Mjerenje uspješnosti investitora na financijskim tržištima

Učitavanje datasetova jednog po jednog te dodavanje godine kao stupac u svaki dataset. Spajanje svih datasetova u jedan veliki i micanje praznih redaka nastalih vjerovatno zbog moje pretvore u csv. Nakon toga svi stupci koji u sebi imaju postotak su pretvoreni u double broj. Znak posto je maknut te broj koji se sada nalazi u tom stupcu označava postotak. Tipa ako piše 10 to onda znači 10 posto. Kolumna koja sadrži godine je potom pretovrena u factor, tj kategorijski podatak. Takav dataset je spremljen u file dionice.Rdata te ga se lako može učitati iz memorije kasnije.

```
# dionice.2009 = read.csv("podaci/mojedionice2009.csv", header = T, sep = ",")
# dionice.2010 = read.csv("podaci/mojedionice2010.csv", header = T, sep = ",")
# dionice.2011 = read.csv("podaci/mojedionice2011.csv", header = T, sep = ",")
# dionice.2012 = read.csv("podaci/mojedionice2012.csv", header = T, sep = ",")
# dionice.2013 = read.csv("podaci/mojedionice2013.csv", header = T, sep = ",")
# dionice.2014 = read.csv("podaci/mojedionice2014.csv", header = T, sep = ",")
# dionice.2015 = read.csv("podaci/mojedionice2015.csv", header = T, sep = ",")
# dionice.2016 = read.csv("podaci/mojedionice2016.csv", header = T, sep = ",")
# dionice.2017 = read.csv("podaci/mojedionice2017.csv", header = T, sep = ",")
#
# dionice.2009$Godina = 2009
# dionice.2010$Godina = 2010
# dionice.2011$Godina = 2011
# dionice.2012$Godina = 2012
# dionice.2013$Godina = 2013
# dionice.2014$Godina = 2014
# dionice.2015$Godina = 2015
# dionice.2016$Godina = 2016
# dionice.2017$Godina = 2017

# problem koji ovdje nastaje je taj da dionice iz 2009 imaju drugačiji naziv za otp indeks

# help("colnames")
# colnames(dionice.2010)[7]
# colnames(dionice.2009)[7] = colnames(dionice.2010)[7]
# dionice.2009$Prinos.iznad..OTP.indeksnog.fonda

# dionice = rbind(
#   dionice.2009,
#   dionice.2010,
#   dionice.2011,
#   dionice.2012,
#   dionice.2013,
#   dionice.2014,
#   dionice.2015,
#   dionice.2016,
#   dionice.2017
# )

#
# dionice = na.omit(dionice)
# dionice
```

```
#
# my_function("5.34%")
# my_function("5,34%")
# my_function = function(string) {
#   s = gsub("%", "", string)
#   s = gsub(",", ".", s)
#   return(as.numeric(s))
# }
#
# dionice$Prinos.iznad..OTP.indeksnog.fonda = unlist(lapply(dionice$Prinos.iznad..OTP.indeksnog.fonda,
# dionice$Prinos = unlist(lapply(dionice$Prinos, my_function))
# dionice$Prinos.bez.div. = unlist(lapply(dionice$Prinos.bez.div., my_function))
#
# save(dionice, file = "dionice.Rdata")
# load("dionice.Rdata")
#
# head(dionice)
#
# class(dionice$Prinos.iznad..OTP.indeksnog.fonda)
# class(dionice$Godina)
#
# levels(dionice$Godina)
# head(dionice)
# dim(dionice)
# names(dionice)
#
# dionice$Godina = as.factor(dionice$Godina)
# levels(dionice$Godina)

# save(dionice, file = "dionice.Rdata")
load("dionice.Rdata")
```

Deskriptivna statistika.

Proučavanje dobivenog dataseta.

Proučavanje po godinama.

Summary po svim godinama. Ne govori nam previše te je dosta teško usporedit podatke.

```
tapply(dionice$Prinos.iznad..OTP.indeksnog.fonda, dionice$Godina, summary)
```

```
## $`2009`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -43.2600  -7.5650  -0.4800  -0.3647   8.4950  26.8100
##
## $`2010`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -58.980 -18.660  -8.910  -8.082   0.950  62.950
##
## $`2011`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -35.26   -9.77   -3.46   -1.98    5.55   42.71
##
```

