Analiza podataka o NBA košarkašima

Marko Haralović, Jan Murić, Dominik Agejev, Ante Perković

2023-12-19

Učitavanje skupa podataka u radnu okolinu

```
nba_data <- read_csv("../dataset/all_seasons.csv", show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
head(nba_data)
```

```
## # A tibble: 6 x 22
##    ...1 player_name  team_abbreviation   age player_height player_weight college
##   <dbl> <chr>        <chr>             <dbl>         <dbl>         <dbl> <chr>
## 1     0 Randy Livin~ HOU                 22           193.          94.8 Louisi~
## 2     1 Gaylon Nick~ WAS                 28           190.          86.2 Northw~
## 3     2 George Lynch VAN                 26           203.         103.  North ~
## 4     3 George McCl~ LAL                 30           203.         102.  Florid~
## 5     4 George Zidek DEN                 23           213.         120.  UCLA
## 6     5 Gerald Wilk~ ORL                 33           198.         102.  Tennes~
## # i 15 more variables: country <chr>, draft_year <chr>, draft_round <chr>,
## #   draft_number <chr>, gp <dbl>, pts <dbl>, reb <dbl>, ast <dbl>,
## #   net_rating <dbl>, oreb_pct <dbl>, dreb_pct <dbl>, usg_pct <dbl>,
## #   ts_pct <dbl>, ast_pct <dbl>, season <chr>
```

Zadaci

Skup sadrži podatke igrača NBA (National Basketball Association) od sezone 1996./1997. do sezone 2022./2023. Neke od varijabli sadrže dob igrača, visinu, težinu, broj zabijenih koševa po sezoni, broj asistencija po sezoni itd.

Istraživačka pitanja:

- Razlikuje li se broj poena igrača po sezoni kroz različita desetljeća?
- Postoji li značajna statistička razlika u visini igrača koji igraju za ekipe zapadne od igrača koji igraju za ekipe istočne konferencije?
- Možemo li predvidjeti prosječni broj poena igrača u sezoni s obzirom na njegove biometrijske podatke?
- Kakva je veza između dobi igrača i prosječnog broja postignutih poena po sezoni?

NBA se sastoji od 30 timova koji su podijeljeni u dvije konferencije (Istočna i Zapadna). Svaka konferencija se sastoji od 3 divizije, što ukupno čini skup od 6 divizija. Imena divizija su: Atlantska, Centralna, Jugoistočna, Sjevernozapadna, Pacifička, Jugozapadna.

Primjetiti ćemo kako u skupu podataka postoji više od 30 jedinstvenih imena timova; naime, neke su franšize, odnosno timovi, mijenjali lokacije (grad/državu) ili naziv tima, što je rezultiralo većim brojem jedinstvenih imena timova.

Osnovni pregled skupa podataka

```
glimpse(nba_data)
```

```
## Rows: 12,844
## Columns: 22
## $ ...1              <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ player_name       <chr> "Randy Livingston", "Gaylon Nickerson", "George Lync~
## $ team_abbreviation <chr> "HOU", "WAS", "VAN", "LAL", "DEN", "ORL", "WAS", "CH~
## $ age               <dbl> 22, 28, 26, 30, 23, 33, 26, 30, 24, 24, 22, 31, 29, ~
## $ player_height     <dbl> 193.04, 190.50, 203.20, 203.20, 213.36, 198.12, 231.~
## $ player_weight     <dbl> 94.80073, 86.18248, 103.41898, 102.05820, 119.74829,~
## $ college           <chr> "Louisiana State", "Northwestern Oklahoma", "North C~
## $ country           <chr> "USA", "USA", "USA", "USA", "USA", "USA", "USA", "US~
## $ draft_year        <chr> "1996", "1994", "1993", "1989", "1995", "1985", "199~
## $ draft_round       <chr> "2", "2", "1", "1", "1", "2", "2", "1", "1", "1", "1~
## $ draft_number      <chr> "42", "34", "12", "7", "22", "47", "30", "4", "1", "~
## $ gp                <dbl> 64, 4, 41, 64, 52, 80, 73, 79, 80, 80, 82, 65, 65, 4~
## $ pts               <dbl> 3.9, 3.8, 8.3, 10.2, 2.8, 10.6, 10.6, 26.8, 21.1, 21~
## $ reb               <dbl> 1.5, 1.3, 6.4, 2.8, 1.7, 2.2, 6.6, 4.0, 6.3, 9.0, 5.~
## $ ast               <dbl> 2.4, 0.3, 1.9, 1.7, 0.3, 2.2, 0.4, 2.0, 3.1, 7.3, 1.~
## $ net_rating        <dbl> 0.3, 8.9, -8.2, -2.7, -14.1, -5.8, 6.9, 3.2, -2.9, 6~
## $ oreb_pct          <dbl> 0.042, 0.030, 0.106, 0.027, 0.102, 0.031, 0.098, 0.0~
## $ dreb_pct          <dbl> 0.071, 0.111, 0.185, 0.111, 0.169, 0.064, 0.217, 0.0~
## $ usg_pct           <dbl> 0.169, 0.174, 0.175, 0.206, 0.195, 0.203, 0.185, 0.2~
## $ ts_pct            <dbl> 0.487, 0.497, 0.512, 0.527, 0.500, 0.503, 0.618, 0.6~
## $ ast_pct           <dbl> 0.248, 0.043, 0.125, 0.125, 0.064, 0.143, 0.024, 0.0~
## $ season            <chr> "1996-97", "1996-97", "1996-97", "1996-97", "1996-97~
```

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

Kako modelirati zadatak: - Podjela skupa podataka po desetljećima: Prvi korak u modeliranju je podijeliti skup podataka na manje dijelove, pri čemu se fokusiramo na određena desetljeća.

- Sumiranje i prosječivanje broja poena po igraču za svako desetljeće:
  - Jedna od načina za pristup ovom zadatku je sumiranje i uprosječivanje broja poena po igraču. Ova metoda uključuje uzimanje prosjeka poena po igraču za svaku utakmicu po sezoni. Metodologija:
    - Umjesto da uklanjamo igrače s malim brojem utakmica, zadržavamo ih u analizi. Budući da sezona ima 82 utakmice, potrebno je odrediti neki prag koji će se smatrati prihvatljivim za analizu.
    - Za svaku sezonu unutar desetljeća, za svakog igrača, množimo broj postignutih poena po utakmici s brojem odigranih utakmica.
    - Zatim sumiramo sve postignute poene svih igrača u toj sezoni i postupak ponavljamo za svaku sezonu unutar desetljeća.
    - Na kraju, ukupan broj poena dijelimo s ukupnim brojem utakmica i brojem igrača kako bismo dobili prosječan broj poena po utakmici za to desetljeće.

Razmatranje uključivanja igrača s malim brojem utakmica: Postavlja se pitanje ima li smisla u statističkoj analizi uključiti igrače koji su odigrali samo mali broj utakmica, s obzirom na to da bi mogli biti potencijalni outlieri. No kako ovdje govorimo o desetljeću te velikom broju utakmica i igrača, smatramo kako nema potrebe za isključivanjem outliera, kako se fiksnim granicama koje bismo postavili postavlja i pitanje greške s naše strane upravo u tom postavljanju granice.

```
# u novi podatkovni okvir spremamo podatke iz origalnog okvira + stupac desetljeće
nba_data_decade <- nba_data<- nba_data %>%
  mutate(Decade = cut(as.numeric(substr(season, 1, 4)),
```

```
                      breaks = seq(1990, 2030, by = 10),
                      labels = c("1990s", "2000s", "2010s", "2020s"),
                      right = FALSE))
head(nba_data_decade)
```

```
## # A tibble: 6 x 23
##    ...1 player_name  team_abbreviation   age player_height player_weight college
##   <dbl> <chr>        <chr>             <dbl>         <dbl>         <dbl> <chr>
## 1     0 Randy Livin~ HOU                 22           193.          94.8 Louisi~
## 2     1 Gaylon Nick~ WAS                 28           190.          86.2 Northw~
## 3     2 George Lynch VAN                 26           203.         103.  North ~
## 4     3 George McCl~ LAL                 30           203.         102.  Florid~
## 5     4 George Zidek DEN                 23           213.         120.  UCLA
## 6     5 Gerald Wilk~ ORL                 33           198.         102.  Tennes~
## # i 16 more variables: country <chr>, draft_year <chr>, draft_round <chr>,
## #   draft_number <chr>, gp <dbl>, pts <dbl>, reb <dbl>, ast <dbl>,
## #   net_rating <dbl>, oreb_pct <dbl>, dreb_pct <dbl>, usg_pct <dbl>,
## #   ts_pct <dbl>, ast_pct <dbl>, season <chr>, Decade <fct>
```

