SAP projekt

Case study *World Happiness Report*

David Kerman, Lorena Lazar, Marta Knežević, Vinko Sabolčec

14.1.2022

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
```

```
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"
```

```
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"
```

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

```
## Loading required package: maps
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

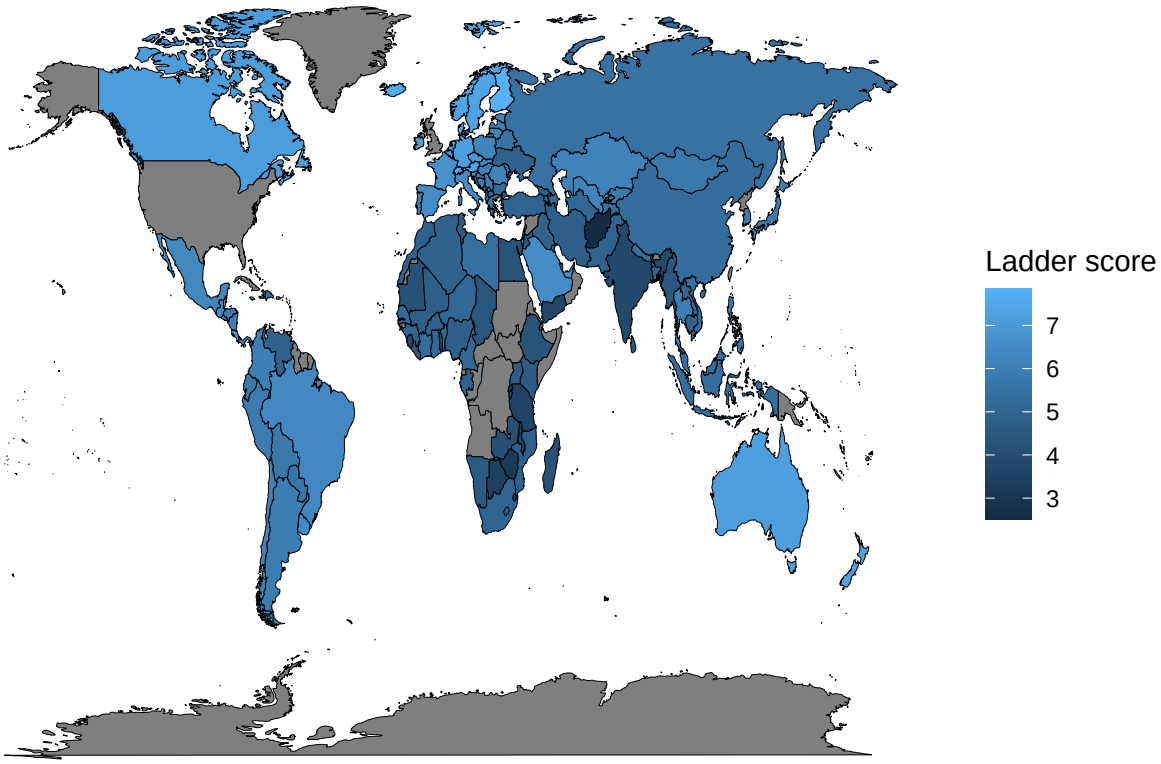
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

