

# 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 sadržani 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 projekt zadrži varijable kao što su BDP po stanovniku, životni vijek, socijalna podrška, percepcija korupcije, doniranje novca u dobrotvorne svrhe, nejednakost dohotka i slično.

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

```
## [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

```
## [1] "Country name" "Regional indicator"
```

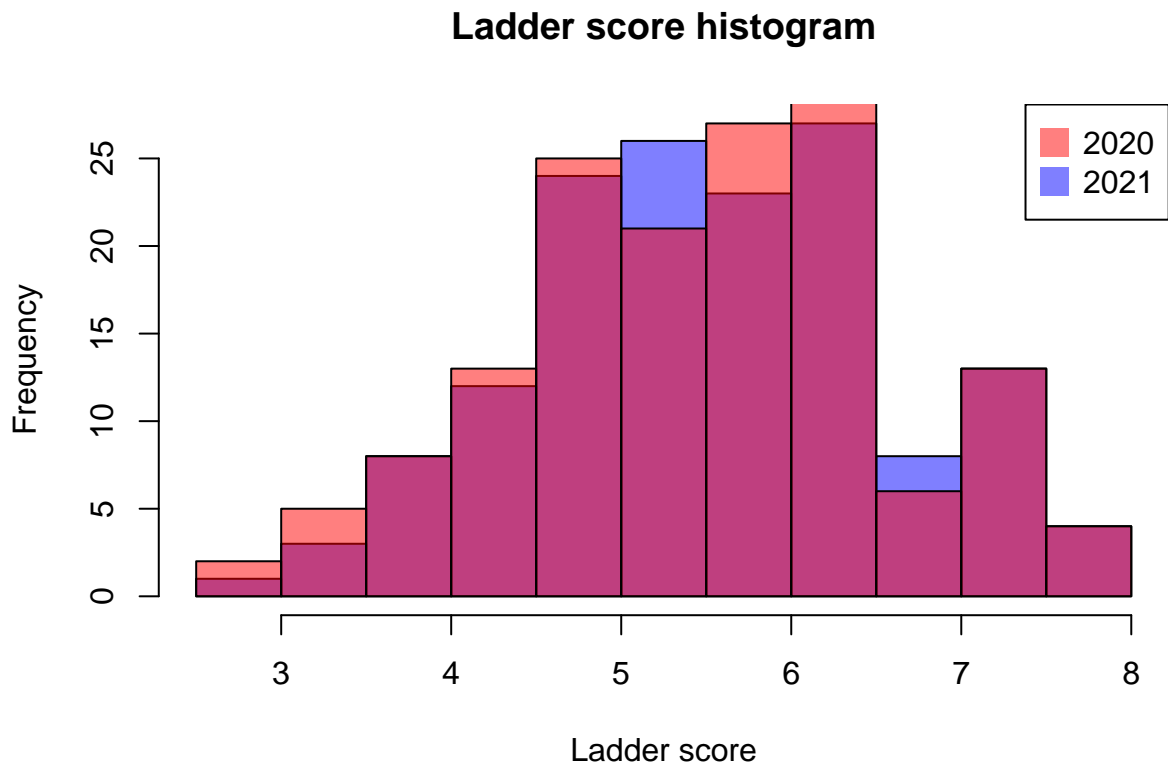
```
## [3] "Ladder score"          "Logged GDP per capita"
## [5] "Social support"        "Healthy life expectancy"
## [7] "Freedom to make life choices" "Generosity"
## [9] "Perceptions of corruption"

## [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

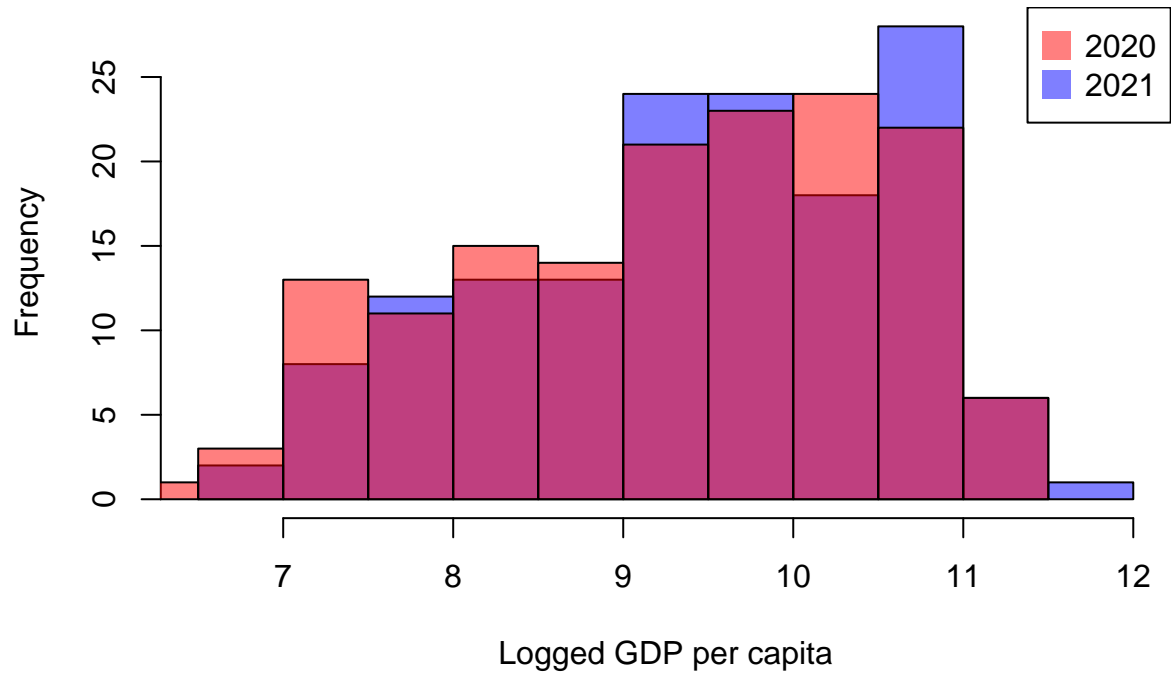
Prikažimo sada histograme usporedbe varijabli za različite godine.

```
#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=TRUE)
  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")
```



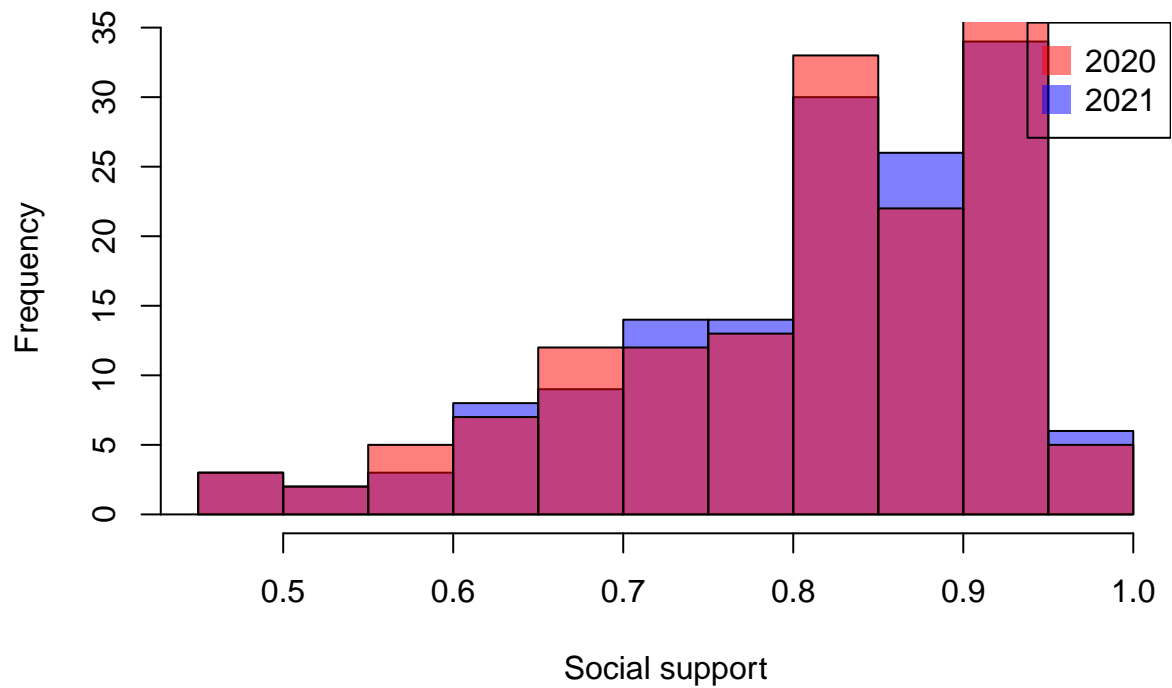
```
plot_by_years("Logged GDP per capita", "Logged GDP per capita histogram")
```

**Logged GDP per capita histogram**



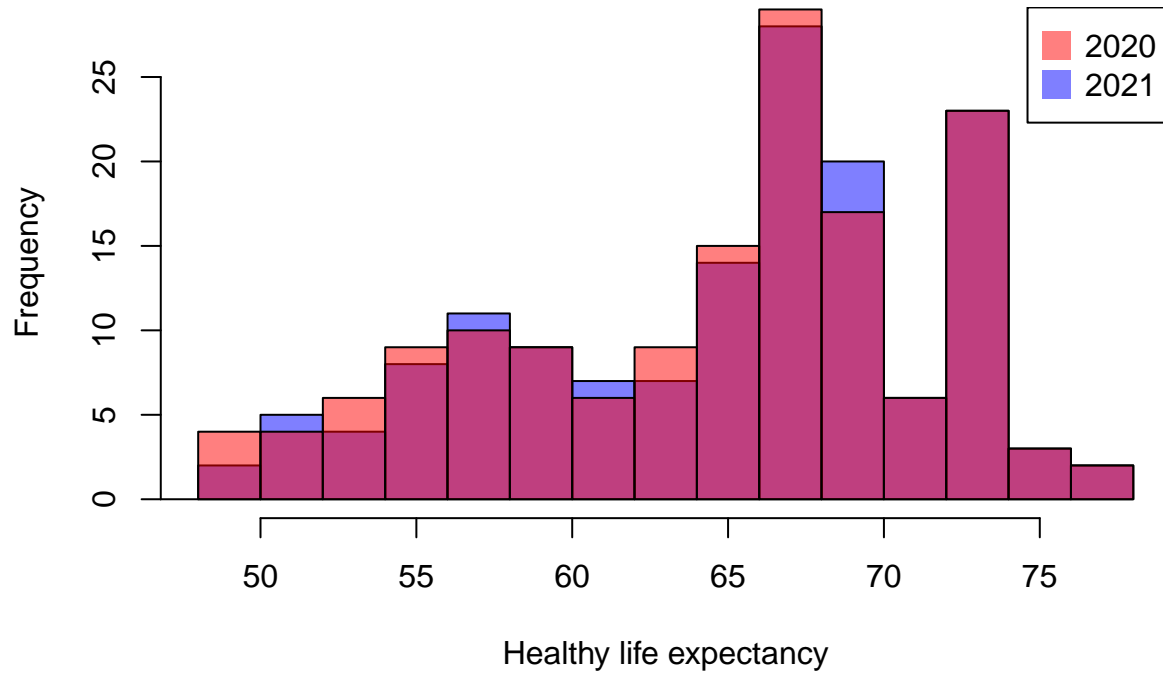
```
plot_by_years("Social support", "Social support histogram")
```

**Social support histogram**



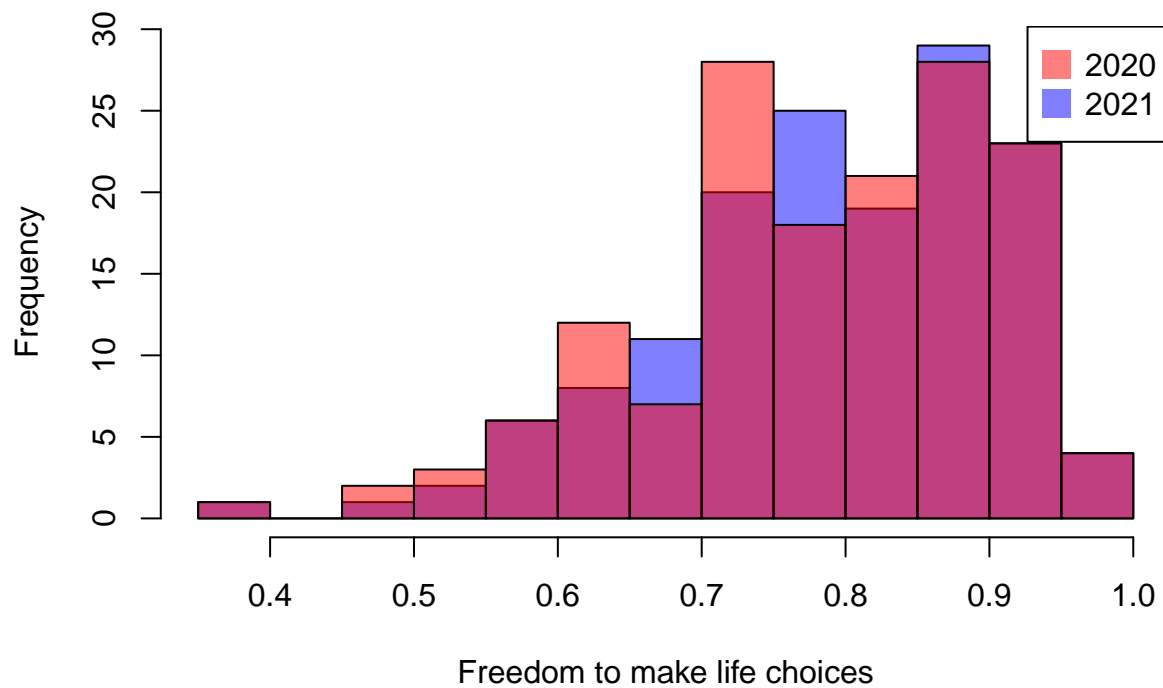
```
plot_by_years("Healthy life expectancy", "Healthy life expectancy histogram")
```

