WhatSap

Opis problema

U kontekstu dioničarskih fondova, biranje dionica kojim će fond “pobijediti tržište” složen je posao, po nekim mišljenjima čak i nemoguć. Iz tog razloga, mnoge su analize rađene na tu temu, te je pravilno upravljanje fondovima danas još uvijek aktualno pitanje. Ovo istraživanje jedna je od takvih analiza. Korištenjem statističkih metoda, cilj ovog projekta bio je pokušati dati odgovore na probleme koji uključuju teme poput konzistentne pobijede fondova u svojoj kategoriji te odnosi pojedinih karakteristika fonda s njegovim dugoročnim uspjehom. Naši rezultati i zaključci, kao i korištena metodologija, nalazi se u nastavku ovog rada.

Opis podataka

Prilikom izrade projekta, koristili smo se podatcima o dioničkim fondovima dobivenim iz sljedećeg skupa podataka.

Osim metapodataka o fondovima, kao što su ime, izdavač i kategorija, svaki fond sadrži širok spektar atributa. U projektu nisu korišteni svi podatci, već se istraživanje fokusira na par ključnih podataka i njihovom povezanošću. Jedan od ključnih podataka vezan za istraživanje je povrat fonda. Budući da je uz skup podataka vezan cijeli niz atributa vezanih uz povrat, mi smo kao mjeru povrata koristili atribut “fund_return_10years”, tj. povrat fonda u zadnjih 10 godina bez uzimanja u obzir rizik fonda. Stoga to treba uzeti u obzir prilikom daljnje analize rada.

Strategije investiranja

Growth, value i blend su tri stila investiranja koje ćemo usporediti u ovom projektu. Growth fondovi ulažu u dionice firmi za koje se pretpostavlja veliki rast. Value fondovi investiraju u dionice čija je tržišna vrijednost ispod stvarne, odnosno u podcijenjene firme. Blend investiranje je investiranje u različite dionice iz istog razreda, npr. S&P500. Možemo ih smatrati kombinacijom value i growth investiranja.

```
## [1] "Growth"
```

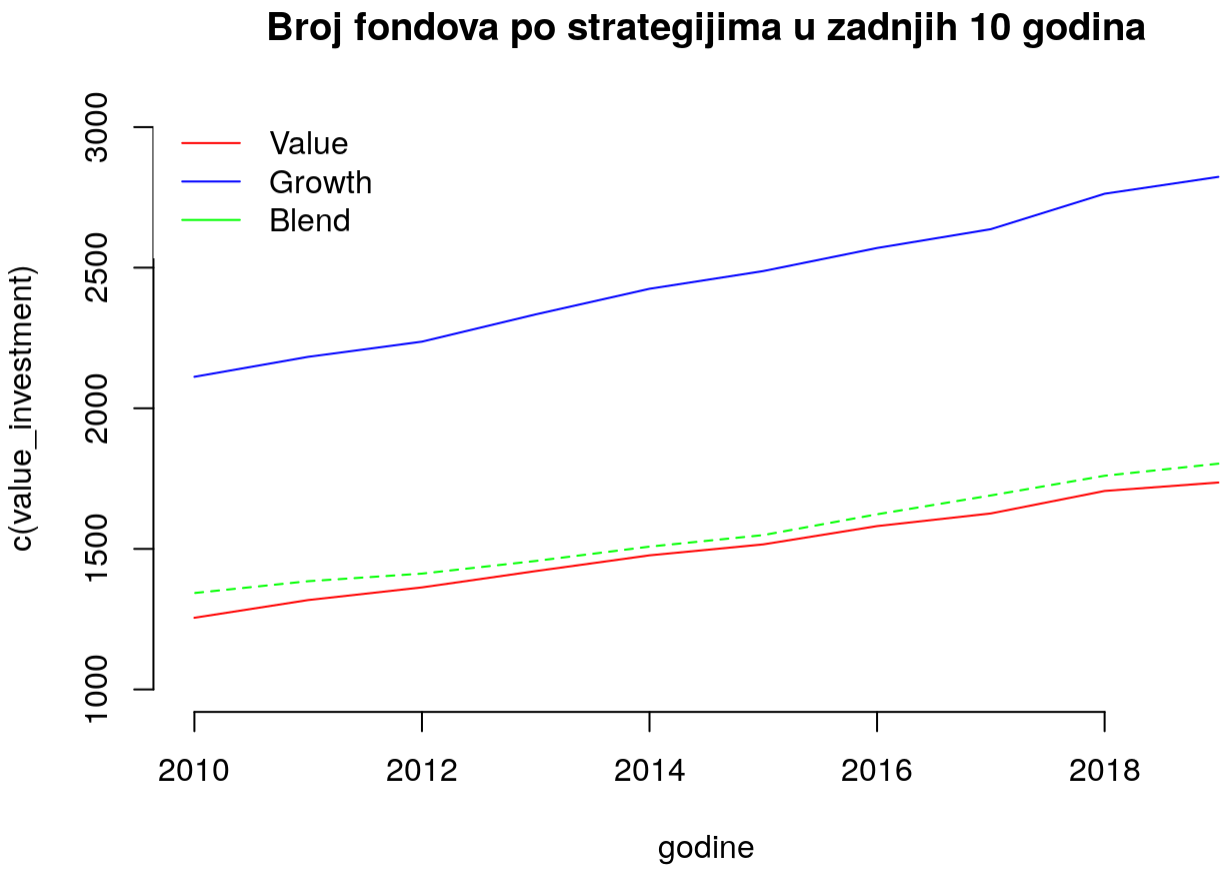
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -7.26   13.33   15.37   15.57   17.62   45.78
```

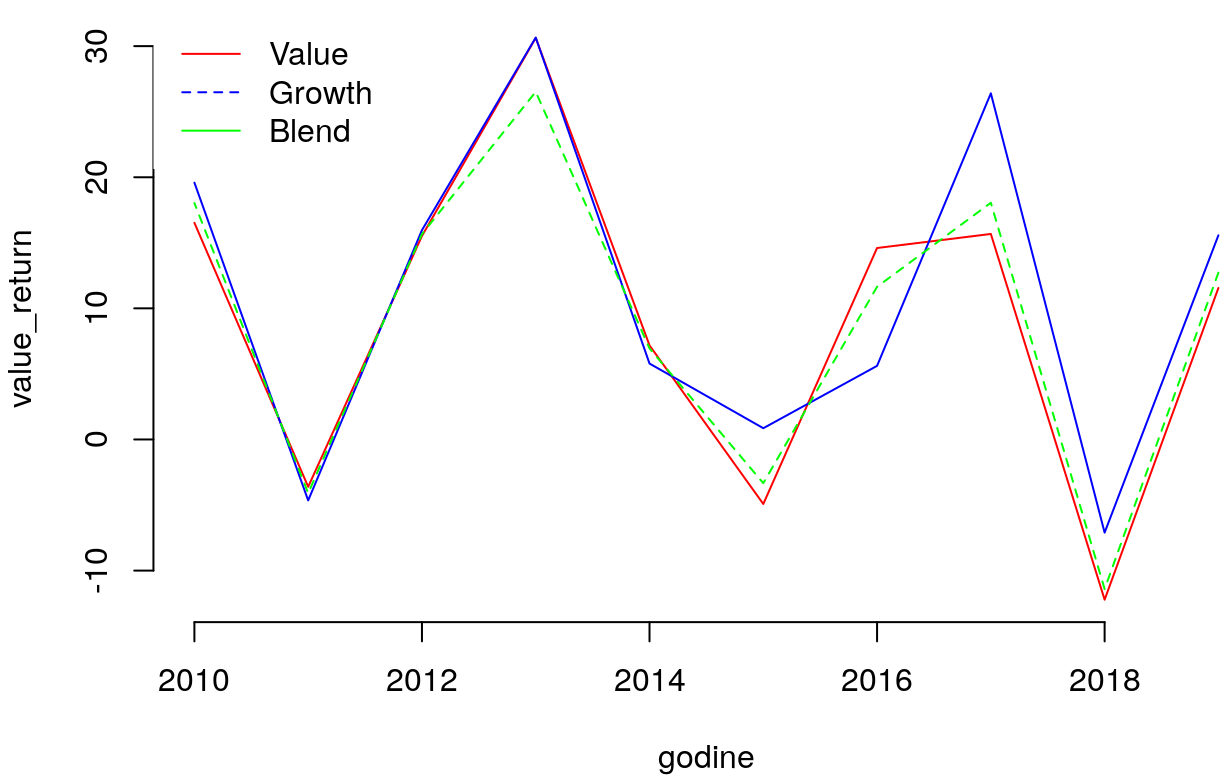
```
## [1] "Value"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -8.94   10.19   11.62   11.55   12.93   21.07
```

```
## [1] "Blend"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -13.65   11.12   12.94   12.72   14.38   41.33       1
```



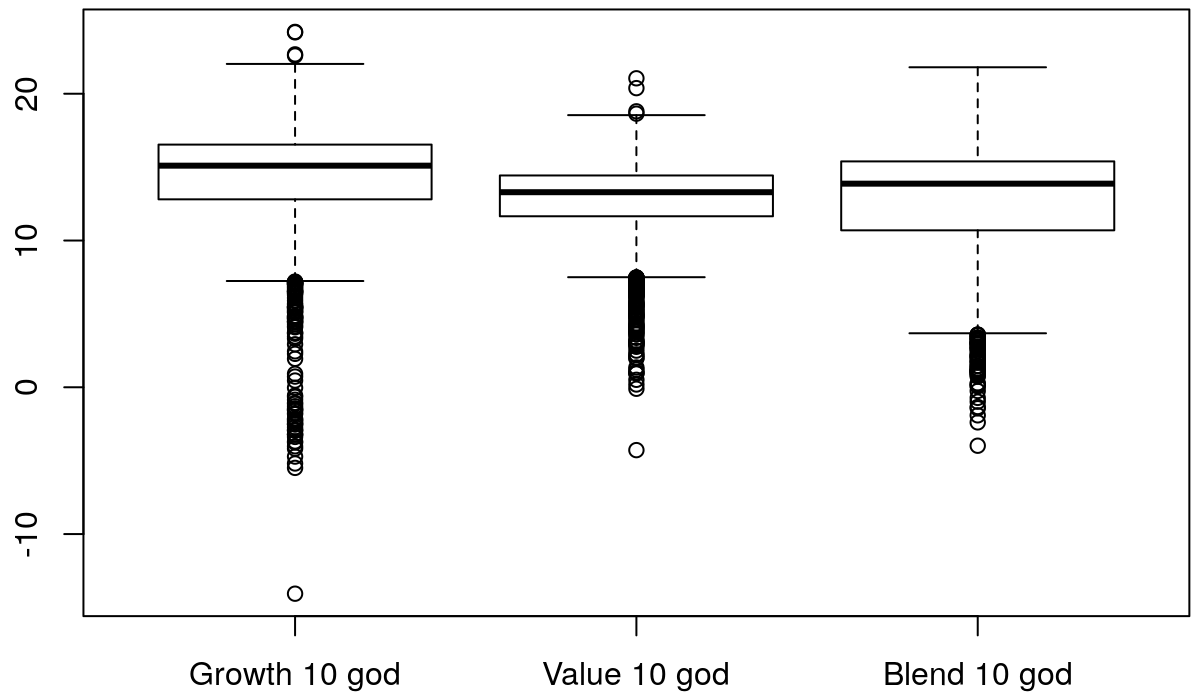


Kao što vidimo, Growth je najučestalija strategija investiranja fondova iz ovog data-seta te možemo vidjeti iz grafa da je za posljednje 3 godine Growth imao najveću vrijednost srednjeg godišnjeg povrata.

Kako bih saznali koji stil investiranja daje najbolje rezultate testirati ćemo njihove povrate na temelju podataka 1-godišnjeg, 3-godišnjeg, 5-godišnjeg i 10-godišnjeg povrata.

Testiranje jednakosti povrata fondova s obzirom na strategiju investiranja

Prikazivat ćemo grafove i testove samo za 10-godišnje povrate jer nam oni daju najbolji dojam o podacima i kako bih smanjili broj prikaza.



Prvo testiramo jednakost povrata strategija investiranja Kruskal-Wallisovim testom. Ovaj neparametarski test koristimo umjesto ANOVA-e jer podaci nisu zadovoljili pretpostavku normalnosti.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fund_return_1year by investment
## Kruskal-Wallis chi-squared = 690.13, df = 2, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fund_return_3years by investment
## Kruskal-Wallis chi-squared = 1476.3, df = 2, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fund_return_5years by investment
## Kruskal-Wallis chi-squared = 1009.7, df = 2, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fund_return_10years by investment
## Kruskal-Wallis chi-squared = 524.83, df = 2, p-value < 2.2e-16
```

Zaključujemo da postoji razlika u povratu fondova s obzirom na stil investiranja koji koriste, međutim ne znamo koji se točno stilovi međusobno razlikuju.

Sada ćemo testirati koji stilovi investiranja se zapravo razlikuju i koji ima najveći povrat.

Koristimo neparametarski Mann-Whitney-Wilcoxov test jer su pretpostavke normalnosti narušene.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  growth_10years and blend_10years
## W = 3221682, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  growth_10years and value_10years
## W = 3386172, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

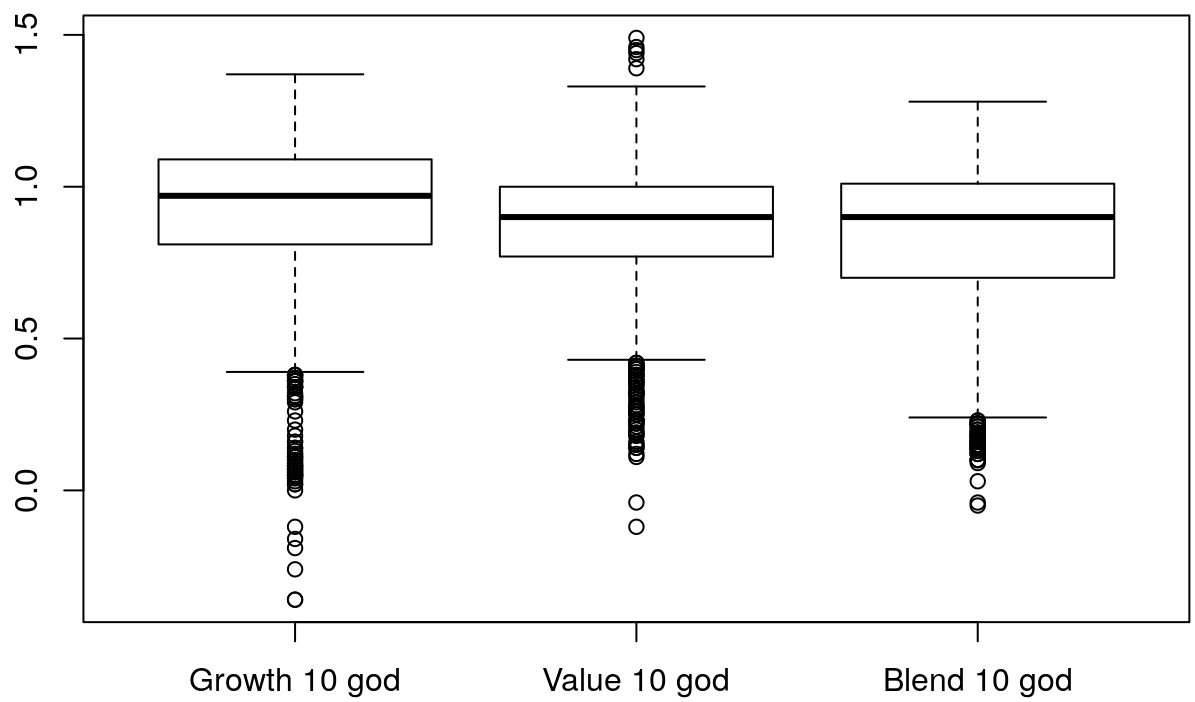
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  blend_10years and value_10years
## W = 1743611, p-value = 2.506e-09
## alternative hypothesis: true location shift is greater than 0
```

Na temelju testova možemo zaključiti sljedeće: Growth investiranje daje najbolje povrate za 1, 3, 5 i 10-godišnje podatke. Sljedeći je Bland način investiranja,te naposljetku Value.

Povrat fondova ovisno o stilu investiranja kada uzmemo rizik u obzir

Utvrdili smo koji stil investiranja je najbolji ako uzmemo u obzir samo povrat fondova. Međutim, tada zanemarujemo rizik koji određeni stil investiranja donosi sa sobom. Sada ćemo testirati povrat fonda ako uzmemo u obzir i rizik. Varijabla koju ćemo koristiti je sharp ratio jer ona dosta dobro prikazuje povrat fonda naspram rizika. Sharp ratio je omjer povrata fonda i standardne devijacije.

Podaci na kojima ćemo provoditi testove su sharp ratio za 3, 5 i 10 godina. Ovdje ćemo također prikazivati grafičke prikaze samo za 10-godišnje podatke.



Prvo testiramo jednakost sharp ratio-a fondova Kruskal-Wallis-ovim testom, ovaj test koristimo jer je narušena pretpostavka normalnosti.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fund_sharpe_ratio_3years by investment
## Kruskal-Wallis chi-squared = 1070.4, df = 2, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fund_sharpe_ratio_5years by investment
## Kruskal-Wallis chi-squared = 632.94, df = 2, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fund_sharpe_ratio_10years by investment
## Kruskal-Wallis chi-squared = 185.02, df = 2, p-value < 2.2e-16
```

Zaključujemo da postoji razlika u sharp ratio-u fondova s obzirom na stil investiranja koji koriste, međutim ne znamo koje se točno srednje vrijednosti međusobno razlikuju.

Sada ćemo testirati koji stil investiranja ima najbolji sharp ratio, te kako se oni međusobno razlikuju.

Koristimo Mann-Whitney-Wilcoxov test jer su pretpostavke normalnosti narušene.

```
## [1] "Growth Blend 10"
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  growth_10years_sharp and blend_10years_sharp
## W = 3041026, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

```
## [1] "Growth Value 10"
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  growth_10years_sharp and value_10years_sharp
## W = 2940494, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

```
## [1] "Blend Value 10"
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  blend_10years_sharp and value_10years_sharp
## W = 1559422, p-value = 0.5612
## alternative hypothesis: true location shift is greater than 0
```

Na temelju p-vrijednosti danih testova zaključujemo sljedeće:

Growth je ostao najbolji stil investiranja kada gledamo sharp ratio fondova, a jedina razlika je to što je 10-godišnji sharp ratio za Value i Blend strategiju jednak

Je li veličina firmi u koje fond investira neovisna o tome koji stil investiranja fond koristi

Sada kad smo ustanovili koji stil investiranja je najbolji, želimo vidjeti postoji li zavisnost između veličine firmi u koje fond investira i stilu investiranja koji koristi fond.

Veličine firme koje se pojavljuju u podacima su sljedeće:

```
##
## <undefined>       Large      Medium       Small
##          23        4017        1439         907
```

Undefined podatke nećemo koristiti, te nam onda preostaje 3 varijable: Large, Medium i Small

Kopiranje podataka kako ne bi promijenili prave vrijednosti

Napravimo kontingencijsku tablicu i dodamo sume redaka i stupaca.

```
##
##          Large Medium Small  Sum
##   Blend   1058    409   337 1804
##   Growth  1804    685   334 2823
##   Value   1155    345   236 1736
##   Sum     4017   1439   907 6363
```

Provjerimo jesu li zadovoljene pretpostavke testa, a pretpostavke su da očekivana frekvencija pojedinog razreda mora biti veća ili jednaka 5.

```
## Očekivane frekvencije za razred  Large - Blend :  1138.876
## Očekivane frekvencije za razred  Large - Growth :  1782.177
## Očekivane frekvencije za razred  Large - Value :  1095.947
## Očekivane frekvencije za razred  Medium - Blend :  407.9767
## Očekivane frekvencije za razred  Medium - Growth :  638.4248
## Očekivane frekvencije za razred  Medium - Value :  392.5985
## Očekivane frekvencije za razred  Small - Blend :  257.1473
## Očekivane frekvencije za razred  Small - Growth :  402.3984
## Očekivane frekvencije za razred  Small - Value :  247.4543
```

Pretpostavke su zadovoljene, pa možemo provesti test.

