SAP projekt

FMMJ

15.01.2024.

Učitavanje podatka i analiza dataseta

```
nba=read.csv("datasets/all_seasons.csv")
dim(nba)
```

```
## [1] 12844    22
```

```
names(nba)
```

```
##  [1] "X"                "player_name"      "team_abbreviation"
##  [4] "age"              "player_height"    "player_weight"
##  [7] "college"          "country"          "draft_year"
## [10] "draft_round"      "draft_number"     "gp"
## [13] "pts"              "reb"              "ast"
## [16] "net_rating"       "oreb_pct"         "dreb_pct"
## [19] "usg_pct"          "ts_pct"           "ast_pct"
## [22] "season"
```

```
print(head(nba, 5))
```

```
##   X      player_name team_abbreviation age player_height player_weight
## 1 0 Randy Livingston               HOU  22        193.04      94.80073
## 2 1 Gaylon Nickerson               WAS  28        190.50      86.18248
## 3 2     George Lynch               VAN  26        203.20     103.41898
## 4 3   George McCloud               LAL  30        203.20     102.05820
## 5 4     George Zidek               DEN  23        213.36     119.74829
##                 college country draft_year draft_round draft_number gp  pts reb
## 1       Louisiana State     USA       1996           2           42 64  3.9 1.5
## 2 Northwestern Oklahoma     USA       1994           2           34  4  3.8 1.3
## 3        North Carolina     USA       1993           1           12 41  8.3 6.4
## 4         Florida State     USA       1989           1            7 64 10.2 2.8
## 5                  UCLA     USA       1995           1           22 52  2.8 1.7
##   ast net_rating oreb_pct dreb_pct usg_pct ts_pct ast_pct  season
## 1 2.4        0.3    0.042    0.071   0.169  0.487   0.248 1996-97
## 2 0.3        8.9    0.030    0.111   0.174  0.497   0.043 1996-97
## 3 1.9       -8.2    0.106    0.185   0.175  0.512   0.125 1996-97
## 4 1.7       -2.7    0.027    0.111   0.206  0.527   0.125 1996-97
## 5 0.3      -14.1    0.102    0.169   0.195  0.500   0.064 1996-97
```

Vidimo da dataset sadrži igrače i sezonu u kojoj su igrali kao i njihove statistike u toj sezoni.

1.Razlikuje li se broj poena igrača po sezoni kroz različita desetljeća?

Prvotno, čistimo podatke i provjeravamo jesu li neki dijelovi neispravni ili nepotrebni. Prvo provjeravamo postoje li nedostajuće vrijednosti. Za odgovor na ovo pitanje nam također ne trebaju sve varijable.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
sum(!complete.cases(nba))
```

```
## [1] 0
```

```
nba1 = select (nba,c("player_name","gp", "pts","season"))
print(head(nba1, 5))
```

```
##         player_name gp  pts  season
## 1 Randy Livingston 64  3.9 1996-97
## 2 Gaylon Nickerson  4  3.8 1996-97
## 3     George Lynch 41  8.3 1996-97
## 4   George McCloud 64 10.2 1996-97
## 5     George Zidek 52  2.8 1996-97
```

Nema nedostajućih vrijednosti.

Prvo ćemo vrijednosti varijable "season", tipa string, odvojiti s operatorom "-" i pretvoriti u integer.

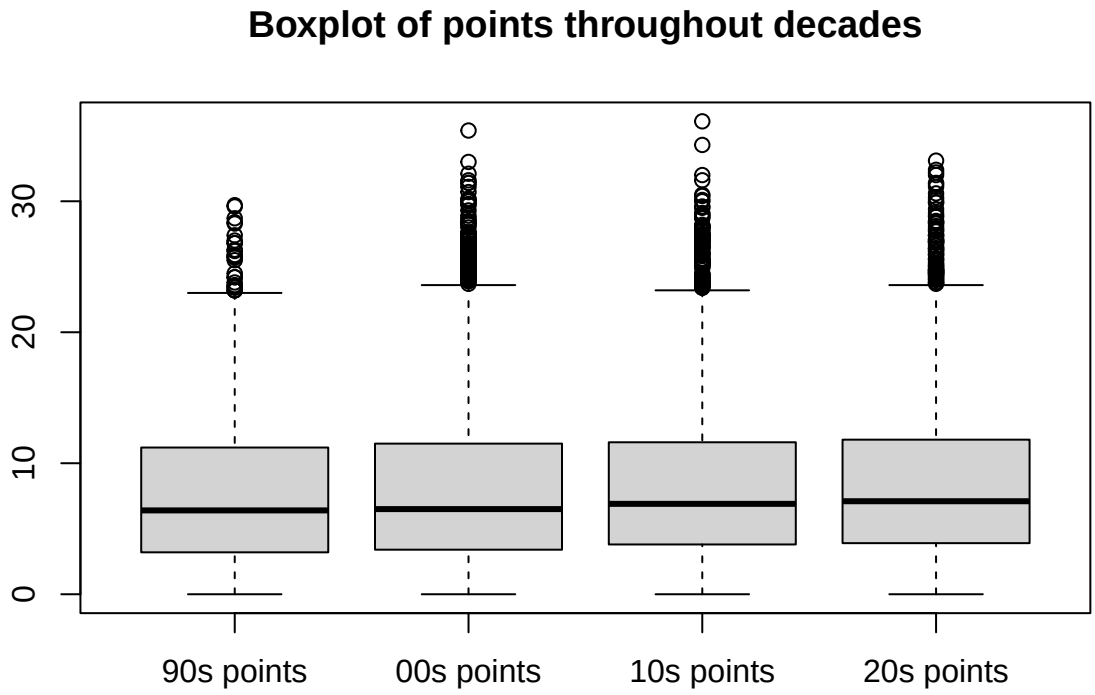
```
x <- list()
year_season <- nba1$season
year_season <- strsplit(year_season,split="-",fixed=T)
for (i in seq_along(year_season)) {
  x<-append(x,year_season[[i]][1])
}
nba1$"year_season"<-strtoi(x) #pretvaramo u integer
```

Sada ćemo odvojiti tablicu u različite tablice po desetljećima kako bi ih mogli uspoređivati.

```
ind= which(nba1$year_season < 2000)
nba90s = nba1[ind,]

ind=which(nba1$year_season > 1999 & nba1$year_season <2010)
nba00s = nba1[ind,]

ind=which(nba1$year_season > 2009 & nba1$year_season <2020)
nba10s = nba1[ind,]

ind=which(nba1$year_season > 2019 & nba1$year_season <2030)
nba20s = nba1[ind,]
```

Idemo usporediti podatke za različita desetljeća.

```
boxplot(nba90s$pts, nba00s$pts, nba10s$pts, nba20s$pts,
        names = c('90s points','00s points','10s points', '20s points'),
        main='Boxplot of points throughout decades')
```



Prema boxplotu vidimo da su 2010te i 2020te imaju ponešto veći prvi kvartil, ali ne možemo zaključiti da je ta razlika značajna. Ove podatke također možemo vidjeti iz deskriptivne statistike.

```
summary(nba90s$pts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.20    6.40    7.83   11.20   29.70
```

```
summary(nba00s$pts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.400   6.500   8.103  11.500  35.400
```

```
summary(nba10s$pts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.800   6.900   8.266  11.600  36.100
```

```
summary(nba20s$pts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.900   7.100   8.747  11.800  33.100
```

S obzirom da uspoređujemo više populacija kojima je glavna razlika desetljeće - sezona igre, koristit ćemo ANOVU (analizu varijance). Pretpostavit ćemo da su populacije nezavisne tj. glavna nezavisna varijabla je desetljeće.

Potrebno je još provjeriti jesu li populacije normalno distribuirane i ispitati homogenost varijanci. Za

nezavisnost koristimo Lillieforsovu inačicu KS testa.

$$H_0 : \text{Podaci su normalno distribuirani.}$$
$$H_1 : \text{Podaci nisu normalno distribuirani.}$$

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(nba90s$pts)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nba90s$pts
## D = 0.1035, p-value < 2.2e-16
```

```
lillie.test(nba00s$pts)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nba00s$pts
## D = 0.10988, p-value < 2.2e-16
```

```
lillie.test(nba10s$pts)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nba10s$pts
## D = 0.10054, p-value < 2.2e-16
```

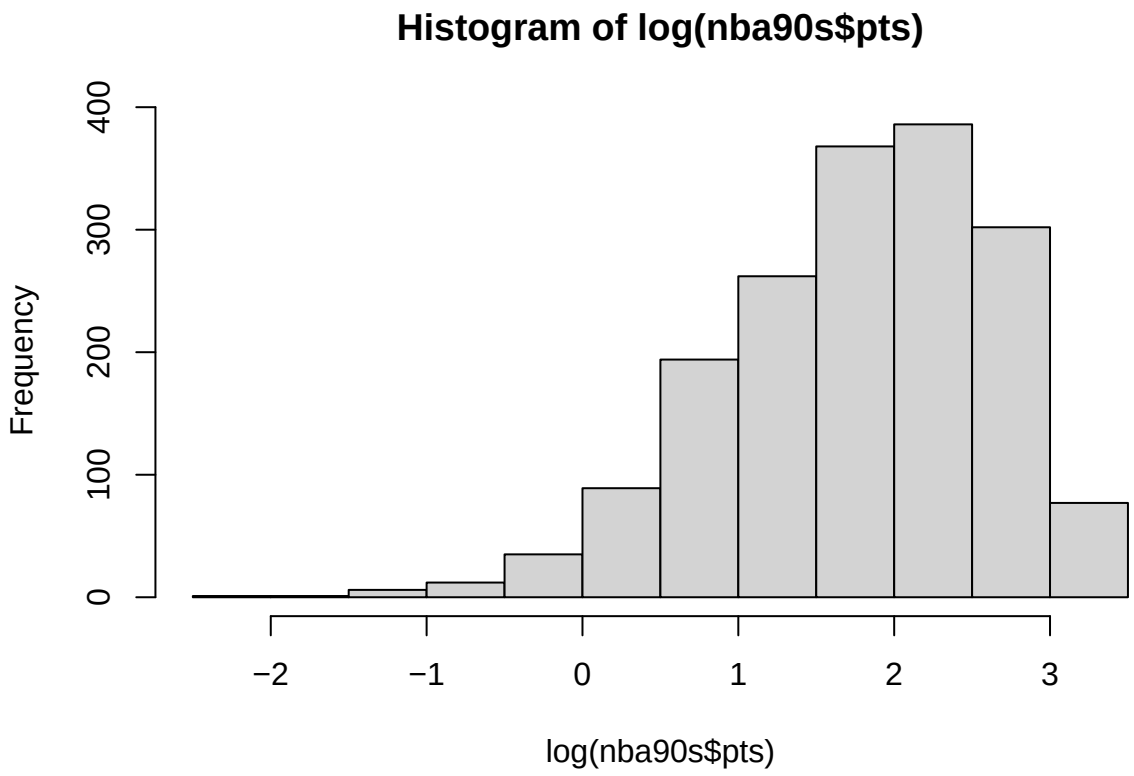
```
lillie.test(nba20s$pts)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nba20s$pts
## D = 0.11262, p-value < 2.2e-16
```

