World Happiness Report

Opis

Podaci kojima se bavimo u ovom projektu su dobiveni kroz ankete koje provode Gallup i Lloyd's Register Foundation. Proučavat ćemo podatke iz 2020. godine koji su sadžani u 9 varijabli te podatke iz 2021. godine koji su sadržani u 11 varijabli. Temeljna varijabla je osjećaj sreće prema Cantrilovoj ljestvici gdje su ispitanici ocjenjivali zadovoljstvo vlastitog života na skali od 0 do 10. Vrijednost varijable je prosjek reprezentativnog uzorka pojedine zemlje. Uz to podaci sadrže varijable kao što su BDP po stanovniku, životni vijek, socijalna podrška, percepcija korupcije, doniranje novca u dobrotvorne svrhe, nejednakost dohotka i slično.

```
# Učitavanje podataka iz csv datoteke:
whr2020 = read.table("WHR_2020.csv", sep = ",")
whr2021 = read.table("WHR_2021.csv", sep = ",")
dim(whr2020)
```

```
## [1] 153   9
```

```
dim(whr2021)
```

```
## [1] 149  11
```

Summary podataka:

```
## [1] "2020: "
```

```
##        V3              V4               V5               V6
##  Min.   :2.567   Min.   : 6.493   Min.   :0.3190   Min.   :45.20
##  1st Qu.:4.724   1st Qu.: 8.351   1st Qu.:0.7370   1st Qu.:58.96
##  Median :5.515   Median : 9.456   Median :0.8290   Median :66.31
##  Mean   :5.473   Mean   : 9.296   Mean   :0.8087   Mean   :64.45
##  3rd Qu.:6.228   3rd Qu.:10.265   3rd Qu.:0.9070   3rd Qu.:69.29
##  Max.   :7.809   Max.   :11.451   Max.   :0.9750   Max.   :76.81
##        V7               V8                V9
##  Min.   :0.3970   Min.   :-0.30100   Min.   :0.1100
##  1st Qu.:0.7150   1st Qu.:-0.12700   1st Qu.:0.6830
##  Median :0.8000   Median :-0.03400   Median :0.7830
##  Mean   :0.7834   Mean   :-0.01454   Mean   :0.7331
##  3rd Qu.:0.8780   3rd Qu.: 0.08500   3rd Qu.:0.8490
##  Max.   :0.9750   Max.   : 0.56100   Max.   :0.9360
```

```
## [1] "2021: "
```

```
##        V3              V4               V5               V6
##  Min.   :2.523   Min.   : 6.635   Min.   :0.4630   Min.   :48.48
##  1st Qu.:4.852   1st Qu.: 8.541   1st Qu.:0.7500   1st Qu.:59.80
##  Median :5.534   Median : 9.569   Median :0.8320   Median :66.60
##  Mean   :5.533   Mean   : 9.432   Mean   :0.8147   Mean   :64.99
##  3rd Qu.:6.255   3rd Qu.:10.421   3rd Qu.:0.9050   3rd Qu.:69.60
##  Max.   :7.842   Max.   :11.647   Max.   :0.9830   Max.   :76.95
##        V7               V8                V9
##  Min.   :0.3820   Min.   :-0.28800   Min.   :0.0820
```

```
##  1st Qu.:0.7180   1st Qu.:-0.12600   1st Qu.:0.6670
##  Median :0.8040   Median :-0.03600   Median :0.7810
##  Mean   :0.7916   Mean   :-0.01513   Mean   :0.7274
##  3rd Qu.:0.8770   3rd Qu.: 0.07900   3rd Qu.:0.8450
##  Max.   :0.9700   Max.   : 0.54200   Max.   :0.9390
```

```r
names(whr2020)
```

```
## [1] "Country name"                 "Regional indicator"
## [3] "Ladder score"                 "Logged GDP per capita"
## [5] "Social support"               "Healthy life expectancy"
## [7] "Freedom to make life choices" "Generosity"
## [9] "Perceptions of corruption"
```

```r
names(whr2021)
```

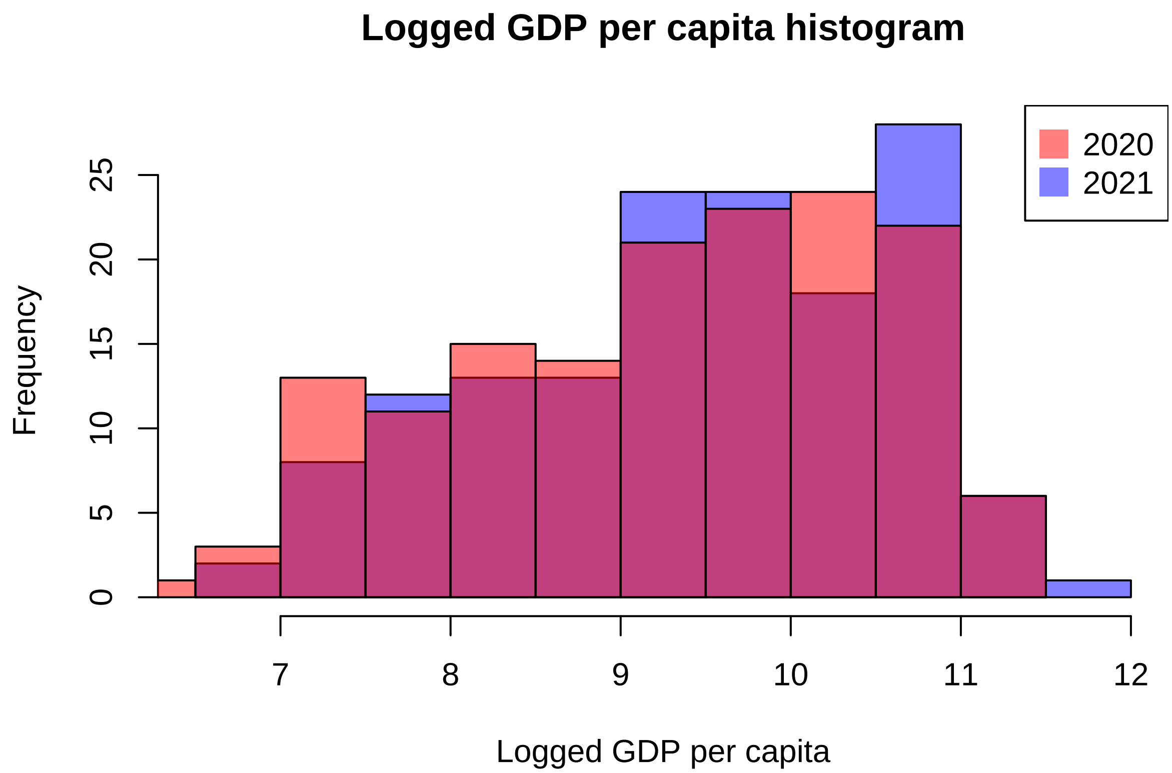
```
##  [1] "Country name"                 "Regional indicator"
##  [3] "Ladder score"                 "Logged GDP per capita"
##  [5] "Social support"               "Healthy life expectancy"
##  [7] "Freedom to make life choices" "Generosity"
##  [9] "Perceptions of corruption"    "Income Gini"
## [11] "Wealth Gini"
```

Deskriptivna statistika

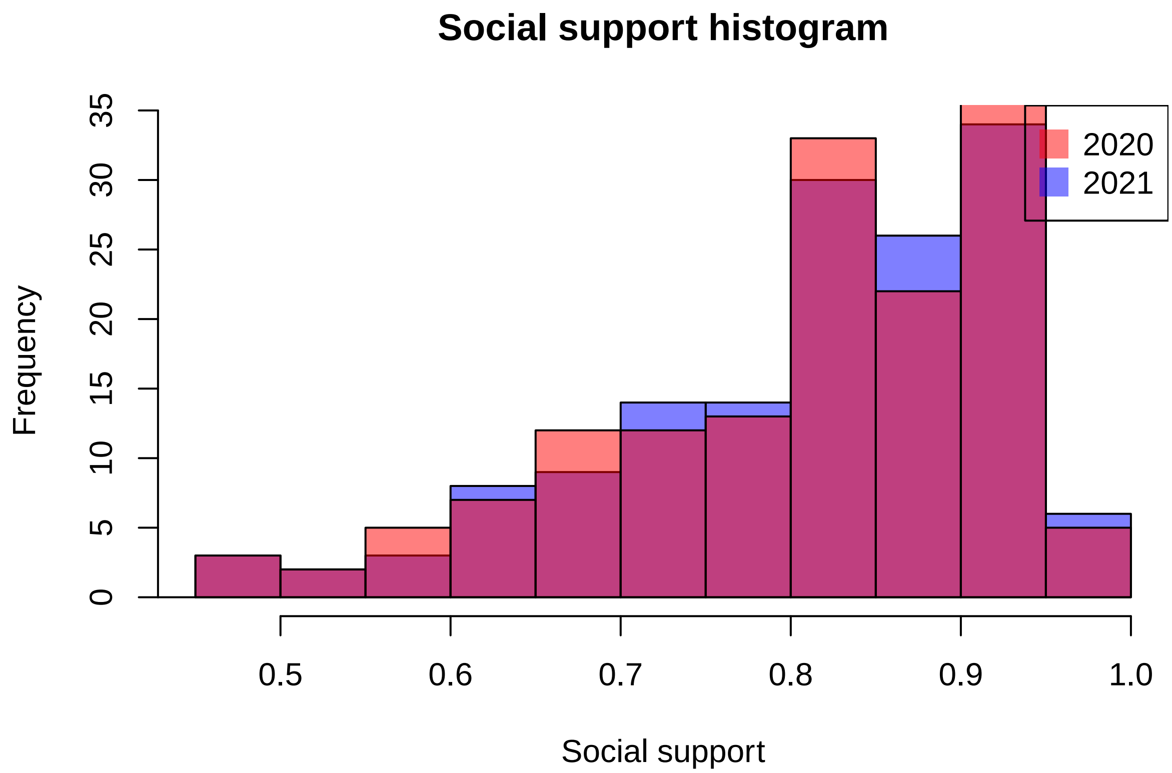
Prikažimo sada histograme usporedbe varijabli za različite godine.

```r
#histogrami varijable s obzirom na godine
plot_by_years <- function(column, main) {
  hist(whr2021[[column]], breaks=15, main=main, xlab=column, ylab="Frequency",col=rgb(0,0,1,0.5))
  hist(whr2020[[column]], breaks=15, main=main, xlab=column, ylab="Frequency", col=rgb(1,0,0,0.5), add=
  legend(x="topright", c("2020", "2021"), col=c(rgb(1,0,0,0.5),
                                                rgb(0,0,1,0.5)), pt.cex = 2, pch = 15)
}
plot_by_years("Ladder score", "Ladder score histogram")
```

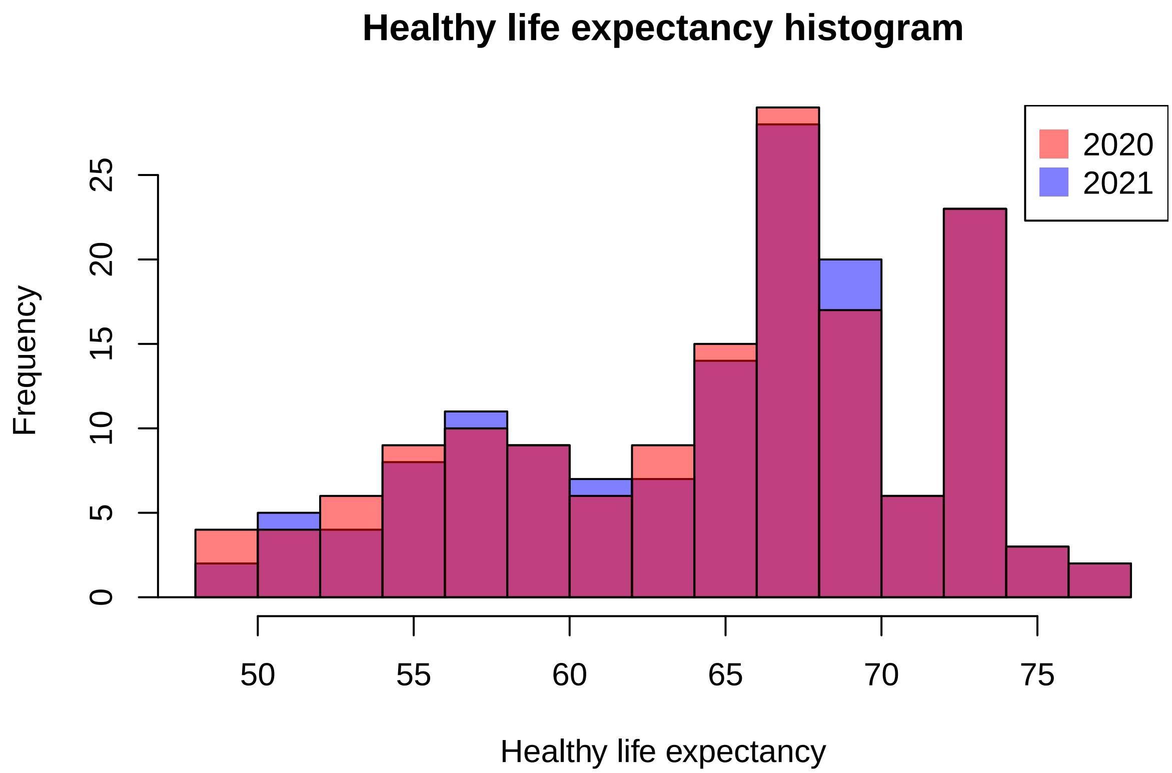

```r
plot_by_years("Logged GDP per capita", "Logged GDP per capita histogram")
```



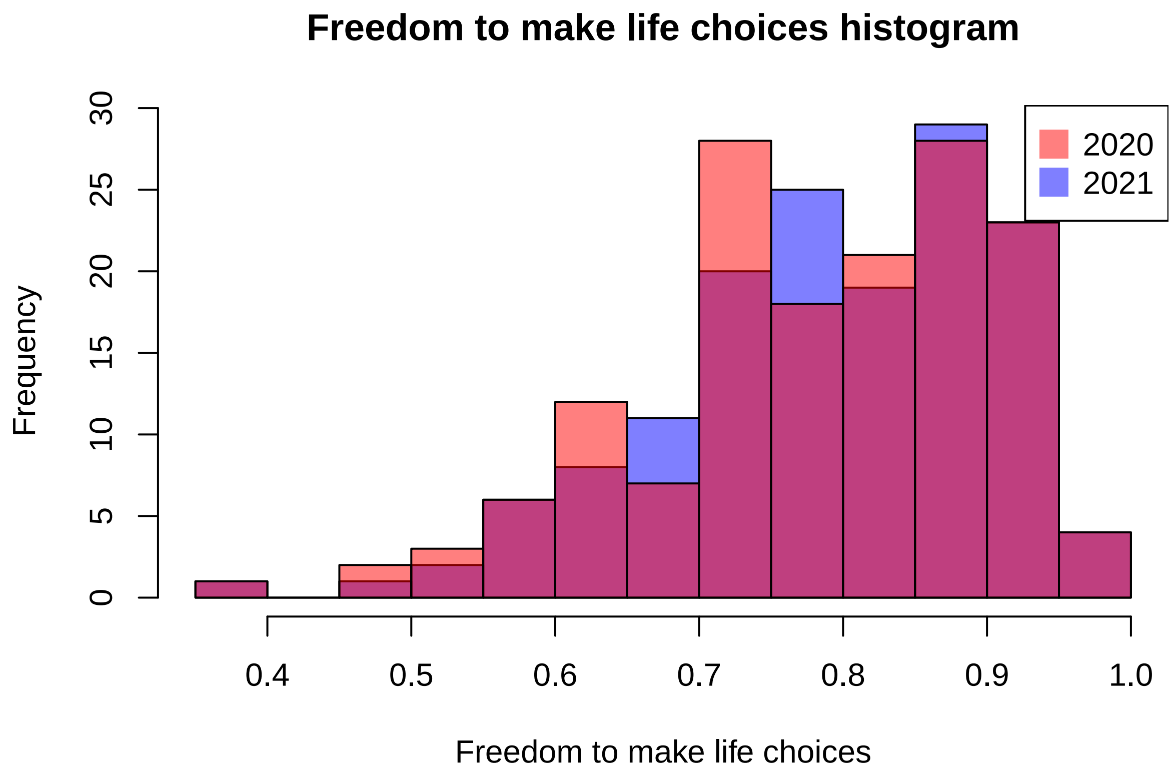

```
plot_by_years("Social support", "Social support histogram")
```



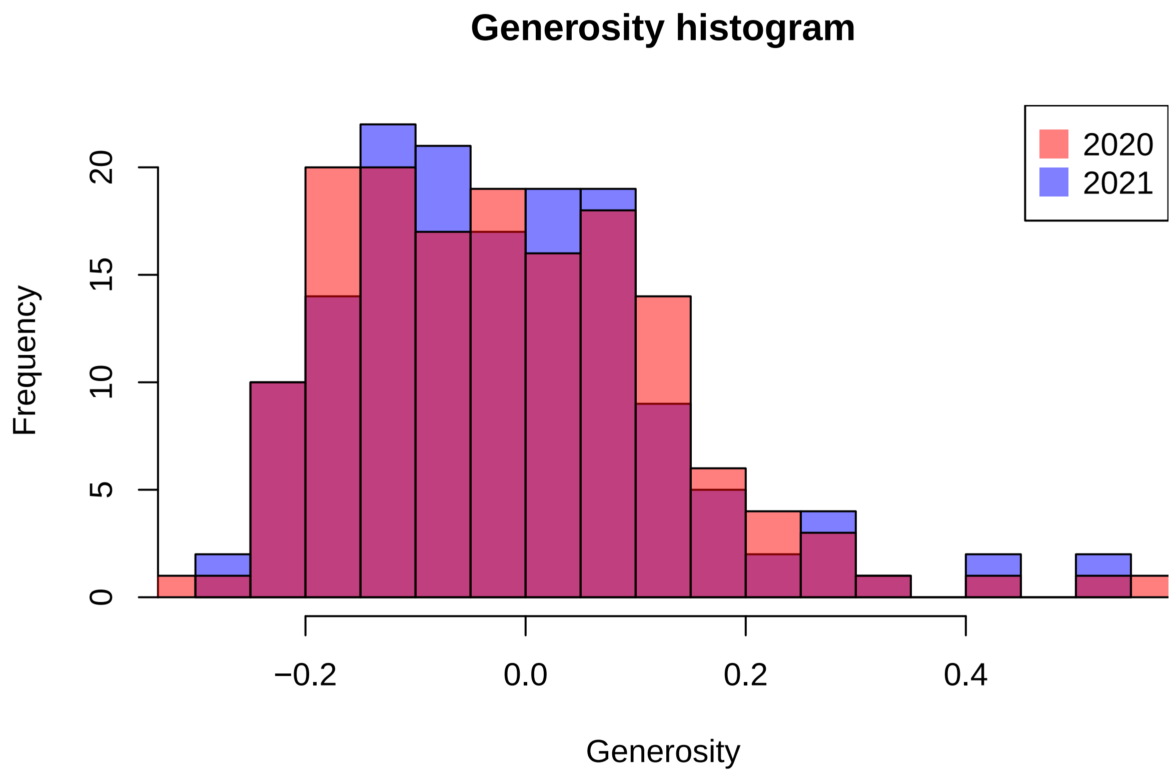

```
plot_by_years("Healthy life expectancy", "Healthy life expectancy histogram")
```



