---

title: “Projekt” author: “Sap-projekt” date: ‘17 01 2021’ output: pdf_document —

Analiza uspješnosti dioničkih fondova:

Opis projekta: Ovaj projekt obavezni je dio izbornog kolegija Statistička analiza podataka Fakulteta elektrotehnike i računarstva. Projekt je poslužio primjeni teorijskih temelja stečenih na predavanjima na skup podataka iz stvarnog svijeta. Kao pomoć u izradi projekta poslužio je programski jezik R koji je pružio potporu za izvođenje testiranja i bolju vizualizaciju podataka. Skup podataka: Korišteni skup podataka sastoji se od velikog broja dioničkih fondova koji su dostupni američkim investitorima i izraženi su u američkim dolarima. U daljnjem tekstu često će se pojavljivati izraz “uspješnost fonda”. Kao uspješnost fonda korišten je srednji povrat fonda u razdoblju od deset godina. Iako je uspješnost moguće definirati na više načina, ovaj je odabran kao standardan način prilikom početka rada na projektu te je kao takav zadržan.

Deskriptivna statistika skupa podataka:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

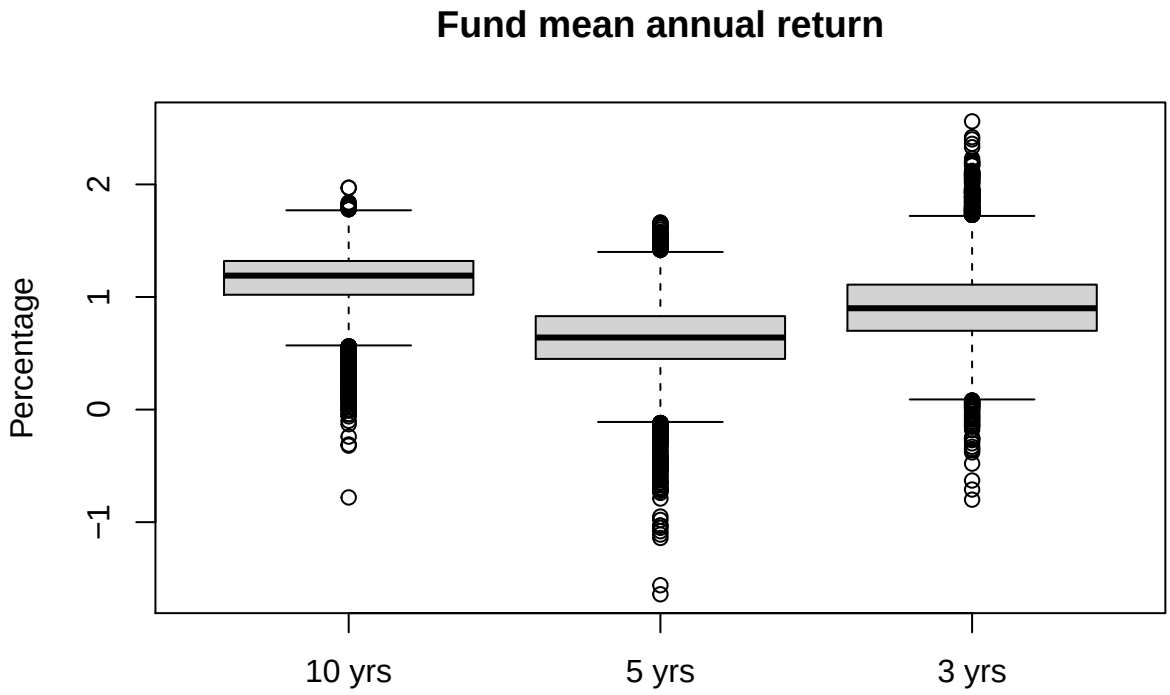
```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
data = read.csv("mutual_funds.csv")
#data = na.omit(data)
```

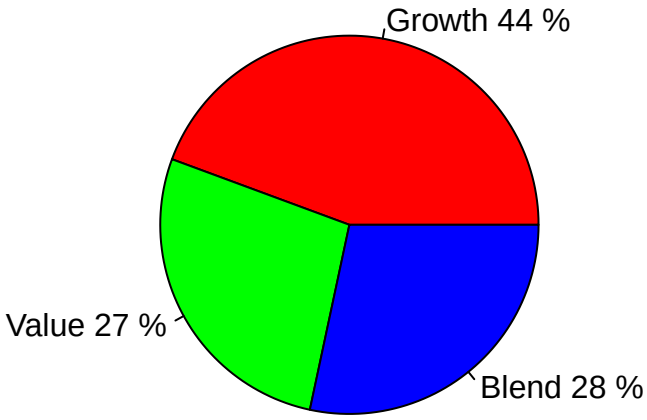
Za glavni kriterij uspješnosti određenog fonda uzimamo srednji godišnji povrat za zadnjih 10 godina. Premda da imamo podatke i o srednjem prošlogodišnjem povratu te srednjem povratu za zadnjih 3 i 5 godina, srednji povrat za 10 godina nam najbolje pokazuje koliko je određeni fond pouzdan/dobar/uspješan jer nam daje podatke o najdužem periodu. Na sljedećem pravokutnom dijagramu prikazani su srednji povrat za 3, 5 i 10 godina. (još malo nadopunit....)

```
success10 = data$fund_mean_annual_return_10years
success5 = data$fund_mean_annual_return_5years
success3 = data$fund_mean_annual_return_3years
boxplot(success10, success5, success3, main="Fund mean annual return", names=c("10 yrs", "5 yrs", "3 yr
```



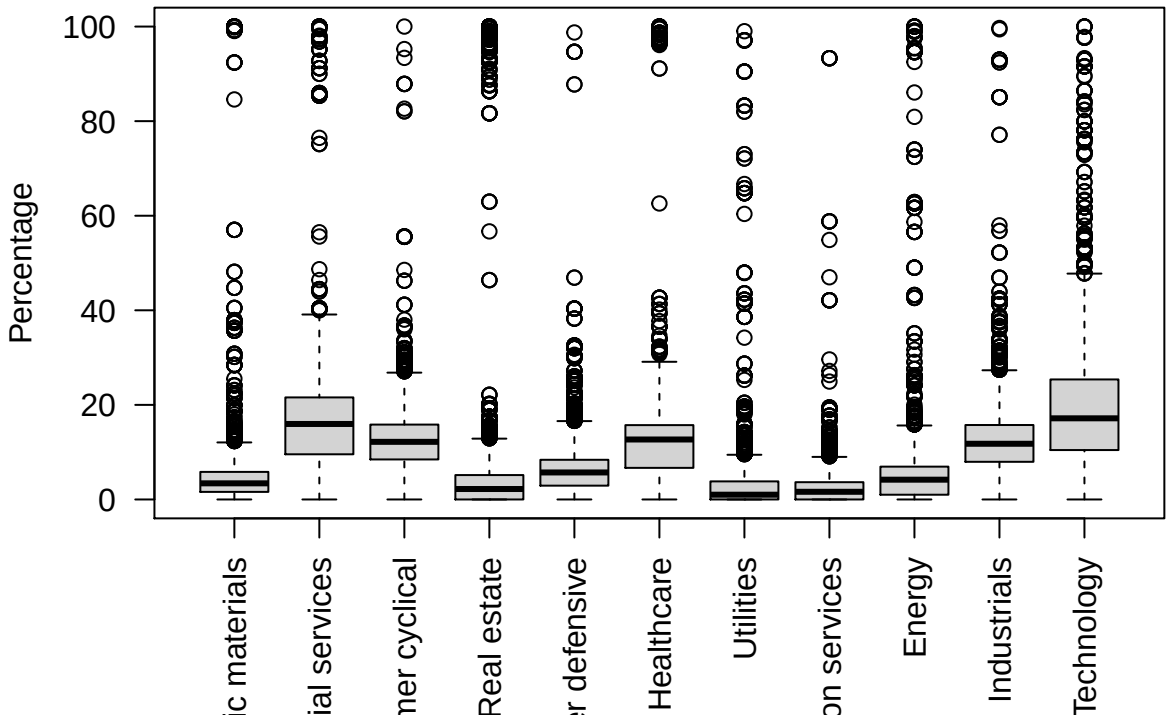
Sljedeći graf prikazuje udjele određenog stila investiranja, u podatcima koje koristimo fondovi imaju 3 različita stila investiranja; Growth, Value, Blend. (mozda u kratko objasnit sta koji znaci)

```
investment = data$investment[data$investment != "<undefined>"]
growth.number = sum(investment == "Growth")
value.number = sum(investment == "Value")
blend.number = sum(investment == "Blend")
values = c(growth.number, value.number, blend.number)
labels = c("Growth", "Value", "Blend")
pct = round(values/sum(values)*100)
labels = paste(labels, pct)
labels = paste(labels,"%")
pie(values, labels=labels, col=rainbow(length(labels)))
```

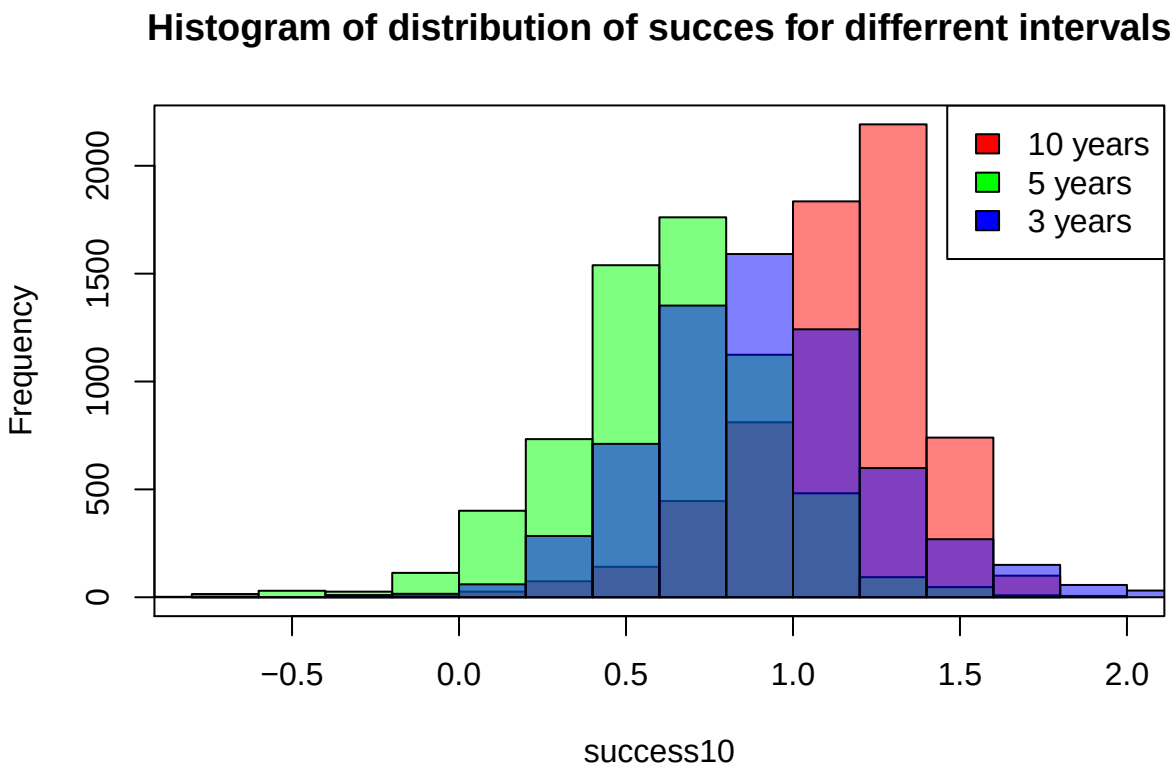



```
materials = data$basic_materials
financial = data$financial_services
cyclical = data$consumer_cyclical
estate = data$real_estate
defensive = data$consumer_defensive
healthcare = data$healthcare
utilities = data$utilities
communication = data$communication_services
energy = data$energy
industrials = data$industrials
technology = data$technology
labels = c("Basic materials",
           "Financial services",
           "Consumer cyclical",
           "Real estate",
           "Consumer defensive",
           "Healthcare",
           "Utilities",
           "Communication services",
           "Energy",
           "Industrials",
           "Technology")
boxplot(materials,
        financial,
        cyclical,
        estate,
```

```
        defensive,
        healthcare,
        utilities,
        communication,
        energy,
        industrials,
        technology,
        names=labels,
        ylab="Percentage",
        las = 2)
```



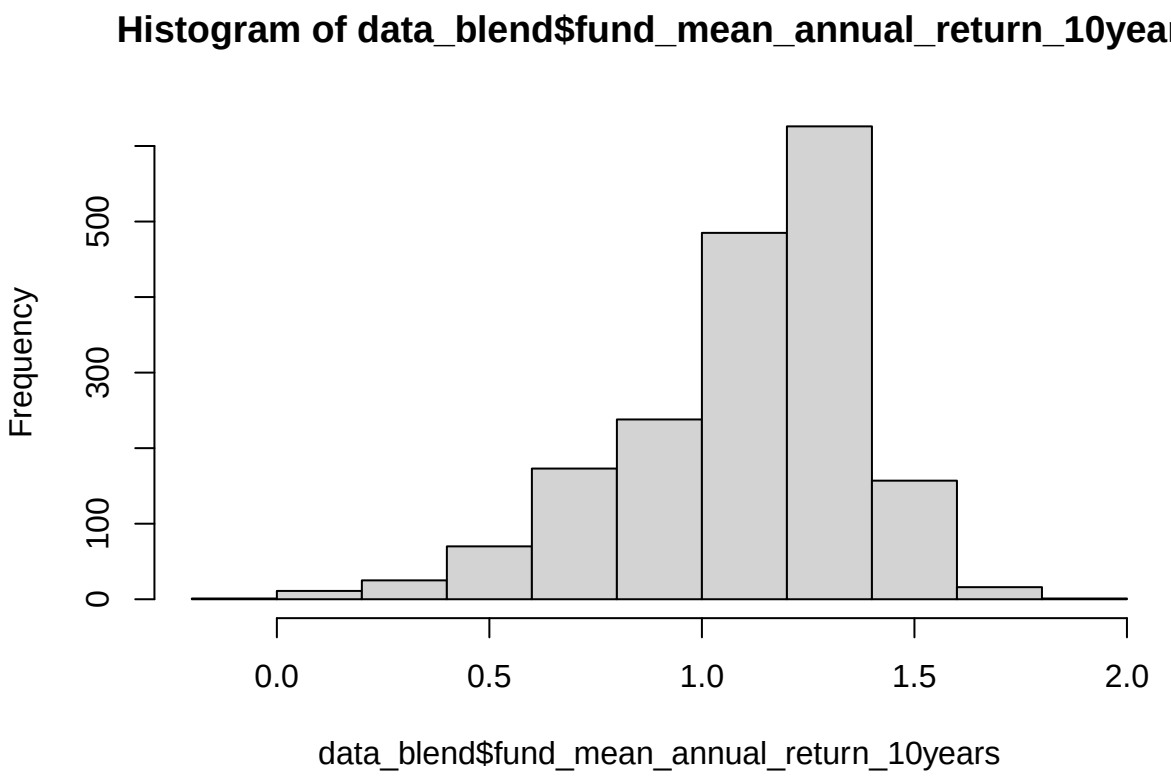

```
hist(success10, col=rgb(1,0,0,0.5), main="Histogram of distribution of succes for differrent intervals")
hist(success5, col=rgb(0,1,0,0.5),add=T)
hist(success3, col=rgb(0,0,1,0.5),add=T)
legend("topright", c("10 years", "5 years", "3 years"), fill=c("red", "green", "blue"))
box()
```



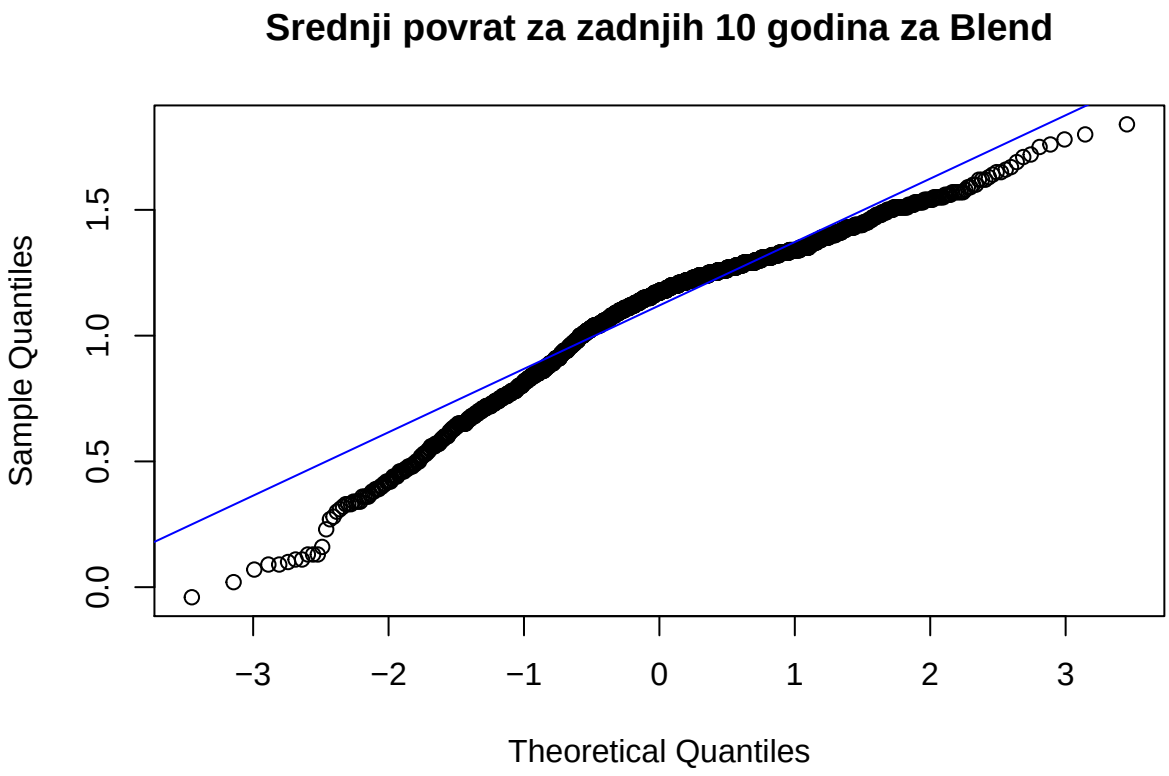
Statističko zaključivanje(mozda neki drugi naslov):

Kao što je već spomenuto svaki fond ima određeni stil investiranja (Growth, Blend, Value). Zanima nas razlikuju li se uspješnosti fondova s obzirom na stil investiranja koji odabiru, odnosno želimo provijeriti imaju li fodnovi s određenim stilom investiranja veće povrate nego ostali. Za početak želimo vidjeti ravnaju li se povrati svake od tih kategorija po normalnoj razdiobi kako bi mogli primjeniti anovu, test o jednakosti sredina.

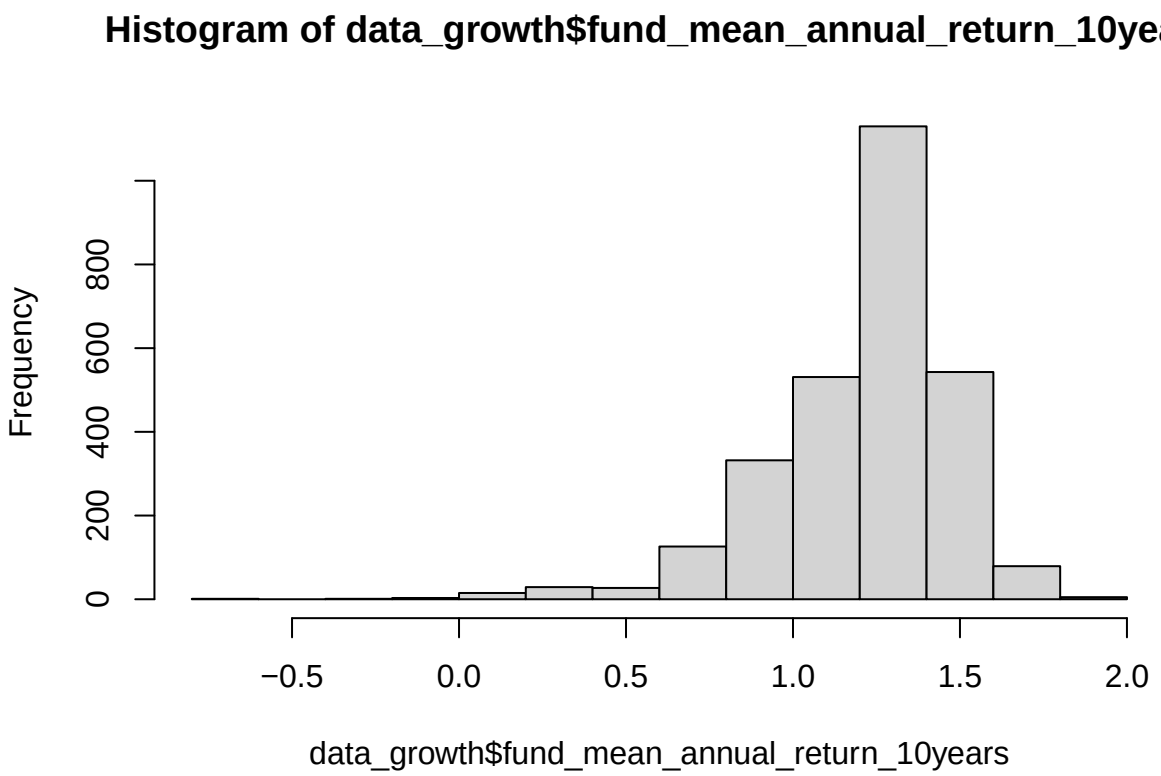
```
data_blend <- data[data$investment == c("Blend"),]
hist(data_blend$fund_mean_annual_return_10years)
```



