Projekt

Analiza podataka o igračima NHL lige

Uvod

U ovom projektu analizirat ćemo podatke o igračima NHL lige. Dobiveni podaci su opširni i sadrže mnogo međusobno povezanih informacija te ćemo se zato fokusirati na manji odabrani podskup. Bavit ćemo se osvojenim bodovima igrača, preferiranom rukom za udarce, pozicijama i plaćama igrača koje ćemo naposlijetku izmodelirati regresijom. Dodati još neke stvari o hokeju jer ja ne znam nista o hokeju

Analiza u ovom radu sastoji se od tri dijela: deskriptivna analiza, testovi sredina i intervalne procjene, te analiza zasnovana na linearnoj regresiji i analizi varijance.

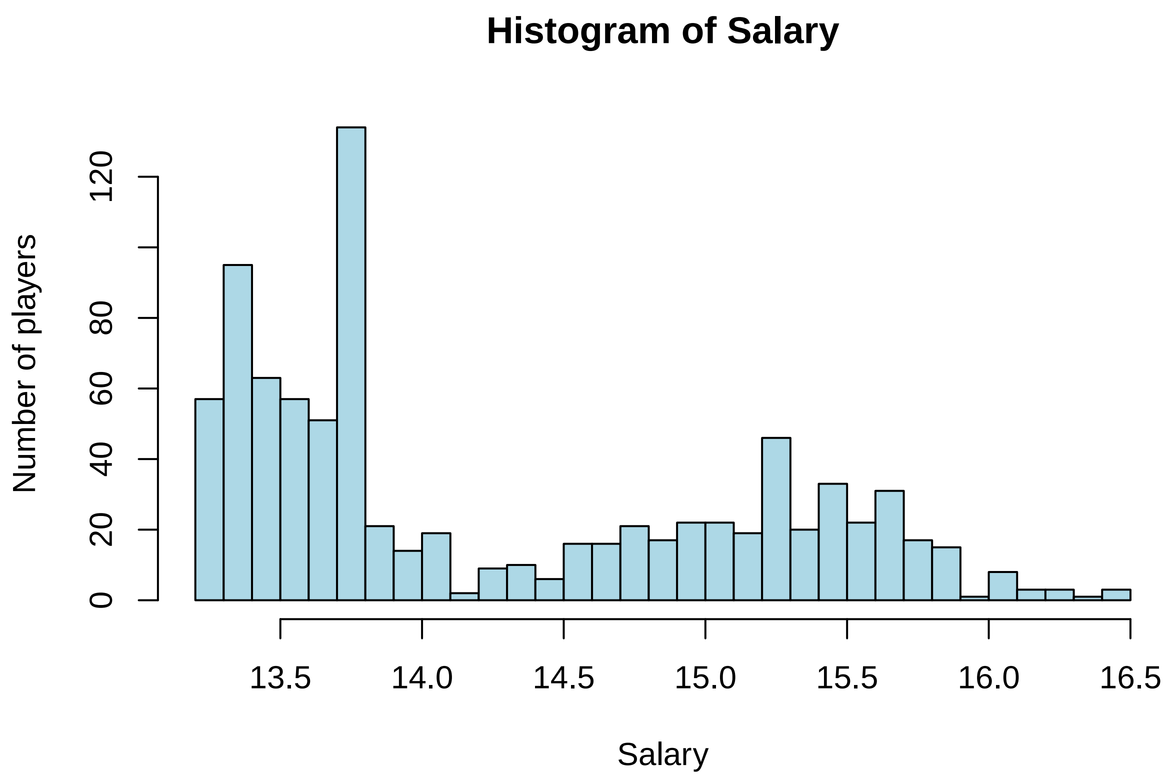
Deskriptivna analiza

Učitavamo podatke i analiziramo kako oni izgledaju.

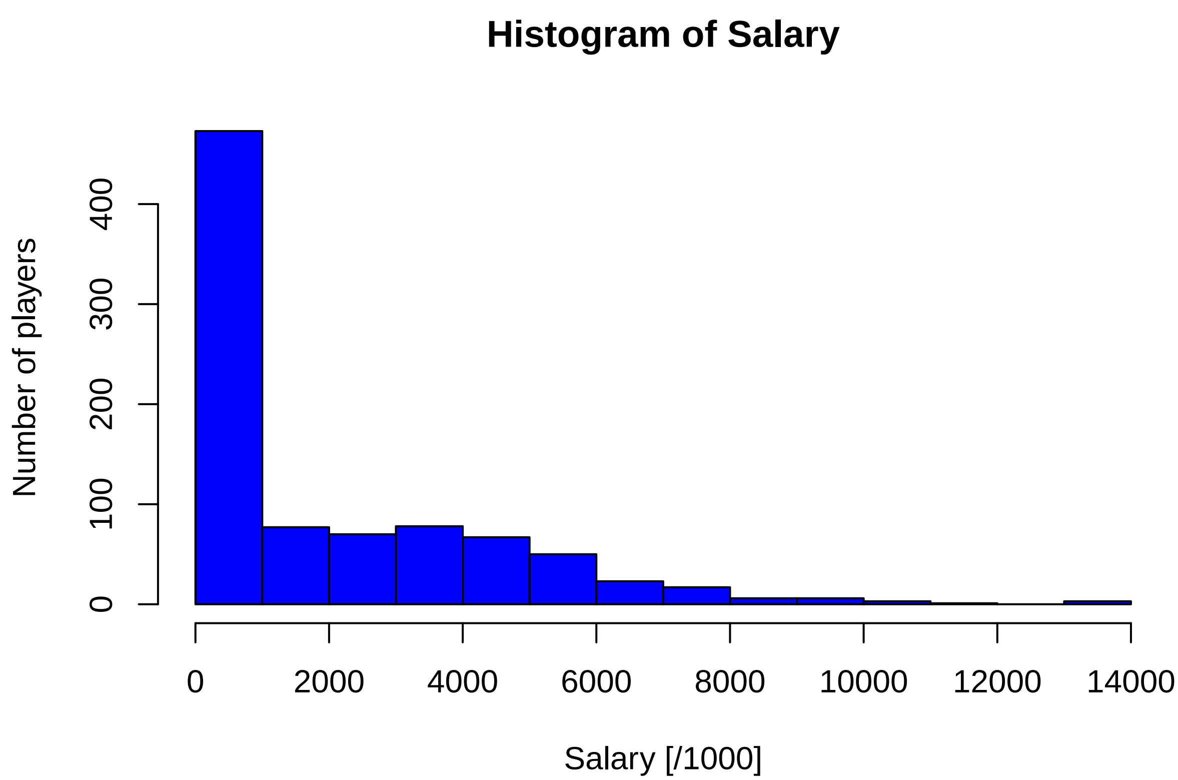
```
s = players.data$Salary
s = as.data.frame.numeric(s)
s <- as.numeric(unlist(s))
```

Histogramima ćemo prikazati razdiobu plaća igrača na logaritamskoj skali i umanjene 1000 puta. Promatranjem prikaza i testom normalnosti zaključujemo da se ne radi o normalnoj distribuciji, nego desno zakrivljenoj.

```
hLog = hist(log(s), main = paste("Histogram of Salary"),
            breaks = 40,
         xlab = "Salary",
         ylab = "Number of players",
         col = "lightblue")
```




```r
h100 = hist(s/1000, main = paste("Histogram of Salary"),
         xlab = "Salary [/1000]",
         ylab = "Number of players",
         col = "blue")
```



```r
cat('Prosječna plaća igrača iznosi ',
    mean(s, na.rm = TRUE),'dolara.\n')
```

```
## Prosječna plaća igrača iznosi  2325289 dolara.
```

```r
#qqnorm(s, pch = 1, frame = FALSE)
#shapiro.test(s)
```