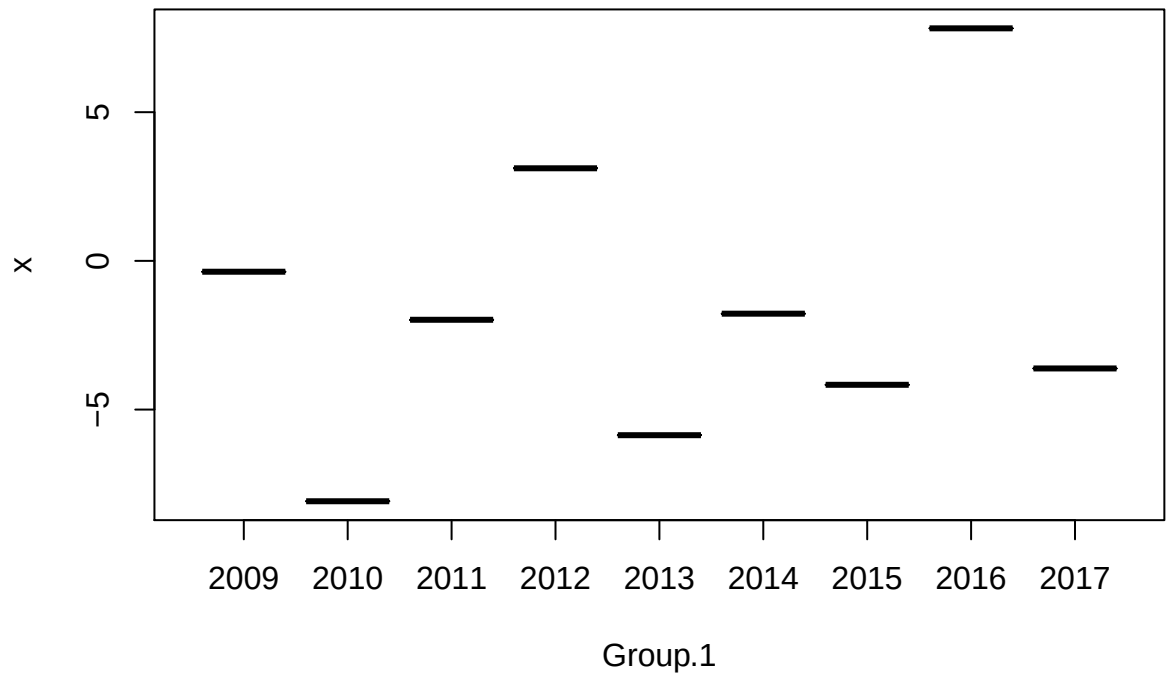
```
## $`2012`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -63.700  -9.020   5.630   3.113  17.230  80.330
##
## $`2013`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -59.690 -16.005  -3.920  -5.862   4.475 131.780
##
## $`2014`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -50.890 -16.670  -3.870  -1.776   9.900  88.830
##
## $`2015`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -49.590 -13.025  -2.870  -4.167   6.425  30.230
##
## $`2016`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -27.890  -1.575   4.940   7.818  16.328  64.080
##
## $`2017`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -84.540 -15.640  -0.590  -3.617  12.325  61.260
```

Graf srednje vrijednosti kroz godine. Većina srednjih vrijednosti se nalazi u blagome minusu uz izuzetak 2016 kada je srednja vrijednost na skoro 8 posto u plusu. 2010 godina je pak u minusu od 8%.

```
mean.by.year = aggregate(dionice[, c(7)], list(dionice$Godina), mean)
mean.by.year
```

```
##   Group.1          x
## 1    2009 -0.3647059
## 2    2010 -8.0820354
## 3    2011 -1.9804580
## 4    2012  3.1134562
## 5    2013 -5.8617871
## 6    2014 -1.7757544
## 7    2015 -4.1666776
## 8    2016  7.8183978
## 9    2017 -3.6173409
```

```
plot(mean.by.year)
```