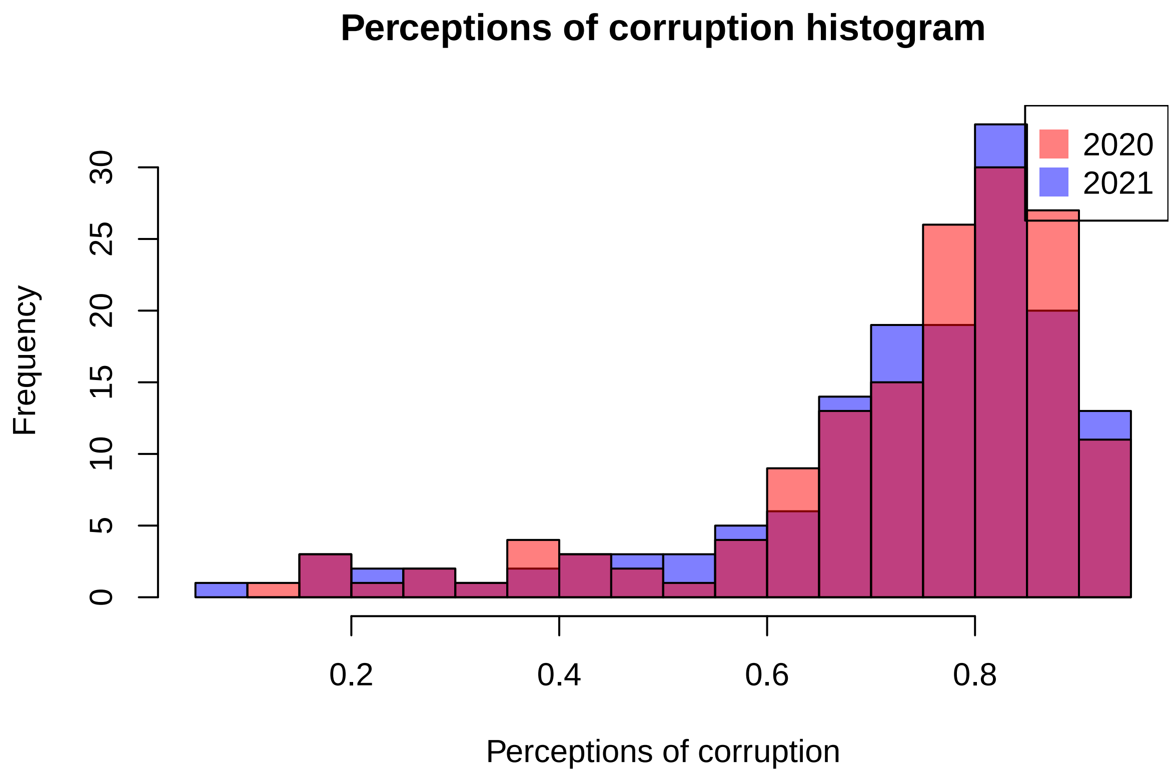

```
plot_by_years("Freedom to make life choices", "Freedom to make life choices histogram")
```



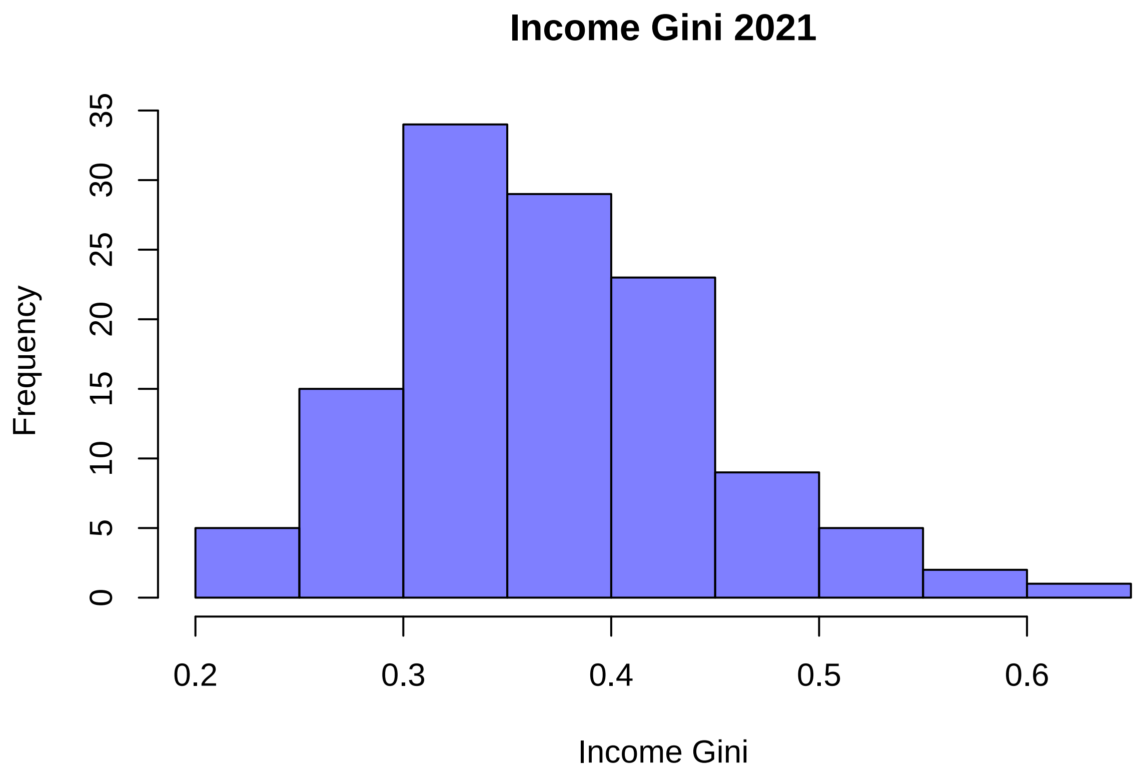

```
plot_by_years("Generosity", "Generosity histogram")
```



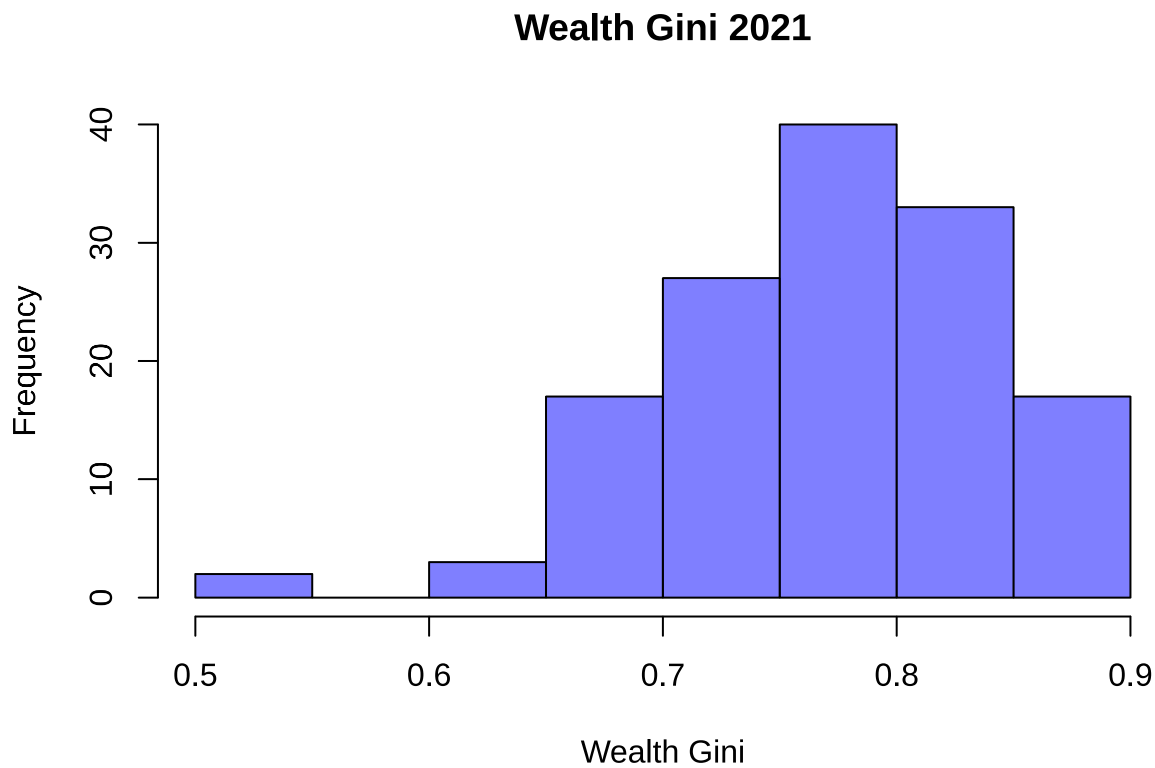

```
plot_by_years("Perceptions of corruption", "Perceptions of corruption histogram")
```




```
hist(whr2021$`Income Gini`, breaks=10, main="Income Gini 2021", xlab="Income Gini", ylab="Frequency",co
```




```
hist(whr2021$`Wealth Gini`, breaks=10, main="Wealth Gini 2021", xlab="Wealth Gini", ylab="Frequency",col
```



Iz dobivenih histograma vidljivo je da postoje promjene u varijablama za različite godine, no raspodjela podataka je veoma slična za obje godine. Također se može naslutiti da većina podataka nije normalno distribuirana.

Izračunajmo srednje vrijednosti i medijane Ladder score-ova po regijama.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
whr2021 %>% group_by(`Regional indicator`) %>% summarise(
          Mean.LadderScore = mean(`Ladder score`),
          Mean.GDP = mean(`Logged GDP per capita`),
          Mean.SocialSupport = mean(`Social support`),
          Mean.LifeExp = mean(`Healthy life expectancy`),
          Mean.Freedom = mean(`Freedom to make life choices`),
          Mean.Generosity = mean(Generosity),
          Mean.Corruption = mean(`Perceptions of corruption`),
          Len = length(`Ladder score`)
           #Mean.IncomeGini = mean(`Income Gini`),
```

```
          #Mean.WealthGini = mean(`Wealth Gini`)
           ) -> summary.result1
summary.result1
```

```
## # A tibble: 10 x 9
##    `Regional indicator`  Mean.LadderScore Mean.GDP Mean.SocialSupp~ Mean.LifeExp
##    <chr>                            <dbl>    <dbl>            <dbl>        <dbl>
##  1 Central and Eastern ~             5.98    10.1             0.887         68.3
##  2 Commonwealth of Inde~             5.47     9.40            0.872         65.0
##  3 East Asia                         5.81    10.4             0.860         71.3
##  4 Latin America and Ca~             5.91     9.37            0.840         67.1
##  5 Middle East and Nort~             5.22     9.67            0.798         65.6
##  6 North America and ANZ             7.13    10.8             0.934         72.3
##  7 South Asia                        4.44     8.68            0.703         62.7
##  8 Southeast Asia                    5.41     9.42            0.820         64.9
##  9 Sub-Saharan Africa                4.49     8.08            0.697         55.9
## 10 Western Europe                    6.91    10.8             0.914         73.0
## # ... with 4 more variables: Mean.Freedom <dbl>, Mean.Generosity <dbl>,
## #   Mean.Corruption <dbl>, Len <int>
```

```
whr2021 %>% group_by(`Regional indicator`) %>% summarise(
          Med.LadderScore = median(`Ladder score`),
          Med.GDP = median(`Logged GDP per capita`),
          Med.SocialSupport = median(`Social support`),
          Med.LifeExp = median(`Healthy life expectancy`),
          Med.Freedom = median(`Freedom to make life choices`),
          Med.Generosity = median(Generosity),
          Med.Corruption = median(`Perceptions of corruption`)
           ) -> summary.result2
summary.result2
```

```
## # A tibble: 10 x 8
##    `Regional indicator`     Med.LadderScore Med.GDP Med.SocialSuppo~ Med.LifeExp
##    <chr>                              <dbl>   <dbl>            <dbl>       <dbl>
##  1 Central and Eastern Eur~            6.08   10.3             0.924        68.6
##  2 Commonwealth of Indepen~            5.47    9.53            0.891        65.1
##  3 East Asia                           5.76   10.6             0.86         71.8
##  4 Latin America and Carib~            5.99    9.45            0.857        67.6
##  5 Middle East and North A~            4.89    9.58            0.826        66.6
##  6 North America and ANZ               7.14   10.8             0.933        73.6
##  7 South Asia                          4.93    8.46            0.693        64.2
##  8 Southeast Asia                      5.38    9.08            0.817        62.2
##  9 Sub-Saharan Africa                  4.62    7.93            0.709        56.2
## 10 Western Europe                      7.08   10.8             0.934        72.7
## # ... with 3 more variables: Med.Freedom <dbl>, Med.Generosity <dbl>,
## #   Med.Corruption <dbl>
```

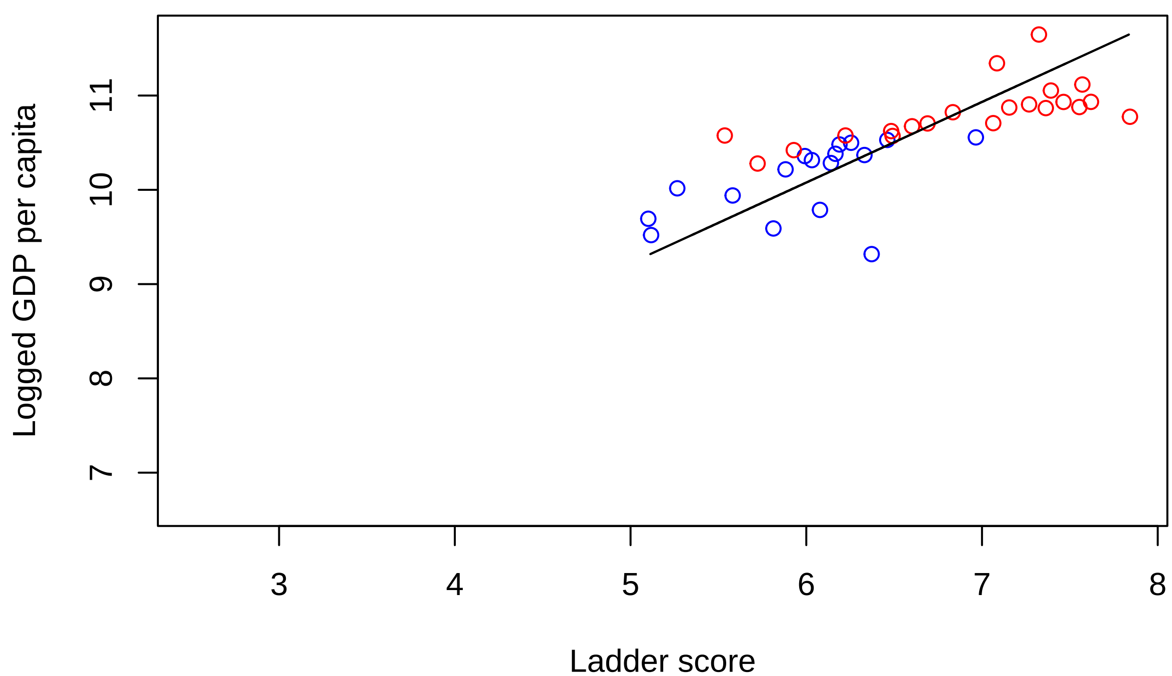
Promatrajući varijable u 2021. godini vidimo da su vrijednosti podataka u svim varijablama (osim kod varijable za percepciju korupcije") veće za Zapadnu Europu u usporedbi s Centralnom i Istočnom Europom.

Povezanost između Ladder score i Logged GDP per capita

Možemo li iz dijagrama raspršenja možda naslutiti kakvu vezu između Ladder score i GDP per capita? Posebno ćemo istaknuti 3 regije na dijagramu (Zapadnu Europu, Srednju i Istočnu Europu i Sub-Saharsku

Afriku).

```
#trebalo bi i ispitat jesu li reziduali iz normalne
ce_europe = whr2021[whr2021$`Regional indicator` == "Central and Eastern Europe",]
w_europe = whr2021[whr2021$`Regional indicator` == "Western Europe",]
europe <- rbind(ce_europe, w_europe)

fitGDPWestEast = lm(europe$`Ladder score` ~ europe$`Logged GDP per capita`)

# Razlikujemo vrste regija:
plot(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Central and Eastern Europe'],
     whr2021$`Logged GDP per capita`[whr2021$`Regional indicator`=='Central and Eastern Europe'],
     col='blue',
     xlim=c(min(whr2021$`Ladder score`),max(whr2021$`Ladder score`)),
     ylim=c(min(whr2021$`Logged GDP per capita`),max(whr2021$`Logged GDP per capita`)),
     xlab='Ladder score',
     ylab='Logged GDP per capita')

points(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Western Europe'],
     whr2021$`Logged GDP per capita`[whr2021$`Regional indicator`=='Western Europe'],col='red')
lines(fitGDPWestEast$fitted.values, europe$`Logged GDP per capita`)
```




```
summary(fitGDPWestEast)
```

```
##
## Call:
## lm(formula = europe$`Ladder score` ~ europe$`Logged GDP per capita`)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04766 -0.28308 -0.09155  0.34764  1.25992
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -5.7879     1.6494  -3.509  0.00123 **
## europe$`Logged GDP per capita`   1.1698     0.1569   7.457 8.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4747 on 36 degrees of freedom
## Multiple R-squared:  0.607,  Adjusted R-squared:  0.5961
## F-statistic: 55.61 on 1 and 36 DF,  p-value: 8.321e-09
```

Iz dijagrama raspršenja vidljiva je moguća povezanosti Ladder score s GDP per capita. Linearnom regresijom potvrđujemo povezanost između varijabli Logged GDP per capita i Ladder score zbog značajnog R-squared i testova o koeficijentima β_0 i β_1. Također vidimo da se na dijagramu razlikuju vrijednosti Zapadne i Srednje i Istočne Europe.

Jesu li ljudi u Zapadnoj Europi sretniji od ljudi u Srednjoj i Istočnoj Europi?