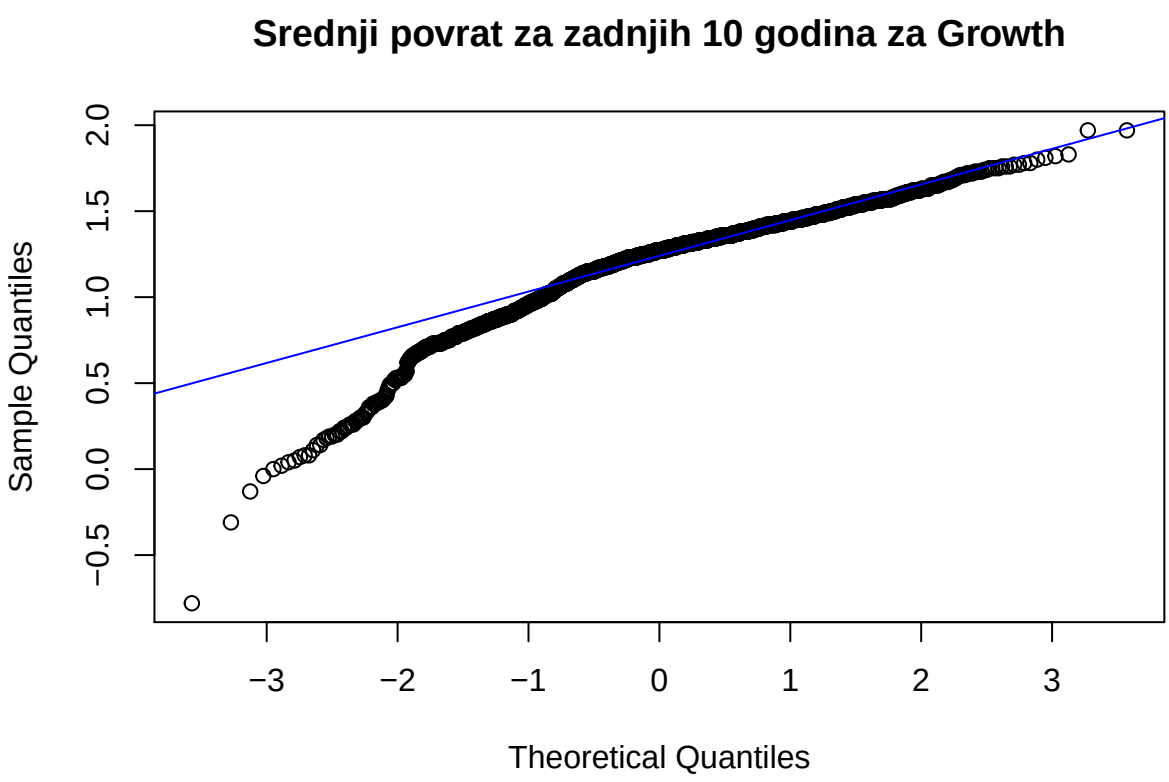

```
qqnorm(data_blend$fund_mean_annual_return_10years, main="Srednji povrat za zadnjih 10 godina za Blend")
qqline(data_blend$fund_mean_annual_return_10years, col="blue")
```



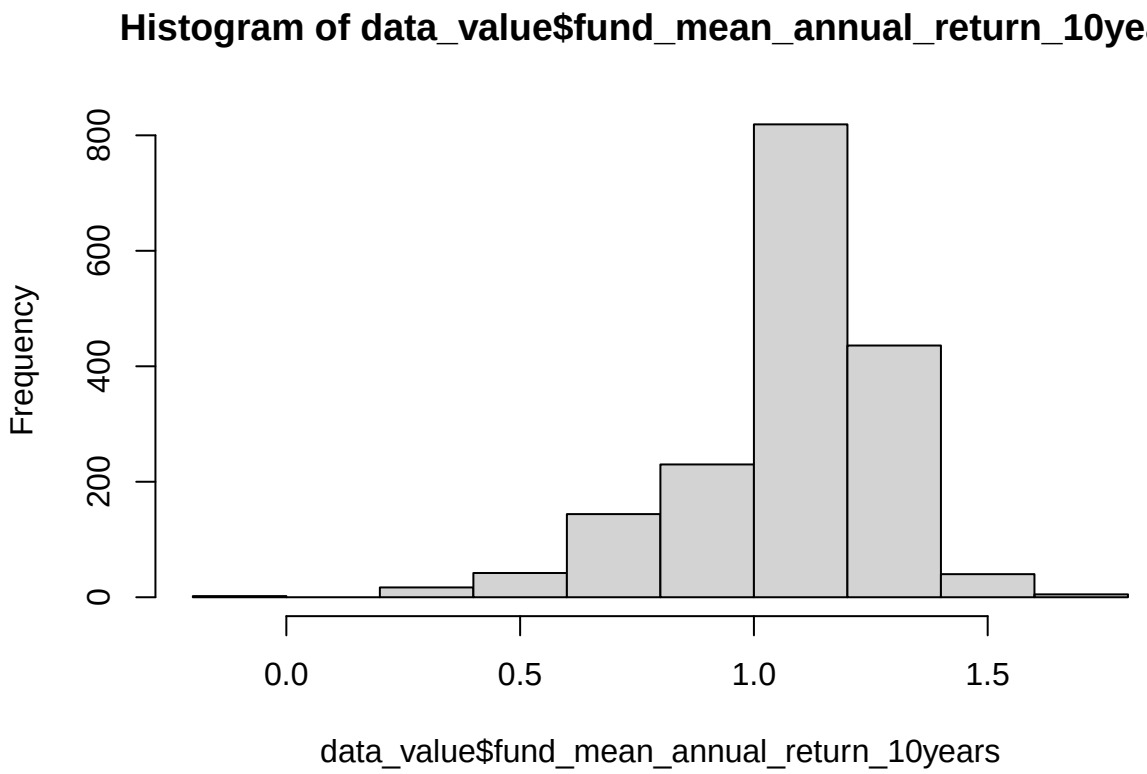

```
#ks.test(data_blend$fund_mean_annual_return_10years, "pnorm")

data_growth <- data[data$investment == c("Growth"),]
hist(data_growth$fund_mean_annual_return_10years)
```



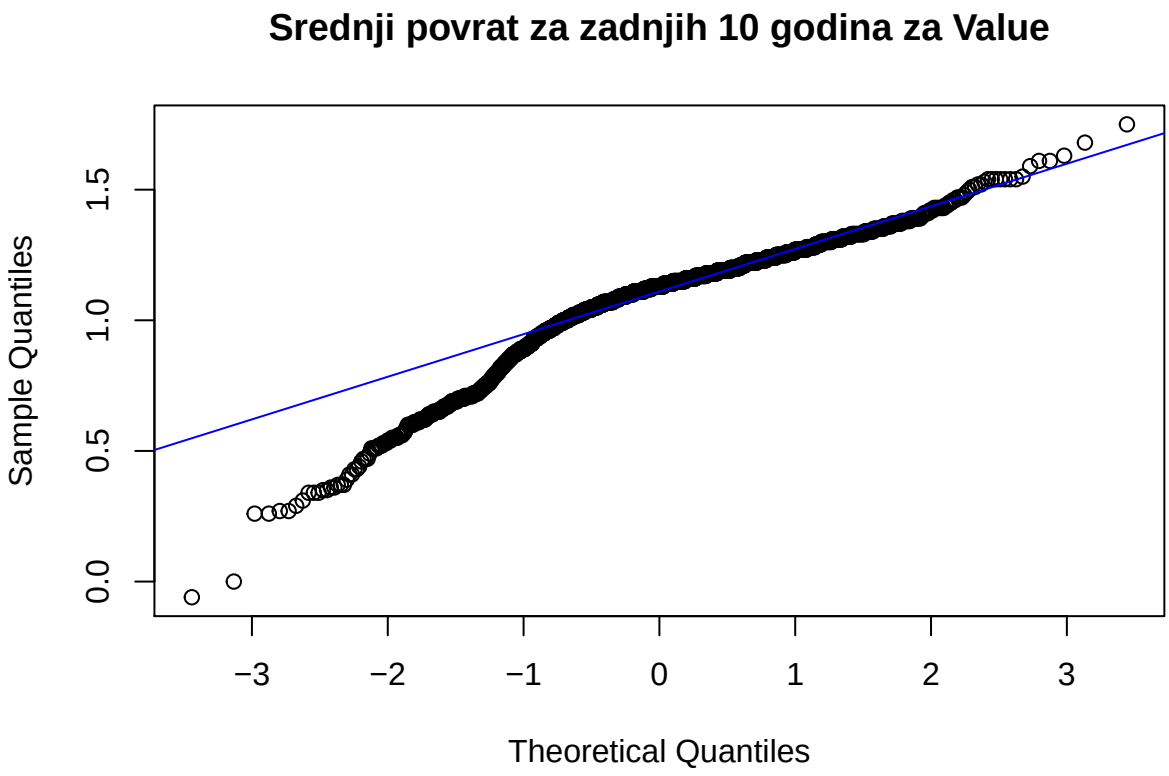

```
qqnorm(data_growth$fund_mean_annual_return_10years, main="Srednji povrat za zadnjih 10 godina za Growth
qqline(data_growth$fund_mean_annual_return_10years, col="blue")
```




```
#ks.test(data_growth$fund_mean_annual_return_10years, "pnorm")


data_value <- data[data$investment == c("Value"),]
hist(data_value$fund_mean_annual_return_10years)
```




```
qqnorm(data_value$fund_mean_annual_return_10years, main="Srednji povrat za zadnjih 10 godina za Value")
qqline(data_value$fund_mean_annual_return_10years, col="blue")
```




```
#ks.test(data_value$fund_mean_annual_return_10years, "pnorm")
```

Iz grafova možemo zaključiti da distribucija ne odstupa drastično od normalne. Uz pretpostavku normalnosti i nezavisnosti podataka TE JEDNAKOSTI VARIJANCI !! možemo provesti anova test.(sta radit s jednakosti varijaci???? bartlet vraca jako malu p vrijednost)

```
data_growth_filtered <- na.omit(data_growth$fund_mean_annual_return_10years)
data_blend_filtered <- na.omit(data_blend$fund_mean_annual_return_10years)
data_value_filtered <- na.omit(data_value$fund_mean_annual_return_10years)

var(data_growth_filtered)
```

```
## [1] 0.07227282
```

```
var(data_blend_filtered)
```

```
## [1] 0.07808235
```

```
var(data_value_filtered)
```

```
## [1] 0.04714453
```

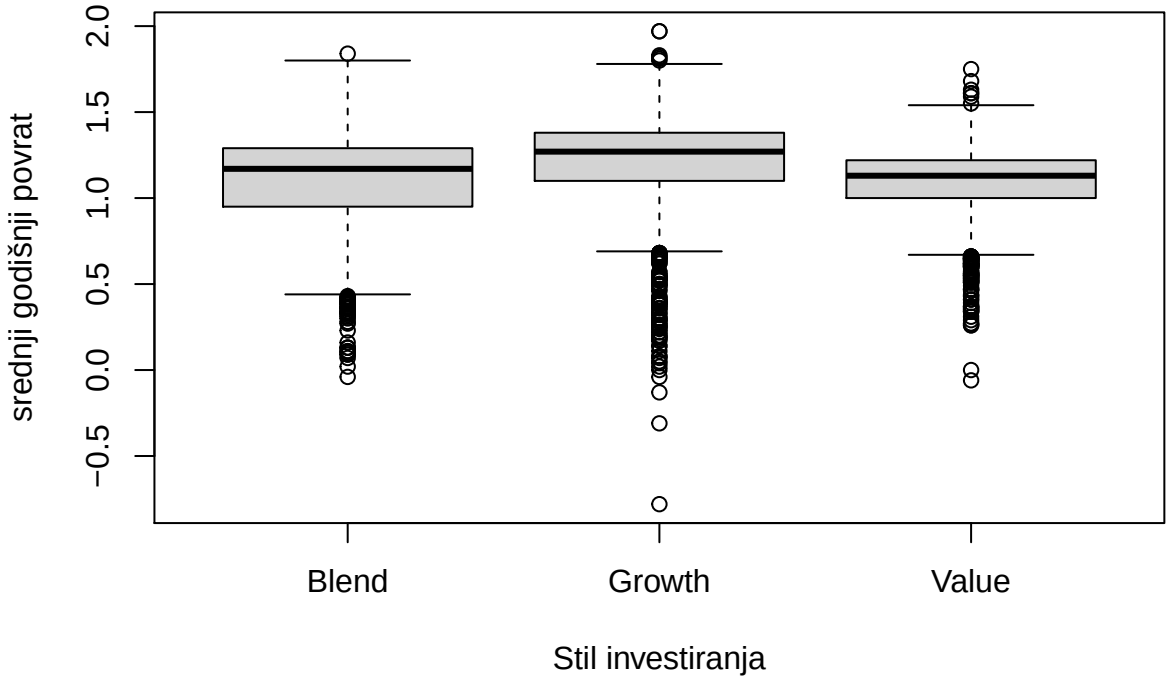
prikažimo box plot za te tri kategorije. Vidimo da su sredine prilično slične dakle možemo se pitat jesu li iste. Budući da se ravnaju po normalnoj te uz pretpostavku nezavisnosti podataka i jednakosti varijanci možemo provesti anova test. Pretpostavka H0 je da su sredine te 3 kategorije jednake.

```
bartlett.test(data$fund_mean_annual_return_10years ~ data$investment)##dal ga provodit??
```

```
##
##  Bartlett test of homogeneity of variances
```

```
##
## data:  data$fund_mean_annual_return_10years by data$investment
## Bartlett's K-squared = 141.49, df = 3, p-value < 2.2e-16
```

```
boxplot(data$fund_mean_annual_return_10years[data$investment != "<undefined>"] ~ data$investment[data$in
        ylab= "srednji godišnji povrat",
        xlab= "Stil investiranja")
```




```
res.aov <- aov(fund_mean_annual_return_10years ~ factor(investment), data = data)
summary(res.aov)
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## factor(investment)    3   30.0  10.013   148.4 <2e-16 ***
## Residuals          6379  430.4   0.067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

```
kruskal.test(data$fund_mean_annual_return_10years ~ data$investment, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$fund_mean_annual_return_10years by data$investment
## Kruskal-Wallis chi-squared = 560.37, df = 3, p-value < 2.2e-16
```

Iz rezultata anove definitivno možemo zaključiti da sredine tih uzoraka nisu jednake te odbacit H0 u korist tvrdnje da su sredine različite. Anova nam samo govori da su njihove međusobne sredine različite no nas

naravno zanima koja od te 3 kategorije prosjcno ima najveci povrat. Iz box plota se može vidjeti da Growth ima nesto veću sredinu nego ostale dvije kategorije te da blend ima mozda malo veću sredinu nego value. Dakle provodimo t test kako bi vidjeli ima li Growth veću sredinu od blenda. Pretpostavke t testa su normalnost i nezavisnot koje smo već potvrdili prije. Prvo nas zanima jesu li varijance kategorija jednake, da bismo to saznali provodima test o jednakosti varijanci.

```
var.test(data_growth_filtered, data_blend_filtered)
```

```
##
##  F test to compare two variances
##
## data:  data_growth_filtered and data_blend_filtered
## F = 0.9256, num df = 2821, denom df = 1802, p-value = 0.06864
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8510221 1.0059219
## sample estimates:
## ratio of variances
##          0.9255974
```

Vidimo da nam test daje p-vrijednost = 0.0686 (ako ommitamo na pocetku ispada 0.0049, treba vidit sta je bolje??)što znači da na razini značajnosti 0.05 ne možemo odbaciti da su varijance jednake. Budući da su varijance jednake radimo t-test za dvije populacije koje imaju jednake varijance.

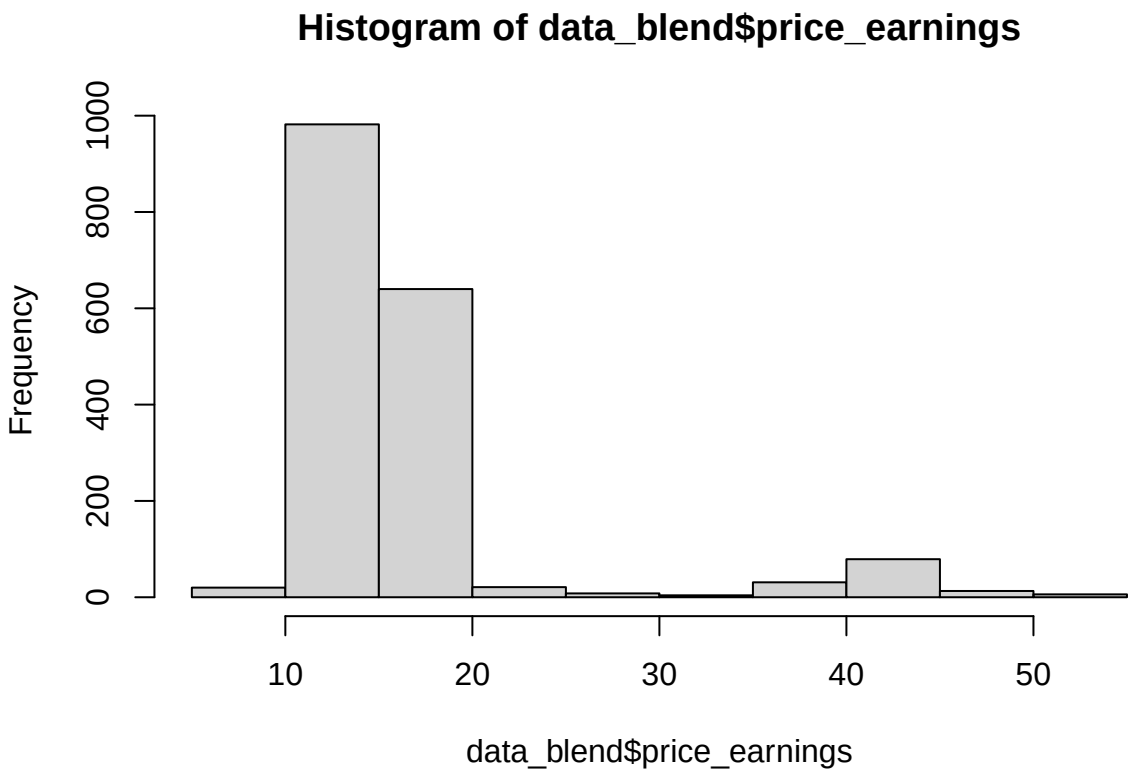
```
t.test(data_growth$fund_mean_annual_return_10years, data_blend$fund_mean_annual_return_10years, alt = "
```

```
##
##  Two Sample t-test
##
## data:  data_growth$fund_mean_annual_return_10years and data_blend$fund_mean_annual_return_10years
## t = 13.387, df = 4623, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.09664701        Inf
## sample estimates:
## mean of x mean of y
##  1.217055  1.106866
```

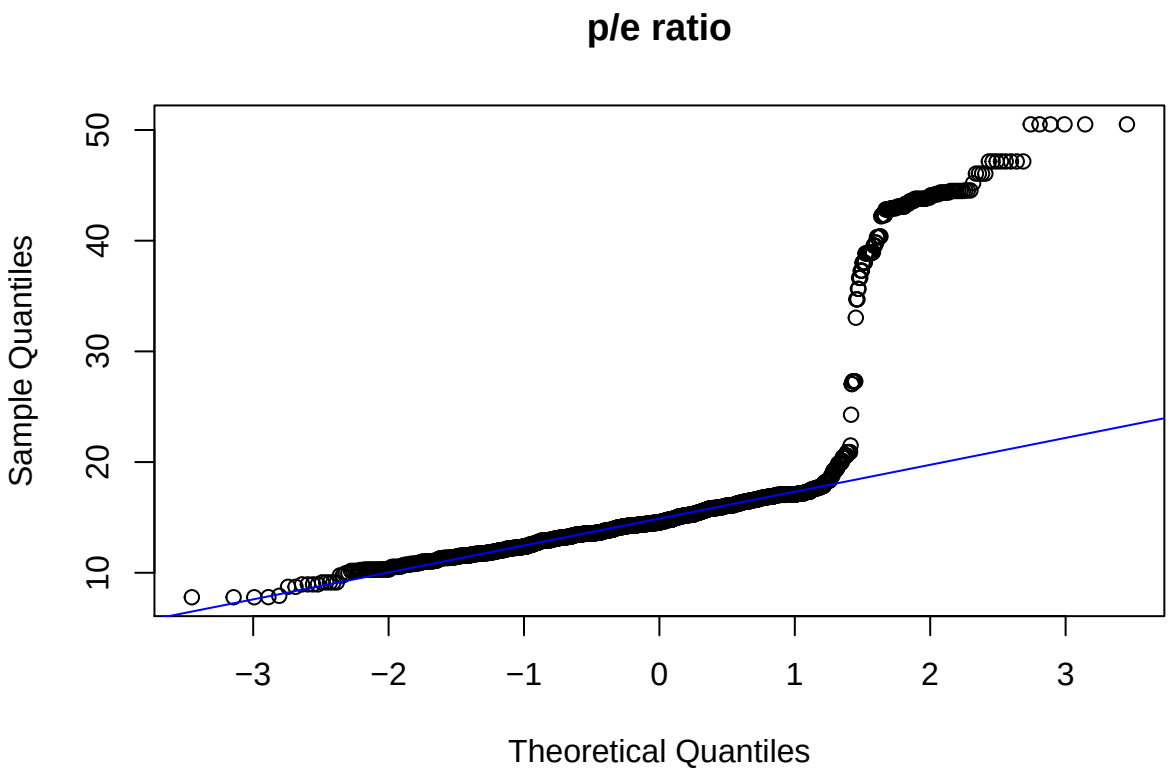
ZAKLJUCAK!!: Vidimo da je p-vrijednost izuzetno mala, što nam govori u prilog odbacivanja nulte hipoteze o jednakosti sredina. Možemo zaključiti da kategorija growth ima veći povrat od kategorije blend, pri čemu je srednja vrijednost prve 1.2171, a druge 1.1069

1.1 P/E ratio GLEDAMO UTJECE LI STIL INVESTIRANJA NA P/E RATIO -> PRETPOSTAVKA JE DA BI TREBAO UTJECATI. Pretpostavke za anovu normalnost, nezavisnost i jednakost varijanci...

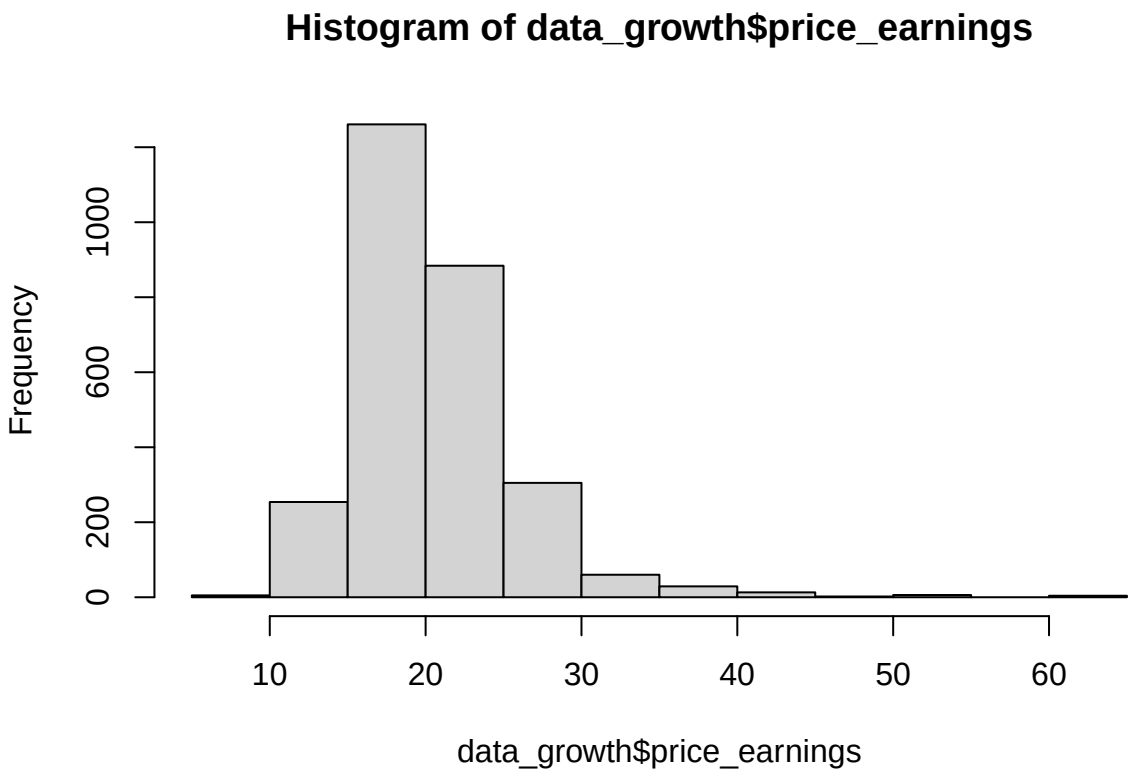
```
data_blend <- data[data$investment == c("Blend"),]
hist(data_blend$price_earnings)
```



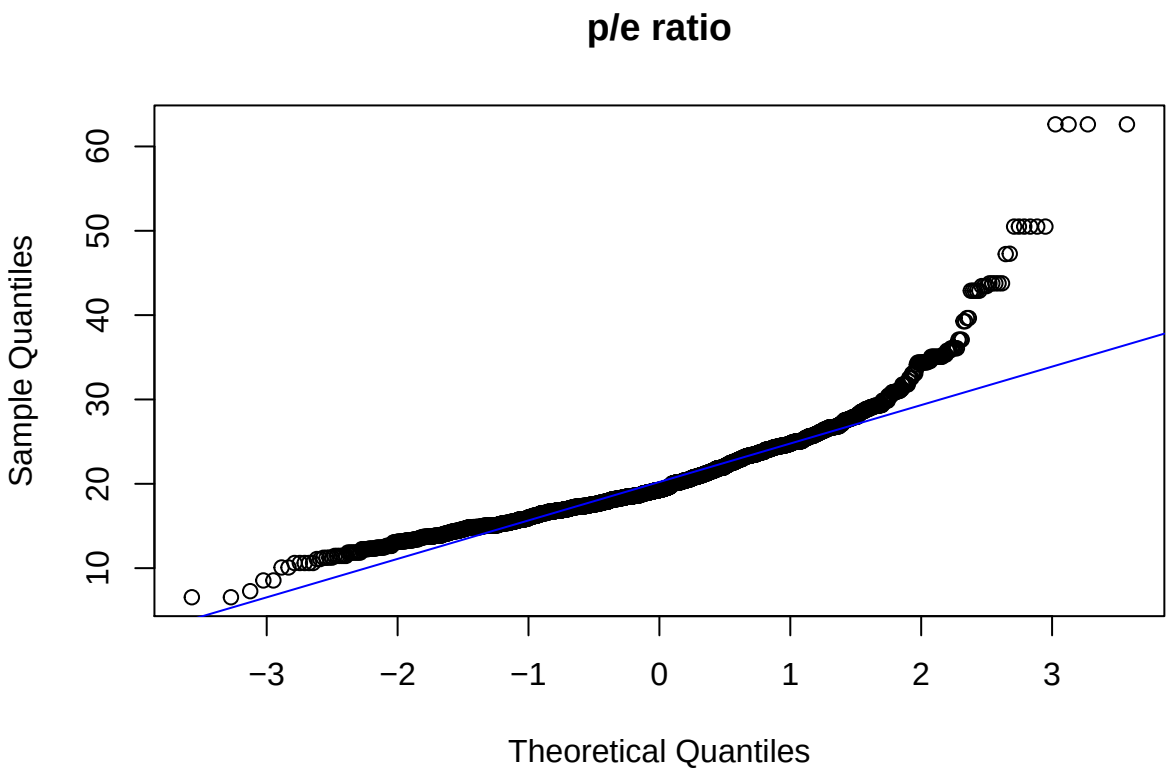