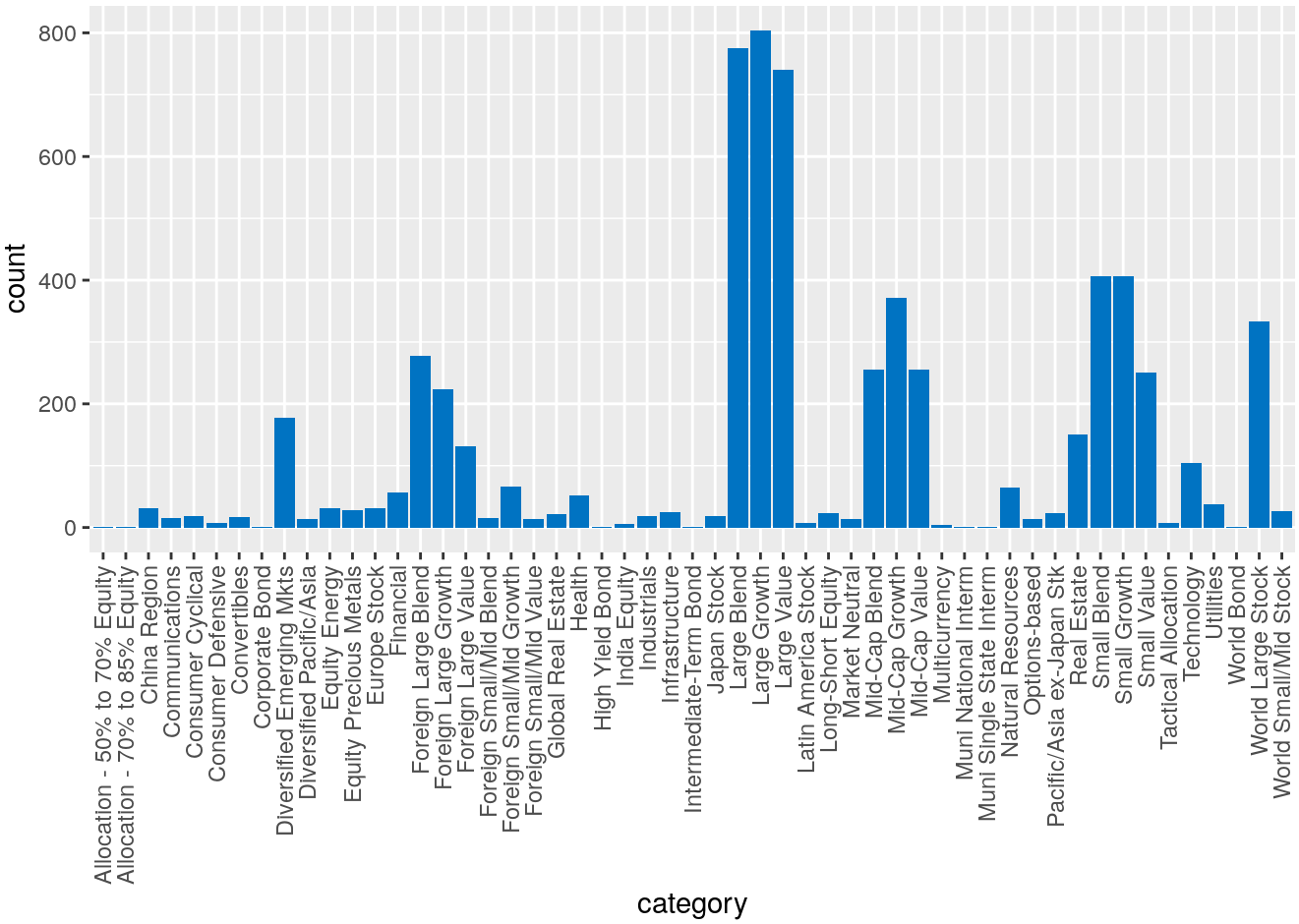
```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 55.317, df = 4, p-value = 2.788e-11
```

Na temelju male p-vrijednosti zaključujemo da je veličina firmi u koju fond ulaže zavisna o stilu investiranja koje fond koristi.

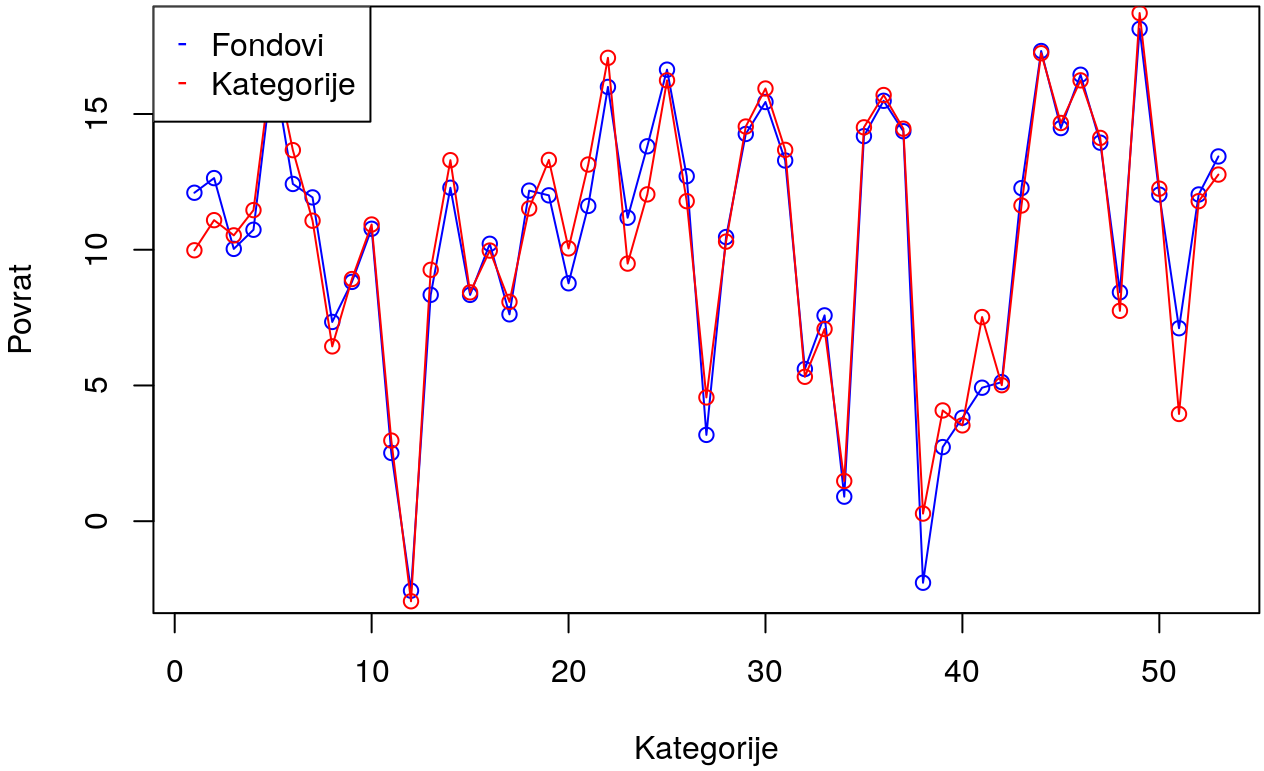
Kategorije fondova

U nastavku ćemo promatrati kategorije fondova. Pogledat ćemo uspijevaju li fondovi pobijediti kategoriju u smislu povrata. Nakon toga ćemo podijeliti dataset u dva dijela. Za prvi dio ćemo uzeti fondove koji uspijevaju pobijediti svoju kategoriju, a za drugi dio one koji ne uspijevaju.

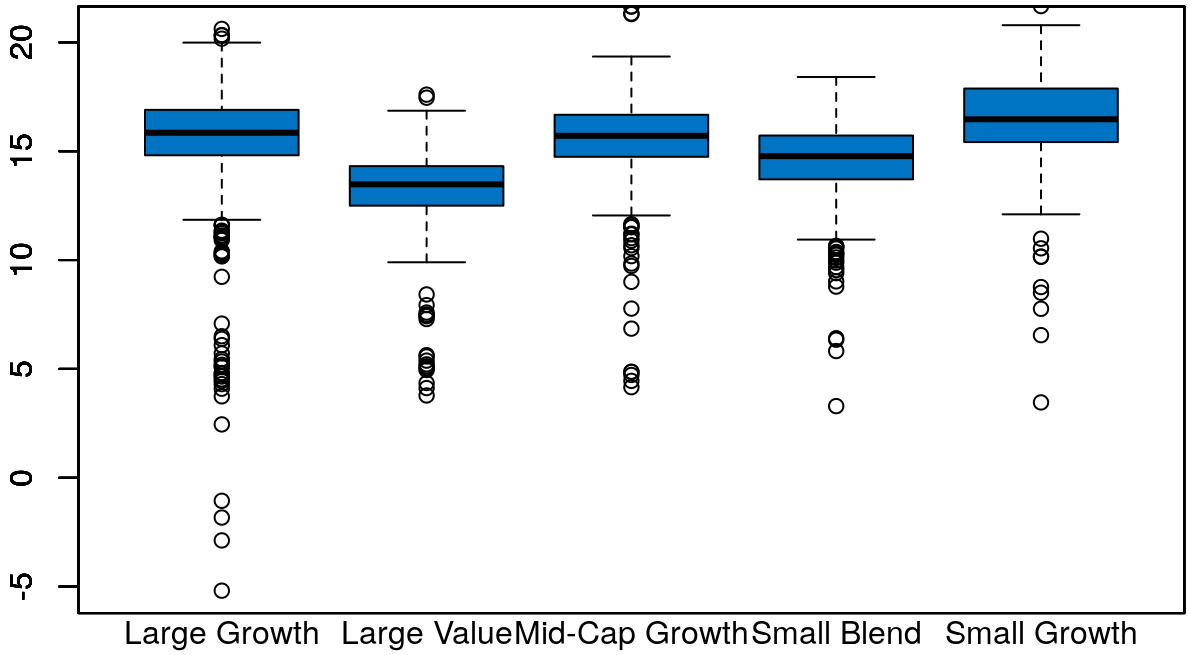
Pogledajmo grafički prikaz broja fondova u svakoj kategoriji.



Pogledajmo prosječnu vrijednost povrata fondova naspram kategorije kojoj oni pripadaju za povrat od 10 godina.



Promotrimo kako se ponašaju podaci unutar nekih od kategorija. Kako bi imali što bolju reprezentaciju, promatrat ćemo kategorije s najviše fondova “Large Growth”, “Large Value”, “Mid-Cap Growth”, “Small Blend” i “Small Growth”. Pogledajmo njihove boxplotove i distribuciju:



Pogledajmo u kojim kategorijama fondovi najviše pobjeđuju. Iskoristit ćemo t-test za svaku kategoriju i prikazati postotak fondova koji su prošli test. Gledamo je li stvarna srednja vrijednost povrata fondova u 10 godina drugačija od povrata kategorije u 10 godina. Pretpostavka H_0 je da fondovi ne pobjeđuju kategorije, tj. mean(fund_return_10years) < category_return_10years.

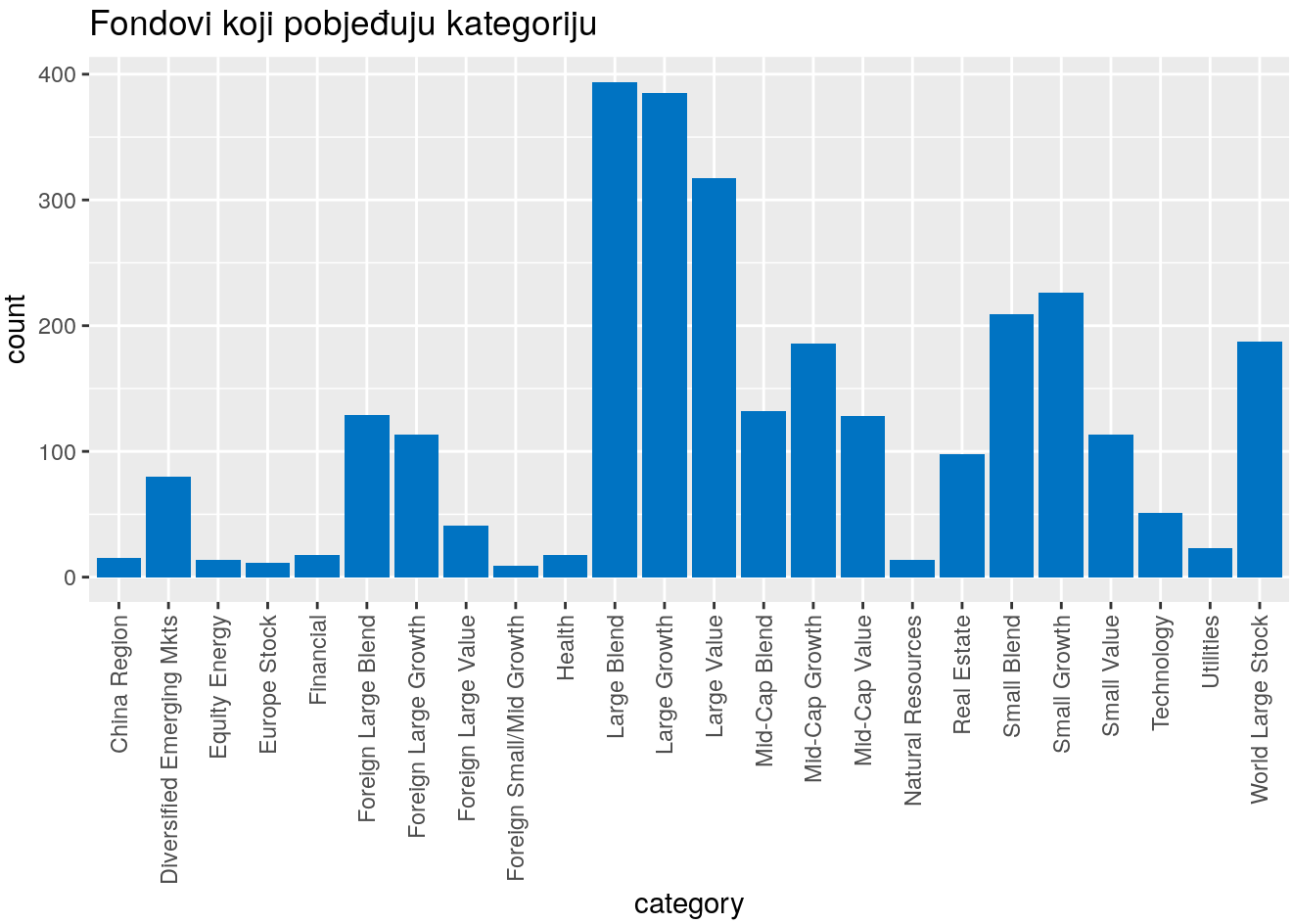
```
##                     category   p10years
## 1               China Region 0.93272520
## 2              Mid-Cap Value 0.84186652
## 3             Mid-Cap Growth 0.96663274
## 4                Large Value 1.00000000
## 5               Large Growth 0.99999989
## 6          World Large Stock 0.01683751
## 7               Small Growth 0.02249687
## 8                Small Value 0.95203541
## 9  Diversified Emerging Mkts 0.77705081
## 10       Foreign Large Value 0.99837641
## 11      Foreign Large Growth 0.01417464
## 12  Foreign Small/Mid Growth 1.00000000
## 13       Foreign Large Blend 0.80997796
## 14               Large Blend 0.99999665
## 15               Small Blend 0.97671891
## 16              Europe Stock 0.99685683
## 17               Real Estate 0.32014452
## 18             Mid-Cap Blend 0.99603032
## 19         Natural Resources 0.99999726
## 20                Technology 0.91376608
## 21             Equity Energy 0.80909944
## 22                 Financial 0.99991963
## 23                 Utilities 0.71370825
## 24                    Health 0.99965388
```

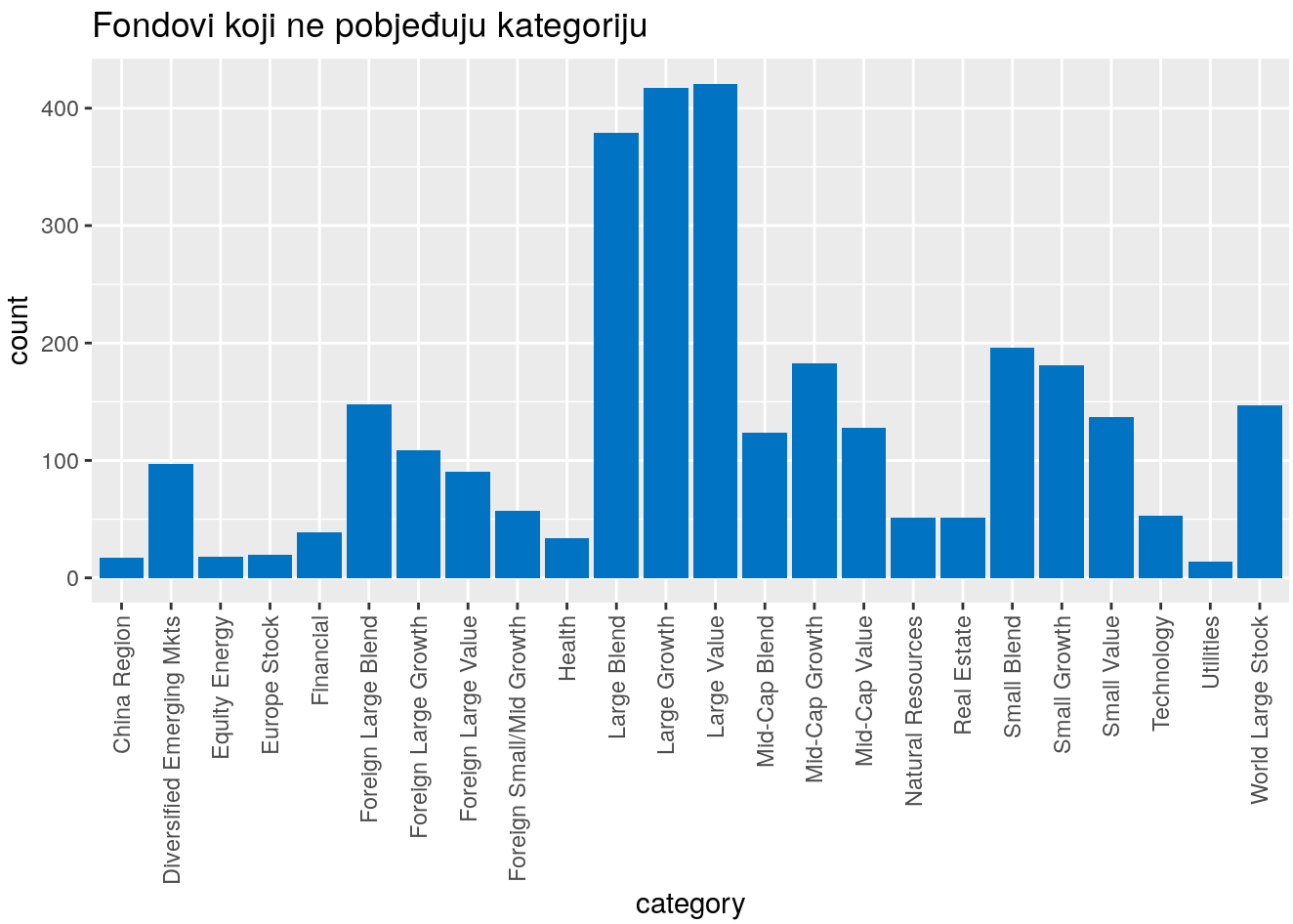
Analizirali smo svaku kategoriju zasebno i dobili smo da u 12.5% fondovi pobjeđuju svoju kategoriju.

Usporedba fondova koji pobjeđuju kategoriju naspram onih koji to ne uspijevaju

Podijelit ćemo fondove u dvije kategorije, one koji pobjeđuju svoju kategoriju i one koji to ne uspijevaju za povrate u 10 godina.

Prikažimo grafički broj fondova za fondove koji uspijevaju pobijediti svoju kategoriju, i one koji ne uspijevaju.


Fondovi koji pobjeđuju kategoriju
count
400
300
200
100
0
China Region
Diversified Emerging Mkts
Equity Energy
Europe Stock
Financial
Foreign Large Blend
Foreign Large Growth
Foreign Large Value
Foreign Small/Mid Growth
Health
Large Blend
Large Growth
Large Value
Mid-Cap Blend
Mid-Cap Growth
Mid-Cap Value
Natural Resources
Real Estate
Small Blend
Small Growth
Small Value
Technology
Utilities
World Large Stock
category



Testirajmo ima li odabrana strategija utjecaj na to hoće li fond pobijediti svoju kategoriju.

Pogledajmo kontingencijsku tablicu.

```
##              Blend Growth Value  Sum
## veci_invest    778   1436   697 2911
## manji_invest   895   1222   987 3104
## Sum           1673   2658  1684 6015
```

Da bi izvršili test, frekvencija za svaki razred nam mora biti veća ili jednaka 5.