Računanje prosječnog broja poena po utakmici na svim podacima

- koristan podatak kako bismo bolje razumjeli značajnosti promejna broja poena po utakmica za svako desetljeće.
- izračunati ćemo i varijancu, standardnu devijaciju , međukvartalni raspon IQR te medijan podataka, kako bismo dobili bolji osječaj o raspršenosti podataka.

```
# Izračun ukupnog prosjeka poena po igri
ukupno_bodova <- sum(nba_data$pts * nba_data$gp, na.rm = TRUE)
ukupno_utakmica <- sum(nba_data$gp, na.rm = TRUE)
overall_average_points_per_game <- ukupno_bodova / ukupno_utakmica

# Ispis prosjeka poena po igri
print(paste("Prosjek poena po igri iznosi:", overall_average_points_per_game))

# Izračunavanje varijance poena po igri
bodovi_po_igri_po_igracu <- nba_data$pts * nba_data$gp / nba_data$gp
varijanca_bodova_po_igri <- var(bodovi_po_igri_po_igracu, na.rm = TRUE)

# Ispis varijance poena po igri
print(paste("Vrijednost varijance poena po igri iznosi:", varijanca_bodova_po_igri))

# Standardna devijacija poena po igri
std_dev_bodova_po_igri <- sd(nba_data$pts, na.rm = TRUE)
print(paste("Standardna devijacija poena po igri iznosi:", std_dev_bodova_po_igri))

# Izračun međukvartilnog raspona (IQR) za poene po igri
IQR_bodova_po_igri <- IQR(nba_data$pts, na.rm = TRUE)
print(paste("Međukvartilni raspon (IQR) poena po igri iznosi:", IQR_bodova_po_igri))

# Izračun mediane apsolutne devijacije (MAD) za poene po igri
MAD_bodova_po_igri <- mad(nba_data$pts, na.rm = TRUE)
print(paste("Medijana apsolutne devijacije (MAD) poena po igri iznosi:",
            MAD_bodova_po_igri))
```

```
## [1] "Prosjek poena po igri iznosi: 9.7938822630528"
```

```
## [1] "Vrijednost varijance poena po igri iznosi: 36.1991564890728"
## [1] "Standardna devijacija poena po igri iznosi: 6.01657348405825"
## [1] "Međukvartilni raspon (IQR) poena po igri iznosi: 7.9"
## [1] "Medijana apsolutne devijacije (MAD) poena po igri iznosi: 5.33736"
```

```
#kreiranje globalnih varijabli preko kojih ćemo računati tražene podatke za zadatak
total_points_decade <- numeric(length(unique(nba_data_decade$Decade)))
print(total_points_decade)

names(total_points_decade) <- unique(nba_data_decade$Decade)
print(names(total_points_decade))

total_games_decade <- numeric(length(unique(nba_data_decade$Decade)))
print(total_games_decade)

names(total_games_decade) <- unique(nba_data_decade$Decade)
print(names(total_games_decade))
```

```
## [1] 0 0 0 0
## [1] "1990s" "2000s" "2010s" "2020s"
## [1] 0 0 0 0
## [1] "1990s" "2000s" "2010s" "2020s"
```

sumiranje i uprosječivanje poena po igraču za svaku sezonu za desetljeće

```
# stupac pts nosi informacije o poenima po utakmici, zaokružene na dvije decimale:
# što dovodi do problema s preciznošću, tako da ukupan broj poena zaokružujem
# na najveću donju granicu

nba_data_decade <- nba_data_decade %>%
  group_by(player_name, season) %>%
  mutate(TotalPointsSeason = floor(gp * pts)) %>%
  ungroup()

print(head(nba_data_decade$TotalPointsSeason))
```

```
## [1] 249  15 340 652 145 848
```

Za svakog igrača odredimo ukupan broj poena i ukupan broj utakmica u desetljeću

```
points_by_decade_player <- nba_data_decade %>%
  group_by(Decade, player_name) %>%
  summarize(TotalPoints = sum(TotalPointsSeason, na.rm = TRUE),
            TotalGames = sum(gp, na.rm = TRUE), .groups = 'drop')

print(head(points_by_decade_player))
```

```
## # A tibble: 6 x 4
##   Decade player_name    TotalPoints TotalGames
##   <fct>  <chr>                <dbl>      <dbl>
## 1 1990s  A.C. Green           1850        297
## 2 1990s  A.J. Bramlett          8          8
## 3 1990s  Aaron McKie          1659        296
## 4 1990s  Aaron Williams       1278        219
## 5 1990s  Acie Earl             188         47
```

```
## 6 1990s  Adam Keefe            1171        248
```

Izračun ukupnog broja poena po desetljeću

- vektorski zapis s n elemenata, gdje je svaki element vektora suma poena igrača u određenom desetljeću

Izračun ukupnog broja utakmica po desetljeću

- vektorski zapis s n elemenata, gdje je svaki element vektora suma poena igrača u određenom desetljeću

n = 4 (podaci od 1990ih do 2020ih)

```
total_points_decade <- sapply(levels(points_by_decade_player$Decade), function(decade) {
  sum(points_by_decade_player$TotalPoints[points_by_decade_player$Decade == decade],
      na.rm = TRUE)
})

print(total_points_decade)

total_games_decade <- sapply(levels(points_by_decade_player$Decade), function(decade) {
  sum(points_by_decade_player$TotalGames[points_by_decade_player$Decade == decade],
      na.rm = TRUE)
})

print(total_games_decade)
```

```
##   1990s   2000s   2010s   2020s
##  821680 2358920 2453891  795885
##  1990s  2000s  2010s  2020s
##  86964 244989 250084  74987
```

Izračun prosječnoj broja poena igrača po utakmici za svako desetljeće

```
#prosječni broj poena po utakmici
average_points_per_game_decade <- total_points_decade / total_games_decade
print(average_points_per_game_decade)
```

```
##     1990s     2000s     2010s     2020s
##  9.448507  9.628677  9.812267 10.613640
```

```
decade_summary <- data.frame(
  Decade = names(total_points_decade),
  TotalPoints = total_points_decade,
  TotalGames = total_games_decade
) %>%
  mutate(AveragePointsPerGame = TotalPoints / TotalGames)

print(decade_summary)
```

```
##       Decade TotalPoints TotalGames AveragePointsPerGame
## 1990s  1990s      821680      86964             9.448507
## 2000s  2000s     2358920     244989             9.628677
## 2010s  2010s     2453891     250084             9.812267
## 2020s  2020s      795885      74987            10.613640
```

Pitanja:

Postoji li značajno odstupanje u sredinama?

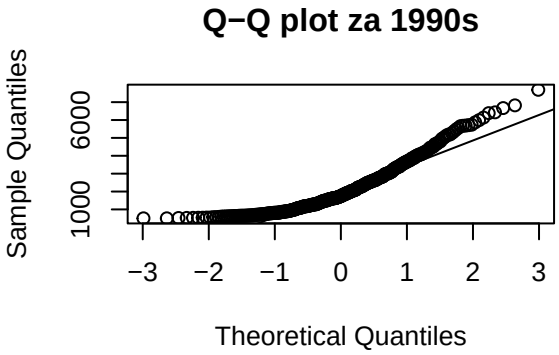
Postoji li neki uzorak koji značajno odstupa od ostalih?

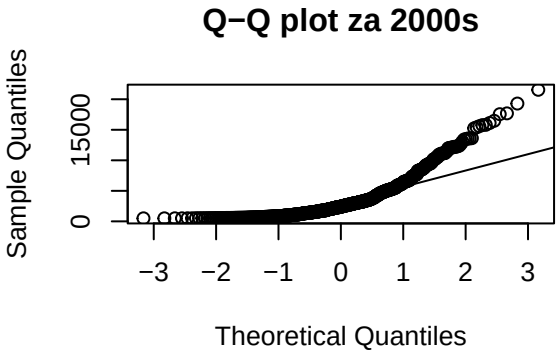
- Odgovore na ova pitanja ćemo pokušati pronaći koristeći ANOVA-u.
- ANOVA se temelji na usporedbi varijance (raspršenosti) između grupa s varijancom unutar grupa. Osnovna premisa ANOVE je da ako postoji značajna razlika između grupa, varijanca između grupa će biti veća od varijance unutar grupa.
- Zavisna varijabla su poeni po desetljeću (kontinuirana varijabla), dok je nezavisna kategorijska varijabla s više od dvije razine varijabla desetljeća.

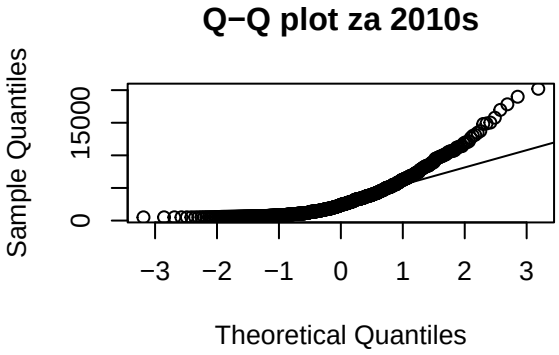
Tri osnovne pretpostavke ANOVA-e su:

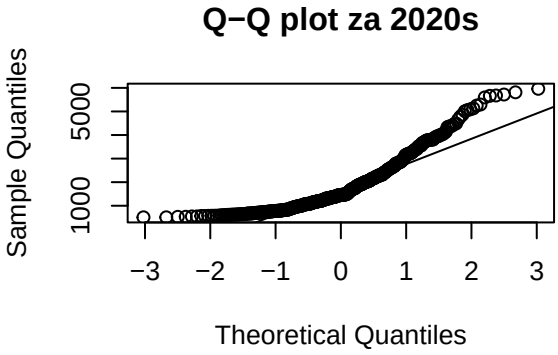
```
- 1. normalnost svake grupe podataka
     (provjera koristeći Q-Q plot i/ili Shapiro-Wilk test normalnosti)
- 2. distribucije imaju identične vraijance
     (kreiranje boxplota i/ili Bartlett Test jednakih varijanci)
- 3. podaci su nezavisni unutar uzorka podataka
     (nema provjere)
```

```r
filtered_data <- points_by_decade_player %>%
  filter(TotalGames >= 50, TotalPoints >= 500)
shapiro_results <- filtered_data %>%
  group_by(Decade) %>%
  summarize(shapiro_p_value = shapiro.test(TotalPoints)$p.value)

print(shapiro_results)

qqPlot <- function(data, group) {
  uniqueGroups <- unique(data[[group]])
  par(mfrow=c(2,2))
  for (grp in uniqueGroups) {
    dataSubset <- data[data[[group]] == grp, ]
    qqnorm(dataSubset$TotalPoints, main = paste("Q-Q plot za", grp))
    qqline(dataSubset$TotalPoints)
  }
}
qqPlot(filtered_data, "Decade")
```










```
## # A tibble: 4 x 2
##   Decade shapiro_p_value
##   <fct>            <dbl>
## 1 1990s         4.49e-15
## 2 2000s         4.85e-28
## 3 2010s         1.41e-27
## 4 2020s         1.53e-17
```

Zaključujemo kako ANOVA test nije prikladan kako podaci nisu noramlno distribuirani, čak i za slučaj da se otklone stršeće vrijednsoti. U ovome smo slučaju postavili da su svi igrači s manje od 50 utakmica ili 500 poena stršeće vrijednosti, kako je to iznimno malena broja u desetljeću u kojemu je 820 ukupno odiranih utakmica regularne sezone.

Iz tih razloga odlučujemo se primjeniti neparametarski test Kruskal-Wallis (neparametarska ANOVA). Uvjet za primjenjivost Kruskal-Wallisovog testa: veličina svakog uzorka je barem 5.