```r
qqnorm(data_blend$price_earnings, main="p/e ratio")
qqline(data_blend$price_earnings, col="blue")
```



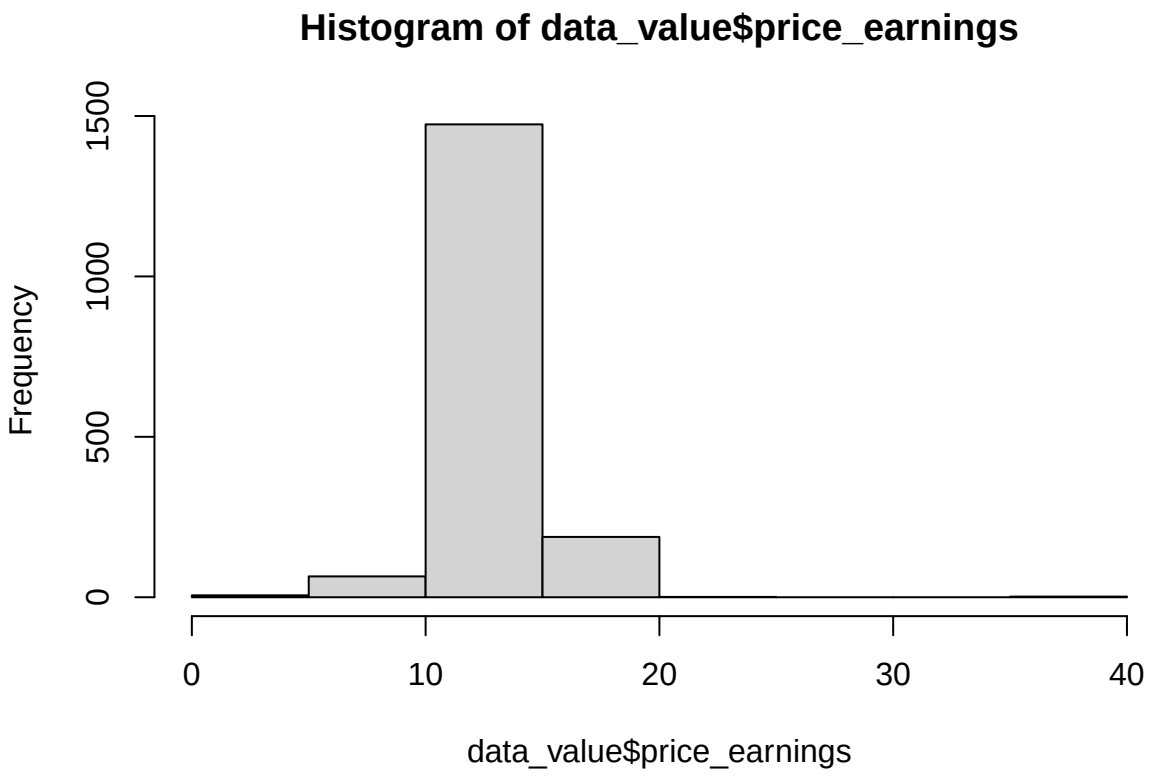

```
#ks.test(data_blend$price_earnings, "pnorm")

data_growth <- data[data$investment == c("Growth"),]
hist(data_growth$price_earnings)
```



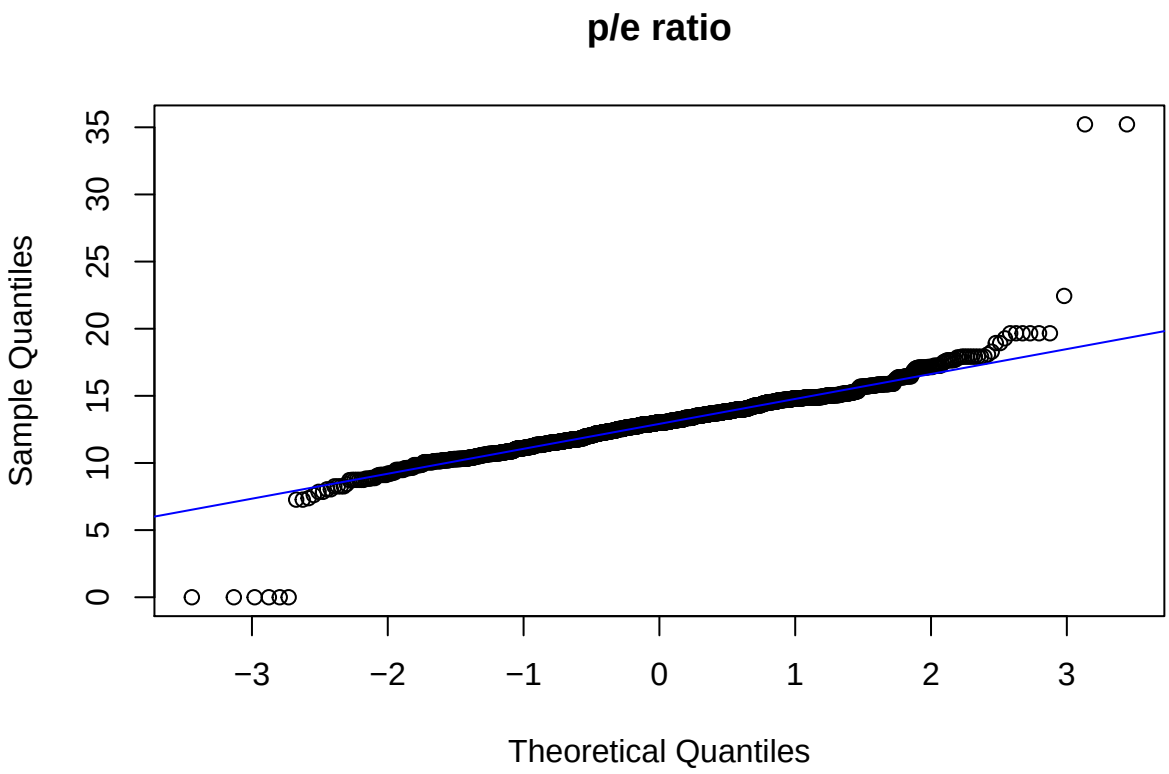

```
qqnorm(data_growth$price_earnings, main="p/e ratio")
qqline(data_growth$price_earnings, col="blue")
```




```
#ks.test(data_growth$fund_mean_annual_return_10years, "pnorm")


data_value <- data[data$investment == c("Value"),]
hist(data_value$price_earnings)
```

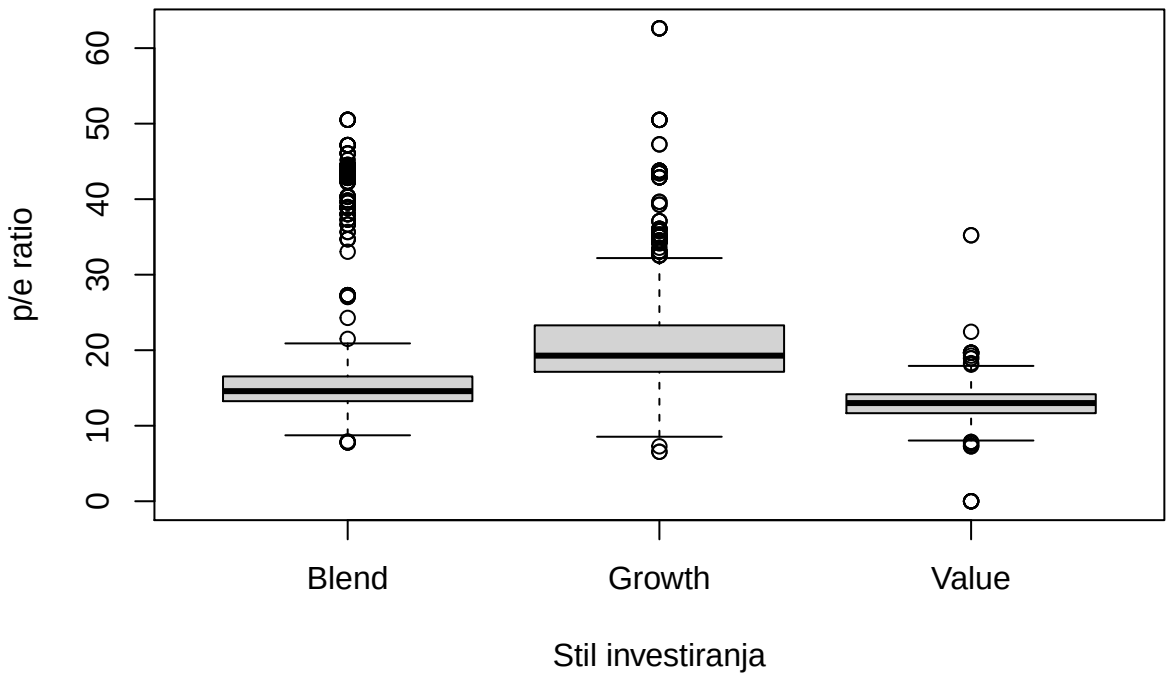



```
qqnorm(data_value$price_earnings, main="p/e ratio")
qqline(data_value$price_earnings, col="blue")
```



Vidimo da se prvi histogram i qqplot bitno razlikuju od ocekivanog za normalnu razdiobu s toga je mozda pametnije koristit kruskal-wallis(ovdje anova i kruskal-wallis)

```
boxplot(data$price_earnings[data$investment != "<undefined>"] ~ data$investment[data$investment != "<und
        ylab= "p/e ratio",
        xlab= "Stil investiranja")
```




```
##res.aov <- aov(price_earnings ~ factor(investment), data = data)
##summary(res.aov)


kruskal.test(data$price_earnings ~ data$investment, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$price_earnings by data$investment
## Kruskal-Wallis chi-squared = 3253.2, df = 3, p-value < 2.2e-16
```

Možemo vidjeti da nam anova(i kruskal-wallis) i ovdje sugerira da možemo odbaciti nultu hipotezu da su sredine jednake te možemo zaključiti da se one razlikuju. Može nas zanimati koja koategorija ima najmanji P/E ratio, odnosno koja kategorija ima najmanji omjer cijene dionice i zarade po dionici.

```
var(data_value$price_earnings)
```

```
## [1] 4.580925
```

```
var(data_blend$price_earnings)
```

```
## [1] 59.24347
```

```
var.test(data_value$price_earning, data_blend$price_earning)
```

```
##
##  F test to compare two variances
##
```

```
## data:  data_value$price_earning and data_blend$price_earning
## F = 0.077324, num df = 1735, denom df = 1803, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.07044245 0.08488436
## sample estimates:
## ratio of variances
##         0.07732371
```

(velika razlika u varijanci provjeri dal se radi o nekim velikim outlierima??) Na prvi pogled možemo zaključit da varijance nisu jednake, a to nam i potvrđuje test o jednakosti varijanci. Sada provodimo t-test. Jel oke proveden??

```
t.test(data_blend$price_earnings, data_value$price_earnings, alt = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_blend$price_earnings and data_value$price_earnings
## t = 19.401, df = 2090.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.344302      Inf
## sample estimates:
## mean of x mean of y
##  16.62401  12.96975
```

ZAKLJUCAK!!: T test nam govori da kategorija value ima manji p/e ratio(malo objasnit, proširit)

2.KATERGORIJA INVESTIRANJA GLEDAMO IMA LI RAZLIKE U POVRATU S OBZIROM NA KATEGORIJU INVESTIRANJA

```
#ima oko 50 kategorija, neznam dal ova anova ima smisla opce
res.aov <- aov(fund_mean_annual_return_10years ~ factor(category), data = data)
summary(res.aov)
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## factor(category)   52  300.6   5.780   228.8 <2e-16 ***
## Residuals        6330  159.9   0.025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

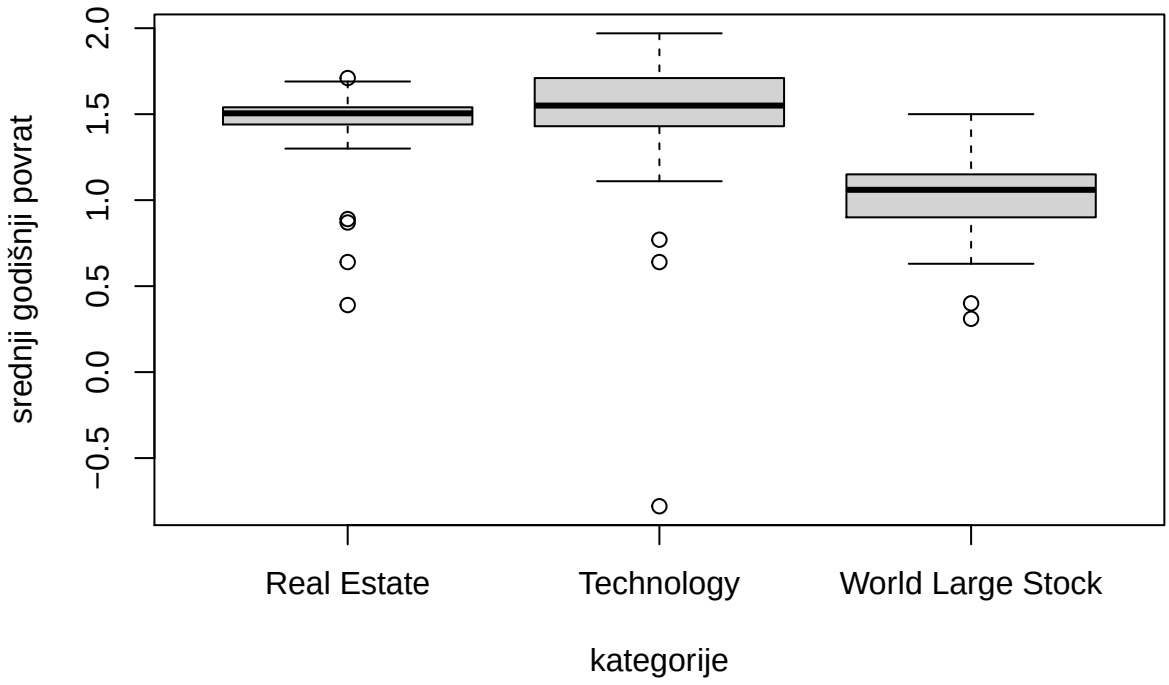
```
data_cat <- data$category
as.data.frame(table(data_cat))
```

```
##                          data_cat Freq
## 1  Allocation - 50% to 70% Equity    1
## 2  Allocation - 70% to 85% Equity    1
## 3                    China Region   32
## 4                  Communications   16
## 5               Consumer Cyclical   19
## 6              Consumer Defensive    8
## 7                    Convertibles   17
## 8                  Corporate Bond    1
## 9       Diversified Emerging Mkts  178
## 10       Diversified Pacific/Asia   13
## 11                  Equity Energy   32
```

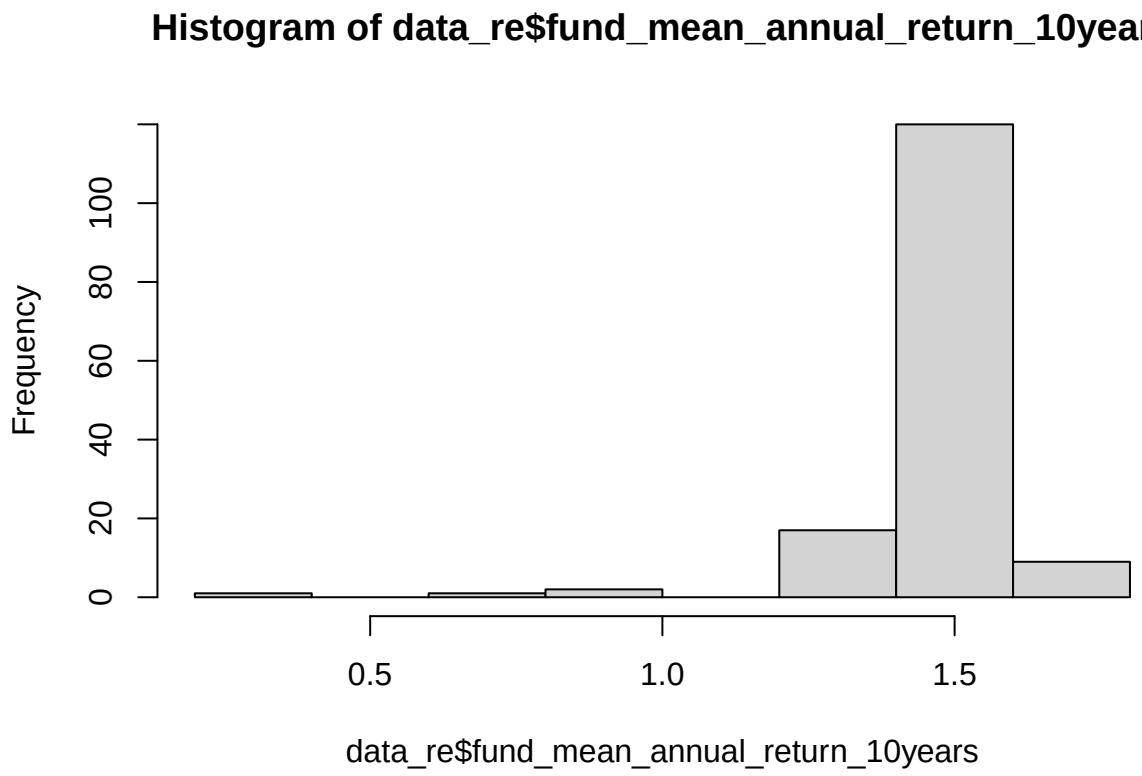
```
## 12         Equity Precious Metals   28
## 13                  Europe Stock   31
## 14                     Financial   57
## 15            Foreign Large Blend  277
## 16           Foreign Large Growth  224
## 17            Foreign Large Value  132
## 18         Foreign Small/Mid Blend   15
## 19        Foreign Small/Mid Growth   66
## 20         Foreign Small/Mid Value   14
## 21              Global Real Estate   22
## 22                         Health   52
## 23                 High Yield Bond    1
## 24                    India Equity    6
## 25                     Industrials   18
## 26                  Infrastructure   25
## 27          Intermediate-Term Bond    1
## 28                     Japan Stock   19
## 29                     Large Blend  776
## 30                    Large Growth  804
## 31                     Large Value  740
## 32             Latin America Stock    8
## 33               Long-Short Equity   23
## 34                  Market Neutral   13
## 35                   Mid-Cap Blend  256
## 36                  Mid-Cap Growth  372
## 37                   Mid-Cap Value  256
## 38                   Multicurrency    4
## 39            Muni National Interm    1
## 40         Muni Single State Interm    1
## 41                Natural Resources   65
## 42                    Options-based   13
## 43      Pacific/Asia ex-Japan Stk   24
## 44                     Real Estate  150
## 45                     Small Blend  406
## 46                    Small Growth  407
## 47                     Small Value  251
## 48             Tactical Allocation    7
## 49                      Technology  104
## 50                       Utilities   37
## 51                     World Bond    1
## 52               World Large Stock  334
## 53            World Small/Mid Stock   27
```

Anova nam ukazuje na to da kategorije nemaju jednaku sredinu, možemo odbaciti hipotezu o jednakosti sredina u korist alternativne hipoteze, a to je da su sredine različite. Budući da postoje 53 različite kategorije, uzeli smo one kategorije u koje puno fondova priprada. Najviše njih pripada u kategorije Large Blend, Large Value, Large Growth, Small Blend, Small Value, Small Growth koje većinom imaju iste stilove (Blend, Growth, Value) koje smo već proučavali gore. Pa ćemo ovjde odabrati neke druge zanimljive kategorije koje također imaju puno fondova.(World Large Stock, Technology, Real Estate)