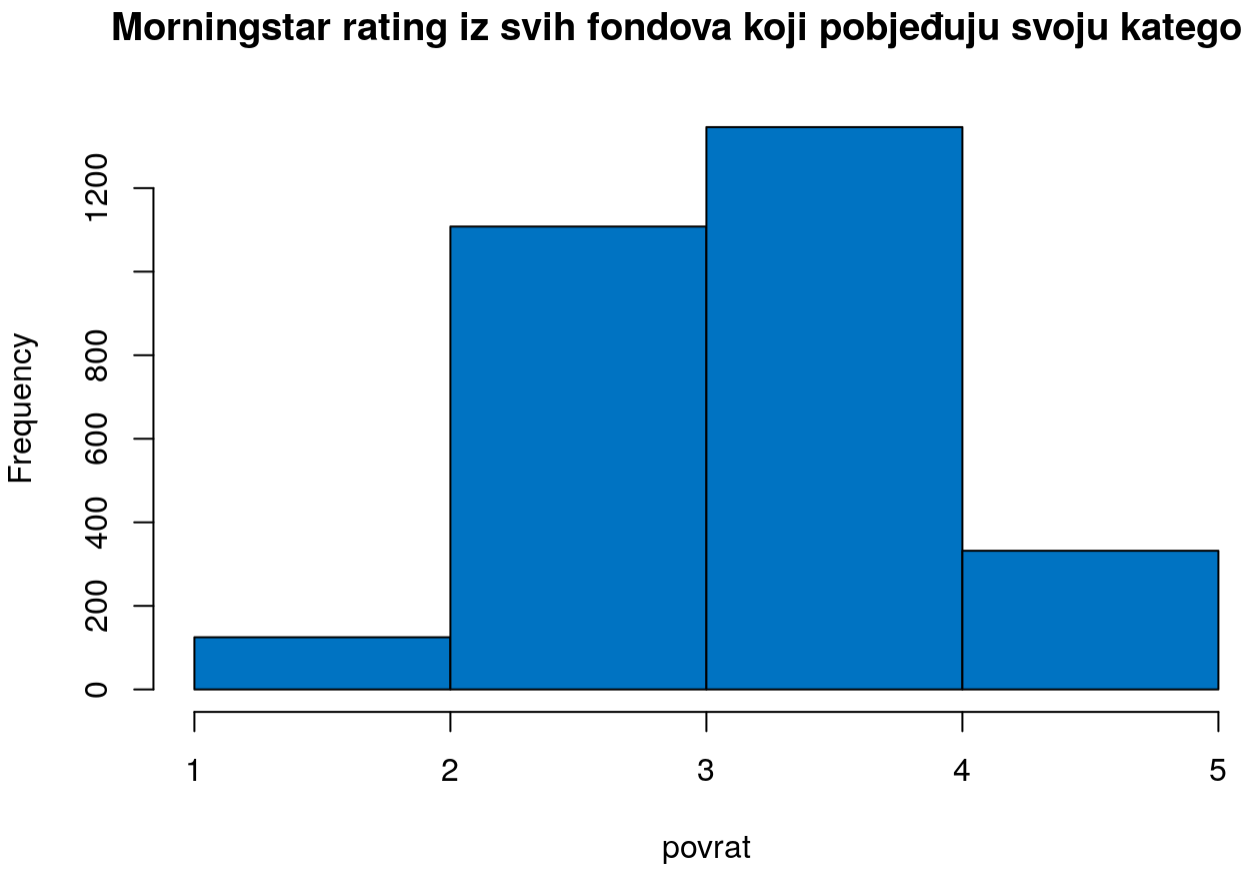
```
## Očekivane frekvencije za razred  Blend - veci_invest :  809.6597
## Očekivane frekvencije za razred  Blend - manji_invest :  863.3403
## Očekivane frekvencije za razred  Growth - veci_invest :  1286.357
## Očekivane frekvencije za razred  Growth - manji_invest :  1371.643
## Očekivane frekvencije za razred  Value - veci_invest :  814.9832
## Očekivane frekvencije za razred  Value - manji_invest :  869.0168
```

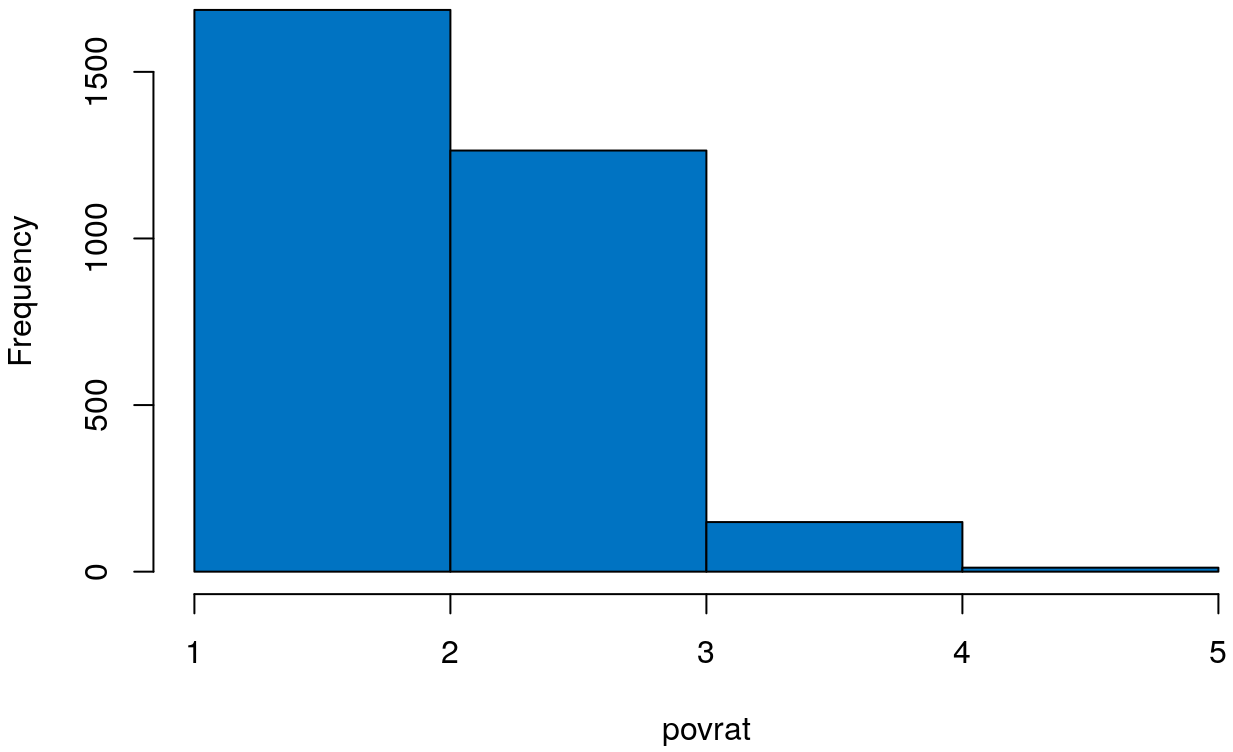
Sve očekivane frekvencije su veće od 5, možemo nastaviti s izvršavanjem testa.

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 69.231, df = 2, p-value = 9.261e-16
```

Vidimo da nam odabrana strategija govori koji fond pobjeđuje kategoriju.

Grafički pregled Morningstar ratinga između fondova koji pobjeđuju i ne pobjeđuju svoju kategoriju. Morningstar je rating nekog fonda zasnovan na velikom broju faktora. Njegova vrijednost ide od 5 (najveće) do 1 (najmanje) vrijednosti. Morningstar je veoma važna metrika uspješnosti fonda.





Prema gore odrađenom grafu, možemo primijetiti da postoji velika razlika između Morningstar ratinga. Ovdje koristimo t-test. Gledamo je li srednja vrijednost Morningstar ratinga kod fonova koji pobjeđuju svoju kategoriju veća od onih koji ne uspijevaju pobijediti.

H_0 Morningstar rating fondova koji pobjeđuju svoju kategoriju nije veći od onih koji ne pobjeđuju svoju kategoriju.