- H0 : medijani distribucija svih uzoraka su jednaki
- H1 :barem dva medijana nisu jednaka

Nivo značajnosti alfa postavljamo na 0.05.

```
kruskal_test_result <- kruskal.test(TotalPoints ~ Decade, data = filtered_data)

# printaj  rezultat  testa  Kruskal-Wallis
print(kruskal_test_result)
```

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

```
## data:  TotalPoints by Decade
## Kruskal-Wallis chi-squared = 83.113, df = 3, p-value < 2.2e-16
```

Rezultati Kruskal-Wallis testa pokazuju vrlo nisku p-vrijednost (manju od 2.2e-16), što ukazuje na to da postoji statistički značajna razlika u ukupnim poenima po igraču između različitih desetljeća.

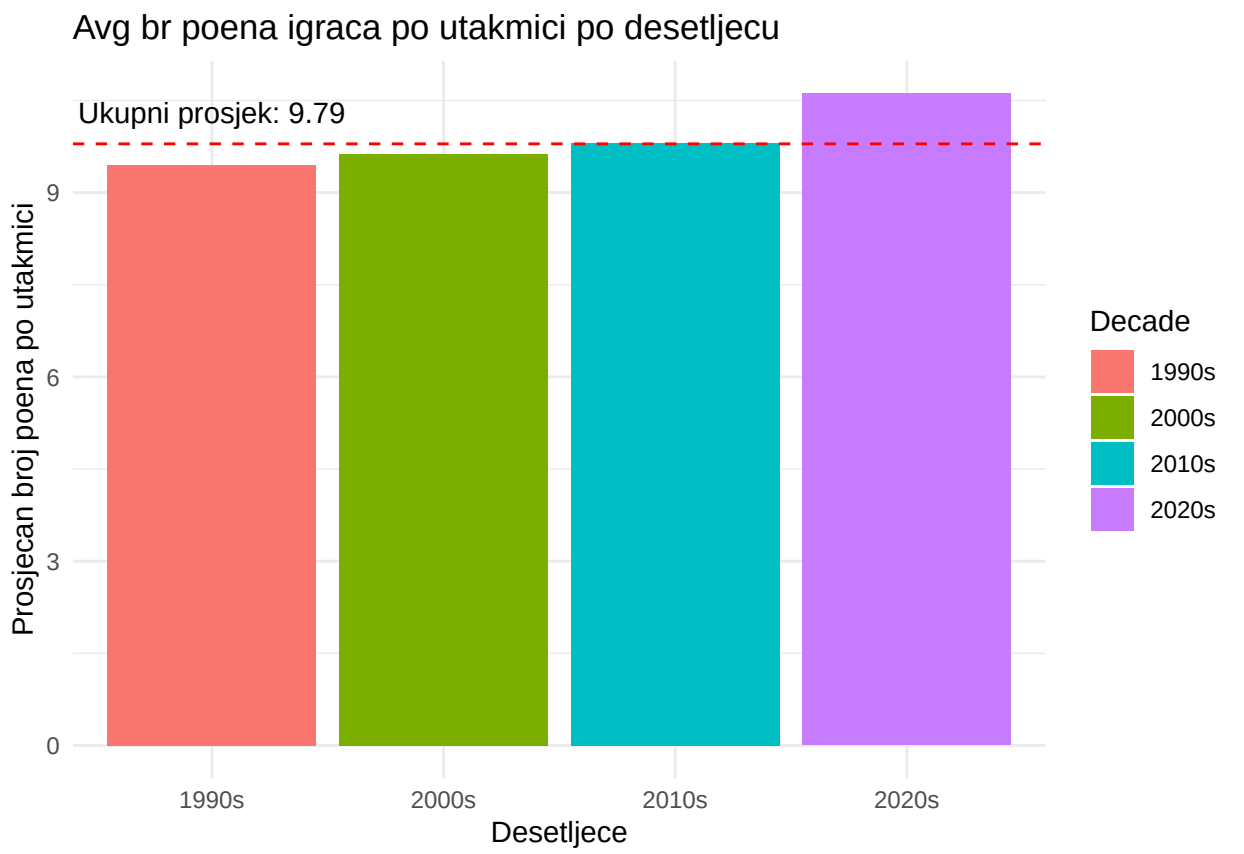
ZAKLJUČAK:

Zaključak je da igrači iz različitih desetljeća imaju statistički značajno različite medijane ukupnih bodova. To može ukazivati na trendove u načinu igranja, pravilima igre, stilovima treninga, ili drugim faktorima koji su se mijenjali tijekom vremena.

Vizualizacija prosjeka poena

Na sljedećem grafu vidimo prosječan broj poena po svakom desetljeću. Crvena isprekidana linija označava prosječan broj poena po svim utakmicama na cijelom skupu podataka.

```
ggplot(decade_summary, aes(x = Decade, y = AveragePointsPerGame, fill = Decade)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept = overall_average_points_per_game, linetype = "dashed",
  color = "red") +
  theme_minimal() +
  labs(title = "Avg br poena igrača po utakmici po desetljeću",
       x = "Desetljeće",
       y = "Prosječan broj poena po utakmici") +
  annotate("text", x = "1990s", y = overall_average_points_per_game,
           label = paste("Ukupni prosjek:",
            round(overall_average_points_per_game, 2)), vjust = -1)
```



ODGOVOR NA PITANJE 2:

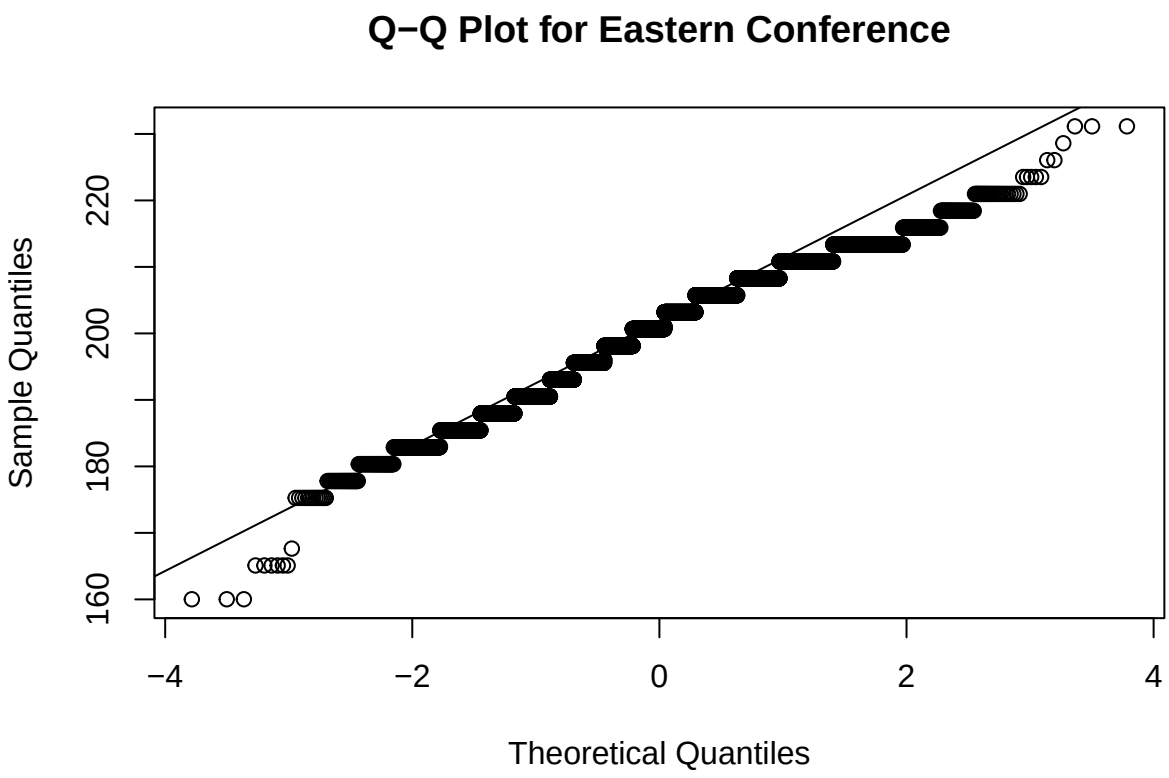
Postoji li značajna statistička razlika u visini igrača koji igraju za ekipe zapadne od igrača koji igraju za ekipe istočne konferencije

```
# istočna i zapadna konferekncija imaju 15 timova, ovdje svakom igraču
nba_data <- nba_data %>%
  mutate(conference = case_when(
    team_abbreviation %in% c("BOS", "BKN", "NYK", "PHI", "TOR",
                             "CHI", "CLE", "DET","IND", "MIL",
                             "ATL", "CHA", "MIA", "ORL", "WAS",
                             "CHH","NJN") ~ "East",
    team_abbreviation %in% c("DEN", "MIN", "OKC", "POR", "UTA",
                             "GSW", "LAC", "LAL", "PHX", "SAC",
                             "DAL", "HOU", "MEM", "NOP", "SAS",
                             "VAN","SEA","NOH","NOK") ~ "West",
    TRUE ~ NA_character_
  ))
```