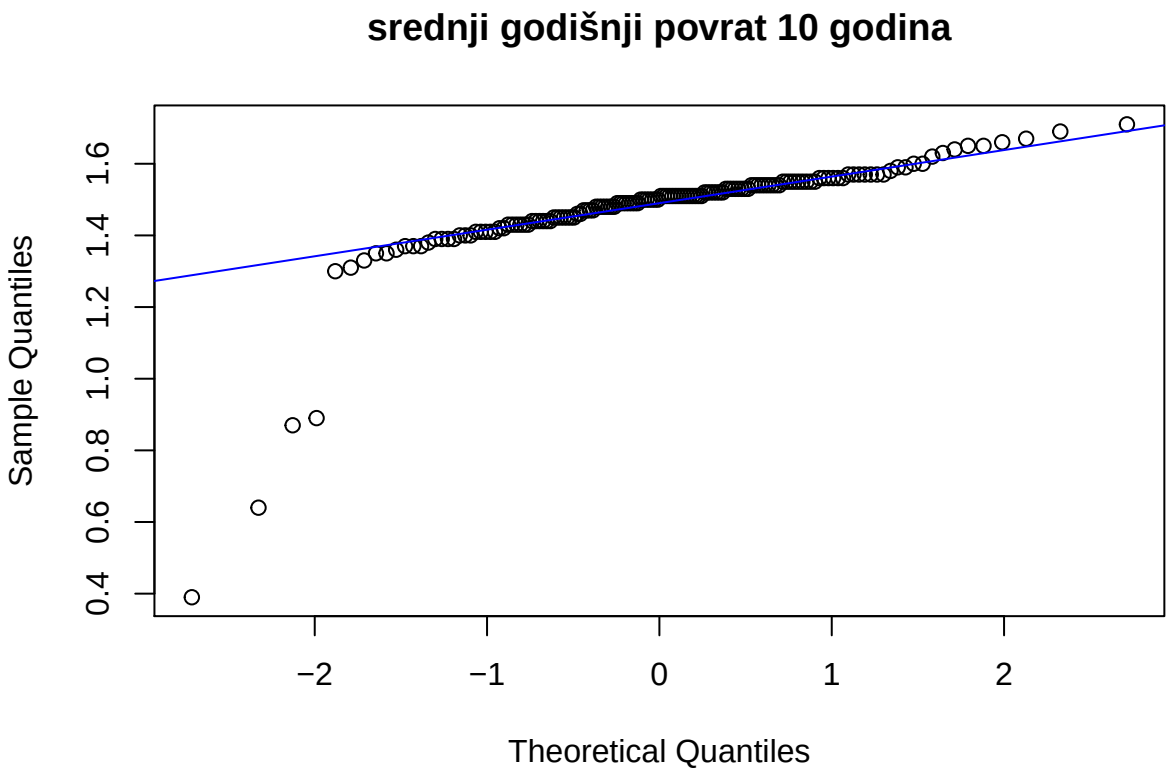
```r
filtered_by_categories <- data[data$category == "World Large Stock" | data$category == "Technology" | da

boxplot(filtered_by_categories$fund_mean_annual_return_10years ~ filtered_by_categories$category,
        ylab = "srednji godišnji povrat",
        xlab = "kategorije")
```



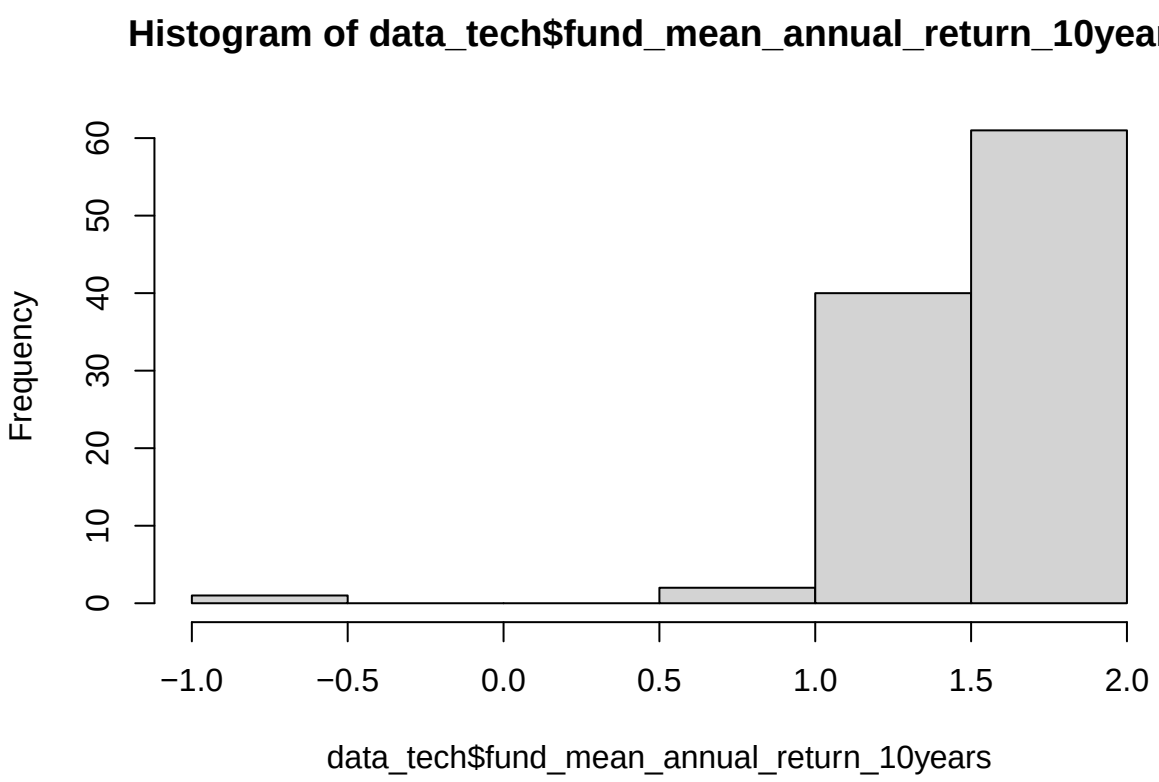

```
data_re <- data[data$category == c("Real Estate"),]
hist(data_re$fund_mean_annual_return_10years)
```



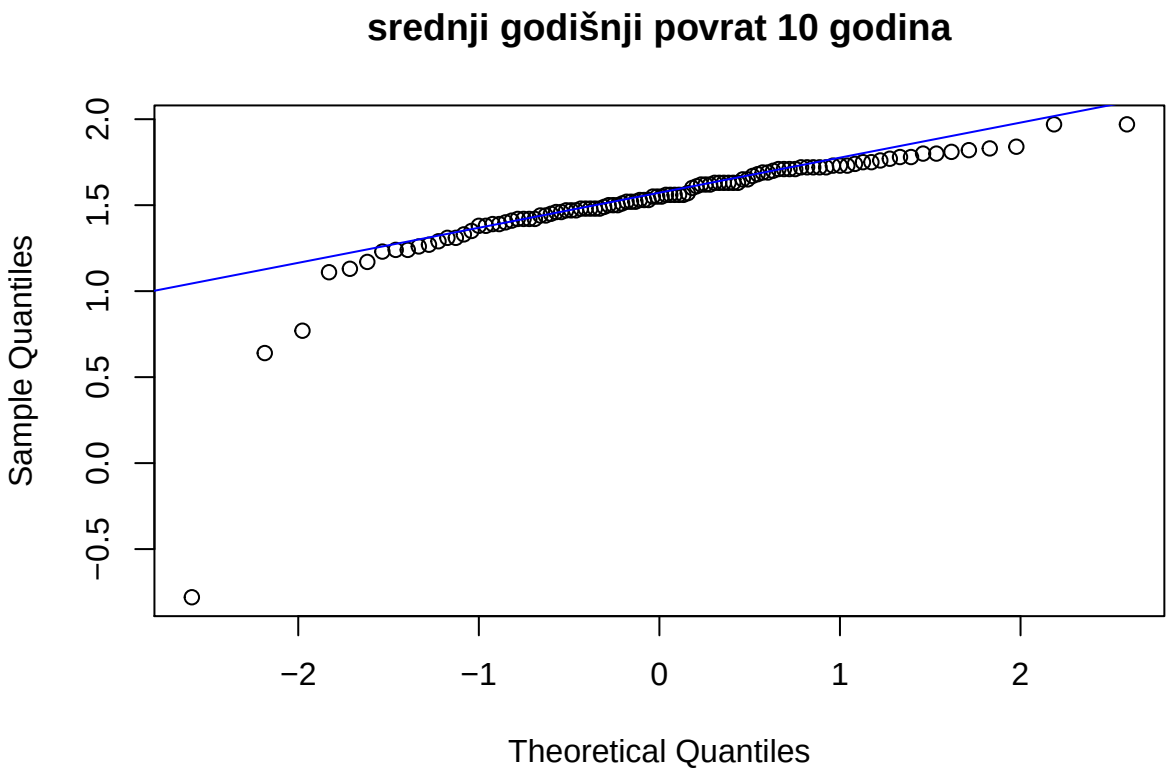

```
qqnorm(data_re$fund_mean_annual_return_10years, main="srednji godišnji povrat 10 godina")
qqline(data_re$fund_mean_annual_return_10years, col="blue")
```



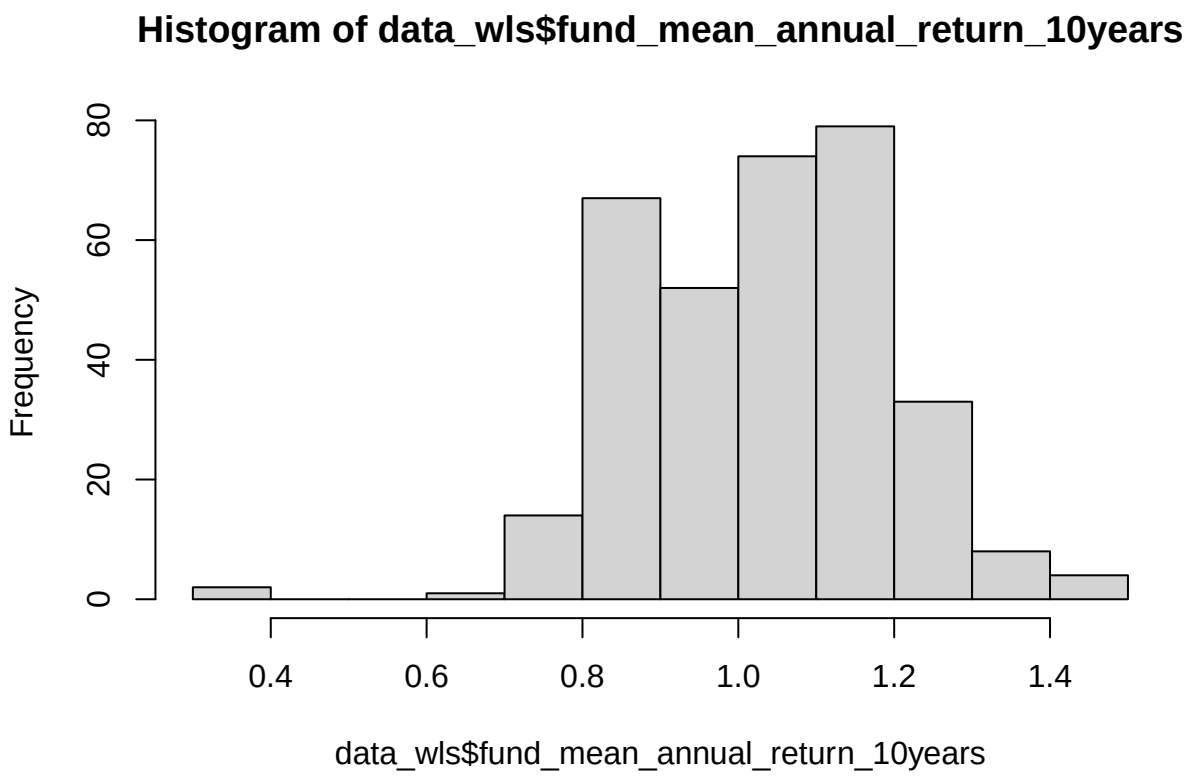

```
#ks.test(data_blend$fund_mean_annual_return_10years, "pnorm")

data_tech <- data[data$category == c("Technology"),]
hist(data_tech$fund_mean_annual_return_10years)
```



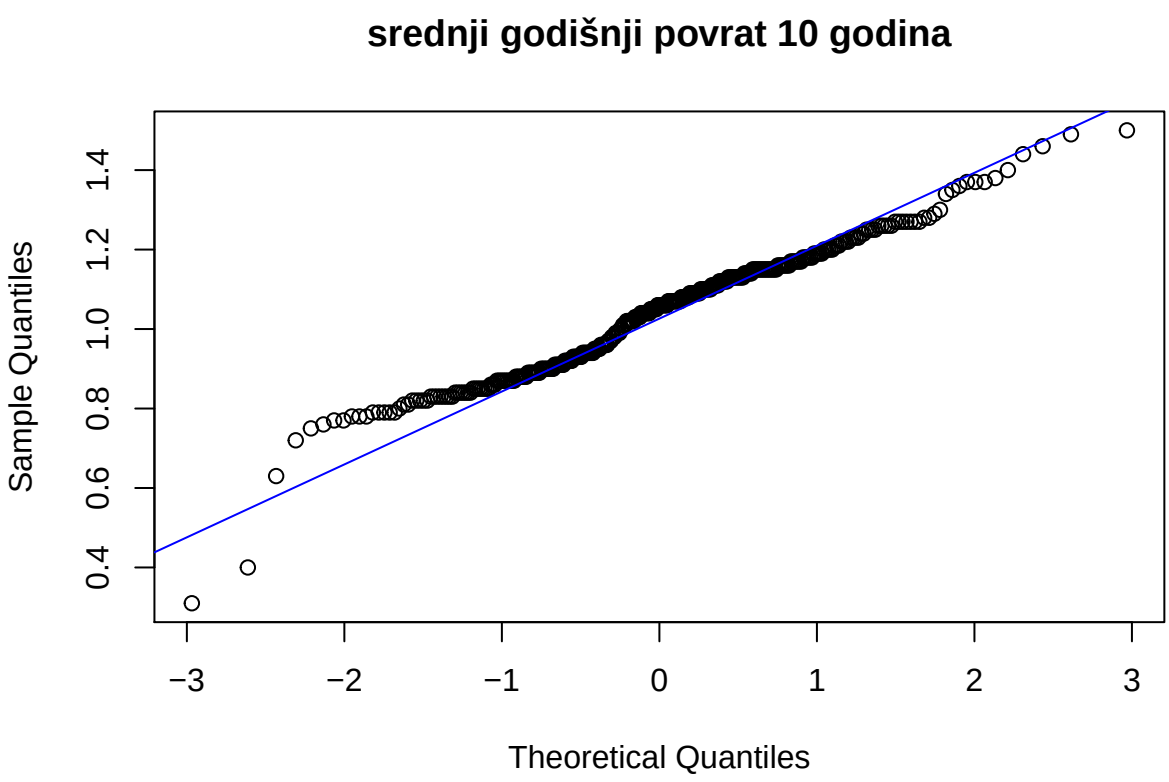

```
qqnorm(data_tech$fund_mean_annual_return_10years, main="srednji godišnji povrat 10 godina")
qqline(data_tech$fund_mean_annual_return_10years, col="blue")
```




```
#ks.test(data_growth$fund_mean_annual_return_10years, "pnorm")


data_wls <- data[data$category == c("World Large Stock"),]
hist(data_wls$fund_mean_annual_return_10years)
```




```
qqnorm(data_wls$fund_mean_annual_return_10years, main="srednji godišnji povrat 10 godina")
qqline(data_wls$fund_mean_annual_return_10years, col="blue")
```



Iz priloženih grafova čini se da podatci ne odskaču drastično od normalne distribucije. Uz pretpostavku nezavisnosti podataka, normalnosti te jednakosti varijanci, možemo provesti anova test.

```
data_re_filtered <- na.omit(data_re$fund_mean_annual_return_10years)
data_tech_filtered <- na.omit(data_tech$fund_mean_annual_return_10years)
data_wls_filtered <- na.omit(data_wls$fund_mean_annual_return_10years)

var(data_re_filtered)
```

```
## [1] 0.02335224
```

```
var(data_tech_filtered)
```

```
## [1] 0.09662187
```

```
var(data_wls_filtered)
```

```
## [1] 0.02697088
```

```
res.aov <- aov(fund_mean_annual_return_10years ~ factor(category), data = filtered_by_categories)
summary(res.aov)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## factor(category)   2  30.13  15.065   393.2 <2e-16 ***
## Residuals        585  22.41   0.038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova nam govori u prilog alternativne hipoteze, a to je da su sredine te tri kategorije različite. Iz box plota se vidi da su sredine kategorija Real Estate i Technology iznad sredine od World Large Stock, stoga ćemo uz

pomoć test testa provjerit imaju li Real Estate i Technology jednake srednje povrate.

```
var.test(data_re_filtered, data_tech_filtered)
```

```
##
##  F test to compare two variances
##
## data:  data_re_filtered and data_tech_filtered
## F = 0.24169, num df = 149, denom df = 103, p-value = 4.167e-15
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1681521 0.3433157
## sample estimates:
## ratio of variances
##          0.2416869
```

Test o jednakosti varijanci odbacuje da su varijance jednake stoga provodimo t-test za nezavisne uzorke za populacije s različitim varijancama.

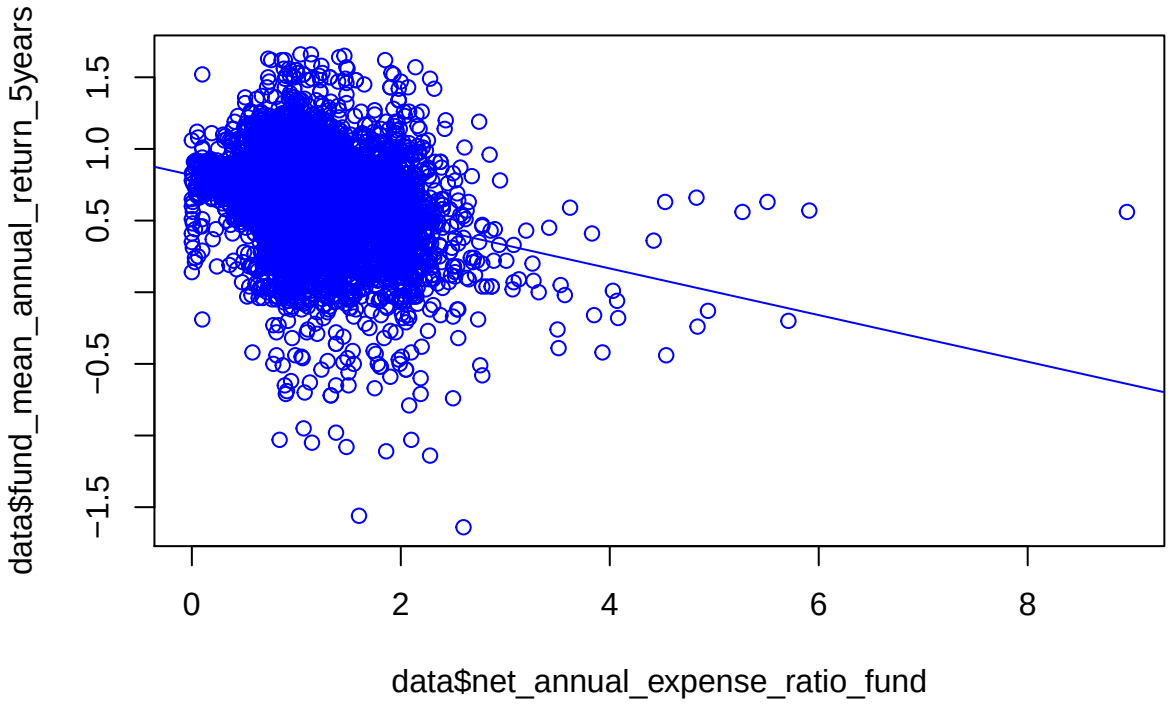
```
t.test(data_tech$fund_mean_annual_return_10years, data_re$fund_mean_annual_return_10years, alt = "great
```

```
##
##  Welch Two Sample t-test
##
## data:  data_tech$fund_mean_annual_return_10years and data_re$fund_mean_annual_return_10years
## t = 1.3054, df = 137.74, p-value = 0.09697
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.0115471        Inf
## sample estimates:
## mean of x mean of y
##  1.519327  1.476333
```

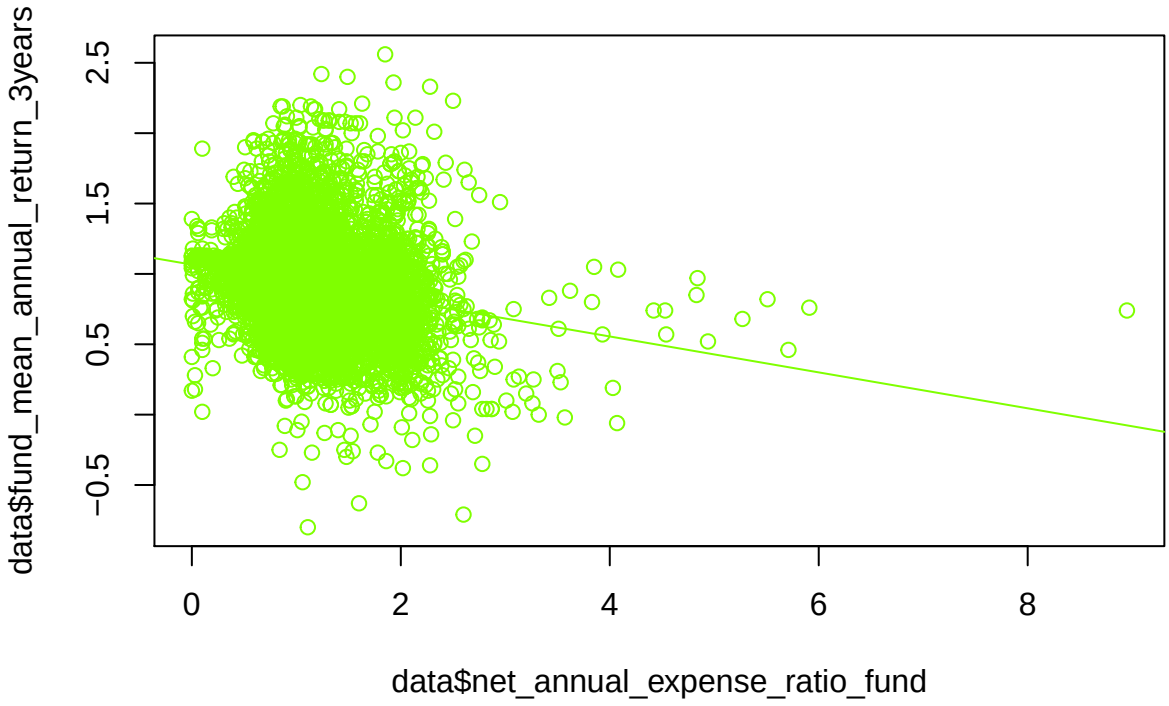
T-test nam vraća p-vrijednost = 0,09697 što znači da ne možemo odbaciti nultu hipotezu da su sredine te dvije skupine jednake.