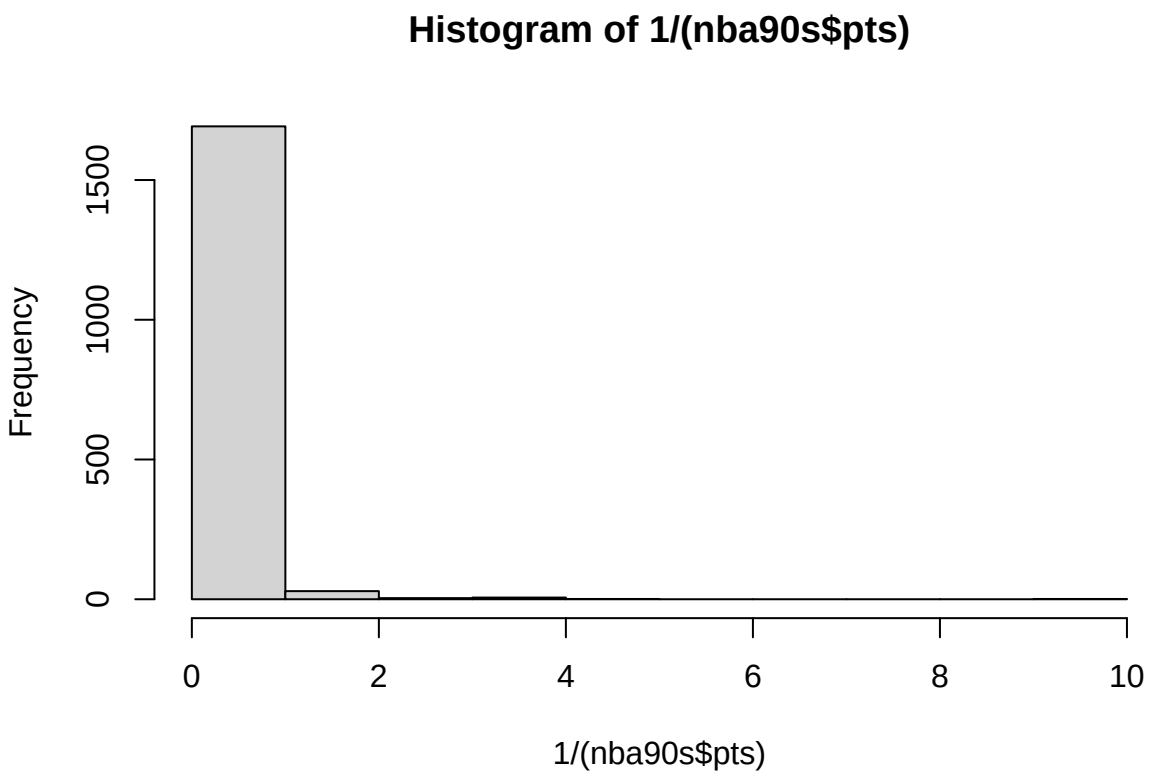
Prema Lillieforsovom testu, p- vrijednost je iznimno mala i zato odbacujemo nul-hipotezu. Populacije nisu normalno distribuirane.

Možemo probati transformacijom podataka doći do homogenosti i normalnosti.

```
hist(log(nba90s$pts))
```




```r
hist(1/(nba90s$pts))
```



Unatoč transformacijama, dataset još uvijek nije normalno distribuiran. S obzirom da nisu ispunjeni uvjeti za običnu ANOVU, pristupamo neparametarskoj ANOVI - Kruskal Wallis testu.

H_0 : Poeni igrača se ne razlikuju kroz različita desetljeća.

H_1 : Poeni igrača se razlikuju kroz različita desetljeća.

```
kruskal_test_result <- kruskal.test(list(nba90s$pts, nba00s$pts, nba10s$pts, nba20s$pts))
print(kruskal_test_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(nba90s$pts, nba00s$pts, nba10s$pts, nba20s$pts)
## Kruskal-Wallis chi-squared = 23.071, df = 3, p-value = 3.903e-05
```

```
kruskal_test_result <- kruskal.test(list(nba90s$pts, nba00s$pts))
print(kruskal_test_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(nba90s$pts, nba00s$pts)
## Kruskal-Wallis chi-squared = 1.6137, df = 1, p-value = 0.204
```

```
kruskal_test_result <- kruskal.test(list(nba00s$pts, nba10s$pts))
print(kruskal_test_result)
```

```
##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  list(nba00s$pts, nba10s$pts)
## Kruskal-Wallis chi-squared = 7.9644, df = 1, p-value = 0.004771
```

```
kruskal_test_result <- kruskal.test(list(nba10s$pts, nba20s$pts))
print(kruskal_test_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(nba10s$pts, nba20s$pts)
## Kruskal-Wallis chi-squared = 2.1872, df = 1, p-value = 0.1392
```

#zaključak

S obzirom da je p-vrijednost mala, postoji značajna razlika u bodovima između desetljeća i odbacujemo nul-hipotezu te prihvaćamo alternativu. Vidimo da je značajna razlika između 2000-ih i 2010-ih i obzirom da 2010-te i 2020-te nemaju značajnu razliku, možemo zaključiti da se broj poena povećava samo u prijelazu s 2000-ih na 2010te.

2.Postoji li značajna statistička razlika u visini igrača koji igraju za ekipe zapadne od igrača koji igraju za ekipe istočne konferencije?

Razdvojimo dataset na "konferencijske datasetove"

```
library(dplyr)

west_conference <- c("DAL", "DEN", "GSW", "HOU", "LAC", "LAL", "MEM",
                     "MIN", "NOP", "OKC", "PHX", "POR", "SAC", "SAS")

east_conference <- c("ATL", "BOS", "BKN", "CHA", "CHI", "CLE", "DET",
                     "IND", "MIA", "MIL", "NYK", "ORL", "PHI", "TOR", "WSH")

all.seasons.west <- nba %>%
  filter(team_abbreviation %in% west_conference) %>%
  group_by(player_name) %>%
  distinct(player_name, .keep_all = TRUE)

all.seasons.east <- nba %>%
  filter(team_abbreviation %in% east_conference) %>%
  group_by(player_name) %>%
  distinct(player_name, .keep_all = TRUE)
```

Napravimo deskriptivnu analizu

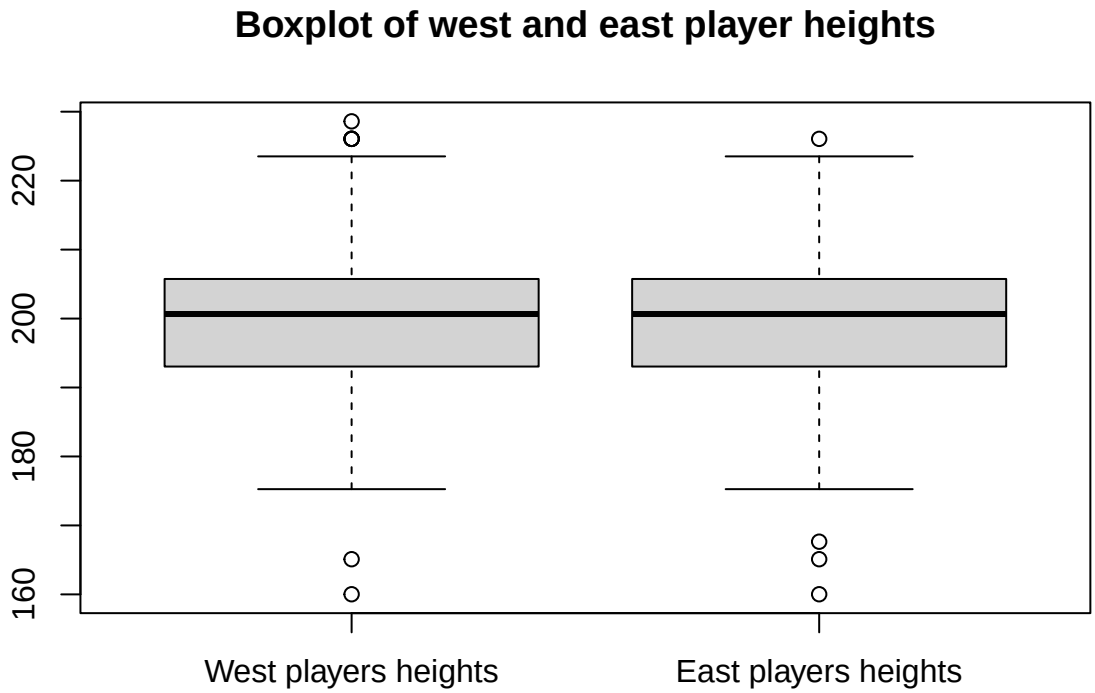
```
cat('Prosječna visina igrača istoka ', mean(all.seasons.east$player_height),'\n')
```

```
## Prosječna visina igrača istoka  200.1878
```