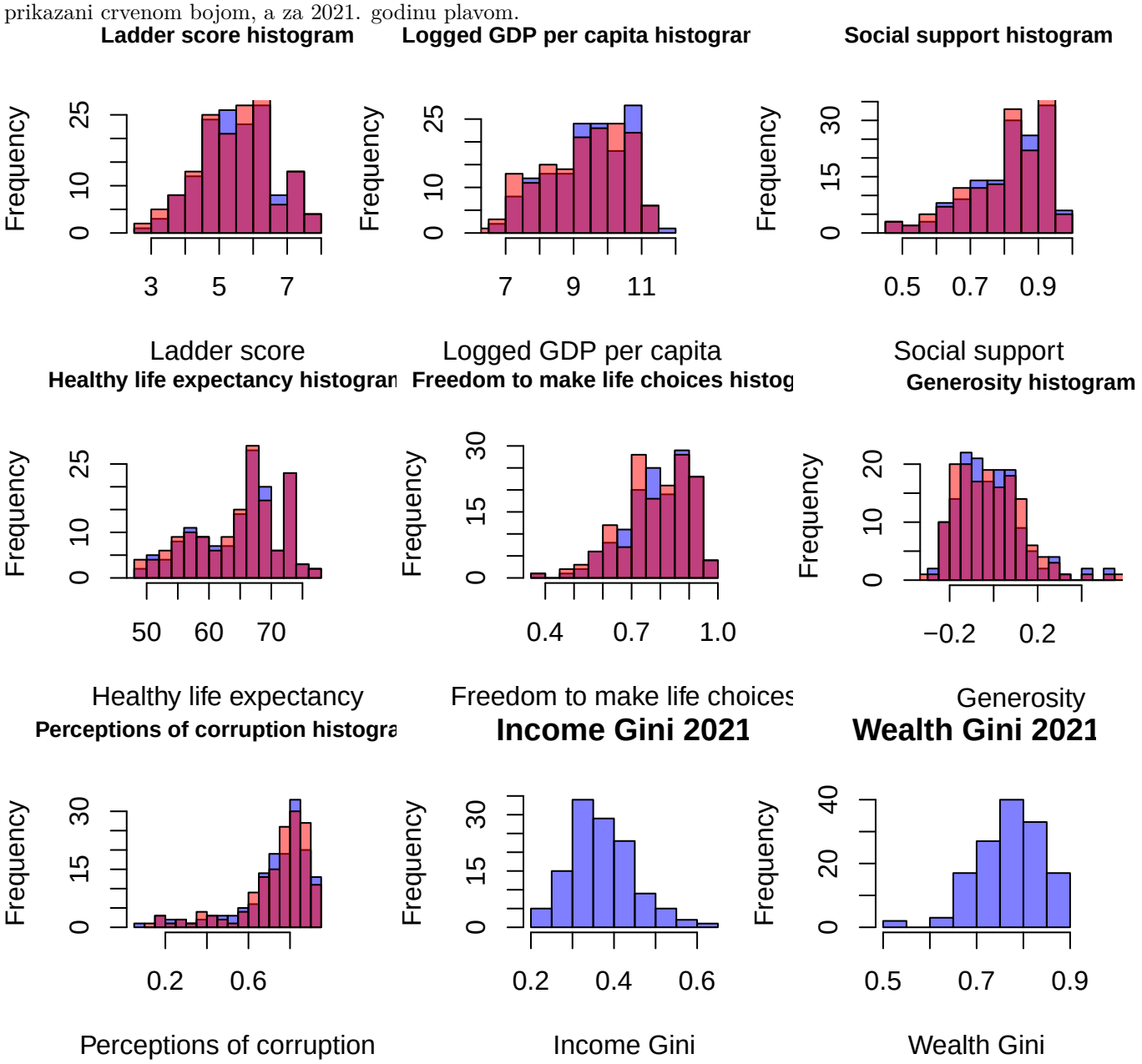
```
data2021 = whr2021[c("Country name", "Ladder score")]
names(data2021)[names(data2021) == "Country name"] = "region"

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
```