```
plot(data$net_annual_expense_ratio_fund, data$fund_mean_annual_return_10years, col = "red")
abline(lm(data$fund_mean_annual_return_10years~data$net_annual_expense_ratio_fund), col="red")
```

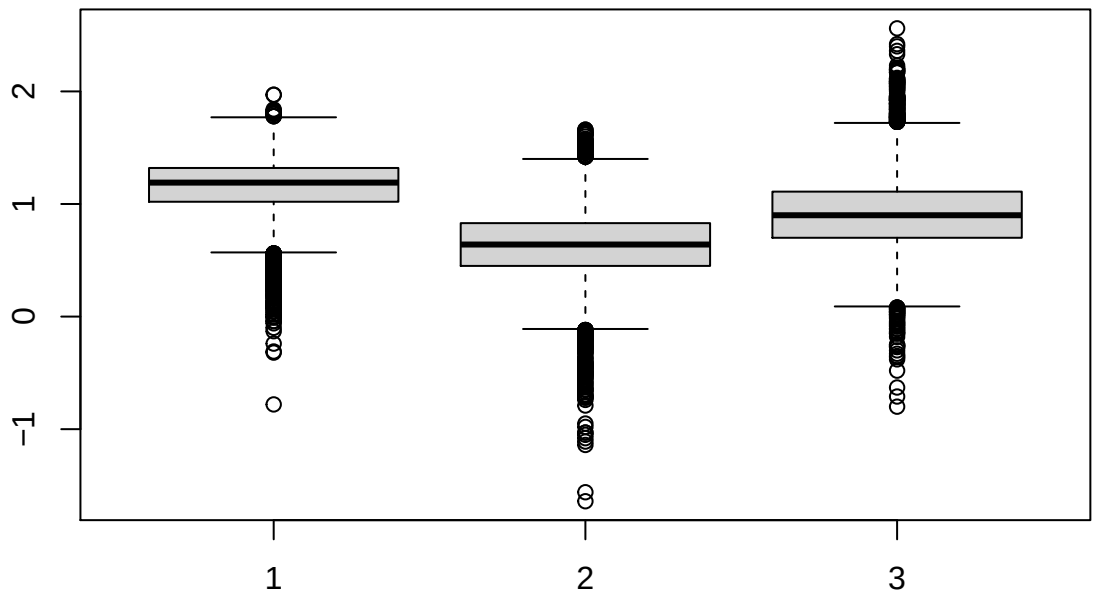

```
plot(data$net_annual_expense_ratio_fund, data$fund_mean_annual_return_5years, col = "blue")
abline(lm(data$fund_mean_annual_return_5years~data$net_annual_expense_ratio_fund), col="blue")
```




```
plot(data$net_annual_expense_ratio_fund, data$fund_mean_annual_return_3years, col = "chartreuse")
abline(lm(data$fund_mean_annual_return_3years ~ data$net_annual_expense_ratio_fund), col="chartreuse")
```

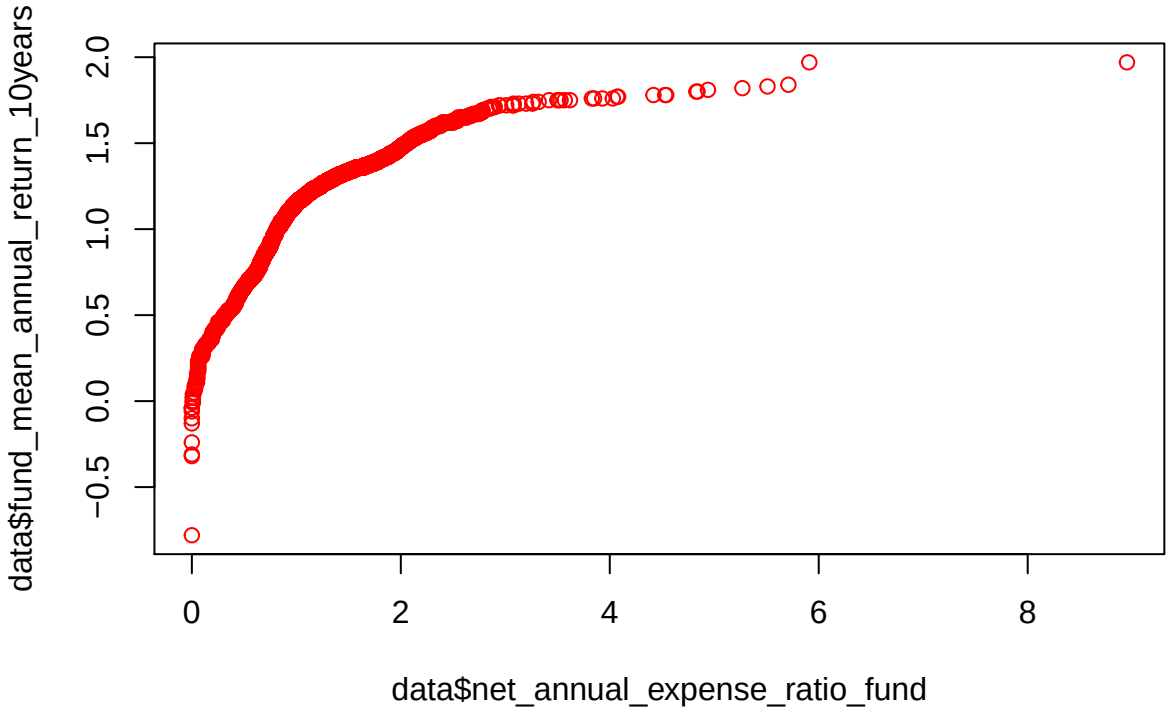



```
boxplot(data$fund_mean_annual_return_10years, data$fund_mean_annual_return_5years, data$fund_mean_annual
```

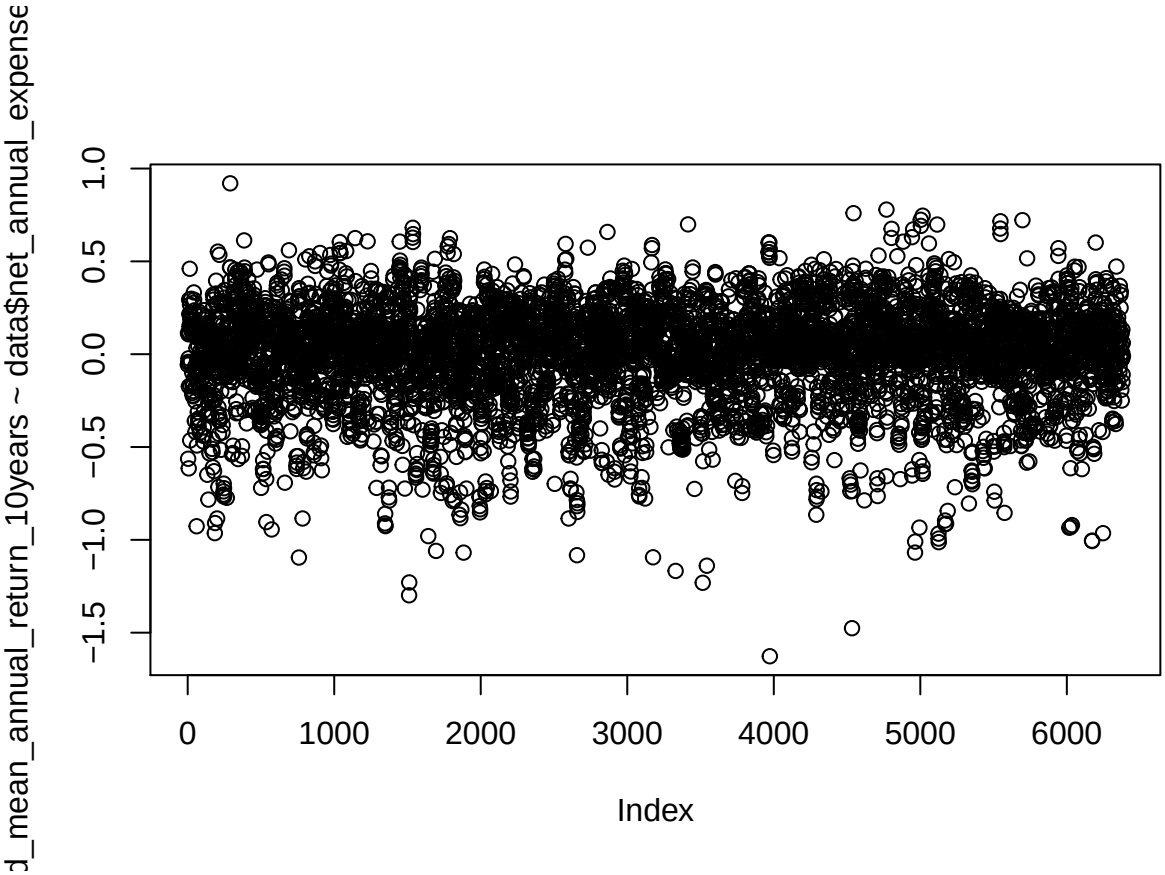


Podaci ne pokazuju nešto pretjerano pametno. Srednji povrat u zadnjih 10 godina regresija pokazuje pad povrata s obzirom na cijenu fonda. To nema smisla. Postoji li koji bolji način za određivanje uspješnosti? Čini se kako skuplji fondovi ne pružaju veću uspješnost, ali garantiraju veću stabilnost.

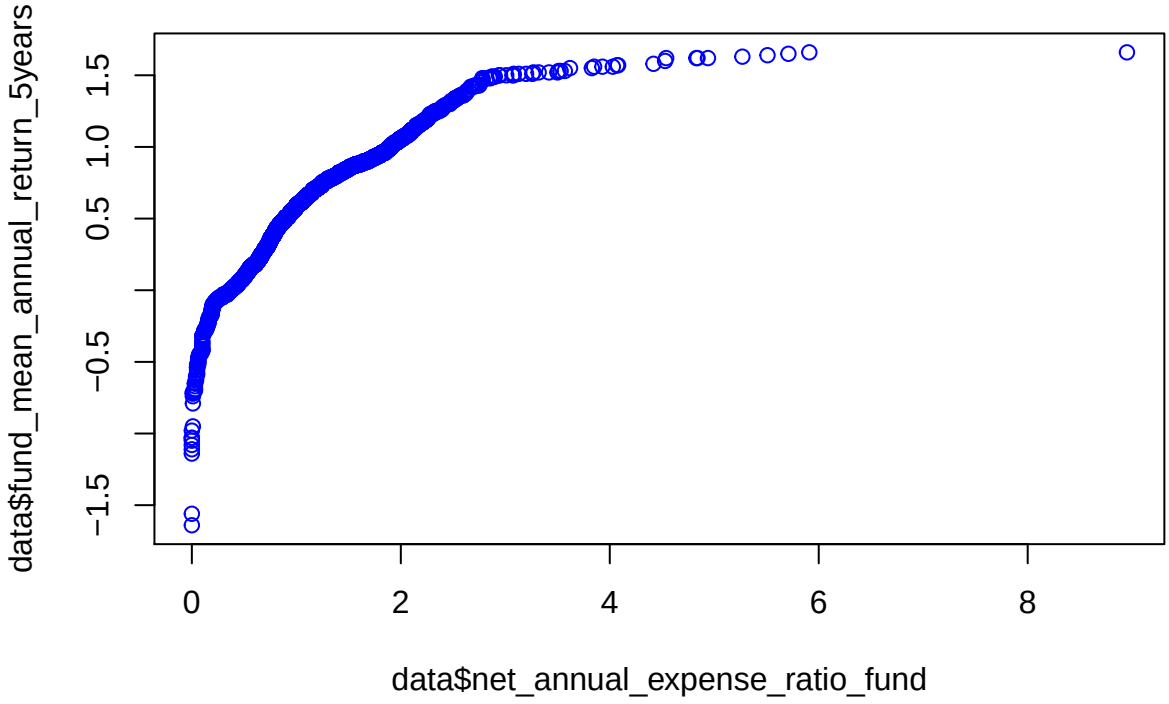
```
qqplot(data$net_annual_expense_ratio_fund, data$fund_mean_annual_return_10years, col = "red")
```



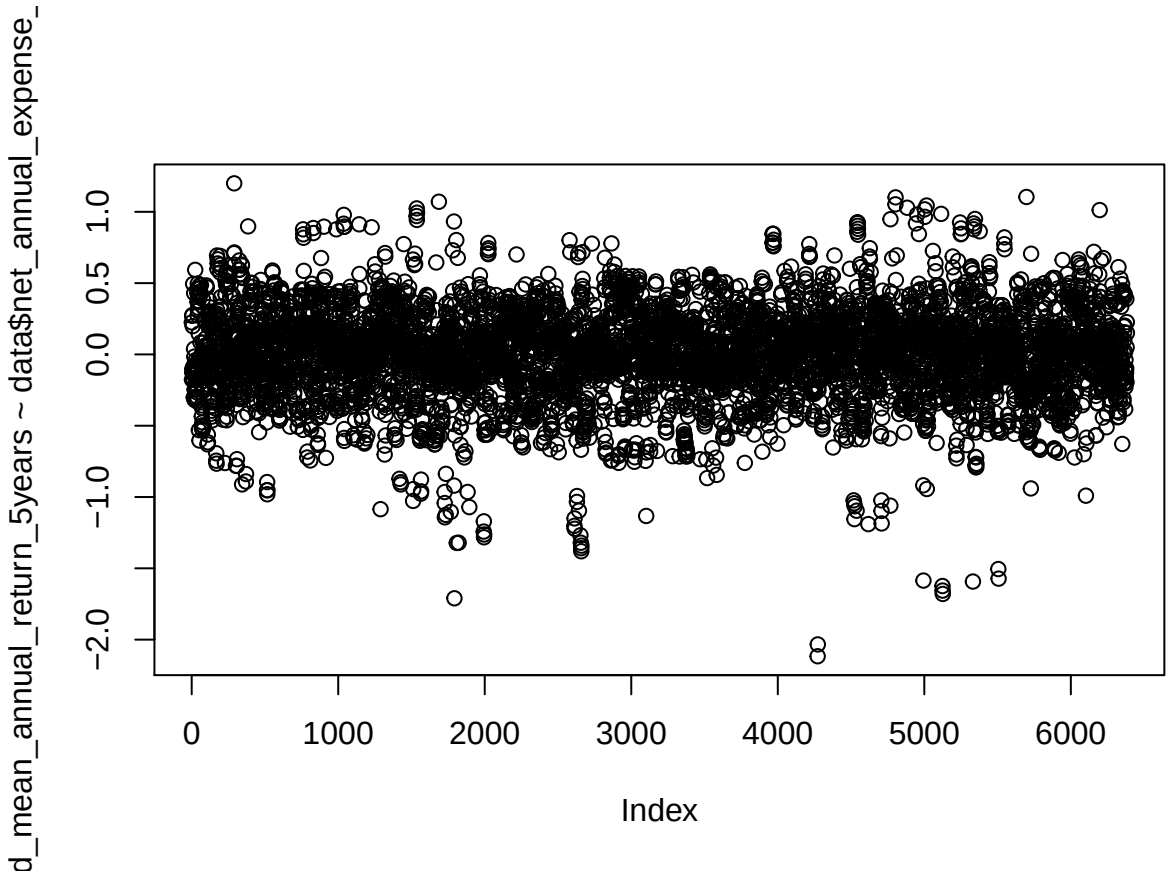

```
plot(residuals(lm(data$fund_mean_annual_return_10years~data$net_annual_expense_ratio_fund), col="red"))
```



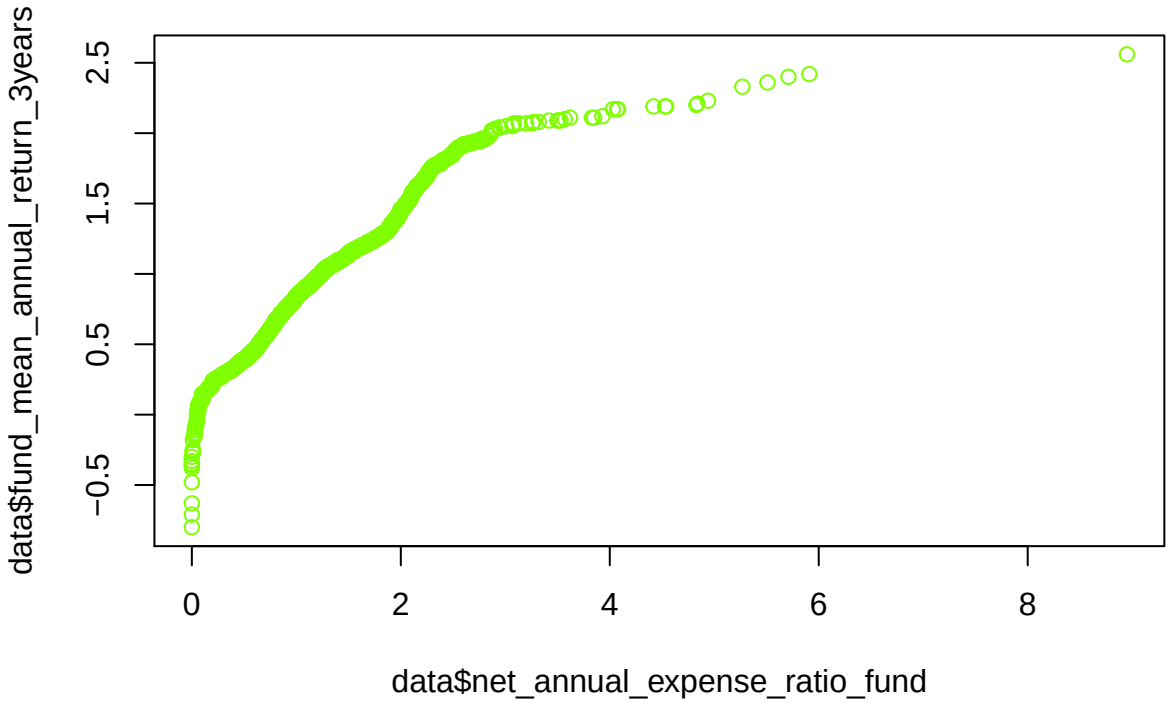

```
qqplot(data$net_annual_expense_ratio_fund, data$fund_mean_annual_return_5years, col = "blue")
```



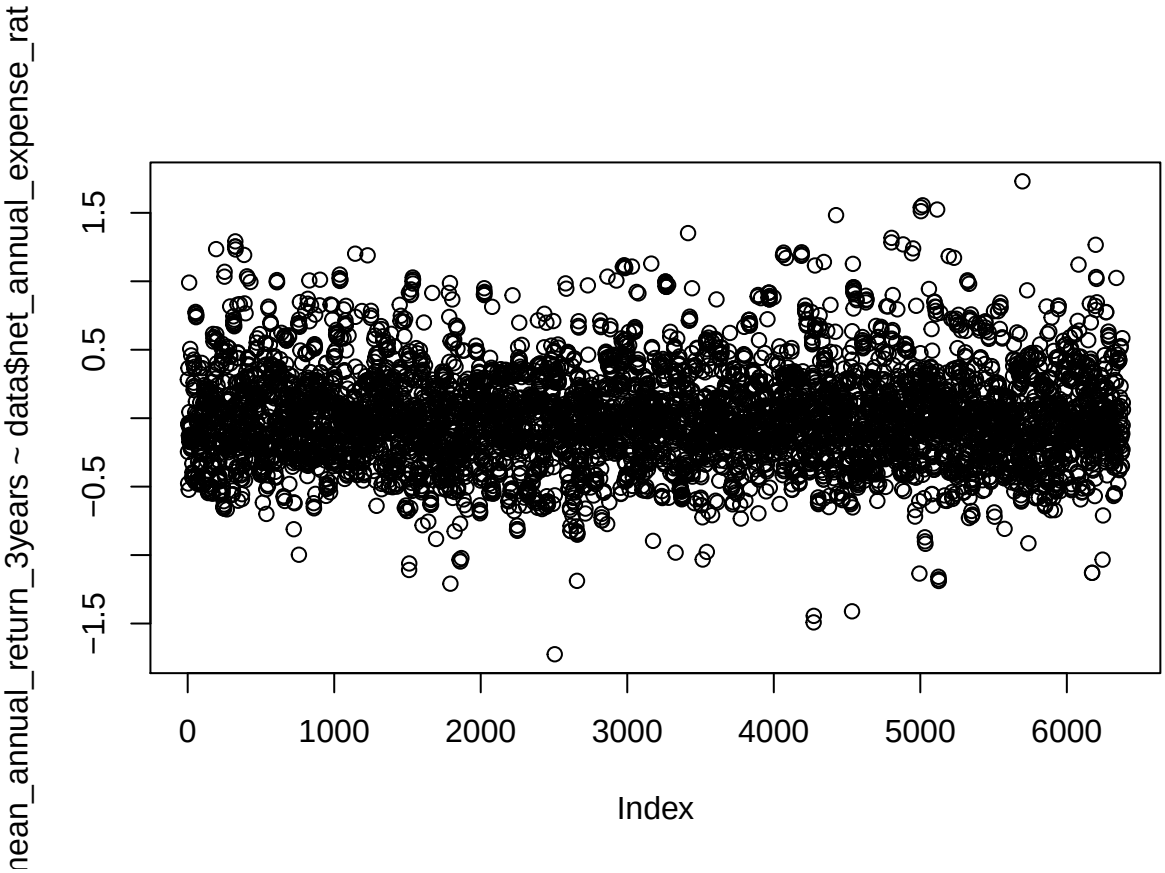

```
plot(residuals(lm(data$fund_mean_annual_return_5years~data$net_annual_expense_ratio_fund), col="blue"))
```




```
qqplot(data$net_annual_expense_ratio_fund, data$fund_mean_annual_return_3years, col = "chartreuse")
```



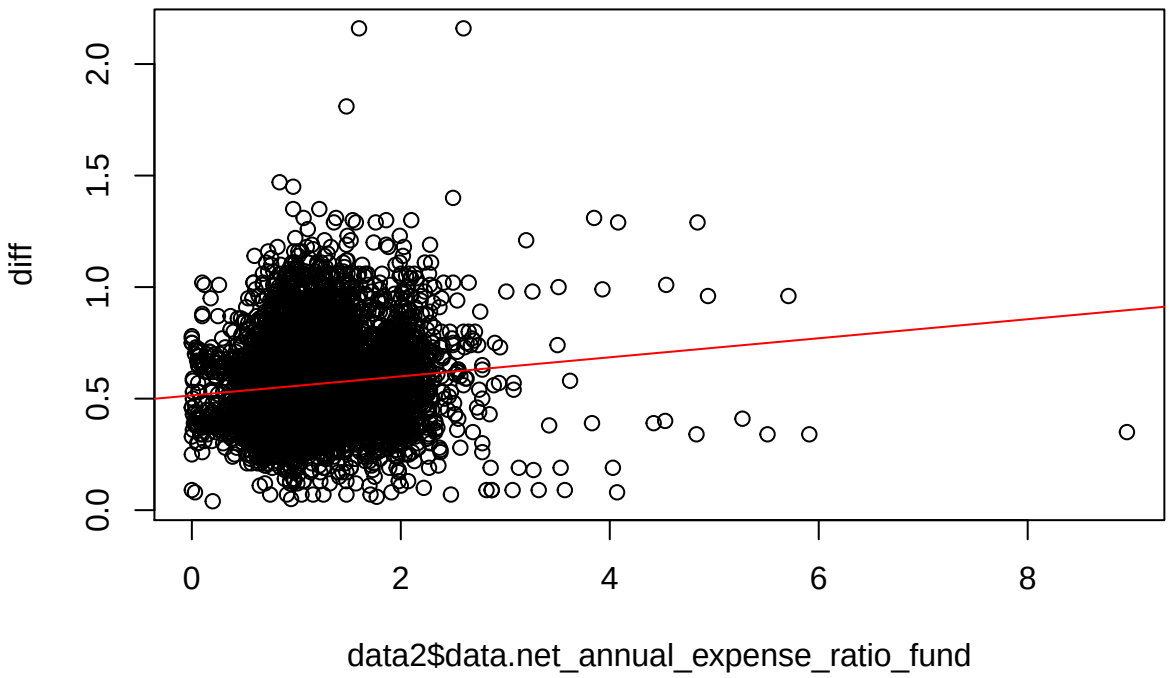