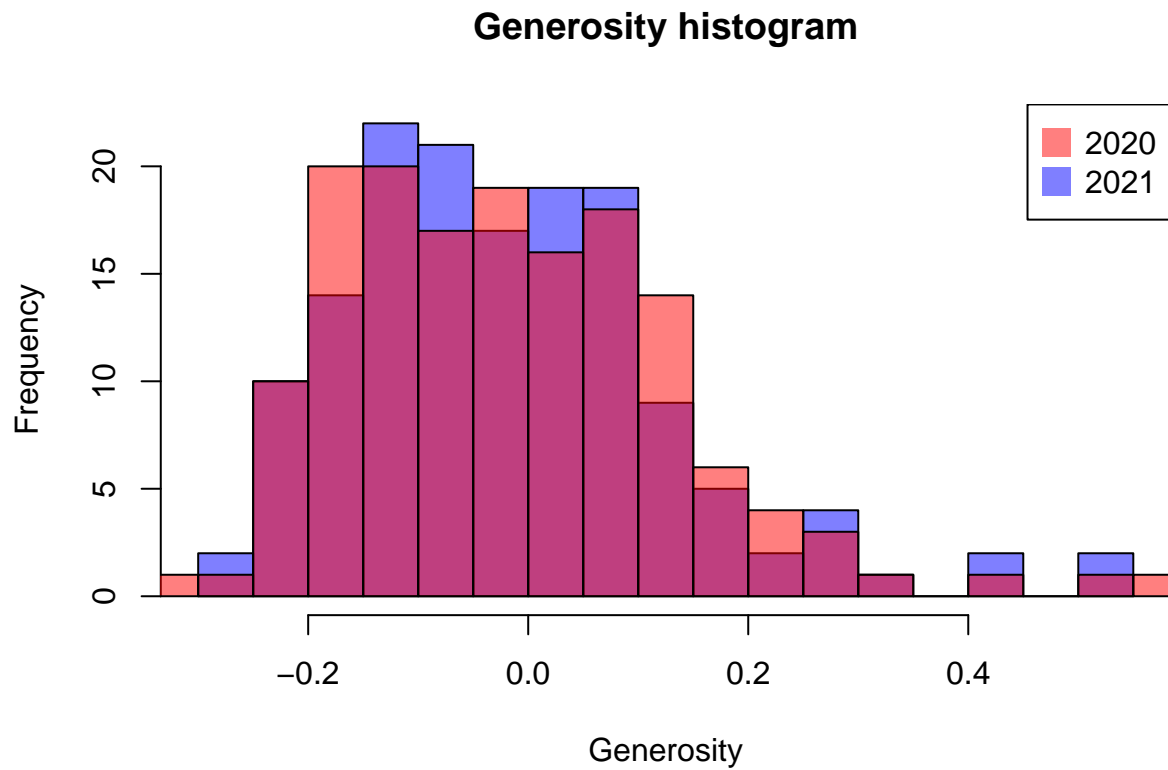
**Healthy life expectancy histogram**



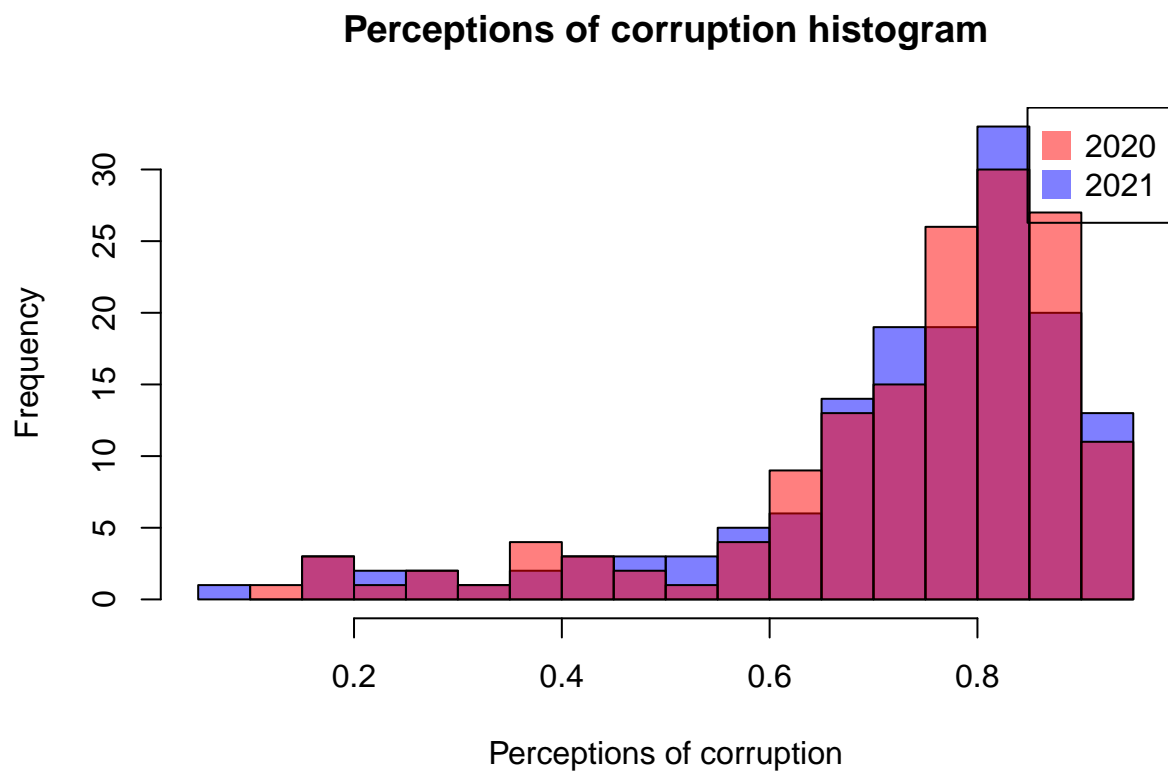
```
plot_by_years("Freedom to make life choices", "Freedom to make life choices histogram")
```

**Freedom to make life choices histogram**

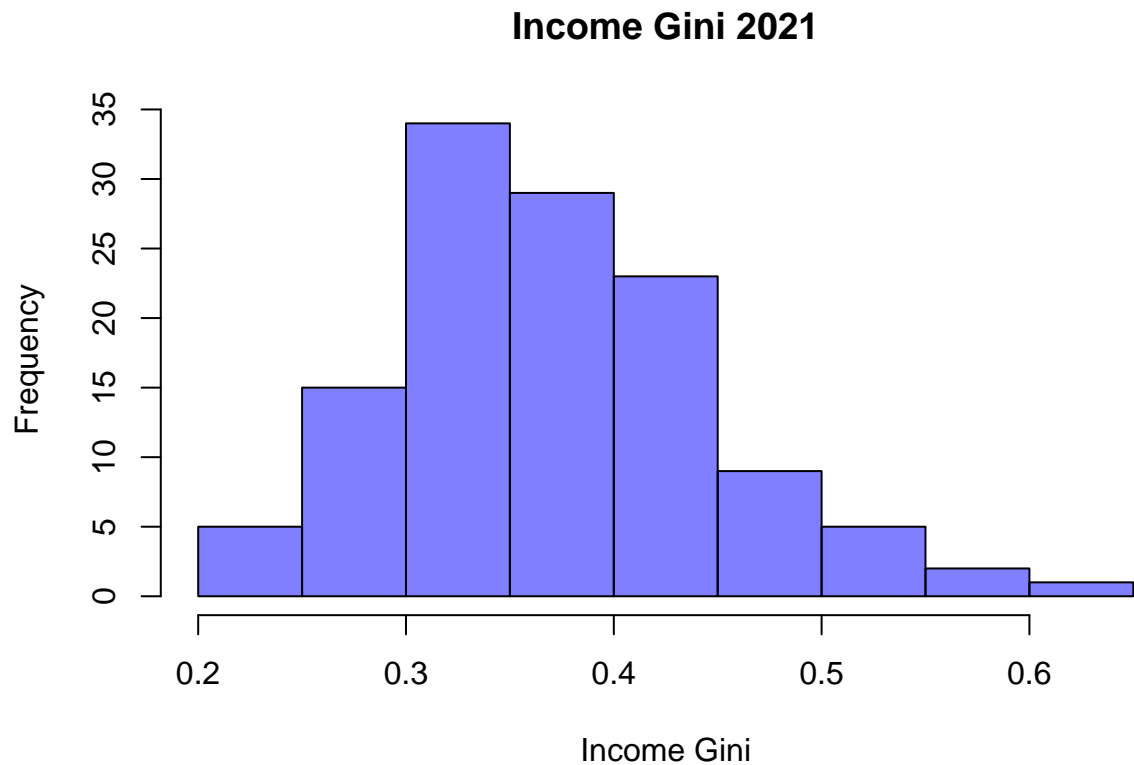




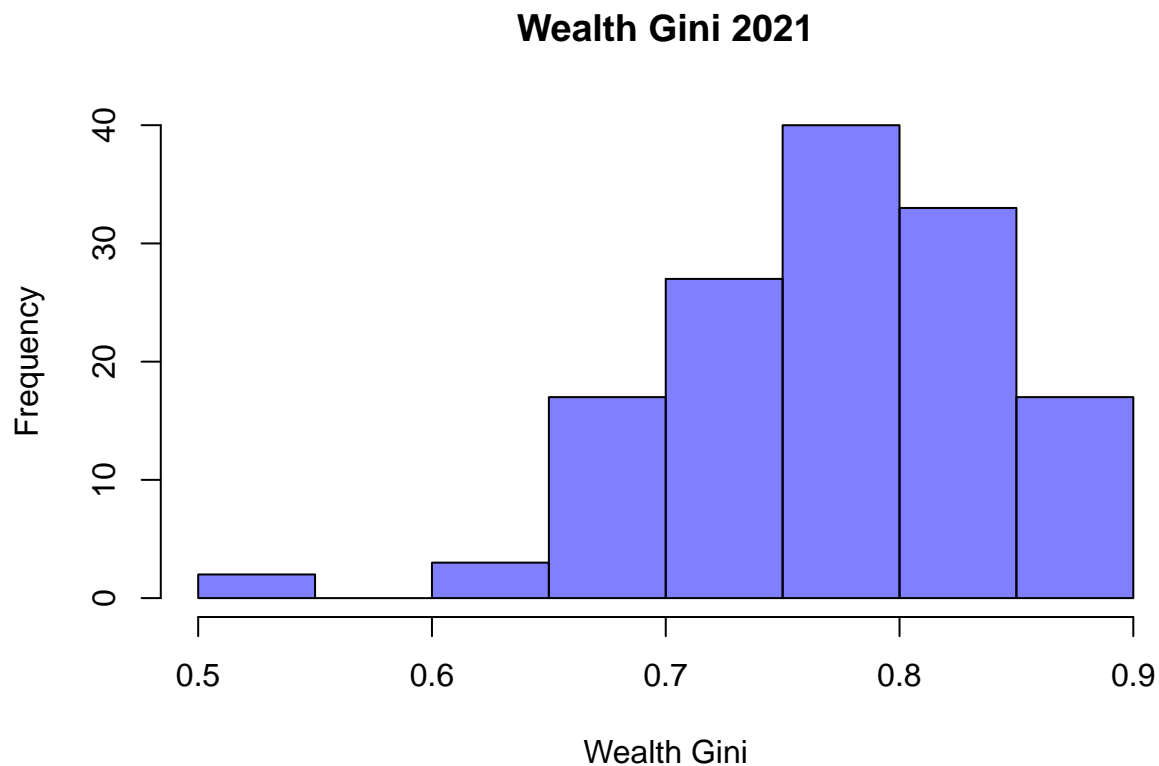
```
plot_by_years("Perceptions of corruption", "Perceptions of corruption histogram")
```



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



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



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

```
## # A tibble: 10 x 8
##   `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 3 more variables: Mean.Freedom <dbl>, Mean.Generosity <dbl>,
## #   Mean.Corruption <dbl>
```

```
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>
```

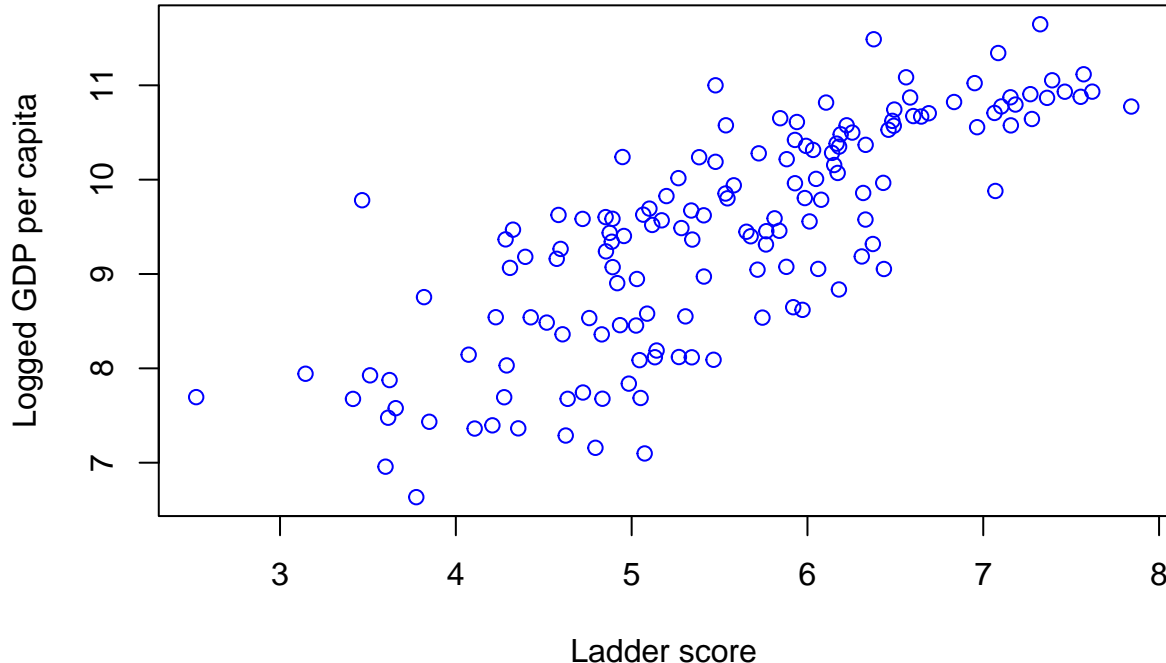
```
## 1 Central and Eastern Eur~      6.08   10.3      0.924      68.6
## 2 Commonwealth of Indepen~    5.47    9.53      0.891      65.1
## 3 East Asia                    5.76   10.6      0.86       71.8
## 4 Latin America and Carib~    5.99    9.45      0.857      67.6
## 5 Middle East and North A~    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).