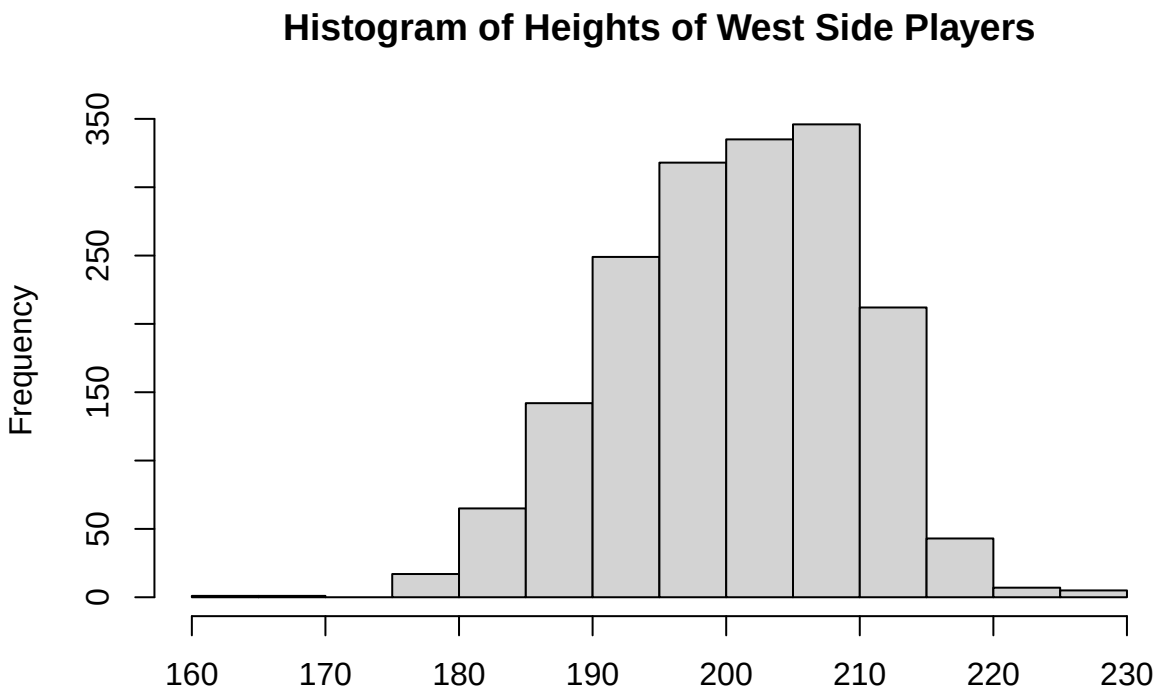
```
cat('Prosječna visina igrača zapada ', mean(all.seasons.west$player_height), '\n')
```

```
## Prosječna visina igrača zapada  200.0159
```

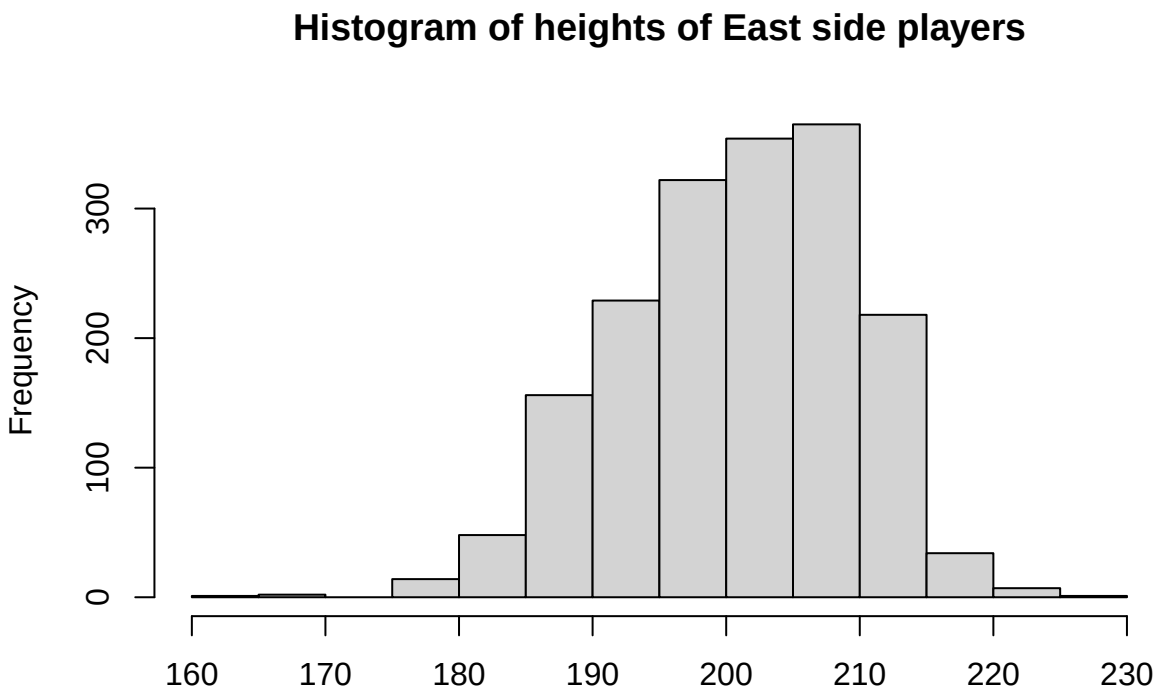
```
boxplot(all.seasons.west$player_height, all.seasons.east$player_height,
        names = c('West players heights','East players heights'),
        main='Boxplot of west and east player heights')
```




```
hist(all.seasons.west$player_height,
     main='Histogram of Heights of West Side Players',
     xlab='')
```




```
hist(all.seasons.east$player_height,
     main='Histogram of heights of East side players',
     xlab='')
```



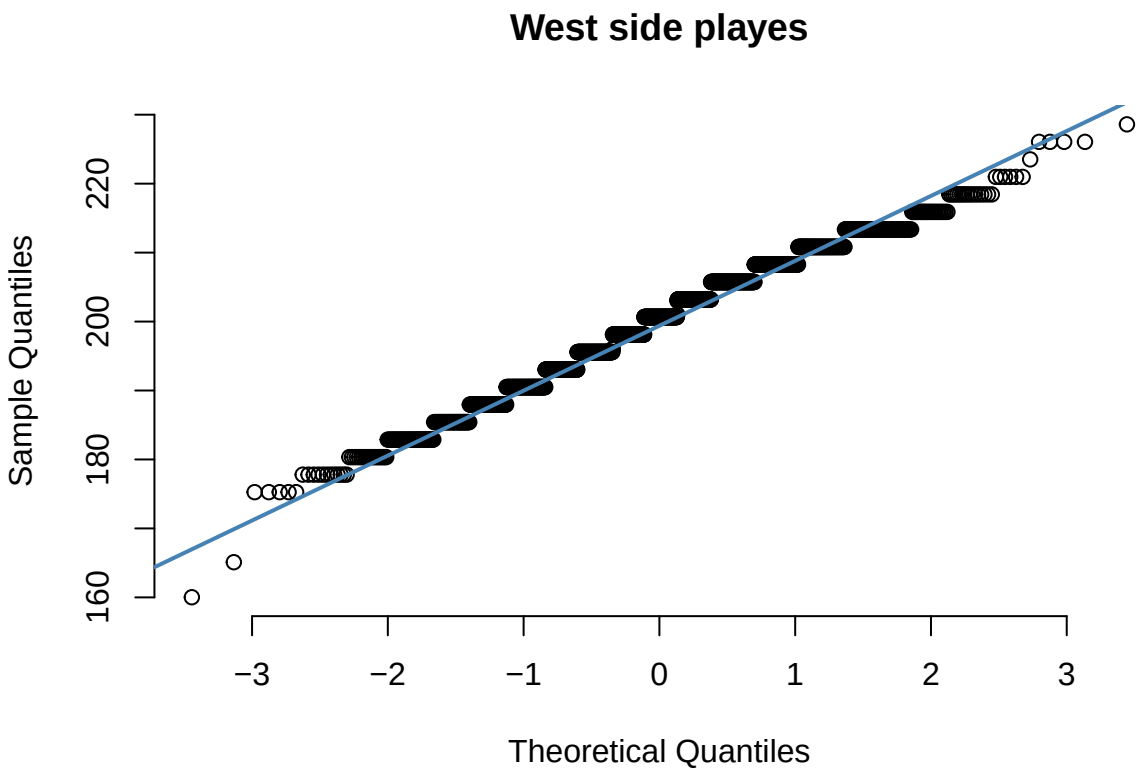
Uspoređivanje srednjih vrijednosti visina igrača istoka i zapada daje indikaciju da nema značajne razlike u visinama igrača.

Najprije provjeravamo pretpostavke nezavisnosti i normalnosti uzorke.