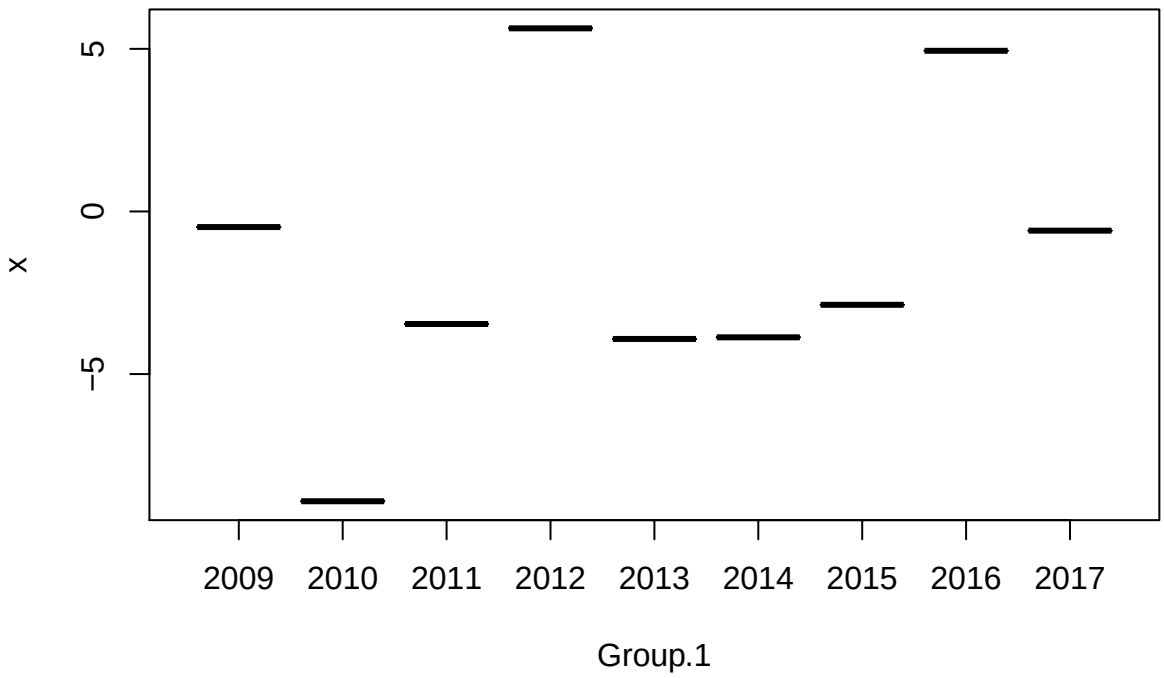


Graf mdeijana kroz godine. Većina medijana se isto nalazi u blagome minusu no zanimljivo je da najveći medijan se ovaj put nalazi u 2012 godini. Medijan za 2012 iznosi 5.63% dok srednja vrijednost za 2012 3.1%. 2010 godina je i dalje u velikom minusu i to skoro za 9% ovaj puta.

```
median.by.year = aggregate(dionice[, c(7)], list(dionice$Godina), median)
median.by.year
```

```
##   Group.1     x
## 1    2009 -0.48
## 2    2010 -8.91
## 3    2011 -3.46
## 4    2012  5.63
## 5    2013 -3.92
## 6    2014 -3.87
## 7    2015 -2.87
## 8    2016  4.94
## 9    2017 -0.59
```

```
plot(median.by.year)
```

