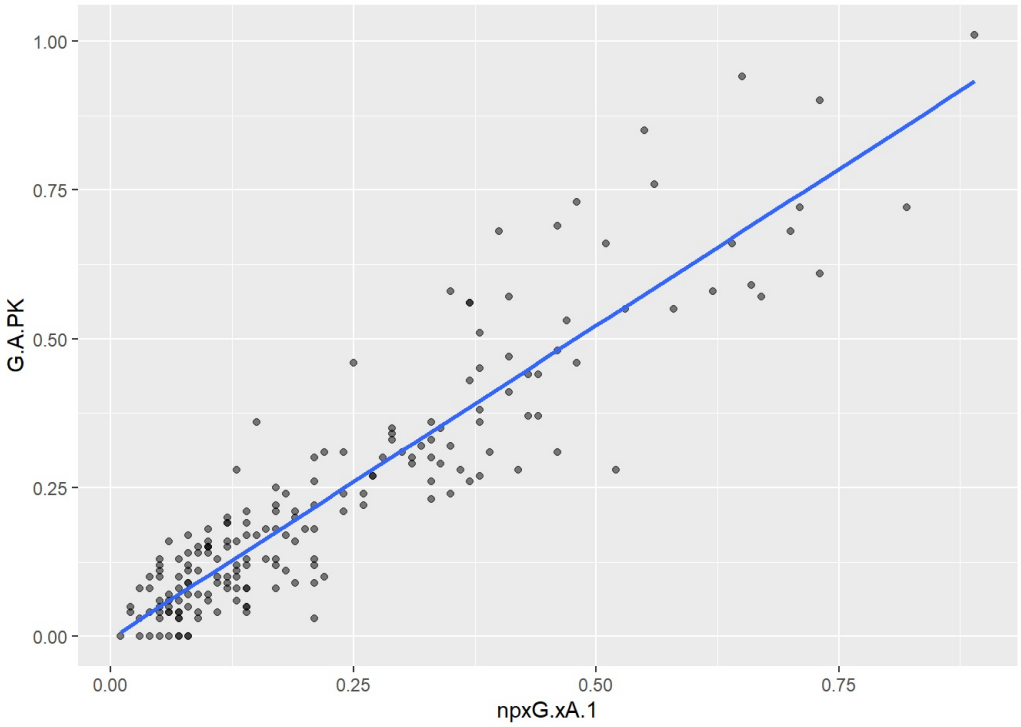
```
##
## Call:
## lm(formula = Gls ~ xG, data = nog)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0076 -0.8704 -0.2742  0.6911  8.3735
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06929    0.17266   0.401    0.689
## xG           1.02456    0.03110  32.941   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.696 on 185 degrees of freedom
## Multiple R-squared:  0.8543, Adjusted R-squared:  0.8536
## F-statistic:  1085 on 1 and 185 DF,  p-value: < 2.2e-16
```

Kao mjeru valjanosti linearne veze razmatramo varijablu R^2. Ona iznosi 0.854 što je dovoljno dobro za reći da mjerom xG relativno dobro možemo odrediti uspješnost igrača.

Određivanje uspješnosti po broju golova i asistencija bez kaznenih udaraca u po 90 min preko mjere očekivanih golova i asistencija bez kaznenih udaraca po 90 min.

G+A/npxG+xA

```
ggplot(nog, aes(x = npxG.xA.1, y = G.A.PK)) + geom_point(alpha=0.5) + stat_smooth(method = lm, formula = y~x, se
= F)
```



Možemo opravdano naslutiti da postoji jaka linearna veza između ovih mjera.

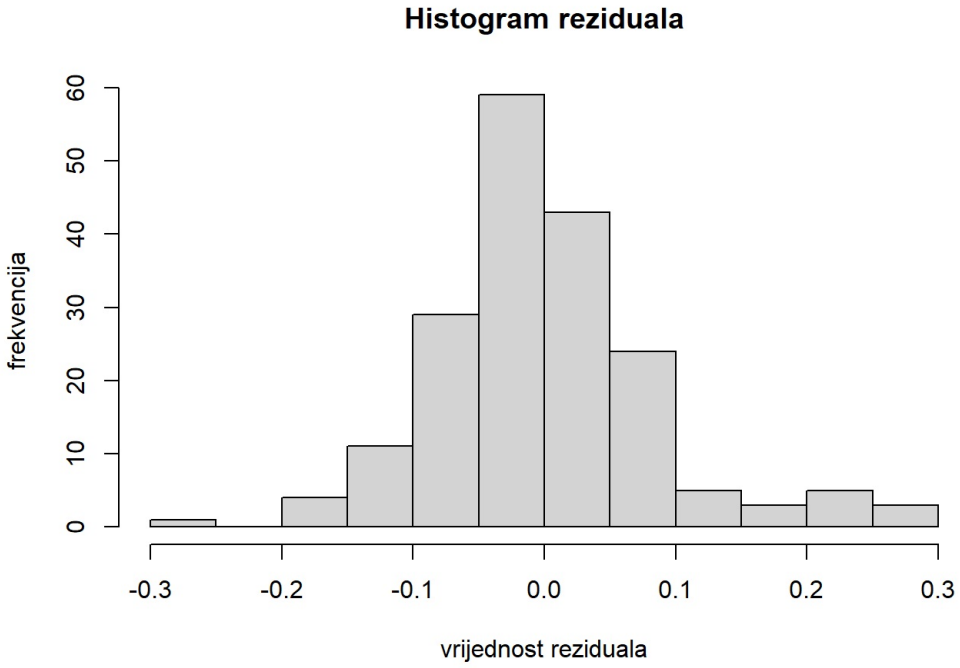
```
fit.ga = lm(G.A.PK~npxG.xA.1,data=nog)
```