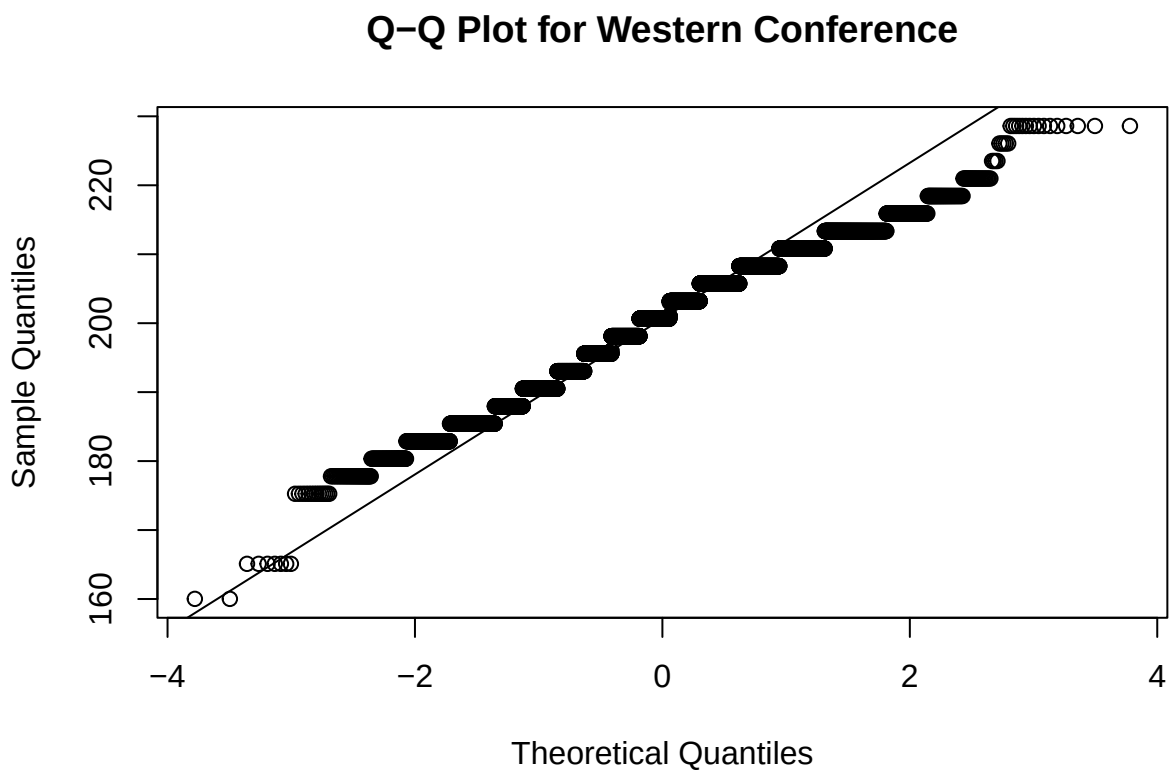
Analiza

a) Testiranje normalnosti Q-Q plotovima

```
# Q-Q plot istočne konferencije
qqnorm(nba_data$player_height[nba_data$conference == "East"],
       main = "Q-Q Plot for Eastern Conference")
qqline(nba_data$player_height[nba_data$conference == "East"])
```



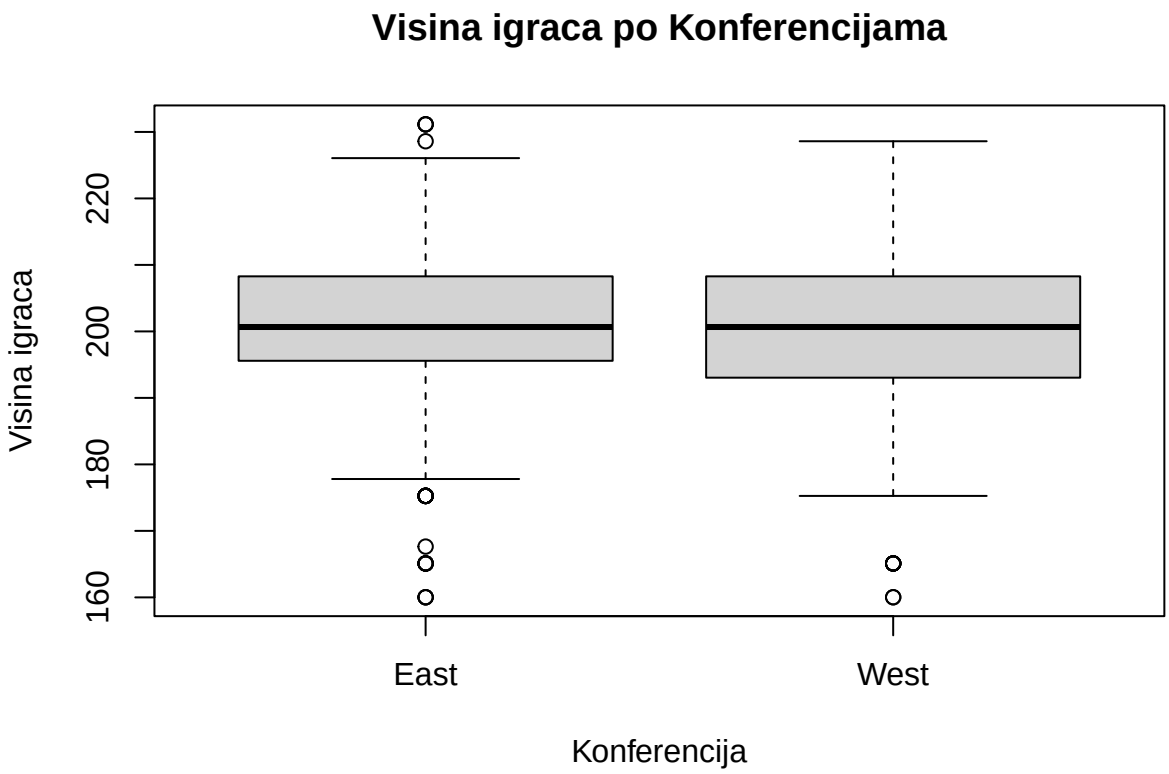
```r
# Q-Q plot zapadne konferencije
qqnorm(nba_data$player_height[nba_data$conference == "West"],
       main = "Q-Q Plot for Western Conference")
qqline(nba_data$player_height[nba_data$conference == "West"])
```



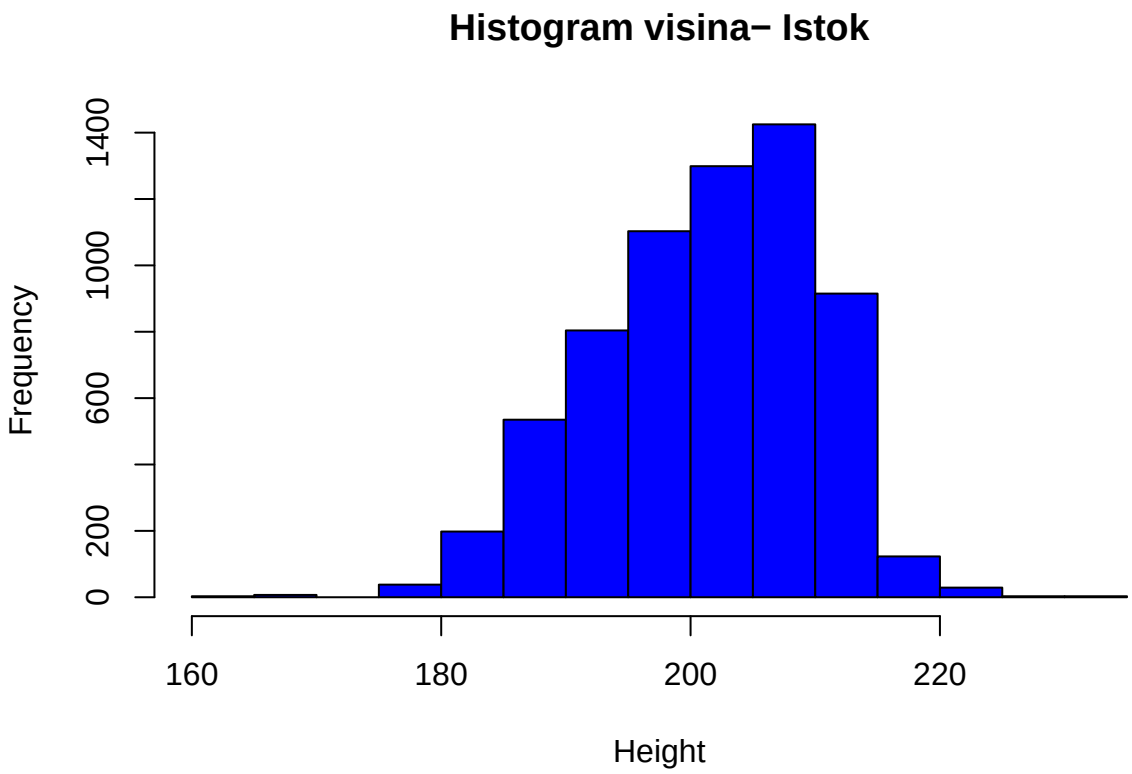
KOMENTAR : Zaključujemo kako Q-Q grafovi obje konferencije prikazuju malenu devijaciju od teoretske linije na donjim i gornjim krajevima, što indicira odstupanje od normalne distribucije. Repovi su teži od normalne distribucije, što znači ili da su podaci dejlomično nakošeni ili da imaju outliera. Trenutan zaključak je kako je ovakva distibucija normalna, odnosno možemo zaključiti kako "dovoljno dobro" odgovara teoretskoj normalnoj distribuciji te ćemo zaključiti kako je ostvarena pretpostavka normalnosti.