Testiranje hipoteza

Pretpostavka: igrači na nekoj određenoj poziciji značajno su više plaćeni od od drugih igrača

Igrače s miješanim pozicijama generalizirat ćemo po prvoj spomenutoj poziciji tako da svaki igrač ima samo jednu poziciju.

```r
for (column_name in c("LW/RW", "LW/C", "LW/C/RW", "LW/RW/C", "LW")){
     players_copy$Position[players_copy$Position == column_name] = "L";
 }
for (column_name in c("RW/C", "C/RW", "RW/LW","LW/C/RW", "RW/C/LW", "RW/LW/C", "C/LW/RW", "C/RW/LW", "RW
     players_copy$Position[players_copy$Position == column_name] = "R";
}
for (column_name in c("LW/C", "C/LW","C/LW/RW", "C/RW/LW", "C", "C/RW", "C/LW/C")){
     players_copy$Position[players_copy$Position == column_name] = "C";
}
for (column_name in c("D/C", "C/D", "D/RW", "D/LW")){
     players_copy$Position[players_copy$Position == column_name] = "D";
}
#for (column_name in c("LW/C/RW", "C/RW/LW", "LW/RW/C", "RW/C/LW", "RW/LW/C", "C/LW/RW", "C/RW/LW")){
  #   players_copy$Position[players_copy$Position == column_name] = "C/LW/RW";
#}
```

Iz slijedećih boxplotova nije vidljiva bitna razlika u plaćama igrača na različitim pozicijama. Zamjećujemo da pozicija "D" ima nešto veći medijan od ostalih pozicija. To možemo pokušati provjeriti ANOVA metodom.

Pretpostavke ANOVA metode su nezavisnost podataka, normalna distribucija i homogenost varijanci, pa ćemo homogenost varijanci provjeriti Bartletovim testom:

$$
\begin{aligned}
& H_0 : \sigma^2_{"L"} = \sigma^2_{"R"} = \sigma^2_{"C"} = \sigma^2_{"D"} \\
& H_1 : \neg H_0.
\end{aligned}
$$

razine značajnosti $\alpha = 0.05$.

```r
l <- subset(players.data.d, Position == "L")
bartlett.test(Salary~Position, players_copy)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Salary by Position
## Bartlett's K-squared = 8.5412, df = 3, p-value = 0.03606
```

Dobivena p vrijednost manja je od razine značajnosti što znači da se odbacuje H_0 pa ne možemo koristiti ANOVA-u. Umjesto ANOVA-e provest ćemo neparametarski test, Kruskal-Wallis test razine značajnosti $\alpha = 0.05$ koji za sobom ne povlači pretpostavke koje dolaze s parametarskim testovima. Kruskal-Wallis test slabiji je od ANOVA-e i uspoređuje medijane, ali ovaj test u kombinaciji s gornjim prikazom dokazat će približnu jednakost plaća po pozicijama igrača.

$$H_0 : M_{"L"} = M_{"R"} = M_{"C"} = M_{"D"}$$
$$H_1 : \neg H_0.$$

```
kruskal.test(Salary~Position, players_copy)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Salary by Position
## Kruskal-Wallis chi-squared = 3.2485, df = 3, p-value = 0.3549
```

Dobivena p-vrijednost veća je od razine značajnosti te iz toga zaključujemo da su plaće igrača ne razlikuju značajno po njihovim pozicijama. Ne odbacujemo H_0.

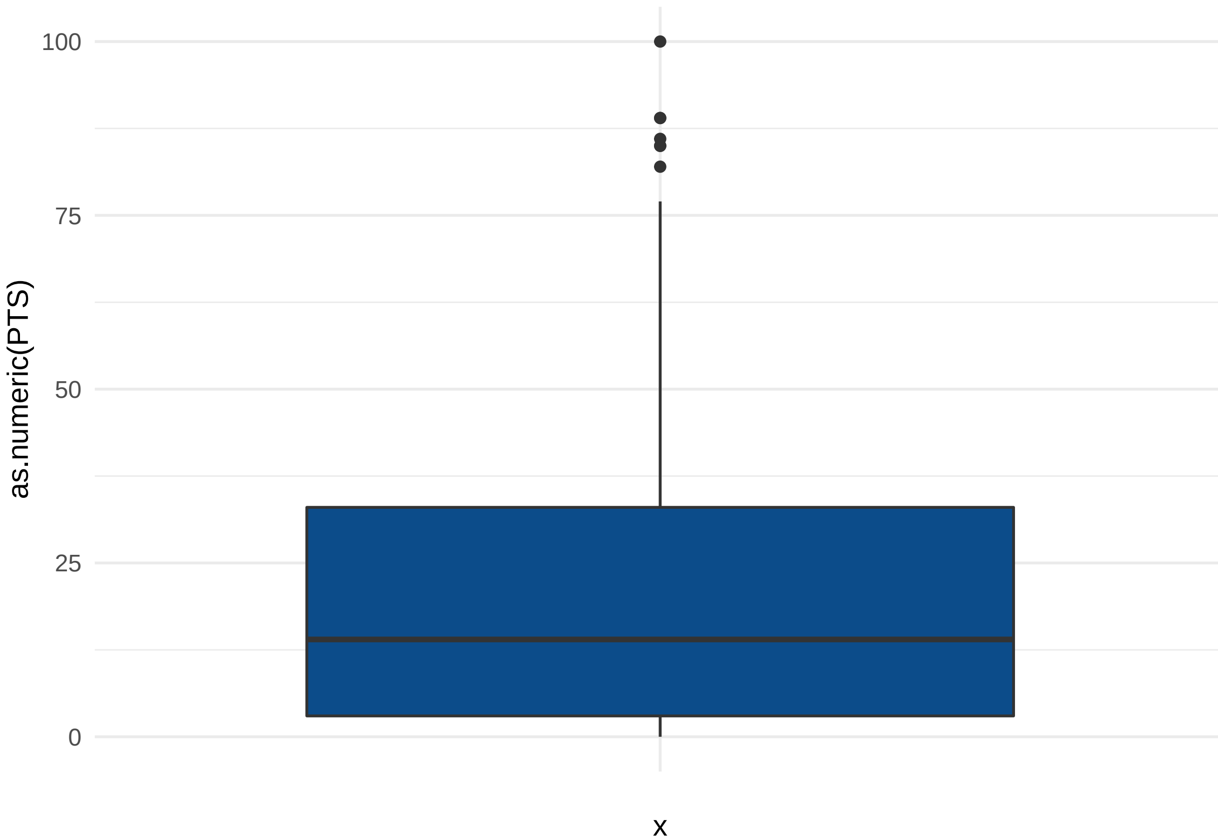
Pretpostavka: igrači na nekoj određenoj poziciji postižu značajno više bodova od drugih igrača

```
require(fastDummies)
players.data.d <- dummy_cols(players_copy,select_columns='Position')
#View(players.data.d)
```