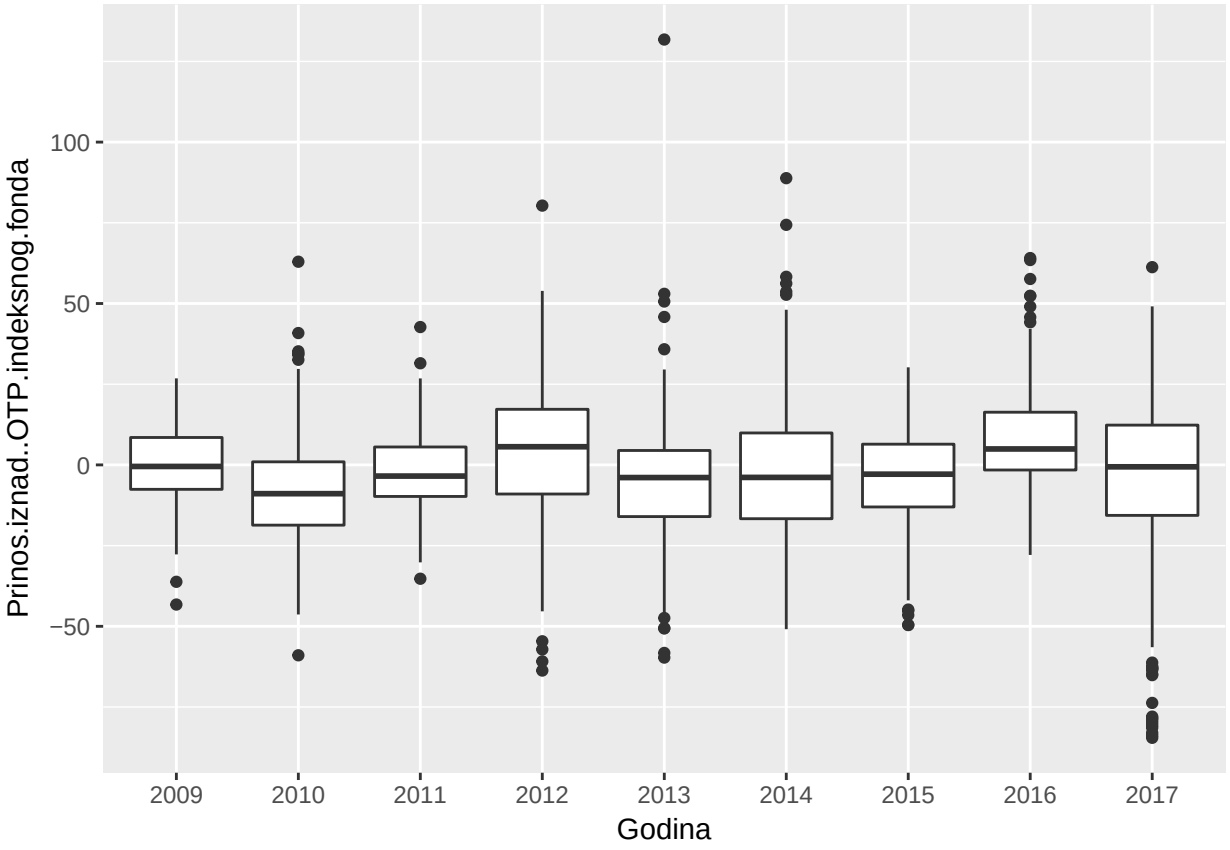


Boxplotovi za svaku godinu pokazani na istom grafu. Vidimo da su svi negdje oko nule te počinjemo sumnjati u hipotezu da je tržište moguće pobijediti. Primijećujemo stršeću pozitivnu vrijednost u 2013 godini. Neko je tad naime bio 131% u plusu. Skrećemo pažnju na donji box u 2012 godini koji je osjetno duđi od gornjeg boxa te zato je takva negativna vrijednost te godine. Paralelno s tim možemo primijetit da je 2016 gornji box osjetno veći od donjeg boxa.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

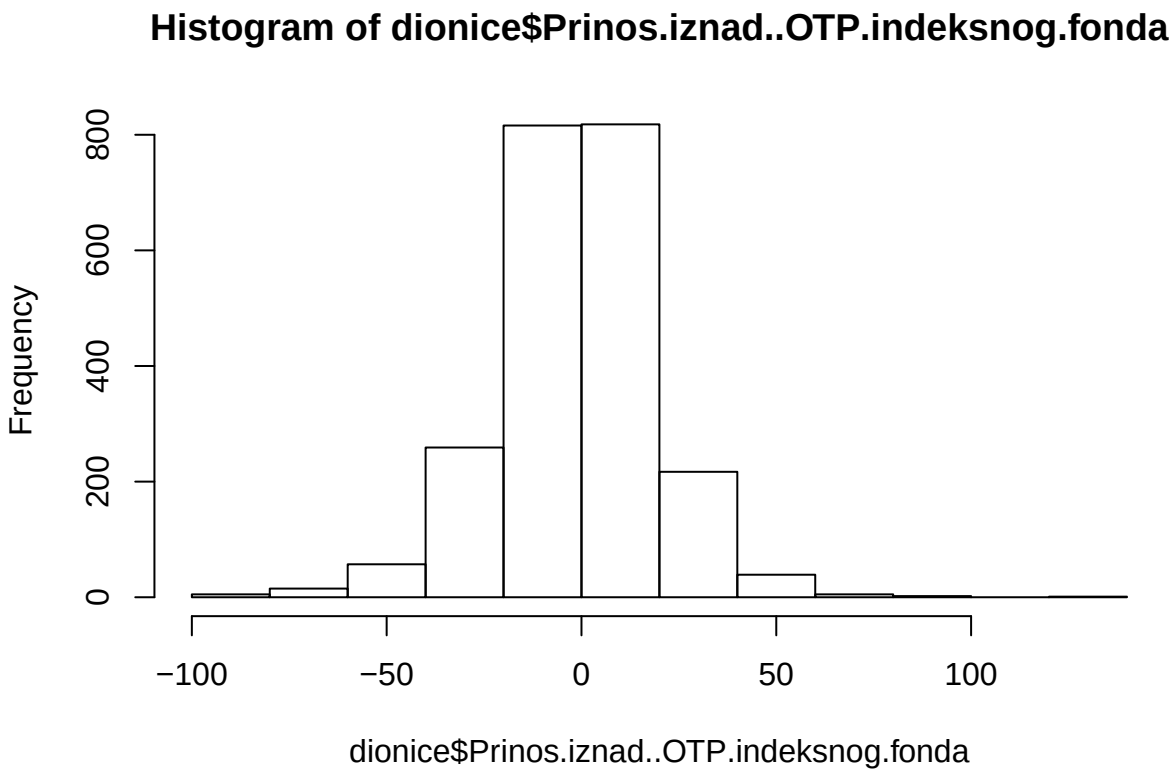
```
ggplot(dionice, aes(x=Godina, y=Prinos.iznad..OTP.indeksnog.fonda)) + geom_boxplot(aes(fill=Prinos.izna
```