b) Crtanje box plota i histograma podataka

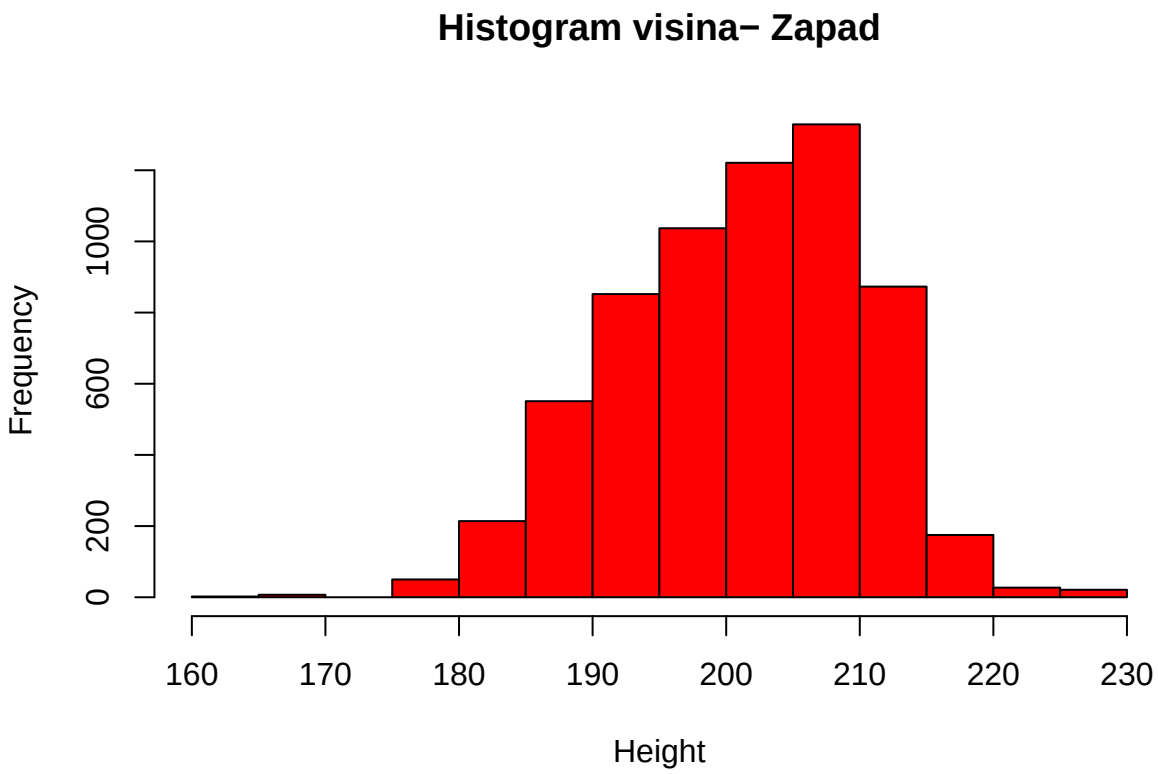
```r
# Box plot
boxplot(player_height ~ conference, data = nba_data,
        main = "Visina igrača po Konferencijama",
        xlab = "Konferencija", ylab = "Visina igrača")
```




```
# Histogram
hist(nba_data$player_height[nba_data$conference == "East"],
     main = "Histogram visina- Istok", xlab = "Height", col = "blue" )
```




```
hist(nba_data$player_height[nba_data$conference == "West"],
     main = "Histogram visina- Zapad", xlab = "Height", col = "red")
```



Na nacrtanim dijagramima vidimo kako postoje stršeće vrijednosti, ali ne u značajnom broju. Medijanu visina obaju konferencija su slični, dok na zapadu je IQR nešto veći u odnosu na istočnu konferenciju.

c) t-test za nezavisne uzorke -> parametarski test T-test je robustan na manja odstupanja od normalnosti. Snažniji je od neparametarskih testova ako njegove pretpostavke nisu povrijeđene. Međutim, prisustvo outliera može utjecati na T-test, čineći ga manje pouzdanim ako su te izvanredne vrijednosti ekstremne. Outliere smo vidjeli na box plotu u b) dijelu, ali Q-Q plotom smo vidjeli minimalno odstupanje od normalne distribucije, tako da smo zaključili kako možemo reći da su podaci normalno distribuirani te ćemo koristiti paramaterski test, odnosno t-test.

```
# izvođenje t-testa da  vidimo postoji li
# značajna razlika u visinama igrača različitih konferencija
t_test_result <- t.test(player_height ~ conference, data = nba_data)

print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  player_height by conference
## t = 0.78218, df = 12794, p-value = 0.4341
## alternative hypothesis: true difference in means between group East and group West is not equal to 0
## 95 percent confidence interval:
##  -0.1894923  0.4411384
## sample estimates:
## mean in group East mean in group West
##           200.6174           200.4916
```

ZAKLJUČAK

Iz grafova smo zaključili kako su podaci normalno distibuirani. Tako, odabiremo t test (parametarski) kako mislimo kako govorimo o dovoljno velikom broju podataka.

H0: nema razlike u srednjim visinama igrača Istočne i Zapadne konferencije na osnovu dostupnih podataka

H1: postoji razlika u srednjim visinama igrača Istočne i Zapadne konferencije na osnovu dostupnih podataka alfa = 0.05

Uzimajući da je alfa = 0.05, a kako p iznosi 0.4341, zaključak je da nema statistički značajne razlike u srednjim vrijednostima visina igrača između dvije konferencije. Drugim riječima, ne možemo odbaciti nultu hipotezu koja kaže da nema razlike u srednjim visinama igrača Istočne i Zapadne konferencije na osnovu dostupnih podataka.

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

Gledati ćemo prosječan broj postignutih poena po utakmici kao i prosječan broj odigranih utakmica igrača po godini, grupiranih po godinama.

Za najboljih 100 strijelaca svake godine izvaditi ćemo statistiku o prosječnoj dobi i broju sezona koje su u prosjeku provedene u NBA-u. Smatramo kako je navedena statistika zanimljiva kako u sezoni prosječno igra oko 500 igrača (30 timova po 18 igrača) pa je ovo značajan udio promatranih igrača koji ipak imaju veći broj psotignutih poena po utakmici te lakše dolazimo do uvida o tome o kojoj se dobi radi kada govorimo o najboljim strijelcima lige. Mogli bismo govoriti o proširenju na 200tinjak igrača, ali kako je broj arbitraran, smatramo da "najboljim" strijelcima možemo proglasiti igrače koji ulaze u ovu statistiku, odnosno sve igrače u top 100 strijelaca sezone (tako uključujemo 2.,3. i 4. napadačku opciju svakog tima).

Nije svaki igrač bio draftom izabran u ligu, tako da smo grupirali podatkovni okvir i izvukli informacije o prvoj godini igranja u ligi za svakog igrača.

```
nba_data <- nba_data %>%
  mutate(season_start_year = as.integer(sub("-.*", "", season)))

# podatkovni okvir s prvog godinom igranja svakog igrača
players_first_year <- nba_data %>%
  group_by(player_name) %>%
  summarise(first_year = min(season_start_year))
#head(players_first_year)

nba_data <- nba_data %>%
  left_join(players_first_year, by = "player_name")

top_scorers_average_age_exp <- nba_data %>%
  mutate(total_points = pts * gp,
         years_in_league = season_start_year - first_year + 1 # izračun broja godina u ligi
         ) %>%
  group_by(season) %>%
  top_n(100, total_points) %>%
  summarise(average_age = mean(age, na.rm = TRUE),
            average_experience = mean(years_in_league, na.rm = TRUE))
# srednja vrijednsost broja godina /iskustvo u ligi

print(top_scorers_average_age_exp)
```

```
## # A tibble: 27 x 3
```

```
##    season  average_age average_experience
##    <chr>         <dbl>              <dbl>
##  1 1996-97        27.9               1
##  2 1997-98        27.9               1.94
##  3 1998-99        28.4               2.8
##  4 1999-00        27.6               3.49
##  5 2000-01        27.3               4.24
##  6 2001-02        27.6               4.77
##  7 2002-03        27.0               5.15
##  8 2003-04        26.6               5.43
##  9 2004-05        27.1               5.91
## 10 2005-06        26.9               6.05
## # i 17 more rows
```

Ovdje je podatkovni okvir bez iskustva, za prvih nekoliko godina nema smisla jer se nisu skupljali raniji podaci

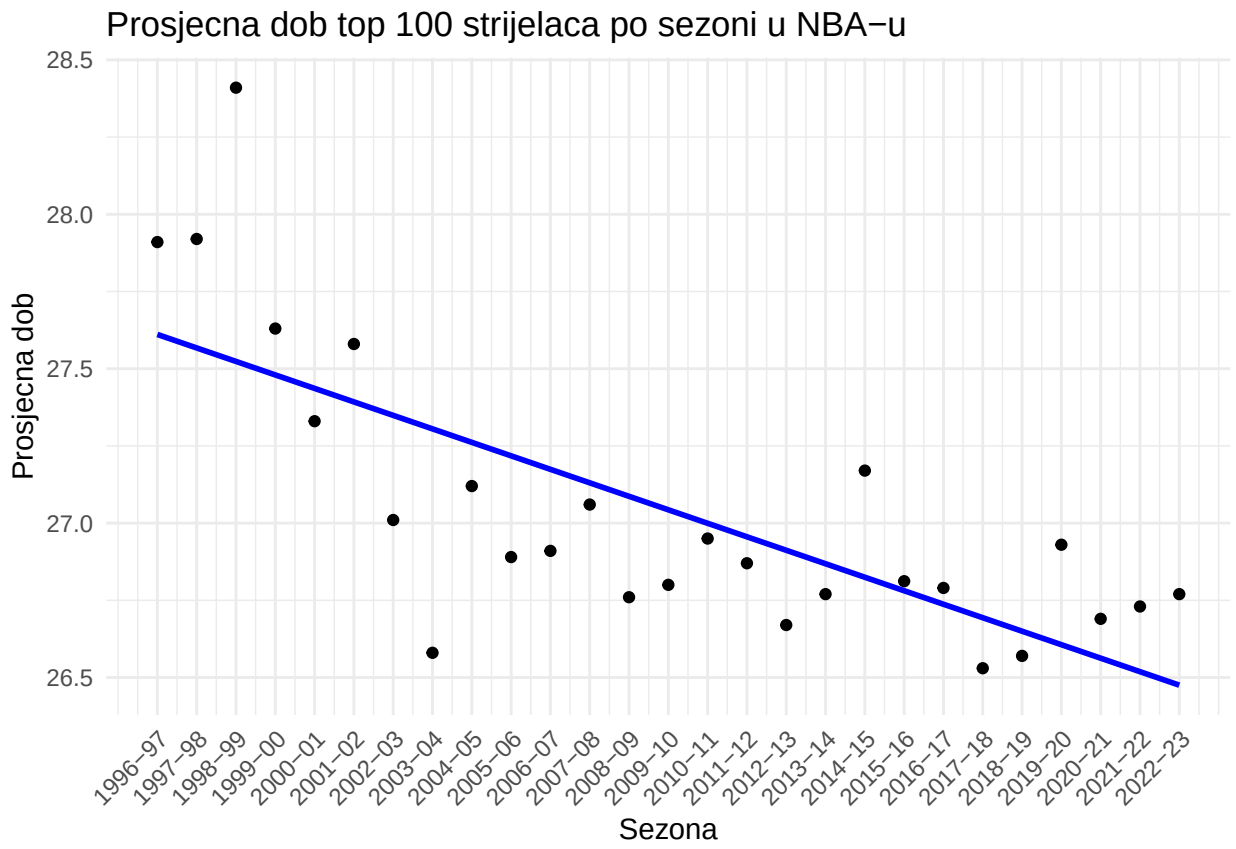
```
top_scorers_average_age <- nba_data %>%
  mutate(total_points = pts * gp) %>%  # izračun ukupnih poena u sezoni
  group_by(season) %>%
  top_n(100, total_points) %>%
  summarise(average_age = mean(age, na.rm = TRUE))

print(top_scorers_average_age)
```

```
## # A tibble: 27 x 2
##    season  average_age
##    <chr>         <dbl>
##  1 1996-97        27.9
##  2 1997-98        27.9
##  3 1998-99        28.4
##  4 1999-00        27.6
##  5 2000-01        27.3
##  6 2001-02        27.6
##  7 2002-03        27.0
##  8 2003-04        26.6
##  9 2004-05        27.1
## 10 2005-06        26.9
## # i 17 more rows
```