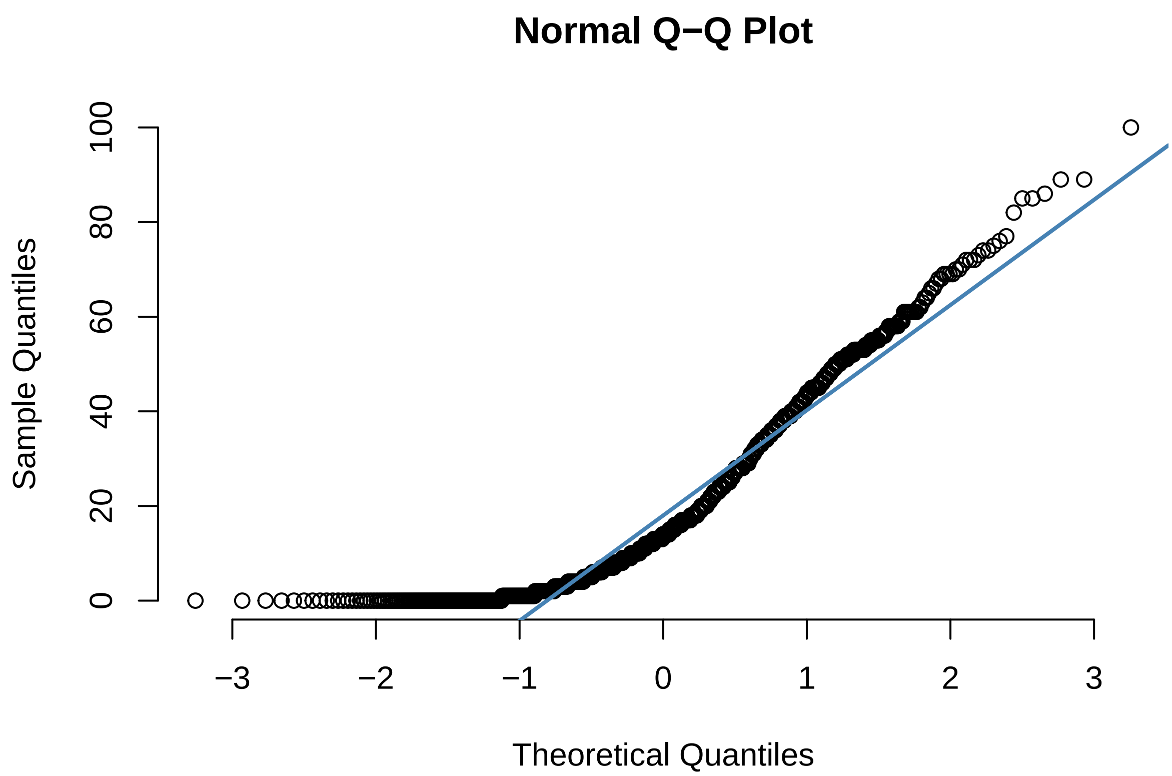
```
#data of potential outliers
players.data.d$PTS <- as.numeric(players.data.d$PTS)
out<-boxplot.stats(players.data.d$PTS)$out
out_ind <- which(players.data.d$PTS %in% c(out))
out_ind
```

```
## [1]  22 150 390 427 486 510 705
```

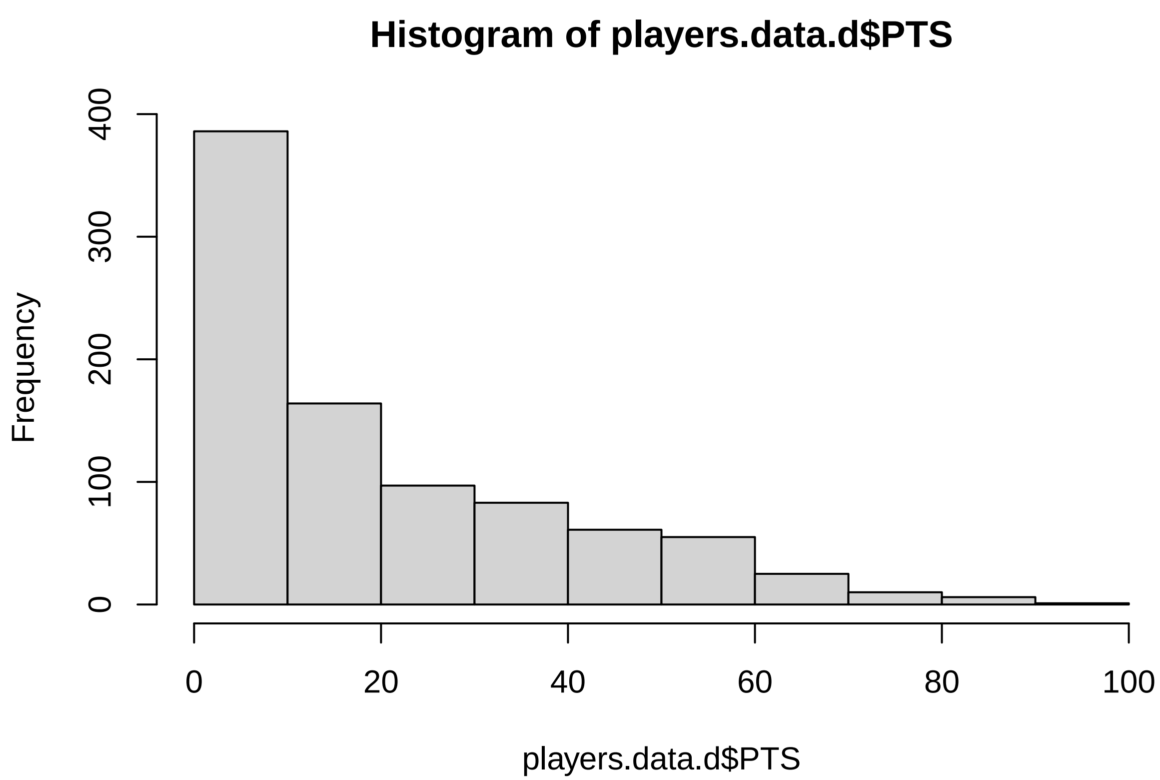
```
#boxplot
ggplot(players.data.d) +
  aes(x = "", y = as.numeric(PTS)) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```




```r
#qq plot
qqnorm(players.data.d$PTS, pch = 1, frame = FALSE)
qqline(players.data.d$PTS, col = "steelblue", lwd = 2)
```