Deskriptivna statistika

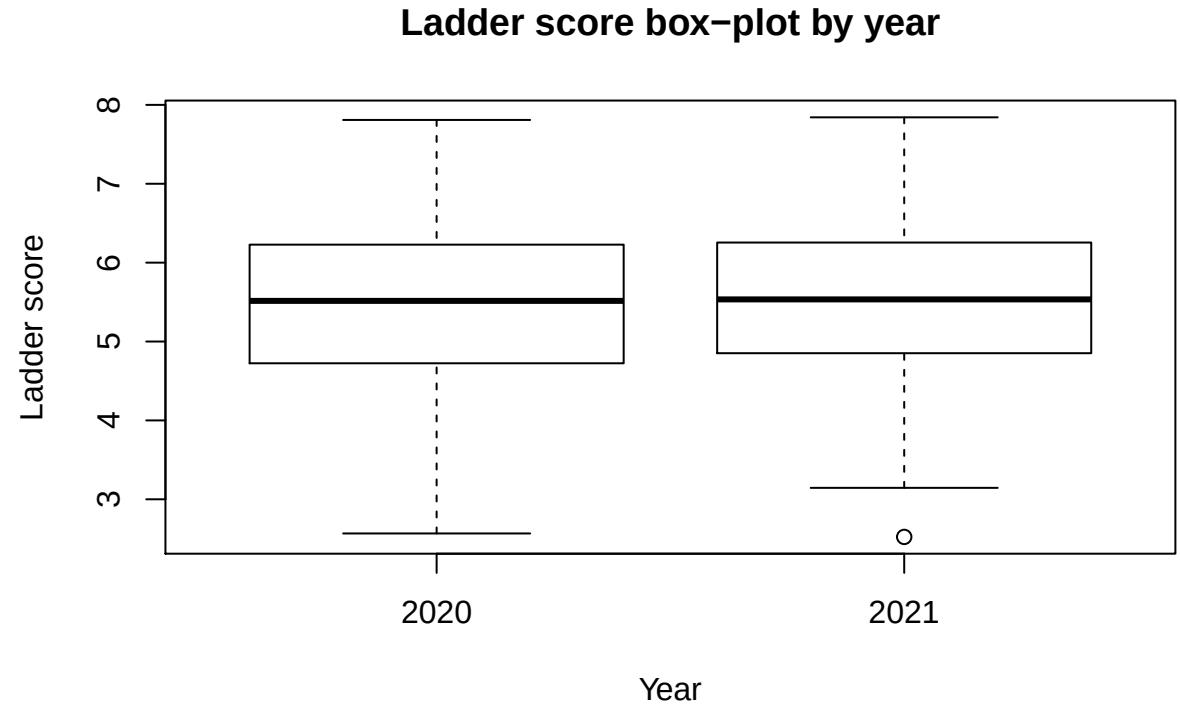
Prikažimo sada histograme usporedbe varijabli za različite godine. Podaci za 2020. godinu na grafovima su prikazani crvenom bojom, a za 2021. godinu plavom.

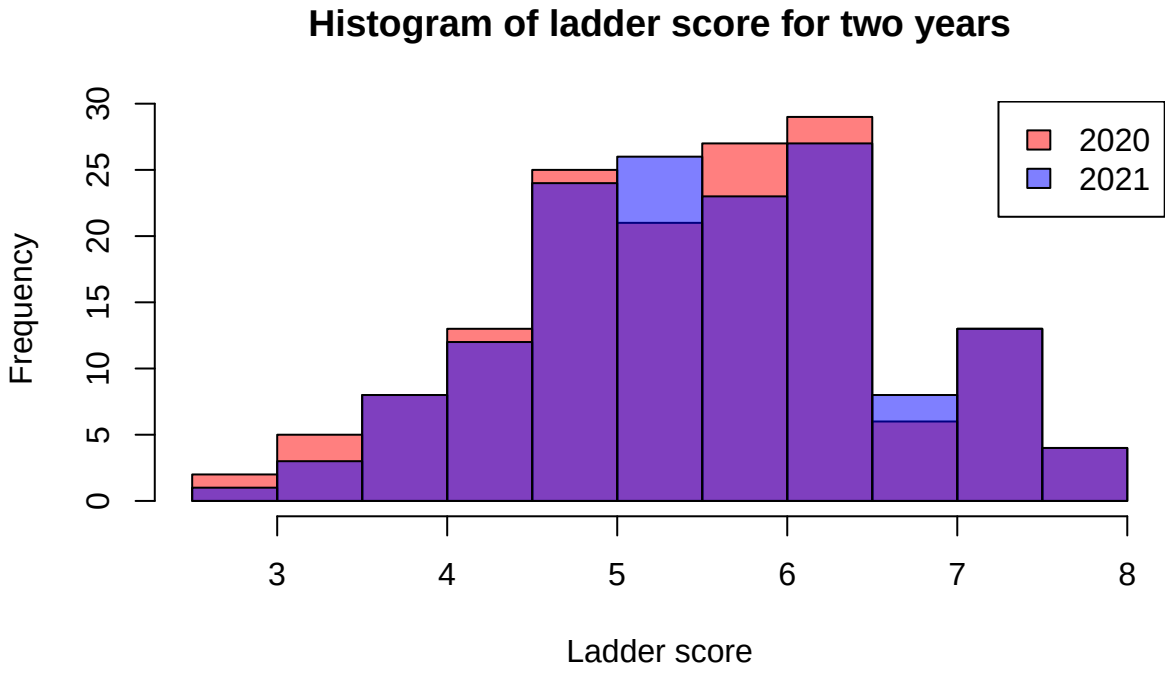


Iz dobivenih histograma vidljivo je da postoje promjene u varijablama za različite godine, no raspodjela podataka je veoma slična za obje godine. Također se može naslutiti da većina podataka nije normalno distribuirana.