```
##
##  Welch Two Sample t-test
##
## data:  v$morningstar_rating and m$morningstar_rating
## t = 63.715, df = 6010, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.213007      Inf
## sample estimates:
## mean of x mean of y
##  3.643422  2.398264
```

Vidimo da naša inicijalna pretpostavka nije bila točna, tj. alternativna hipoteza je točna. Fondovi s većim Morningstar rating generalno pobjeđuju svoje kategorije više nego oni s manjim.

Uspješnost fondova

Sljedeći korak našeg istraživanja bio je odrediti “uspješnost” fondova.

Prije početka rada potrebno je pregledati kako se ponašaju varijable koje smo izabrali.

Vidimo da se atributi većinski sastoje od “numeric”, “integer” i “character” podatka. Posebno je zanimljiv atribut “morningstar_rating”. On predstavlja rejting agencije “Morningstar” i može biti u rasponu vrijednosti od 1 do 5. Budući da možemo primijetiti da pojedini fondovi imaju rating 0, te fondove ćemo izostaviti iz daljnje analize. Osim toga, atribut “morningstar_rating” koristit ćemo kao kategorijsku varijablu te je pretvaramo iz tipa “integer” u “factor”.

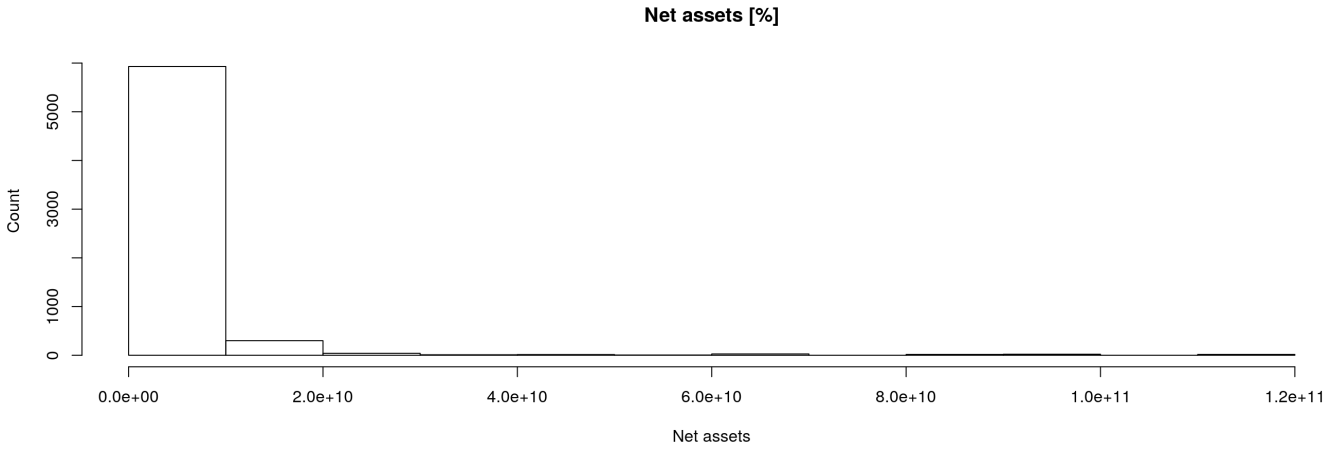
```
##      fund_name      category net_annual_expense_ratio_fund
## 1512     FDPIX Multicurrency                          1.78
## 1513     FDPSX Multicurrency                          2.78
## 4966     RDPIX Multicurrency                          1.78
## 4967     RDPSX Multicurrency                          2.78
##      net_annual_expense_ratio_category fund_return_10years
## 1512                              1.29               -3.13
## 1513                              1.29               -4.07
## 4966                              1.29               -0.44
## 4967                              1.29               -1.43
##      category_return_10years morningstar_rating net_assets  investment
## 1512                    0.28                  0    3950000 <undefined>
## 1513                    0.28                  0    3950000 <undefined>
## 4966                    0.28                  0   15870000 <undefined>
## 4967                    0.28                  0   15870000 <undefined>
##      price_earnings
## 1512              0
## 1513              0
## 4966              0
## 4967              0
```

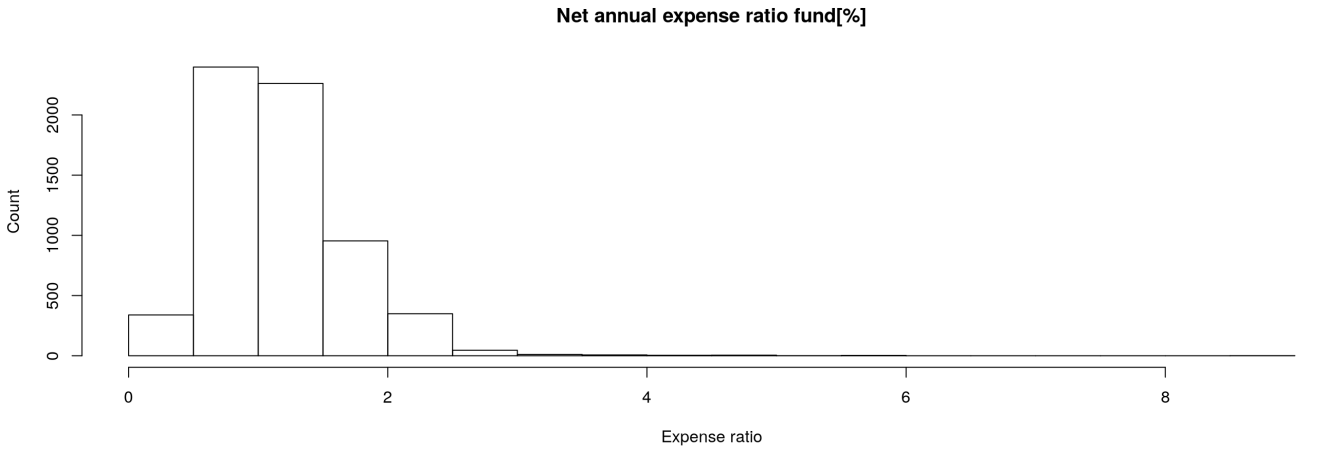
Nakon što smo izbacili fondove s nevaljanim rejtingom, kojih na sreću nije bilo puno, potrebno je pregledati preostale atribute svih fondova te potencijalno izbaciti fondove koji ne posjeduju podatke o atributima važnim za našu analizu.

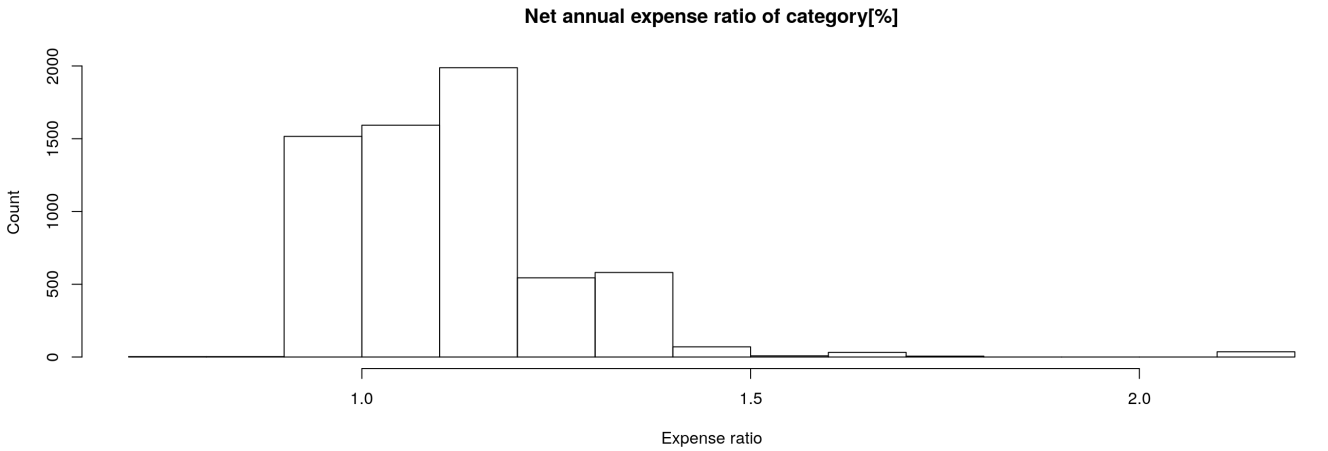
```
## Ukupno nedostajućih vrijednosti za varijablu  net_assets :  2
```

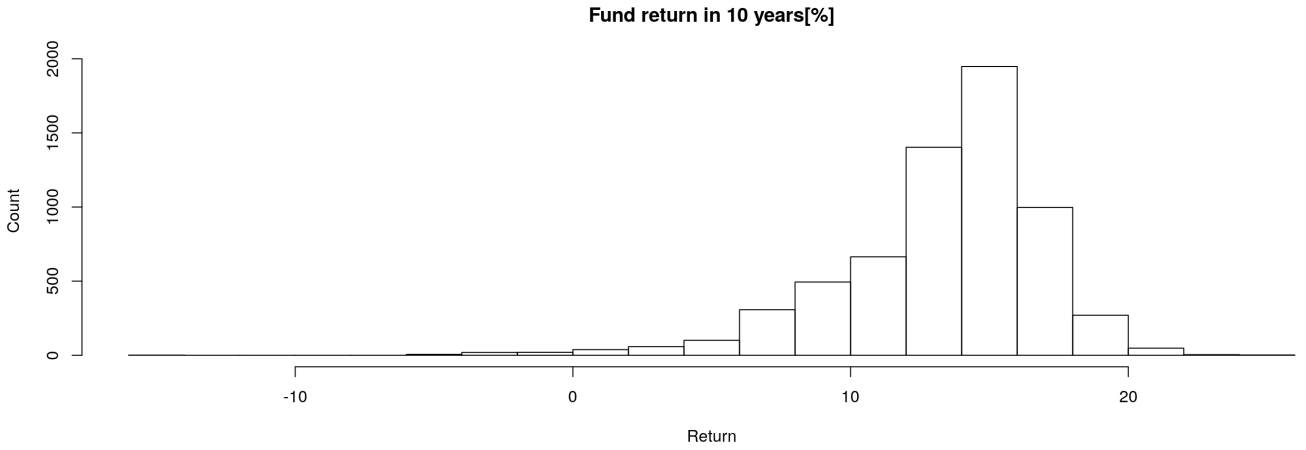
Vidimo da postoje dva fonda koji nemaju definiranu varijablu net_assets. Budući da izbacivanje takvog malog broja fondova neće znatno utjecati na našu statistiku, to možemo napraviti.

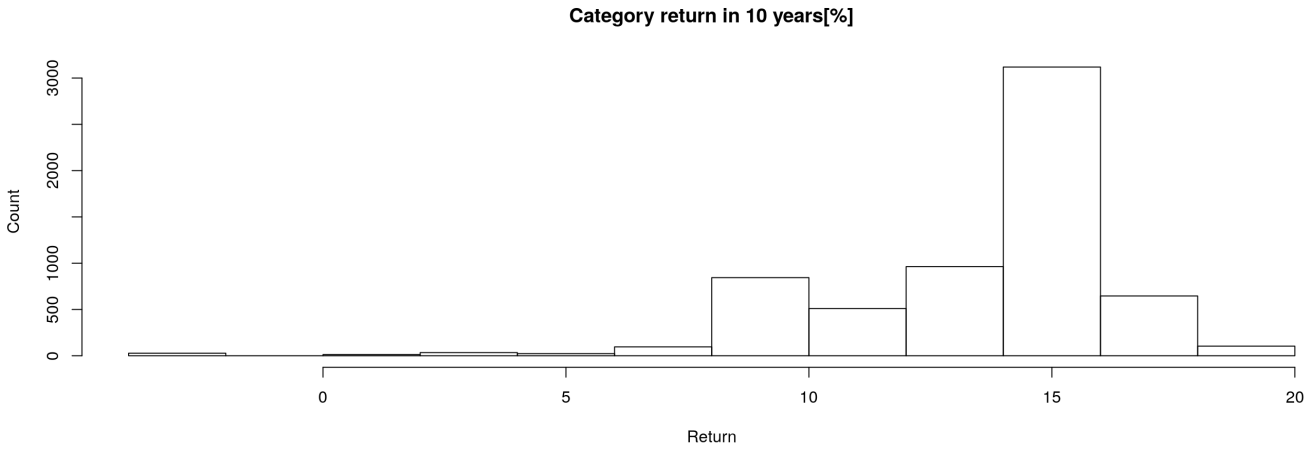
Kako bismo dobili bolji uvid u podatke s kojima radimo, prikazat ćemo ih grafički. Pritom ćemo definirati novi atribut “success” kao razliku između prosječnog povrata fonda u zadnjih 10 godina i prosječnog povrata svih fondova u toj kategoriji. Tu mjeru koristit ćemo za većinu daljnje analize, te nam je ponašanje tog atributa posebno zanimljivo.

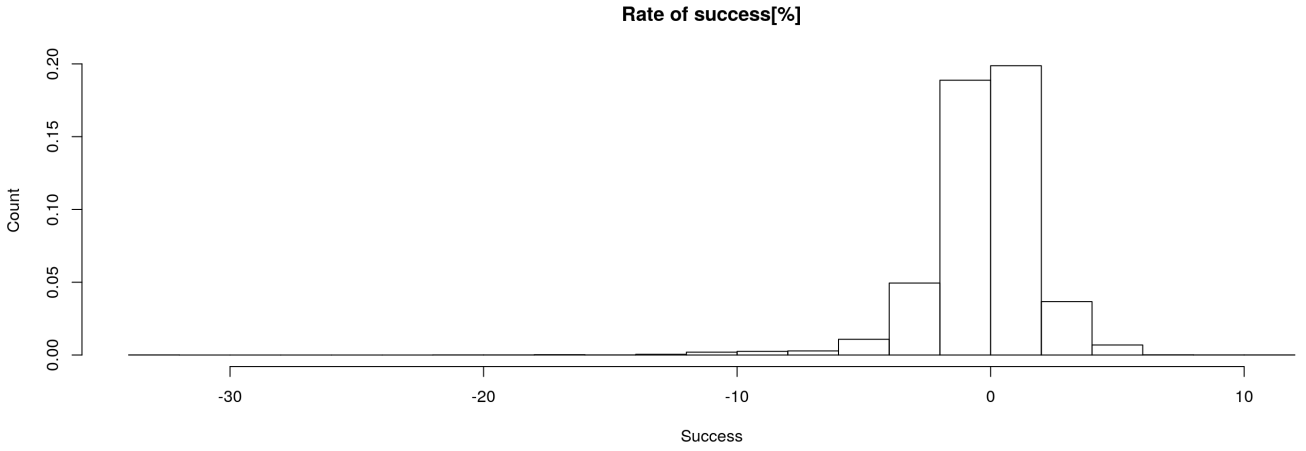

Net assets [%]
Count
0
1000
3000
5000
0.0e+00
2.0e+10
4.0e+10
6.0e+10
8.0e+10
1.0e+11
1.2e+11
Net assets


Net annual expense ratio fund[%]
Count
0
500
1000
1500
2000
0
2
4
6
8
Expense ratio

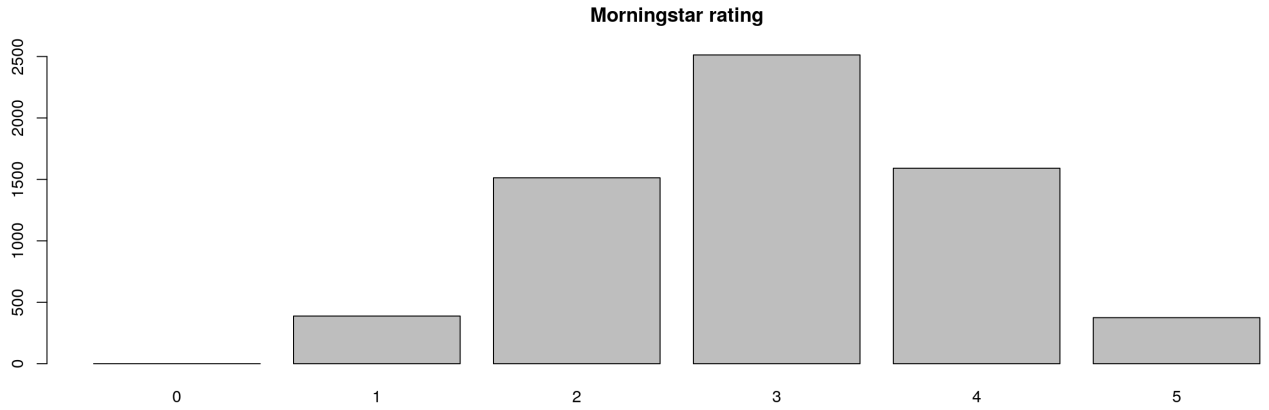

Net annual expense ratio of category[%]
Count
0
500
1000
1500
2000
1.0
1.5
2.0
Expense ratio







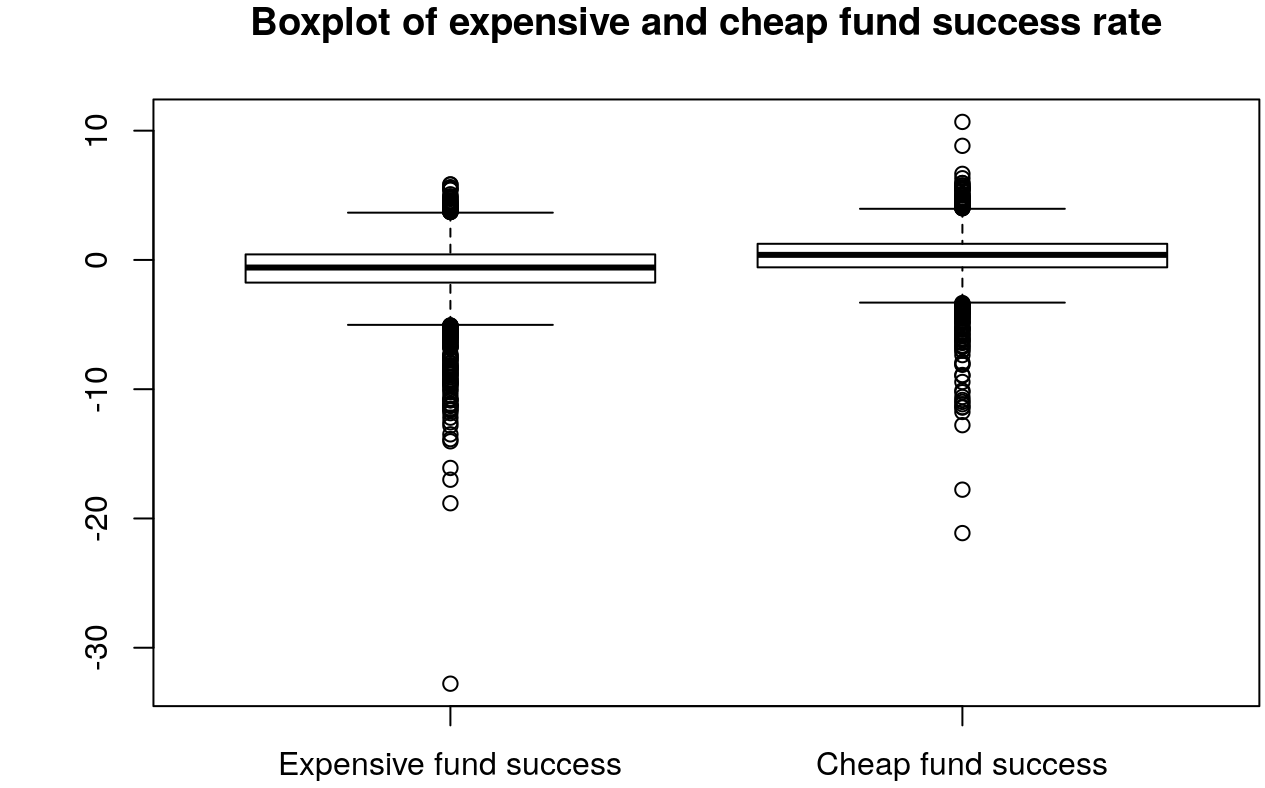
Kao i za numeričke podatke, deskriptivnu statistiku napravit ćemo i nad kategorijskim podatkom “morningstar_rating”.

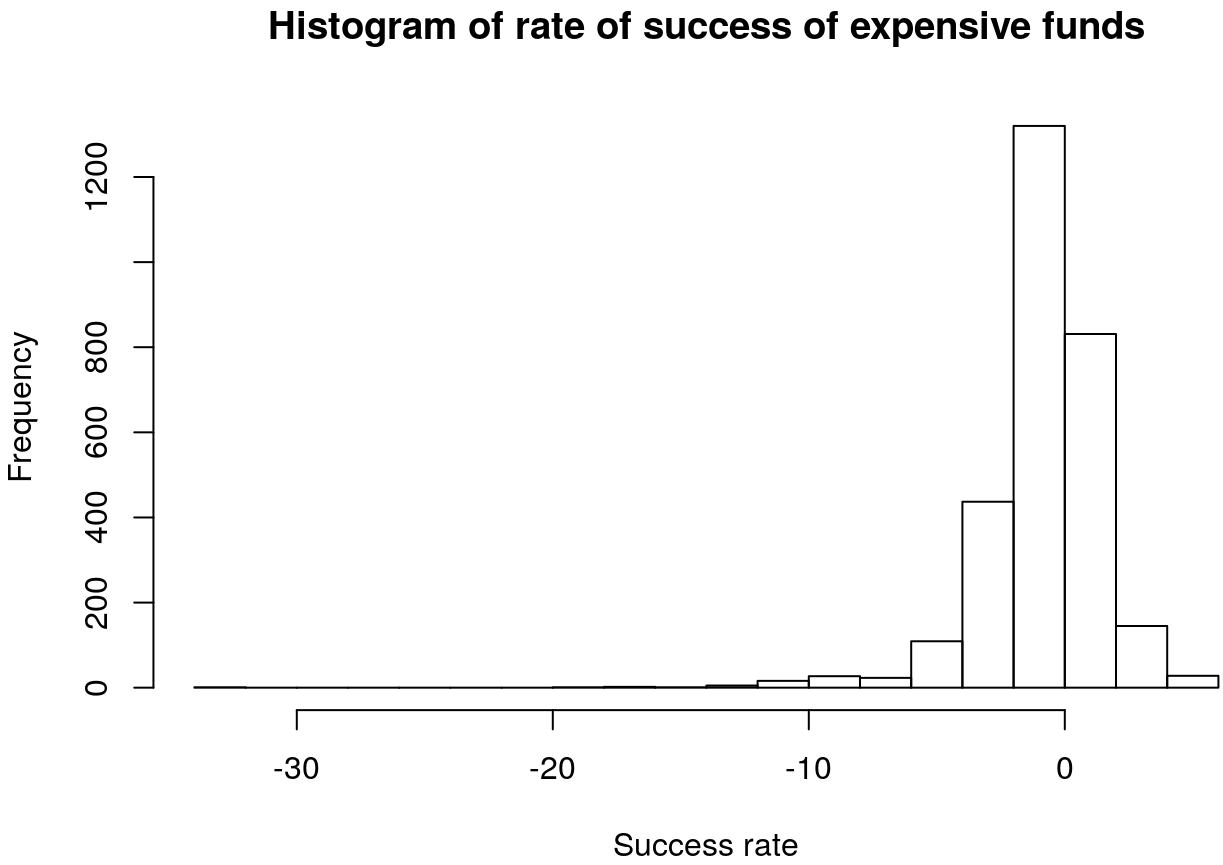


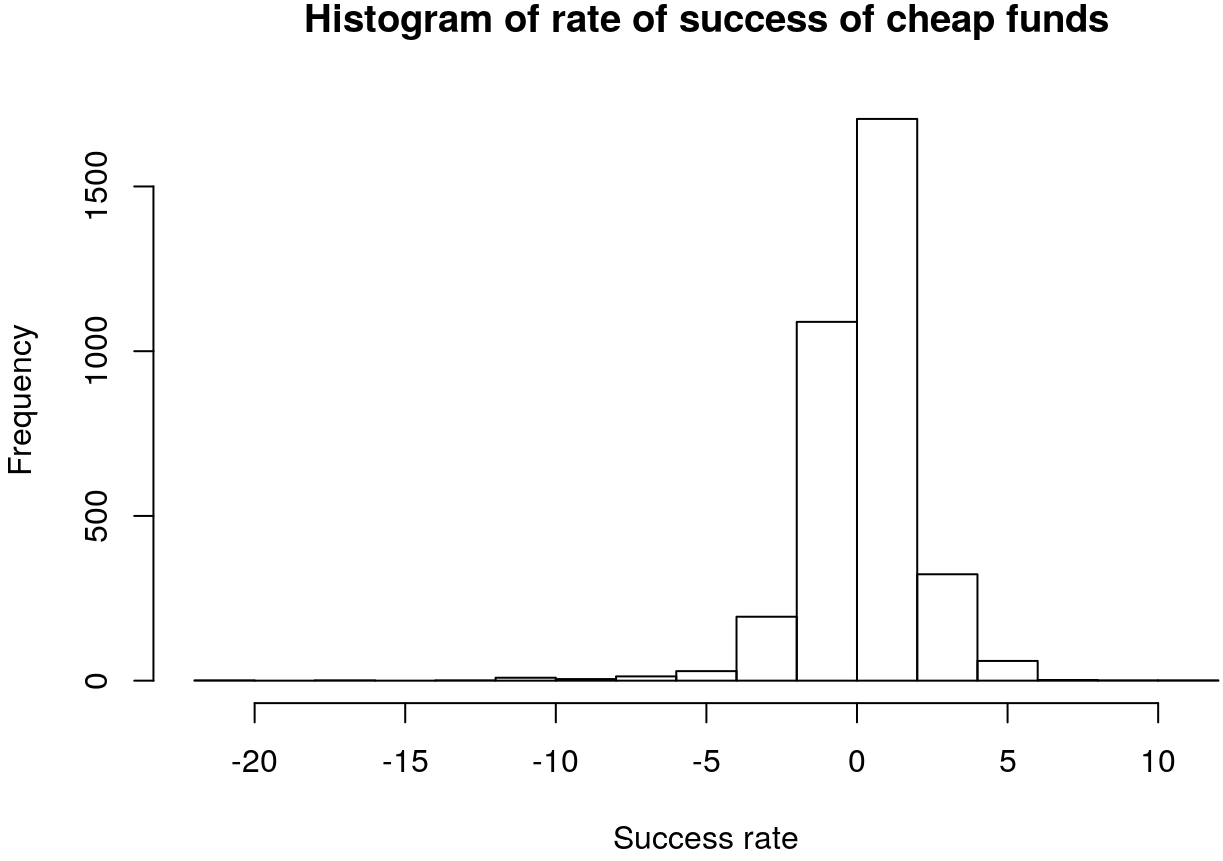
Kao što je već spomenuto, naše istraživanje bavit će se uspjehom pojedinih fondova. Specifično, u sljedećem koraku istraživanja zanima nas jesu li skuplji fondovi uspješniji od jeftinijih te vrijedi li ista ta tvrdnja i za veće i manje fondove. Stoga ćemo fondove podijeliti na skuplje, ako im je godišnji postotni trošak upravljanja veći od prosječnog godišnjeg postotnog troška fondova u toj kategoriji, te na jeftinije, ako je godišnji postotni trošak manji ili jednak. Na sličan način, fondovi će biti veći ako im je ukupna imovina pod upravljanje veća od medijana podataka, a manji ako vrijedi suprotno. U istom koraku napravit ćemo pregled pojedinih skupova podataka.

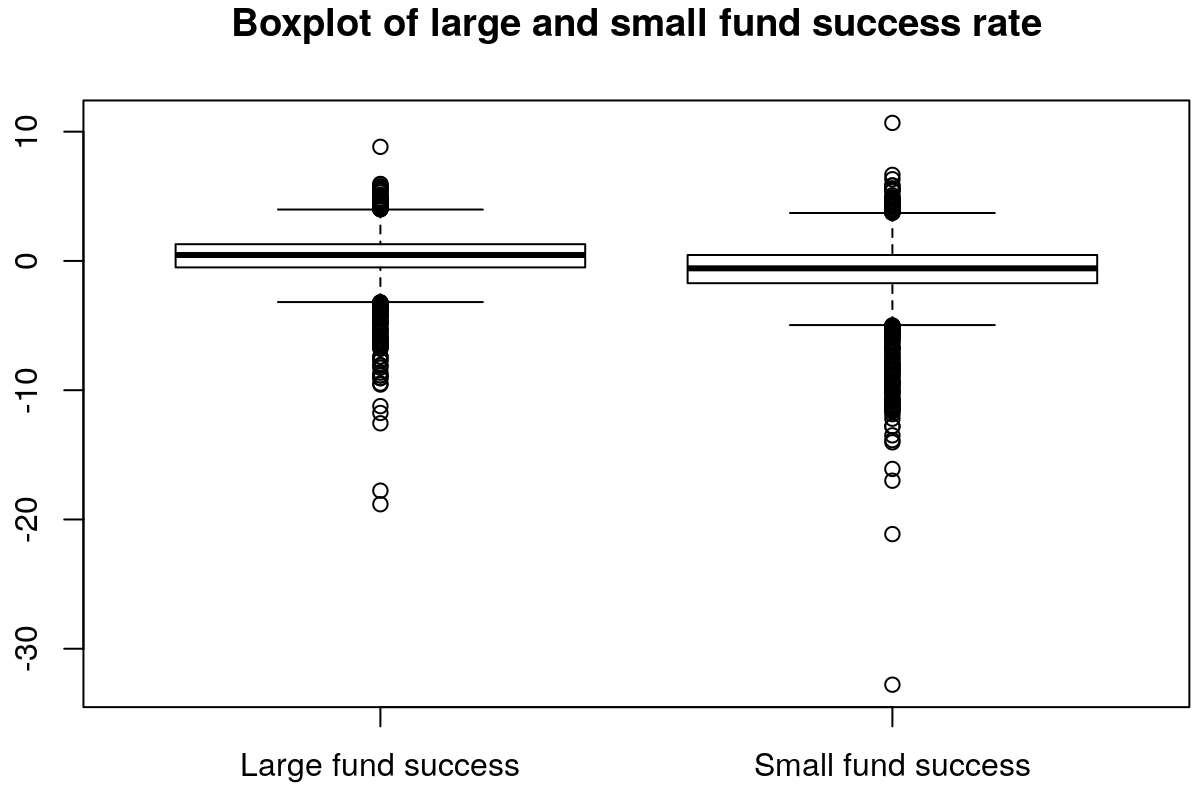
Već iz inicijalnog pregleda dobivenih podataka možemo dobiti generalnu ideju o tome koji će skup podataka biti uspješniji. To možemo vidjeti i ako usporedimo srednje vrijednosti uspjeha dobivenih skupova te modova njihovih rejtinga.

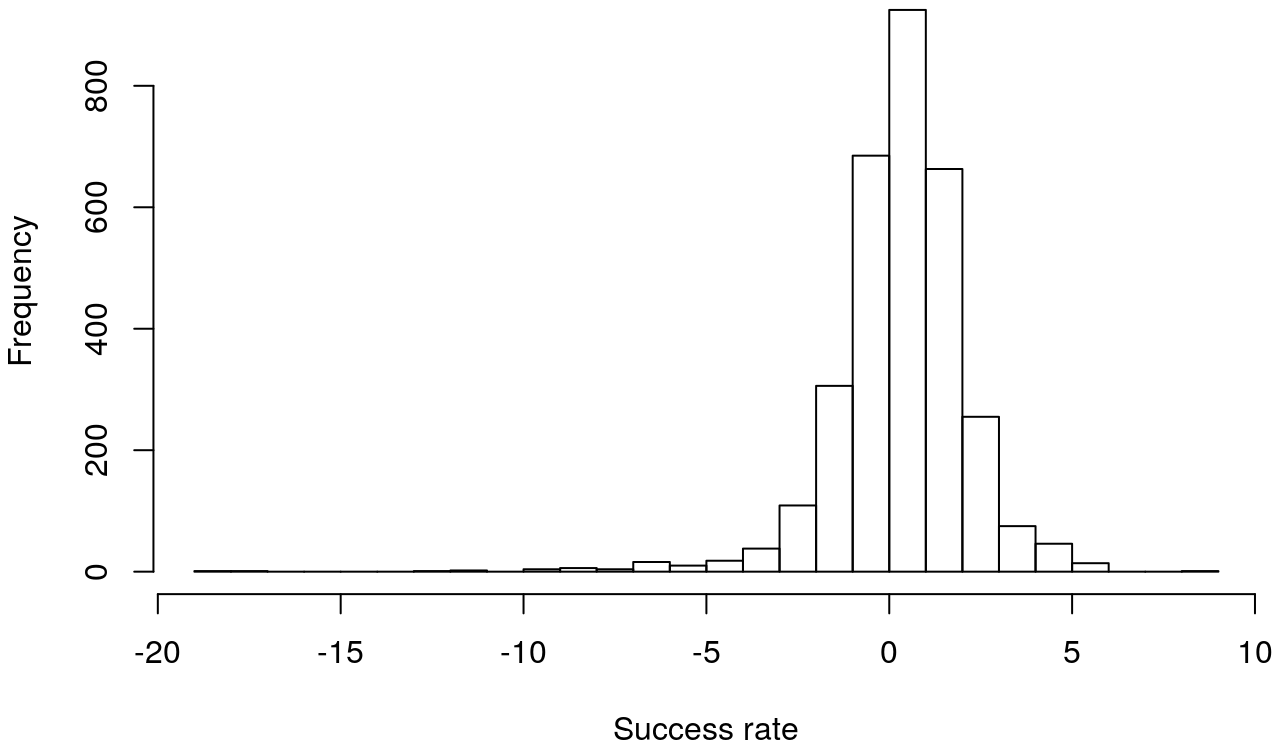
Inicijalnom procjenom podataka mogli bi zaključiti da će jeftiniji fondovi biti uspješniji od skupljih te da će veći fondovi biti uspješniji od manjih. Kako bi još lakše mogli analizirati podatke, prikazat ćemo ih i grafički pomoću box-plotova i histograma.

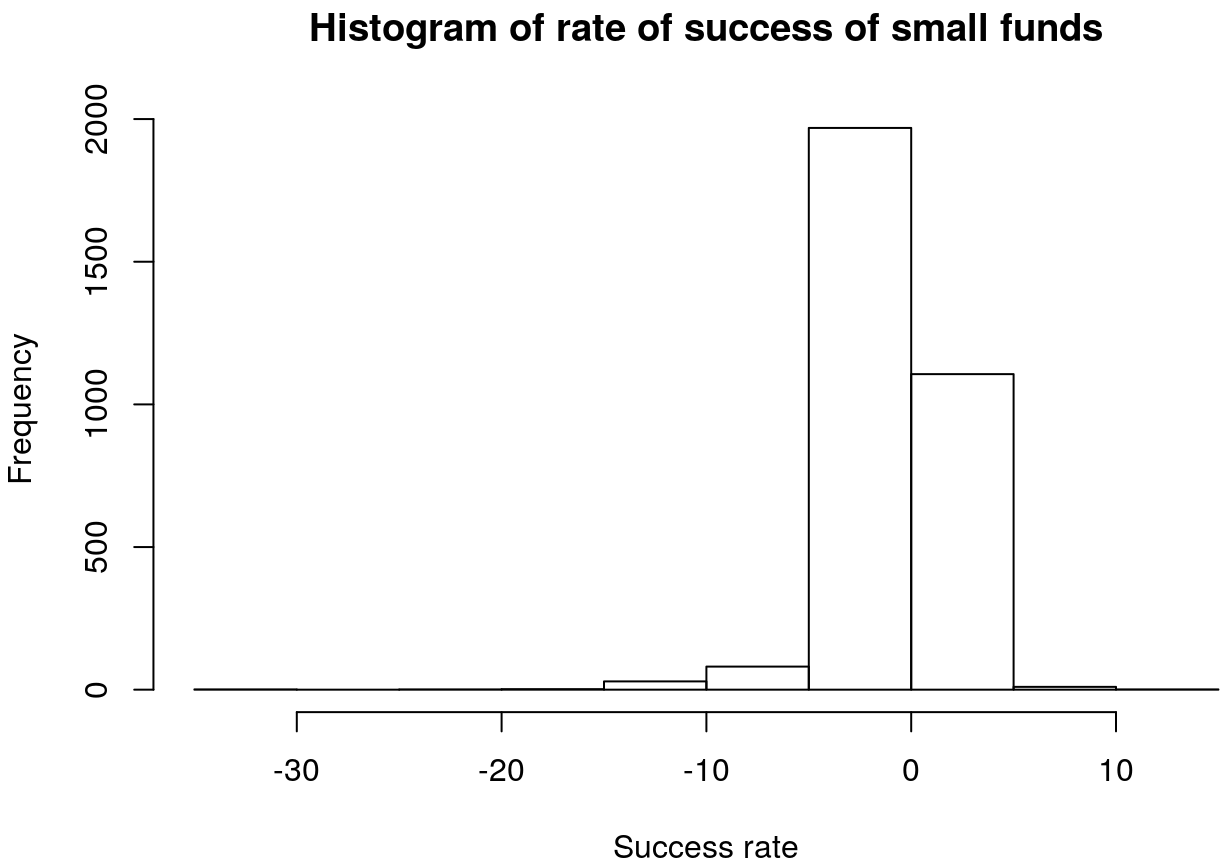

Boxplot of expensive and cheap fund success rate
10
0
-10
-20
-30
Expensive fund success
Cheap fund success


Histogram of rate of success of expensive funds
Frequency
0
200
400
600
800
1200
-30
-20
-10
0
Success rate

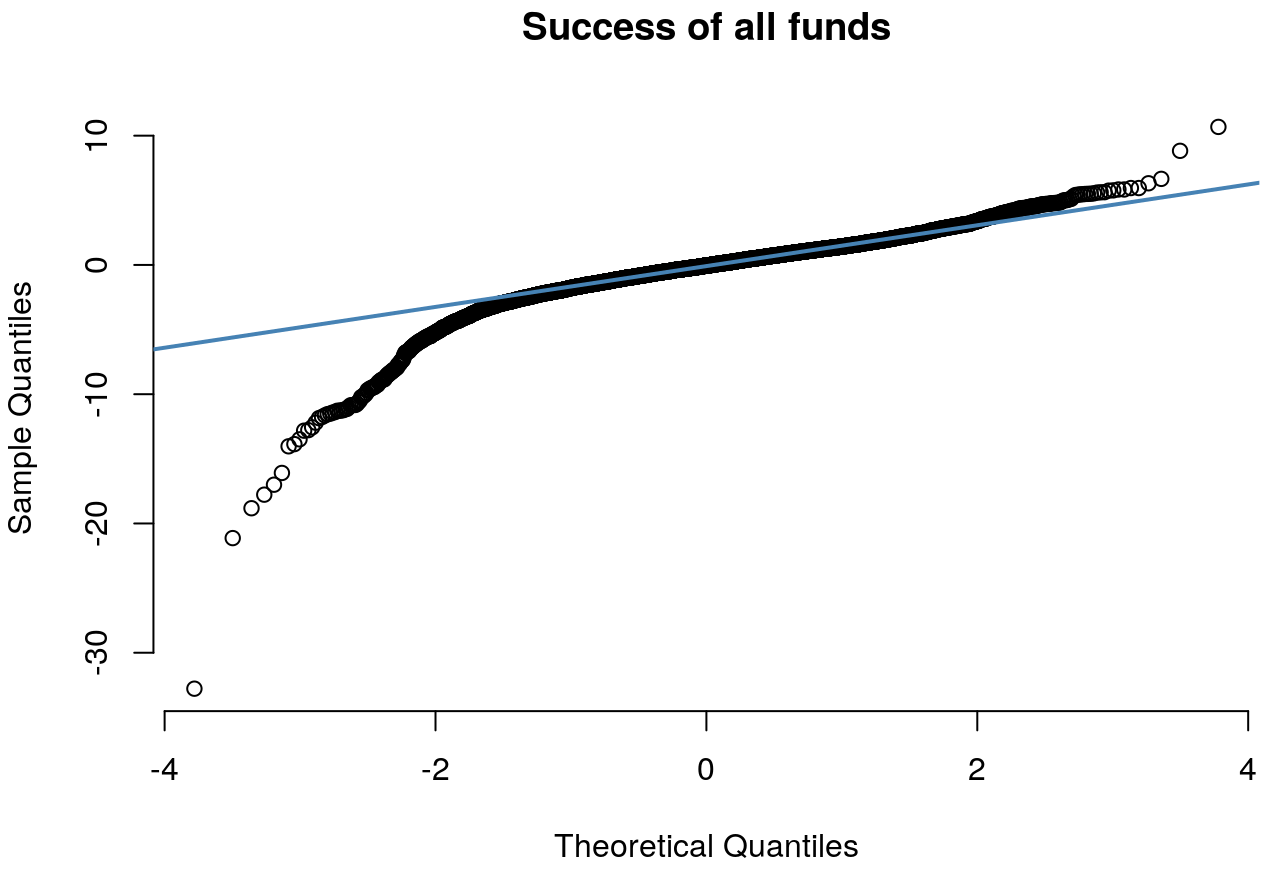

Histogram of rate of success of cheap funds
Frequency
0
500
1000
1500
-20
-15
-10
-5
0
5
10
Success rate


Boxplot of large and small fund success rate
10
0
-10
-20
-30
Large fund success
Small fund success


Histogram of rate of success of large funds
Frequency
0
200
400
600
800
-20
-15
-10
-5
0
5
10
Success rate


Histogram of rate of success of small funds
Frequency
0
500
1000
1500
2000
-30
-20
-10
0
10
Success rate

Iz grafičkih podataka još bolje možemo zaključiti da se jeftiniji i veći fondovi na prvi pogled čine uspješnijima. Sada je potrebno odrediti jesu li te razlike statistički značajne. Kako bismo mogli odrediti koje statističke testove koristiti, prvo je potrebno ispitati normalnost dobivenih podataka. To ćemo ispitati upotrebom qq plotova te upotrebom Lillieforsove inačice Kolmogorov-Smirnovljevog testa. Bitno je napomenuti i prisutnost velikog broja stršećih vrijednosti. Uzevši u obzir kontekst podataka s kojima baratamo, odlučili smo takve vrijednosti zadržati u analizi, budući da su nam takve vrijednosti posebno zanimljive.