Potrebno je provjeriti jesu li narušene osnovne pretpostavke o rezidualima prije nego nastavimo dalje. Pretpostavke reziduala su normalnost i homogenost varijance.

Normalnost

Normalnost možemo provjeriti grafički pomoću histograma.

```
hist(fit.ga$residuals,
     main = "Histogram reziduala",
     xlab = "vrijednost reziduala",
     ylab = "frekvencija")
```



Statistički ju možemo provjeriti pomoću Kolmogorov-Smirnovljevog testa.

```
require(nortest)
lillie.test(fit.ga$residuals)
```

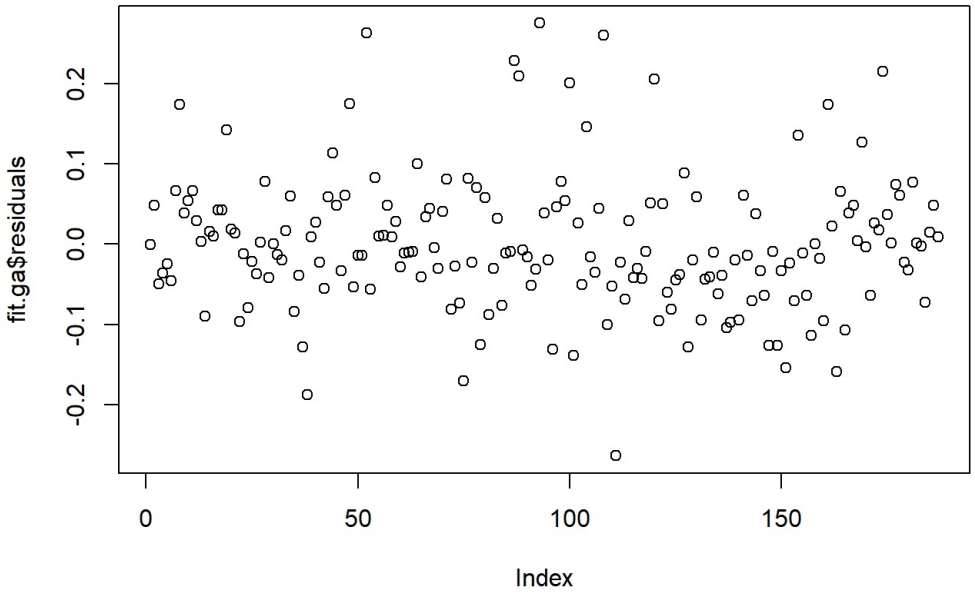
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fit.ga$residuals
## D = 0.079875, p-value = 0.005499
```

Budući da reziduali ne pokazuju preveliko odstupanje od normalnosti (u smislu zakrivljenosti ili drugih razlika u distribuciji) te je poznato da je t-test robustan na (ne)normalnost - u analizi podataka se u ovakvim slučajevima i dalje mogu donositi statistički zaključci iz regresijskih modela.

Homogenost

Homogenost varijance provjerit ćemo grafički prikazom reziduala. Bitno nam je da se reziduali ne šire povećanjem y.