```r
#histogram
hist(players.data.d$PTS)
```



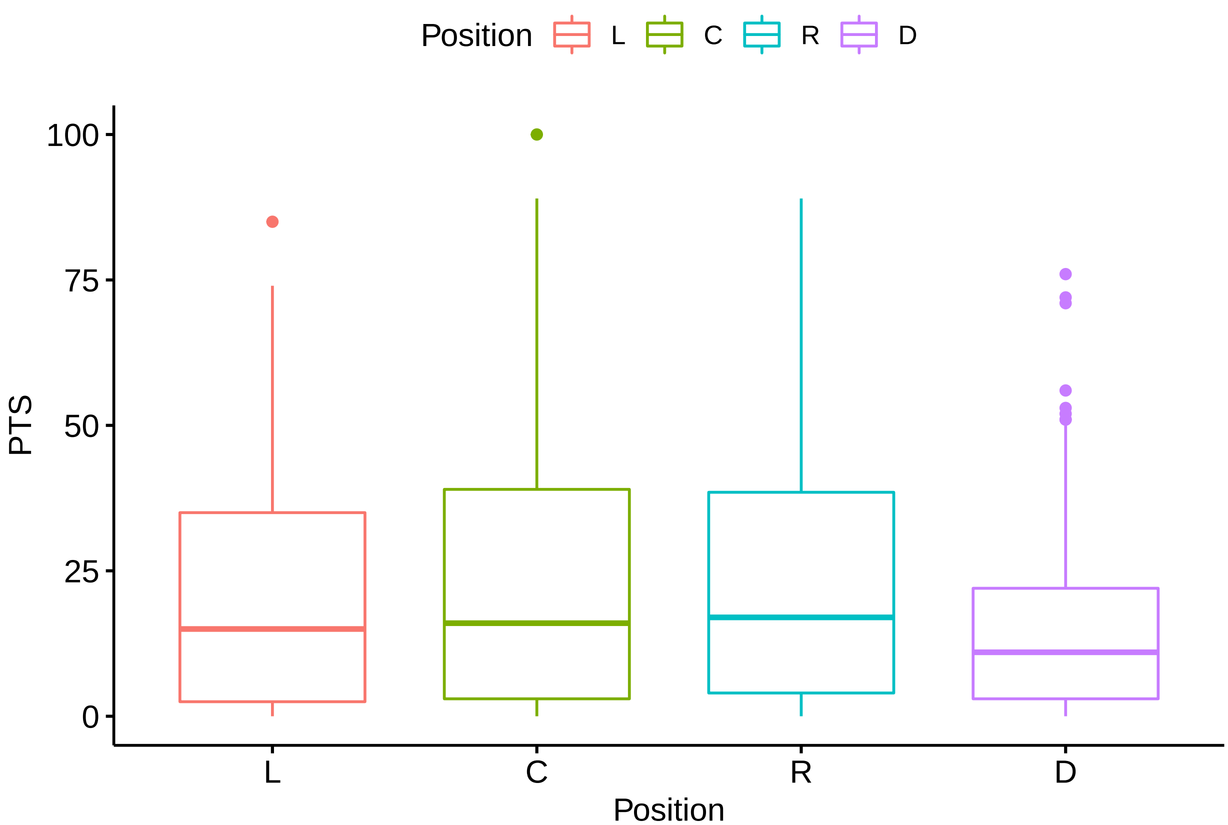
Iz histograma zaključujemo da zavisna varijabla(PTS) nije normalno distribuirana. Zato koristimo Wilcoxon-Mann-Whitney test

```r
library(dplyr)
group_by(players.data.d, Position) %>%
  summarise(
    count = n(),
    median = median(PTS, na.rm = TRUE),
    IQR = IQR(PTS, na.rm = TRUE)
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 4
##   Position count median   IQR
##   <chr>    <int>  <dbl> <dbl>
## 1 C          201     16  36
## 2 D          300     11  19
## 3 L          175     15  32.5
## 4 R          212     17  34.5
```

```r
library("ggpubr")
View(players.data.d)
ggboxplot(players.data.d, x = "Position", y = "PTS",
          color = "Position",
          ylab = "PTS", xlab = "Position")
```




```r
res2 <-cor.test(players.data.d$Position_C, players.data.d$Position_R,  method = "spearman")
```

```
## Warning in cor.test.default(players.data.d$Position_C,
## players.data.d$Position_R, : Cannot compute exact p-value with ties
```

```r
res2
```

```
##
##  Spearman's rank correlation rho
##
## data:  players.data.d$Position_C and players.data.d$Position_R
## S = 152055334, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.3029104
```

```r
res1 <-cor.test(players.data.d$Position_L, players.data.d$Position_R,  method = "spearman")
```

```
## Warning in cor.test.default(players.data.d$Position_L,
## players.data.d$Position_R, : Cannot compute exact p-value with ties
```

```r
res1
```

```
##
##  Spearman's rank correlation rho
##
## data:  players.data.d$Position_L and players.data.d$Position_R
## S = 149082805, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2774399
```

Prema tablici možemo zaključiti da najveći broj bodova postižu igrači na lijevom i desnom krilu, s medijanom 17.0, odnosno 15.5, što nam i potvrđuje boxplot.

```r
kruskal.test(players.data.d$Position, players.data.d$PTS)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  players.data.d$Position and players.data.d$PTS
## Kruskal-Wallis chi-squared = 111.56, df = 81, p-value = 0.01379
```

```r
#f test
var.test(players.data.d$Position_L, players.data.d$Position_R, conf.level = 0.95, paired = T)
```

```
##
##  F test to compare two variances
##
## data:  players.data.d$Position_L and players.data.d$Position_R
## F = 0.87065, num df = 887, denom df = 887, p-value = 0.03928
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7632155 0.9932141
## sample estimates:
## ratio of variances
##          0.8706528
```

Iz provedenog F testa vidimo da se hipoteza o jednakosti varijanci ne može odbaciti na razini znacajnosti od 5%, stoga cemo u t-testu postaviti parametar var.equal na True.

```r
#t test
t.test(players.data.d$Position_L, players.data.d$Position_R, conf.level = 0.95, paired = T)
```

```
##
##  Paired t-test
##
## data:  players.data.d$Position_L and players.data.d$Position_R
## t = -1.8835, df = 887, p-value = 0.05996
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.085083838  0.001750505
## sample estimates:
## mean of the differences
##             -0.04166667
```

Iz provedenog t-testa uocavamo da je p vrijedost izuzetno malena, stoga hipotezu da su aritmeticke sredine kategorija position_L i position_R jednake odbacujemo na razini znacajnosti od 5%. U nastavku cemo provjeriti imaju li te dvije kategorije jednaku razdiobu.

```r
wilcox.test(players.data.d$Position_L, players.data.d$Position_R, conf.level = 0.95, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  players.data.d$Position_L and players.data.d$Position_R
## V = 33950, p-value = 0.06003
## alternative hypothesis: true location shift is not equal to 0
```

Nakon provođenja Wilcoxonova testa dobivamo jednake rezultate kao i u t-testu, dakle odbacujemo nultu hipotezu.

```r
ks.test(players.data.d$Position_L, players.data.d$Position_R, conf.level = 0.95, var.equal = T, paired =
```