```
western_europe = whr2021[whr2021$`Regional indicator` == "Western Europe",]
central_eastern_europe = whr2021[whr2021$`Regional indicator` == "Central and Eastern Europe",]
```

```
cat('Prosječan Ladder score zemalja iz Zapadne Europe ', mean(western_europe$`Ladder score`),'\n')
```

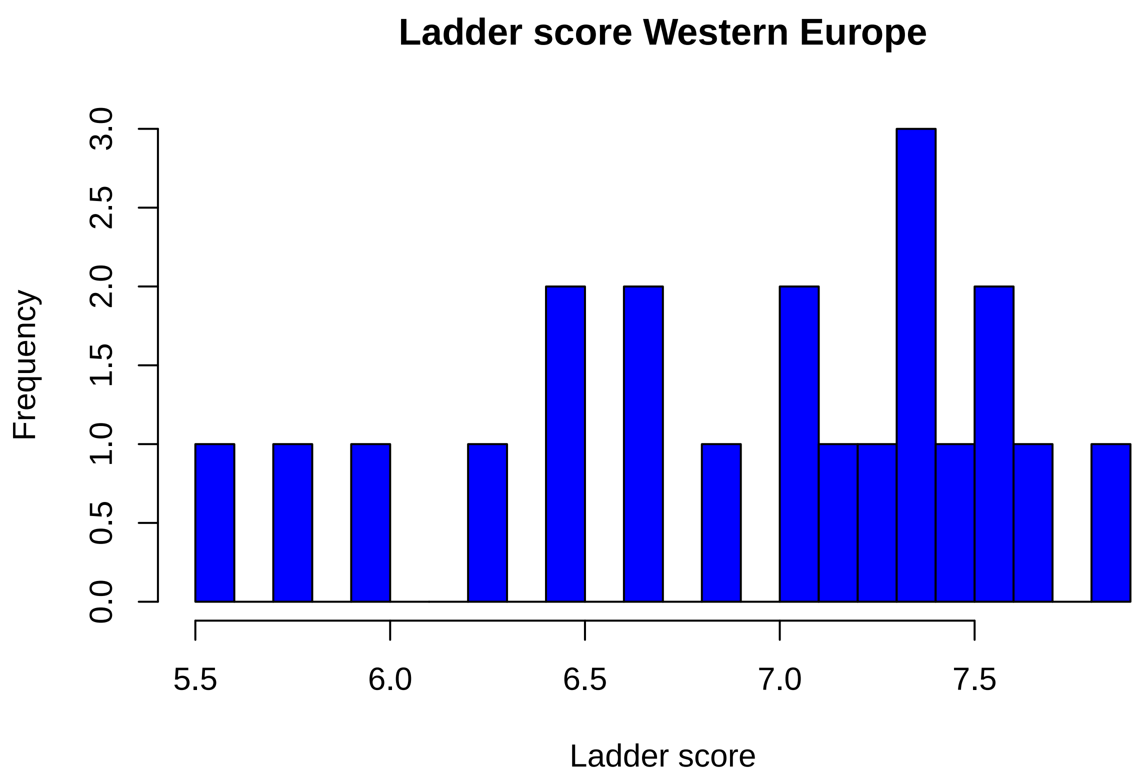
```
## Prosjecan Ladder score zemalja iz Zapadne Europe  6.914905
```

```
cat('Prosječan Ladder score zemalja iz Srednje i Istočne Europe', mean(central_eastern_europe$`Ladder sc
```

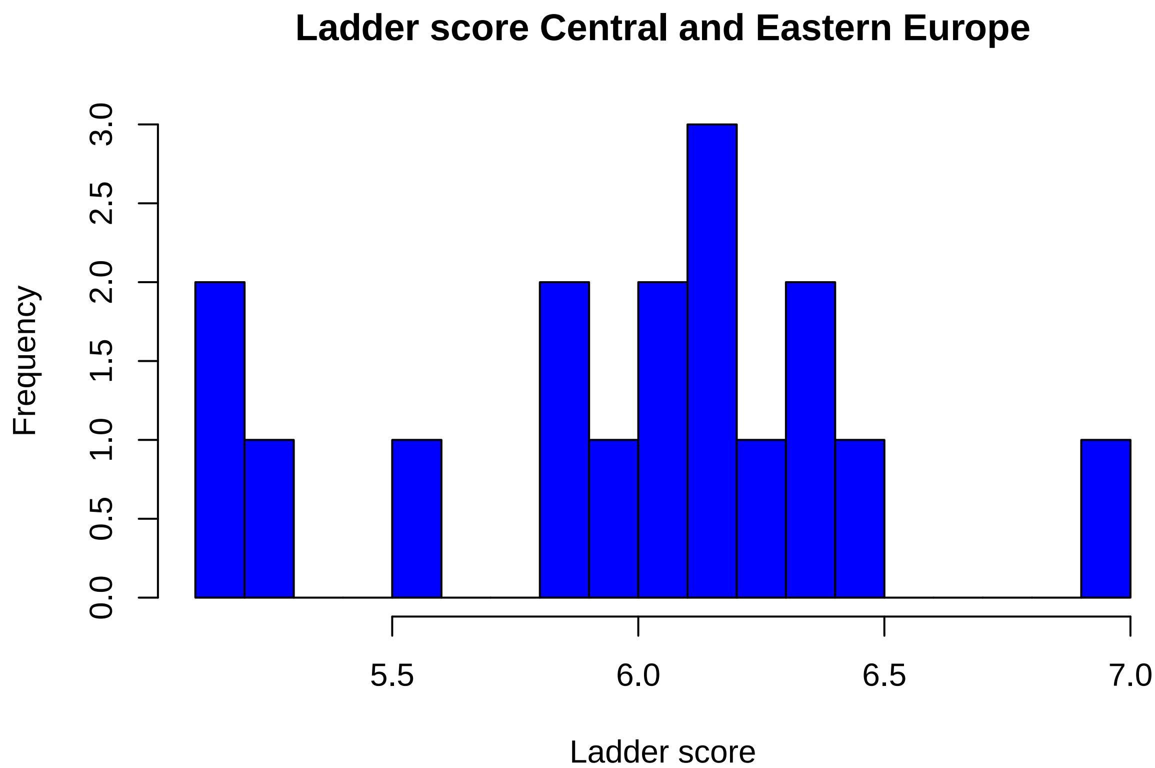
```
## Prosjecan Ladder score zemalja iz Srednje i Istocne Europe 5.984765
```

Histogrami za za Zapadnu i Centralnu/Istočnu Europu:

```
h = hist(western_europe$`Ladder score`,
         main="Ladder score Western Europe",
         xlab="Ladder score",
         ylab='Frequency',
         col="blue",
         breaks = 20
         )
```

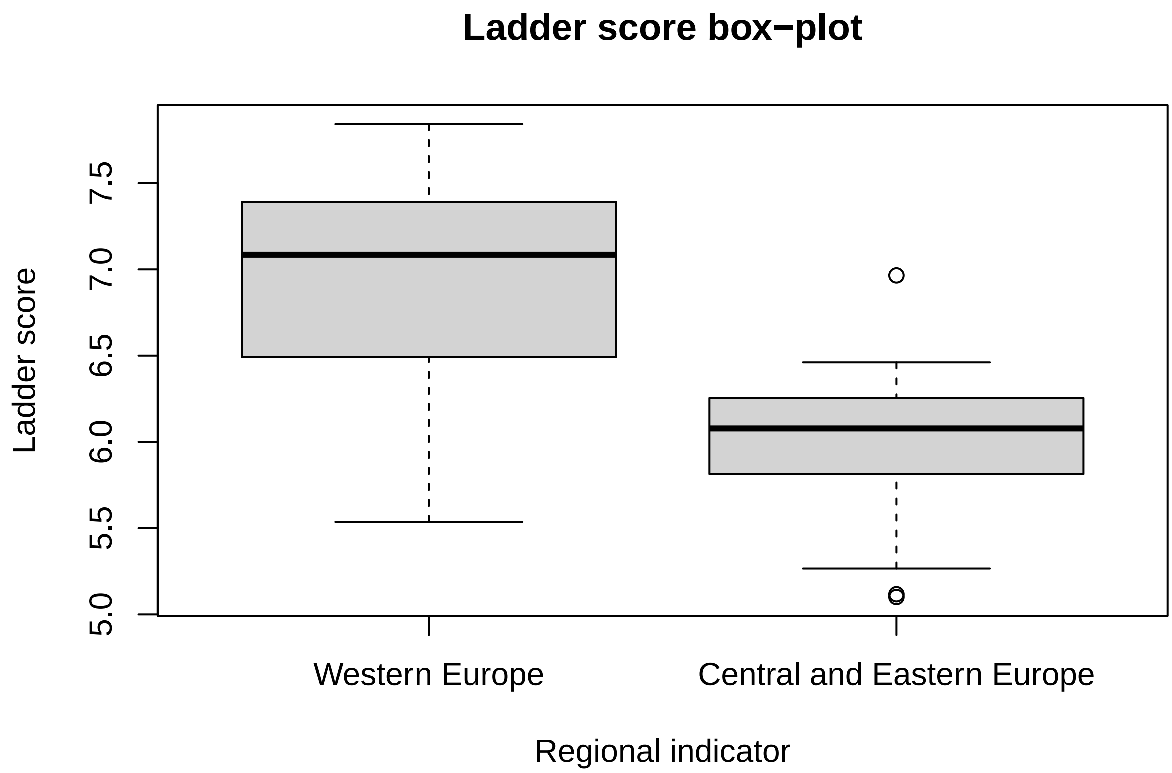



```
h = hist(central_eastern_europe$`Ladder score`,
         main="Ladder score Central and Eastern Europe",
         xlab="Ladder score",
         ylab='Frequency',
         col="blue",
         breaks = 20
         )
```



Pravokutni dijagram za Zapadnu i Centralnu/Istočnu Europu:

```
boxplot(western_europe$`Ladder score`,central_eastern_europe$`Ladder score`,
        main='Ladder score box-plot',
        ylab='Ladder score', xlab="Regional indicator", names = c("Western Europe", "Central and Eastern
```

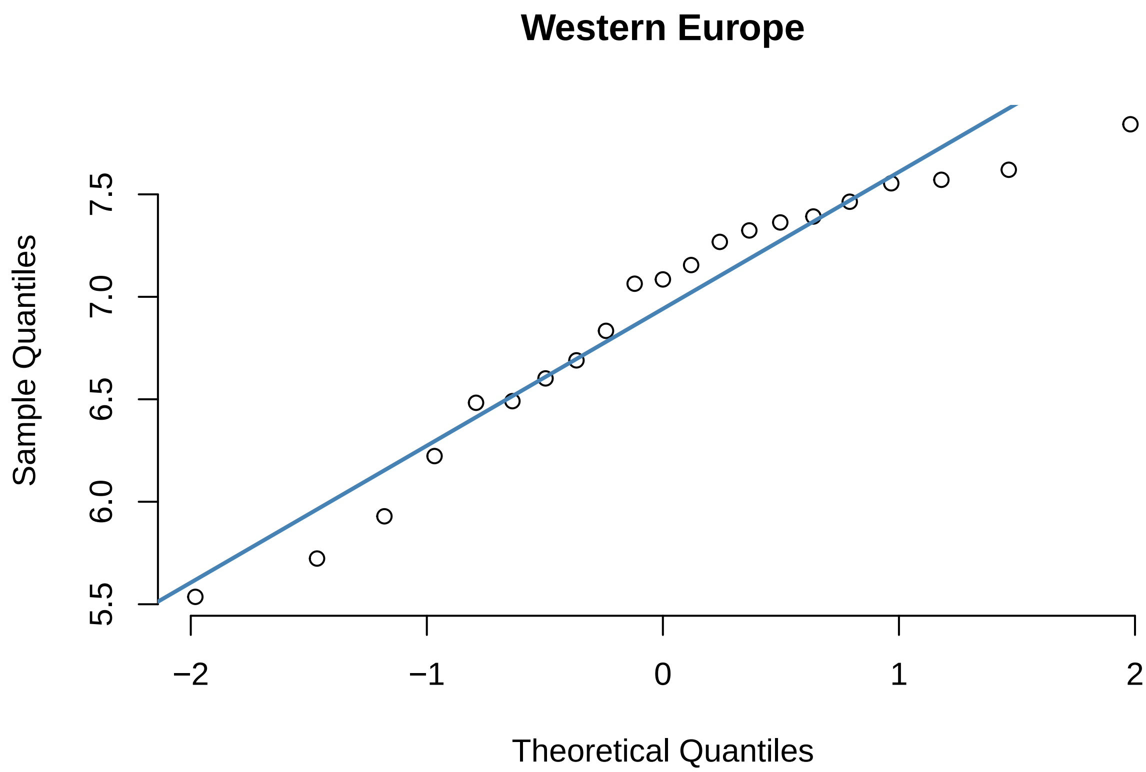


Postoje indikacije da bi ljudi iz zemalja Zapadne Europe trebali biti sretniji od ljudi iz zemalja Srednje i Istočne Europe.

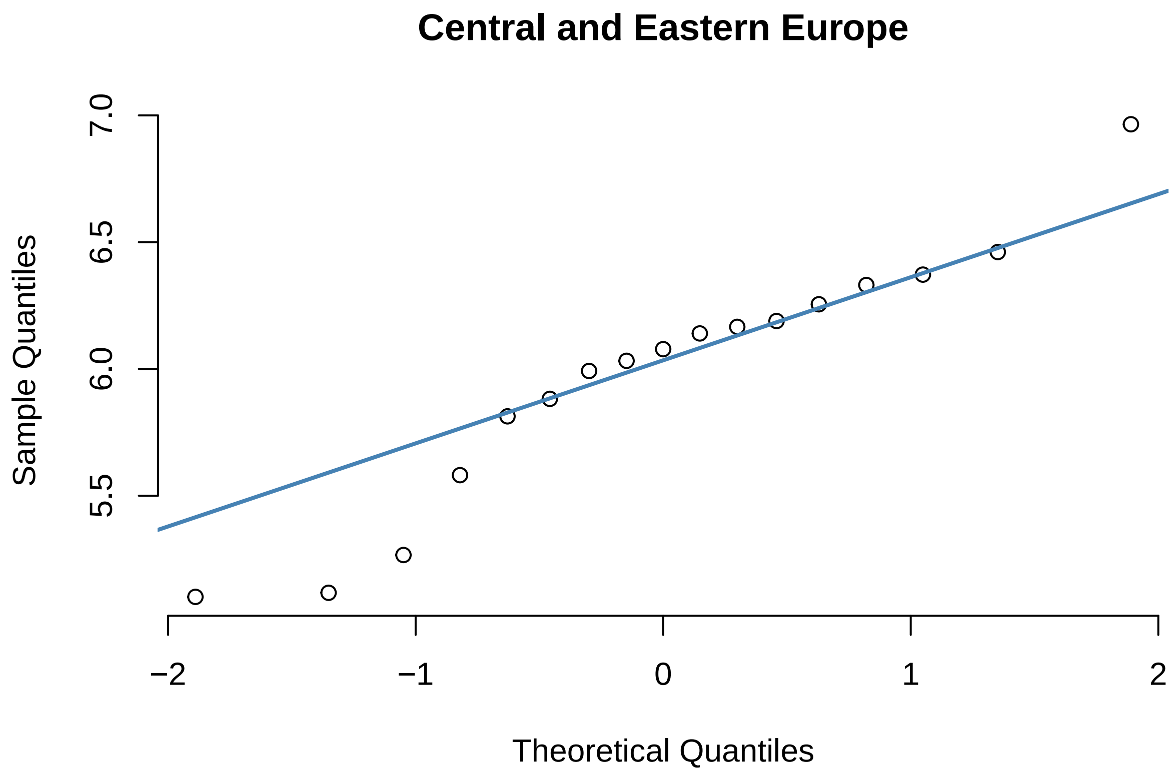
Postavimo sljedeće hipoteze:
H_0: Ladder score je jednak za Zapadnu i Srednju i Istočnu Europu
H_1: Ladder score je veći u Zapadnoj Europi od onog u Srednjoj i Istočnoj Europi

Ovakvo ispitivanje možemo provesti t-testom. Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo dva uzoraka iz dvije različite regije, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka koju ćemo provjeriti qq-plotom i KS testom.

```
qqnorm(western_europe$`Ladder score`, pch = 1, frame = FALSE,main='Western Europe')
qqline(western_europe$`Ladder score`, col = "steelblue", lwd = 2)
```




```
qqnorm(central_eastern_europe$`Ladder score`, pch = 1, frame = FALSE,main='Central and Eastern Europe')
qqline(central_eastern_europe$`Ladder score`, col = "steelblue", lwd = 2)
```



Koristimo Lillieforsovu inačicu testa normalnosti jer srednju vrijednost i varijancu računamo iz uzorka.

```
library(nortest)
lillie.test(western_europe$`Ladder score`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  western_europe$`Ladder score`
## D = 0.16126, p-value = 0.1645
```

```
lillie.test(central_eastern_europe$`Ladder score`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  central_eastern_europe$`Ladder score`
## D = 0.15291, p-value = 0.3622
```

Iz qq-plota ne možemo zaključiti normalnost podataka. Velika p-vrijednost kod Lillieforsovog testa govori kako ne možemo odbaciti hipotezu da podaci dolaze iz normalne distribucije.

Pogledajmo vrijednost varijanci oba uzorka.

```
var(western_europe$`Ladder score`)
```

```
## [1] 0.4310178
```

```
var(central_eastern_europe$`Ladder score`)
```

```
## [1] 0.2433699
```

```r
#Jesu li varijance značajno različite
var.test(western_europe$`Ladder score`, central_eastern_europe$`Ladder score`)
```

```
##
##  F test to compare two variances
##
## data:  western_europe$`Ladder score` and central_eastern_europe$`Ladder score`
## F = 1.771, num df = 20, denom df = 16, p-value = 0.2498
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6606402 4.5100231
## sample estimates:
## ratio of variances
##            1.77104
```

p-vrijednost od 0.2498 nam govori da ne odbacujemo hipotezu da su varijance uzoraka jednake.

Provedimo sada t-test uz pretpostavku jednakosti varijanci.