```
# Ne razlikujemo vrste regija:
plot(whr2021$Ladder score`, whr2021$`Logged GDP per capita`,
     col="blue",
     xlab='Ladder score',
     ylab='Logged GDP per capita')
```

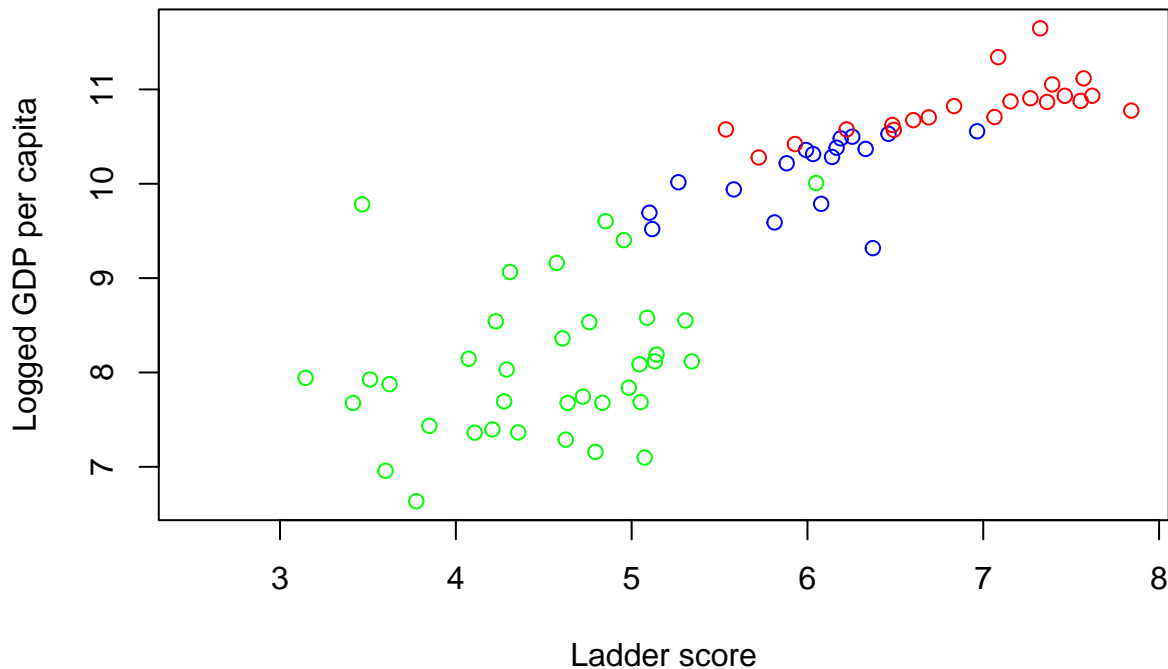


```
# Razlikujemo vrste regija:
plot(whr2021$Ladder score`[whr2021$`Regional indicator`=='Central and Eastern Europe'],
     whr2021$`Logged GDP per capita`[whr2021$`Regional indicator`=='Central and Eastern Europe'],
     col='blue',
     xlim=c(min(whr2021$Ladder score`),max(whr2021$Ladder score`)),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



```
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')
points(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Sub-Saharan Africa'],
whr2021$`Logged GDP per capita`[whr2021$`Regional indicator`=='Sub-Saharan Africa'],col='green')
```



Iz dijagrama raspršenja vidljiva je moguća povezanosti Ladder score s GDP per capita. Vidi se da što je veći GDP per capita, to je i razina sreće veća iskazana s Ladder score. Također vidimo da se na dijagramu razlikuju vrijednosti Zapadne, Srednje i Istočne Europe, te Sub-Saharske Afrike.

**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')

## Prosječan Ladder score zemalja iz Zapadne Europe 6.914905

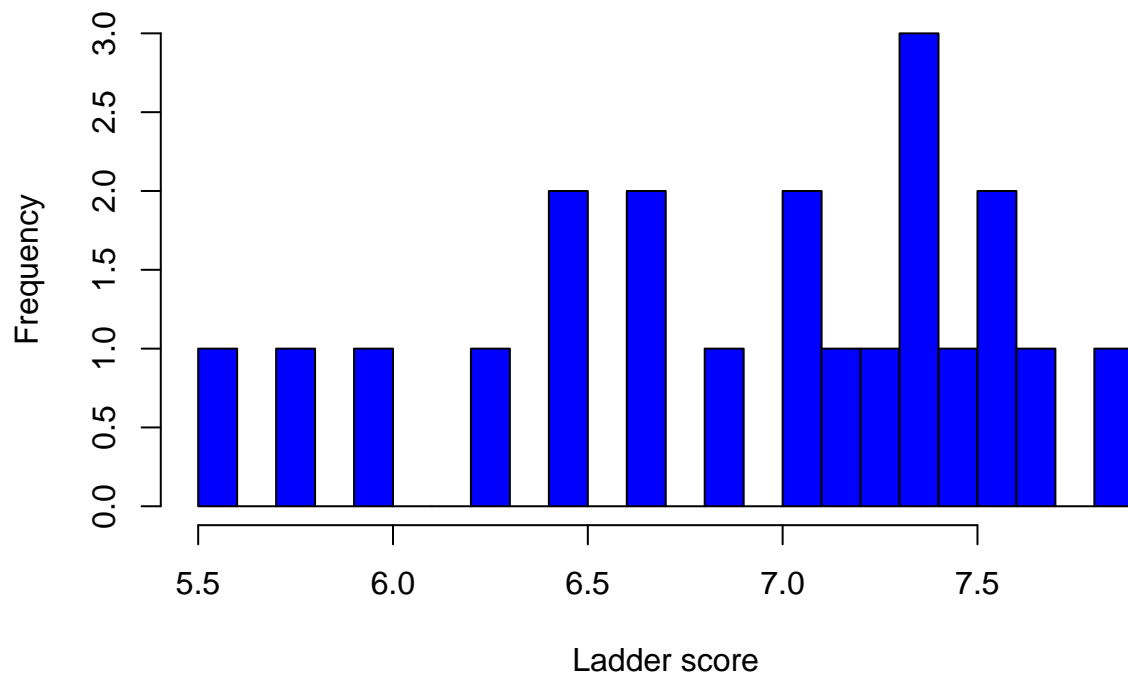
cat('Prosječan Ladder score zemalja iz Srednje i Istočne Europe', mean(central_eastern_europe$`Ladder score`), '\n')

## Prosječan Ladder score zemalja iz Srednje i Istočne Europe 5.984765
```

Histogrami za za Zapadnu i Centralnu/Istočnu Europu:

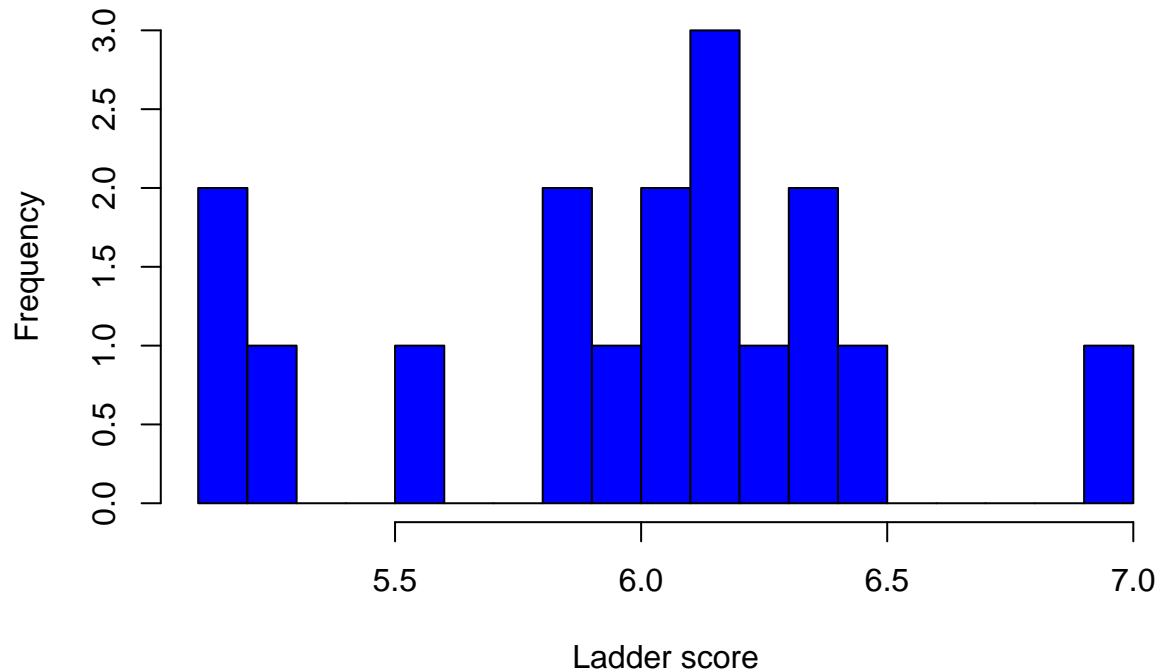
```
h = hist(western_europe$`Ladder score`,
main="Ladder score Western Europe",
xlab="Ladder score",
ylab='Frequency',
col="blue",
breaks = 20
)
```

## Ladder score Western Europe



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

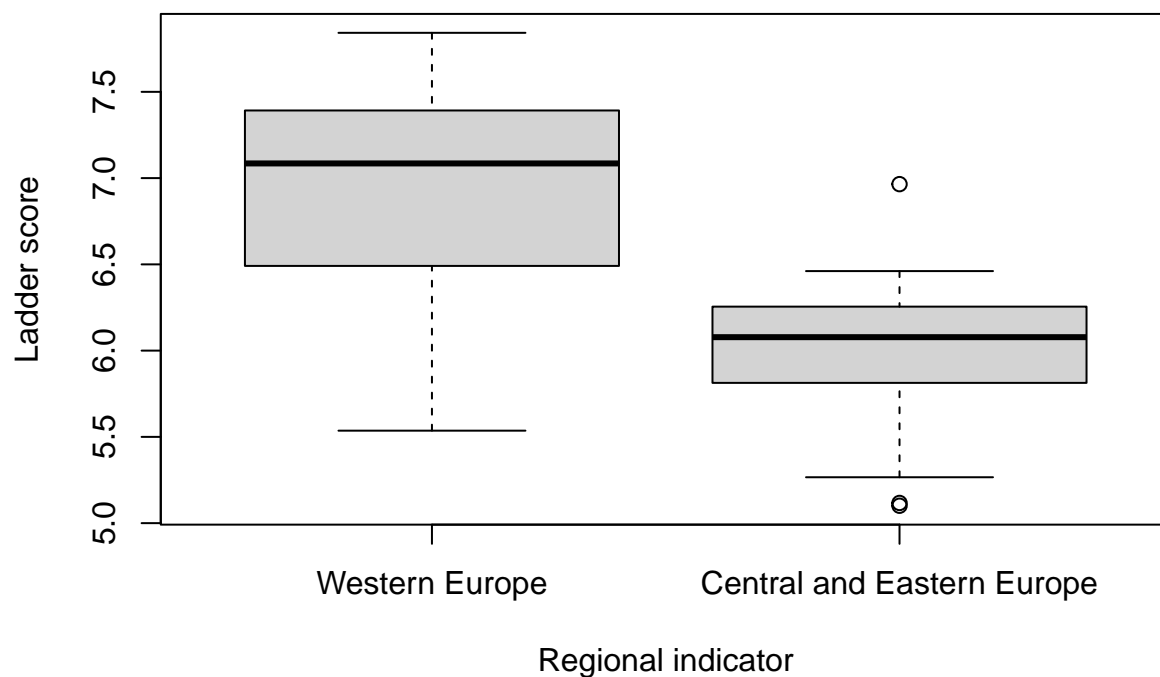
## Ladder score Central and Eastern Europe



vokutni 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 Europe"))
```