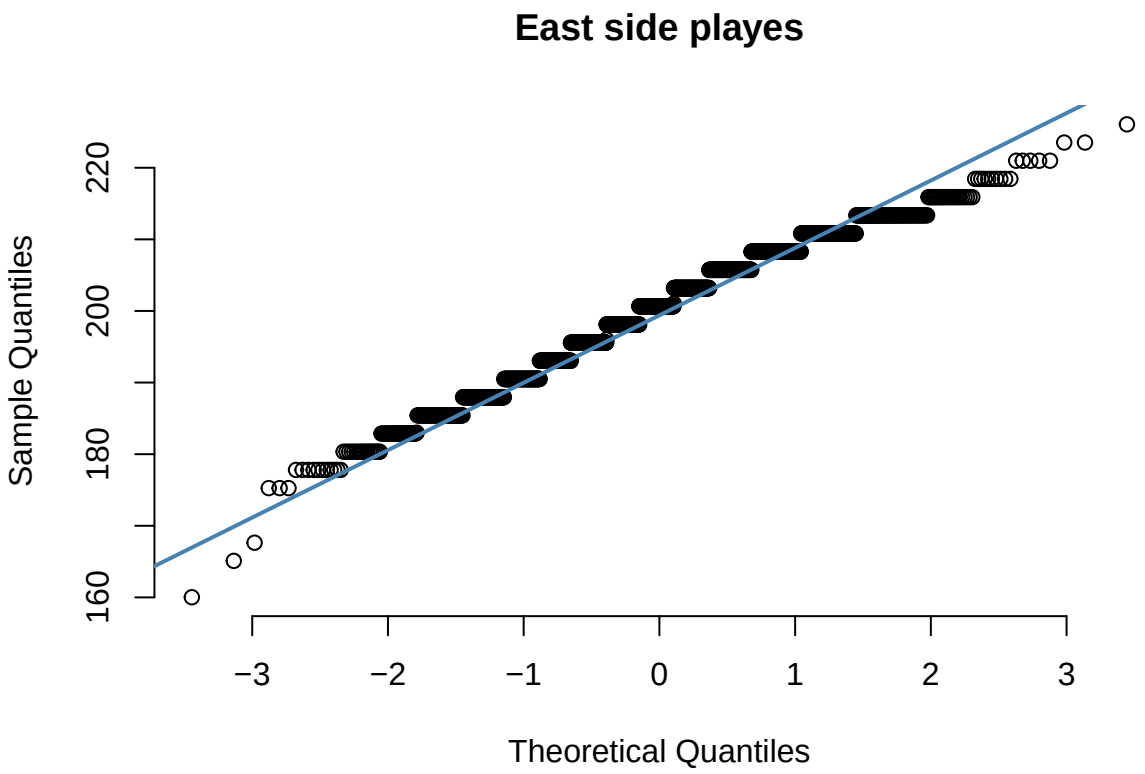
Nezavisnost možemo pretpostaviti obzirom na to da se radi o dva uzorka iz različitih dijelova zemlje.

Za provjeru normalnosti koristimo Q-Q plot-ove.

```
qqnorm(all.seasons.west$player_height, pch = 1, frame = FALSE, main='West side playes')
qqline(all.seasons.west$player_height, col = "steelblue", lwd = 2)
```




```
qqnorm(all.seasons.east$player_height, pch = 1, frame = FALSE, main='East side playes')
qqline(all.seasons.east$player_height, col = "steelblue", lwd = 2)
```



Na Q-Q plotu primjećujemo manja odstupanja od teoretske linije, no smatramo ih zanemarivima te zaključujemo da su podaci dovoljno blizu očekivane normalne distribucije.

Sljedeće uspoređujemo varijance naših uzoraka.

```
var.test(all.seasons.west$player_height, all.seasons.east$player_height)
```

```
##
##  F test to compare two variances
##
## data:  all.seasons.west$player_height and all.seasons.east$player_height
## F = 1.0703, num df = 1740, denom df = 1750, p-value = 0.1562
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.974377 1.175604
## sample estimates:
## ratio of variances
##           1.070265
```

Na temelju p-vrijednosti iz provedenog testa zaključujemo da su varijance jednake te nastavljamo dalje s t-testom uz pretpostavku jednakosti varijanci.

H_0 : nema razlike u srednjim visinama igrača Istočne i Zapadne konferencije.

H_1 : postoji razlika u srednjim visinama igrača Istočne i Zapadne konferencije.

```
t.test(all.seasons.west$player_height, all.seasons.east$player_height, var.equal = TRUE)
```

```
##
```

```
##  Two Sample t-test
##
## data:  all.seasons.west$player_height and all.seasons.east$player_height
## t = -0.56292, df = 3490, p-value = 0.5735
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7705822  0.4268005
## sample estimates:
## mean of x mean of y
##  200.0159  200.1878
```

#zaključak

Nakon izvođenja t-testa, možemo zaključiti da nema statistički značajne razlike u visini između igrača koji predstavljaju timove zapada i onih koji predstavljaju timove istoka. Drugim riječima, nemamo dovoljno dokaza da odbacimo nultu hipotezu u korist alternativne hipoteze na razini značajnosti od 5%.

3.Možemo li predvidjeti prosječni broj poena igrača u sezoni s obzirom na njegove biometrijske podatke?

```
library(dplyr)
```

#Jednostavna regresija

Za procjenu utjecaja biometrijskih podataka na broj poena treba napraviti tri modela jednostavne regresije, po jedan za svaki biometrijski podatak: visina, težina i dob.

```
data_perheight <- nba %>%
  group_by(player_height) %>%
  summarize(avg_pts = mean(pts, na.rm = TRUE))

data_perweight <- nba %>%
  group_by(player_weight) %>%
  summarize(avg_pts = mean(pts, na.rm = TRUE))

data_perage <- nba %>%
  group_by(age) %>%
  summarize(avg_pts = mean(pts, na.rm = TRUE))

fit.height = lm(avg_pts~player_height,data=data_perheight)

fit.weight = lm(avg_pts~player_weight,data=data_perweight)

fit.age = lm(avg_pts~age,data=data_perage)
```

Ako pogledamo sve mjere nad modelima možemo zaključiti koji biometrijski podatci su značajni, a koji nisu:

```
summary(fit.height)
```

```
##
## Call:
## lm(formula = avg_pts ~ player_height, data = data_perheight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0572 -1.5911  0.7371  1.6949  4.4089
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    3.94271    5.28114   0.747    0.462
## player_height  0.01534    0.02651   0.579    0.567
##
## Residual standard error: 2.762 on 28 degrees of freedom
## Multiple R-squared:  0.01182,    Adjusted R-squared:  -0.02347
## F-statistic: 0.335 on 1 and 28 DF,  p-value: 0.5674
```

```
summary(fit.weight)
```

```
##
## Call:
## lm(formula = avg_pts ~ player_weight, data = data_perweight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1622 -1.6845 -0.0269  1.6269 18.8048
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.45778    1.39498   4.629 7.71e-06 ***
## player_weight  0.01472    0.01306   1.127    0.261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.638 on 155 degrees of freedom
## Multiple R-squared:  0.00813,    Adjusted R-squared:  0.001731
## F-statistic: 1.271 on 1 and 155 DF,  p-value: 0.2614
```

```
summary(fit.age)
```

```
##
## Call:
## lm(formula = avg_pts ~ age, data = data_perage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2331 -0.9082  0.0634  1.3108  2.0252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.02113    1.27642   8.634 5.69e-09 ***
## age         -0.13082    0.03993  -3.276  0.00308 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.616 on 25 degrees of freedom
## Multiple R-squared:  0.3003, Adjusted R-squared:  0.2723
## F-statistic: 10.73 on 1 and 25 DF,  p-value: 0.003084
```

Vidi se da je statistički značajan samo podatak dobi jer on jedini ima malu p-vrijednost. Također, koeficijent determinacije objašnjava 30.03% varijance podataka.

Statističku značajnost dobi potvrđuje i korelacijski test:

```
cor.test(data_perage$age,data_perage$avg_pts)
```

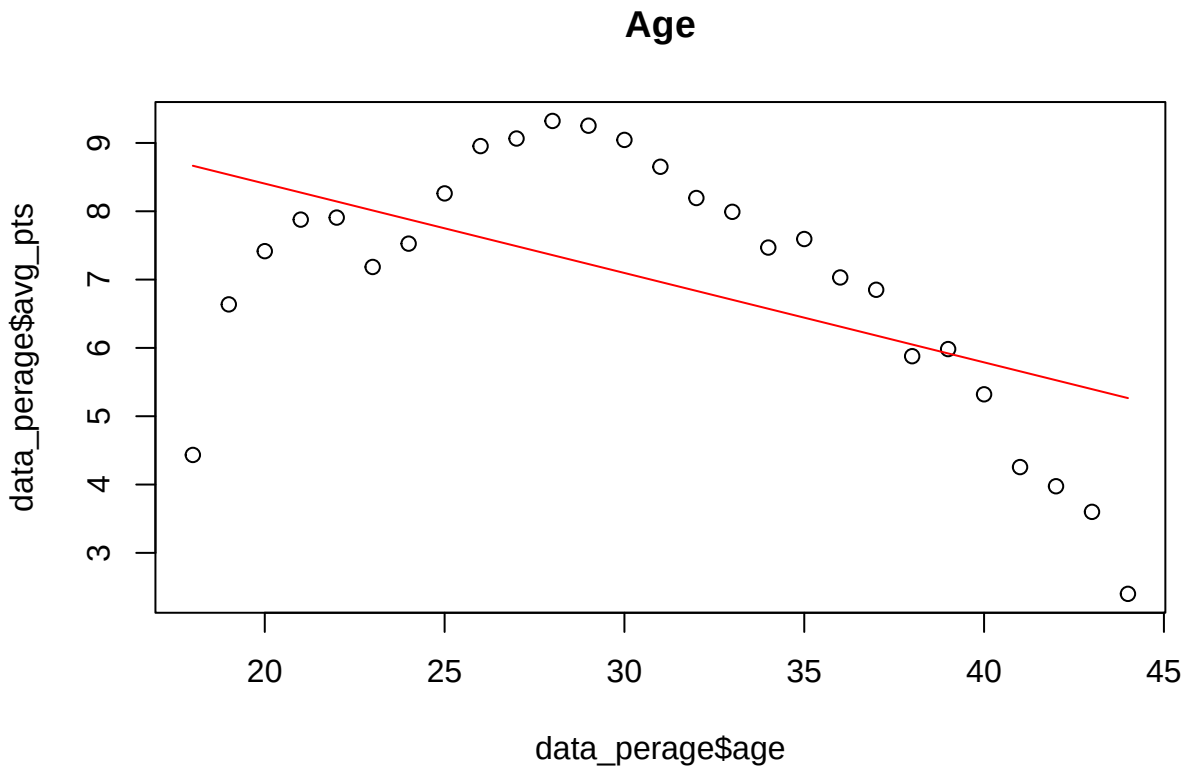
```
##
```

```
##  Pearson's product-moment correlation
##
## data:  data_perage$age and data_perage$avg_pts
## t = -3.2758, df = 25, p-value = 0.003084
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7680799 -0.2122023
## sample estimates:
##       cor
## -0.5480237
```

P-vrijednost 0.003084 je jako mala što dovodi do odbacivanja nulte hipoteze, tj. postoji značajna nenulta korelacija između dobi igrača i njegovih postignutih bodova.