```r
t.test(western_europe$`Ladder score`, central_eastern_europe$`Ladder score`, alt = "greater", var.equal
```

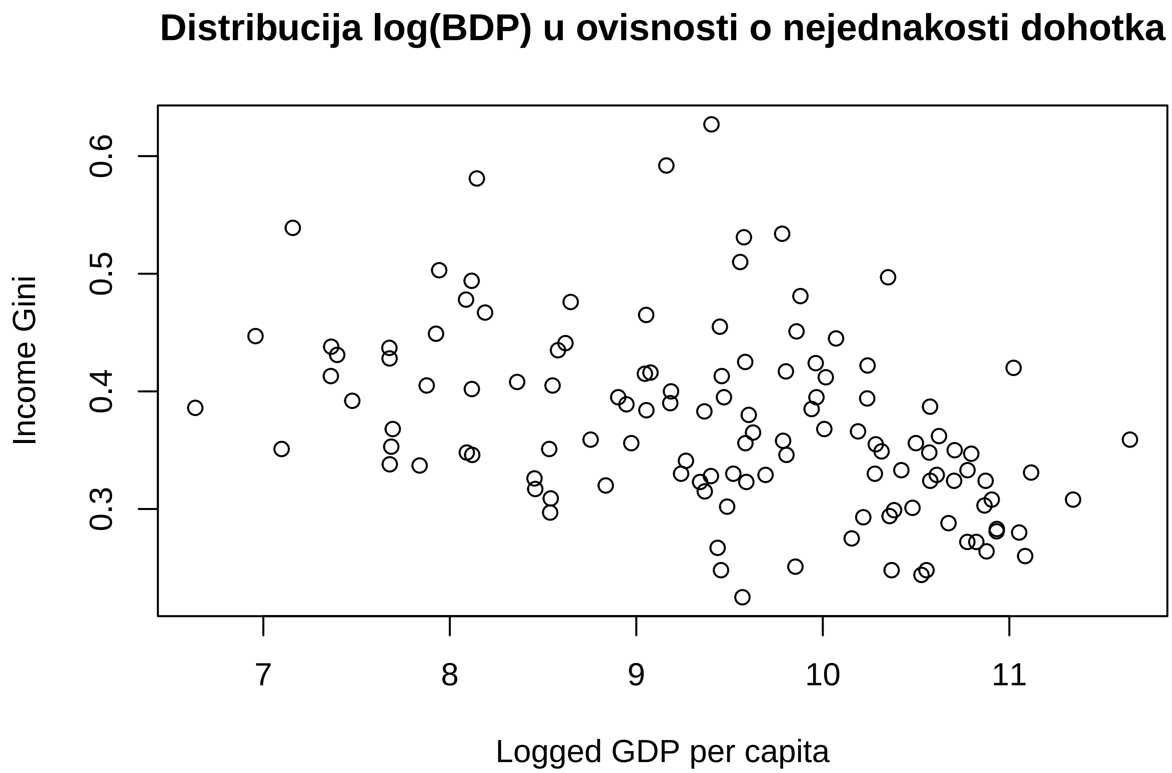
```
##
##  Two Sample t-test
##
## data:  western_europe$`Ladder score` and central_eastern_europe$`Ladder score`
## t = 4.8355, df = 36, p-value = 1.241e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.6053832       Inf
## sample estimates:
## mean of x mean of y
##  6.914905  5.984765
```

Zbog male p-vrijednosti možemo odbaciti hipotezu H_0 u korist alternative da je Ladder score veći u Zapadnoj Europi od onog u Srednjoj i Istočnoj Europi.

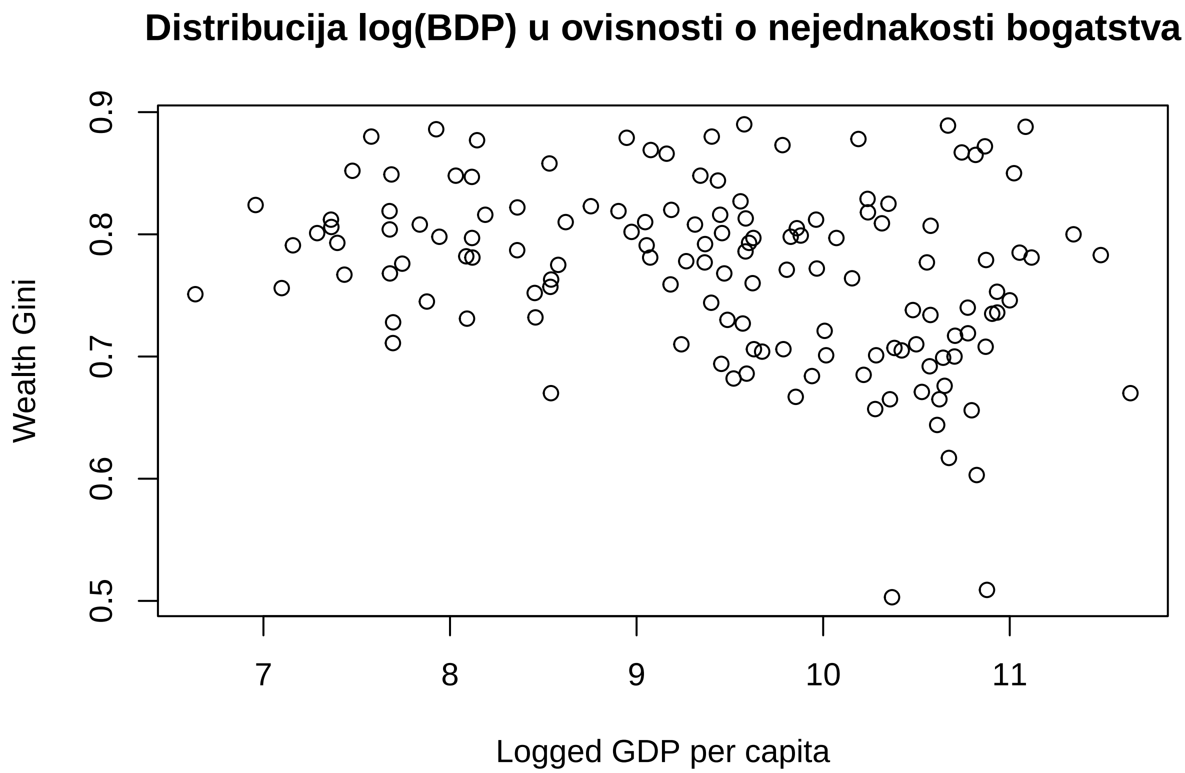
Povezanost između Logged GDP per capita i Gini koeficijenata

Pogledajmo distribuciju prirodnog logaritma bruto domaćeg proizvoda po stanovniku prema paritetu kupovne moći za nejednakost dohotka i nejednakost bogatstva.

```r
plot(whr2021$`Logged GDP per capita`, whr2021$`Income Gini`, xlab = "Logged GDP per capita", ylab = "In
     main = "Distribucija log(BDP) u ovisnosti o nejednakosti dohotka")
```




```
plot(whr2021$`Logged GDP per capita`, whr2021$`Wealth Gini`, xlab = "Logged GDP per capita", ylab = "Wea
     main = "Distribucija log(BDP) u ovisnosti o nejednakosti bogatstva")
```



Iz grafova vidimo da podaci ne slijede lijepi linerarni trend te bi mogli pretpostaviti da ne postoji značajna zavisnost između prirodnog logaritma bruto domaćeg proizvoda po stanovniku s nejednakostima dohotka i bogatstva.

Izračunajmo sada srednje vrijednosti i medijane za nejednokost bogatstva po regijama:

```
library(tidyverse)

whr2021 %>% group_by(`Regional indicator`) %>% summarise(
          Mean.WealthGini = mean(`Wealth Gini`),
          Median.WealthGini = median(`Wealth Gini`)
            ) -> summary.WealthGiniRegion
summary.WealthGiniRegion
```

```
## # A tibble: 10 x 3
##    `Regional indicator`               Mean.WealthGini Median.WealthGini
##    <chr>                                        <dbl>             <dbl>
##  1 Central and Eastern Europe                  NA                NA
##  2 Commonwealth of Independent States          NA                NA
##  3 East Asia                                    0.704             0.706
##  4 Latin America and Caribbean                 NA                NA
##  5 Middle East and North Africa                NA                NA
##  6 North America and ANZ                        0.731             0.709
##  7 South Asia                                   0.769             0.768
##  8 Southeast Asia                               0.796             0.787
##  9 Sub-Saharan Africa                          NA                NA
## 10 Western Europe                              NA                NA
```

!Primjećujemo da nedostaju podaci za neke države te zbog toga nisu prikazani rezultati za sve regije.!

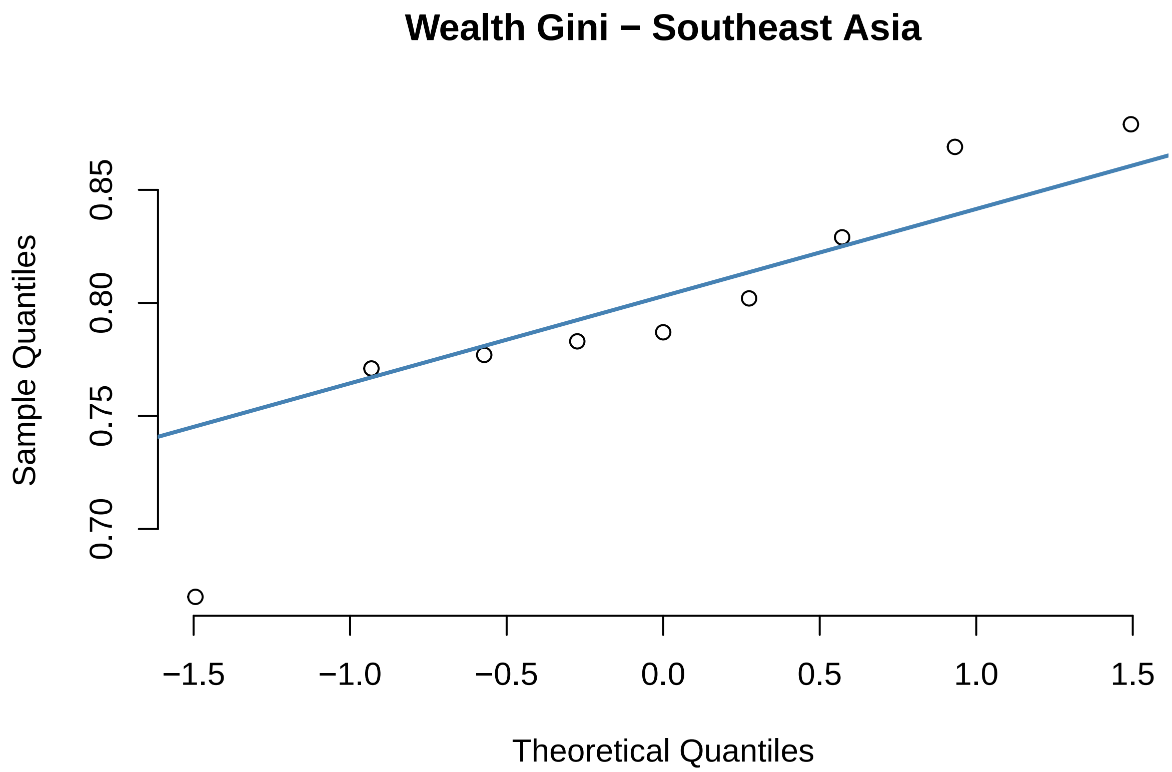
Najveća razlika srednje vrijednosti i medijana vidljiva je između Istočne Azije i Jugoistočne Azije. Postoje indikacije da je nejednakost bogatstva veća u Jugoistočnoj Aziji u odnosu na Istočnu Aziju.

##Nejednakost bogatstva Istočna Azija vs Jugoistočna Azija Postavimo sljedeće hipoteze: H_0: Nejednakost bogatstva je jednaka u Istočnoj i Jugoistočnoj Aziji H_1: Nejednakost bogatstva je veća u Jugoistočnoj Aziji u odnosu na Istočnu Aziju

Ovakvo ispitivanje možemo provesti t-testom. Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo uzorke država različitih regija, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka koju ćemo provjeriti qq-plotom i Lillieforsovim testom.

```
library(nortest)
southeast_asia = whr2021[whr2021$`Regional indicator` == "Southeast Asia",]
east_asia = whr2021[whr2021$`Regional indicator` == "East Asia",]

qqnorm(east_asia$`Wealth Gini`, pch = 1, frame = FALSE,main='Wealth Gini - East Asia')
qqline(east_asia$`Wealth Gini`, col = "steelblue", lwd = 2)
```


```
qqnorm(southeast_asia$`Wealth Gini`, pch = 1, frame = FALSE,main='Wealth Gini - Southeast Asia')
qqline(southeast_asia$`Wealth Gini`, col = "steelblue", lwd = 2)
```




```r
lillie.test(east_asia$`Wealth Gini`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  east_asia$`Wealth Gini`
## D = 0.18032, p-value = 0.783
```

```r
lillie.test(southeast_asia$`Wealth Gini`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  southeast_asia$`Wealth Gini`
## D = 0.22957, p-value = 0.1867
```

Iz qq-plota ne možemo pretpostaviti normalnost podataka. Velika p-vrijednost kod Lillieforsovog testa govori kako ne možemo odbaciti hipotezu da podaci dolaze iz normalne distribucije.

Pogledajmo vrijednost varijanci oba uzorka.

```r
var(east_asia$`Wealth Gini`)
```

```
## [1] 0.001552667
```

```r
var(southeast_asia$`Wealth Gini`)
```

```
## [1] 0.00380675
```

```r
#Jesu li varijance značajno različite
var.test(east_asia$`Wealth Gini`, southeast_asia$`Wealth Gini`)
```

```
##
##  F test to compare two variances
##
## data:  east_asia$`Wealth Gini` and southeast_asia$`Wealth Gini`
## F = 0.40787, num df = 5, denom df = 8, p-value = 0.3381
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.0846686 2.7560611
## sample estimates:
## ratio of variances
##           0.407872
```

p-vrijednost od 0.3381 nam govori da ne odbacujemo hipotezu da su varijance uzoraka jednake.

Provedimo sada t-test uz pretpostavku jednakosti varijanci.

```r
t.test(southeast_asia$`Wealth Gini`, east_asia$`Wealth Gini`, alt = "greater", var.equal = TRUE)
```

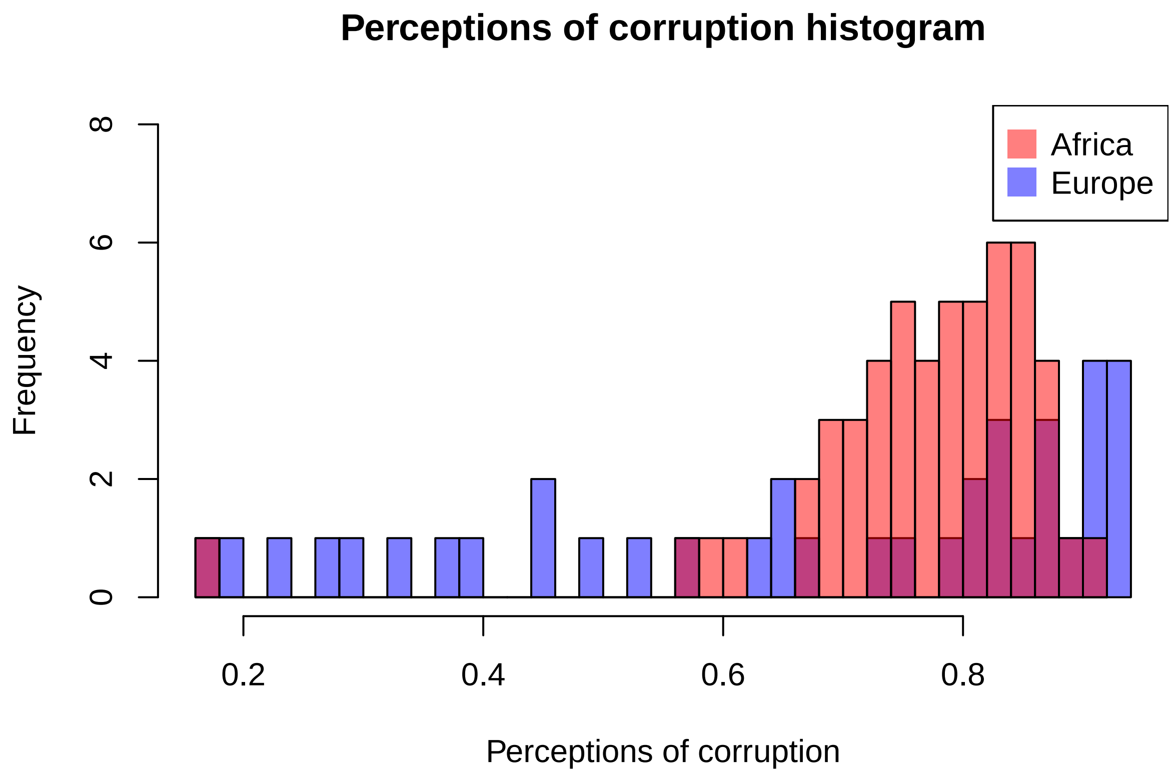
```
##
##  Two Sample t-test
##
## data:  southeast_asia$`Wealth Gini` and east_asia$`Wealth Gini`
## t = 3.2428, df = 13, p-value = 0.003208
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0420598       Inf
## sample estimates:
## mean of x mean of y
## 0.7963333 0.7036667
```

Zbog male p-vrijednosti možemo odbaciti hipotezu H_0 u korist alternative da je nejednakost bogatstva u Jugoistočnoj Aziji u prosjeku veća od nejednakosti bogatstva u Istočnoj Aziji.

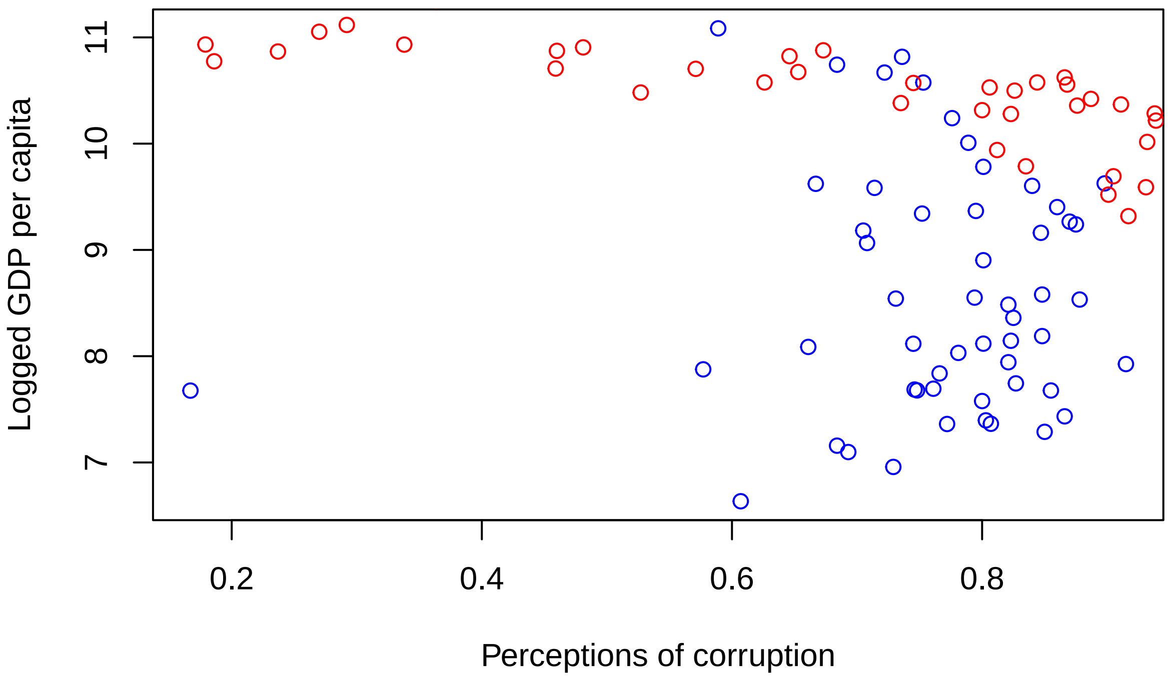
Zastupljenost korupcije u zemljama Europe i Afrike

Pokušajmo sada zaključiti nešto o korupciji. Promatrat ćemo zemlje Europe i Afrike te želimo saznati gdje je korupcija zastupljenija. Ispitat ćemo zavisnost percepcije korupcije o logaritmu BDP-a po stanovniku.

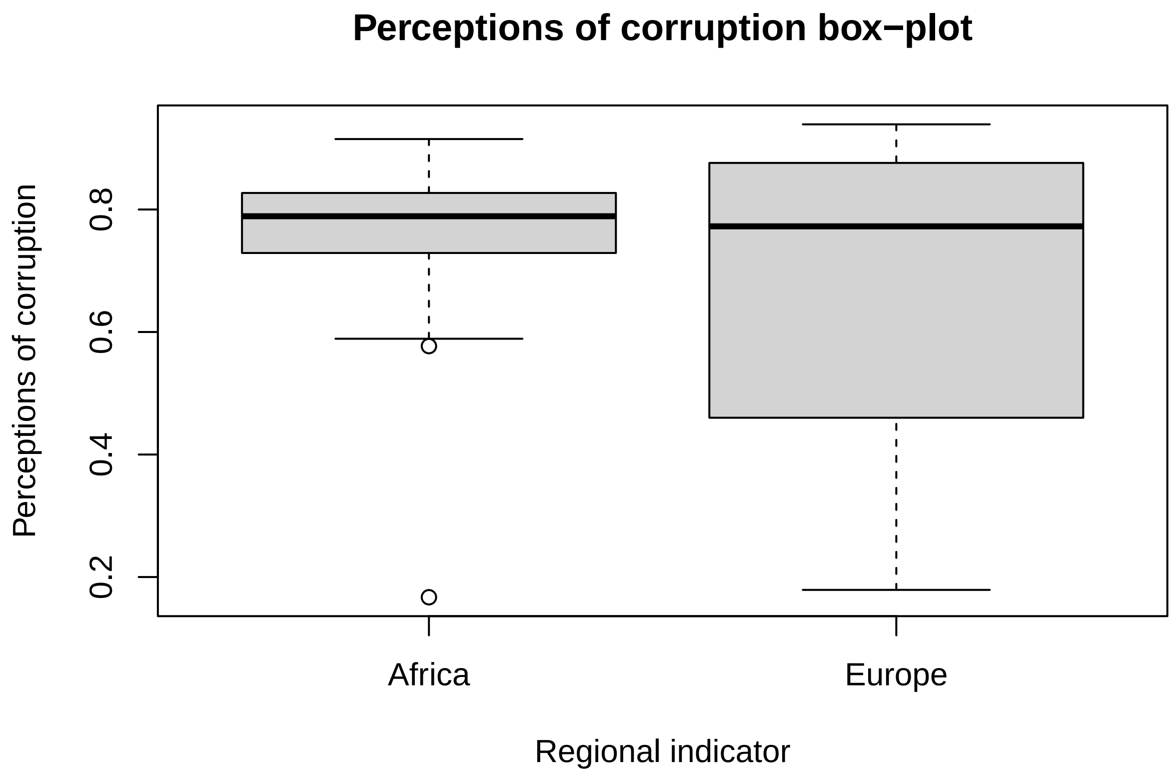
```r
ce_europe = whr2021[whr2021$`Regional indicator` == "Central and Eastern Europe",]
w_europe = whr2021[whr2021$`Regional indicator` == "Western Europe",]
europe <- rbind(ce_europe, w_europe)
men_africa = whr2021[whr2021$`Regional indicator` == "Middle East and North Africa",]
ss_africa = whr2021[whr2021$`Regional indicator` == "Sub-Saharan Africa",]
africa <- rbind(men_africa, ss_africa)