## Ladder score box-plot



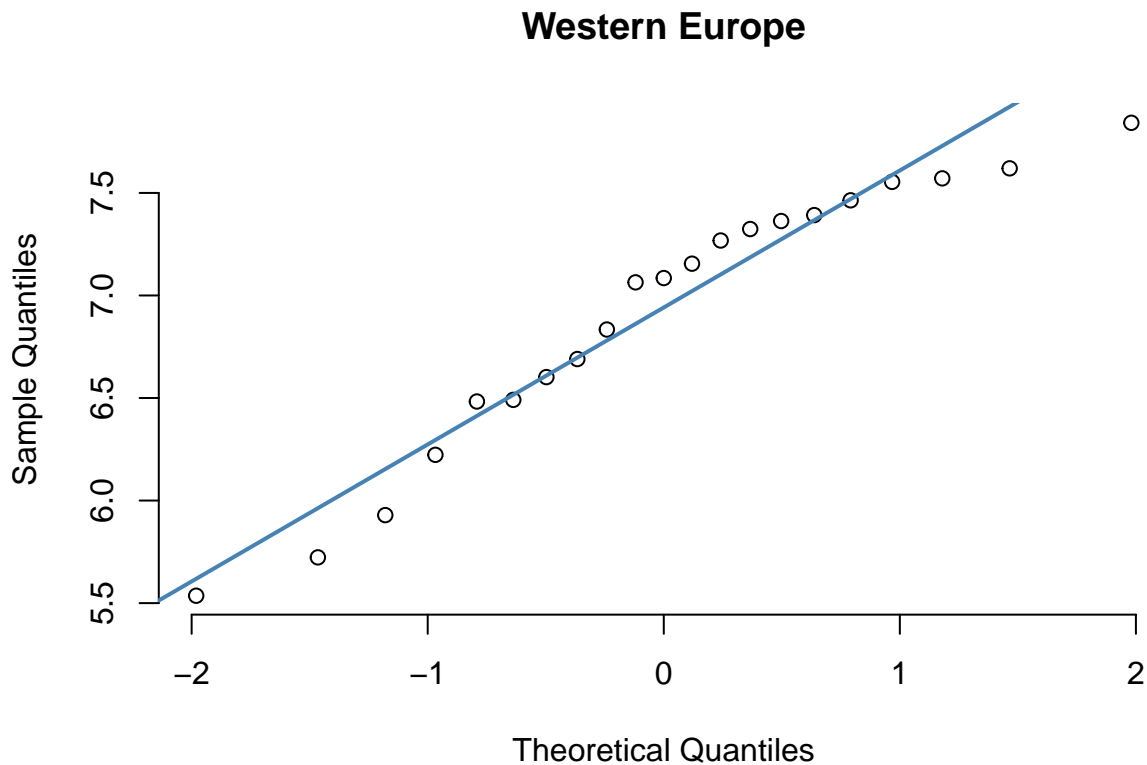
toje indikacije da bi ljudi iz zemalja Zapadne Europe trebali biti sretniji od ljudi iz zemalja Srednje i Istočne

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

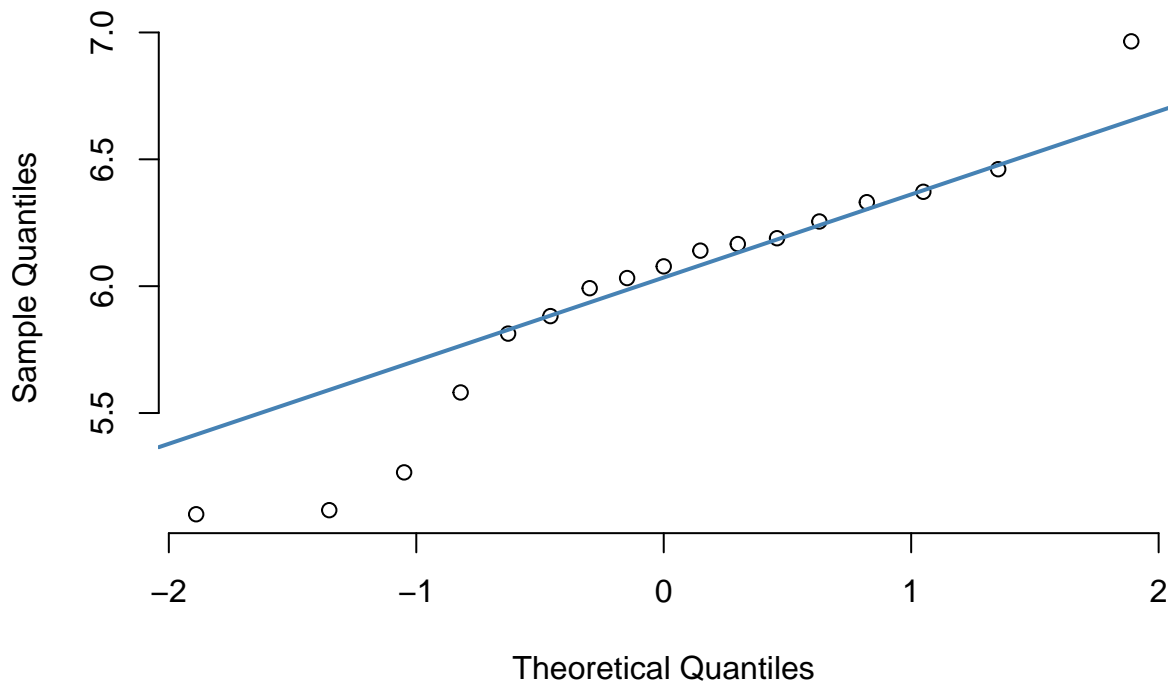
Ovakvo ispitivanje možemo provesti t-testom. Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo dva uzoraka iz dvije različite regije, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka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

## Central and Eastern Europe



Koris-

timo 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

#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.

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

```
##
## 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
```

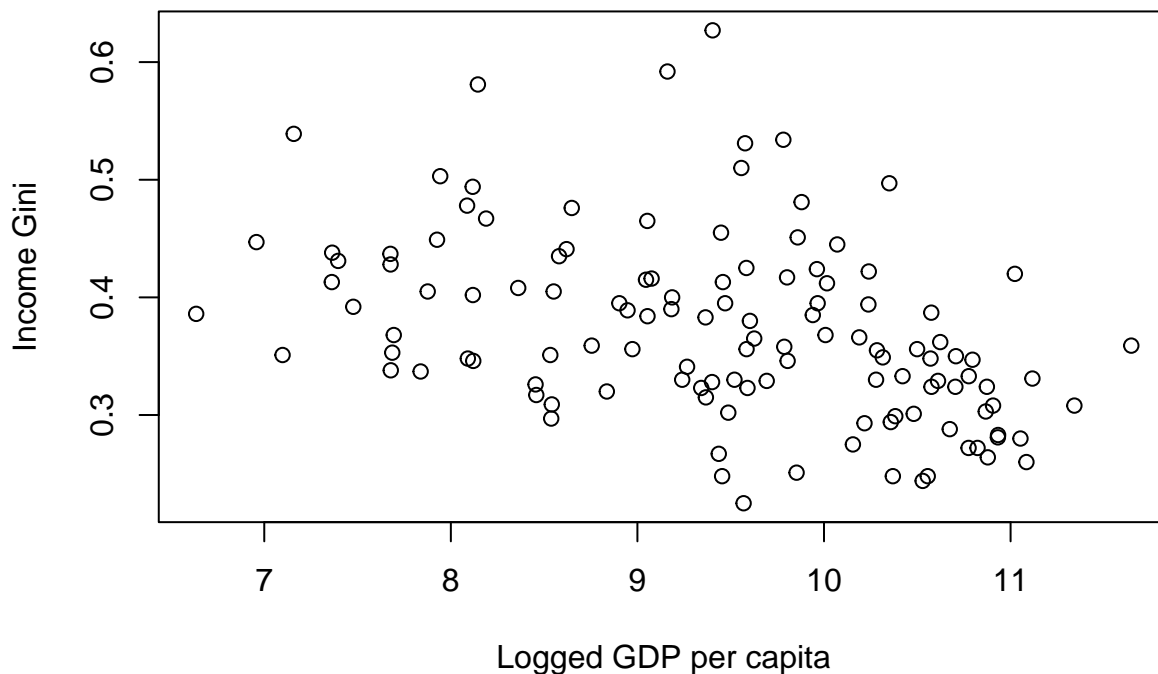
Zbog male p-vrijednosti možemo odbaciti hipotezu  $H_0$  u korist alternative da je Ladder score veći u Zapadnoj Europi od onog u Srednjoj i Istočnoj Europi.

## 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.

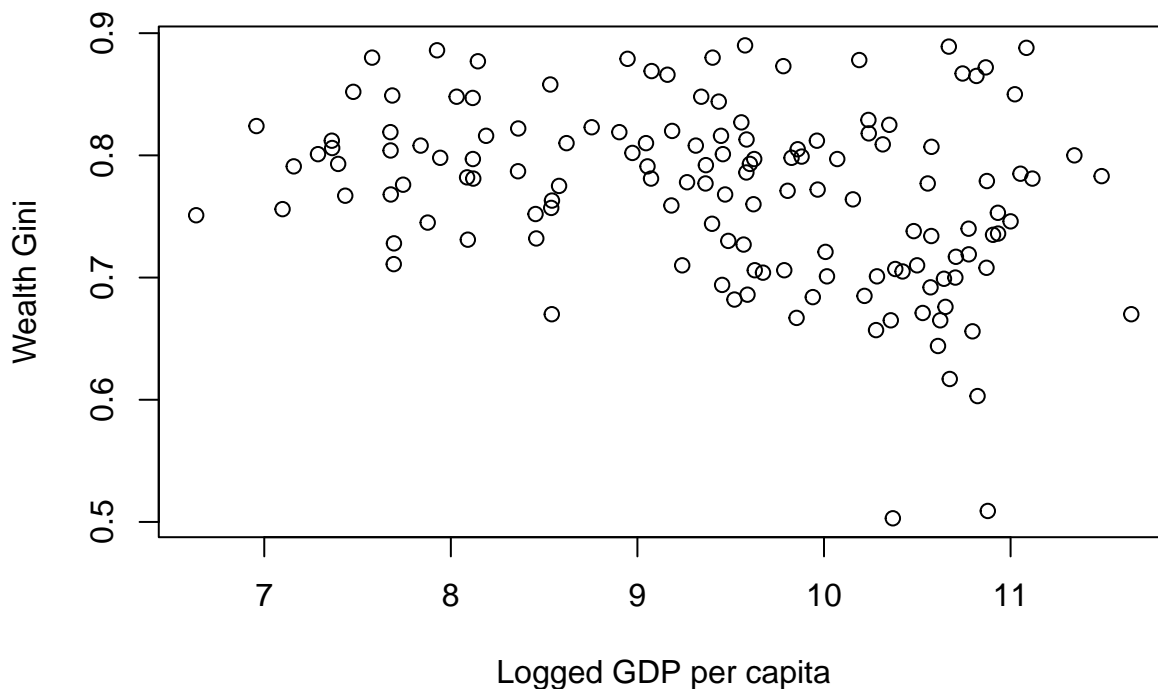
```
plot(whr2021$`Logged GDP per capita`, whr2021$`Income Gini`, xlab = "Logged GDP per capita", ylab = "Income Gini",
     main = "Distribucija log(BDP) u ovisnosti o nejednakosti dohotka")
```

## Distribucija log(BDP) u ovisnosti o nejednakosti dohotka



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

## 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 nejednakost 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

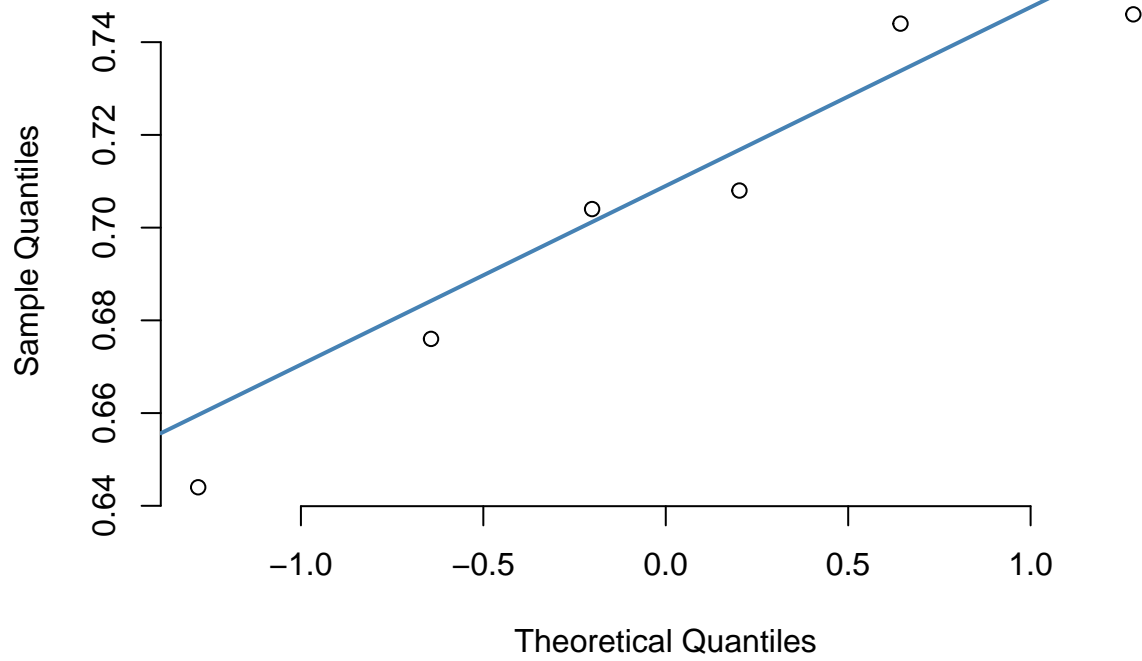
Ovakvo ispitivanje možemo provesti t-testom. Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo uzorke država različitih regija, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka koju ćemo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