Sve godine gledane zajedno.

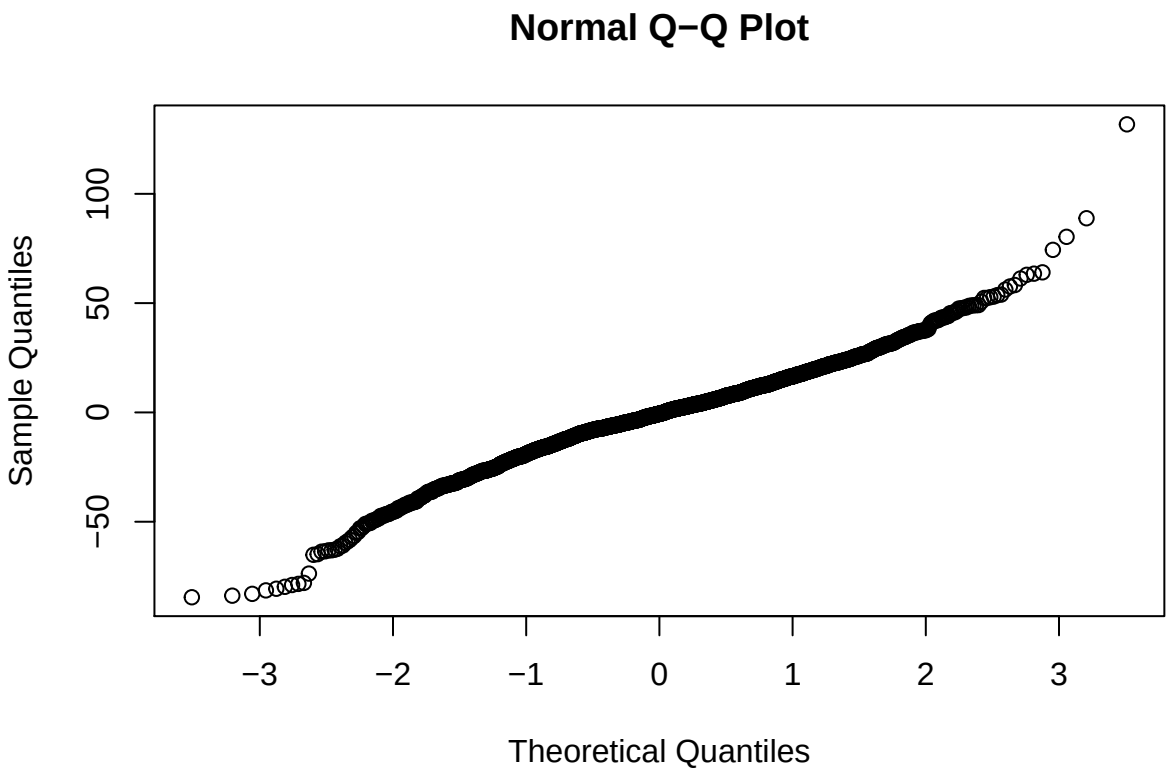
Možemo pretpostaviti na temelju histograma da se stvarno radi o normalnoj distribuciji.

```
hist(dionice$Prinos.iznad..OTP.indeksnog.fonda)
```



QQplot je dosta blizu pravca.

```
qqnorm(dionice$Prinos.iznad..OTP.indeksnog.fonda)
```



Najlošiji postotak je onaj od -84% dok je najbolji od plus 131%. Medijan i srednja vrijednost su jako blizu nule.

```
summary(dionice$Prinos.iznad..OTP.indeksnog.fonda)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -84.540 -11.750  -0.605  -1.171  10.637 131.780
```