plot_by_gini <- function(column, main) {
  hist(europe[[column]], breaks=30, main=main, xlab=column, ylab="Frequency", ylim = c(0,8),col=rgb(0,0
  hist(africa[[column]], breaks=30, main=main, xlab=column, ylab="Frequency", col=rgb(1,0,0,0.5), add=T
  legend(x="topright", c("Africa", "Europe"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch =
}
plot_by_gini("Perceptions of corruption", "Perceptions of corruption histogram")
```




```
plot(africa$`Perceptions of corruption`,
     africa$`Logged GDP per capita`,
     col='blue',
     ylab='Logged GDP per capita',
     xlab='Perceptions of corruption')
points(europe$`Perceptions of corruption`,
     europe$`Logged GDP per capita`,col='red')
```




```
boxplot(africa$`Perceptions of corruption`,europe$`Perceptions of corruption`,
        main='Perceptions of corruption box-plot',
        ylab='Perceptions of corruption', xlab="Regional indicator", names = c("Africa", "Europe"))
```



Iz histograma vidimo da je percepcija korupcije u Africi bitno veća nego u Europi. Iz drugog grafa vidimo da je logaritam BDP-a po stanovniku relativno visok za sve države Europe te neovisno o njemu ljudi različito percipiraju korupciju. Za države Afrike prevladava visok stupanj percepcije korupcije neovisno o BDP-u. Iz box-plota vidimo veliki rang podataka za Europu, no medijan je otprilike jednak za oba kontinenta. Izračunajmo sada srednju vrijednost percepcije korupcije za Europu i Afriku.

```
mean_europe = mean(europe$`Perceptions of corruption`)
mean_africa = mean(africa$`Perceptions of corruption`)
print(mean_europe)
```

```
## [1] 0.6695789
```

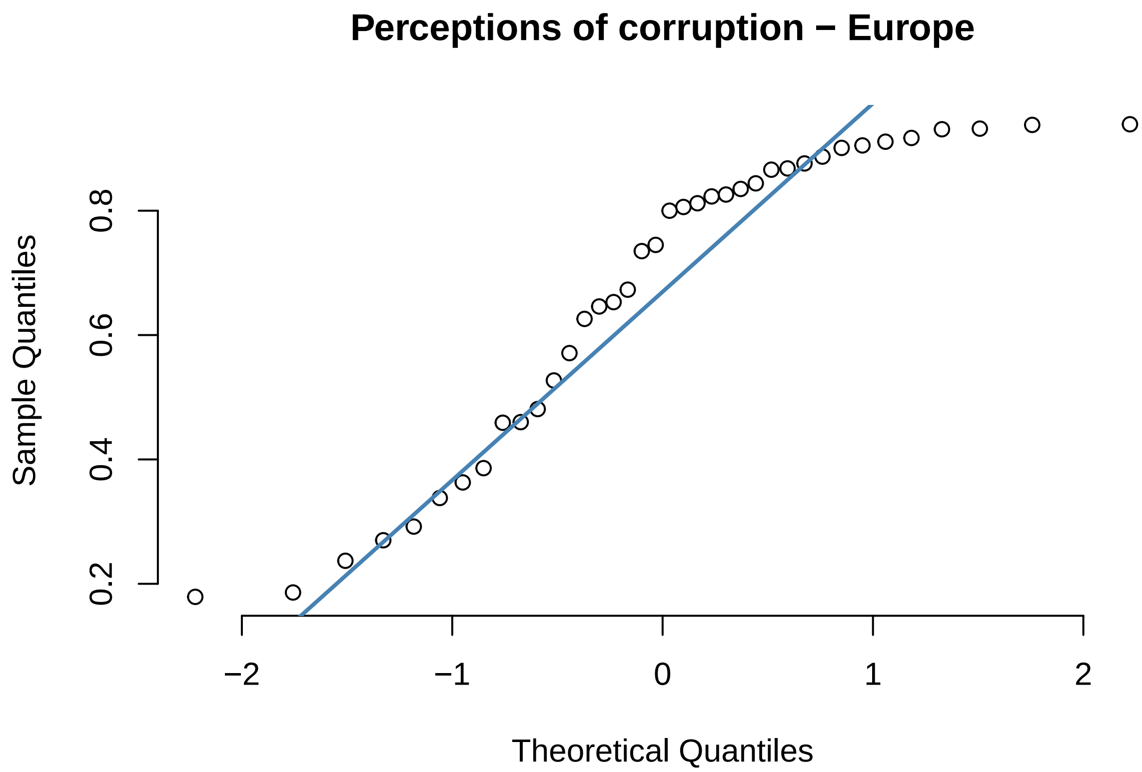
```
print(mean_africa)
```

```
## [1] 0.7647547
```

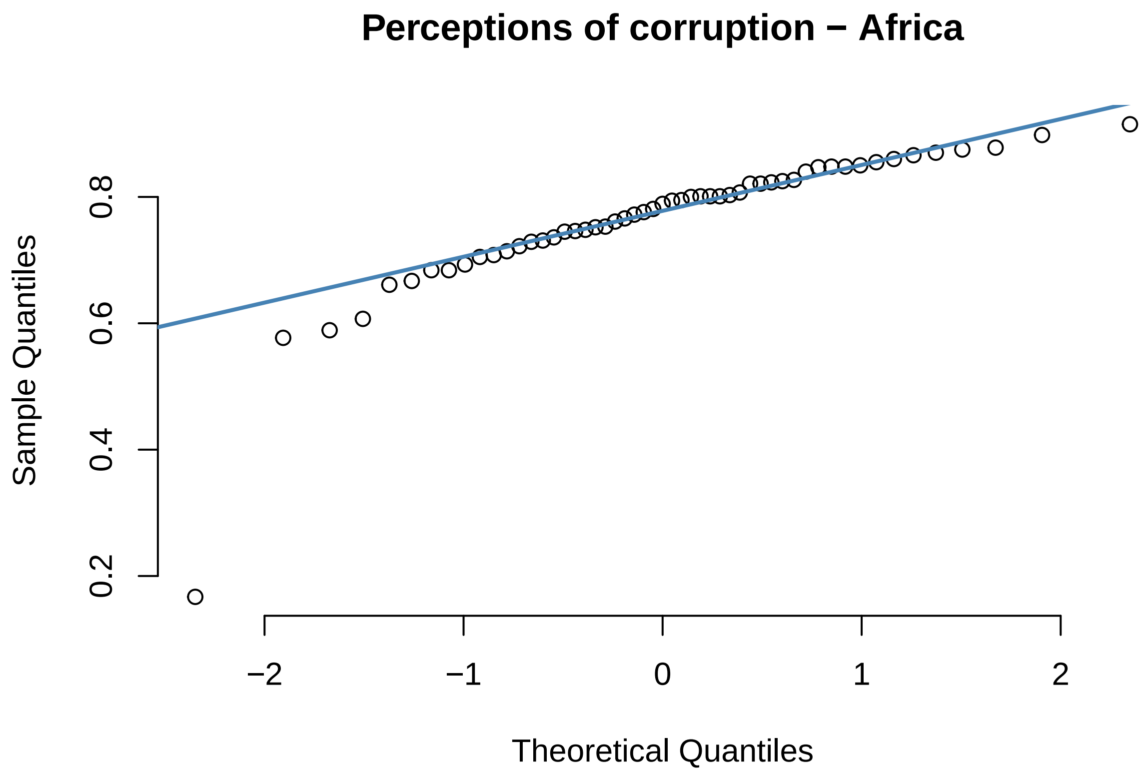
Možemo li na temelju analiza zaključiti da je percepcija korupcije manja u Europi?

Postavimo hipoteze: H_0: srednja vrijednost percepcije korupcije za Europu i Afriku je jednaka H_1: srednja vrijednost percepcije korupcije za Europu je manja od srednje vrijednosti za Afriku Ovakvo ispitivanje možemo provesti t-testom. Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo uzorke država različitih kontinenta, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka koju ćemo provjeriti qq-plotom.

```
qqnorm(europe$`Perceptions of corruption`, pch = 1, frame = FALSE,main='Perceptions of corruption - Eur
qqline(europe$`Perceptions of corruption`, col = "steelblue", lwd = 2)
```




```
qqnorm(africa$`Perceptions of corruption`, pch = 1, frame = FALSE,main='Perceptions of corruption - Afri
qqline(africa$`Perceptions of corruption`, col = "steelblue", lwd = 2)
```



Iz dobivenih grafova možemo naslutiti normalnost podataka za Afiku uz male izuzetke na repovima dok normalnost podataka za Europu nije vidljiva pa ne možemo provesti t-test. Već iz prethodnog histograma se dalo naslutiti da podaci za Europu ne slijede normalnu distibuciju.

Testirajmo li podatke Lillieforsovim testom dolazimo do istog zaključka.

```
lillie.test(africa$`Perceptions of corruption`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  africa$`Perceptions of corruption`
## D = 0.13097, p-value = 0.02386
```

```
lillie.test(europe$`Perceptions of corruption`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  europe$`Perceptions of corruption`
## D = 0.20137, p-value = 0.0004698
```

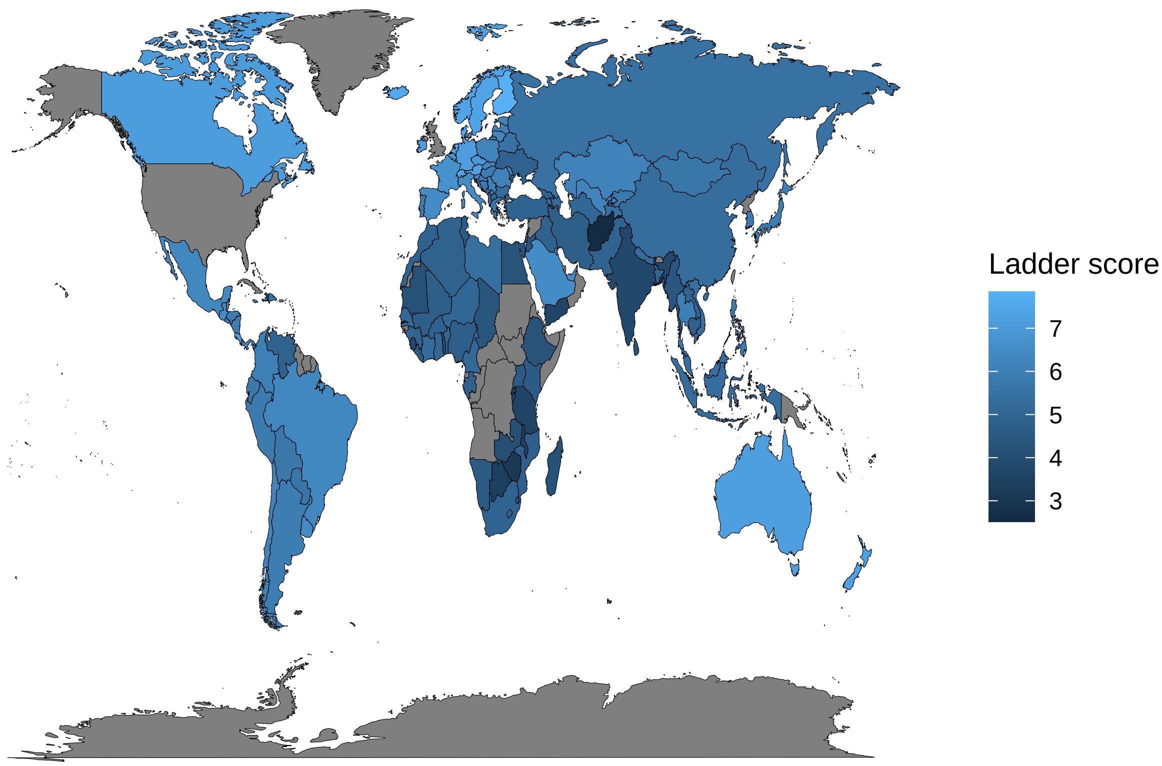
Zbog male p-vijednosti možemo odbaciti hiptezu H_0 da podaci dolaze iz normalne distibucije. Ne možemo provesti t-test. Jedan od mogućih rješenja je transformirati podatke i provesti jackknife.

Usporedba razina sreće u 2020. i 2021. godini.

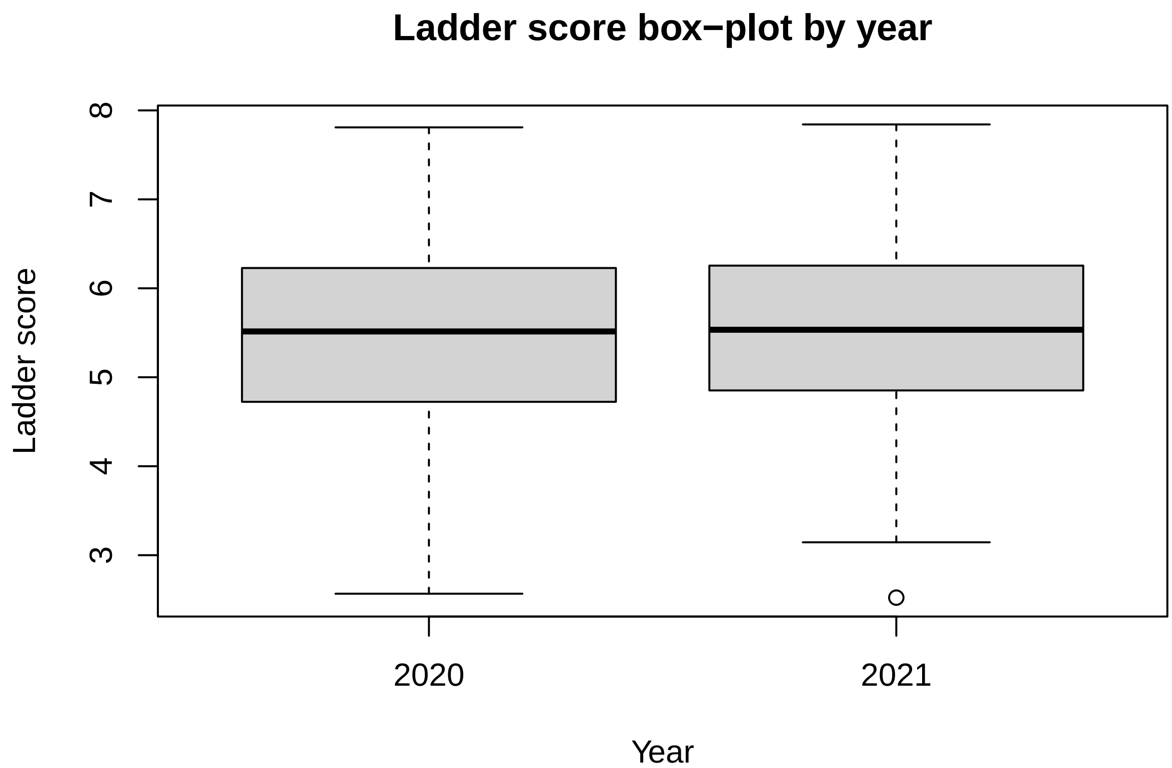
```
library(ggplot2)
require(maps)
```