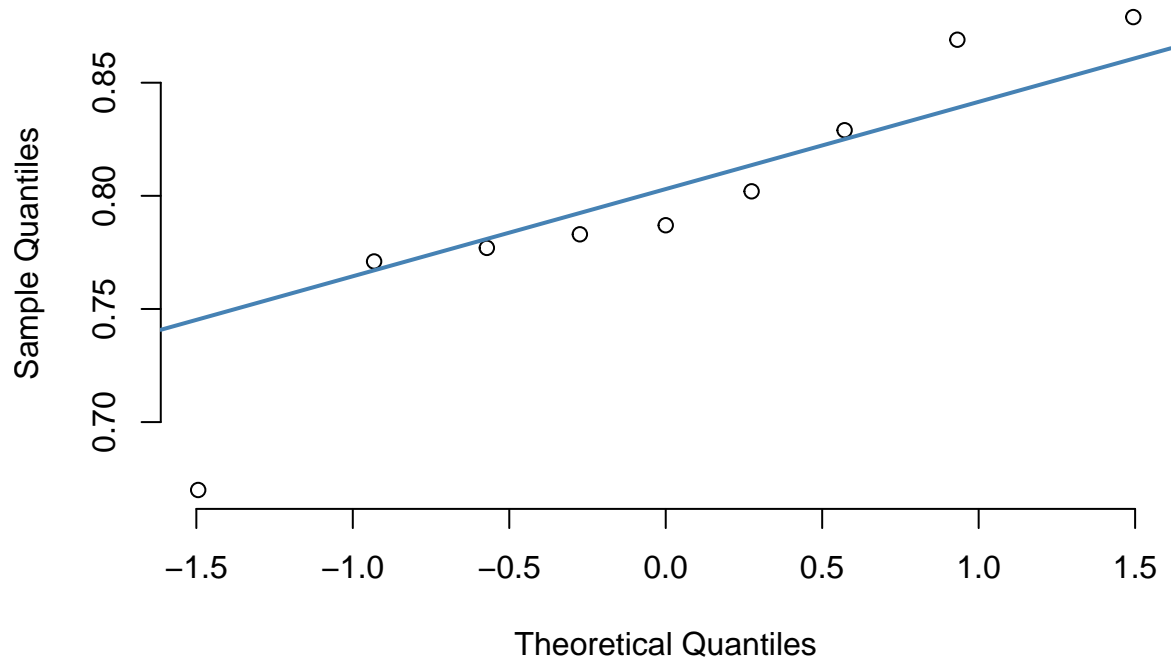


## Wealth Gini – East Asia



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

## Wealth Gini – Southeast Asia



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

##

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

```
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.

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

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

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

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

```
#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.

```
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

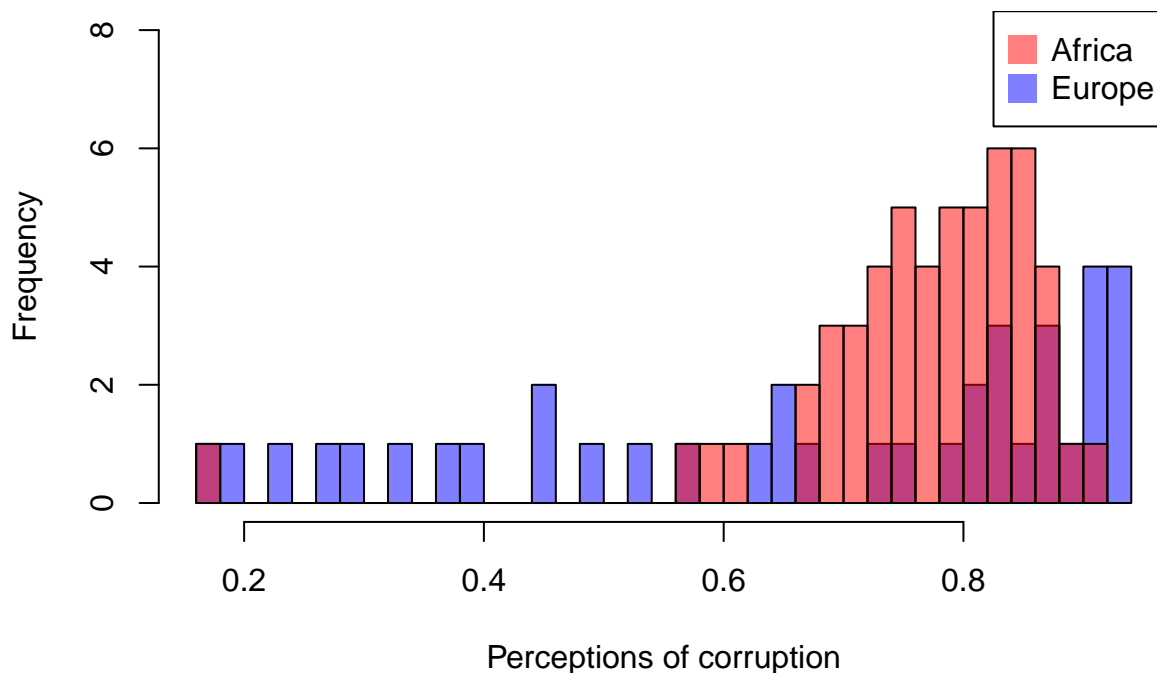
Pokušajmo sada zaključiti nešto o korupciji. Promatrat ćemo zemlje Europe i Afrike te želimo saznati gdje je korupcija zastupljenija. Ispitat ćemo zavisnost percepcije korupcije o logaritmu BDP-a po stanovniku.

```
ce_europe = whr2021[whr2021$`Regional indicator` == "Central and Eastern Europe",]
w_europe = whr2021[whr2021$`Regional indicator` == "Western Europe",]
men_africa = whr2021[whr2021$`Regional indicator` == "Middle East and North Africa",]
ss_africa = whr2021[whr2021$`Regional indicator` == "Sub-Saharan Africa",]

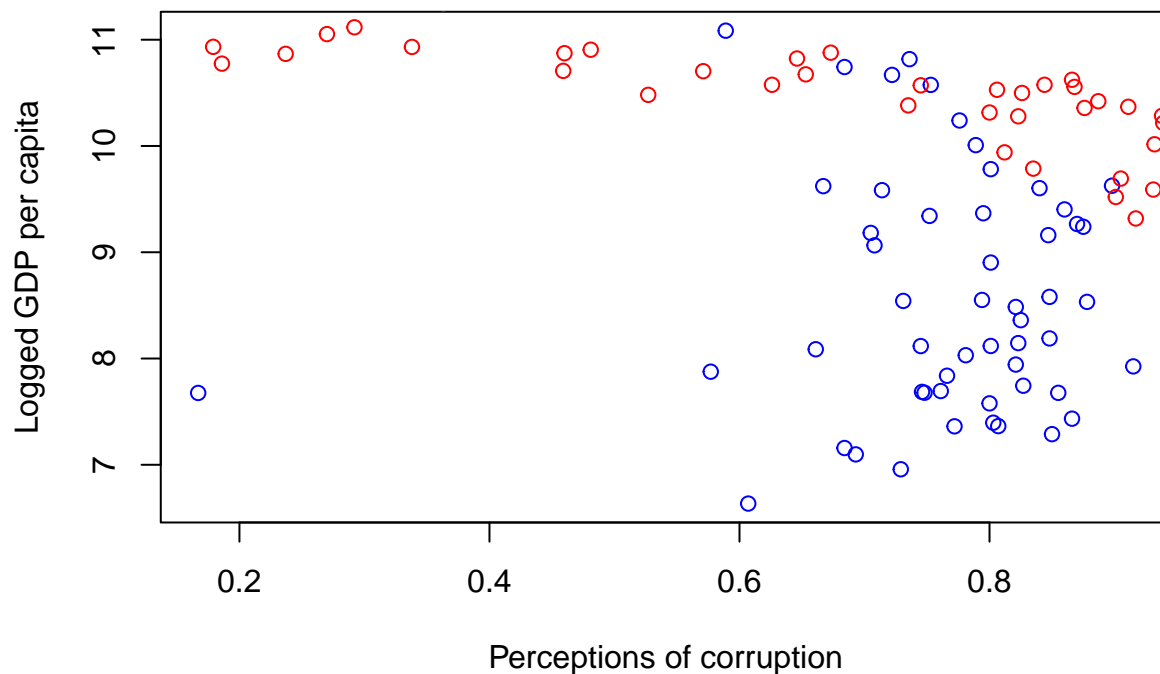
europe <- rbind(ce_europe, w_europe)
africa <- rbind(men_africa, ss_africa)

plot_by_gini <- function(column, main) {
  hist(europe[[column]], breaks=30, main=main, xlab=column, ylab="Frequency", ylim = c(0,8), col=rgb(0,0,1,0.5), add=T)
  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 = 1)
}
plot_by_gini("Perceptions of corruption", "Perceptions of corruption histogram")
```

Perceptions of corruption histogram

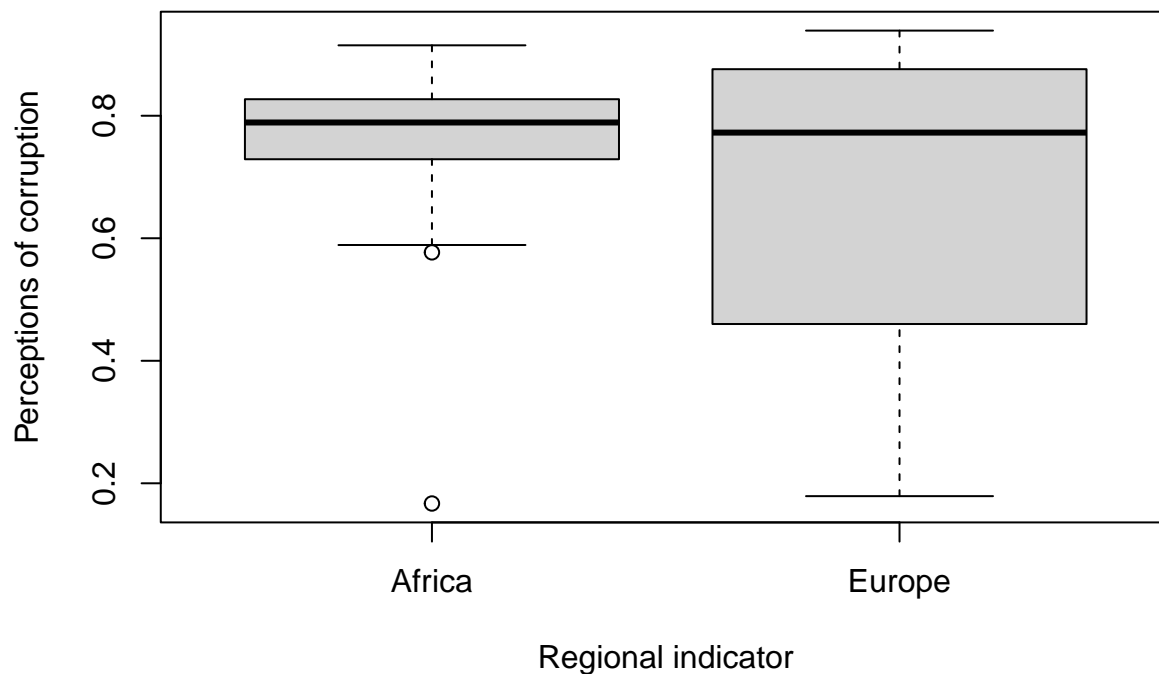


```
plot(africa$`Perceptions of corruption`,
     africa$`Logged GDP per capita`,
     col='blue',
     ylab='Logged GDP per capita',
     xlab='Perceptions of corruption')
points(europe$`Perceptions of corruption`,
       europe$`Logged GDP per capita`, col='red')
```



```
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"))
```

**Perceptions of corruption box-plot**



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
```