```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mutual_funds_outliers$success
## D = 0.086301, p-value < 2.2e-16
```

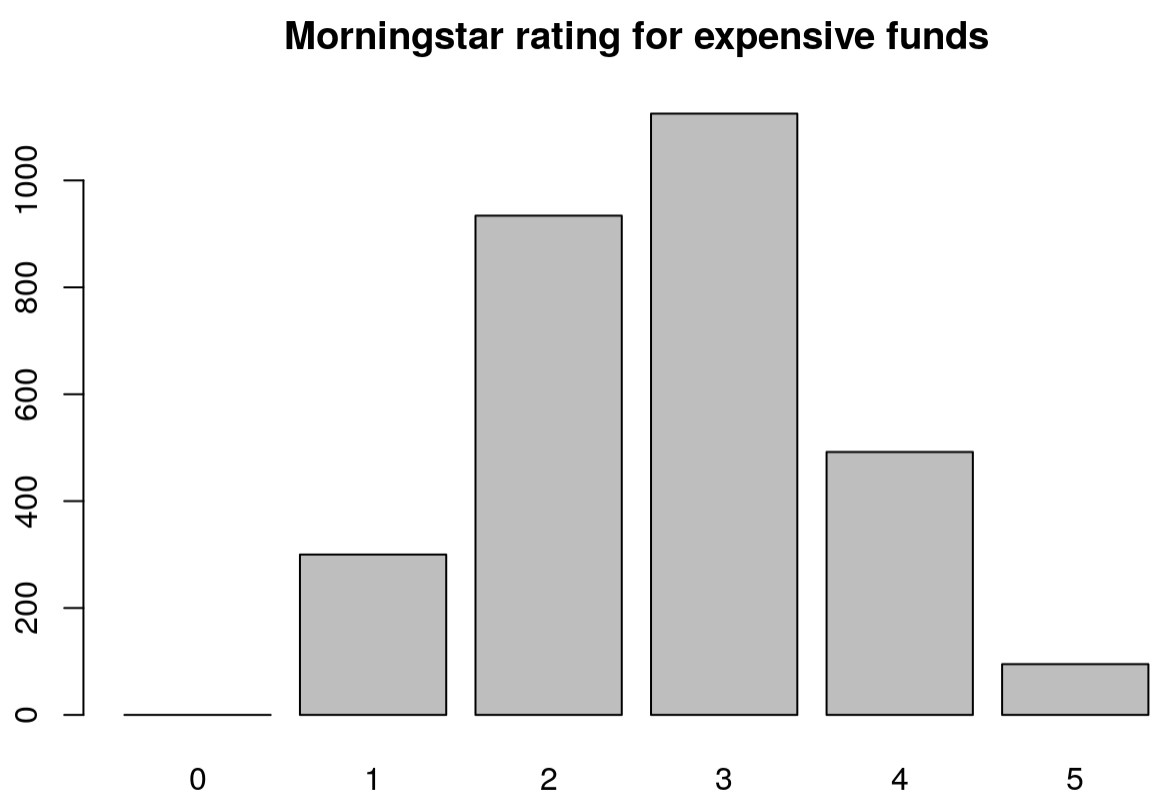
Iz rezultata je vidljivo da se dobivene distribucije značajno razlikuju od normalne.Stoga nećemo koristiti t-test kako bismo usporedili srednje vrijednosti tih skupova, nego ćemo koristiti njegovu neparametarsku inačicu, tj. Mann-Whitney-Wilcoxov test.

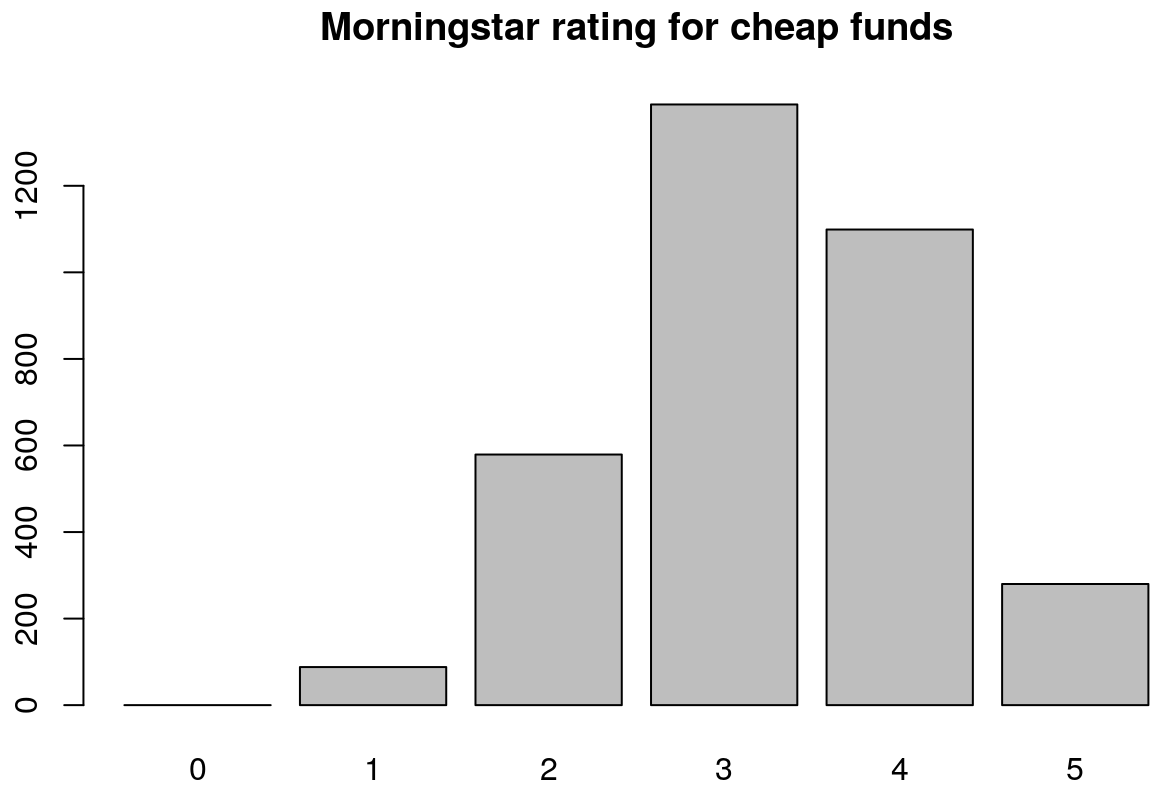
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mutual_funds_outliers_cheap$success and mutual_funds_outliers_expensive$suc
cess
## W = 6785542, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

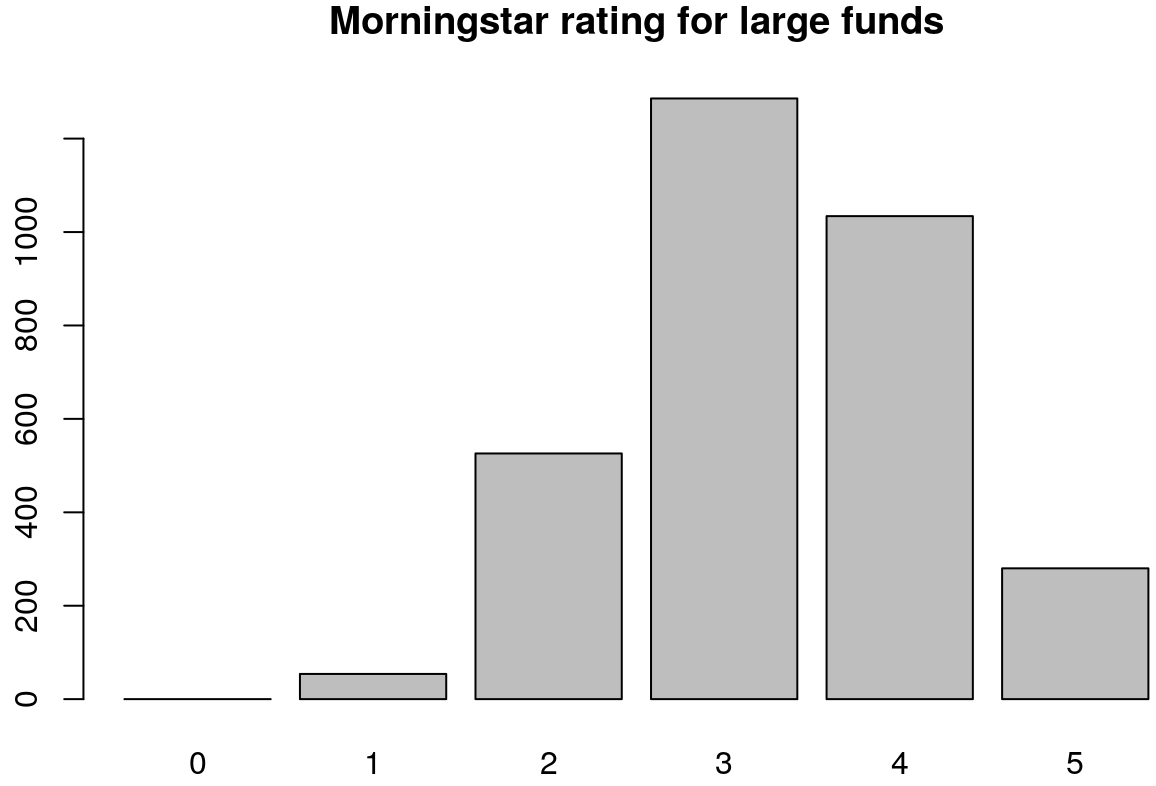
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mutual_funds_large$success and mutual_funds_small$success
## W = 6898014, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

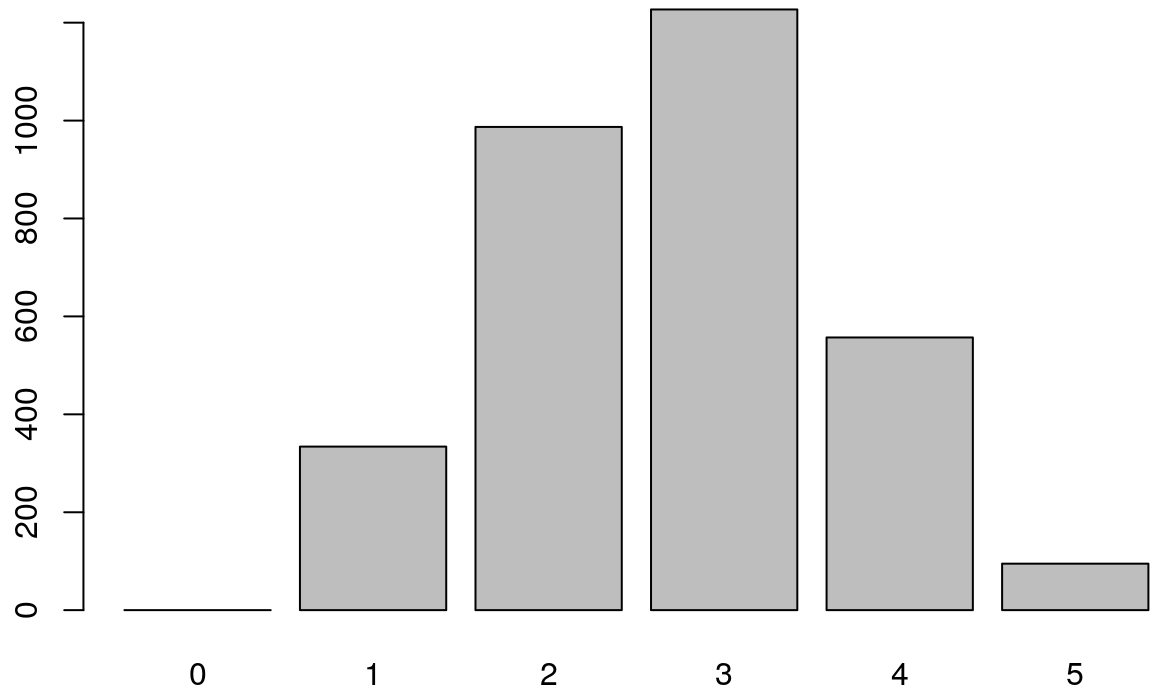
Iz dobivenog rezultata možemo očitati vrlo malu p-vrijednost, što nam govori da možemo odbaciti hipotezu H0 u korist H1. Drugim riječima, možemo reći da su jeftiniji fondovi u prosjeku uspješniji.

Varijabla “morningstar_rating” služit će nam kao druga mjera uspjeha fondova. Stoga ćemo analizirati i te podatke za sve skupove




Morningstar rating for cheap funds
0
200
400
600
800
1200
0
1
2
3
4
5


Morningstar rating for large funds
0
200
400
600
800
1000
0
1
2
3
4
5



Iz podataka je vidljivo da jeftiniji fondovi imaju veći udio bolje rangiranih fondova. Ista pretpostavka vrijedi i za veće fondove. Te pretpostavke provjerit ćemo tako da usporedimo proporcije visoko rangiranih fondova (s rejtingom 4 ili 5) skupova. Za to nam može poslužiti test proporcija s dvije varijable.

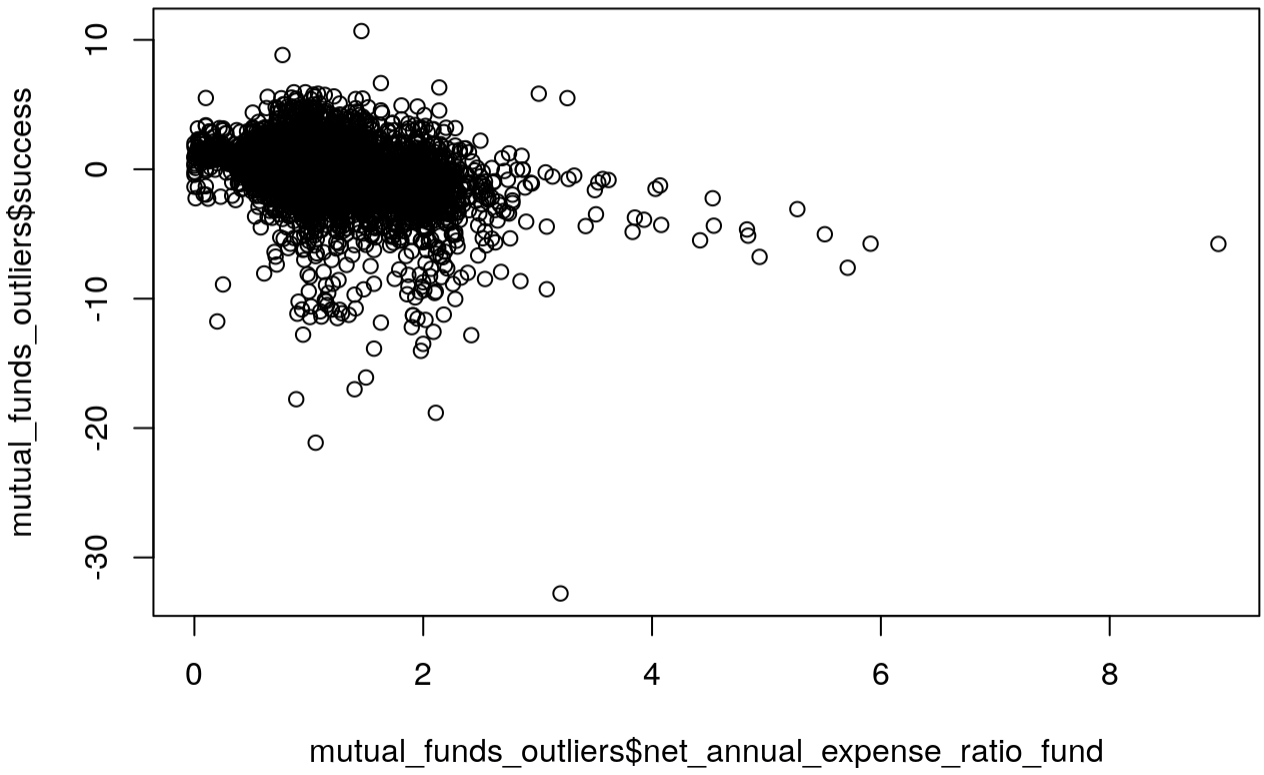
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  rating_proportions out of rating_sum
## X-squared = 303.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1836774 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.4015725 0.1992532
```