```
## Loading required package: maps

##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

```
data2021 = whr2021[c("Country name", "Ladder score")]
names(data2021)[names(data2021) == "Country name"] = "region"

mapdata2021 = map_data("world")
mapdata2021 = left_join(mapdata2021, data2021, by = "region")

map2021 = ggplot(mapdata2021, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = `Ladder score`), color = "black", size = 0.1) + theme(axis.text.x = element_bl
                          axis.text.y = element_blank(),
                          axis.ticks = element_blank(),
                          axis.title.y = element_blank(),
                          axis.title.x = element_blank(),
                          rect = element_blank())

map2021
```

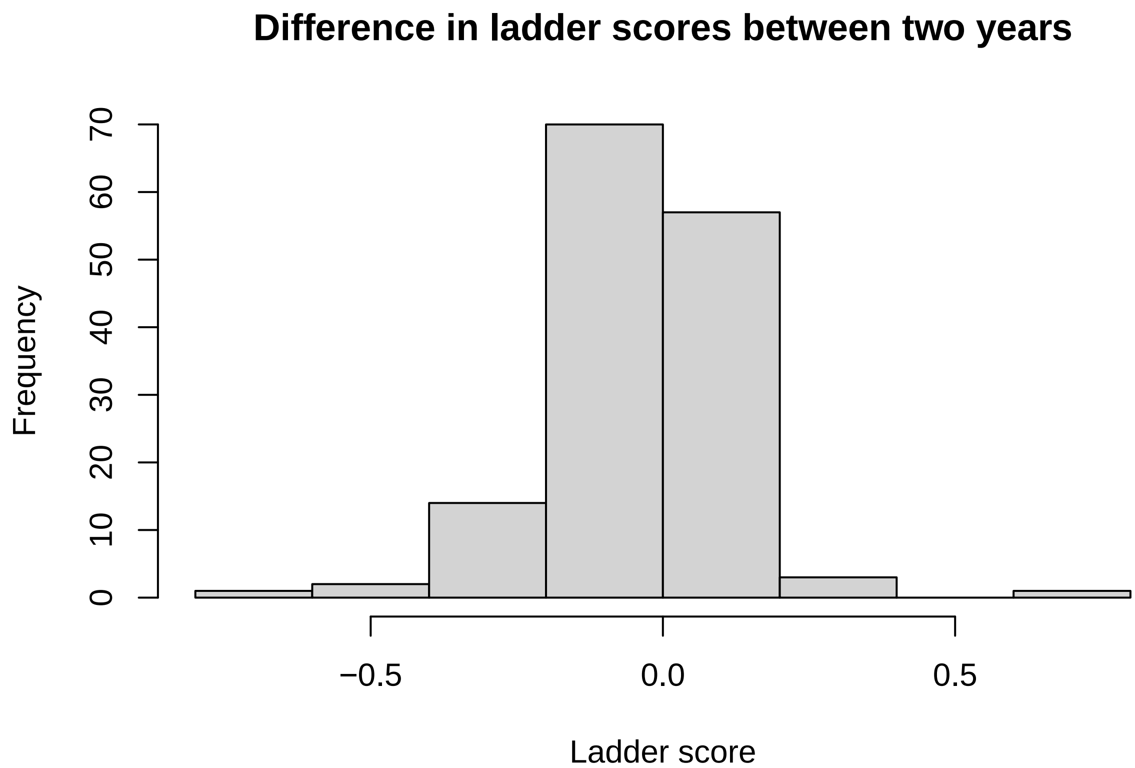


Pravokutni dijagram Ladder score-ova za 2020. i 2021. godinu.

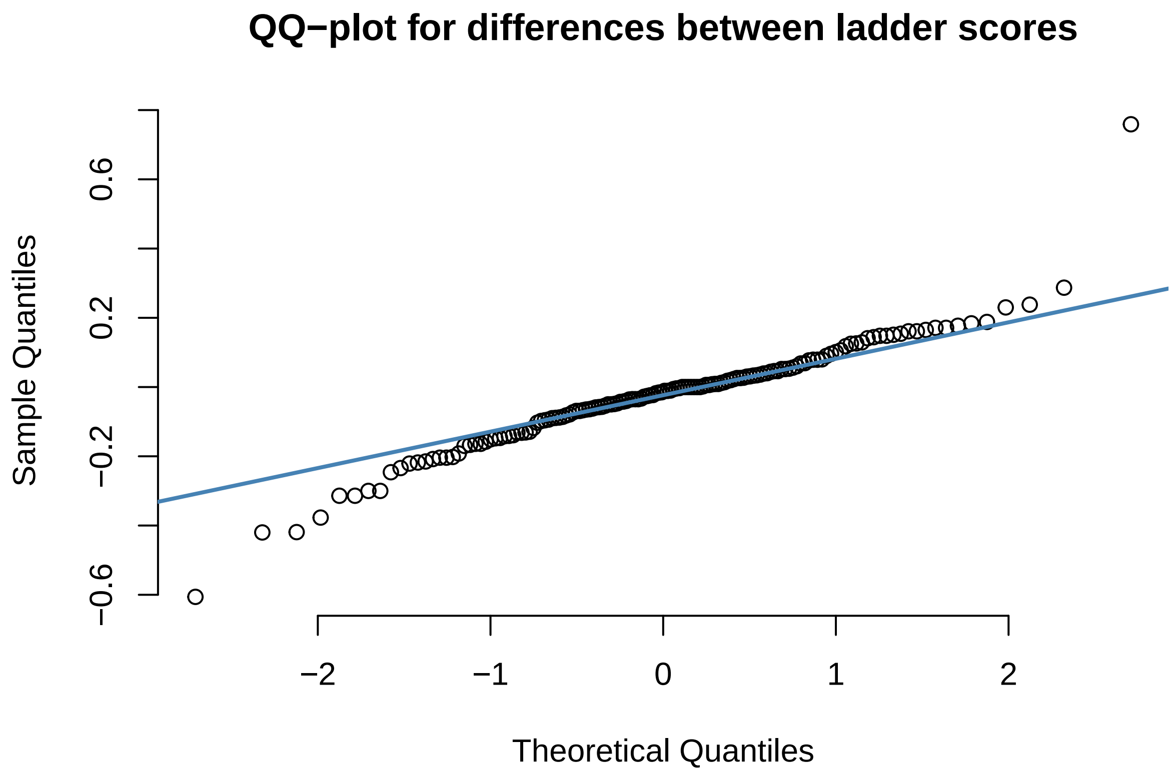

Ladder score box−plot by year
Ladder score
8
7
6
5
4
3
2020
2021
Year

Spojimo podatke iz dvije godine te na histogramu prikažimo razlike razina sreće za dvije godine.

```
mergedData = merge(whr2020, whr2021, by="Country name", suffixes = c(".20",".21"))

hist(mergedData$`Ladder score.20`-mergedData$`Ladder score.21`,
     main=paste('Difference in ladder scores between two years'),
     xlab='Ladder score')
```




```r
qqnorm(mergedData$`Ladder score.20`-mergedData$`Ladder score.21`,
       pch = 1,
       frame = FALSE,
       main=paste('QQ-plot for differences between ladder scores'))
qqline(mergedData$`Ladder score.20`-mergedData$`Ladder score.21`,
       col = "steelblue", lwd = 2)
```



Histogram razlika nam sugerira normalnost podataka, dok iz qq-plota vidimo malo odstupanje lijevog repa. Testiramo normalnost podataka o razlici razina sreće za dvije države. Koristimo Lillieforsovu inačicu KS testa.

```
lillie.test(mergedData$`Ladder score.20`-mergedData$`Ladder score.21`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mergedData$`Ladder score.20` - mergedData$`Ladder score.21`
## D = 0.084528, p-value = 0.01157
```

Unatoč maloj p-vrijednosti Lillieforsovog testa, nastavljamo s testom o uparmin podacima, jer na razini značajnosti od 1% ipak ne možemo odbaciti hipotezu da podaci dolaze iz normalne razdiobe. Pod pretpostavkom da su podatci normalni, koristimo upareni t-test. Postavljamo hipoteze:
H_0: $\mu_{2020} = \mu_{2021}$ H_1: $\mu_{2020} < \mu_{2021}$

```
t.test(mergedData$`Ladder score.20`,
       mergedData$`Ladder score.21`,
       paired = TRUE,
       alt = "less")
```

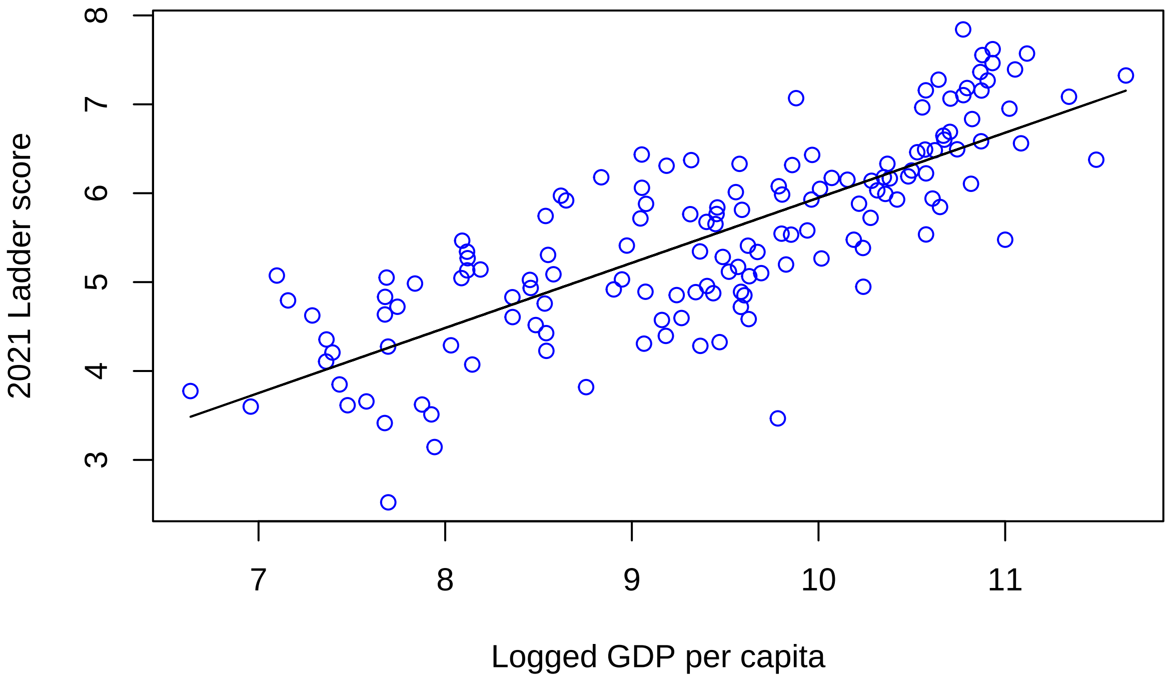
```
##
##  Paired t-test
##
## data:  mergedData$`Ladder score.20` and mergedData$`Ladder score.21`
## t = -2.0749, df = 147, p-value = 0.01987
## alternative hypothesis: true difference in means is less than 0
```

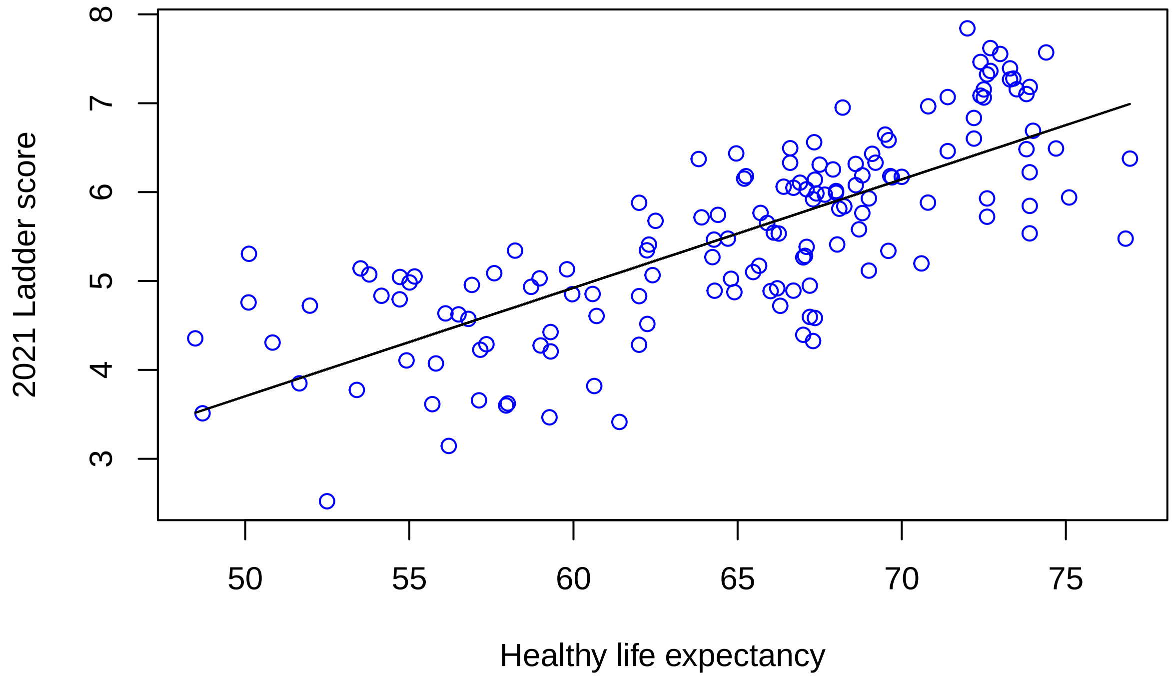
```
## 95 percent confidence interval:
##          -Inf -0.005247129
## sample estimates:
## mean of the differences
##             -0.02594595
```

Jako mala p-vrijednost nam ukazuje da postoji statistički značajna razlika u "ladder score-u" u dvije godine. Postoje značajne razlike u sreći država u dvije godine tj. u periodu

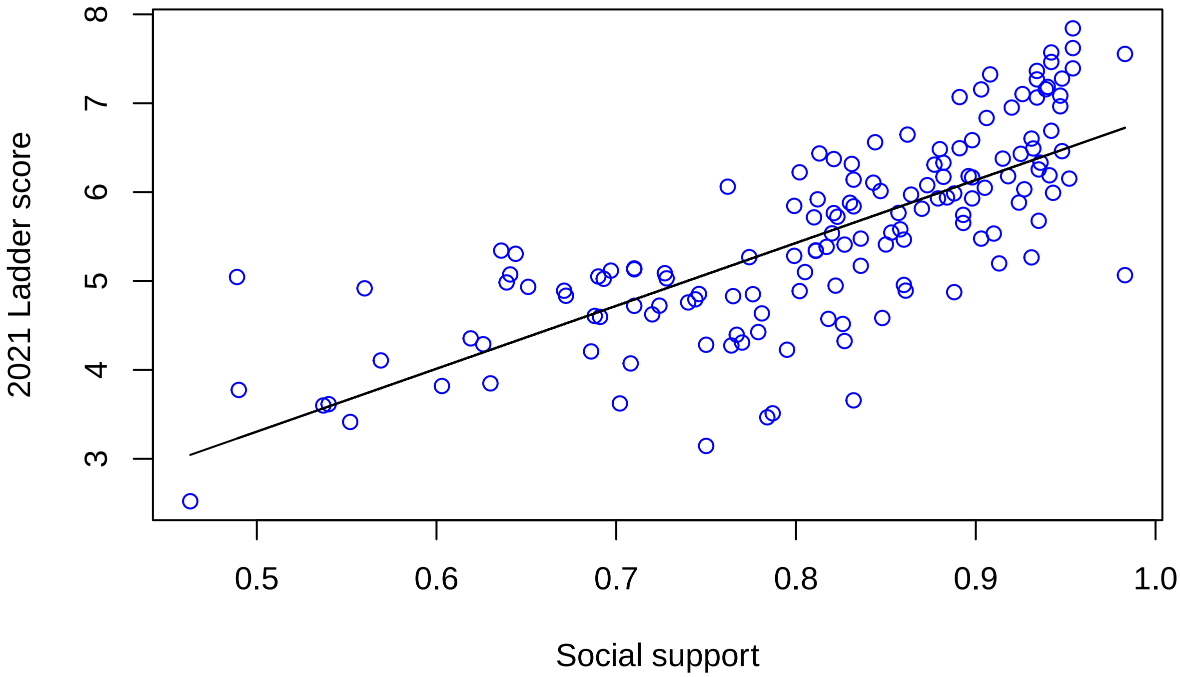
Ovisnost razine sreće o drugim varijablama u 2021. godini

Možemo li iz dijagrama raspršenja naslutiti vezu između varijabli iz dataset-a i Ladder score-a?