Interkvartilni rang iznosi 22 te vidimo gore da se nalazi između -10 i plus 10. Očekivano je da se 50 posto podataka nalazi oko nule uz ovakve medijane i srednje vrijednosti. Varijanca je 403. Što vjerovatno znači da podaci dosta variraju.

```
otp = dionice$Prinos.iznad..OTP.indeksnog.fonda
IQR(otp); var(otp)
```

```
## [1] 22.3875
```

```
## [1] 403.0616
```

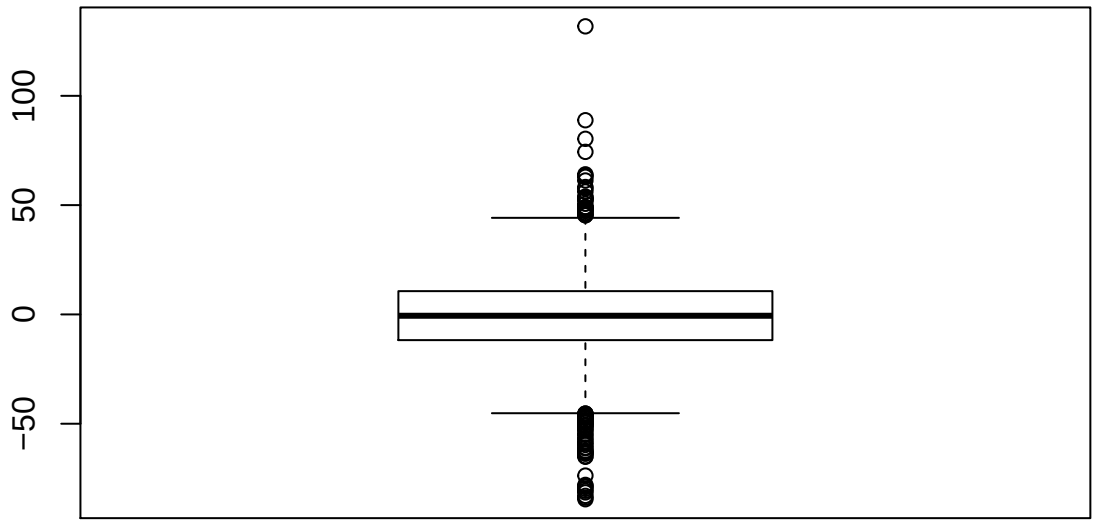
Ako probamo izračunat mean ali maknemo 20 posto najmanjih i 20 posto najvećih srednja vrijednost je i dalje oko nule. Znači da stršeće vrijednosti definitivno ne utječu previše.

```
mean(otp, trim=0.2)
```

```
## [1] -0.5620864
```

Boxplot

```
boxplot(otp)
```



Testiranje srednje vrijednosti otp-a

Testiramo null hipotezu da je srednja vrijednost jendaka nuli, s alternativom da nije jednaka nuli. Alpha je 5% a p vrijednost je 0.58% što znači da možemo odbaciti null hipotezu. Gledajući 95%-tni interval pouzdanosti koji ide od [-2.0, -0.33] možemo vidjet da nula definitvno nije unutra te da je srednja vrijednost otp-a populacije negativna. Možemo zaključiti kako pojedinci generalno gube a tržište pobjeđuje.

```
t.test(otp, mu=0, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  otp
## t = -2.7571, df = 2233, p-value = 0.005879
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.0040814 -0.3381478
## sample estimates:
## mean of x
## -1.171115
```

Ide li ljudima sve bolje i bolje. Po ovom dolje izgleda da ide pošto možemo odbaciti null hipotezu. Ili sam ja nešto opako grdo napravio jer mean 2017 je manji od meana 2009.

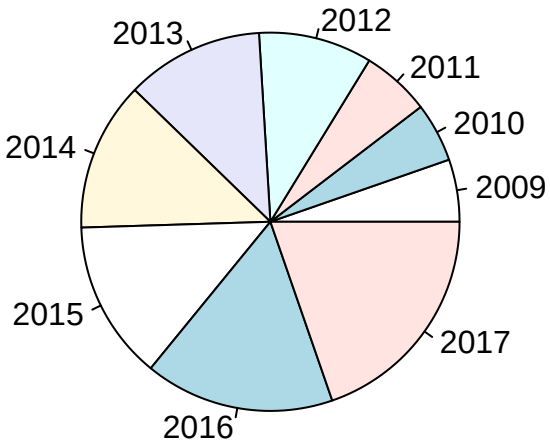
```
otp_2017 = dionice[dionice$Godina == 2017, ]$Prinos.iznad..OTP.indeksnog.fonda
otp_2009 = dionice[dionice$Godina == 2009, ]$Prinos.iznad..OTP.indeksnog.fonda
```

```r
t.test(otp_2009, otp_2017, alt="less")
```