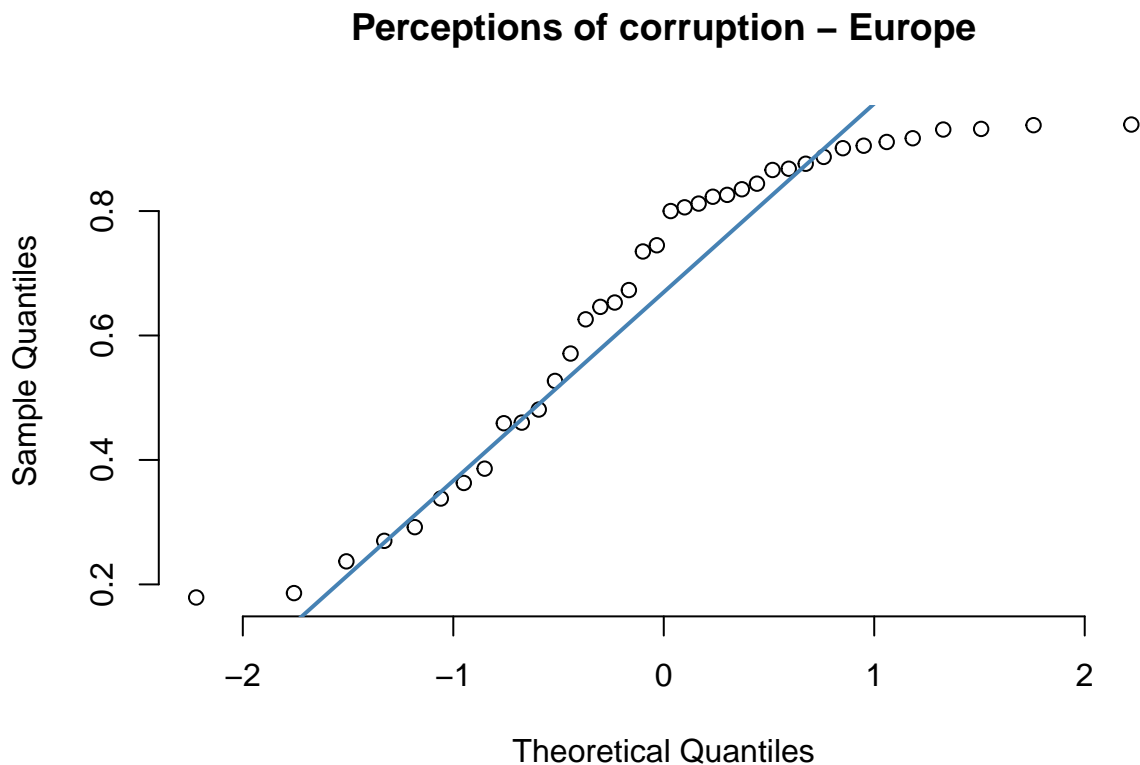
```
print(mean_africa)
```

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

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

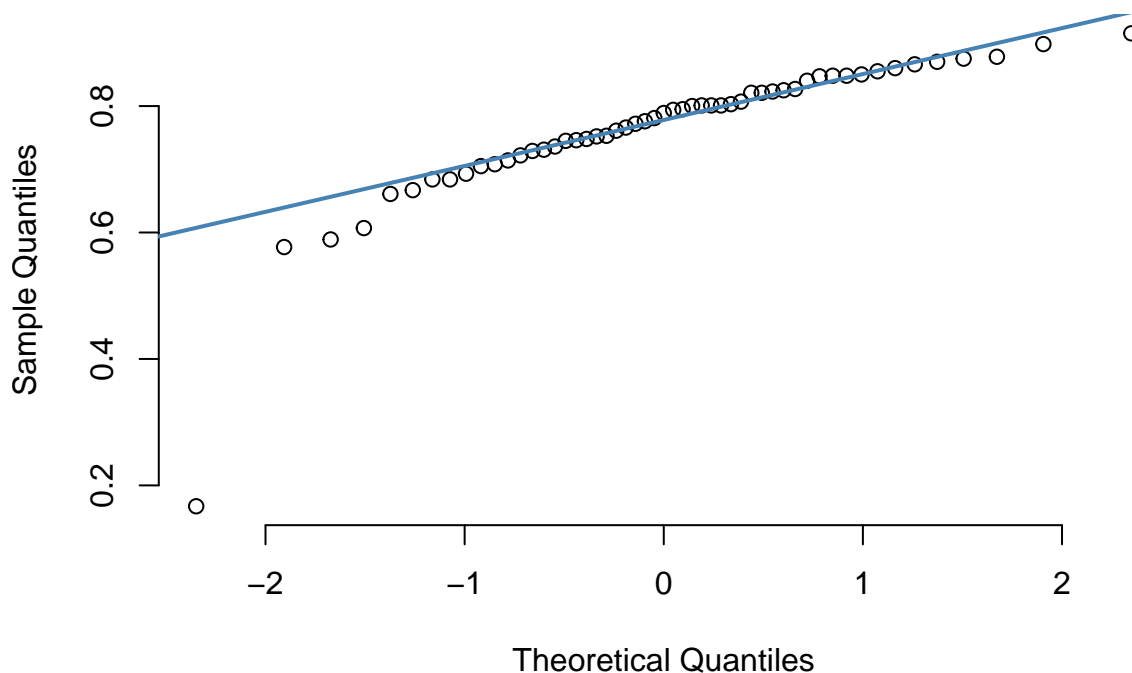
Postavimo hipoteze:  $H_0$ : srednja vrijednost percepcije korupcije za Europu i Afriku je jednaka  $H_1$ : srednja vrijednost percepcije korupcije za Europu je manja od srednje vrijednosti za Afriku. Ovakvo ispitivanje možemo provesti t-testom. Kako bi mogli provesti test, moramo najprije provjeriti 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$`Perceptions of corruption`, pch = 1, frame = FALSE, main='Perceptions of corruption - Europe')
qqline(europe$`Perceptions of corruption`, col = "steelblue", lwd = 2)
```



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

## Perceptions of corruption – Africa



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 distribuciju.

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 hipotezu  $H_0$  da podaci dolaze iz normalne distribucije. Ne možemo provesti t-test. Jedan od mogućih rješenja je transformirati podatke i provesti jackknife.

## Usporedba razina sreće u 2020. i 2021. godini.

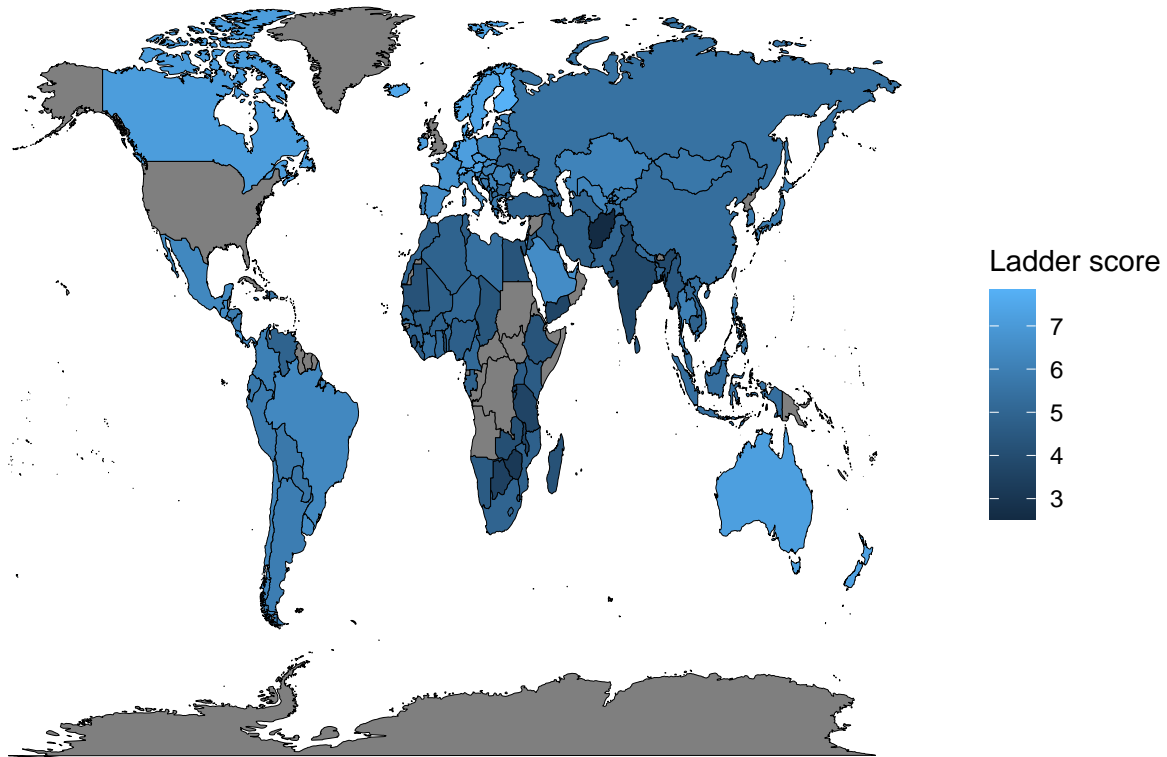
```
library(ggplot2)

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

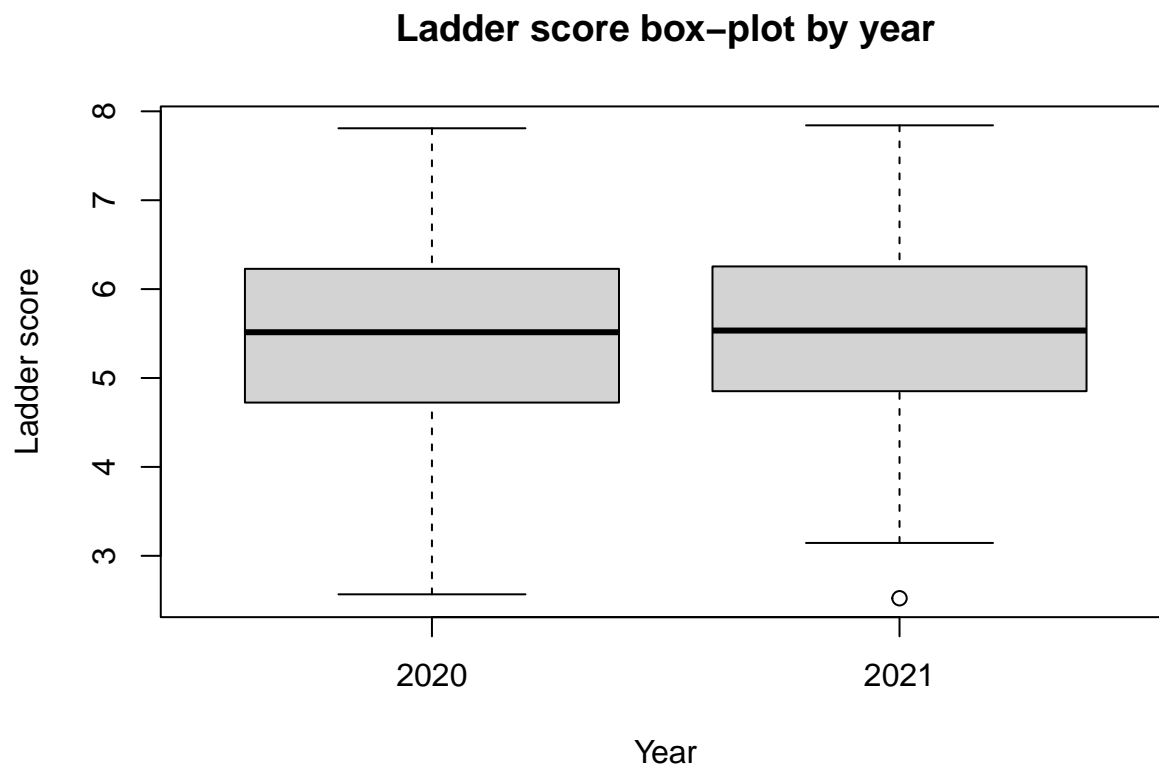
```
mapdata2021 = map_data("world")
mapdata2021 = left_join(mapdata2021, data2021, by = "region")

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_b
    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.



Provest ćemo test o jednakosti aritmetičkih sredina za različite godine. Hipoteze su:

$H_0$ : aritmetičke sredine su jednake

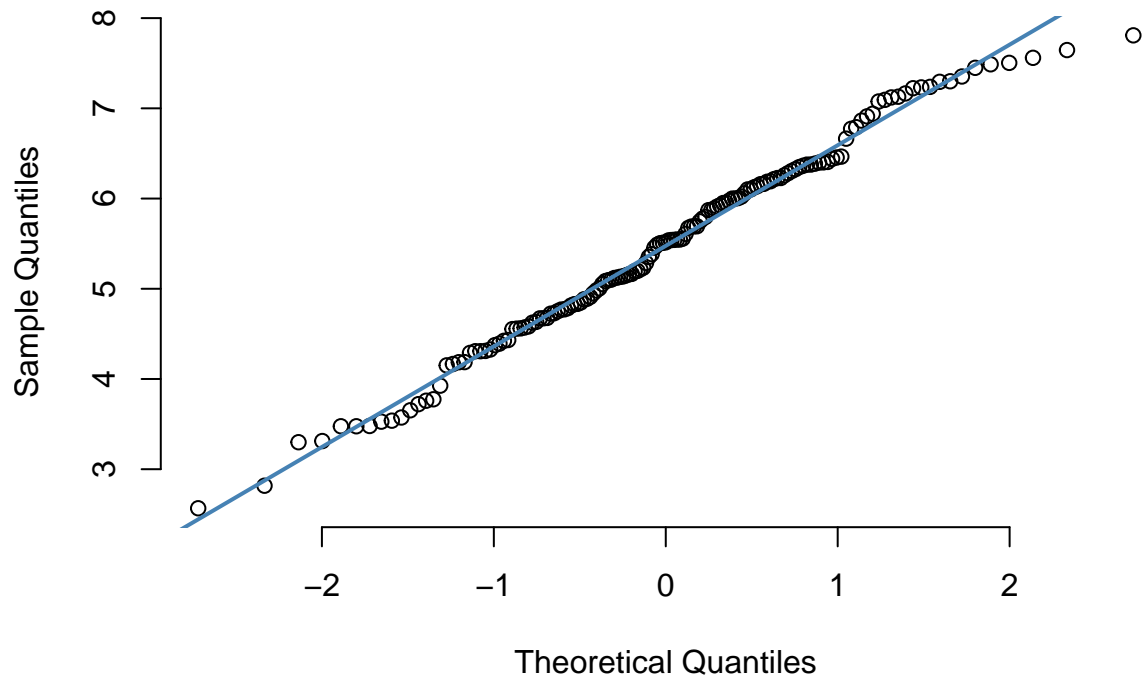
$H_1$ : aritmetičke sredine nisu jednake

Prije provedbe t-testa provjeravamo pretpostavke normalnosti uzorka. S obzirom na to da razmatramo dva uzoraka iz dvije različite regije, možemo pretpostaviti njihovu nezavisnost. Normalnost podataka provjeravamo qq-plotom.