```
plot(residuals(lm(data$fund_mean_annual_return_3years ~ data$net_annual_expense_ratio_fund), col="chart
```



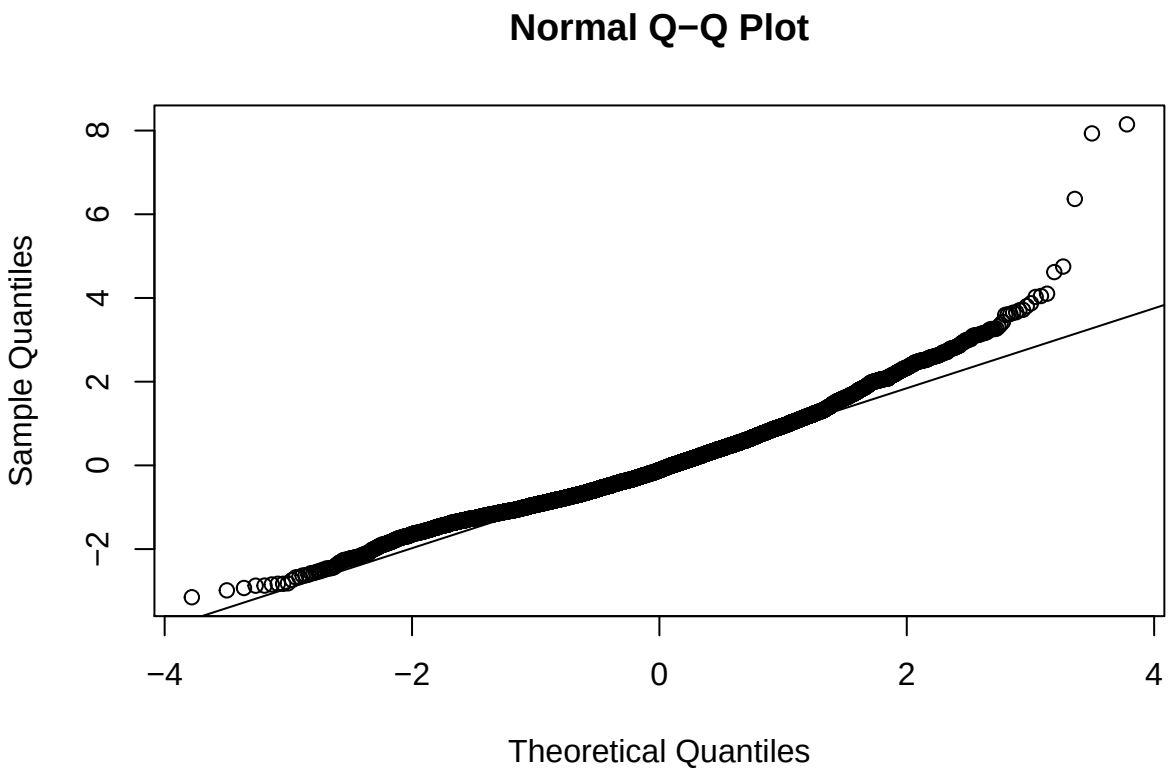
Regresija ne daje neke odgovore na to mijenja li se udio ulaganja u pojedine sektore u odnosu na godine osnivanja fonda. Jedinu zamjetnu razliku pokazuje sektor nekretnina kod kojega je krivulja rastuća prema 2000 godini pad udjela za zdravstvo.

H0: razlika je jednaka za skuplje i jeftinije fondove H1: razlika je manja za skuplje fondove

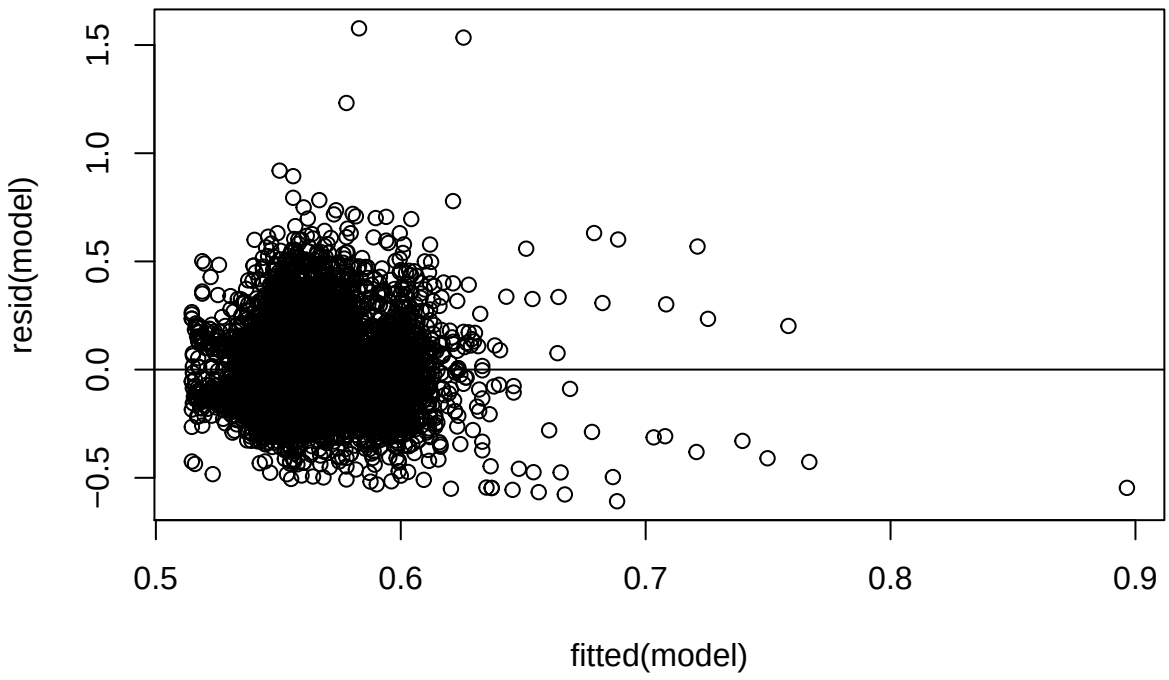
```
data1 = data.frame(data$fund_mean_annual_return_10years, data$fund_mean_annual_return_5years, data$fund
data2 = data.frame(data$net_annual_expense_ratio_fund)
data2$Min <-apply(data1,1,FUN=min)
data2$Max <-apply(data1,1,FUN=max)
diff = data2$Max - data2$Min
model = lm(diff ~ data2$data.net_annual_expense_ratio_fund)
plot(data2$data.net_annual_expense_ratio_fund, diff)
abline(model, col = "red")
```




```
qqnorm(rstandard(model))
qqline(rstandard(model))
```



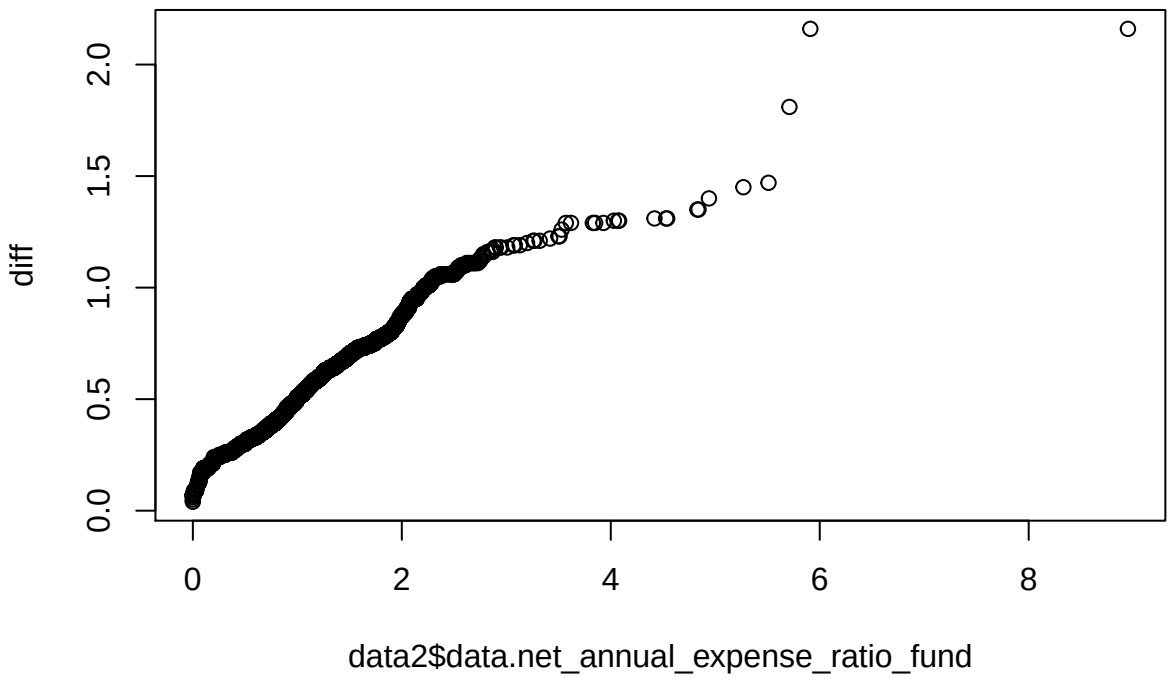

```
plot(fitted(model), resid(model))
abline(0,0)
```




```
summary(model)
```

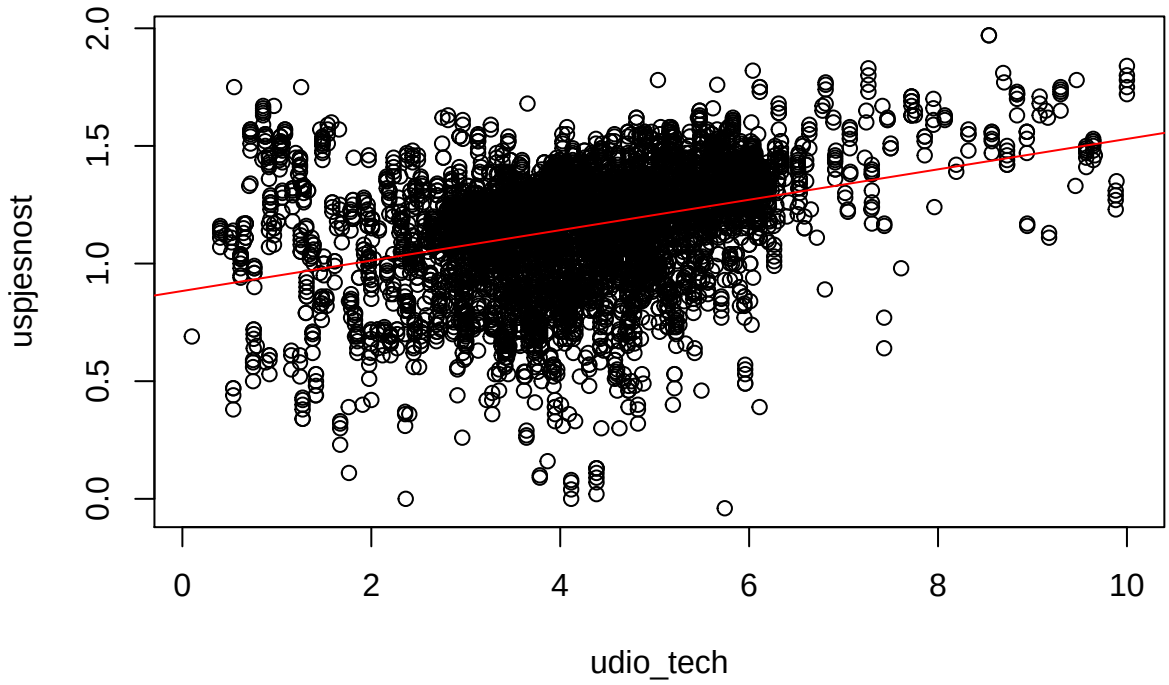
```
##
## Call:
## lm(formula = diff ~ data2$data.net_annual_expense_ratio_fund)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60833 -0.13780 -0.02247  0.11160  1.57705
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              0.514684   0.005893  87.338   <2e-16
## data2$data.net_annual_expense_ratio_fund 0.042665   0.004556   9.365   <2e-16
##
## (Intercept)                              ***
## data2$data.net_annual_expense_ratio_fund ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1936 on 6381 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.01356,    Adjusted R-squared:  0.0134
## F-statistic:  87.7 on 1 and 6381 DF,  p-value: < 2.2e-16
```

```
qqplot(data2$data.net_annual_expense_ratio_fund, diff)
```

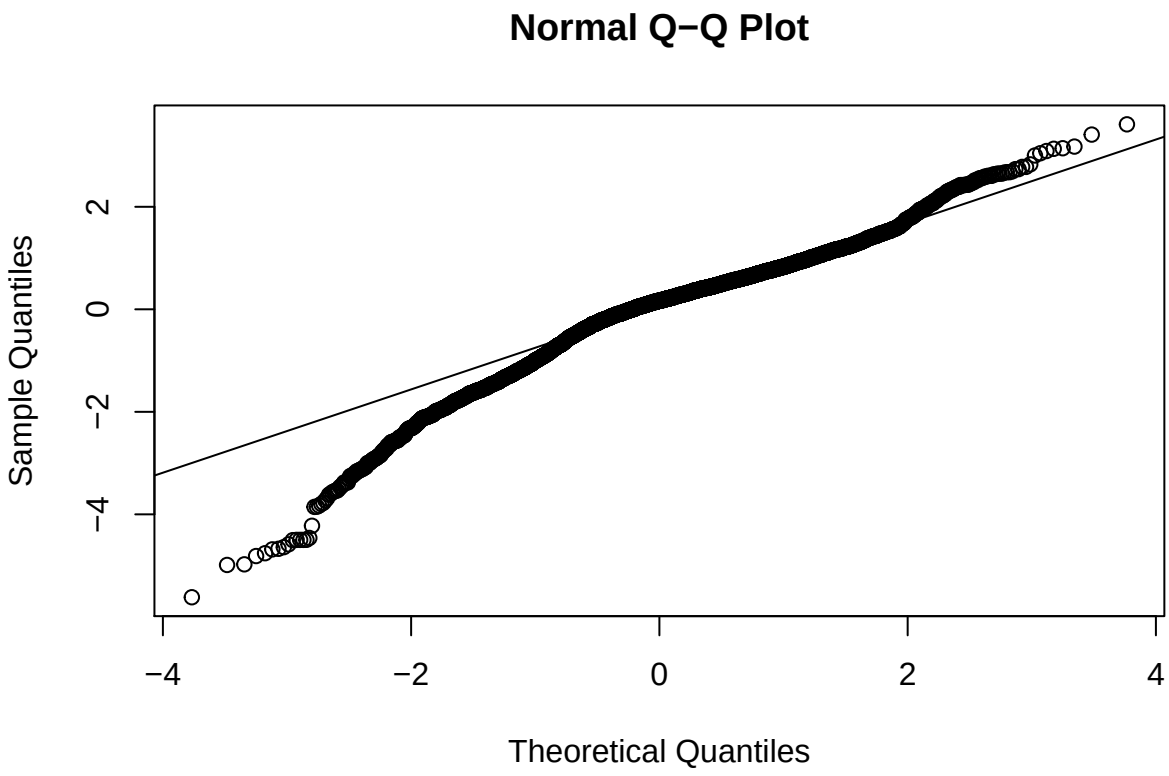


H0: fondovi imaju jednaku uspješnost bez obzira na udio koji ulažu u tehnologiju H1: fondovi imaju veću uspješnost ako imaju veći udio ulaganja u tehnologiju

```
data1 = subset(data, technology > 0)
uspjesnost = data1$fund_mean_annual_return_10years
udio_tech = sqrt(data1$technology)
model = lm(uspjesnost~udio_tech)
plot(udio_tech, uspjesnost)
abline(model, col = "red")
```




```
qqnorm(rstandard(model))
qqline(rstandard(model))
```




```
summary(model)
```

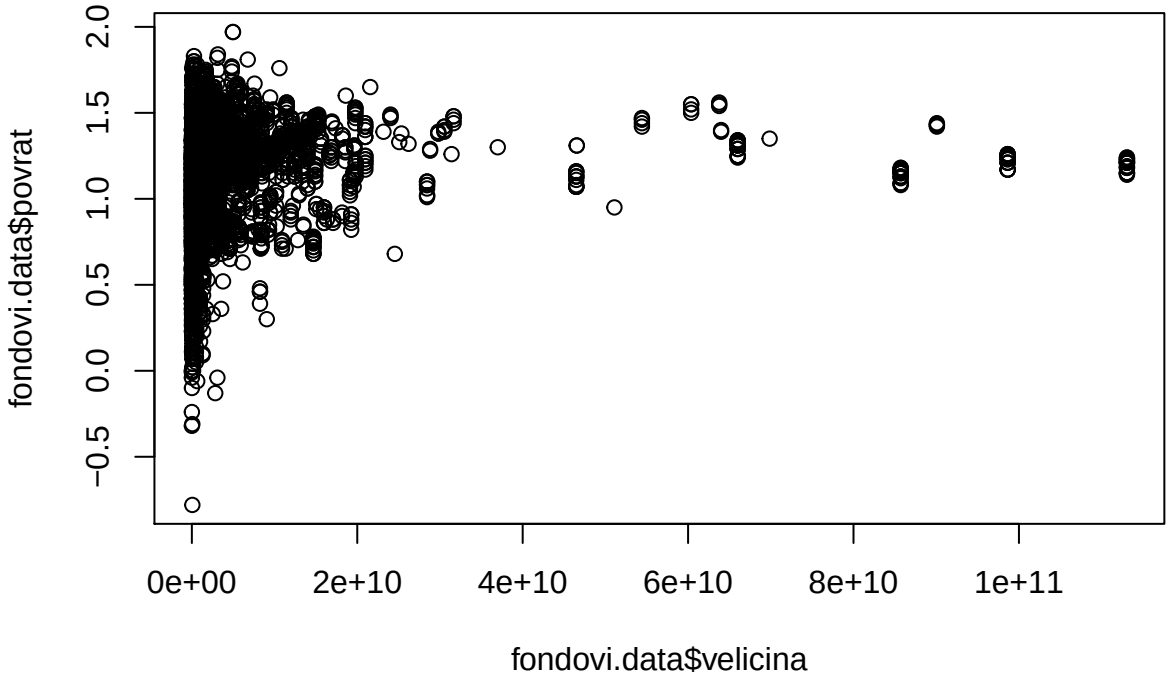
```
##
## Call:
## lm(formula = uspjesnost ~ udio_tech)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29455 -0.11198  0.03838  0.14063  0.83113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.883469   0.009609   91.94   <2e-16 ***
## udio_tech   0.064626   0.002149   30.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2305 on 6043 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.1302, Adjusted R-squared:   0.13
## F-statistic: 904.3 on 1 and 6043 DF,  p-value: < 2.2e-16
```

```
fondovi.data = data[,c("net_assets","fund_mean_annual_return_10years")]
colnames(fondovi.data) = c("velicina", "povrat")
```

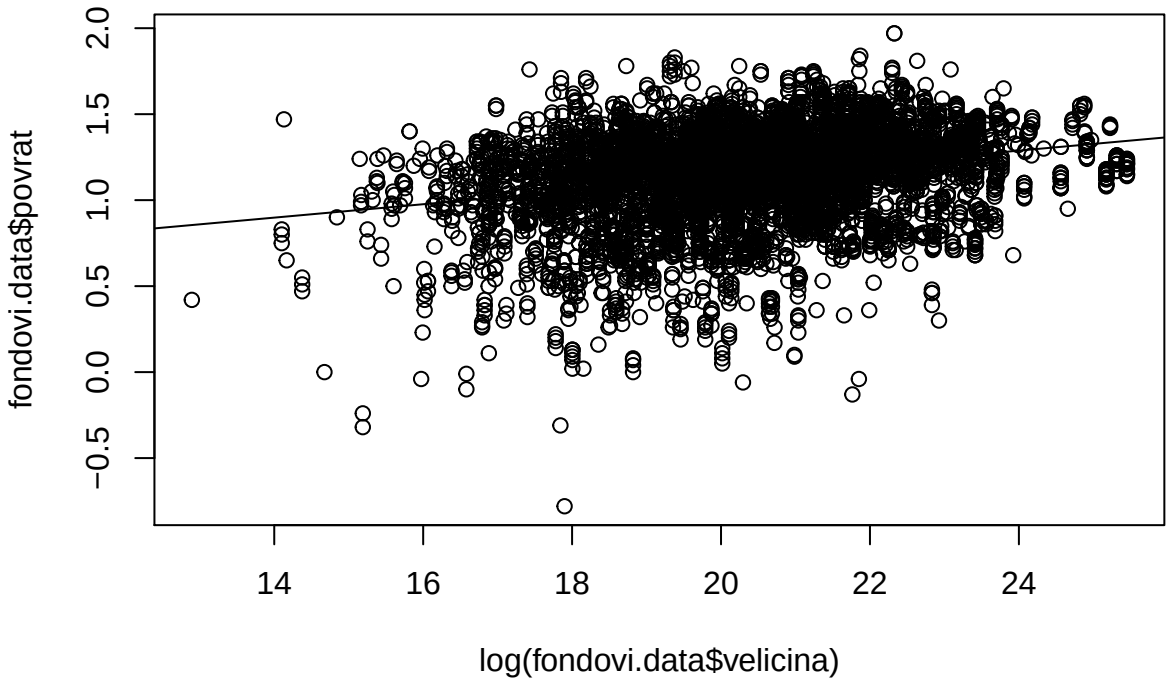
```
fondovi.data = na.omit(fondovi.data)
summary(fondovi.data)
```

```
##     velicina             povrat
##  Min.   :3.977e+05   Min.   :-0.780
##  1st Qu.:2.197e+08   1st Qu.: 1.020
##  Median :7.930e+08   Median : 1.190
##  Mean   :3.497e+09   Mean   : 1.148
##  3rd Qu.:2.410e+09   3rd Qu.: 1.320
##  Max.   :1.131e+11   Max.   : 1.970
```