Vizualizacija po godinama prosječne dobi najboljih 100 strijelaca

```
# računanj vrijednosti broja godina najboljih 100 strijelaca lige
top_scorers_average_age <- top_scorers_average_age %>%
  mutate(season_start_year = as.numeric(sub("-.*", "", season)))

ggplot(top_scorers_average_age, aes(x = season_start_year, y = average_age)) +
  geom_point() +  #
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  theme_minimal() +
  labs(x = "Sezona", y = "Prosječna dob",
       title = "Prosječna dob top 100 strijelaca po sezoni u NBA-u") +
  scale_x_continuous(breaks = top_scorers_average_age$season_start_year,
                     labels = top_scorers_average_age$season) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Iz grafa se jasno vidi kako dob najboljih strijelaca lige generalno opada po sezonama, odnosno moguće je zaključiti kako postoji trend u podacima za najbolkjih 100 strijelaca, a taj jest kako se prosječna dob najboljih 100 strijelaca kontinuirano smanjuje. Naravno, nije potpuno linearna ovisnost, ali jasno je moguće zaključiti kako se dob najboljih streijalca smanjila kroz godine.

Nastavljamo promatrati sve igrače (ne samo 100 ponajboljih strijalca svake sezone)

- Gledamo prosječan broj postignutih poena po starosti igrača.
- Odrediti ćemo minimalnu i maksimalnu dob cijelog skupa podataka, za svaku godinu dodavati poene i utakmice, izračunati prosječan broj poena igrača ovisno o njegovoj starosti.

```
points_by_age <- nba_data %>%
  group_by(age) %>%
  summarize(TotalPoints = sum(gp * pts, na.rm = TRUE),
            TotalGames = sum(gp, na.rm = TRUE), .groups = 'drop')
print(points_by_age)
```

```
## # A tibble: 27 x 3
##      age TotalPoints TotalGames
##    <dbl>       <dbl>      <dbl>
##  1    18         798.        162
##  2    19       29475.       3553
##  3    20      143367.      15864
##  4    21      260154.      27379
##  5    22      389562.      40309
##  6    23      511265.      55768
```

```
##  7    24     596214.      62668
##  8    25     602889.      59642
##  9    26     584614.      55901
## 10    27     574892.      53864
## # i 17 more rows
```

```
total_points_age <- points_by_age$TotalPoints
total_games_age <- points_by_age$TotalGames

average_points_per_game_age <- total_points_age / total_games_age

age_summary <- data.frame(
  Age = points_by_age$age,
  TotalPoints = total_points_age,
  TotalGames = total_games_age,
  AveragePointsPerGame = average_points_per_game_age
)

print(age_summary)
```

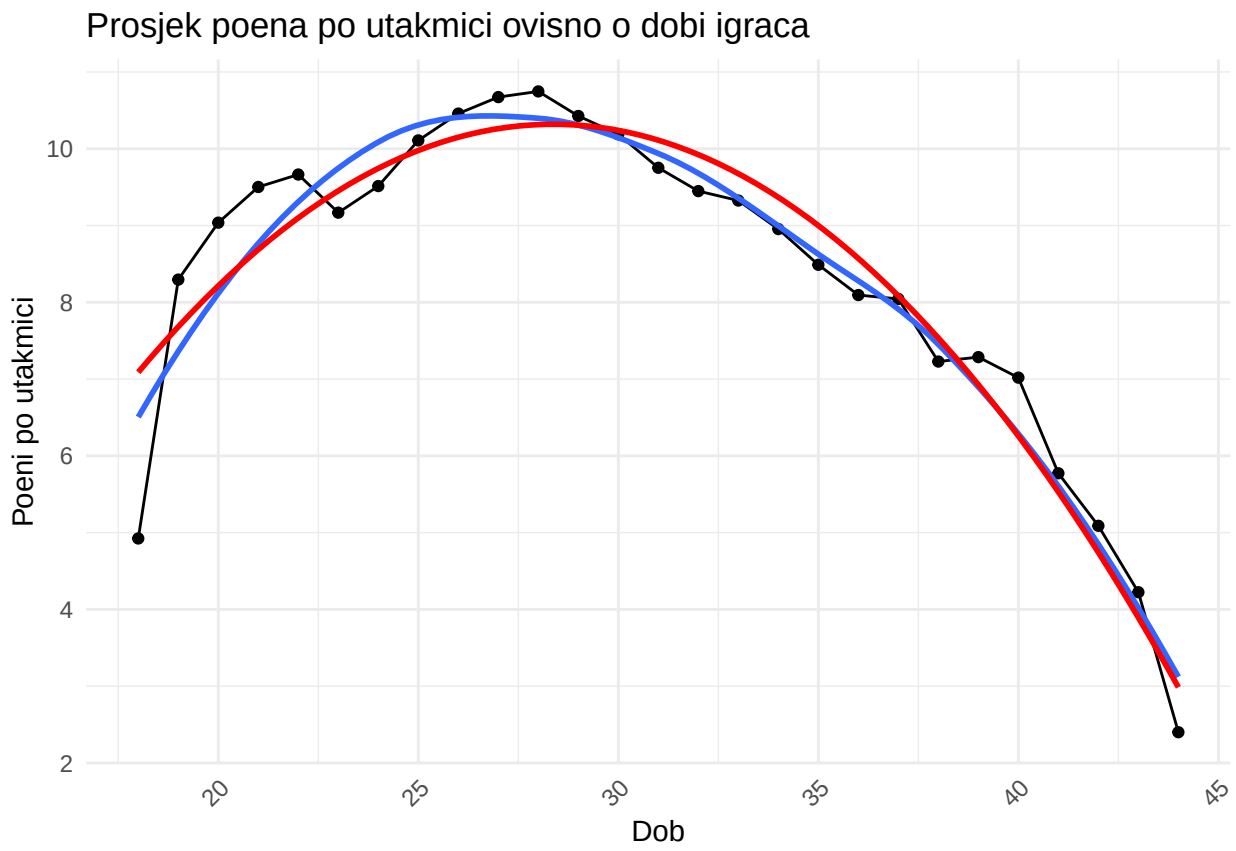
```
##    Age TotalPoints TotalGames AveragePointsPerGame
## 1   18       797.7        162             4.924074
## 2   19     29475.2       3553             8.295863
## 3   20    143367.1      15864             9.037260
## 4   21    260154.5      27379             9.501972
## 5   22    389561.8      40309             9.664388
## 6   23    511265.1      55768             9.167714
## 7   24    596214.3      62668             9.513856
## 8   25    602888.7      59642            10.108459
## 9   26    584614.5      55901            10.458033
## 10  27    574892.3      53864            10.673034
## 11  28    525864.8      48932            10.746849
## 12  29    463260.1      44430            10.426741
## 13  30    418328.2      41060            10.188217
## 14  31    339755.1      34839             9.752148
## 15  32    278378.3      29462             9.448724
## 16  33    227657.9      24412             9.325655
## 17  34    168837.1      18856             8.954025
## 18  35    116815.5      13761             8.488882
## 19  36     84199.0      10402             8.094501
## 20  37     58189.5       7234             8.043890
## 21  38     30626.4       4237             7.228322
## 22  39     17967.3       2466             7.286010
## 23  40      8394.4       1196             7.018729
## 24  41      1997.4        346             5.772832
## 25  42       798.9        157             5.088535
## 26  43       502.6        119             4.223529
## 27  44        12.0          5             2.400000
```

Sada ćemo vizualizirati ovisnost o starosti igrača i prosjeku poena po utakmici. Na x osi nalaze se podaci o godinama igrača, a na y osi nalaze se prosjeci poena po utakmici po godinama.

```
ggplot(age_summary, aes(x = Age, y = AveragePointsPerGame)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE) +
```

```
  geom_smooth(method = "lm", formula = y ~ poly(x, 2),
  color = "red", se = FALSE) +
  theme_minimal() +
  labs(x = "Dob", y = "Poeni po utakmici",
       title = "Prosjek poena po utakmici ovisno o dobi igrača") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Model plave boje na vizualizaciji i koji opisuje prosjek poena po utakmici ovisno o dobi igrača je loess, što je kratica za lokalnu regresiju.

Lokalna je regresija ne parametarski pristup koji prilagođava više regresora u lokalnom okruženju, što je utoliko korisnije kako znamo minimalnu i maksimalnu dob igrača u našem skupu podataka.

Krivulja crvene boje je vizualizacija prilagodbe polinomnog modela drugog stupnja.

ZAKLJUČAK:

- a) postoji ne linearna veza između dobi igrača i performansi na terenu u kontekstu postignutih poena na utakmici. Performanse ne opadaju niti rastu linearno, već prate zakrivljenu trajektoriju.
- b) vrhunac karijere: u sportskom žargonu, najbolje godine igrača kada isti nastupa ponajbolje u svojoj karijeri i u odnosu na ostatak lige. Zaključak je kako postoji trenutak kada rast poena, prisutan od rookie (prve) sezone, krene stagnirati; upravo ta stagnacija gdje igrači postižu najviše poena tokom godina predstavlja vrhunac karijere
- c) Loess vs polinomna prilagodba: loess krivulja blisko prati stvarne podatke, pružajući fleksibilnu prilagodbu lokalnim varijacijama podataka. S druge strane, polinomna krivulja isto dobro

objašnjava globalni trend, ali je glađa i ne prilagođava se lokalnim fluktuacijama kao loess krivulja, ali obje krivulje objašnjavaju sličan trend u podacima.