```
## Warning in ks.test(players.data.d$Position_L, players.data.d$Position_R, : p-
## value will be approximate in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  players.data.d$Position_L and players.data.d$Position_R
## D = 0.041667, p-value = 0.4239
## alternative hypothesis: two-sided
```

Rezultati provedenog Kolmogorov-Smirnovljeva idu u prilog odbacivanju hipoteze da kategorije lijevog i desnog krila imaju jednaku razdiobu.

##Ovisnost pozicije o preferiranoj ruci udarca

Nulta hipoteza je da varijable nemaju linearnu ovisnost, te ako dobijemo p vrijednost manju od 0.05 odbacit ćemo nultu hipotezu i prihvatiti alternativnu - varijable imaju linearnu ovisnost.

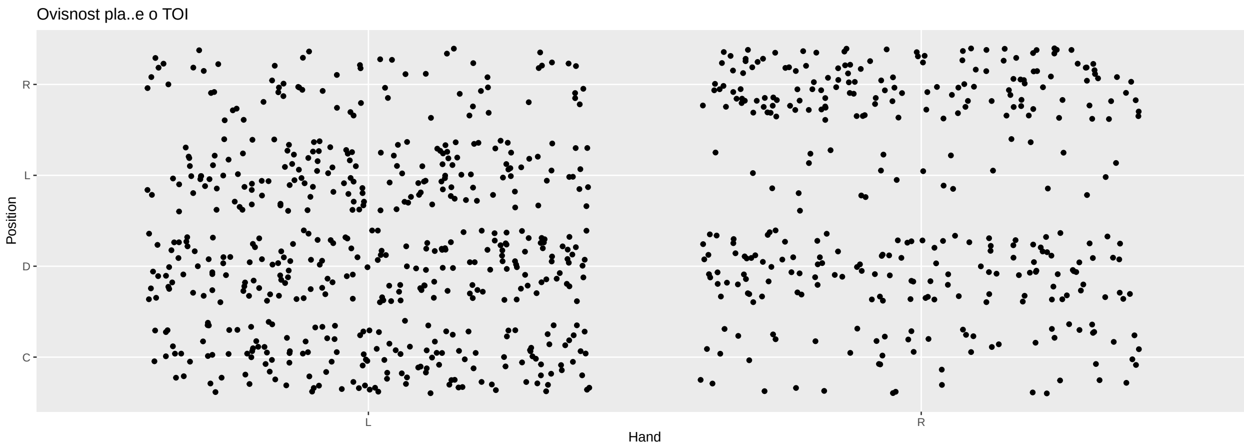
```r
ggplot(players_copy, aes(x =Hand , y = Position)) +
  geom_jitter() +
  geom_smooth(method = "lm") +
  labs(title = "Ovisnost plaće o TOI")
```

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

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
```

```
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <87>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <87>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <87>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <87>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <87>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <87>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <87>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
## for <c4>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Ovisnost plaće o TOI' in 'mbcsToSbcs': dot substituted
```

```
## for <87>
```




```r
tbl = table(players_copy$Position,
            players_copy$Hand)

added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##          L   R Sum
##   C    148  53 201
##   D    175 125 300
##   L    150  25 175
##   R     64 148 212
##   Sum  537 351 888
```

```r
for (col_names in colnames(added_margins_tbl)){
  for (row_names in rownames(added_margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,': ',(added_margins_tbl[row_names,
    }
  }
}
```

```
## Očekivane frekvencije za razred  L - C :  121.5507
## Očekivane frekvencije za razred  L - D :  181.4189
## Očekivane frekvencije za razred  L - L :  105.8277
## Očekivane frekvencije za razred  L - R :  128.2027
## Očekivane frekvencije za razred  R - C :  79.44932
## Očekivane frekvencije za razred  R - D :  118.5811
## Očekivane frekvencije za razred  R - L :  69.1723
## Očekivane frekvencije za razred  R - R :  83.7973
```

```r
#chi^2 test nezavisnosti
test <- chisq.test(tbl,correct=F)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 143.12, df = 3, p-value < 2.2e-16
```

```r
print("p-vrijednost chi^2 testa:")
```

```
## [1] "p-vrijednost chi^2 testa:"
```

```r
test$p.value
```

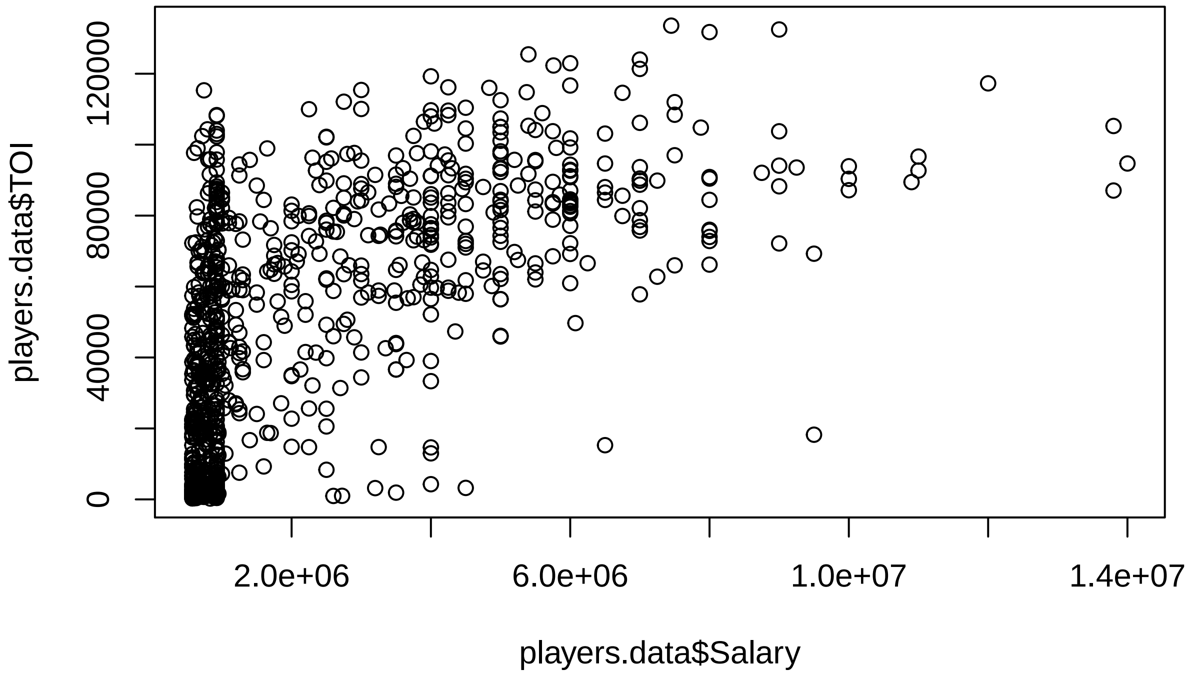
```
## [1] 8.019989e-31
```

Dobivena p-vrijednost je jako mala, što ide u prilog odbacivanju H0 hipoteze koja pretpostavlja nezavisnost varijabli pozicija i preferirane ruke.

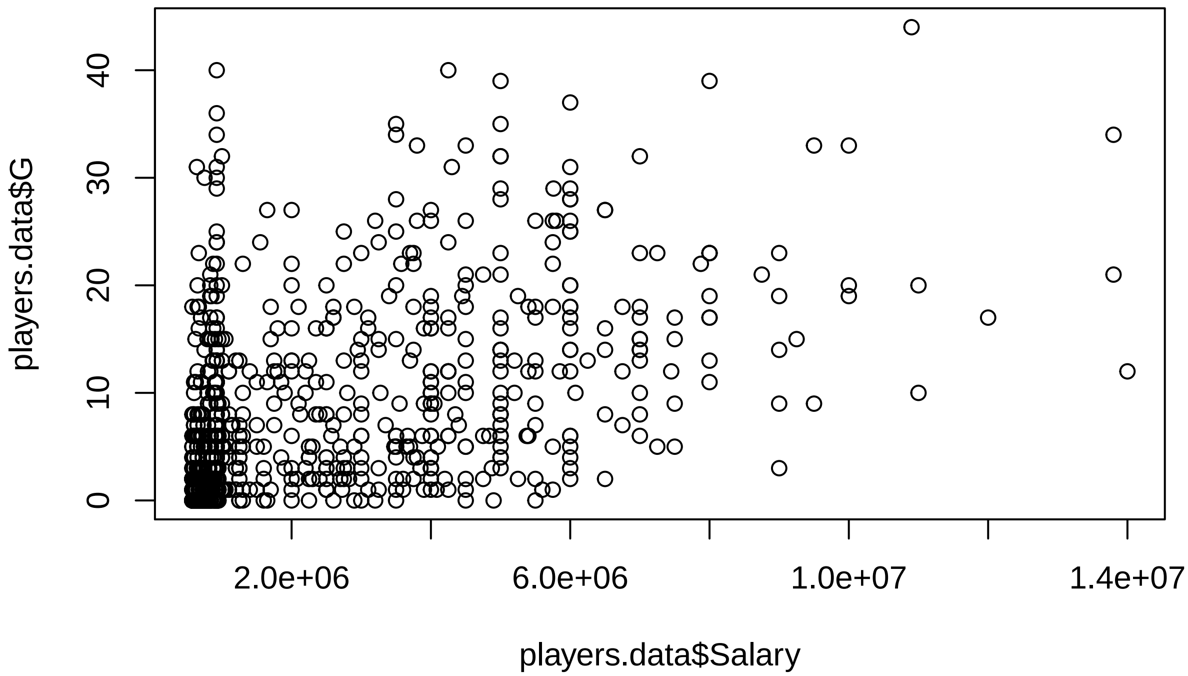
##Ovisnost plaće igrača o više varijabli

Za problem određivanja veze plaća igrača i ostalih faktora koji utječu na plaća napravit ćemo kompleksniji model koji će koristiti više regresora - model višestruke regresije. Kao regresore uzet ćemo ?? i ??, a za reakciju plaće igrača.