```
plot(fondovi.data$velicina, fondovi.data$povrat)
```



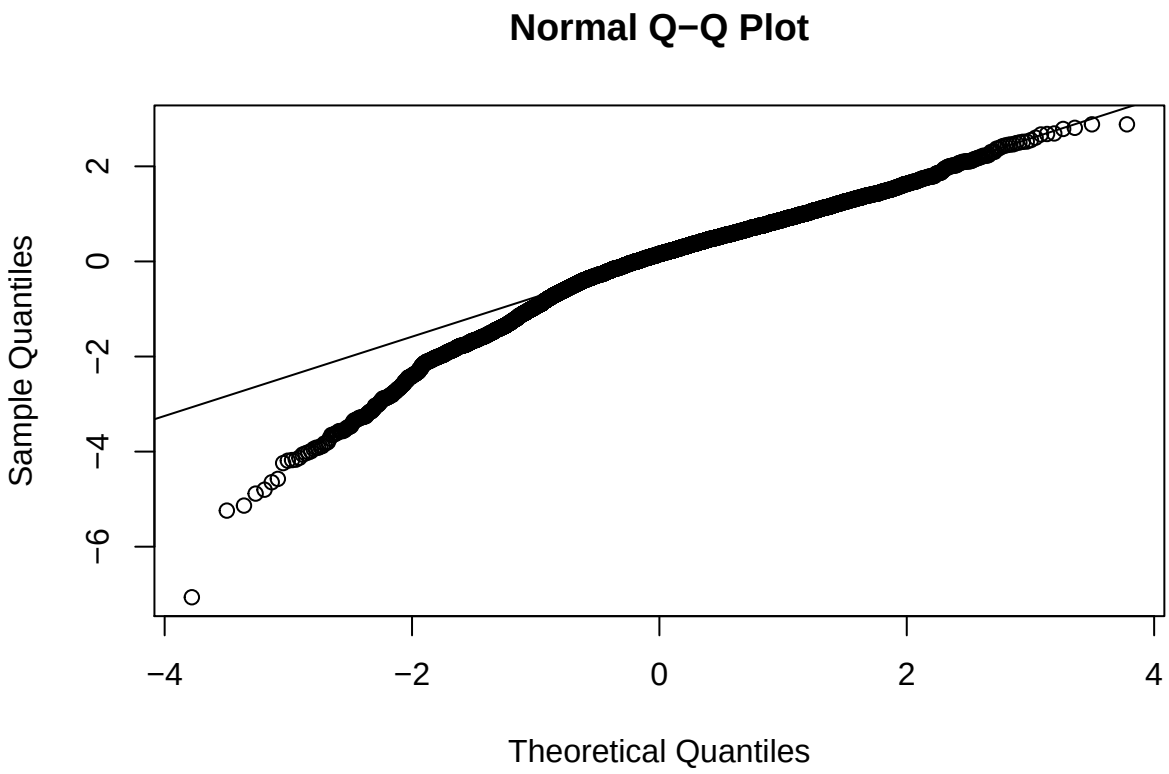

```
log_velicina = log(fondovi.data$velicina)

plot(log(fondovi.data$velicina), fondovi.data$povrat)

fit.velicine = lm(fondovi.data$povrat~log_velicina)

abline(fit.velicine)
```



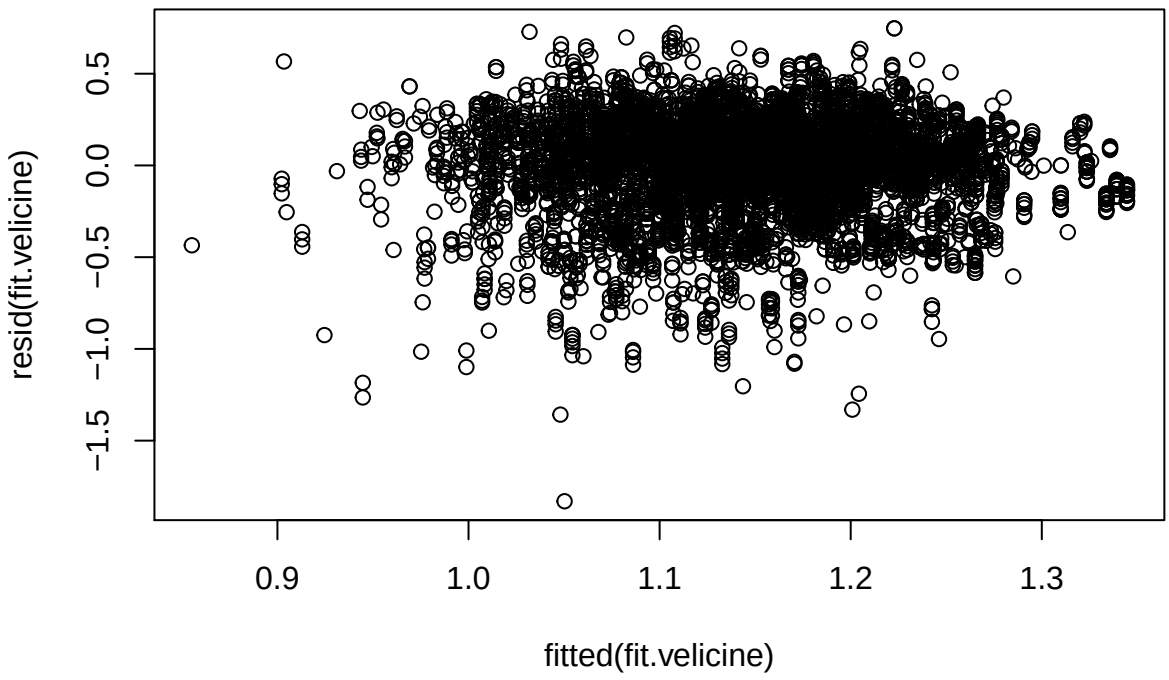

```
summary(fit.velicine)
```

```
##
## Call:
## lm(formula = fondovi.data$povrat ~ log_velicina)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83026 -0.12369  0.04054  0.16779  0.74719
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.352773   0.036848   9.574   <2e-16 ***
## log_velicina 0.038969   0.001798  21.670   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2592 on 6380 degrees of freedom
## Multiple R-squared:  0.06856,    Adjusted R-squared:  0.06841
## F-statistic: 469.6 on 1 and 6380 DF,  p-value: < 2.2e-16
```

```
qqnorm(rstandard(fit.velicine))
qqline(rstandard(fit.velicine))
```




```
plot(fitted(fit.velicine), resid(fit.velicine))
```



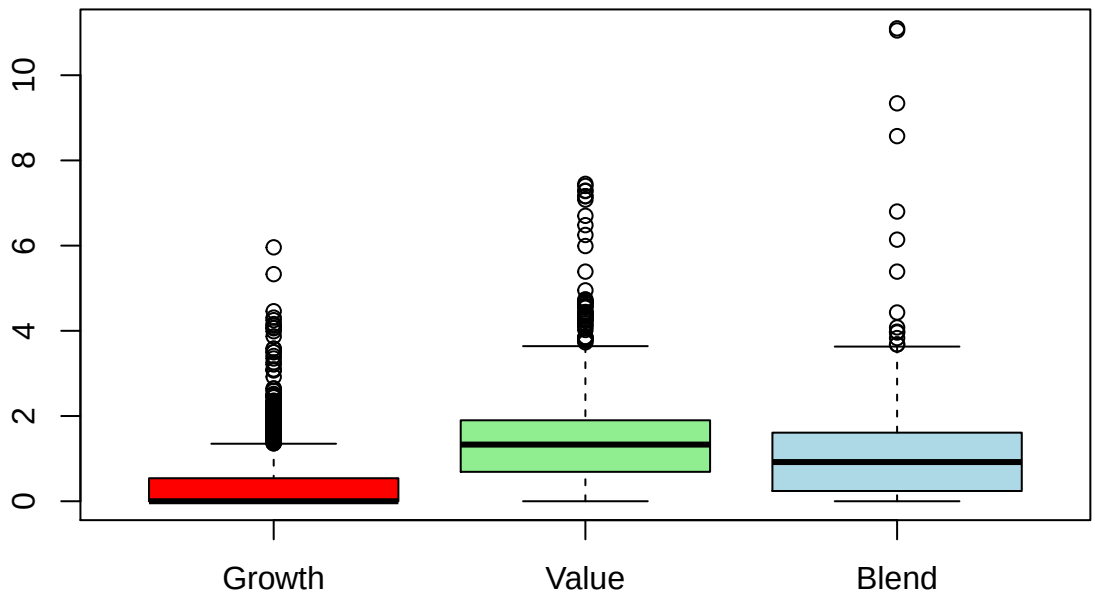

```
matrix_coef <- summary(fit.velicine)$coefficients
my_estimates <- matrix_coef[ , 1]
my_estimates
```

```
##  (Intercept) log_velicina
##   0.35277292   0.03896851
```

povrat = 0.3528 + 0.0389 * log velicina fonda Za 10 puta veću imovinu pod upravljanjem fonda dobiva se 3.89% posto veći povrat.

Analiziraju se dividende ovisno o stilu investiranja fonda.

```
returnsComplete = funds[complete.cases(funds[,c("fund_return_10years","category_return_10years","fund_sl

growthFunds = funds[funds$investment == "Growth",]
valueFunds = funds[funds$investment == "Value",]
blendFunds = funds[funds$investment == "Blend",]

boxplot(growthFunds$fund_yield,
        valueFunds$fund_yield,
        blendFunds$fund_yield, names=c("Growth","Value","Blend"), col=c("Red", "Light green", "Light blu
```



Box plot dividendi po stilu investiranja ukazuje na veliku zakrivljenost kod dividendi Growth fondova. Zbog te činjenice ne možemo koristiti testove koji se oslanjaju na pretpostavku normalnosti. Za analizu jednakosti sredina koristit ćemo Kruskal-Wallisov test umjesto ANOVA testa. Dodatno testiramo razliku dividendi između Value i Blend fondova. Prije samog testiranja razlike provodimo Kolmogorov-Smirnovljev test kako bismo utvrdili možemo li za testiranje razlike koristiti t-test ili moramo koristiti neki od neparametarskih testova. Tvrdnja koju testiramo je jesu li dividende value fondova veće od dividendi blend fondova

```
kruskal.test(funds$fund_yield~funds$investment, data=funds);
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  funds$fund_yield by funds$investment
## Kruskal-Wallis chi-squared = 1853, df = 3, p-value < 2.2e-16
```

```
#jedan od fondova ima različitu dividendu

shapiro.test(valueFunds$fund_yield)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  valueFunds$fund_yield
## W = 0.9045, p-value < 2.2e-16
```

```
shapiro.test(blendFunds$fund_yield)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  blendFunds$fund_yield
## W = 0.8325, p-value < 2.2e-16
```

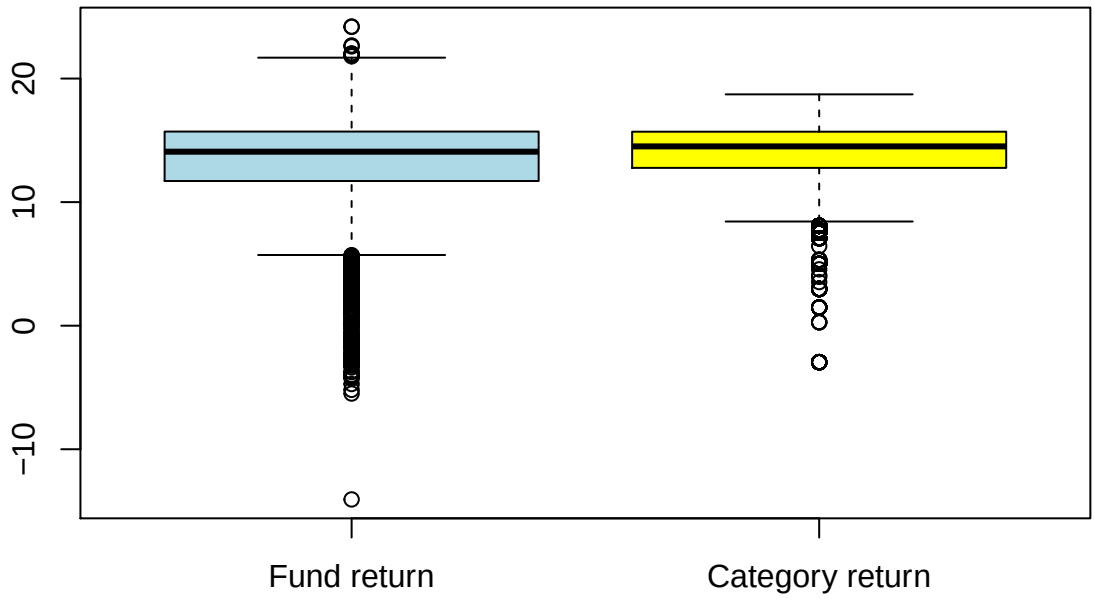
```
wilcox.test(valueFunds$fund_yield, blendFunds$fund_yield, paired = FALSE, var.equal = FALSE, alternativ
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  valueFunds$fund_yield and blendFunds$fund_yield
## W = 1916463, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

Kolmogorov-Smirnovljev testovi ukauzju na činjenicu da fondovi nemaju normalnu razdiobu te se stoga koristi neparametarski Wilcoxonov test predznačenih rangova. Navedeni test daje zaključak da value fondovi imaju veće dividende od blend fondova.

Za istraživanje tvrdnje da fondovi pobjeđuju svoje kategorije, analizirali smo povrate fondova i njihovih kategorija nakon 10 godina. Boxplot za povrate fondova i kategorija mogao bi ukazivati na činjenicu da su povrati fondova i kategorija jednaki s razlikom u većoj varijaciji kod fondova.

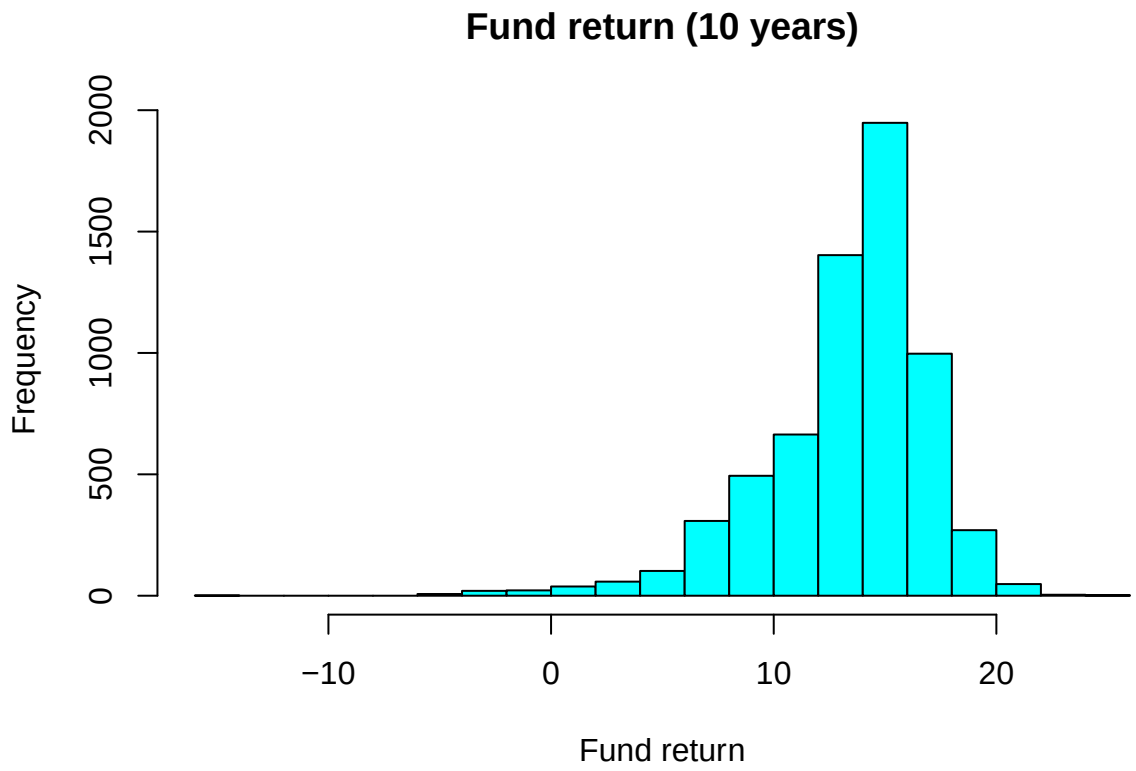
```
returns_10years = funds[c("fund_return_10years","category_return_10years")]
ind = which(!is.na(returns_10years$fund_return_10years) & !is.na(returns_10years$category_return_10years
returns_10years = returns_10years[ind,]

boxplot(funds$fund_return_10years, funds$category_return_10years, names=c("Fund return","Category retur
```



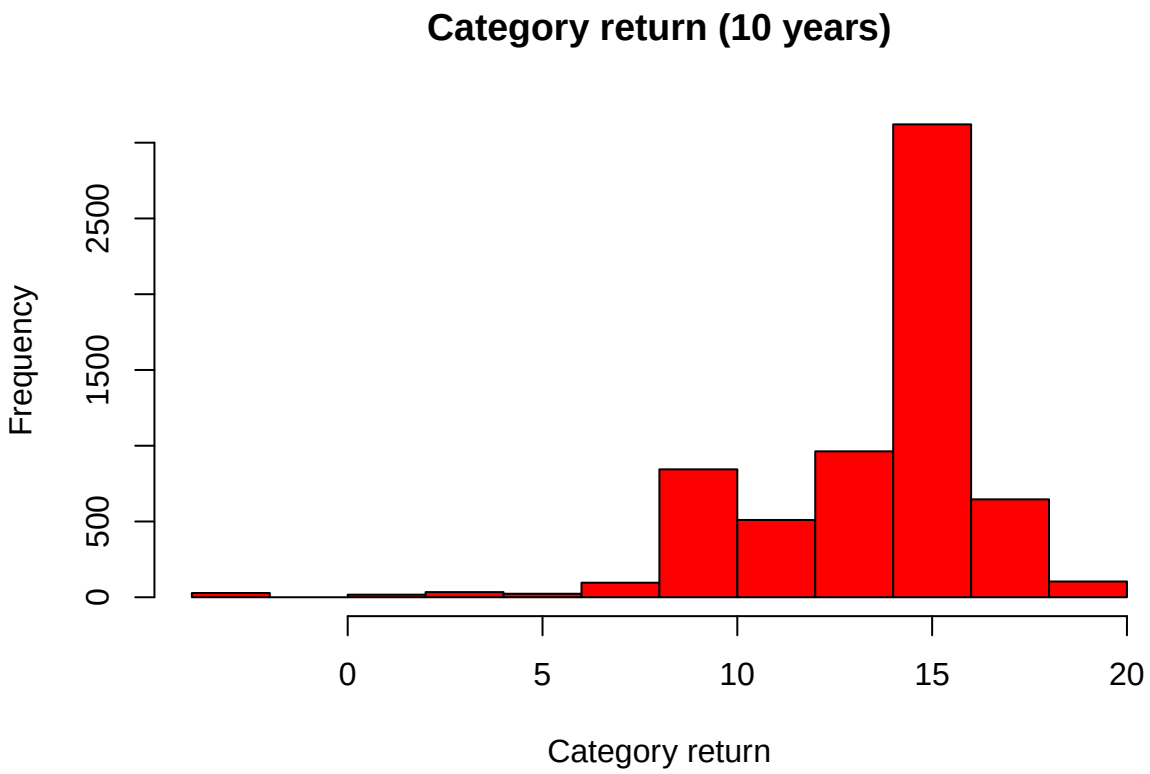
Analizom histograma povrata fondova i kategorija uočavaju se blage zakrivljenosti podataka, ali u histogramu

razlike povrata fonda i povrata kategorije ne uočava se veća zakrivljenost podataka. Za analizu razlike sredina koristit ćemo upareni t-test jer svaki fond ima pridruženu odgovarajuću kategoriju. T-test je robustan na manje zakrivljenosti u podacima te je bolja opcija u odnosu na neparametarske testove (pretežno zbog svoje veličine) kod velikih uzoraka. Nulta hipoteza testa je da fondovi imaju manji ili jednak povrat od svojih kategorija. Alternativna hipoteza testa je da fondovi imaju veći povrat od svojih kategorija.

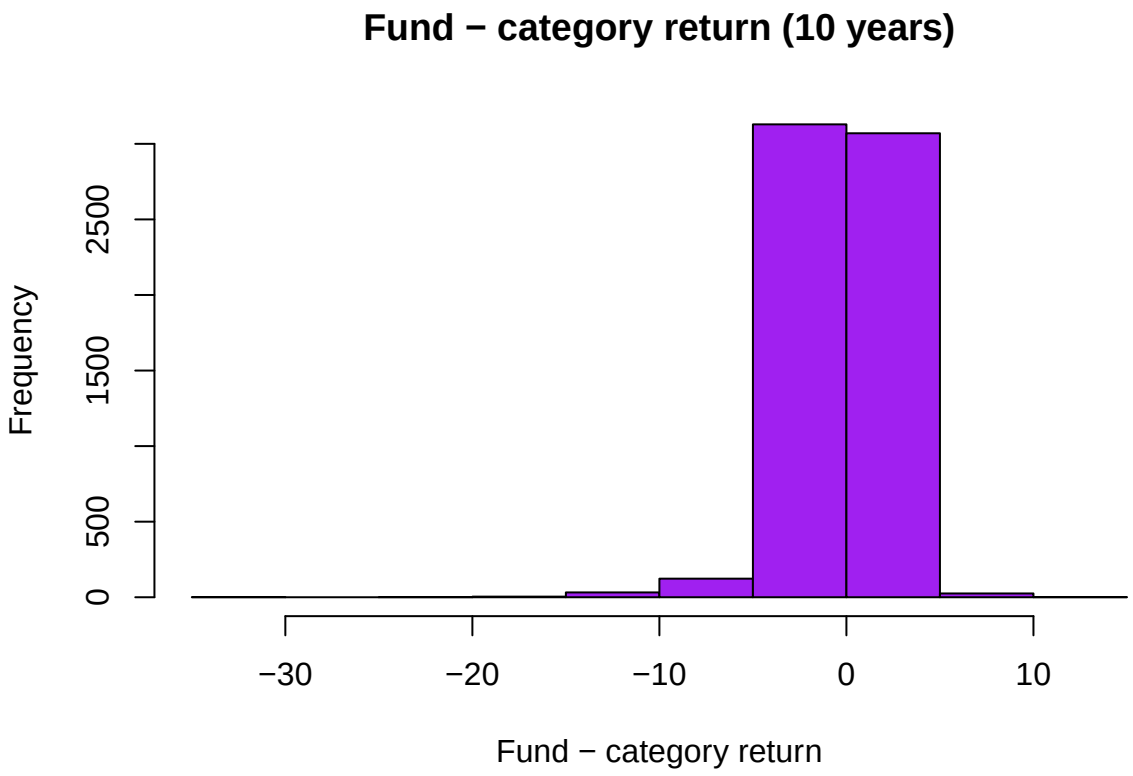
```
hist(returns_10years$fund_return_10years, main="Fund return (10 years)"
     ,col="cyan"
     ,xlab = "Fund return")
```




```
hist(returns_10years$category_return_10years, main="Category return (10 years)", col="red", xlab = "Cat
```




```
hist(returns_10years$fund_return_10years - returns_10years$category_return_10years, main="Fund - categ
```




```
print("Wilcoxonov test nad razlikom povrata fonda i povrata kategorije u razdoblju od 10 godina.")
```

```
## [1] "Wilcoxonov test nad razlikom povrata fonda i povrata kategorije u razdoblju od 10 godina."
```

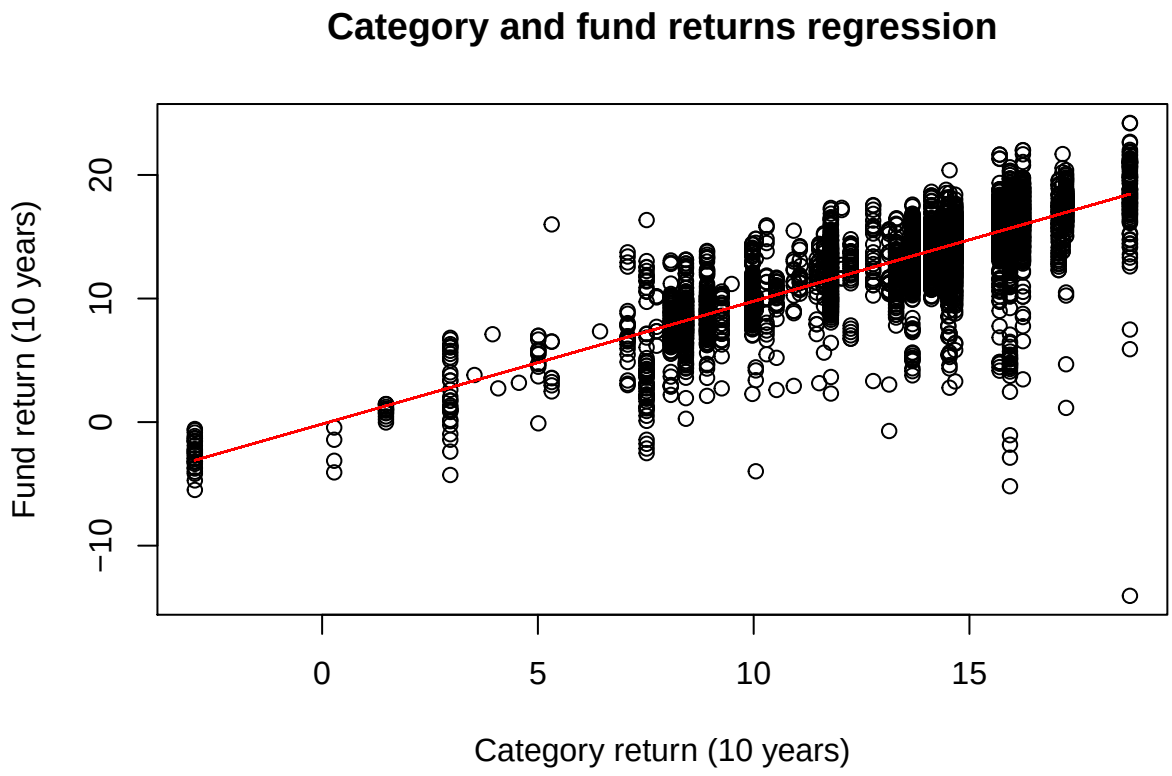
```
t.test(x = funds$fund_return_10years,
            y = funds$category_return_10years,
            paired = TRUE,
            alternative = "greater",
            conf.level = 0.99)
```

```
##
##  Paired t-test
##
## data:  funds$fund_return_10years and funds$category_return_10years
## t = -8.8946, df = 6385, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  -0.3032769        Inf
## sample estimates:
## mean of the differences
##              -0.2403883
```

Zaključak testiranja je da ne možemo odbaciti tvrdnju da fondovi imaju manji ili jednak povrat od svojih kategorija, odnosno nismo uspijeli dokazati da fondovi imaju veći povrat od svojih kategorija.