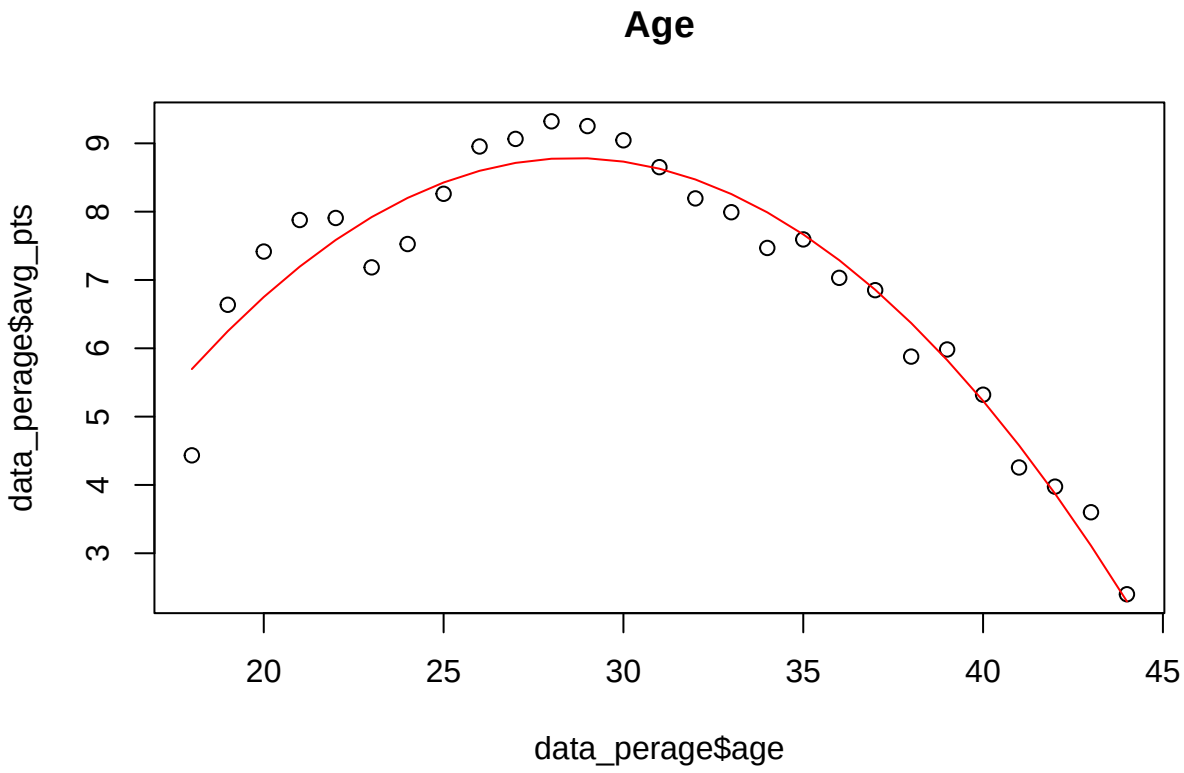
Ako se grafički prikažu podatci tog modela može se naznačiti i nagib pravca linearne regresije koji pokazuje negativan utjecaj varijable dobi na broj postignutih poena:

```
plot(data_perage$age,data_perage$avg_pts, main = "Age")
lines(data_perage$age,fit.age$fitted.values,col='red')
```



Ali, postoji i bolji način prilagodbe modela podatcima koristeći kvadratnu regresiju:

```
fit.age.sq <- lm(avg_pts ~ poly(age, 2, raw = TRUE), data = data_perage)

plot(data_perage$age, data_perage$avg_pts, main = "Age")
lines(data_perage$age, predict(fit.age.sq), col = 'red')
```




```
summary(fit.age.sq)
```

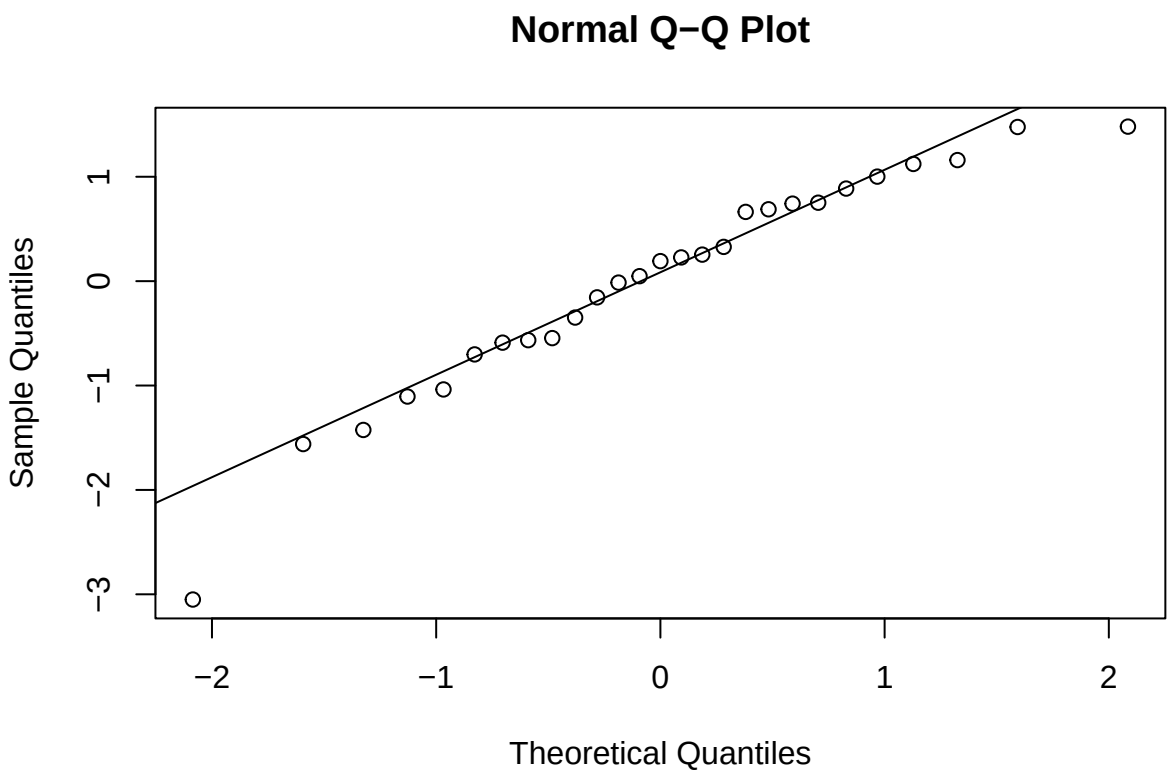
```
##
## Call:
## lm(formula = avg_pts ~ poly(age, 2, raw = TRUE), data = data_perage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26264 -0.27149  0.08976  0.35353  0.68243
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -13.665646   1.617494  -8.449 1.19e-08 ***
## poly(age, 2, raw = TRUE)1    1.569198   0.108819  14.420 2.55e-13 ***
## poly(age, 2, raw = TRUE)2   -0.027420   0.001744 -15.720 3.89e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4908 on 24 degrees of freedom
## Multiple R-squared:  0.9381, Adjusted R-squared:  0.9329
## F-statistic: 181.8 on 2 and 24 DF,  p-value: 3.185e-15
```

Grafički se vidi da model bolje prikazuje podatke, a koristeći summary vidimo i da objašnjava 93.81% varijance. Budući da taj model najbolje objašnjava podatke, taj model će se koristiti i u idućem podzadatku.

Moraju se potvrditi i pretpostavke o regresorima i residualima. U multivarijatnoj regresiji regresori ne smiju biti međusobno jako korelirani, normalnost reziduala i homogenost varijance. Dobra provjera normalnosti

reziduala je i kvantil-kvantil plot:

```
qqnorm(rstandard(fit.age.sq))
qqline(rstandard(fit.age.sq))
```



Kvantil-kvantil plot je relativno ravan što ukazuje na normalnu razdiobu reziduala, ali za potvrdu koristan je i KS test na normalnost:

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

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.age.sq)
## D = 0.11685, p-value = 0.8137
## alternative hypothesis: two-sided
```

Zbog velike p-vrijednosti (p-value=0.8137) ne odbacuje se nul-hipoteza tj. razdioba reziduala je normalna. Vidi se po svim grafovima i testu da reziduali imaju normalnu razdiobu.

#Višestruka regresija

Što ako se koristi više biometrijskih podataka u jednom modelu?

Za višestruku regresiju treba se prvo provjeriti da parovi varijabli nisu previše korelirani.

```
cor(cbind(nba$player_height,nba$player_weight,nba$age))
```

```
##              [,1]       [,2]         [,3]
## [1,]  1.000000000 0.82214119 -0.007903669
## [2,]  0.822141192 1.00000000  0.063560952
```

```
## [3,] -0.007903669 0.06356095  1.000000000
```

Iz tablice se vidi da su visina i težina igrača previše korelirane varijable (cor=0.82214), a visina i dob najmanje korelirane (cor=-0.0079). To znači da se za višestruku regresiju moe koristiti model s visinom i dobi, ali i model s težinom i dobi jer je i njihova korelacija relativno mala (cor=0.06356).