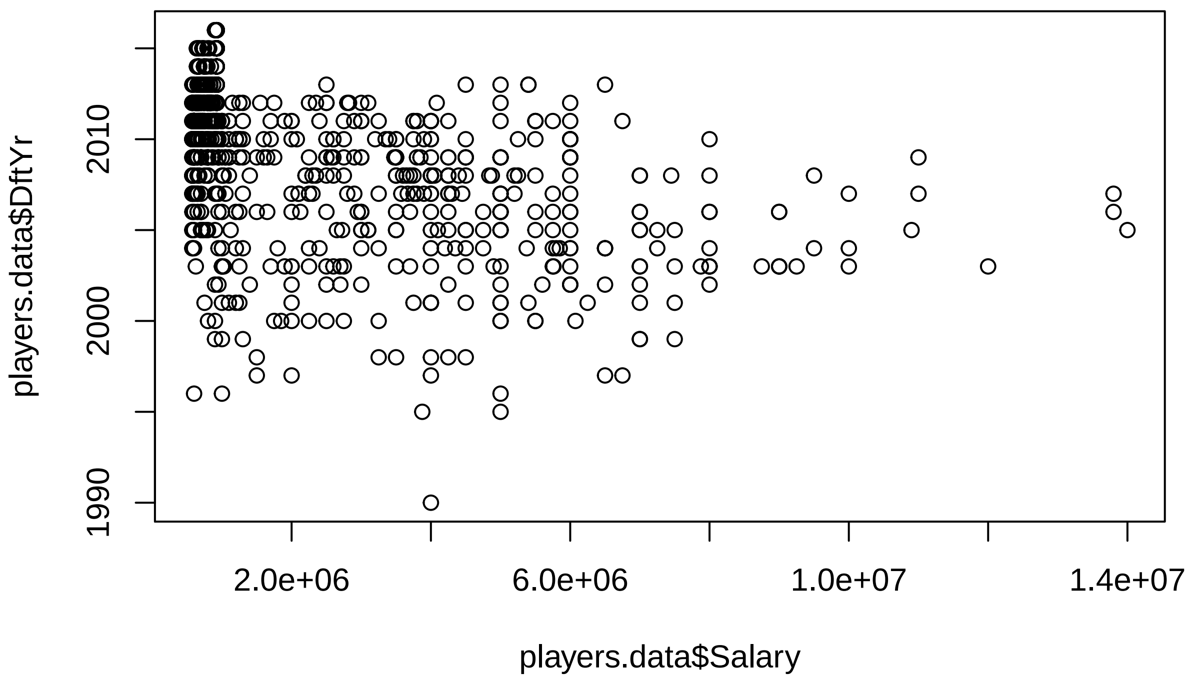
```r
plot(players.data$Salary, players.data$TOI) #salary vs time on ice
```



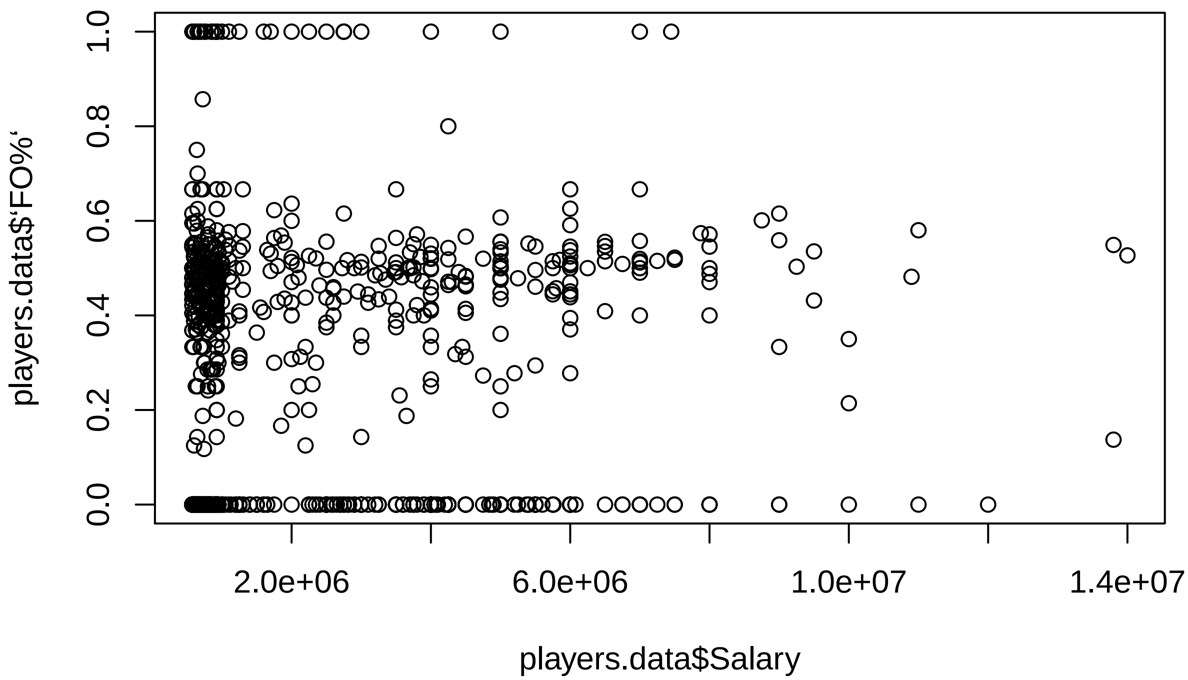

```r
plot(players.data$Salary, players.data$G) #salary vs goals
```