```
plot(fit.ga$residuals)
```



Pogledajmo rezultat analize…

```
summary(fit.ga)
```

```
##
## Call:
## lm(formula = G.A.PK ~ npxG.xA.1, data = nog)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26343 -0.04521 -0.01004  0.04428  0.27504
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003017   0.009901  -0.305    0.761
## npxG.xA.1    1.050857   0.033811  31.080   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08462 on 185 degrees of freedom
## Multiple R-squared:  0.8393, Adjusted R-squared:  0.8384
## F-statistic:   966 on 1 and 185 DF,  p-value: < 2.2e-16
```

Kao mjeru valjanosti linearne veze razmatramo varijablu R^2. Ona iznosi 0.84 što opravdava naše izvorne pretpostavke.

7. Doprinose li sveukupnom uspjehu svoga tima više “domaći” igrači (tj. igrači engleske nacionalnosti) ili strani igrači?

Svi koji prate nogomet malo detaljnije znaju čiji igrači se cijene. Brazilci su najbolji dribleri, Španjolci najbolji u tiki-taki, Hrvati najbolji u penalima, ali u Engleskoj su najbolji Englezi. Javnost to zove *“English tax”* i time se cilja na činjenicu kako engleski klubovi skuplje plaćaju i prodaju domaće igrače u odnosu na strane. Je li to opravdano, pokazat će nam ANOVA. Koristit ćemo ju jer ćemo imati dvije skupine (strani i domaći igrači) gdje ćemo pretpostaviti jednakost te ćemo napokon saznati doprinose li oni sveukupnom uspjehu tima ili je to još jedna preuveličana engleska nogometna bajka. …

Prvi korak koji moramo napraviti je razdvojiti igrače po nacionalnosti, tj. odvojiti domaće igrače od stranih.

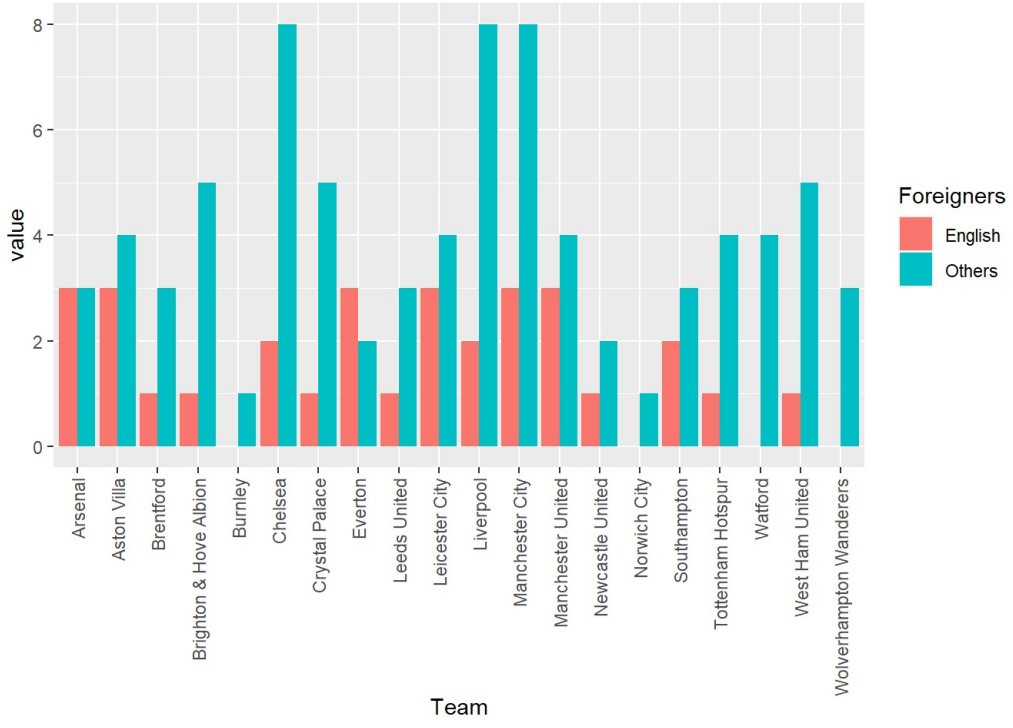
```
nogometasi$Foreigners <- ifelse(nogometasi$Nation=="ENG", "ENG", "Other")
```

Pošto kod procjene uspješnosti možemo samo koristiti golove i asistencije jer nam ostali podaci nisu dostupni, a želimo vidjeti uspješnost, u obzir ćemo uzeti igrače koji su imali barem pet golova ili asistencija u ligi i koji su odigrali barem 4 utakmice.