```
fit.heightage = lm(pts ~ player_height + poly(age, 2, raw = TRUE), nba)
summary(fit.heightage)
```

```
##
## Call:
## lm(formula = pts ~ player_height + poly(age, 2, raw = TRUE),
##     data = nba)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.420 -4.525 -1.492  3.199 27.120
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -8.184458   2.227244  -3.675 0.000239 ***
## player_height              -0.032914   0.005791  -5.684 1.35e-08 ***
## poly(age, 2, raw = TRUE)1   1.654738   0.132189  12.518  < 2e-16 ***
## poly(age, 2, raw = TRUE)2 -0.028996   0.002328 -12.456  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.972 on 12840 degrees of freedom
## Multiple R-squared:  0.01508,    Adjusted R-squared:  0.01485
## F-statistic: 65.52 on 3 and 12840 DF,  p-value: < 2.2e-16
```

```
fit.weightage = lm(pts ~ player_weight + poly(age, 2, raw = TRUE), nba)
summary(fit.weightage)
```

```
##
## Call:
## lm(formula = pts ~ player_weight + poly(age, 2, raw = TRUE),
##     data = nba)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.094 -4.534 -1.503  3.227 27.260
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -14.112127   1.888012  -7.475 8.25e-14 ***
## player_weight               -0.012406   0.004253  -2.917  0.00354 **
## poly(age, 2, raw = TRUE)1   1.693453   0.132150  12.815  < 2e-16 ***
## poly(age, 2, raw = TRUE)2  -0.029631   0.002327 -12.732  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.977 on 12840 degrees of freedom
## Multiple R-squared:  0.01325,    Adjusted R-squared:  0.01302
## F-statistic: 57.48 on 3 and 12840 DF,  p-value: < 2.2e-16
```

Ako usporedimo ove modele s modelom koji koristi samo dob možemo zaključiti da makar imaju manju p-vrijednost, objašnjavaju manje varijance podataka pa nisu bolji od modela koji korissti samo dob za predikciju.

#BMI

Što ako se uzme BMI igrača?

```
nba$player_BMI <- nba$player_weight / (nba$player_height ^ 2)

data_perBMI <- nba %>%
  group_by(player_BMI) %>%
  summarize(avg_pts = mean(pts, na.rm = TRUE))

fit.BMI = lm(avg_pts~player_BMI,data=data_perBMI)

summary(fit.BMI)
```

```
##
## Call:
## lm(formula = avg_pts ~ player_BMI, data = data_perBMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3893 -3.1325 -0.6725  2.1610 23.6866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.094      1.499   2.064  0.03928 *
## player_BMI  1783.942    596.464   2.991  0.00286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.595 on 886 degrees of freedom
## Multiple R-squared:  0.009995,   Adjusted R-squared:  0.008878
## F-statistic: 8.945 on 1 and 886 DF,  p-value: 0.002859
```

Mjere modela ukazuju na to da je BMI statistički značajan podatak. Ali ono što se još mora provjeriti je normalnost reziduala.

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

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.BMI)
## D = 0.090628, p-value = 9.247e-07
## alternative hypothesis: two-sided
```

P-vrijednost KS testa je jako mala (9.247e-07) što znači da reziduali nisu normalno razdijeljeni.

#Zaključak

Najbolji regresijski model koji je nađen je kvadratni model linearne regresije broja poena i dobi igrača. On objašnjava 93.81% varijance broja poena što znači da možemo koristiti taj model kako bi predvidjeli broj poena igrača.

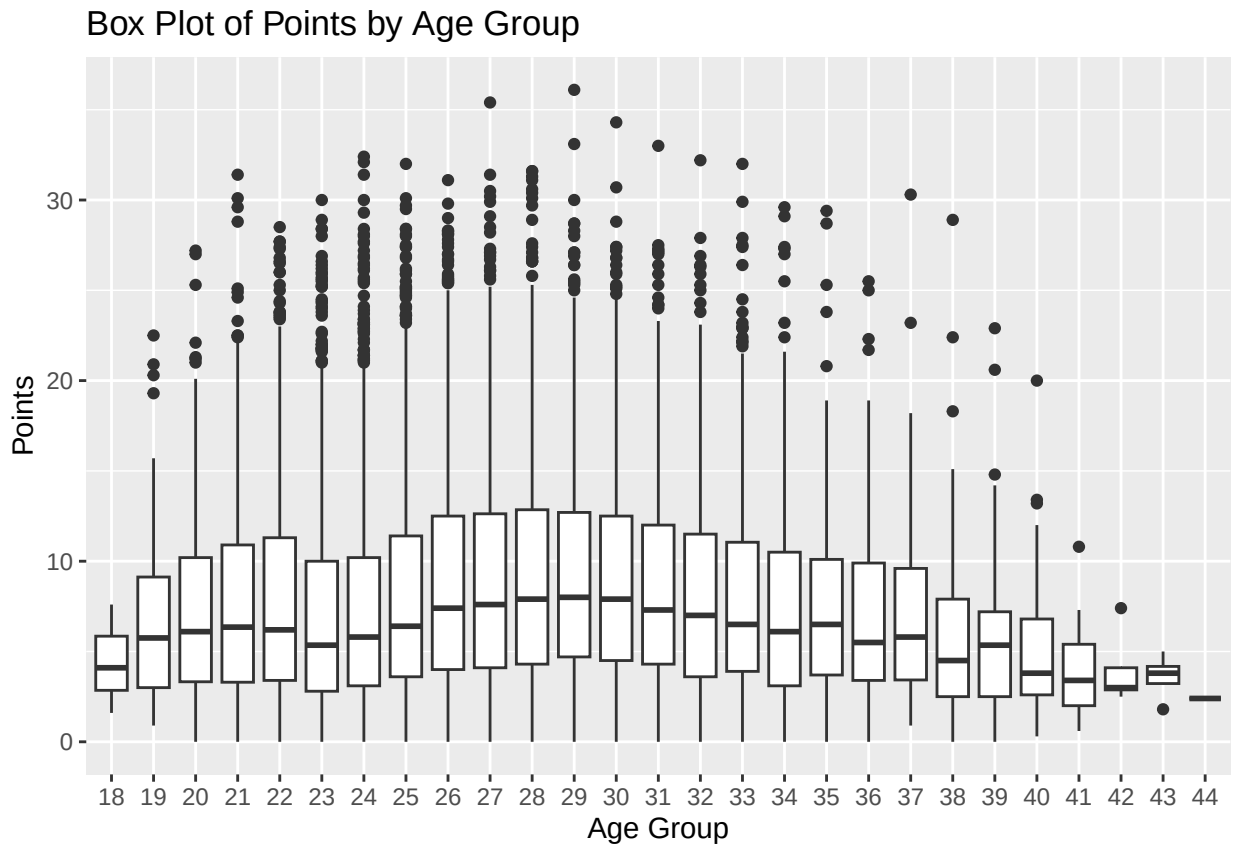
4.Kakva je veza izmedu dobi igrača i prosječnog broja postignutih poena po sezoni?

Jedine varijable koje će nam biti potrebne su “age” i “pts”, odnosno dob i broj poena po utakmici. Nakon što ih izdvojimo, napravimo box plot po godištima.

```
NBA <- nba[,c("age","pts")]
summary(NBA)
```

```
##       age             pts        
##  Min.   :18.00   Min.   : 0.000  
##  1st Qu.:24.00   1st Qu.: 3.600  
##  Median :26.00   Median : 6.700  
##  Mean   :27.05   Mean   : 8.213  
##  3rd Qu.:30.00   3rd Qu.:11.500  
##  Max.   :44.00   Max.   :36.100
```

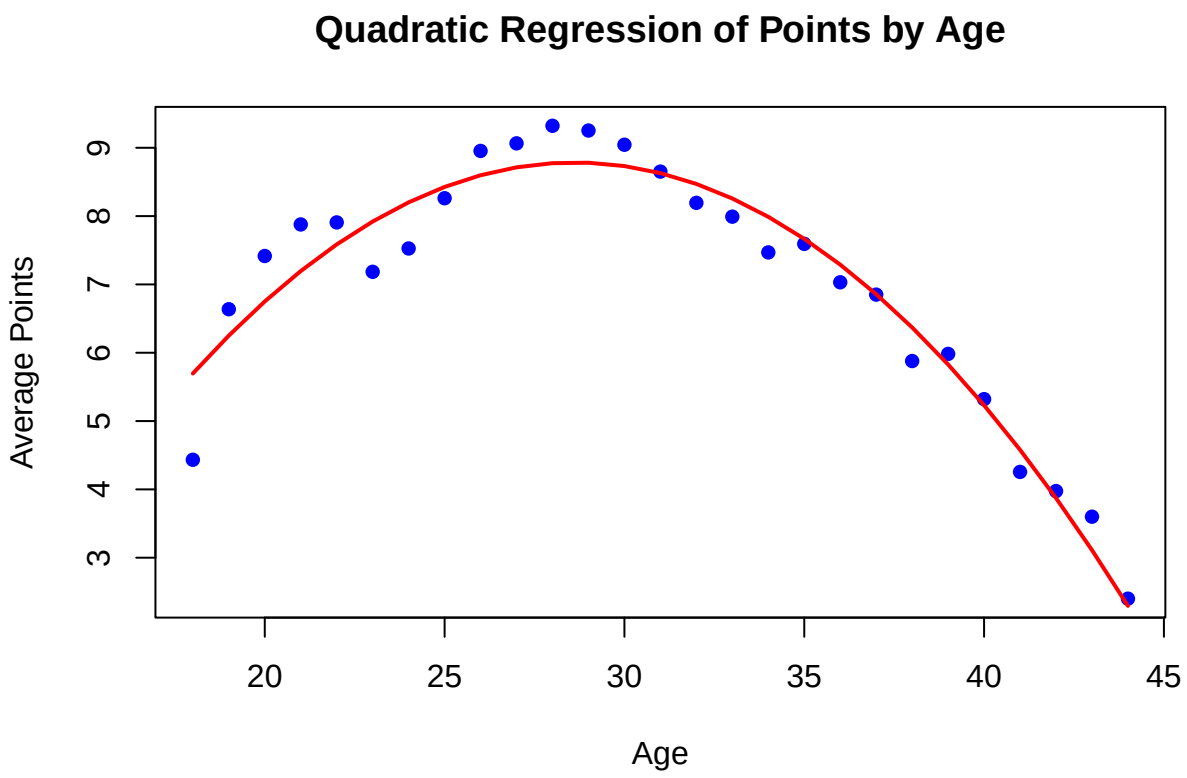
```
library(ggplot2)

# box plot
ggplot(NBA, aes(x = factor(age), y = pts)) +
  geom_boxplot() +
  labs(title = "Box Plot of Points by Age Group",
       x = "Age Group",
       y = "Points")
```




```
NBA$age <- as.factor(NBA$age)
```