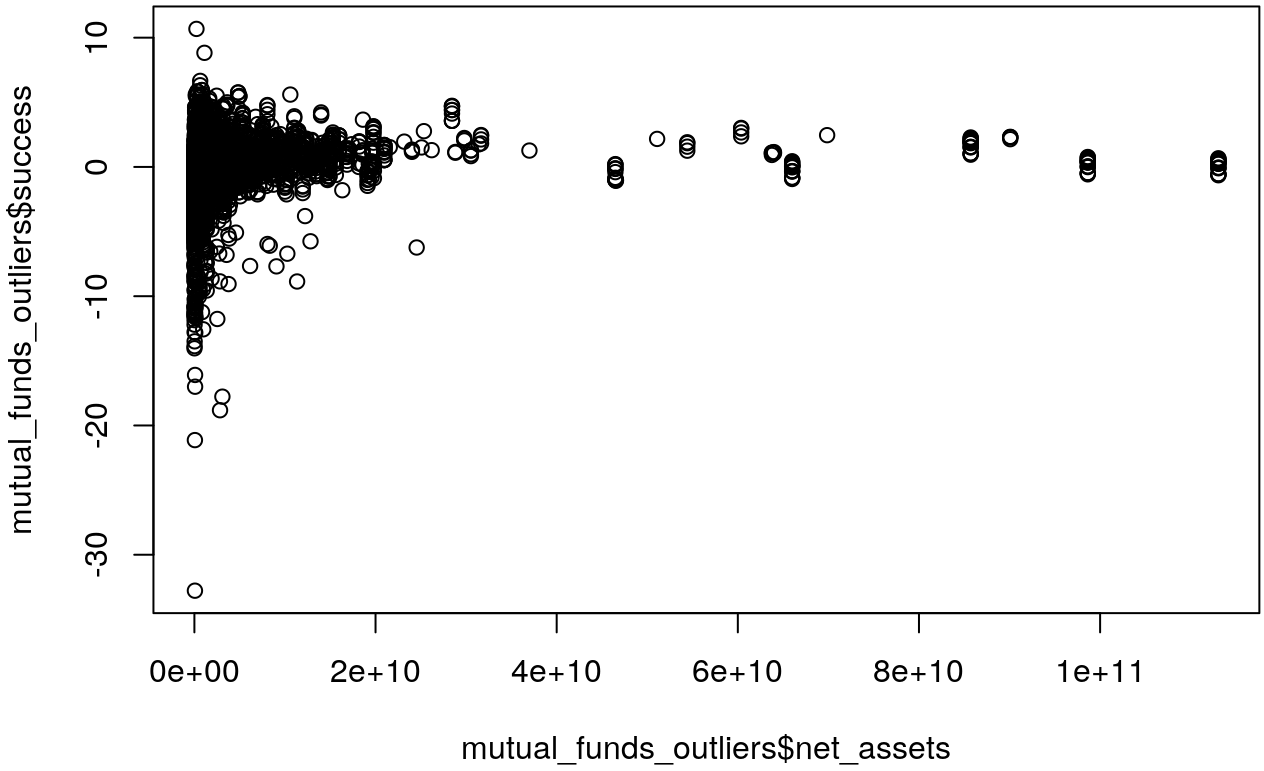
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  rating_proportions out of rating_sum
## X-squared = 327.25, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1906114 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.4132075 0.2037500
```

Male p vrijednosti govore nam da možemo odbaciti hipotezu H_0 u oba slučaja, tj. možemo zaključiti da jeftiniji fondovi imaju veću proporciju visoko rangiranih fondova od skupljih te da isto vrijedi za veće i manje fondove. Ti rezultati poklapaju se i s rezultatima dobivenim gledanjem atributa “success”. Na taj način imamo dvije mjere uspjeha koje daju iste rezultate, što daje veću potporu valjanosti naših metoda.

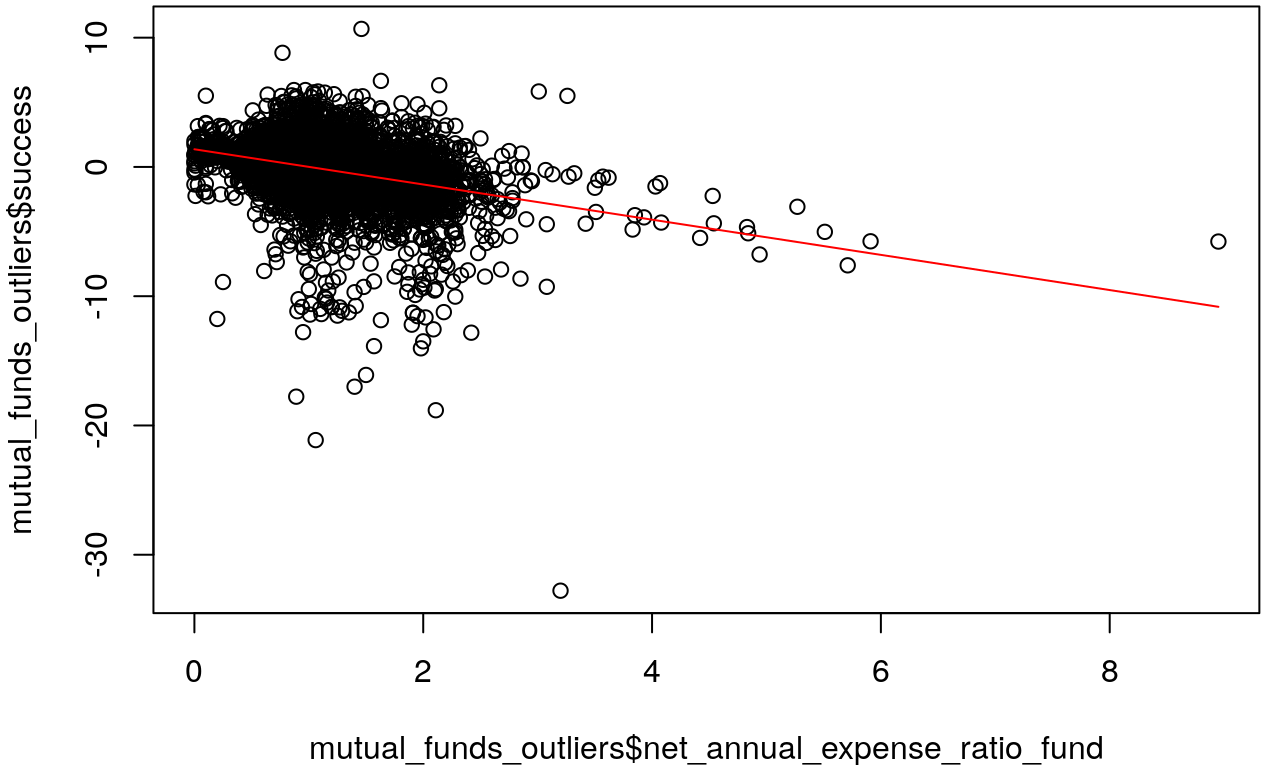
##Pronalazak regresijskog modela

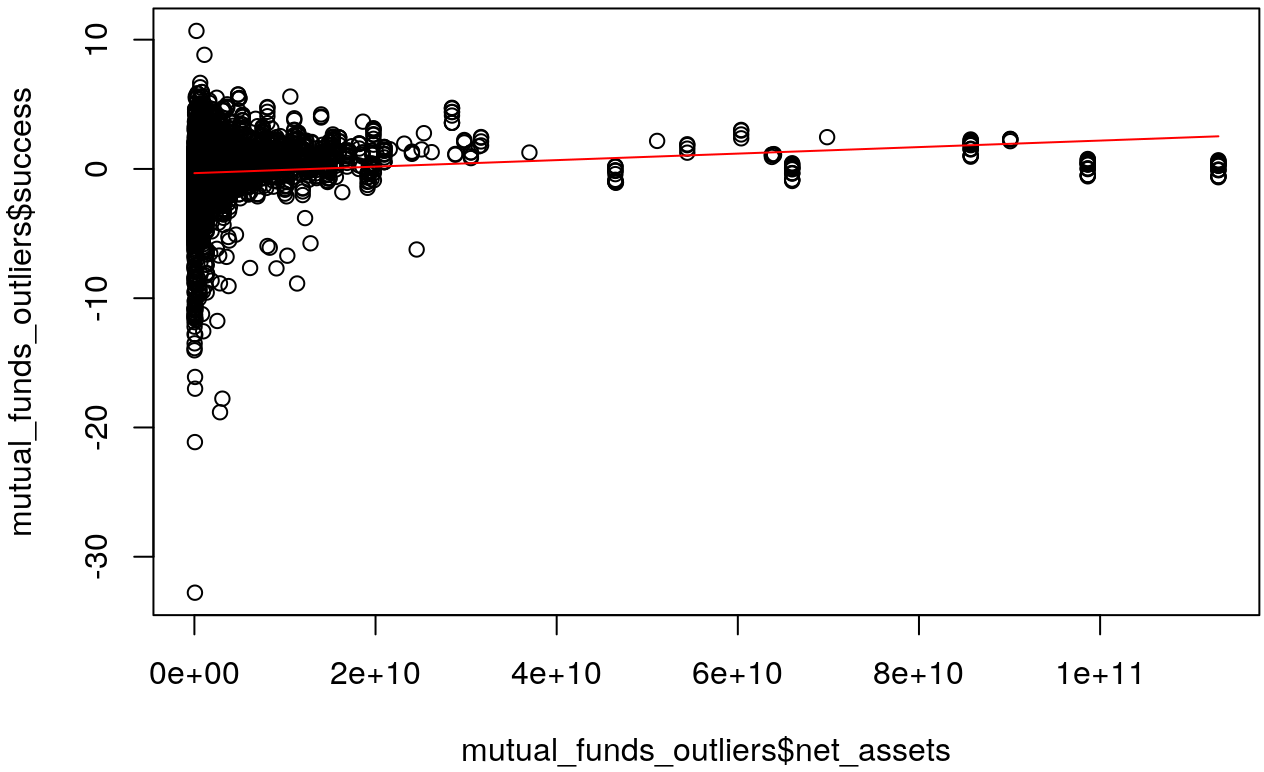
Sljedeće pitanje koje nam se postavlja je to utječu li trošak fonda i njegova veličina direktno na uspjeh fonda. Iz tog razloga želimo istražiti možemo li izgraditi regresijski model koji podržava tu tvrdnju. U prvom koraku te analize pokušat ćemo izgraditi modele jednostavne regresije koristeći prosječni godišnji trošak i ukupnu imovinu pod upravljanje kao nezavisne varijable, te uspjeh fonda kao zavisnu varijablu. Kako bi dobili bolji uvid na način na koji te varijable utječu na uspjeh, koristit ćemo se scatter plotom.





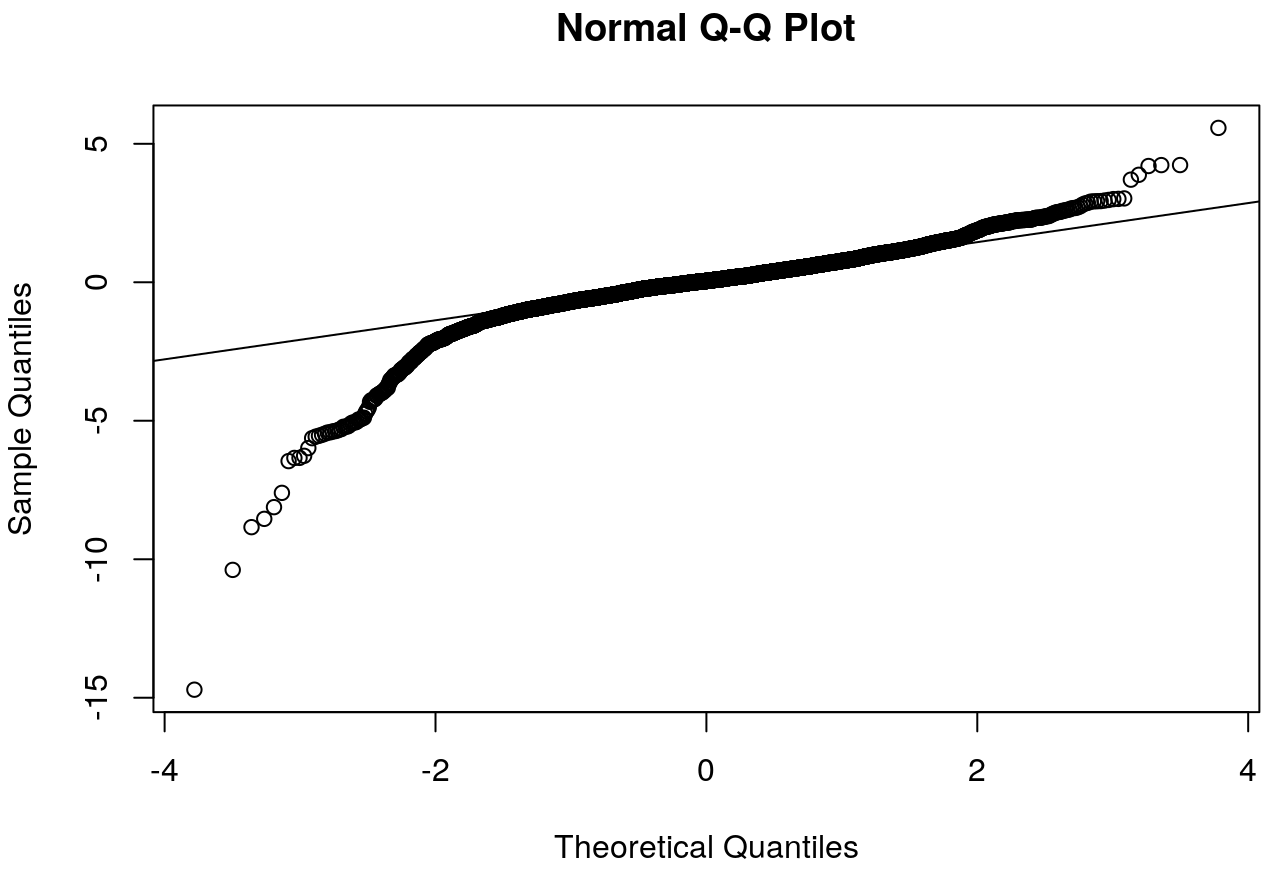
Gledajući dobiveni graf, na prvi pogled bi mogli zaključiti da postoji određeni utjecaj troška na uspjeh. Iz tog razloga gradimo jednostavni regresijski model koji uzima prosječni godišnji trošak kao nezavisnu varijablu, a uspjeh kao zavisnu. Količina imovine ne daje nam neke očite naznake, ali ćemo probati i za nju izgraditi model.

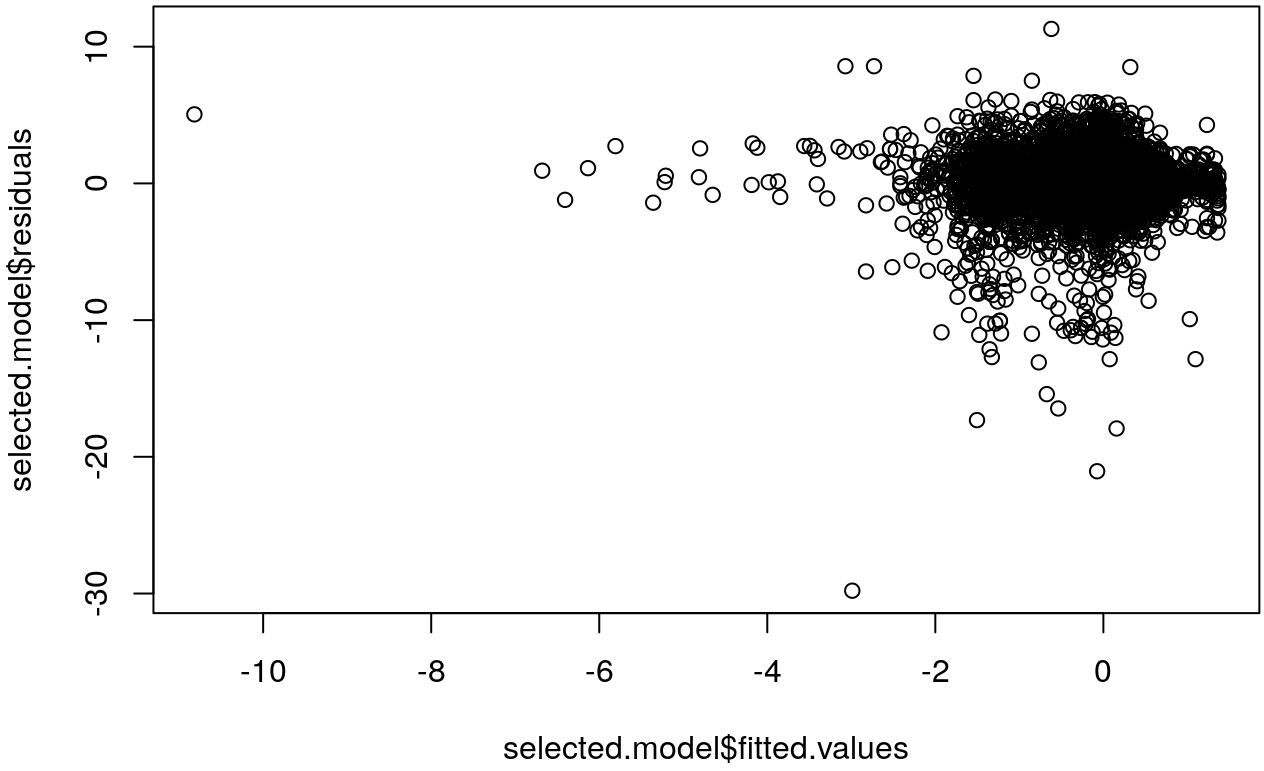

mutual_funds_outliers$success
10
0
-10
-20
-30
0
2
4
6
8
mutual_funds_outliers$net_annual_expense_ratio_fund


mutual_funds_outliers$success
10
0
-10
-20
-30
0e+00
2e+10
4e+10
6e+10
8e+10
1e+11
mutual_funds_outliers$net_assets

Dobiveni nagib pravca potvrđuje tvrdnju o utjecaju troška, no kako bi mogli nastaviti analizu modela, potrebno je ispitati pretpostavke modela, tj. normalnost reziduala i homogenost varijance. Sličan slučaj vrijedi i za veličinu

Prvo analiziramo normalnost reziduala pomoću histograma, qq plotova i Lillieforsove inačice Kolmogorov-Smirnovljevog testa.



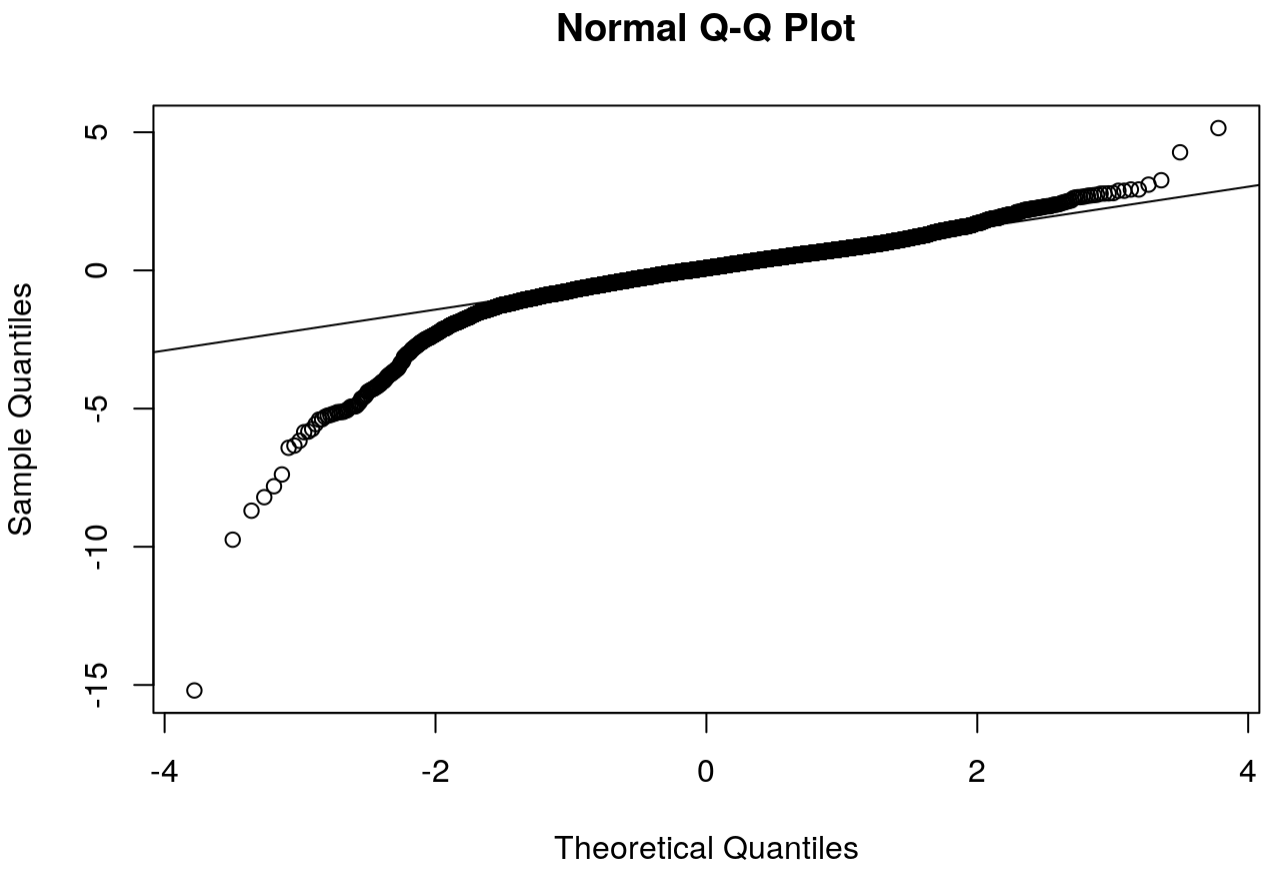


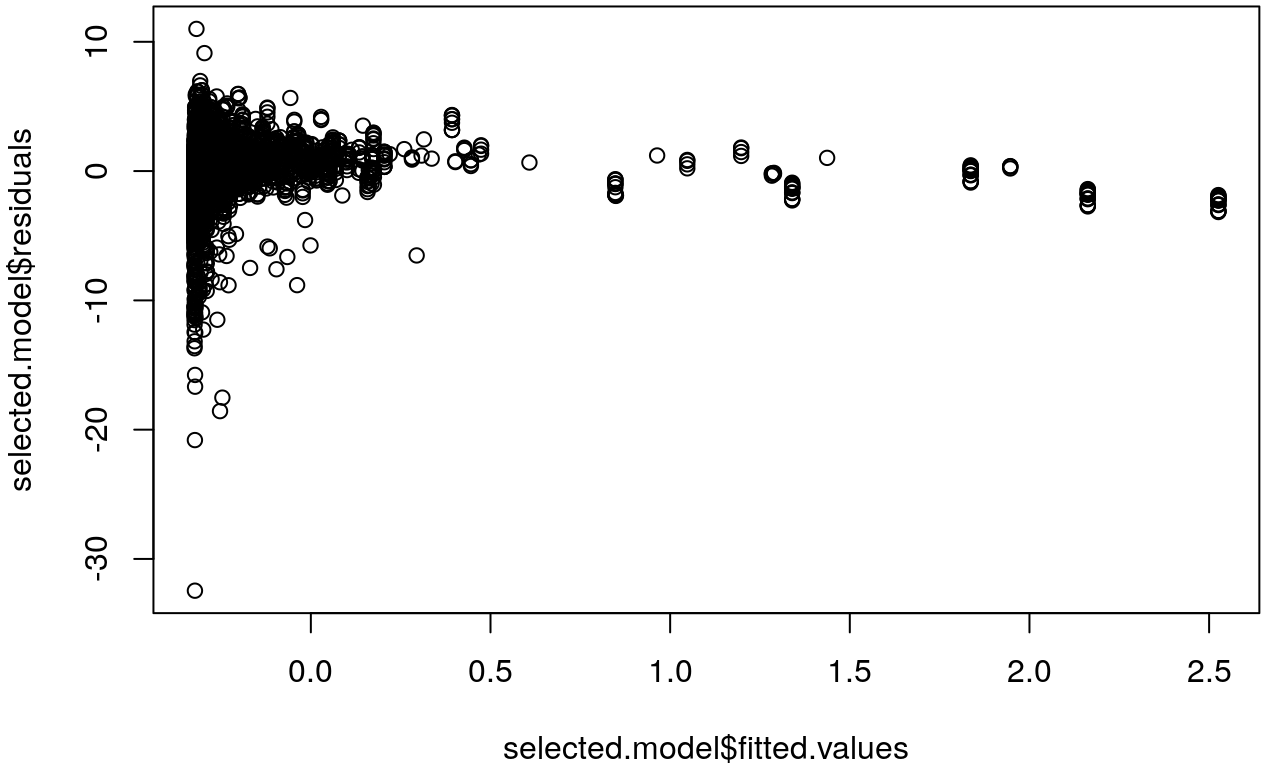