- d) Igrači su u najboljim godine, ukoliko je gruba granica 10 poena po utakmici (predefinirana s naše strane), od svoje 25. do svoje 30. godine života
- e) Postepeni rast poena do dobi od 25. te postepeni pad nakon 30. godine mogu biti objašnjeni, osim neiskustvom za ranije godine te padom fizičke spreme nakon 30. godine i ulogom u timu, ali kako nemamo podatke o tome, ne možemo zaključiti ništa po tom pitanju.

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

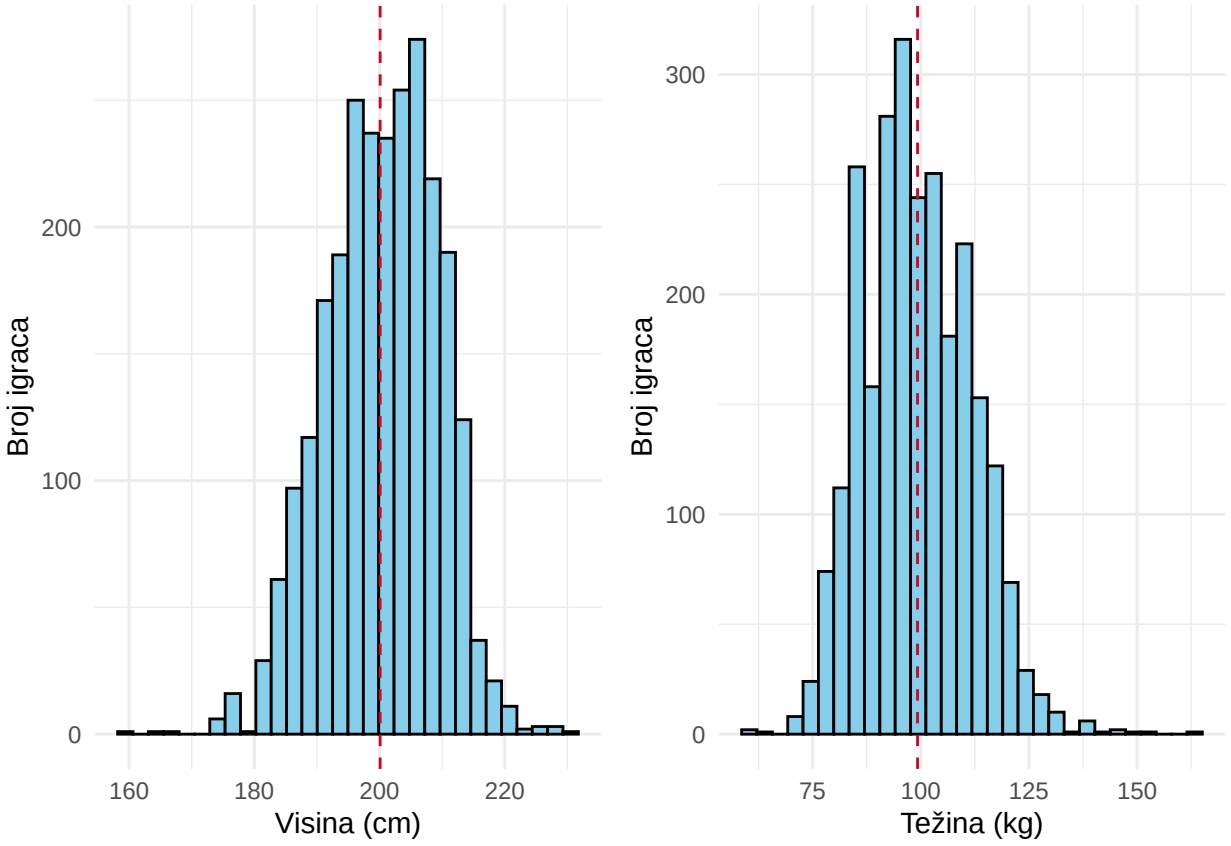
BONUS: Prije samog modeliranja ovisnosti poena o biometrijskim podacima, provjeravamo zavisnost ostalih važnih komponenata, odnosno skokova i asistencija, o biometrijskim podacima
Pretpostavke:

- niži igraći imaju više asistencija i ukradenih lopti te manje skokova
- viši igrači manje asistencija te više skokova i blokada
- težina, visina i dob dobro opisuju igračev prosjek poena.

```
težine_visine <- nba_data %>%
  group_by(player_name) %>%
  summarise(player_height = mean(player_height, na.rm = TRUE),
            player_weight = mean(player_weight, na.rm = TRUE))
```

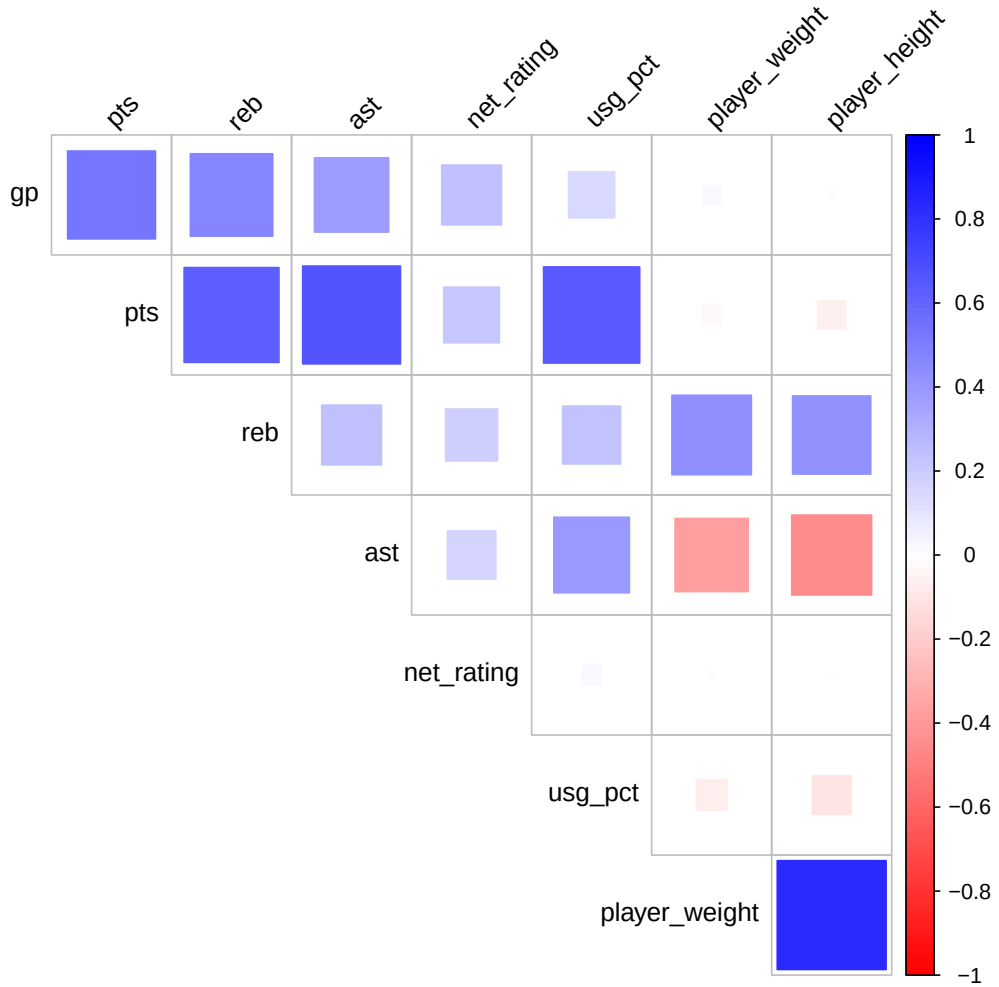
Vizualizacija težina i visina igrača

```
p1 <- ggplot(težine_visine, aes(x = player_height)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  geom_vline(aes(xintercept = mean(player_height, na.rm = TRUE)),
             color = '#c9082a', linetype = "dashed") +
  labs(y = "Broj igrača", x = "Visina (cm)") +
  theme_minimal()

p2 <- ggplot(težine_visine, aes(x = player_weight)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  geom_vline(aes(xintercept = mean(player_weight, na.rm = TRUE)),
             color = '#c9082a', linetype = "dashed") +
  labs(y = "Broj igrača", x = "Težina (kg)") +
  theme_minimal()

grid.arrange(p1, p2, ncol = 2)
```




```r
df_corr <- nba_data %>%
  filter(season != '2019-01-01') %>%
  select(gp, pts, reb, ast, net_rating, usg_pct, player_weight, player_height)

corr <- cor(df_corr, use = "complete.obs")

corrplot(corr, method = "square", type = "upper",
         tl.col = "black", tl.srt = 45, diag = FALSE,
         col = colorRampPalette(c("red", "white", "blue"))(200))
```