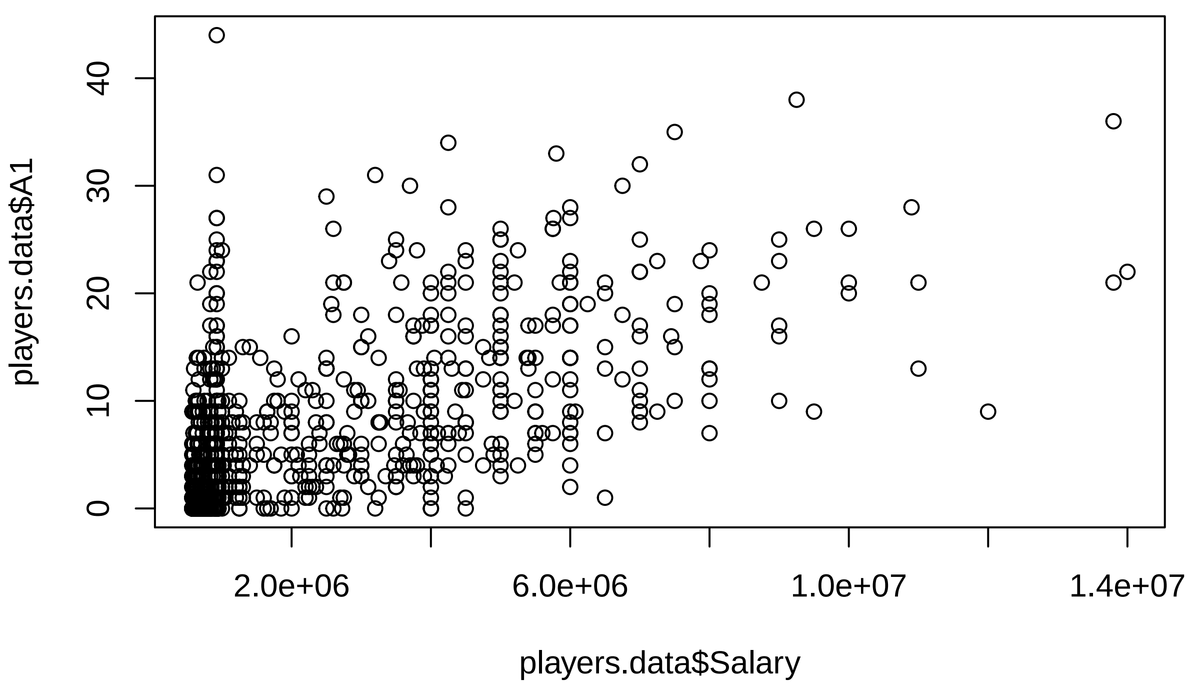

```r
plot(players.data$Salary, players.data$DftYr)#salary vs year drafted
```



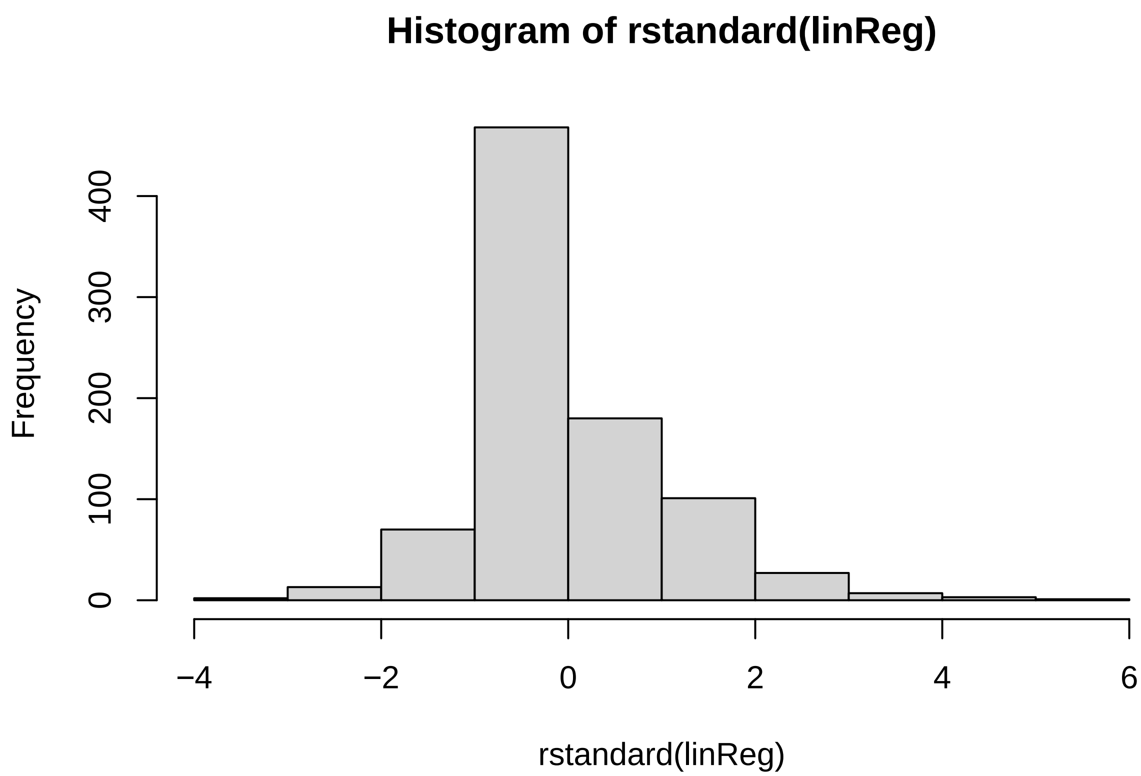

```r
plot(players.data$Salary, players.data$`FO%`) #salary vs faceoff wins
```



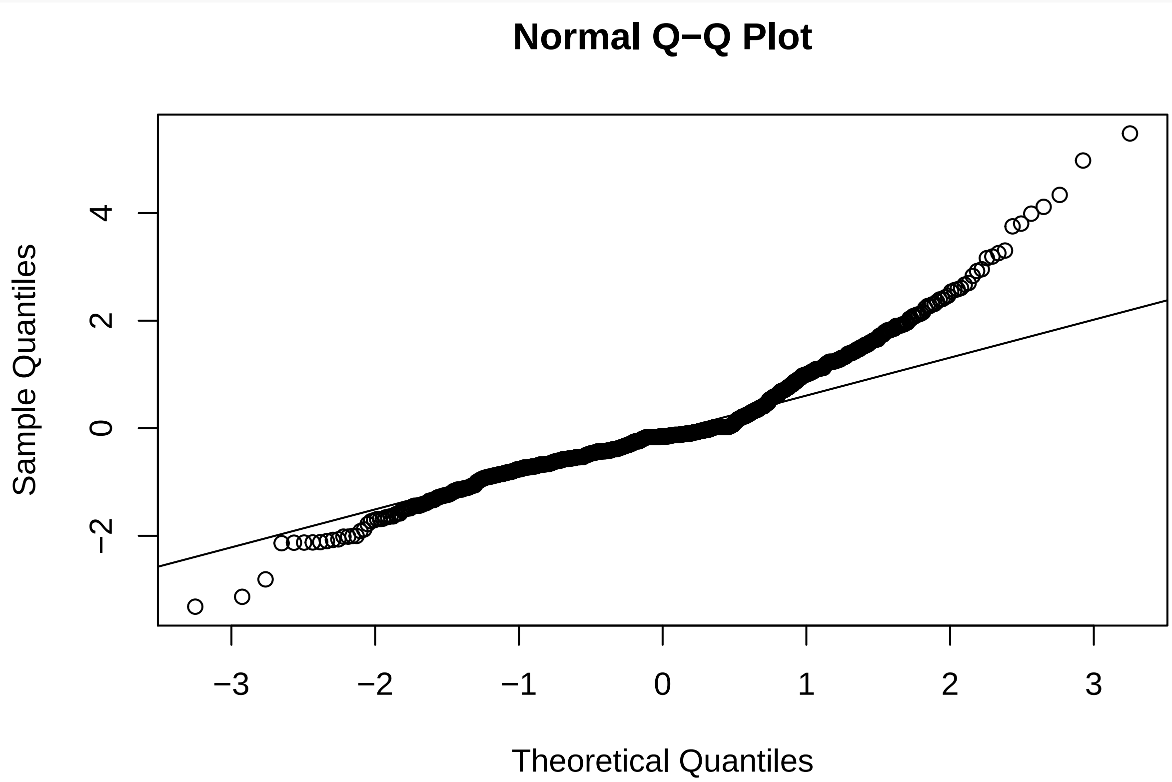

```r
plot(players.data$Salary, players.data$`A1`) #salary vs first assist
```