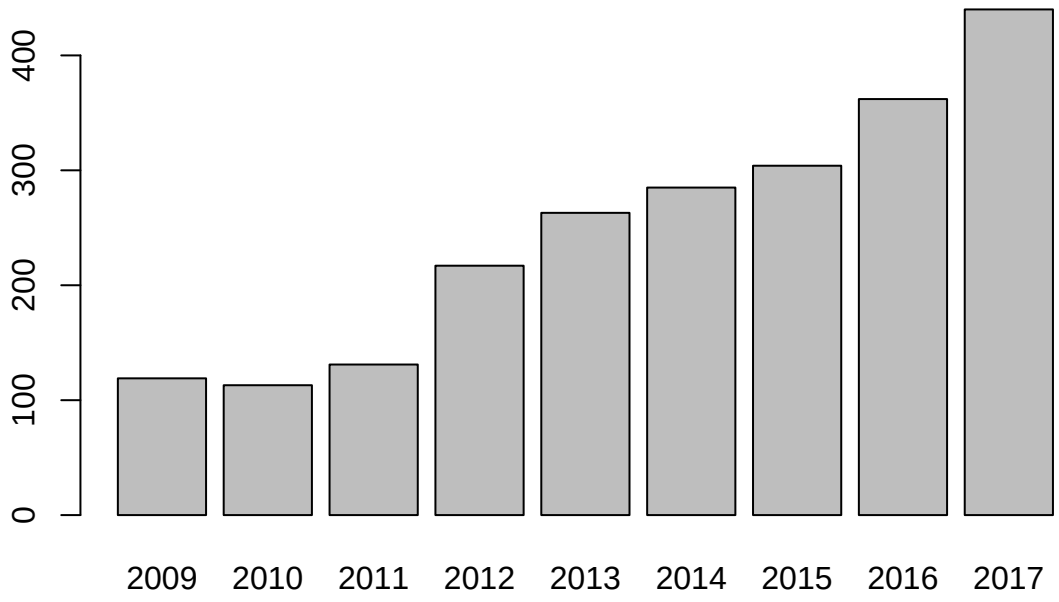
```
##
##  Welch Two Sample t-test
##
## data:  otp_2009 and otp_2017
## t = 1.9301, df = 366.48, p-value = 0.9728
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf 6.03153
## sample estimates:
##  mean of x  mean of y
## -0.3647059 -3.6173409
```

FUN FACT: Broj prijavljenih ekipa raste tokom godina

```r
#pie chart
pie(table(dionice$Godina))
```




```r
#barplot
barplot(table(dionice$Godina))
```



Broj promjena portfelja

Parsiranje stupca s dionicama kako bi izvukli broj promjena tokom natjecanja. Taj stupac je potom dodan kao stupac Broj.promjena

```
parser = function(string) {
  s = strsplit(as.character(string), "\\s+")[[1]][1]
  s = gsub("[(]", "", s)
  s = gsub("[)]", "", s)
  return(as.numeric(s))
}
dionice$Broj.promjena = dionice$Dionice
dionice$Broj.promjena = unlist(lapply(dionice$Broj.promjena, parser))

save(dionice, file = "dionice.Rdata")
```

Korelacija između stupca broj promjena i otp postotka je nažalost jako malena(pozitivna) te ne možemo tvrditi da broj promjena povećava šanse za uspjeh.

```
cor(dionice$Prinos.iznad..OTP.indeksnog.fonda, dionice$Broj.promjena);
```

```
## [1] 0.1022367
```

```
# cor(dionice$Prinos.iznad..OTP.indeksnog.fonda, dionice$Broj.promjena, method = "kendall")
# cor(dionice$Prinos.iznad..OTP.indeksnog.fonda, dionice$Broj.promjena, method = "spearman")
```

Test usporedbe srednje vrijednosti onih s 0 promjena i onih s >=1 promjenom. Alternativna hipoteza je ta da onima s više promjena ide bolje. ### Radim nešto krivo kod ovog testa s 2 uzorka.

```
nula_promjena = dionice[dionice$Broj.promjena > 0, ]$Prinos.iznad..OTP.indeksnog.fonda
više_od_nula_promjena = dionice[dionice$Broj.promjena == 0, ]$Prinos.iznad..OTP.indeksnog.fonda
length(nula_promjena) + length(više_od_nula_promjena) == length(otp)
```

```
## [1] TRUE
```

```
t.test(nula_promjena, više_od_nula_promjena, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  nula_promjena and više_od_nula_promjena
## t = 2.4091, df = 1460.9, p-value = 0.9919
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 3.766639
## sample estimates:
##  mean of x  mean of y
##  0.2162309 -2.0215523
```