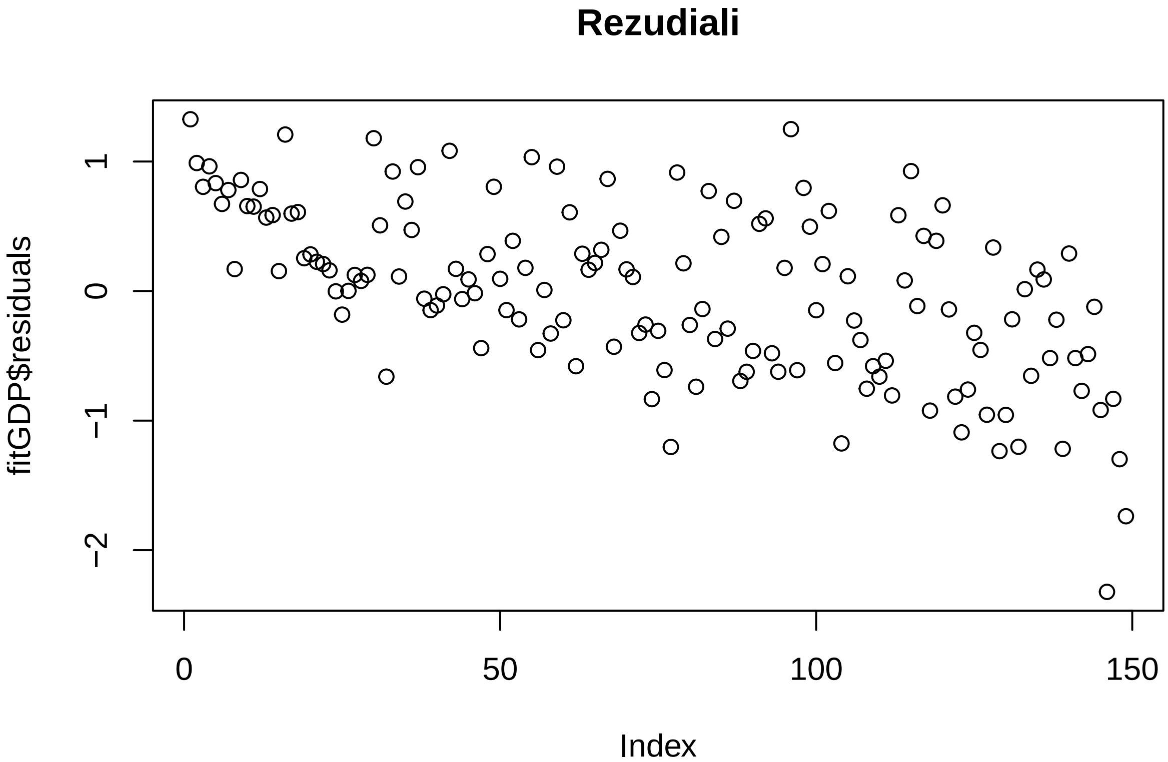


2021 Ladder score
3
4
5
6
7
8
50
55
60
65
70
75
Healthy life expectancy

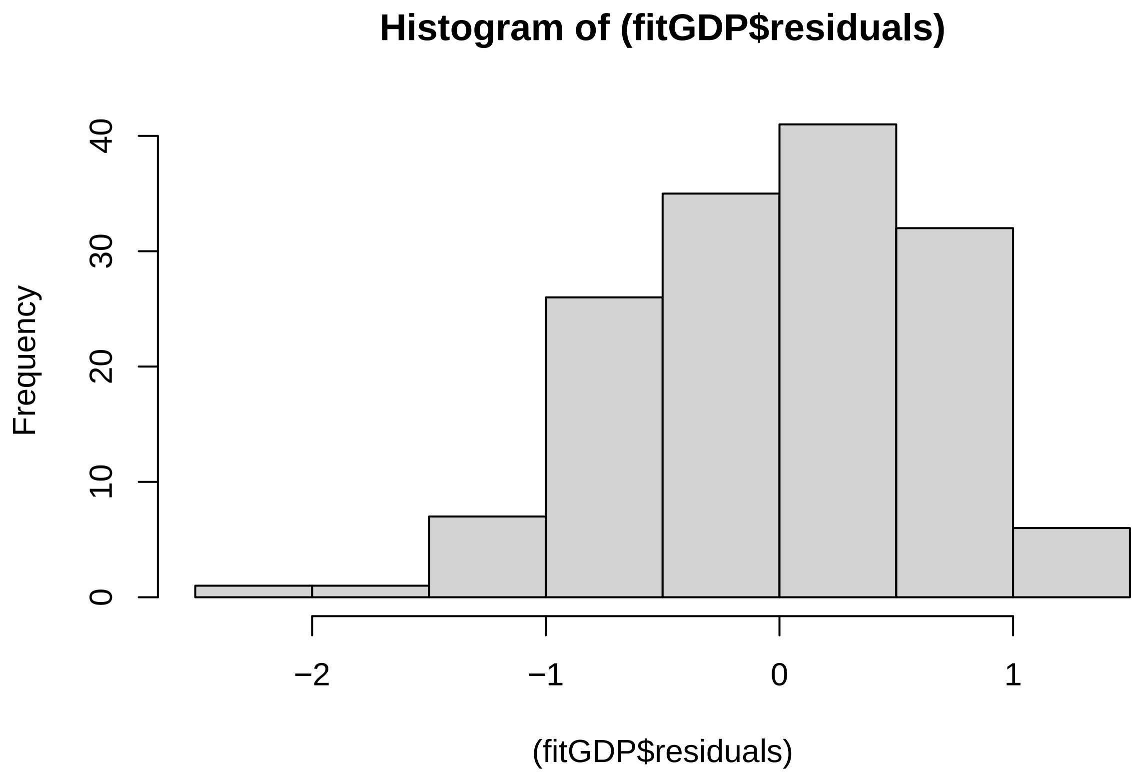


Iz dobivenih grafova bi mogli naslutiti da postoji veza između ulaznih varijabli i izlazne. Da bi nastavili daljnju analizu potrebno je provjeriti pretpostavke modela o regresorima i rezidualima. One ne smiju biti jako narušene. Mora vrijediti normalnost reziduala i homogenost varijance te regresori ne smiju biti jako korelirani kada imamo više regresora. Provjerimo prvo normalnost reziduala i homogenost varijance.

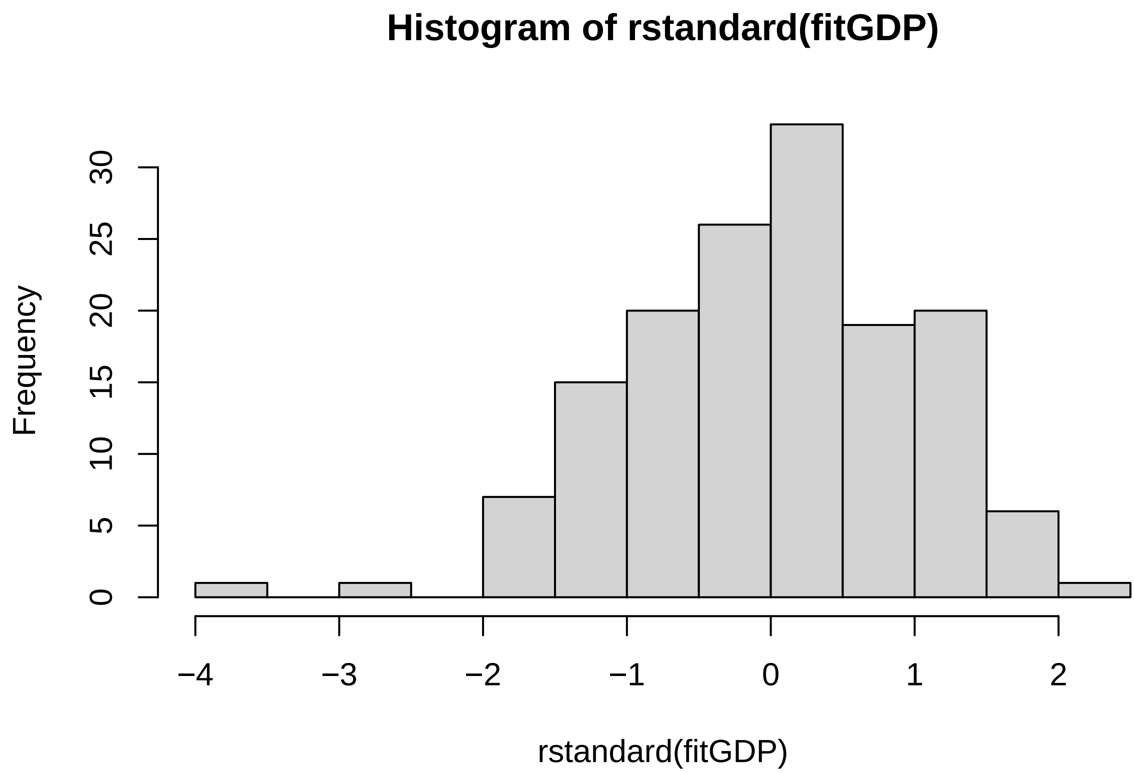
```
plot(fitGDP$residuals, main = "Rezudiali")
```



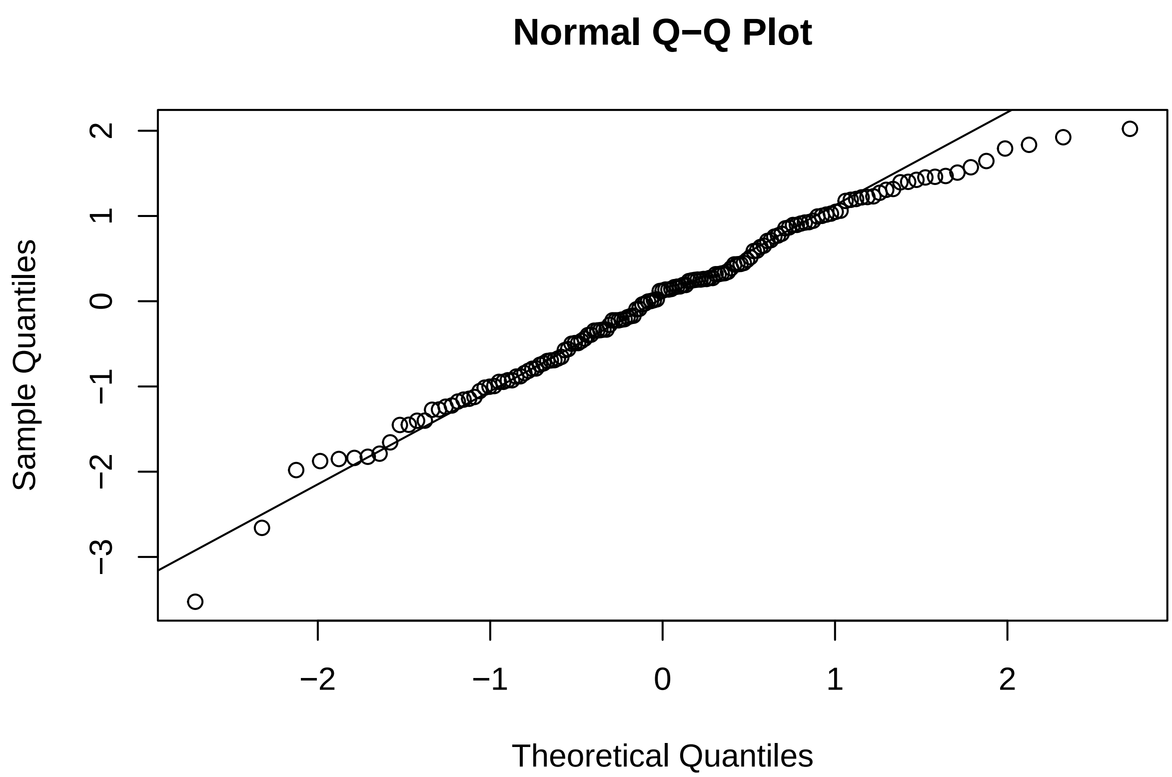

```
hist((fitGDP$residuals))
```




```
hist(rstandard(fitGDP))
```




```
qqnorm(rstandard(fitGDP))
qqline(rstandard(fitGDP))
```




```r
require(nortest)
lillie.test(rstandard(fitGDP))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fitGDP)
## D = 0.057186, p-value = 0.2733
```

Iz samog prikaza reziduala, teško je doći do nekog zaključka o normalnosti. Histogrami nam prikazuju izgled distibucije reziduala te vidimo da distibucija donekle nalikuje normalnoj. Također, iz qq grafa možemo naslutiti normalnost reziduala. Velika p-vrijednost kod Lillieforsovog testa govori kako ne možemo odbaciti hipotezu da podaci dolaze iz normalne distribucije

Izračunajmo sada mjere za model jednostavne linearne regresije za ulaznu varijabu "Logged GDP per capita" i izlaznu varijabu "Ladder score".

```r
summary(fitGDP)
```

```
##
## Call:
## lm(formula = whr2021$`Ladder score` ~ whr2021$`Logged GDP per capita`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32190 -0.46198  0.08206  0.50740  1.32618
##
## Coefficients:
```

```
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.3719     0.4456  -3.079  0.00248 **
## whr2021$`Logged GDP per capita`   0.7320     0.0469  15.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.661 on 147 degrees of freedom
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6212
## F-statistic: 243.7 on 1 and 147 DF,  p-value: < 2.2e-16
```

R-kvadrat (koeficijent determinacije) za dobiveni model iznosi 0.6237 što nam govori koliki postotak varijance u izlaznoj varijabli("Ladder score") je estimirani linearni model opisao. F-statistika nam služi za ispitivanje signifikantnosti modela.

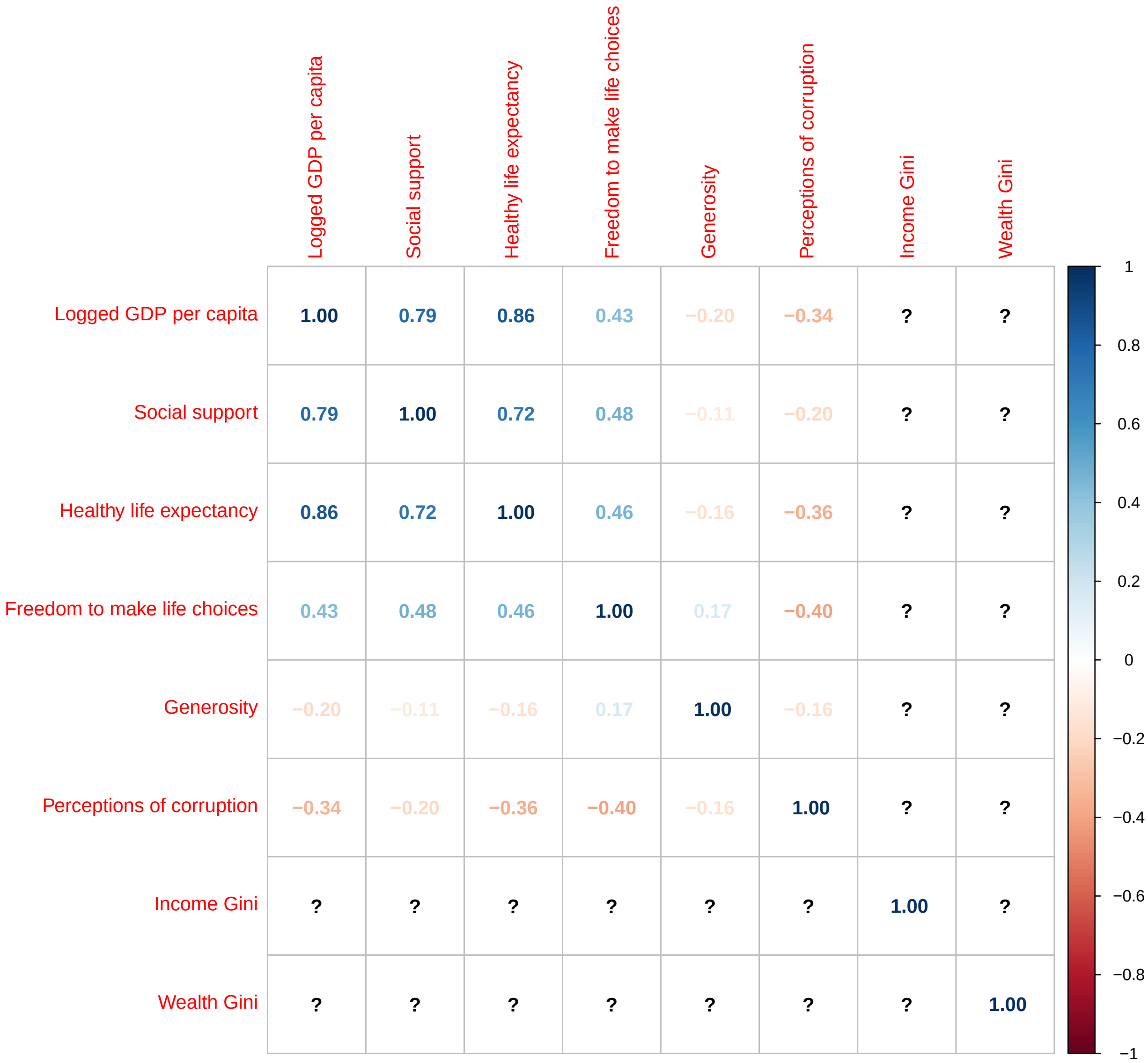
Pogledajmo koliko iznose R-kvadrat i F-statistika za preostale jednostavne modele linearne regresije.

	R-kvadrat	**F-statistika**	**Lillieforsov test normalnosti (p-vrijednost)**
Logged GDP per capita	0.6237	243.7	0.2733
Social support	0.5729	197.2	0.1273
Healthy life expectancy	0.59	211.5	0.2764
Freedom to make life choices	0.3694	86.1	0.4493
Generosity	0.0003168	0.04659	0.74
Perceptions of corruption	0.1774	31.69	0.000541
Income Gini	0.1595	22.96	0.3773
Wealth Gini	0.1003	15.28	0.2915

Prema vrijednostima R-kvadrat i F-statistike kao tri najznačajnija regresora su redom "Logged GDP per capita", "Healthy life expectancy" i "Social support". Varijabla "Generosity" se pokazala kao najmanje značajna te ju vjeroatno ni nećemo koristiti u višestrukoj linearnoj regresiji.

Prije nego što krenemo s višestukom linearnom regresijom moramo provjeriti korelaciju među ulaznim varijablama.

```
library("corrplot")
korelacija <- whr2021[,(names(whr2021) %in% c("Logged GDP per capita", "Social support", "Healthy life e
num <- cor(korelacija)
corrplot(num, method="number")
```

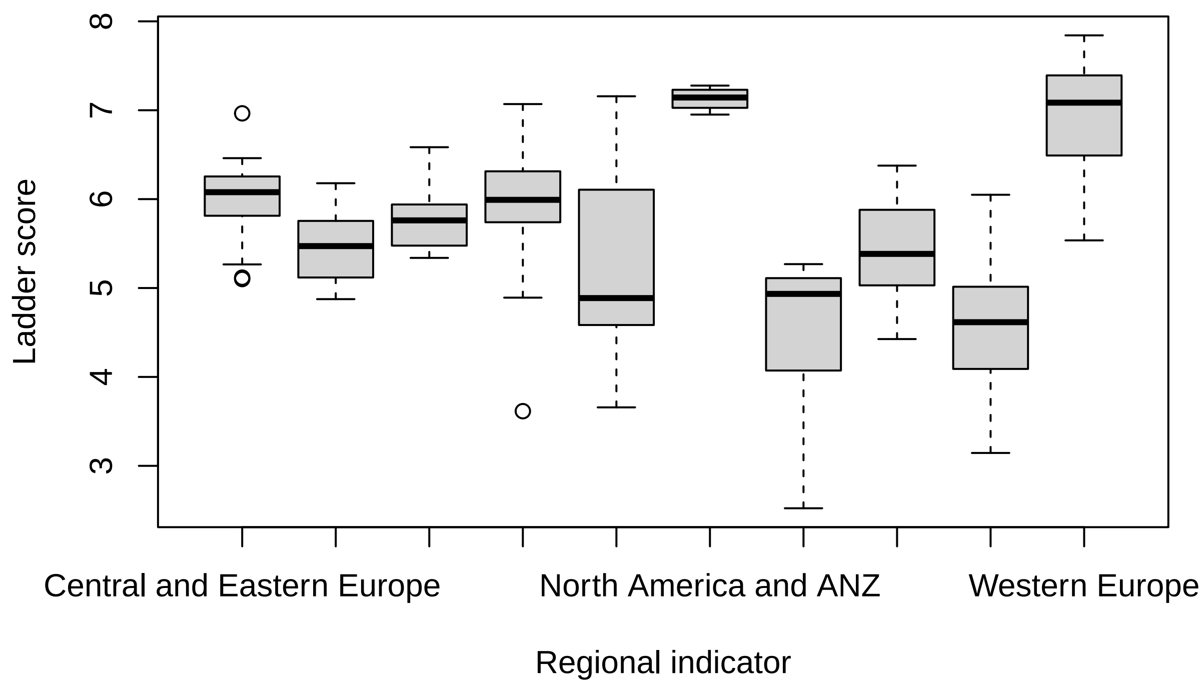


Varijabla "Logged GDP per capita" koja je u jednostavnoj linearnoj regresiji bila najznačajnija je jako korelirana s druge dvije najznačajnije ("Social support" i "Healthy life expectancy"). Zbog toga pri izgradnji modela višestruke linearne regresije ne smijemo koristiti sve 3 navedene varijable. Pokušajmo izgraditi model tako da R-kvadrat i F-statistika budu najveći. Primjetimo također da za Gini nedostaju neki podatci pa nam ova funkcija ne izračunava korelaciju.