Usporedba razina sreće u 2020. i 2021. godini.

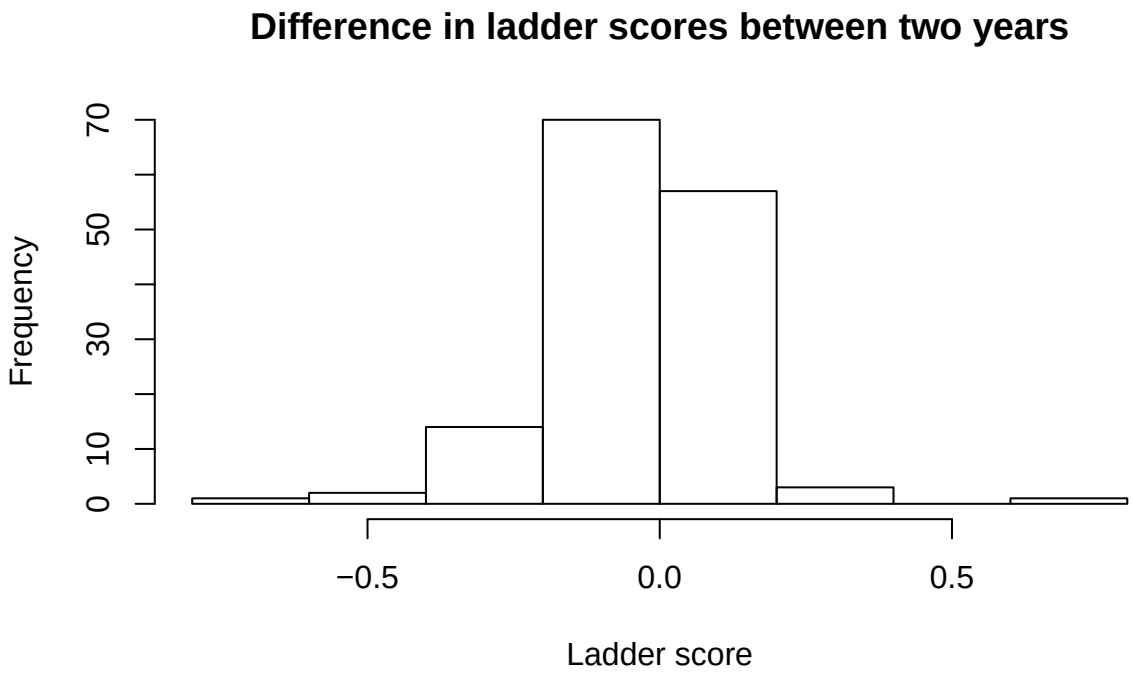
Pravokutni dijagram Ladder score-ova za 2020. i 2021. godinu.