Ekipe koje se ponavljaju tokom godina

Nalazimo one čije se ime barem 2 puta ponovilo.

```
name.occurence = data.frame(table(dionice$Naziv.portfelja))
more.than.2.times = name.occurence[name.occurence$Freq > 1, ]
```

U 2234 podataka ima samo 1347 jedinstevnih igrača te od tog ima 430 igrača koji su bili barem dva puta.

```
length(name.occurence$Var1); length(more.than.2.times$Var1)
```

```
## [1] 1347
```

```
## [1] 430
```

Donosi li iskustvo prednost?

Za one koji imaju više sudjelovanja, sveli smo im otp na srednju vrijednost svih pokušaja. Razdvojimo dataset na starosjedioce i guštere.

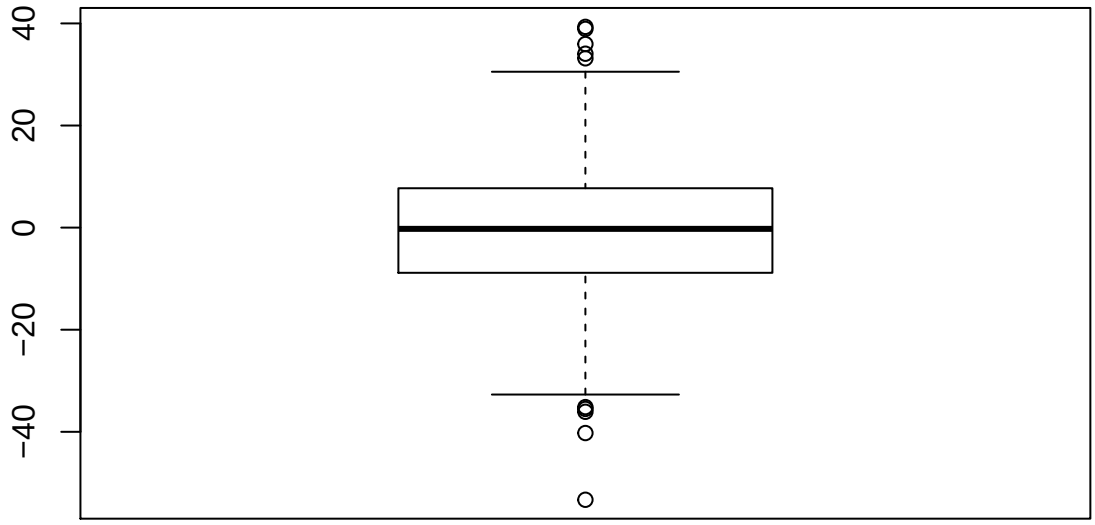
```
spojeni_otp = aggregate(dionice$Prinos.iznad..OTP.indeksnog.fonda, by=list(dionice$Naziv.portfelja), FU
colnames(spojeni_otp) = c("ID", "otp")

gusteri = spojeni_otp[!(spojeni_otp$ID %in% more.than.2.times$Var1), ]
starosjedioci = spojeni_otp[spojeni_otp$ID %in% more.than.2.times$Var1, ]
length(gusteri$ID); length(starosjedioci$ID)
```

```
## [1] 916
```

```
## [1] 430
```

Idemo vidjet malo proučit starosjedioce. Izgleda da nisu ništa posebno. Srednja vrijednost i medijan blago negativni opet.

```
boxplot(starosjedioci$otp)
```



Summary

```
summary(starosjedioci)
```

```
##          ID           otp
##  ac_dc    :  1   Min.   :-53.310
##  ALL IN   :  1   1st Qu.: -8.841
##  Barbus   :  1   Median : -0.245
##  batistuta:  1   Mean   : -1.038
##  bravo    :  1   3rd Qu.:  7.697
##  bujto    :  1   Max.   : 39.325
##  (Other)  :424
```

```
# t test na arrayu od mediana
# plotaj tocke koji na x osi imaju broj promjena
# starosjedioci with best of or last otp
# mijenjaju li starosjedioci češće portfelj
# korelacija između broja dolaska i agregated meana

# together = merge(agregated.otp.more.than.2.times, more.than.2.times,by="ID")
# together[with(together, order(-x)), ]
# save(together, file="best_teams.Rdata")
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click

the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.