Nadalje, probat ćemo iskoristiti i kategorijsku varijablu "Regional indicator", no prije moramo provjeriti: - radi li se o varijabli na nominalnoj ili ordinalnoj skali, - ima li varijabla linearan efekt na izlaznu varijablu, - predstavlja li određena kategorijska varijabla nešto što je određenom metričkom varijablom već predstavljeno.

U slučaju varijable "Regional indicator", ona je na nominalnoj skali,te nije predstavljena nekom metričkom varijablom. Za provjeru linearanog efekta iskoristit ćemo box-plot.

```r
boxplot(`Ladder score`~`Regional indicator`,data=whr2021)
```



Iz priloženog box-plota ne vidimo neki linearan trend, te ćemo zasad zanemariti ovu varijablu.

```r
fitAll= lm(`Ladder score` ~ `Logged GDP per capita` + `Social support` + `Healthy life expectancy` + `F
summary((fitAll))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support` +
##     `Healthy life expectancy` + `Freedom to make life choices` +
##     Generosity + `Perceptions of corruption` + `Income Gini` +
##     `Wealth Gini`, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67398 -0.24034  0.05907  0.32531  1.16407
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.69329    1.15052  -1.472 0.143994
## `Logged GDP per capita`         0.22094    0.10649   2.075 0.040394 *
## `Social support`                2.84833    0.78465   3.630 0.000435 ***
## `Healthy life expectancy`       0.04096    0.01708   2.398 0.018194 *
## `Freedom to make life choices`  1.45431    0.59195   2.457 0.015611 *
## Generosity                      0.35180    0.35211   0.999 0.319974
## `Perceptions of corruption`    -0.87515    0.33706  -2.596 0.010731 *
```

```
## `Income Gini`                     -0.29481    0.83542  -0.353 0.724855
## `Wealth Gini`                     -0.27628    0.88205  -0.313 0.754716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5393 on 108 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.7699, Adjusted R-squared:  0.7529
## F-statistic: 45.18 on 8 and 108 DF,  p-value: < 2.2e-16
```

Na temelju regresije sa svim varijablama, možemo zaključiti da Social support, Freedom to make life choices i Perception of corruption najviše djeluju na osjećaj sreće. Treba pronaći model koji opisuje veći postotak varijance, ali uz što manji broj regresora.

```
fitm1= lm(`Ladder score` ~ `Social support`+`Freedom to make life choices` + `Perceptions of corruption
summary((fitm1))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Social support` + `Freedom to make life choices` +
##     `Perceptions of corruption`, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87491 -0.34264  0.09363  0.42610  1.34028
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.0779     0.5594   0.139    0.889
## `Social support`                 5.6256     0.4980  11.297  < 2e-16 ***
## `Freedom to make life choices`   2.2271     0.5397   4.127 6.18e-05 ***
## `Perceptions of corruption`     -1.2254     0.3052  -4.015 9.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6094 on 145 degrees of freedom
## Multiple R-squared:  0.6845, Adjusted R-squared:  0.678
## F-statistic: 104.9 on 3 and 145 DF,  p-value: < 2.2e-16
```

```
fitm2= lm(`Ladder score` ~ `Healthy life expectancy`+`Freedom to make life choices` + `Perceptions of co
summary((fitm2))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Healthy life expectancy` + `Freedom to make life choices` +
##     `Perceptions of corruption`, data = whr2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3506 -0.3385  0.1023  0.4190  1.3682
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -2.535503   0.695521  -3.645 0.000371 ***
## `Healthy life expectancy`       0.095401   0.008663  11.013  < 2e-16 ***
## `Freedom to make life choices`  2.816597   0.525520   5.360 3.21e-07 ***
```

```
## `Perceptions of corruption`    -0.497107   0.316567  -1.570 0.118523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6166 on 145 degrees of freedom
## Multiple R-squared:  0.677,  Adjusted R-squared:  0.6703
## F-statistic: 101.3 on 3 and 145 DF,  p-value: < 2.2e-16
```

```
fitm3= lm(`Ladder score` ~ `Logged GDP per capita`+`Freedom to make life choices` + `Perceptions of cor
summary((fitm3))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Freedom to make life choices` +
##     `Perceptions of corruption`, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32565 -0.37867  0.07027  0.41682  0.94506
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.87303    0.60081  -3.117   0.0022 **
## `Logged GDP per capita`         0.58456    0.04642  12.593  < 2e-16 ***
## `Freedom to make life choices`  2.85474    0.48681   5.864 2.93e-08 ***
## `Perceptions of corruption`    -0.50531    0.29542  -1.710   0.0893 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5775 on 145 degrees of freedom
## Multiple R-squared:  0.7167, Adjusted R-squared:  0.7108
## F-statistic: 122.3 on 3 and 145 DF,  p-value: < 2.2e-16
```

```
fitm4= lm(`Ladder score` ~ `Logged GDP per capita` + `Healthy life expectancy`+`Freedom to make life ch
summary((fitm4))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Healthy life expectancy` +
##     `Freedom to make life choices` + `Perceptions of corruption`,
##     data = whr2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0602 -0.3593  0.1125  0.3693  0.8756
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -2.51551    0.63881  -3.938 0.000128 ***
## `Logged GDP per capita`         0.41678    0.07891   5.282 4.63e-07 ***
## `Healthy life expectancy`       0.03590    0.01379   2.603 0.010210 *
## `Freedom to make life choices`  2.65281    0.48366   5.485 1.81e-07 ***
## `Perceptions of corruption`    -0.43433    0.29099  -1.493 0.137737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5664 on 144 degrees of freedom
## Multiple R-squared:  0.7294, Adjusted R-squared:  0.7219
## F-statistic: 97.04 on 4 and 144 DF,  p-value: < 2.2e-16
```

```
fitm5= lm(`Ladder score` ~ `Logged GDP per capita` + `Social support`+`Freedom to make life choices` + `
summary((fitm5))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support` +
##     `Freedom to make life choices` + `Perceptions of corruption`,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.13669 -0.32296  0.05636  0.39667  1.03170
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.56273    0.57685  -2.709  0.00757 **
## `Logged GDP per capita`         0.38713    0.06605   5.862 3.00e-08 ***
## `Social support`                2.69946    0.67143   4.020 9.33e-05 ***
## `Freedom to make life choices`  2.25500    0.48662   4.634 7.96e-06 ***
## `Perceptions of corruption`    -0.74279    0.28723  -2.586  0.01070 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5495 on 144 degrees of freedom
## Multiple R-squared:  0.7453, Adjusted R-squared:  0.7382
## F-statistic: 105.3 on 4 and 144 DF,  p-value: < 2.2e-16
```

```
fitm6= lm(`Ladder score` ~ `Social support` + `Healthy life expectancy`+`Freedom to make life choices` +
summary((fitm6))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Social support` + `Healthy life expectancy` +
##     `Freedom to make life choices` + `Perceptions of corruption`,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65601 -0.27080  0.00865  0.38516  1.35552
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.96256    0.63719  -3.080  0.00248 **
## `Social support`                3.48214    0.60533   5.752 5.08e-08 ***
## `Healthy life expectancy`       0.05602    0.01041   5.383 2.90e-07 ***
## `Freedom to make life choices`  2.01063    0.49574   4.056 8.15e-05 ***
## `Perceptions of corruption`    -0.78943    0.29092  -2.714  0.00747 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.558 on 144 degrees of freedom
## Multiple R-squared:  0.7373, Adjusted R-squared:   0.73
## F-statistic: 101.1 on 4 and 144 DF,  p-value: < 2.2e-16
```

```
fitm7= lm(`Ladder score` ~ `Logged GDP per capita`+`Social support` + `Healthy life expectancy`+`Freedo
summary((fitm7))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support` +
##     `Healthy life expectancy` + `Freedom to make life choices` +
##     `Perceptions of corruption`, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93303 -0.29768  0.06863  0.33924  1.02304
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -2.11039    0.62112  -3.398 0.000880 ***
## `Logged GDP per capita`         0.26400    0.08584   3.075 0.002518 **
## `Social support`                2.50670    0.66835   3.751 0.000256 ***
## `Healthy life expectancy`       0.02936    0.01332   2.204 0.029095 *
## `Freedom to make life choices`  2.13266    0.48342   4.412 2.01e-05 ***
## `Perceptions of corruption`    -0.66778    0.28549  -2.339 0.020718 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5423 on 143 degrees of freedom
## Multiple R-squared:  0.7536, Adjusted R-squared:  0.745
## F-statistic: 87.49 on 5 and 143 DF,  p-value: < 2.2e-16
```

Iz priloženog vidimo, uključivši svih 5 značajnih varijabli dobivamo najveći R-squared. No, približno jednak rezultat dobivamo ako ne uključimo "Healthy life expectancy", što je posljedica koreliranosti između varijabli "Logged GDP per capita", "Social support" i "Healthy life expectancy". Također uočimo da u slučaju kada koristimo "Logged GDP per capita" i "Social support" u odnosu na "Logged GDP per capita" i "Healthy life expectancy" dobivamo bolji R-squared, dok je R-squared kod jednostavne regresije pojedinačnih varijabli veći u slučaju "Healthy life expectancy" nego "Social support". Razlog opet leži u većoj koreliranosti.

```
fit1= lm(`Ladder score` ~ `Logged GDP per capita` + `Social support`+`Freedom to make life choices` + `
summary((fit1))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support` +
##     `Freedom to make life choices` + `Perceptions of corruption`,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.13669 -0.32296  0.05636  0.39667  1.03170
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.56273    0.57685  -2.709  0.00757 **
## `Logged GDP per capita`         0.38713    0.06605   5.862 3.00e-08 ***
```

```
## `Social support`                2.69946    0.67143   4.020 9.33e-05 ***
## `Freedom to make life choices`  2.25500    0.48662   4.634 7.96e-06 ***
## `Perceptions of corruption`    -0.74279    0.28723  -2.586  0.01070 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5495 on 144 degrees of freedom
## Multiple R-squared:  0.7453, Adjusted R-squared:  0.7382
## F-statistic: 105.3 on 4 and 144 DF,  p-value: < 2.2e-16
```

```
fit2= lm(`Ladder score` ~ `Logged GDP per capita` + `Social support`+`Freedom to make life choices` , da
summary((fit2))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support` +
##     `Freedom to make life choices`, data = whr2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2334 -0.3487  0.0519  0.4296  1.0608
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.6143     0.4171  -6.268 3.96e-09 ***
## `Logged GDP per capita`          0.4361     0.0645   6.761 3.12e-10 ***
## `Social support`                 2.3424     0.6698   3.497 0.000625 ***
## `Freedom to make life choices`   2.6849     0.4662   5.759 4.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5602 on 145 degrees of freedom
## Multiple R-squared:  0.7334, Adjusted R-squared:  0.7279
## F-statistic:   133 on 3 and 145 DF,  p-value: < 2.2e-16
```

```
fit3= lm(`Ladder score` ~ `Logged GDP per capita` + `Social support`+ `Perceptions of corruption`, data
summary((fit3))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support` +
##     `Perceptions of corruption`, data = whr2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9998 -0.3330  0.0655  0.4087  1.1832
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.19581    0.52957  -0.370    0.712
## `Logged GDP per capita`      0.38414    0.07055   5.445 2.16e-07 ***
## `Social support`             3.65325    0.68273   5.351 3.34e-07 ***
## `Perceptions of corruption` -1.19748    0.28838  -4.152 5.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.587 on 145 degrees of freedom
## Multiple R-squared:  0.7073, Adjusted R-squared:  0.7012
## F-statistic: 116.8 on 3 and 145 DF,  p-value: < 2.2e-16
```

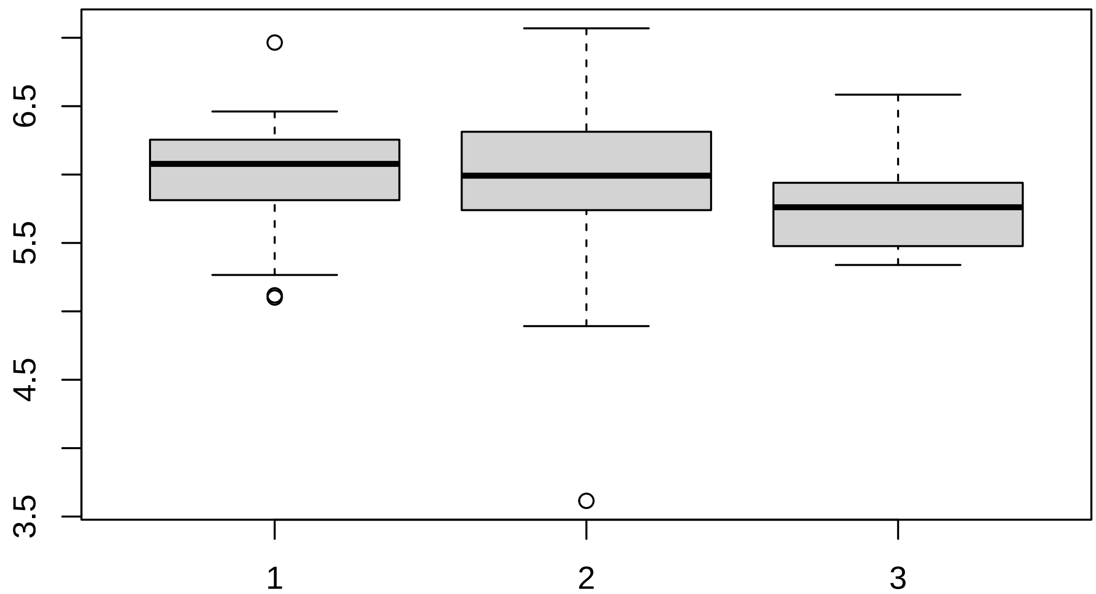
```
fit4= lm(`Ladder score` ~ `Logged GDP per capita` + `Social support`, data = whr2021)
summary((fit4))
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support`,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12862 -0.40577  0.02927  0.46460  1.23356
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.63939    0.42112  -3.893  0.00015 ***
## `Logged GDP per capita`  0.47246    0.07091   6.663 5.12e-10 ***
## `Social support`         3.33340    0.71511   4.661 7.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6188 on 146 degrees of freedom
## Multiple R-squared:  0.6725, Adjusted R-squared:  0.668
## F-statistic: 149.9 on 2 and 146 DF,  p-value: < 2.2e-16
```

Iz ovoga proizlazi da originalan model ne možemo reducirati jer gubimo u R-squared. Konačan model se sastoji od "Logged GDP per capita", "Social support", "Freedom to make life choices" i "Perceptions of corruption".

ANOVA

```
ce_europe = whr2021[whr2021$`Regional indicator` == "Central and Eastern Europe",]
l_america = whr2021[whr2021$`Regional indicator` == "Latin America and Caribbean",]
e_asia = whr2021[whr2021$`Regional indicator` == "East Asia",]
regions = whr2021[whr2021$`Regional indicator` == "Central and Eastern Europe" | whr2021$`Regional indi



boxplot(ce_europe$`Ladder score`, l_america$`Ladder score`, e_asia$`Ladder score`)
```



Želimo testirati jednakost sredina razina sreće u regijama srednje i istočne Europe, Latinske Amerike i Kariba i istočne Azije. S obzirom da je pretpostavka ANOVA-e normalnost podataka, normalnost ćemo testirati Lillieforceovom inačicom KS testa.

```
lillie.test(regions$`Ladder score`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  regions$`Ladder score`
## D = 0.12133, p-value = 0.1154
```

```
lillie.test(ce_europe$`Ladder score`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ce_europe$`Ladder score`
## D = 0.15291, p-value = 0.3622
```

```
lillie.test((l_america$`Ladder score`))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (l_america$`Ladder score`)
## D = 0.20652, p-value = 0.02522
```

```r
lillie.test(e_asia$`Ladder score`)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  e_asia$`Ladder score`
## D = 0.21742, p-value = 0.5012
```

Na temelju rezultata Lillieforceovih testova, možemo zaključiti da na razini značajnosti od 1% sve populacije dolaze iz normalne razdiobe. Nadalje, provest ćemo test homogenosti varijaci populacija:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$
$$H_1 : \neg H_0.$$

```r
bartlett.test(regions$`Ladder score` ~ regions$`Regional indicator`)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  regions$`Ladder score` by regions$`Regional indicator`
## Bartlett's K-squared = 2.6094, df = 2, p-value = 0.2713
```

```r
aggregate(regions$`Ladder score`, by=list(regions$`Regional indicator`), FUN=var)
```

```
##                       Group.1         x
## 1  Central and Eastern Europe 0.2433699
## 2                  East Asia 0.1935239
## 3 Latin America and Caribbean 0.4808964
```

Na temelju rezultata Bartlettovog testa, zaključujemo da su varijance populacija homogene.

Provodimo test o jednakosti sredina populacija.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \neg H_0.$$

```r
a = aov(regions$`Ladder score` ~ regions$`Regional indicator`)
summary(a)
```

```
##                             Df Sum Sq Mean Sq F value Pr(>F)
## regions$`Regional indicator`  2  0.145  0.0727   0.208  0.813
## Residuals                   40 13.999  0.3500
```

Na kraju provedenog testa, možemo zaključiti da su sredine razina sreće u prethodno navedene tri regije jednake. Na temelju rezultata testa ne možemo odbaciti hipotezu H_0