```
## Warning in ks.test(rstandard(fit.expense), "pnorm"): ties should not be present
## for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.expense)
## D = 0.08726, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.expense)
## D = 0.087287, p-value < 2.2e-16
```

```
##
## Call:
## lm(formula = success ~ net_annual_expense_ratio_fund, data = mutual_funds_outlier
s)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.7915  -0.8854   0.0966   1.0433  11.2991
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.36905    0.06178   22.16   <2e-16 ***
## net_annual_expense_ratio_fund -1.36173    0.04779  -28.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.028 on 6378 degrees of freedom
## Multiple R-squared:  0.1129, Adjusted R-squared:  0.1128
## F-statistic:   812 on 1 and 6378 DF,  p-value: < 2.2e-16
```






```
## Warning in ks.test(rstandard(fit.size), "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.size)
## D = 0.083758, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.size)
## D = 0.083783, p-value < 2.2e-16
```

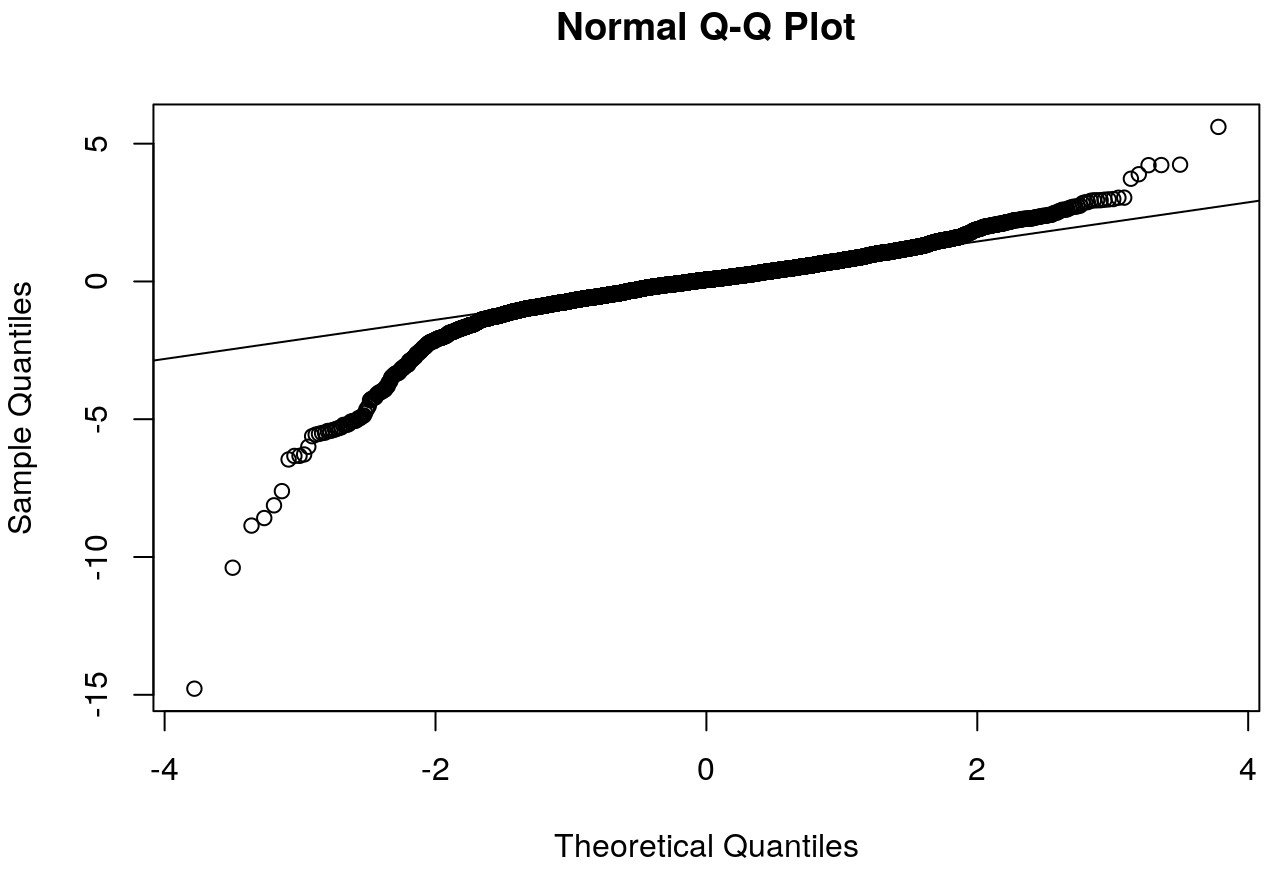
```
##
## Call:
## lm(formula = success ~ net_assets, data = mutual_funds_outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.457  -0.928   0.180   1.206  10.998
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.241e-01  2.809e-02  -11.54   <2e-16 ***
## net_assets   2.521e-11  2.460e-12   10.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.136 on 6378 degrees of freedom
## Multiple R-squared:  0.0162, Adjusted R-squared:  0.01604
## F-statistic:   105 on 1 and 6378 DF,  p-value: < 2.2e-16
```

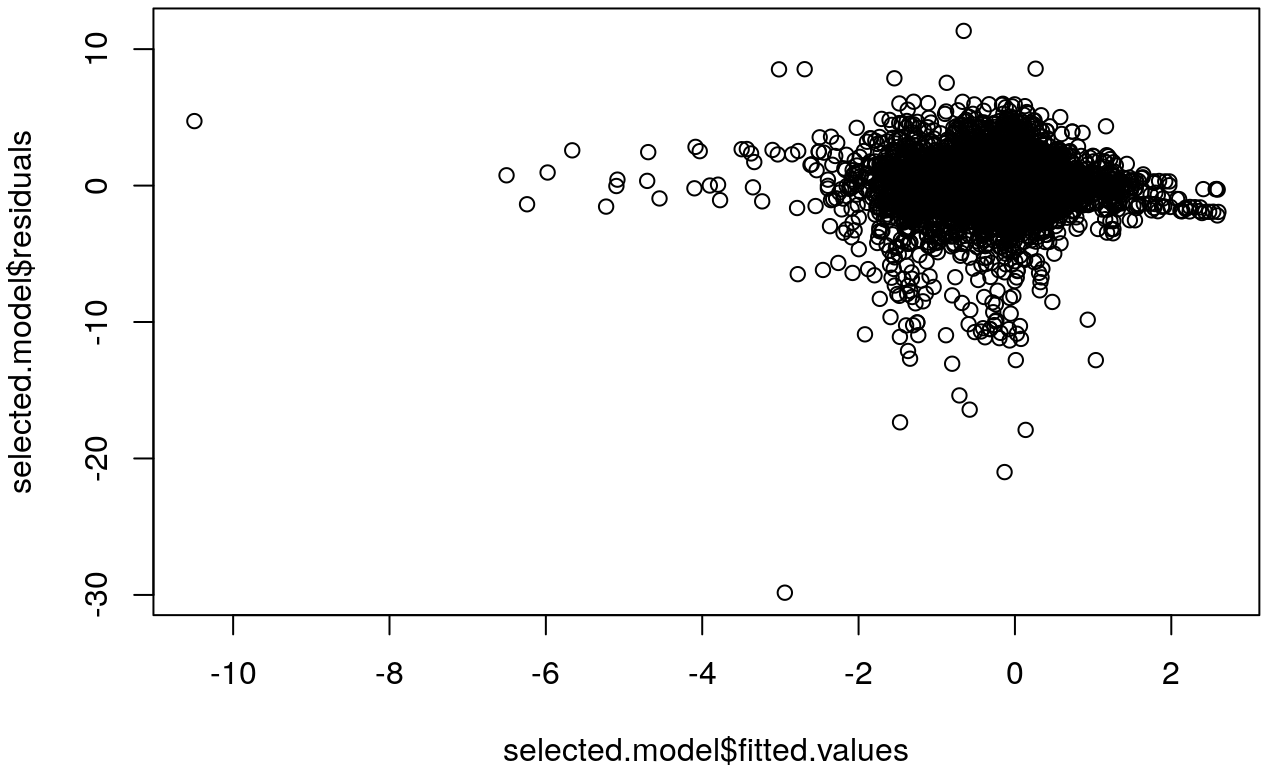
Iz rezultata vidimo da ne vrijedi pretpostavka o normalnosti reziduala. To nas sprječava da radimo daljnju analizu linearne regresije. Osim toga, ako promotrimo dobivenu R^2 vrijednost, možemo vidjeti da je ona vrlo mala, što znači da naš model ni u kojem slučaju ne bi predstavljao veći dio podataka.

Slične zaključke kao i za trošak fonda dobili smo analizom ukupne imovine pod upravljanjem fondova. Stoga smo pokušali napraviti i model regresije s te dvije varijable. Pritom je nužno utvrditi da one nisu međusobno jako korelirane.

```
## [1] -0.1550149
```

Budući da je korelacija između varijabli mala, možemo nastaviti s postupkom višestruke regresije, pritom koristeći sličan postupak kao za linearnu regresiju


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
5
0
-5
-10
-15
-4
-2
0
2
4


selected.model$residuals
selected.model$fitted.values
10
0
-10
-20
-30
-10
-8
-6
-4
-2
0
2

```
## Warning in ks.test(rstandard(fit.multi), "pnorm"): ties should not be present
## for the Kolmogorov-Smirnov test
```

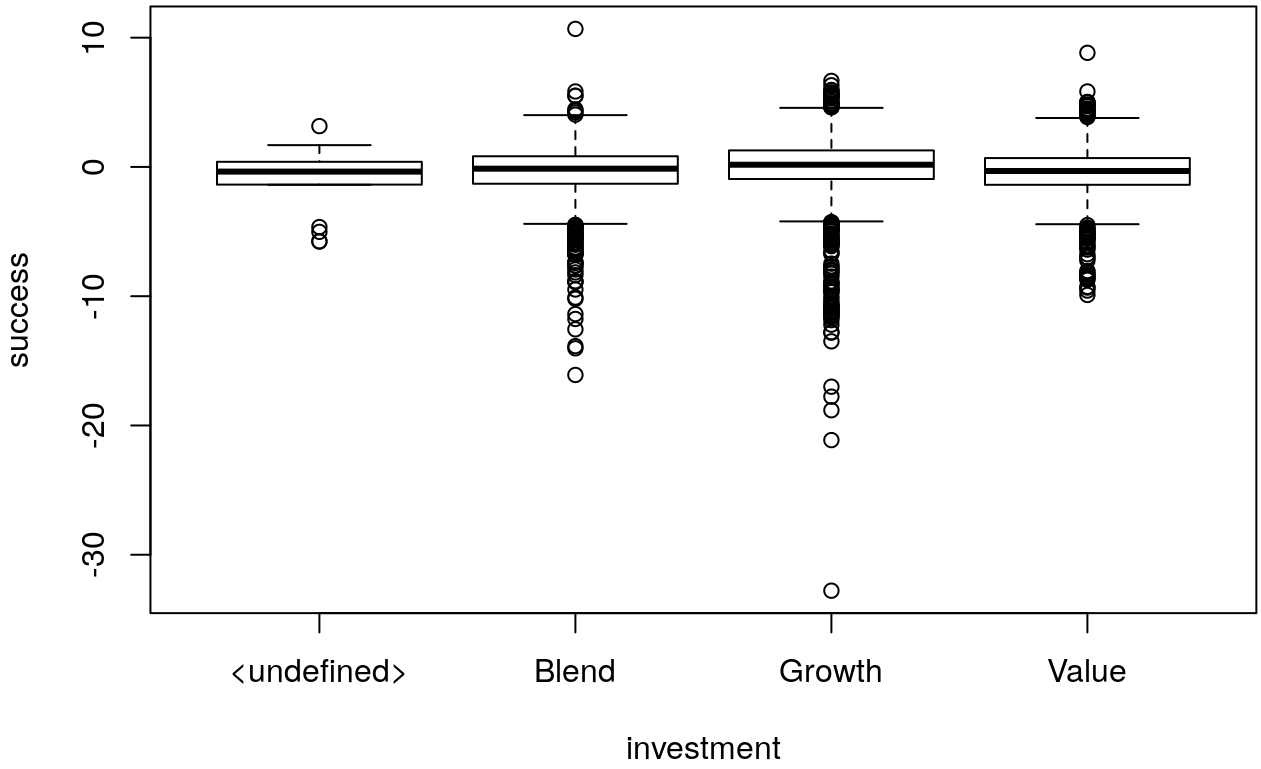
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.multi)
## D = 0.084866, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.multi)
## D = 0.084897, p-value < 2.2e-16
```

```
##
## Call:
## lm(formula = success ~ net_annual_expense_ratio_fund + net_assets,
##     data = mutual_funds_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.8368  -0.9075   0.1196   1.0297  11.3354
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.259e+00  6.390e-02  19.697  < 2e-16 ***
## net_annual_expense_ratio_fund -1.313e+00  4.822e-02 -27.239  < 2e-16 ***
## net_assets                     1.525e-11  2.357e-12   6.472 1.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.021 on 6377 degrees of freedom
## Multiple R-squared:  0.1187, Adjusted R-squared:  0.1185
## F-statistic: 429.6 on 2 and 6377 DF,  p-value: < 2.2e-16
```

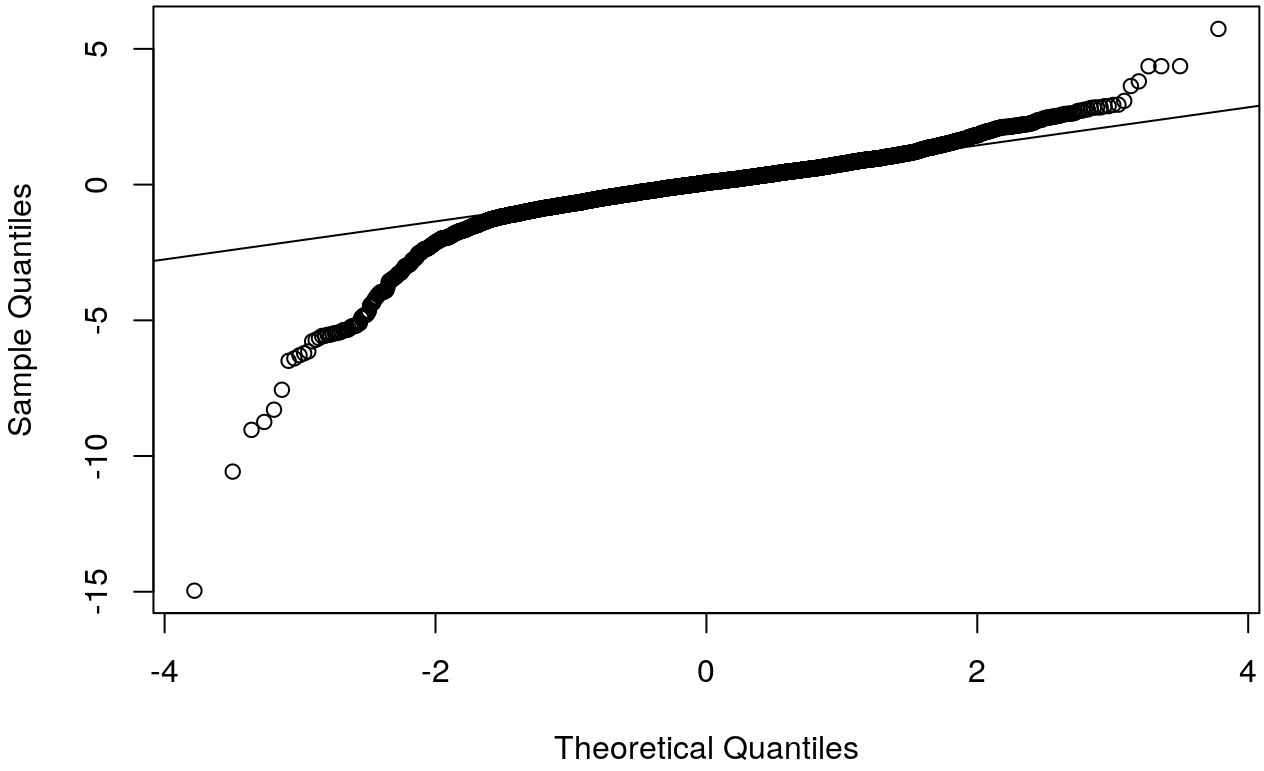
Iz rezultata vidimo da ni reziduali kod višestruke regresije nisu normalno distribuirani i da je R^2 još uvijek mali, pa stoga i ovaj model regresije možemo odbaciti