Isprobajmo nad prosjekom bodova po godištu regrsiju 2.stupnja, obzirom da obična linearna regresija ne izgleda kao rješenje i obzirom da je u prošlom podzadatku takav zaključak.

```r
avg_ppg_by_age <- aggregate(pts ~ age, data = NBA, FUN = mean)

avg_ppg_by_age$age <- as.numeric(as.character(avg_ppg_by_age$age))

# Fit modela kvadratne regresije
fit.pts.sq <- lm(pts ~ poly(age, 2, raw = TRUE), data = avg_ppg_by_age)

# Plotting
plot(avg_ppg_by_age$age, avg_ppg_by_age$pts, col = 'blue', pch = 16, main = 'Quadratic Regression of Poi
lines(avg_ppg_by_age$age, predict(fit.pts.sq), col = 'red', lw = 2)
```

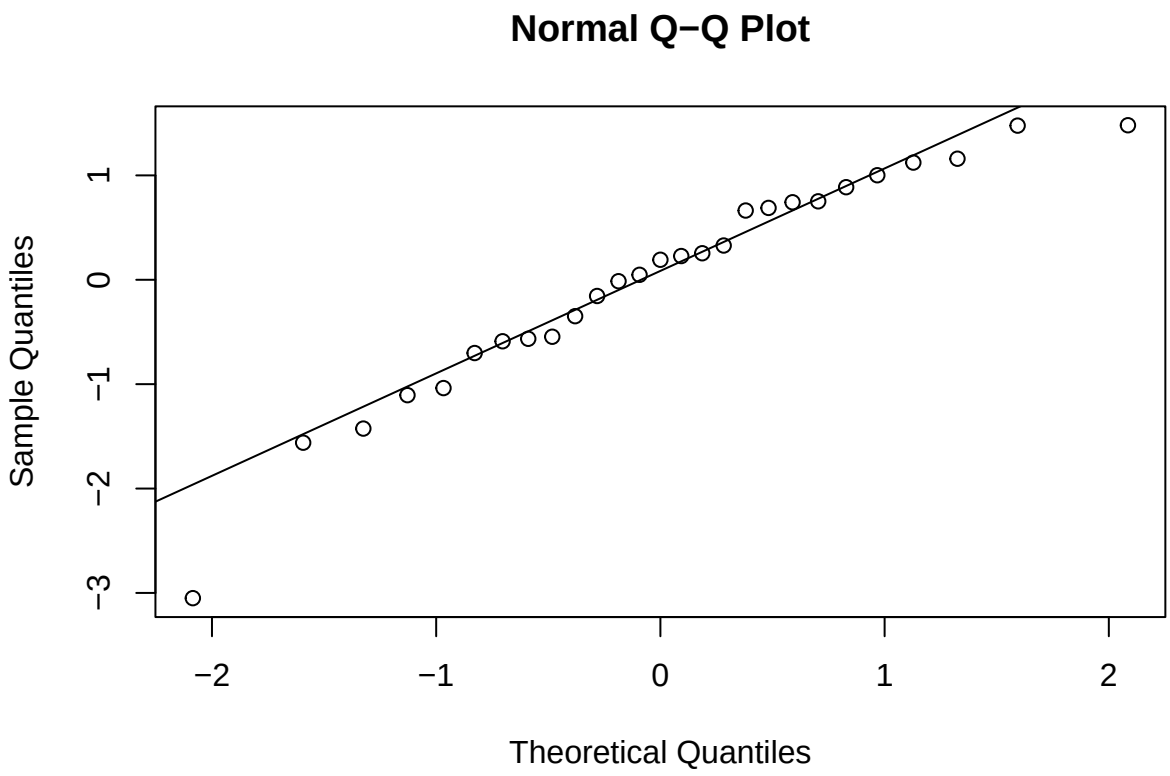



```r
# Summary
summary(fit.pts.sq)
```

```
##
## Call:
## lm(formula = pts ~ poly(age, 2, raw = TRUE), data = avg_ppg_by_age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26264 -0.27149  0.08976  0.35353  0.68243
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -13.665646   1.617494  -8.449 1.19e-08 ***
## poly(age, 2, raw = TRUE)1    1.569198   0.108819  14.420 2.55e-13 ***
```

```
## poly(age, 2, raw = TRUE)2  -0.027420   0.001744 -15.720 3.89e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4908 on 24 degrees of freedom
## Multiple R-squared:  0.9381, Adjusted R-squared:  0.9329
## F-statistic: 181.8 on 2 and 24 DF,  p-value: 3.185e-15
```

```r
# QQ Plot
qqnorm(rstandard(fit.pts.sq))
qqline(rstandard(fit.pts.sq))
```



Vidimo da smo dobili identičan graf kao u prošlom podzadatku. Po velikoj R squared vrijednosti i niskoj p vrijednosti možemo zaključiti da je ovaj model dobar. Deriviramo li dobivene koeficijente dobivamo da u prosjeku bodovi igrača rastu do 28.61 godina te zatim kreću padati. Isprobajmo još jednu metodu ANOVA

Razdvojimo grupe u individualne grupe po godinama i nazovimo ih age18,age19,age20, te provjerimo koliko instanci ima u njima itd.

```r
# Split grupa
age_groups <- split(NBA, NBA$age)

# Napravi individualne grupe
for (age_group_name in names(age_groups)) {
  age_group_data <- age_groups[[age_group_name]]
  assign(paste0("age", age_group_name), age_group_data)
}
```

Obzirom na mali broj određenih godišta, kombinirajmo sve datasetove u 3 velika dataseta(igrači do 24 godina,

igrači između 25 i 30 godina i igrači stariji od 30 godina)

```
young <- rbind(age18, age19, age20, age21, age22, age23, age24)

prime <- rbind(age25, age26, age27, age28, age29, age30)

old <- rbind(age31, age32, age33, age34, age35, age36, age37, age38, age39, age40, age41, age42, age43,
summary(young)
```

```
##       age            pts
##  24     :1333   Min.   : 0.00
##  23     :1218   1st Qu.: 3.10
##  22     : 817   Median : 5.90
##  21     : 534   Mean   : 7.52
##  20     : 318   3rd Qu.:10.30
##  19     :  76   Max.   :32.40
##  (Other):   3
```

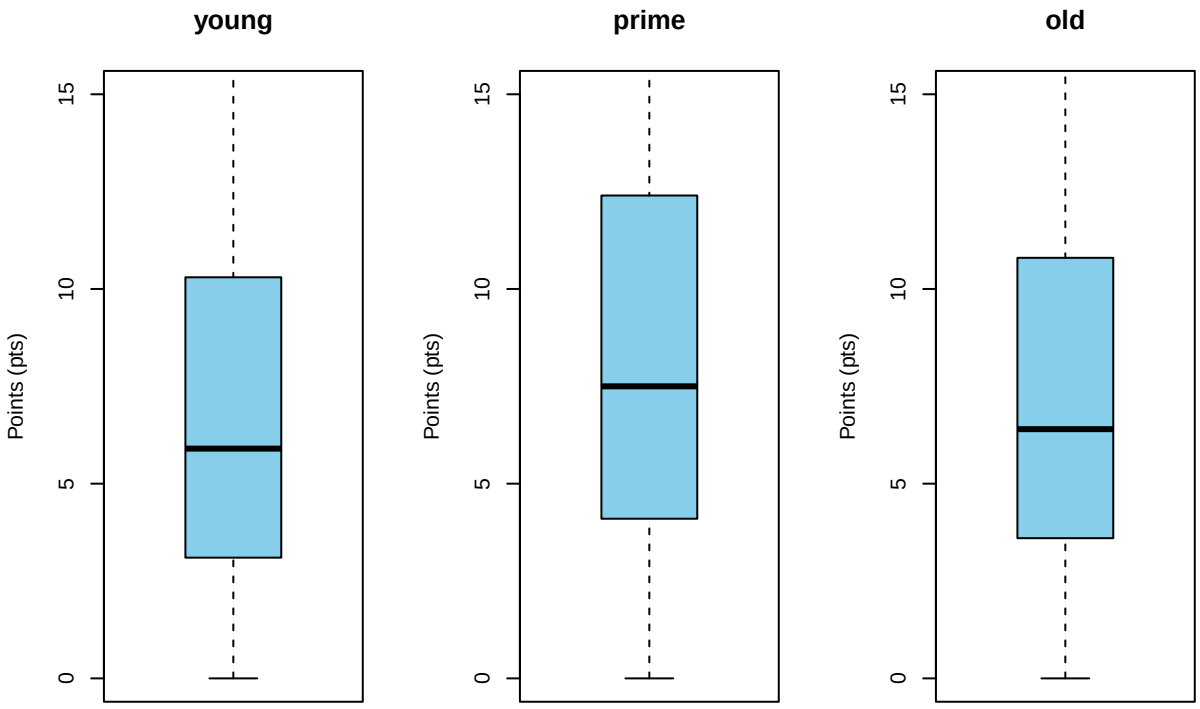
```
summary(prime)
```

```
##       age            pts
##  25     :1195   Min.   : 0.000
##  26     :1058   1st Qu.: 4.100
##  27     :1016   Median : 7.500
##  28     : 891   Mean   : 8.941
##  29     : 813   3rd Qu.:12.400
##  30     : 741   Max.   :36.100
##  (Other):   0
```

```
summary(old)
```

```
##       age           pts
##  31     :643   Min.   : 0.000
##  32     :564   1st Qu.: 3.600
##  33     :467   Median : 6.400
##  34     :367   Mean   : 7.796
##  35     :257   3rd Qu.:10.800
##  36     :205   Max.   :33.000
##  (Other):328
```

```
all_datasets <- list(young = young, prime = prime, old = old)

par(mfrow = c(1, 3))  # Arrange the plots in one row with three columns
for (i in 1:length(all_datasets)) {
  boxplot(all_datasets[[i]]$pts,
          main = names(all_datasets)[i],
          ylab = "Points (pts)",
          col = "skyblue",
          border = "black",
          ylim = c(0, 15))
}
```