Vizualizacija korelacijske matrice između značanjih varijabli

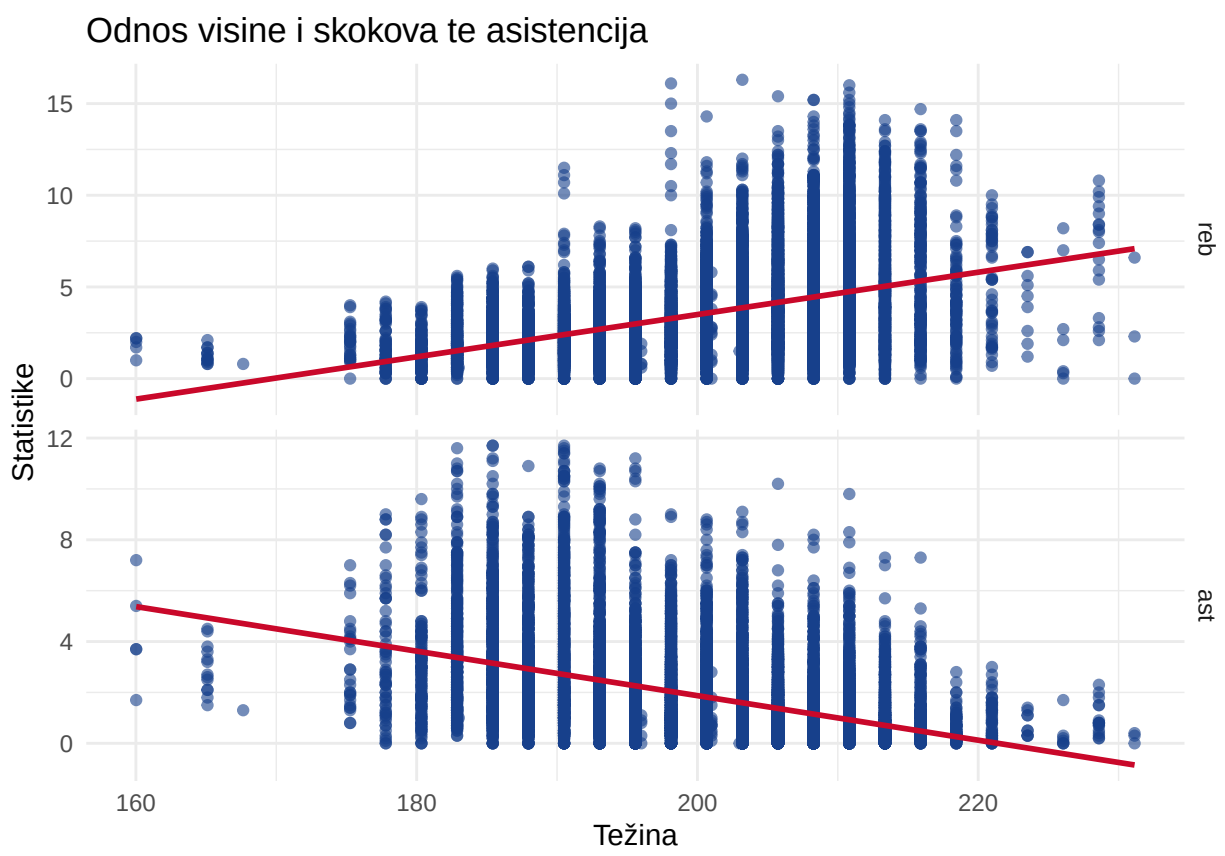



```
df_melted <- reshape2::melt(df_corr, id.vars = c("player_height", "player_weight"))

df_filtered <- df_melted %>%
  filter(variable %in% c("reb", "ast"))

ggplot(df_filtered, aes(x = player_height, y = value)) +
  geom_point(color = '#17408b', alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = '#c9082a') +
  facet_grid(variable ~ ., scales = "free") +
  theme_minimal() +
  labs(x = "Težina", y = "Statistike") +
  theme(strip.text.x = element_text(size = 12)) +
```

```
  ggtitle("Odnos visine i skokova te asistencija")
```

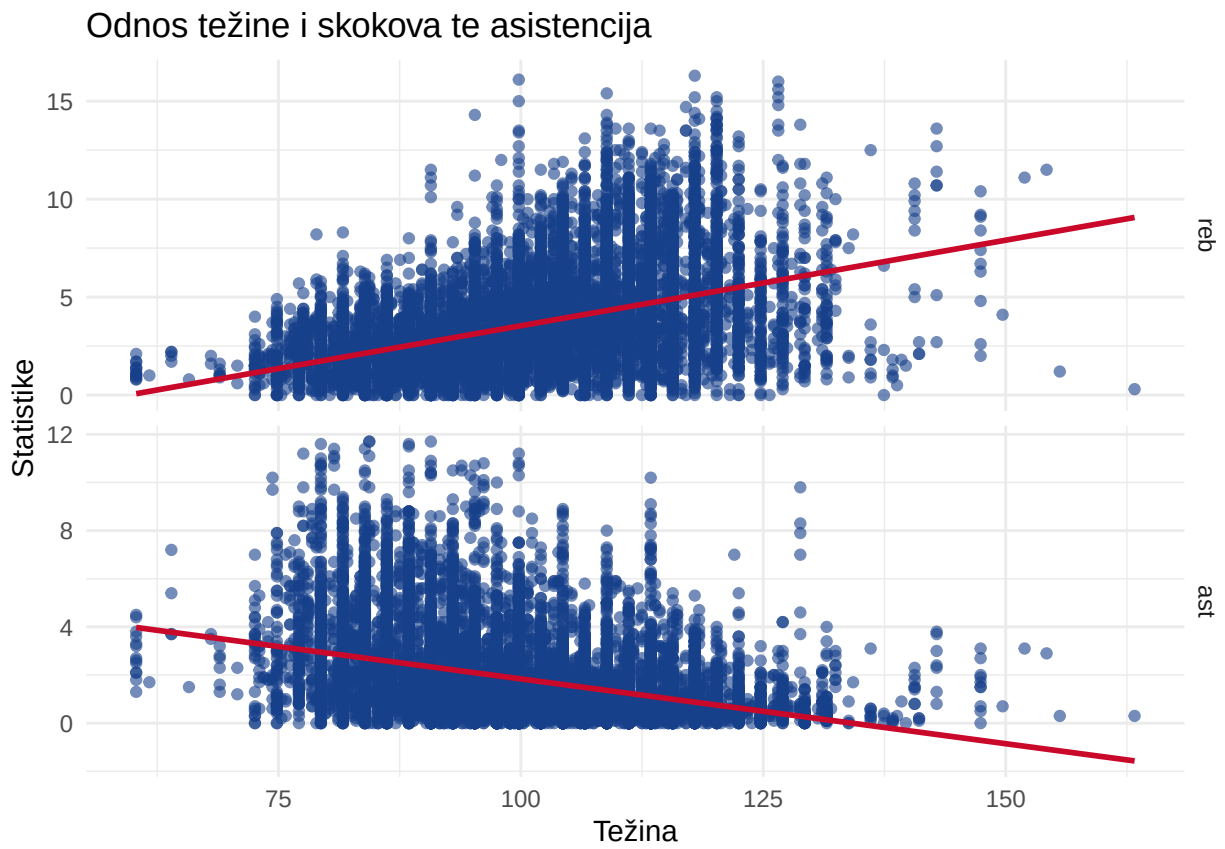
Vizualizacija ovisnosti broja skokova i asistencija ovisno o igračevoj težini, odnosno visini.

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(df_filtered, aes(x = player_weight, y = value)) +
  geom_point(color = '#17408b', alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = '#c9082a') +
  facet_grid(variable ~ ., scales = "free") +
  theme_minimal() +
  labs(x = "Težina", y = "Statistike") +
  theme(strip.text.x = element_text(size = 12)) +
  ggtitle("Odnos težine i skokova te asistencija")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



ZAKLJUČAK : dvije naše pretpostavke su na temelju grafova opravdane; teži/viši igrači imaju veći prosjek skokova po utakmici te manji broj asistencija po utakmici u odnosu na niže igraće, koji imaju više asistencija i manje skokova u odnosu na više igrače.

Sada ćemo promatrati broj poena po utakmici ovisno o visini i težini igrača.

Pomoću modela linerane regresije probati ćemo opisati ovisnost poena igrača o težini i visini igrača, kao i o dobi igrača.

```
# Treniranje modela na cijelom datasetu
model <- lm(pts ~ player_height + player_weight + age, data = nba_data)

# Prikaz sažetka modela
summary(model)

# Predviđanja modela na istom datasetu
predictions <- predict(model, newdata = nba_data)

# Izračunavanje Mean Squared Error (MSE)
MSE <- mean((predictions - nba_data$pts)^2)
print(paste("MSE:", MSE))

# Izračunavanje R kvadrat (R²) - koeficijent determinacije
SST <- sum((nba_data$pts - mean(nba_data$pts))^2)
SSR <- sum((predictions - nba_data$pts)^2)
R_squared <- 1 - (SSR/SST)
print(paste("R²:", R_squared))
```

```
##
## Call:
## lm(formula = pts ~ player_height + player_weight + age, data = nba_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.978 -4.596 -1.507  3.237 27.536
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.997214   1.563165  12.153  < 2e-16 ***
## player_height -0.069920   0.010272  -6.807 1.04e-11 ***
## player_weight  0.029829   0.007546   3.953 7.77e-05 ***
## age            0.009152   0.012304   0.744    0.457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.004 on 12840 degrees of freedom
## Multiple R-squared:  0.004387,   Adjusted R-squared:  0.004154
## F-statistic: 18.86 on 3 and 12840 DF,  p-value: 3.36e-12
##
## [1] "MSE: 36.0375462010074"
## [1] "R²: 0.00438696081898304"
```

ZAKLJUČAK:

Model linearne regresije analizirao je utjecaj visine (player_height), težine (player_weight) i dobi (age) igrača na broj poena po utakmici (pts). Prema modelu, odsječak na y-osi iznosi 18.9972, što sugerira da bi teoretski igrač s visinom 0 i težinom 0 u prosjeku postigao oko 18.997214 poena po utakmici. Ova statistika nije praktično relevantna jer ne postoji igrač s takvim karakteristikama.

Koeficijenti za visinu i težinu su statistički značajni (p-vrijednosti < 0.05), što znači da oni imaju statistički značajan utjecaj na broj postignutih poena po utakmici. Koeficijent za dob nije statistički značajan (p-vrijednost > 0.05), što ukazuje da dob, unutar ovog modela, nema značajan utjecaj na broj postignutih poena.

Koeficijent R^2 iznosi 0.004387, što znači da model objašnjava samo oko 0.44% varijabilnosti u broju postignutih poena. Ovo je vrlo nizak postotak, što ukazuje na to da model ne objašnjava dobro varijabilnost ciljane varijable, odnosno da možda postoje drugi faktori koji bolje objašnjavaju broj postignutih poena.

Rezidualna standardna pogreška je oko 6, što sugerira da su predviđanja modela u prosjeku udaljena za 6 poena od stvarnih vrijednosti.

Ukupno gledano, iako su visina i težina statistički značajni prediktori, model ukazuje na to da oni sami po sebi ne pružaju snažno objašnjenje za varijabilnost u broju postignutih poena po utakmici. Drugi faktori, koji nisu uključeni u model, mogli bi dati bolje objašnjenje i trebali bi biti razmotreni u dubljim analizama.