```
qqnorm(whr2020$Ladder score`, pch = 1, frame = FALSE, main='Ladder score for 2020')
qqline(whr2020$Ladder score`, col = "steelblue", lwd = 2)
```

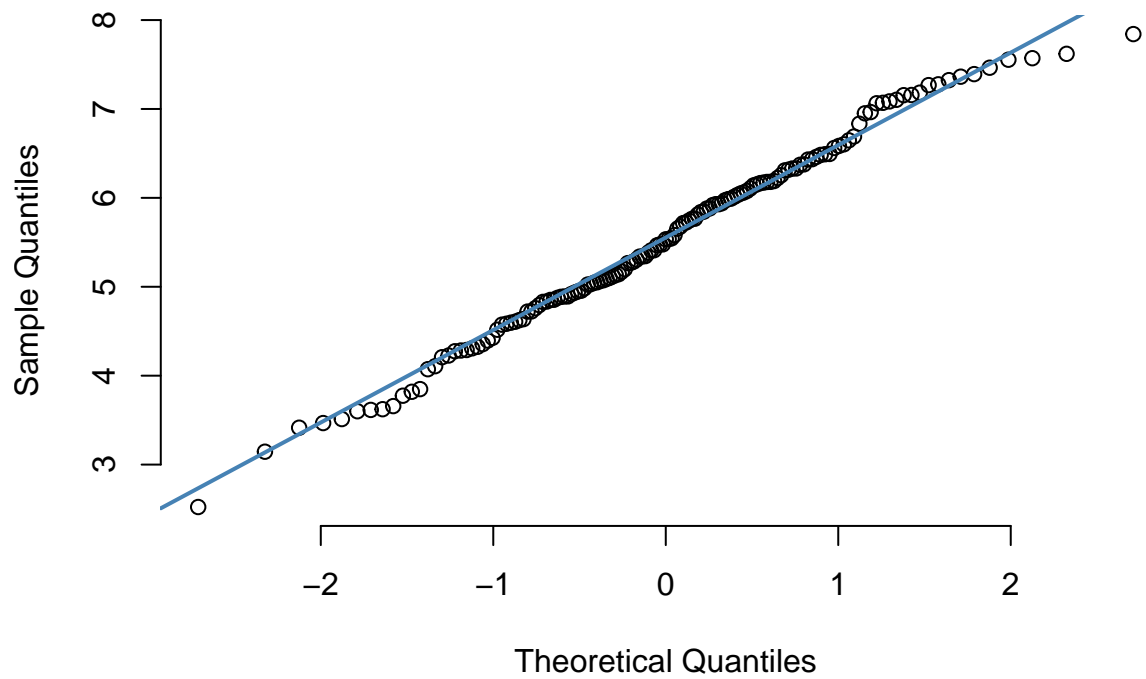


### Ladder score for 2020



```
qqnorm(whr2021$`Ladder score`, pch = 1, frame = FALSE, main='Ladder score for 2021')  
qqline(whr2021$`Ladder score`, col = "steelblue", lwd = 2)
```

### Ladder score for 2021



sugerira na normalnost.

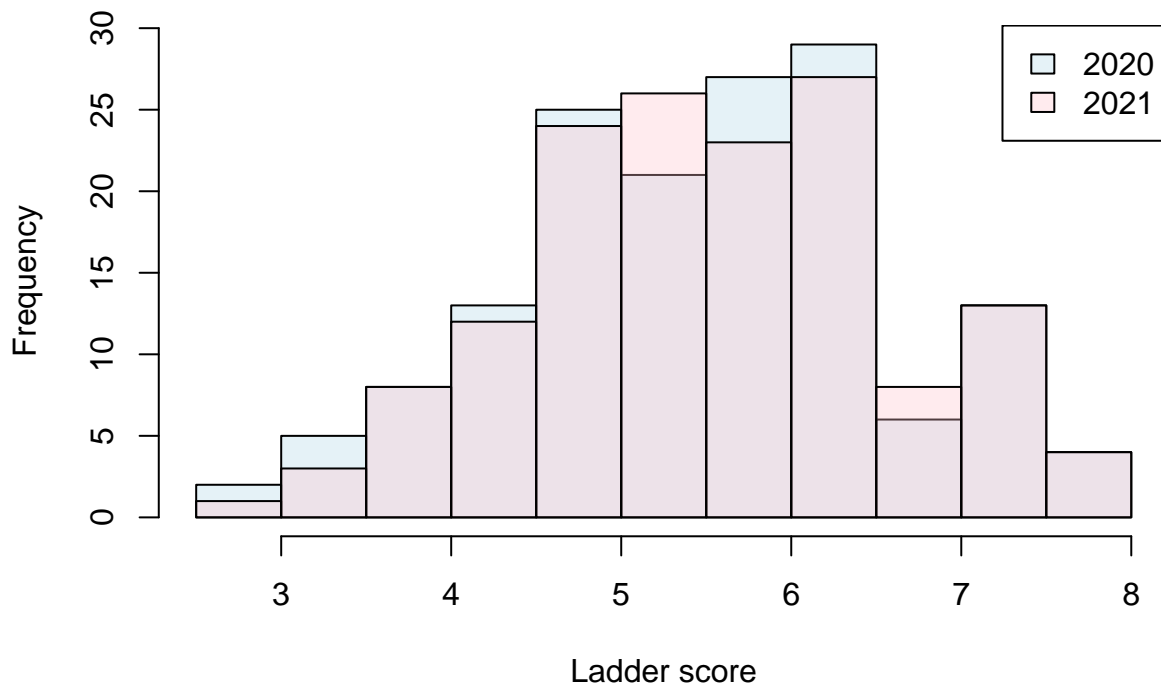
QQ-plot

```
t.test(whr2020$`Ladder score`, whr2021$`Ladder score`, alternative = "two.sided", var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: whr2020$`Ladder score` and whr2021$`Ladder score`
## t = -0.47341, df = 300, p-value = 0.6363
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.3072685 0.1881005
## sample estimates:
## mean of x mean of y
## 5.473255 5.532839
```

Na temelju rezultata možemo zaključiti da su aritmetičke sredine razine sreće za dvije godine jednake.

### Histogram of ladder score for two years

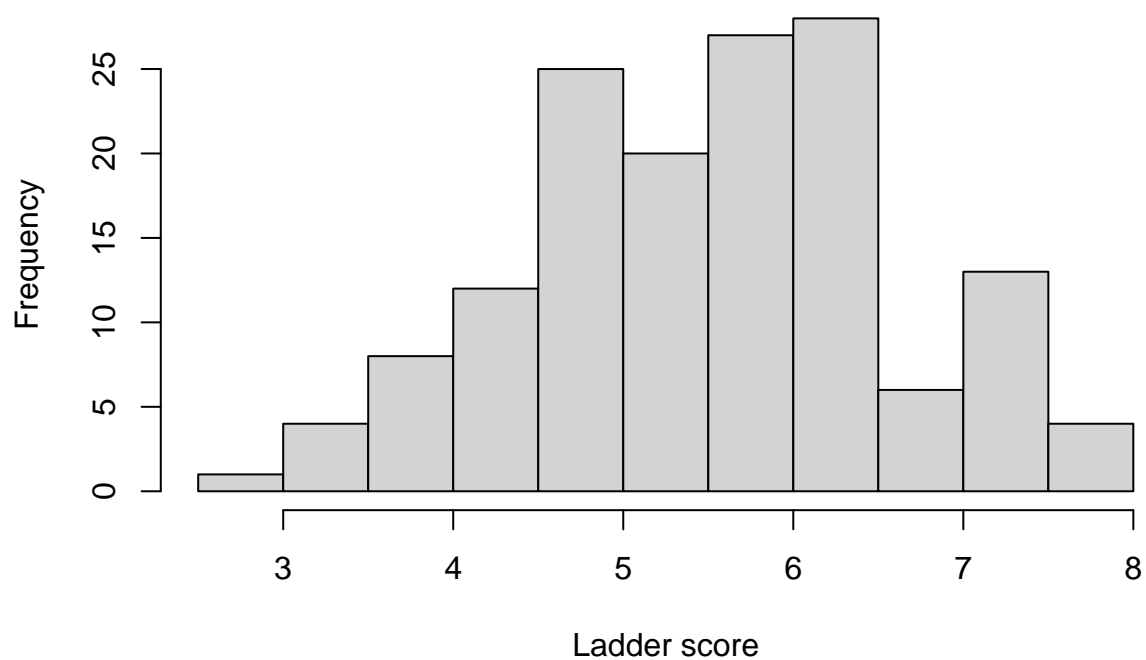


jimo podatke iz dvije godine:

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

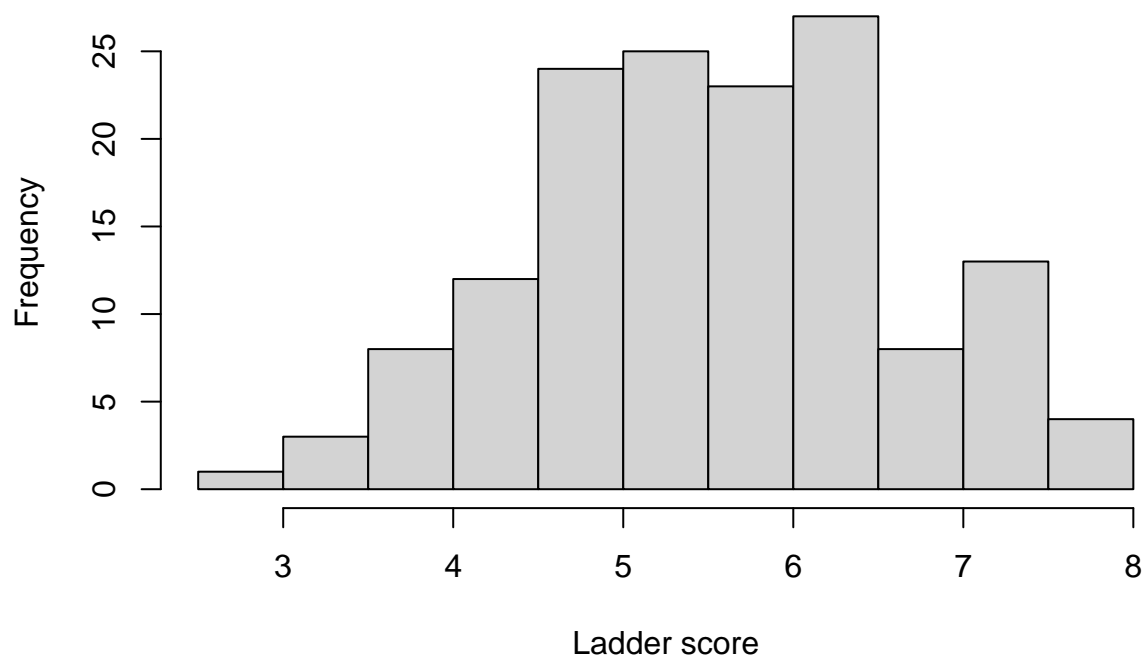
hist(mergedData$`Ladder score.20`,
     main=paste('Histogram of ladder score in 2020'),
     xlab='Ladder score')
```

### Histogram of ladder score in 2020



```
hist(mergedData$`Ladder score.21`,  
     main=paste('Histogram of ladder score in 2021'),  
     xlab='Ladder score')
```