Vidimo da igrači srednje dobi zabijaju više poena i od mladih i starijih igrača. Potrebno je još provjeriti jesu li populacije normalno distribuirane i ispitati homogenost varijanci. Za nezavisnost koristimo Lillieforsovu inačicu KS testa.

H_0 : Podaci su normalno distribuirani.

H_1 : Podaci nisu normalno distribuirani.

```
require(nortest)
lillie.test(young$pts)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  young$pts
## D = 0.12095, p-value < 2.2e-16
```

```
lillie.test(prime$pts)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  prime$pts
## D = 0.097253, p-value < 2.2e-16
```

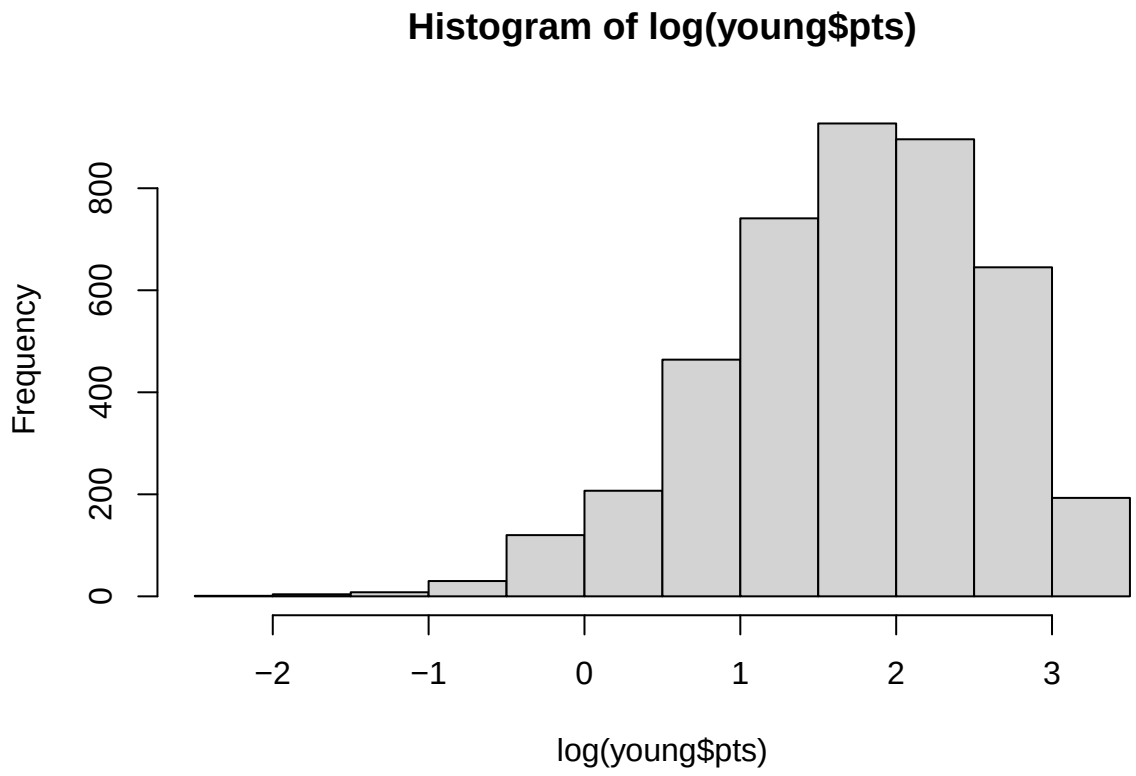
```
lillie.test(old$pts)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  old$pts
```

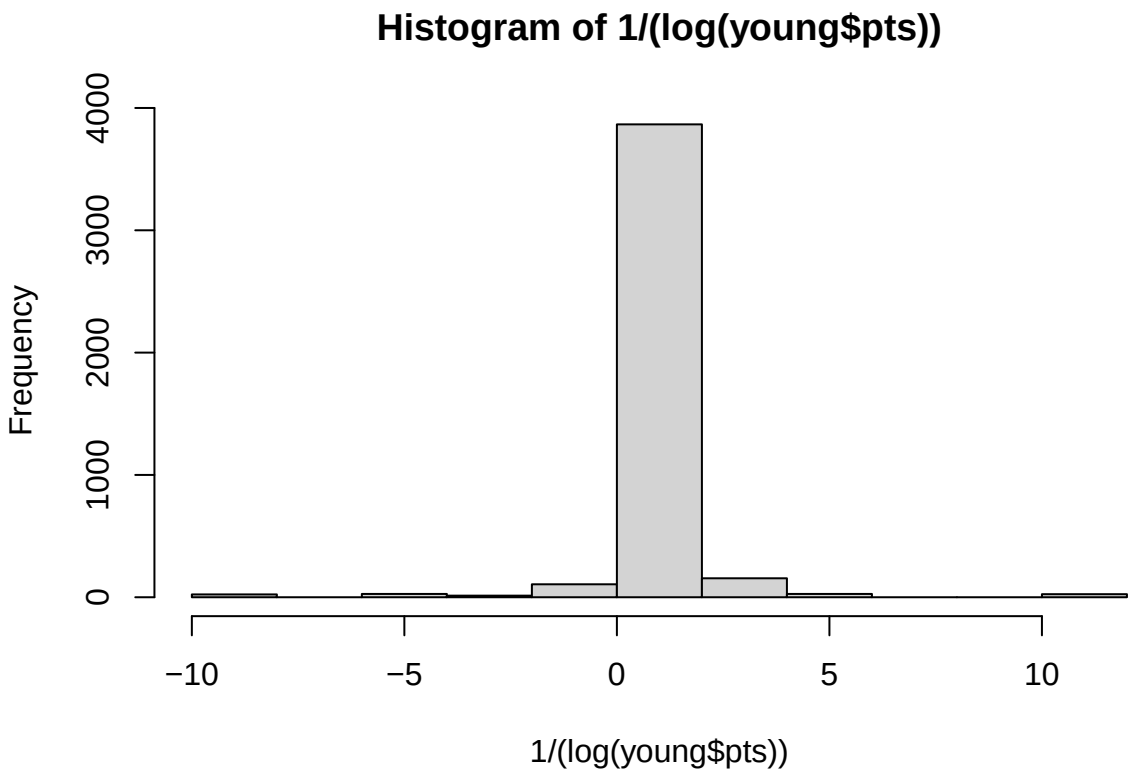
```
## D = 0.10459, p-value < 2.2e-16
```

Vidljivo je da zbog ekstremno niske p vrijednosti distribucije nisu normalne, isprobajmo 2 transformacije podataka kako bismo došli bliže normalnoj distribuciji

```
hist(log(young$pts))
```




```
hist(1/(log(young$pts)))
```



Čak i nakon 2 transformacije vidimo da nismo došli bliže normalnoj distribuciji. Isprobajmu inačicu ANOVE - Kruskal Wallis test koji se koristi u slućaju ne normalnih populacija.

H_0 : Nema značajne razlike u prosjeku poena po utakmici između populacija.

H_1 : Ima značajne razlike u prosjeku poena po utakmici između populacija

```
kruskal_test_result <- kruskal.test(list(young$pts, prime$pts, old$pts))
print(kruskal_test_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(young$pts, prime$pts, old$pts)
## Kruskal-Wallis chi-squared = 168.07, df = 2, p-value < 2.2e-16
```

Zbog niske p vrijednosti zaključujemo da populacije nemaju isti prosjek bodova. Isprobajmo kruskal test na mladim igračima i igračima koji su u svome vrhuncu.

```
kruskal_test_result <- kruskal.test(list(young$pts, prime$pts))
print(kruskal_test_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(young$pts, prime$pts)
## Kruskal-Wallis chi-squared = 156.02, df = 1, p-value < 2.2e-16
```

Vidimo zbog niske p vrijednosti i gledajući box plot i summary da je očito da igrači u dobi od 25 do 30 godina imaju veći prosjek poena od mladih igrača. Isprobajmo usporedbu sa starim igračima

```r
kruskal_test_result <- kruskal.test(list(prime$pts, old$pts))
print(kruskal_test_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(prime$pts, old$pts)
## Kruskal-Wallis chi-squared = 59.096, df = 1, p-value = 1.502e-14
```

#Zaključak

Niska p vrijednost ukazuje ponovno na to da postoji razlika u zabijanju između igrača od 25 do 30 godina i starih igrača. Box plot i summary nam ukazuju da igrači u dobi od 25 do 30 godina zabijaju više od starih igrača

Iz ova 2 testa zaključujemo je da prosjek poena po utakmici igrača raste dok igrač ne uđe u period od 25. do 30. godine, nakon čega opada kako igrač dalje stari.