U potrazi za varijablama koje bi mogle utjecati na uspjeh, ispitali smo i regresijski model između stila investiranja i uspjeha, no ni taj proces nije dao značajne rezultate.




```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.investment)
## D = 0.091056, p-value < 2.2e-16
```

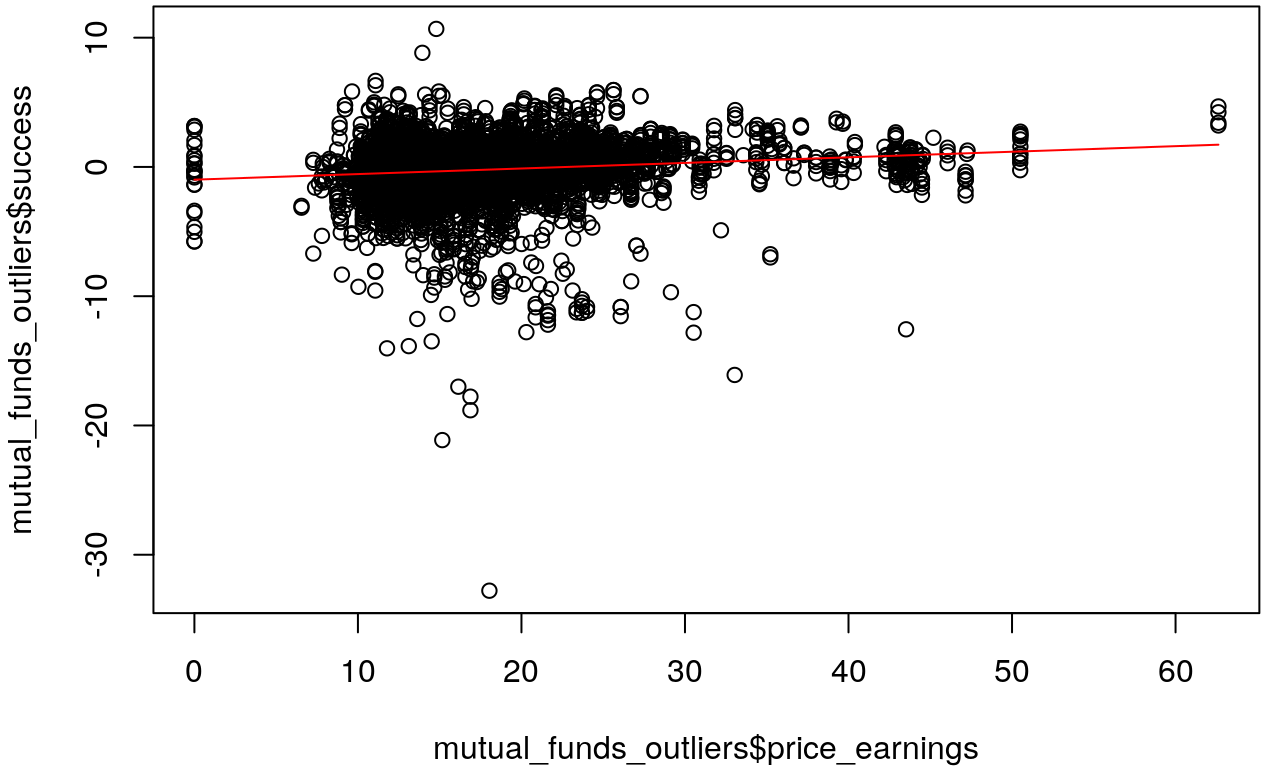
```
## Loading required package: fastDummies
```

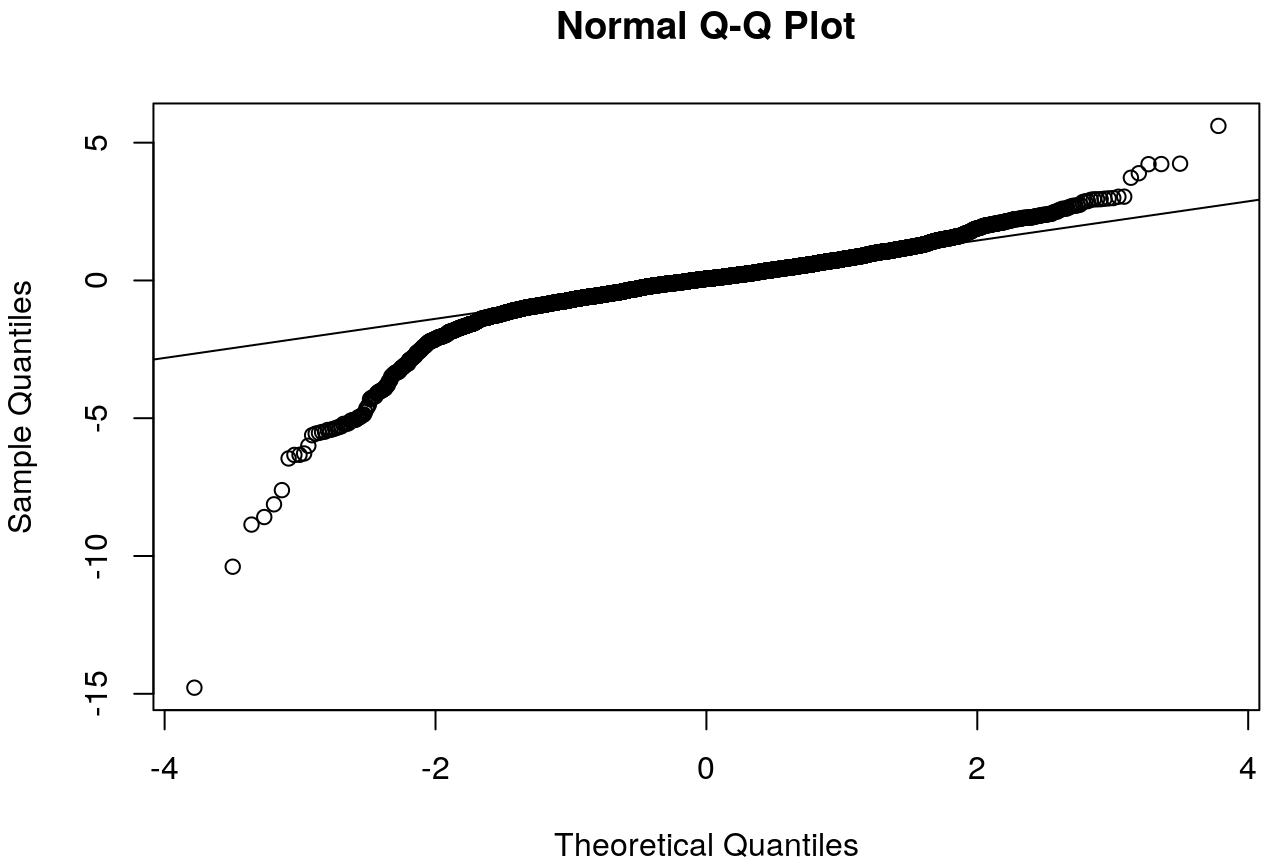


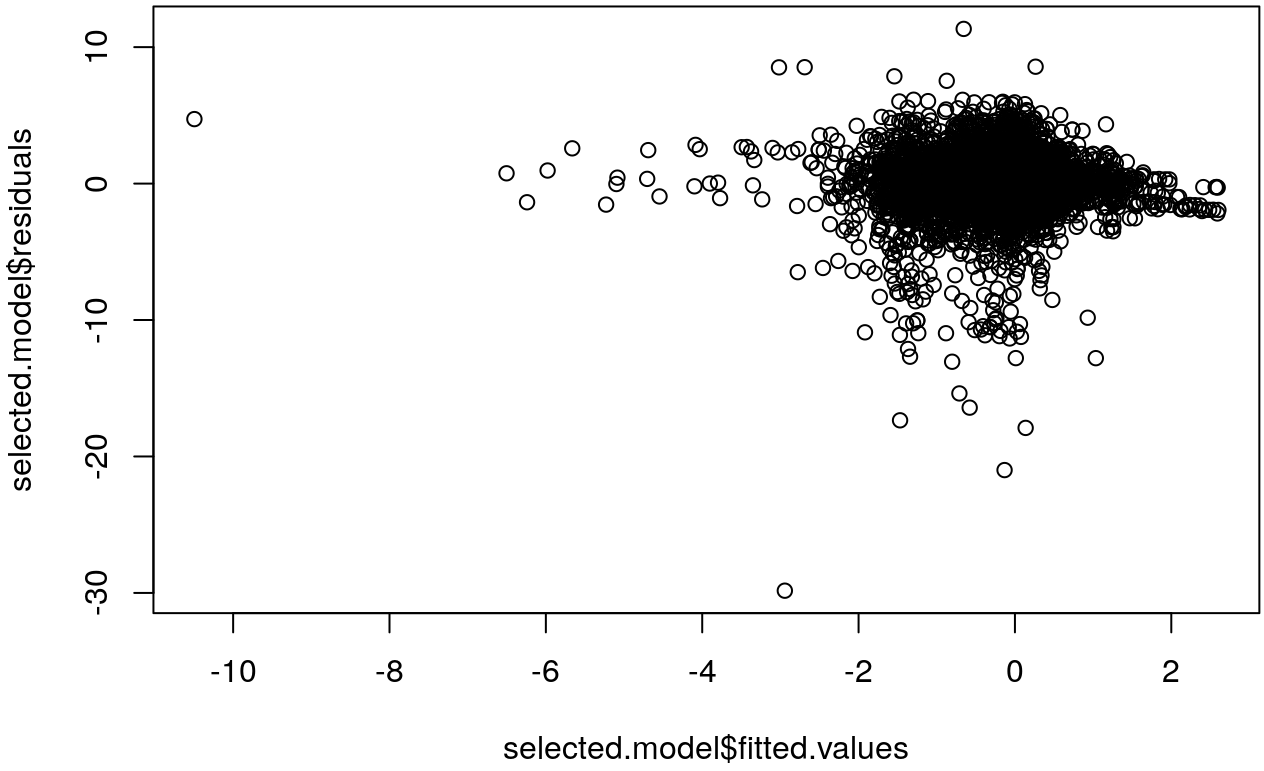

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.investment.dummy)
## D = 0.087943, p-value < 2.2e-16
```

```
##
## Call:
## lm(formula = success ~ net_annual_expense_ratio_fund + net_assets +
##     investment_Blend + investment_Growth + investment_Value,
##     data = mutual_funds_outliers_dummies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.0332  -0.8567   0.1394   1.0417  11.5250
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.730e+00  4.715e-01   3.670 0.000245 ***
## net_annual_expense_ratio_fund -1.339e+00  4.820e-02 -27.790  < 2e-16 ***
## net_assets                    1.389e-11  2.349e-12   5.912 3.56e-09 ***
## investment_Blend             -6.229e-01  4.654e-01  -1.338 0.180814
## investment_Growth            -1.920e-01  4.644e-01  -0.413 0.679343
## investment_Value             -6.439e-01  4.656e-01  -1.383 0.166714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.01 on 6374 degrees of freedom
## Multiple R-squared:  0.1291, Adjusted R-squared:  0.1284
## F-statistic:   189 on 5 and 6374 DF,  p-value: < 2.2e-16
```

Konačnu varijablu koju smo ispitali je P/E omjer, tj. omjer trenutačne tržišne cijene i zarada po dionici prethodne godine.Međutim, ni taj pokušaj nije urodio plodom.


mutual_funds_outliers$success
10
0
-10
-20
-30
0
10
20
30
40
50
60
mutual_funds_outliers$price_earnings


Normal Q-Q Plot
Sample Quantiles
5
0
-5
-10
-15
-4
-2
0
2
4
Theoretical Quantiles




```
## Warning in ks.test(rstandard(fit.pe), "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

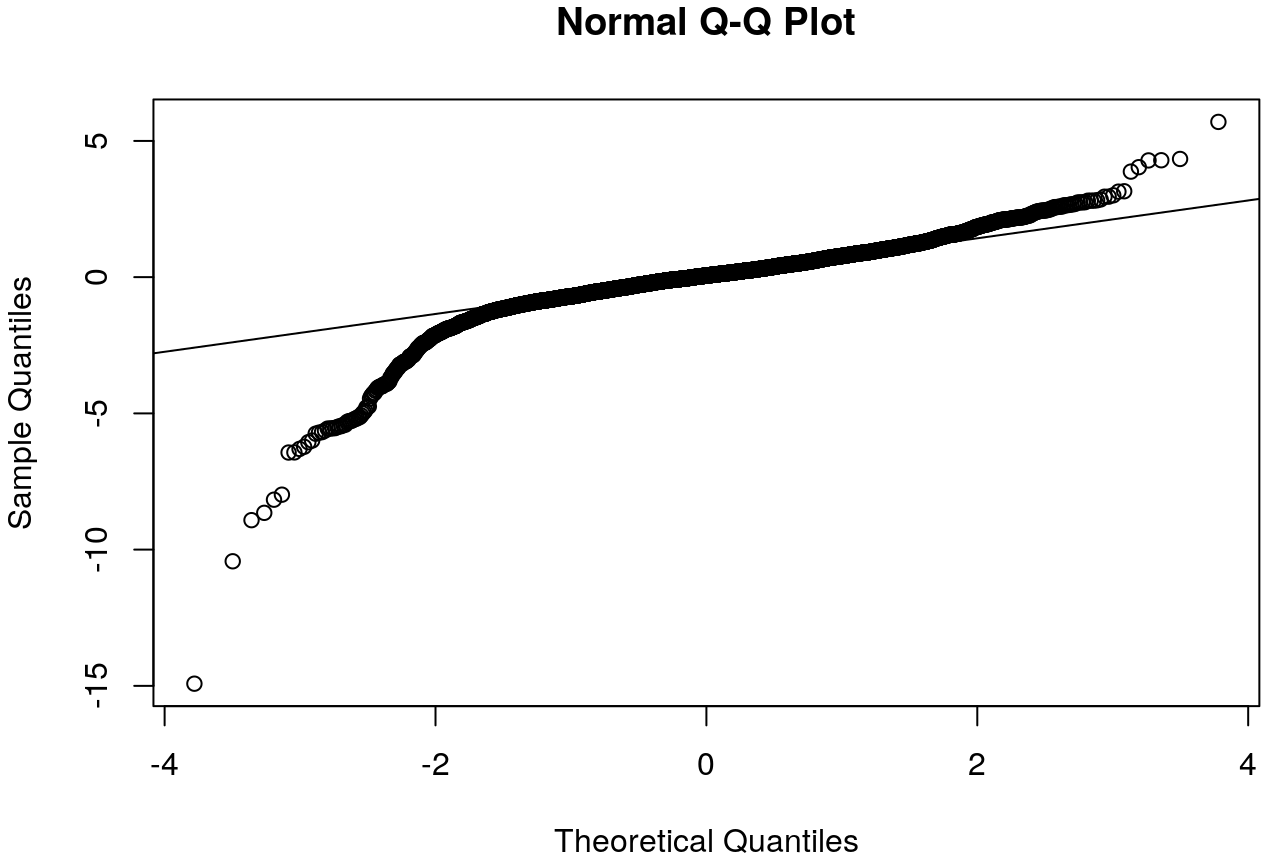
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.pe)
## D = 0.08931, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

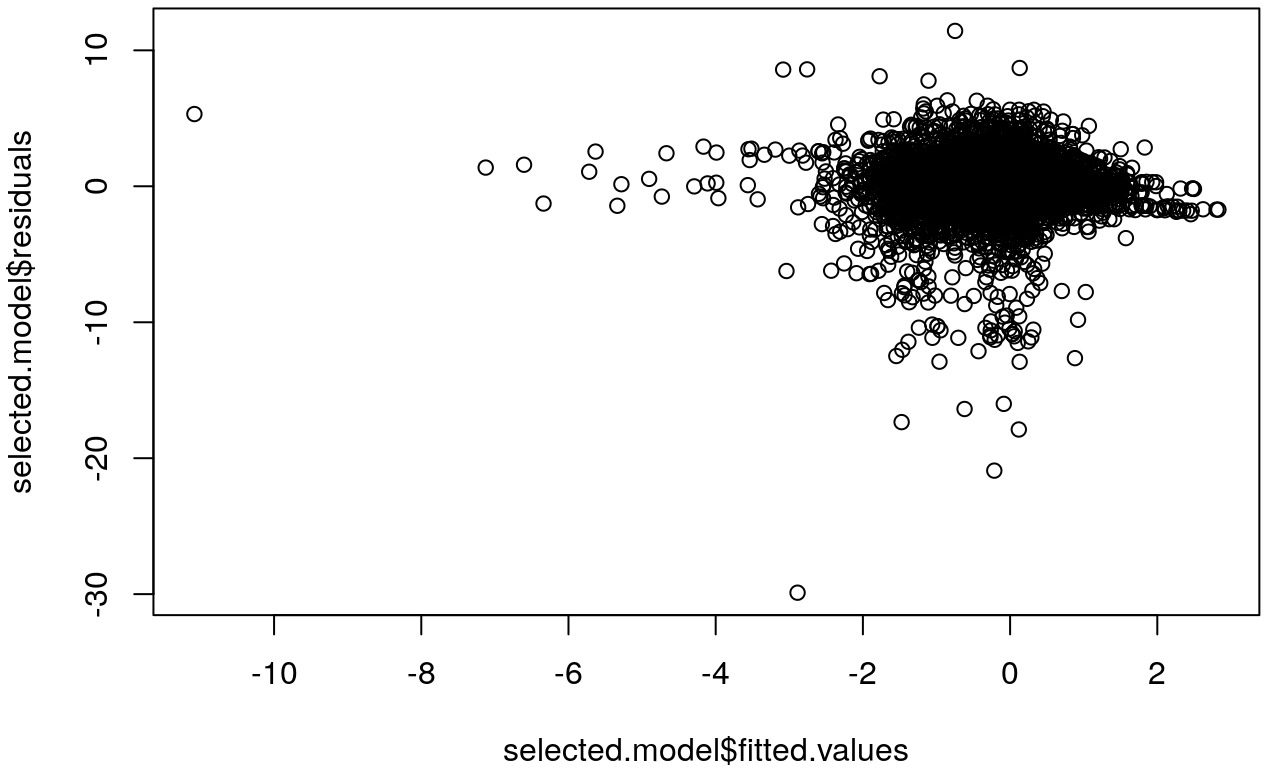
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.pe)
## D = 0.089324, p-value < 2.2e-16
```

```
##
## Call:
## lm(formula = success ~ price_earnings, data = mutual_funds_outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.577  -0.894   0.176   1.193  11.023
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.982414   0.076828  -12.79   <2e-16 ***
## price_earnings  0.043202   0.004168   10.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.135 on 6378 degrees of freedom
## Multiple R-squared:  0.01656,    Adjusted R-squared:  0.01641
## F-statistic: 107.4 on 1 and 6378 DF,  p-value: < 2.2e-16
```

Slične rezultate dobili smo ponovnim pokušajem višestruke regresije s novim dodanim parametrom.

```
##             [,1]        [,2]        [,3]
## [1,]  1.00000000 -0.15501489 -0.02883586
## [2,] -0.15501489  1.00000000  0.03444437
## [3,] -0.02883586  0.03444437  1.00000000
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
5
0
-5
-10
-15
-4
-2
0
2
4


selected.model$residuals
selected.model$fitted.values
10
0
-10
-20
-30
-10
-8
-6
-4
-2
0
2

```
## Warning in ks.test(rstandard(fit.multi), "pnorm"): ties should not be present
## for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.multi)
## D = 0.088475, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.multi)
## D = 0.088502, p-value < 2.2e-16
```

```
##
## Call:
## lm(formula = success ~ net_annual_expense_ratio_fund + price_earnings +
##     net_assets, data = mutual_funds_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.8922  -0.8613   0.1188   1.0165  11.4275
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.696e-01  9.358e-02   6.087 1.21e-09 ***
## net_annual_expense_ratio_fund -1.302e+00  4.786e-02 -27.204  < 2e-16 ***
## price_earnings                 3.924e-02  3.919e-03  10.012  < 2e-16 ***
## net_assets                     1.454e-11  2.340e-12   6.215 5.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.006 on 6376 degrees of freedom
## Multiple R-squared:  0.1324, Adjusted R-squared:  0.132
## F-statistic: 324.3 on 3 and 6376 DF,  p-value: < 2.2e-16
```

Zaključak

U konačnici, ispitivanjem raznih dostupnih varijabli nismo uspjeli pronaći model koji zadovoljava sve pretpostavke regresije i koji bi mogao dobro opisati dobivene podatke.Iako se naše testiranje na prvi pogled čini uzaludnim, to nam zapravo govori da je određivanje faktora koji čine fond uspješnim složen postupak. Da smo uspjeli pronaći jednu takvu jednostavnu regresiju, to bi nam reklo da bi u praksi odabir uspješnog fonda bio puno jednostavniji nego što uistinu je. Za kraj, trebalo bi uzeti u obzir da vjerojatno postoji još faktora koje nismo uzeli u obzir tijekom istraživanja. Razlog tomu je ograničenost podataka kojima smo baratali, a i potencijalna složenost baratanja takvim varijablama koja nadilazi naše iskustvo. Iz tog razloga, iako su rezultati dobiveni u

ovom radu zanimljivi, trebali bi se koristit ne kao činjenice, već kao poziv na detaljnije istraživanje i razradu problema kojima smo se bavili.