```
korisni <- nogometasi %>% filter(Gls+Ast>5 & X90s > 4)
```

Prvo želimo prikazati odnos količine stranih i domaćih igrača po klubovima

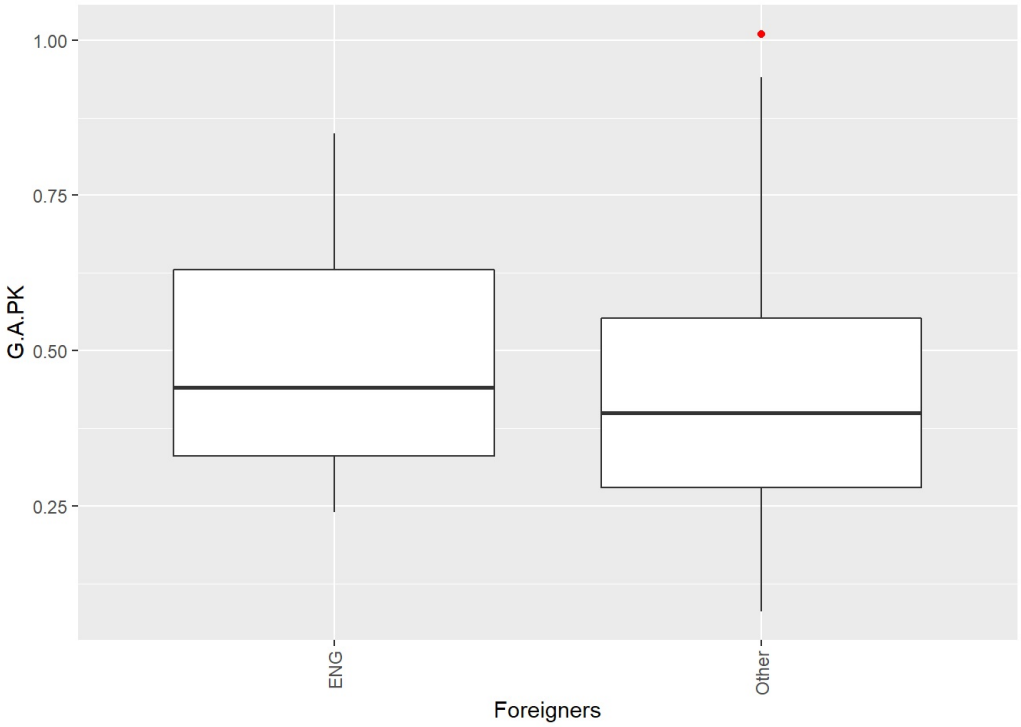
```
nogometasi_nat <- korisni %>% summarise(English = ifelse(Nation == "ENG", 1, 0), Team) %>% group_by(Team) %>% sum
marise(English = sum(English, na.rm = T), Others = n() - English) %>% pivot_longer(cols = English:Others, names_t
o = "Foreigners")

ggplot(nogometasi_nat, aes(x=Team, y=value, fill=Foreigners)) +
    geom_bar(stat="identity", position="dodge") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



Vidimo da većina klubova ima veći broj stranih igrača (uz već spomenute kriterije), što je i očekivano kada se u obzir uzme kako Engleska ima puno manji broj stanovnika od ostatka svijeta. Kada se maknu naši kriteriji, više nije tolika razlika, ali je još uvijek jasno vidljiva.

Pogledajmo koliko u prosjeku domaći i strani igrači imaju doprinos u golovima i asistencijama po utakmici:

```
ggplot(korisni, aes(x = Foreigners, y = G.A.PK)) +
  geom_boxplot(outlier.color = "red") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



Možemo vidjeti kako su vrijednosti slične. Sljedeći korak je testiranje homogenosti varijance raspodjele stranih i domaćih igrača u odnosu sa golovima i asistencijama:

```
bartlett.test(korisni$G.A.PK ~ korisni$Foreigners)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  korisni$G.A.PK by korisni$Foreigners
## Bartlett's K-squared = 0.63283, df = 1, p-value = 0.4263
```

Utvrdili smo da su rezultati homogeni. Sada je potrebno testirati normalnost distribucije golova i asistencija po nacionalnost (domaći <-> strani) :

```
require(nortest)
lillie.test(korisni$G.A.PK[korisni$Foreigners == "ENG"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  korisni$G.A.PK[korisni$Foreigners == "ENG"]
## D = 0.14143, p-value = 0.1178
```

```
lillie.test(korisni$G.A.PK[korisni$Foreigners == "Other"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  korisni$G.A.PK[korisni$Foreigners == "Other"]
## D = 0.11201, p-value = 0.01467
```

Sada kada smo pretpostavili homogenost varijance, normalnost i nezavisnost provest ćemo ANOVA test: **Nulta hipoteza:** Engleski nogometaši imaju jednak doprinos uspjehu svojeg tima kao i strani nogometaši. Za potrebe testa ćemo koristiti alpha = 0.05.

```
anova(lm(G.A.PK ~ Foreigners, data = korisni))
```

```
## Analysis of Variance Table
##
## Response: G.A.PK
##             Df Sum Sq  Mean Sq F value Pr(>F)
## Foreigners   1 0.0349 0.034853  0.9312 0.3367
## Residuals  109 4.0795 0.037427
```

Zaključci:

Ne odbacujemo nultu hipotezu da su engleski nogometaši uspješniji od stranih. Dobivena p-vrijednost je velika i to nam govori kako je naša prvotna pretpostavka bila točna.