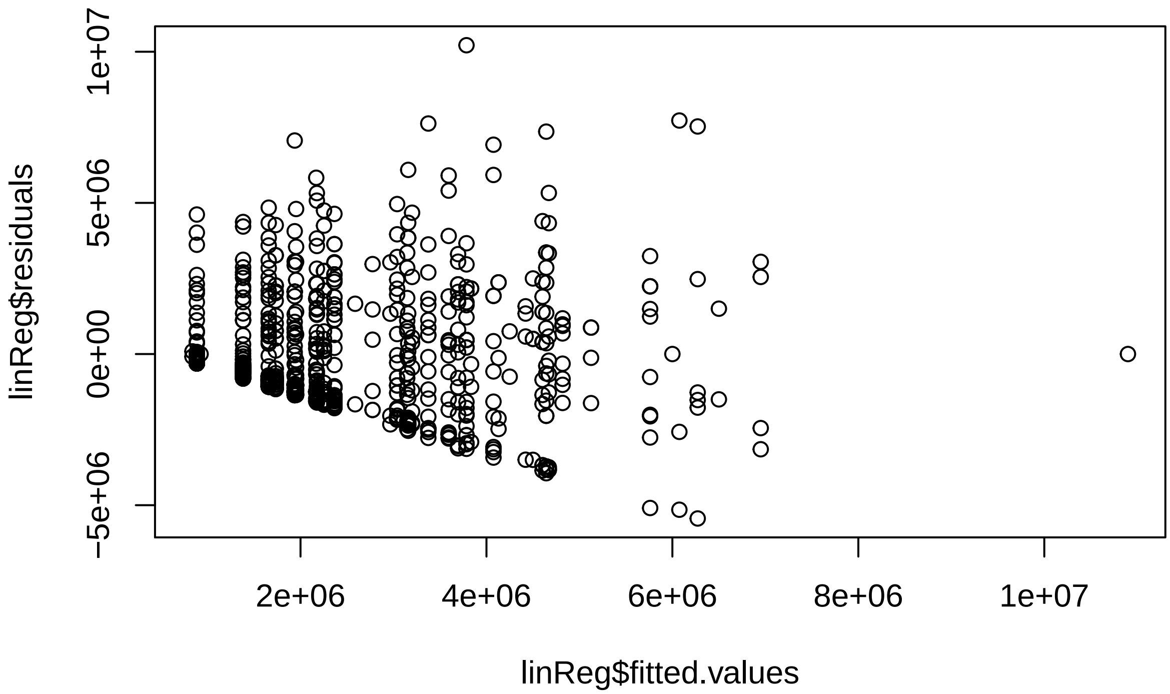

```r
linReg = lm(Salary ~ G , players.data)
#summary(linReg)
#plot(linReg$residuals)
#hist(linReg$residuals)
hist(rstandard(linReg))
```




```r
qqnorm(rstandard(linReg))
qqline(rstandard(linReg))
```



```r
plot(linReg$fitted.values,linReg$residuals) #reziduale je dobro prikazati u ovisnosti o procjenama mode
```

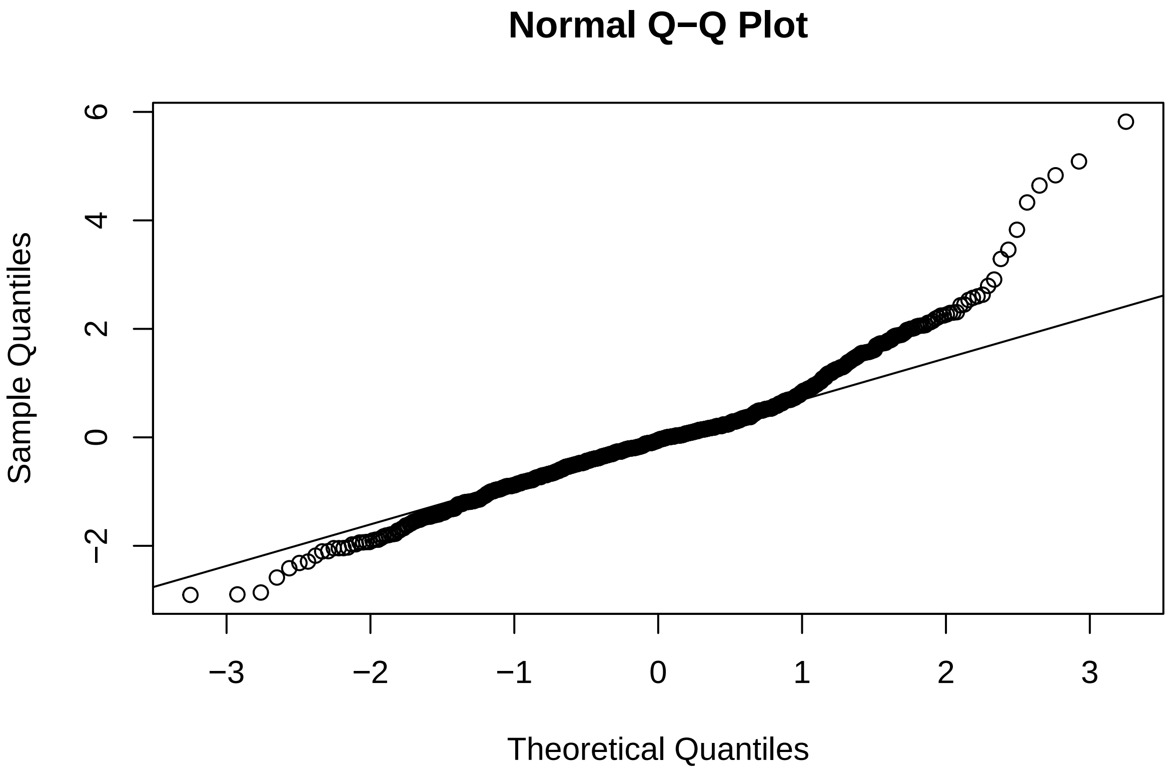



```r
year <- format(as.Date(players.data$Born, format="%Y-%m-%d"), "%Y")


#year
#plot(as.numeric(year),linReg$residuals)
lg = lm(players.data$Salary ~ year + players.data$G)
lg$terms
```

```
## players.data$Salary ~ year + players.data$G
## attr(,"variables")
## list(players.data$Salary, year, players.data$G)
## attr(,"factors")
##                      year players.data$G
## players.data$Salary    0              0
## year                   1              0
## players.data$G         0              1
## attr(,"term.labels")
## [1] "year"            "players.data$G"
## attr(,"order")
## [1] 1 1
## attr(,"intercept")
## [1] 1
## attr(,"response")
## [1] 1
## attr(,".Environment")
## <environment: R_GlobalEnv>
## attr(,"predvars")
## list(players.data$Salary, year, players.data$G)
## attr(,"dataClasses")
## players.data$Salary                year          players.data$G
```

```
##          "character"          "character"          "character"
```

```r
qqnorm(rstandard(lg))
qqline(rstandard(lg))
```




```r
#players.lm = lm(Hand ~ Position, players_copy)
#summary(payers.lm)
```