Linearnom regresijom provjerit će se ovisi li povrat fonda o povratu kategorije. Drugim riječima analizira se prate li fondovi svoje kategorije u smislu povrata.

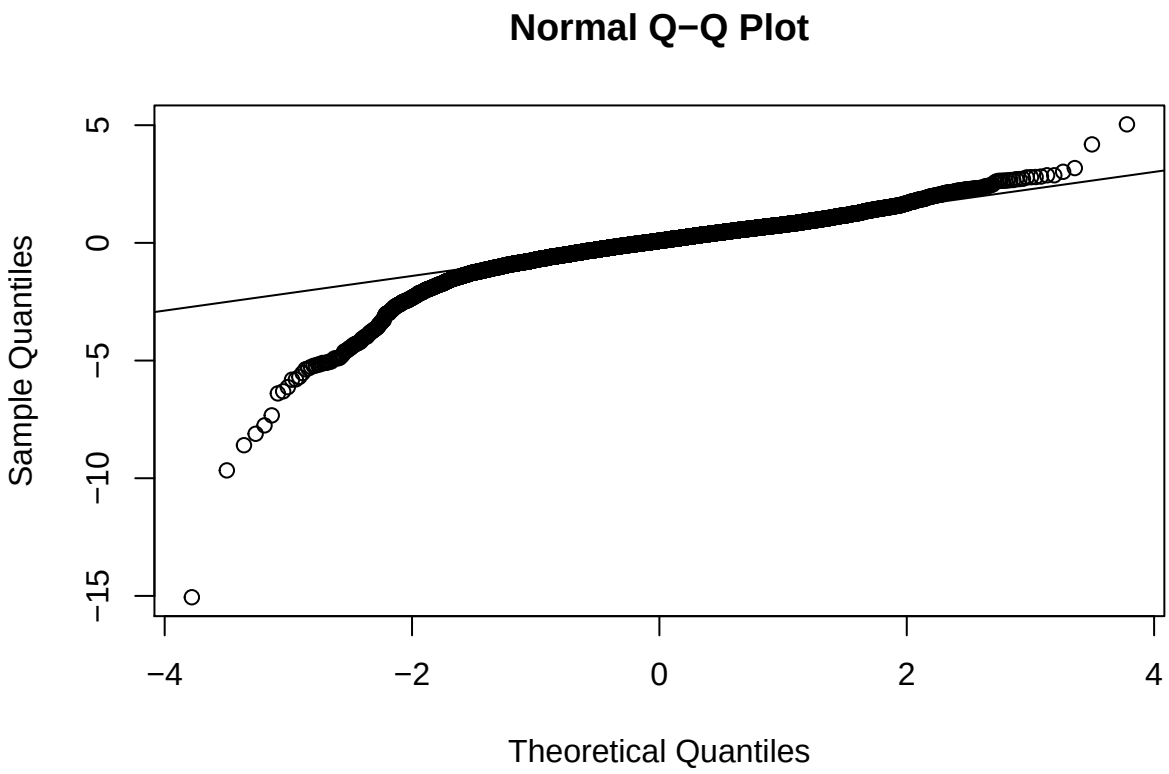
```
returnsComplete = funds[complete.cases(c("fund_return_10years","category_return_10years")),]
x = returnsComplete$category_return_10years
y = returnsComplete$fund_return_10years

fit.return = lm(y~x, data=returnsComplete)
plot(x, y, main = "Category and fund returns regression", xlab="Category return (10 years)", ylab="Fund
lines(x, fit.return$fitted.values, col="red")
```



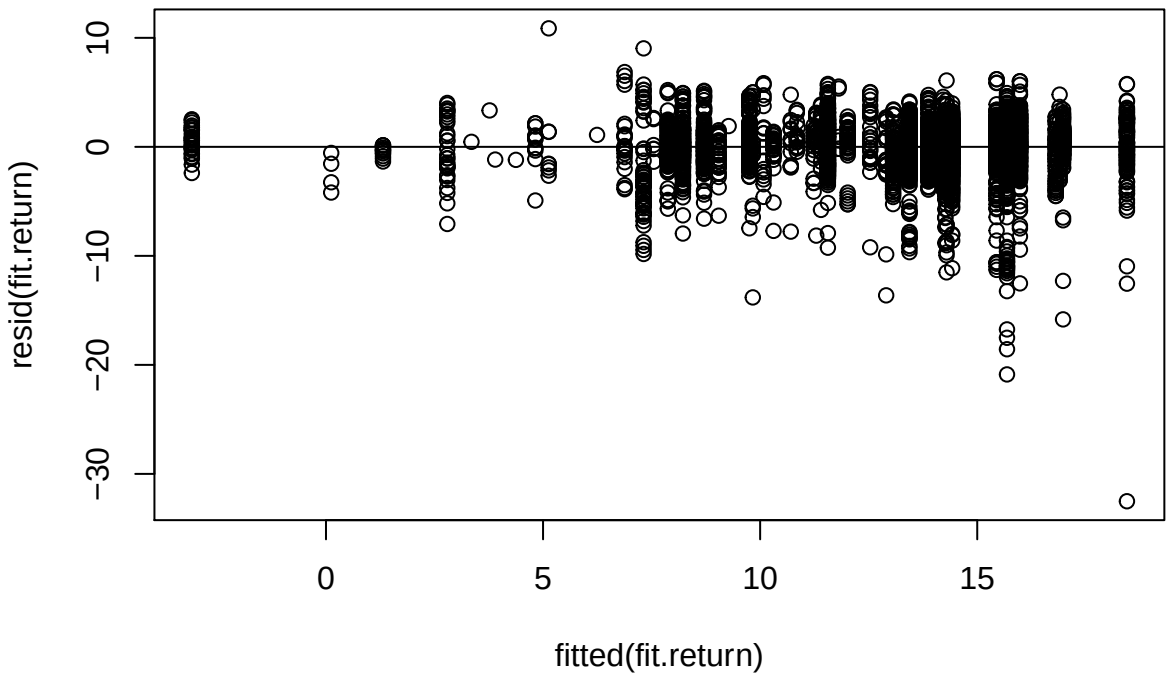
Uočava se potencijalna zavisnost povrata fonda o povratu kategorije te je potrebno provesti analizu reziduala. Također će se odrediti Pearsonov koeficijent korelacije koji će predstalvljati jačinu linearne veze, kao i koeficijent determinacije koji određuje kvalitetu modela

Analiza reziduala:

```
qqnorm(rstandard(fit.return))
qqline(rstandard(fit.return))
```




```
plot(fitted(fit.return), resid(fit.return))
abline(0,0)
```




```
#c("Koeficijent determinacije: ",rsq(x,y))
c("Pearsonov koeficijent korelacije:", cor(x,y,method="pearson"))
```

```
## [1] "Pearsonov koeficijent korelacije:" "0.808657328513655"
```

```
summary(fit.return)
```

```
##
## Call:
## lm(formula = y ~ x, data = returnsComplete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.508  -0.923   0.189   1.223  10.870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.157373   0.125952  -1.249    0.212
## x            0.993893   0.009049 109.831   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.16 on 6384 degrees of freedom
## Multiple R-squared:  0.6539, Adjusted R-squared:  0.6539
## F-statistic: 1.206e+04 on 1 and 6384 DF,  p-value: < 2.2e-16
```

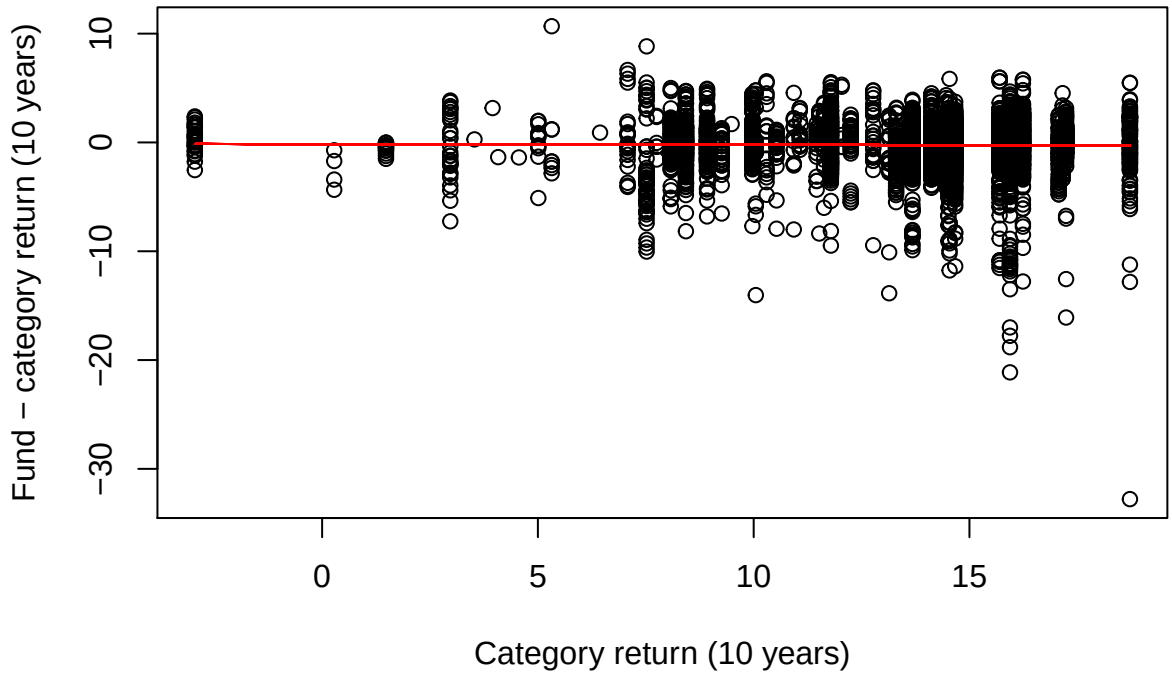
```
coefficients(fit.return)
```

```
## (Intercept)           x
##  -0.1573729   0.9938934
```

Pearsonov koeficijent i koeficijent determinacije ukazuju na jaku linearnu vezu između povrata kategorije i povrata fonda. Reziduali ne pokazuju nikakvu vezu što je nužno za provođenje zaključaka o linearnog regresiji. Zaključuje se da postoji jaka linearna veza između povrata fondova i povrata njihovih kategorija. Procjenjeni koeficijenti iznose: b1 = 0.993893 i b0 = -0.157373. Ovom regresijom ne možemo donositi zaključke pobjeđuje li fond svoju kategoriju, već samo donosimo zaključak da fondovi prate svoje kategorije u smislu povrata.

Za provjere hipoteze o pobjeđivanju fondova analizirati će se zavisnost razlike povrata fonda i povrata kategorije o povratu kateogrije.

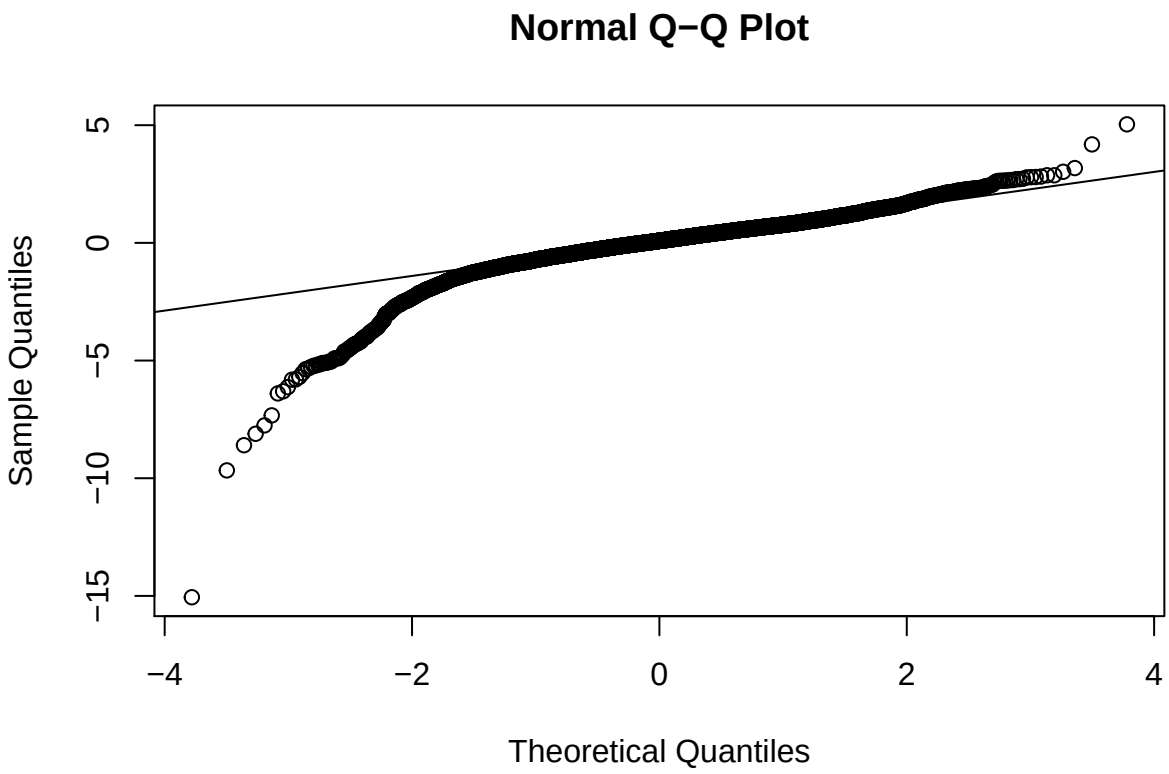
```
returnsComplete = funds[complete.cases(c("fund_return_10years","category_return_10years")),]
x = returnsComplete$category_return_10years
y = returnsComplete$fund_return_10years - returnsComplete$category_return_10years

fit.return = lm(y~x, data=returnsComplete)
plot(x, y, xlab="Category return (10 years)", ylab="Fund - category return (10 years)")
lines(x, fit.return$fitted.values, col="red")
```



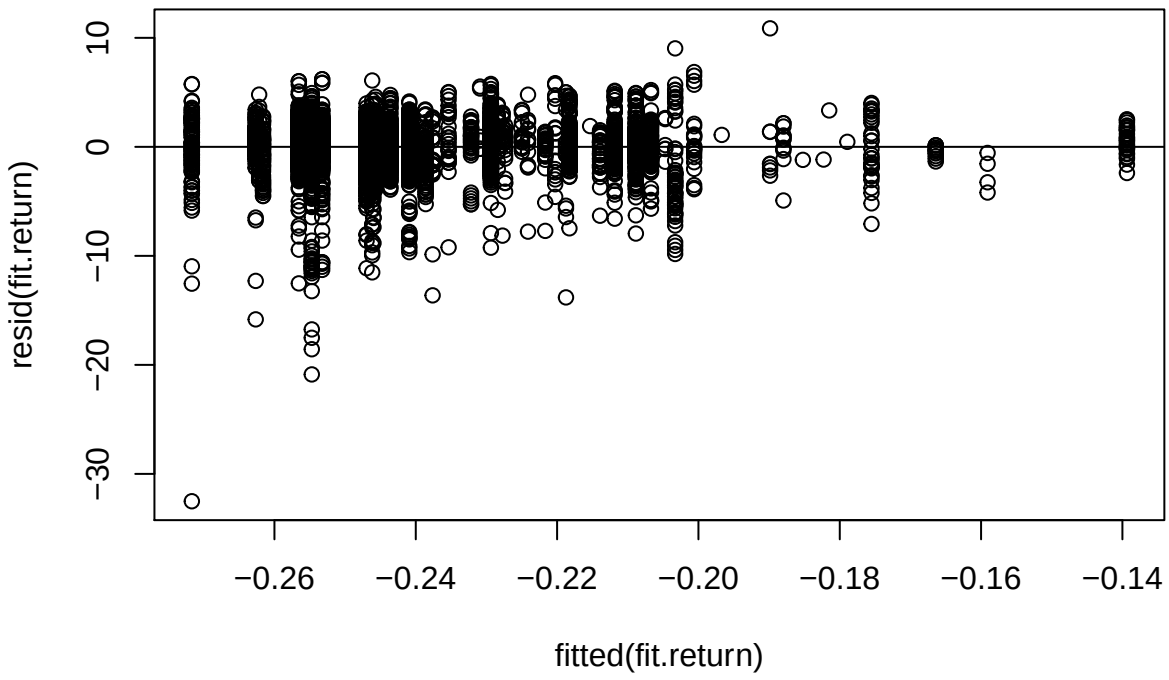
Uočava se vrlo slaba linearna zavisnost između razlike povrata fonda i kategorije te same kategorije.

Analiza reziduala:

```
qqnorm(rstandard(fit.return))
qqline(rstandard(fit.return))
```




```
plot(fitted(fit.return), resid(fit.return))
abline(0,0)
```




```
#c("Koeficijent determinacije: ",rsq(x,y))
c("Pearsonov koeficijent korelacije:", cor(x,y,method="pearson"))
```

```
## [1] "Pearsonov koeficijent korelacije:" "-0.00844555250251895"
```

```
summary(fit.return)
```

```
##
## Call:
## lm(formula = y ~ x, data = returnsComplete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.508  -0.923   0.189   1.223  10.870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.157373   0.125952  -1.249    0.212
## x           -0.006107   0.009049  -0.675    0.500
##
## Residual standard error: 2.16 on 6384 degrees of freedom
## Multiple R-squared:  7.133e-05,  Adjusted R-squared:  -8.53e-05
## F-statistic: 0.4554 on 1 and 6384 DF,  p-value: 0.4998
```

```
coefficients(fit.return)
```

```
##  (Intercept)            x
## -0.157372861 -0.006106648
```

Pearsonov koeficijent korelacije te koeficijent determinacije svojim niskim vrijednostima ukazuju na gotovo nikakvu linearnu zavisnost između razlike povrata fonda i kategorije i povrata kategorije. Koeficijenti procijenjenog pravca su isto približno jednaki nuli što ukazuje na činjenicu da će razlika povrata fonda i kategorije unutar svake kategorije ravnomjerno varirati, neovisno o povratu kategorije.

#dodati analizu koreliranosti varijabli

Višestrukom regresijom pokušati ćemo pronaći kauzalnu vezu između razlike povrata fonda i kategorije i drugih pokazatelja fonda. Parametri koji bi mogli utjecati na razliku povrata fonda i povrata kategorije su: ukupna imovina pod upravljanje (pretpostavljamo da fondovi s većom imovinom pod upravljanje imaju veći vjerojatnost pobijediti svoje kategorije), razlika godišnjeg troška upravljanja fonda i kategorije (fondovi koji uzimaju veći postotak za upravljanje fondom imaju veću vjerojatnost pobijediti svoju kategoiju), medijalna tržišna kapitalizacija (fondovi koji ulažu u tvrtke s većom tržišnom kapitalizacijom će vjerojatnije pobijediti svoje kategorije).

```
returnsComplete = funds[complete.cases(funds[,c("fund_return_10years","category_return_10years","fund_sl

x1 = log(returnsComplete$net_assets)
x2 = returnsComplete$net_annual_expense_ratio_fund - returnsComplete$net_annual_expense_ratio_category
x3 = returnsComplete$median_market_cap

cor(cbind(x1,x2,x3))
```

```
##            x1          x2          x3
## x1  1.0000000 -0.29043724  0.18342861
## x2 -0.2904372  1.00000000 -0.04490357
## x3  0.1834286 -0.04490357  1.00000000
```

Analiziranjem koreliranosti varijabli primjećuje se slaba koreliranost između varijabli

```
y = returnsComplete$fund_return_10years - returnsComplete$category_return_10years


fit.return = lm(y~x1 + x2 + x3, data=returnsComplete)
summary(fit.return)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = returnsComplete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.6037  -0.8129   0.0828   0.9971  11.1332
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.845e+00  2.961e-01 -23.114  < 2e-16 ***
## x1           3.311e-01  1.459e-02  22.695  < 2e-16 ***
## x2          -1.162e+00  5.110e-02 -22.745  < 2e-16 ***
## x3          -2.256e-06  6.608e-07  -3.413 0.000646 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.95 on 6066 degrees of freedom
## Multiple R-squared:  0.1932, Adjusted R-squared:  0.1928
## F-statistic: 484.2 on 3 and 6066 DF,  p-value: < 2.2e-16
```

Uočava se vrlo slaba linearna zavisnost između razlike povrata fonda i kategorije te same samih parametara.