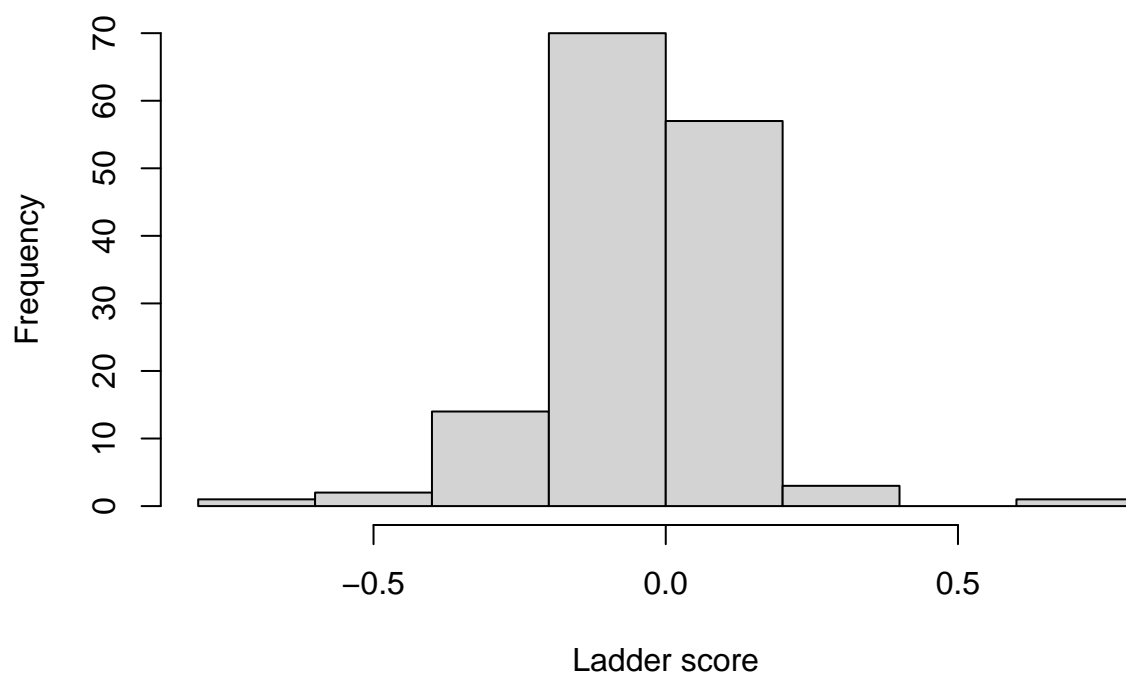
### Histogram of ladder score in 2021



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

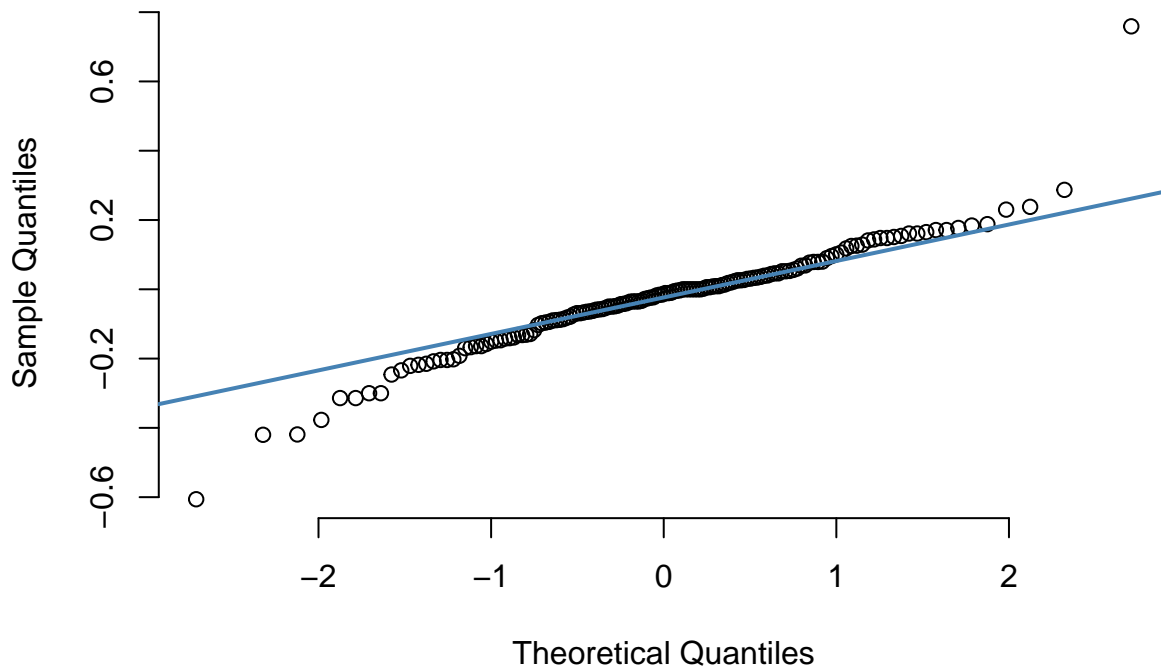
```
xlab='Ladder score')
```

## Difference in ladder scores between two years



```
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)
```

## QQ-plot for differences between ladder scores



Histogram razlika nam sugerira normalnost podataka, dok iz qq-plota vidimo malo odstupanje lijevog repa. Pod pretpostavkom da su podaci normalni, koristimo upareni t-test.

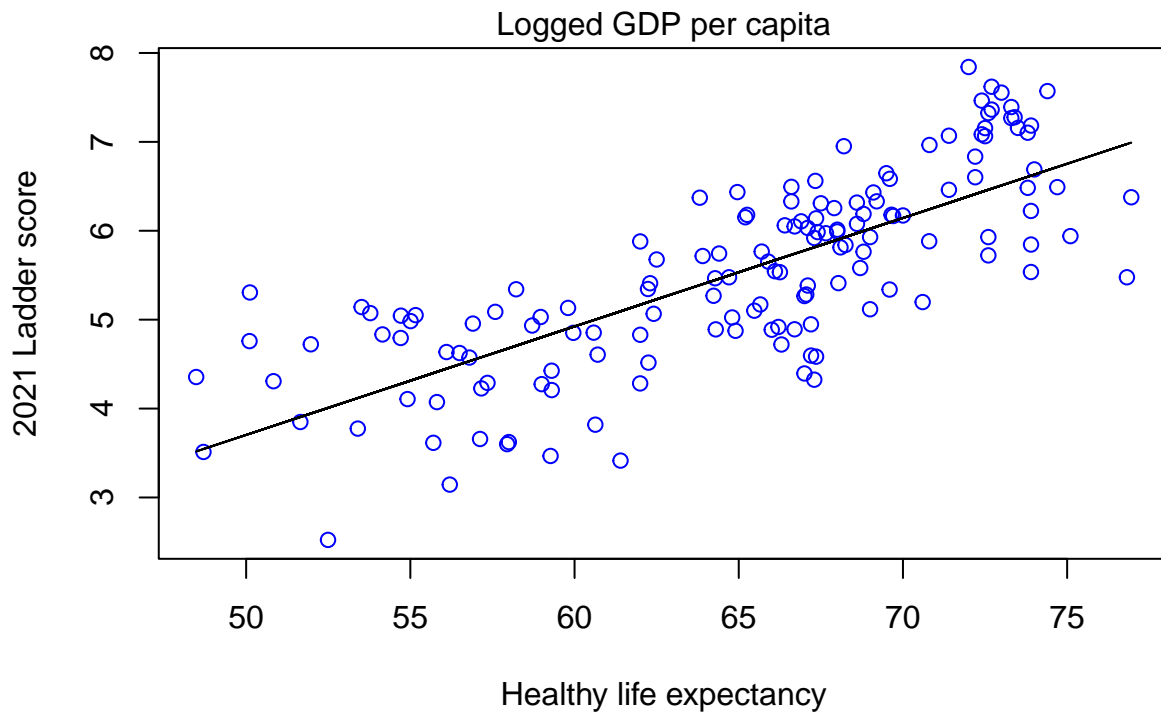
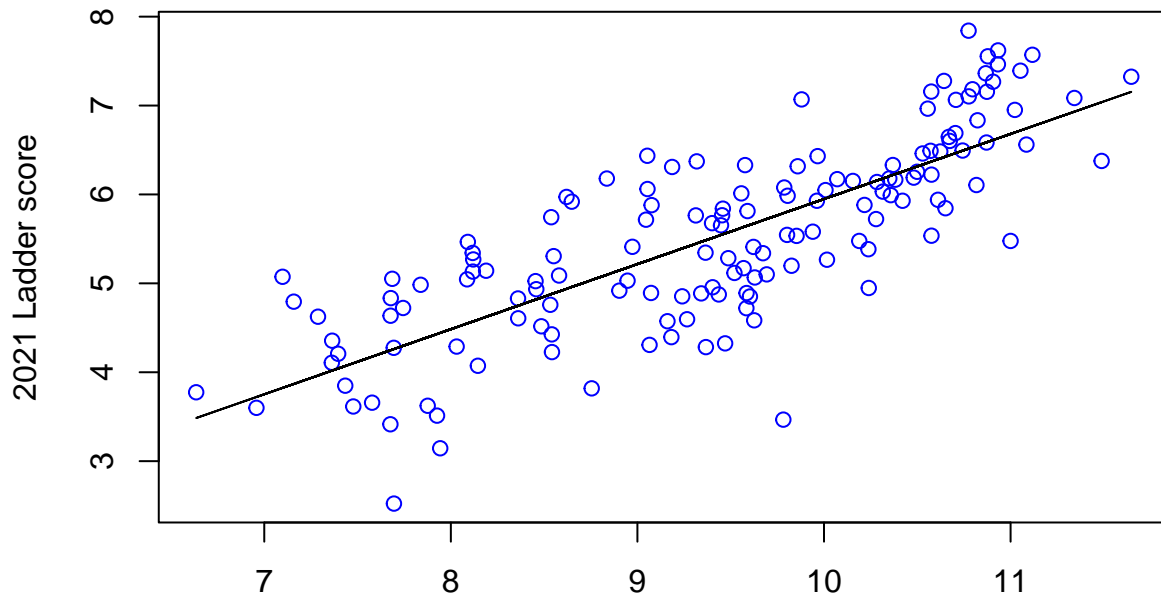
```
t.test(mergedData$`Ladder score.20`,
       mergedData$`Ladder score.21`,
       paired = TRUE,
       alt = "less")

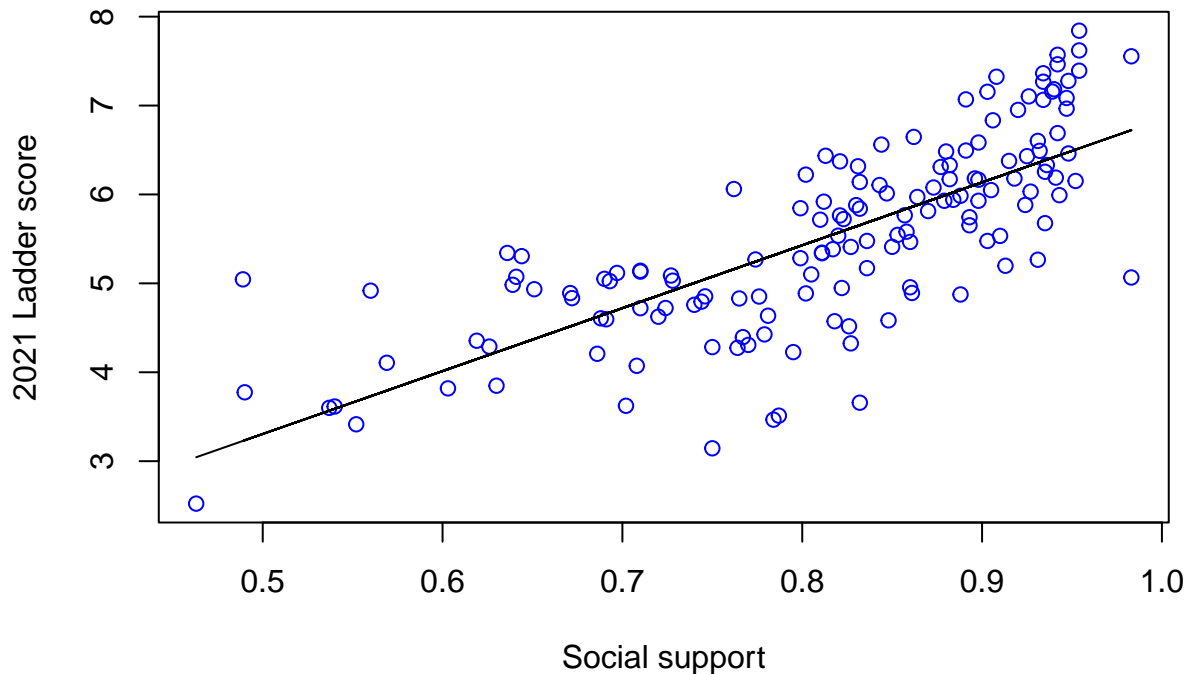
##
## Paired t-test
##
## data: mergedData$`Ladder score.20` and mergedData$`Ladder score.21`
## t = -2.0749, df = 147, p-value = 0.01987
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -0.005247129
## sample estimates:
## mean of the 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 pojedinih država.

## Ovisnost razine sreće o drugim varijablama u 2021. godini

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





Izračunavamo koeficijente linearne regresije.

```
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
```

```
summary(fitHealth)
```

```
##
## Call:
## lm(formula = whr2021$Ladder score ~ whr2021$Healthy life expectancy)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-1.67956	-0.49054	0.09232	0.56360	1.58818

```
##
## Coefficients:
```

```

##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.395426   0.548061  -4.371 2.33e-05 ***
## whr2021$`Healthy life expectancy` 0.121987   0.008388  14.544 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.69 on 147 degrees of freedom
## Multiple R-squared:  0.59, Adjusted R-squared:  0.5872
## F-statistic: 211.5 on 1 and 147 DF, p-value: < 2.2e-16

fitGDP_WGini = lm(whr2021$Ladder score` ~ whr2021$`Logged GDP per capita` + whr2021$`Wealth Gini`)
fitGDP_Freedom = lm(whr2021$Ladder score` ~ whr2021$`Logged GDP per capita` + whr2021$`Freedom to make
summary(fitGDP_WGini)

##
## Call:
## lm(formula = whr2021$Ladder score` ~ whr2021$`Logged GDP per capita` +
##     whr2021$`Wealth Gini`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20258 -0.42782  0.04803  0.51928  1.31363
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.5478     0.8859  -0.618   0.537
## whr2021$`Logged GDP per capita`  0.7273     0.0489  14.872 <2e-16 ***
## whr2021$`Wealth Gini`          -1.0272     0.8078  -1.272   0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6463 on 136 degrees of freedom
## (10 observations deleted due to missingness)
## Multiple R-squared:  0.6574, Adjusted R-squared:  0.6524
## F-statistic: 130.5 on 2 and 136 DF, p-value: < 2.2e-16

summary(fitGDP_Freedom)

##
## Call:
## lm(formula = whr2021$Ladder score` ~ whr2021$`Logged GDP per capita` +
##     whr2021$`Freedom to make life choices`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37656 -0.40059  0.01935  0.45091  1.01389
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.59081     0.43280  -5.986 1.59e-08 ***
## whr2021$`Logged GDP per capita`  0.60079     0.04574  13.135 < 2e-16 ***
## whr2021$`Freedom to make life choices` 3.10374     0.46758   6.638 5.83e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```



```
##
## Residual standard error: 0.5813 on 146 degrees of freedom
## Multiple R-squared: 0.711, Adjusted R-squared: 0.707
## F-statistic: 179.6 on 2 and 146 DF, p-value: < 2.2e-16
```

Na temelju malog povećanja vrijednosti R-squared kod regresije s varijablama GDP per capita i Wealth Gini u odnosu na vrijednost R-squared kod regresije samo s varijablom GDP per capita vidimo da su varijable jako korelirane, dok su (na temelju istog zaključka) GDP per capita i Freedom to make choices manje korelirane.

```
fitAll= lm(`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)
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.