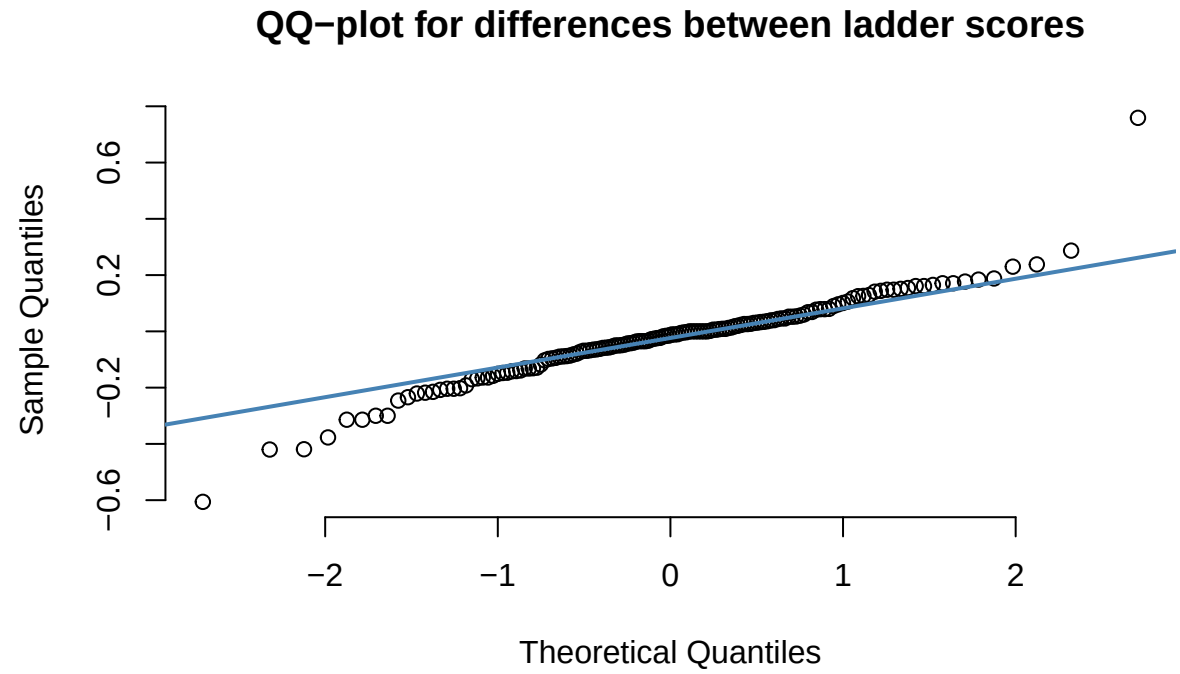


Spojimo podatke iz dvije godine te na histogramu prikažimo razlike razina sreće za dvije godine.

```
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')
```




```
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.

```
require(nortest)
```

```
## Loading required package: nortest
```

```
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 uparenim 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")
```

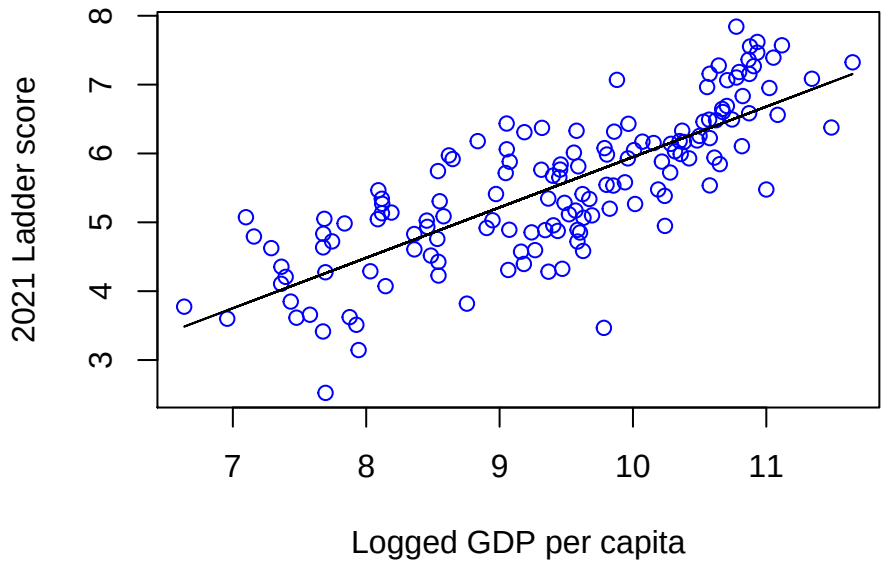
```
##
##  Paired t-test
##
## data:  mergedData$`Ladder score.20` and mergedData$`Ladder score.21`
## t = -2.0749, df = 147, p-value = 0.01987
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.005247129
## sample estimates:
## mean of the differences
##             -0.02594595
```

Jako mala p-vrijednost nam ukazuje da postoji statistički značajna razlika u "Ladder score-u" u dvije godine.Na razini značajnosti od 5% možemo odbaciti hipotezu H_0. Postoje statistički značajne razlike u sreći država u dvije godine tj. države u 2021. godini su sretnije nego u 2020.

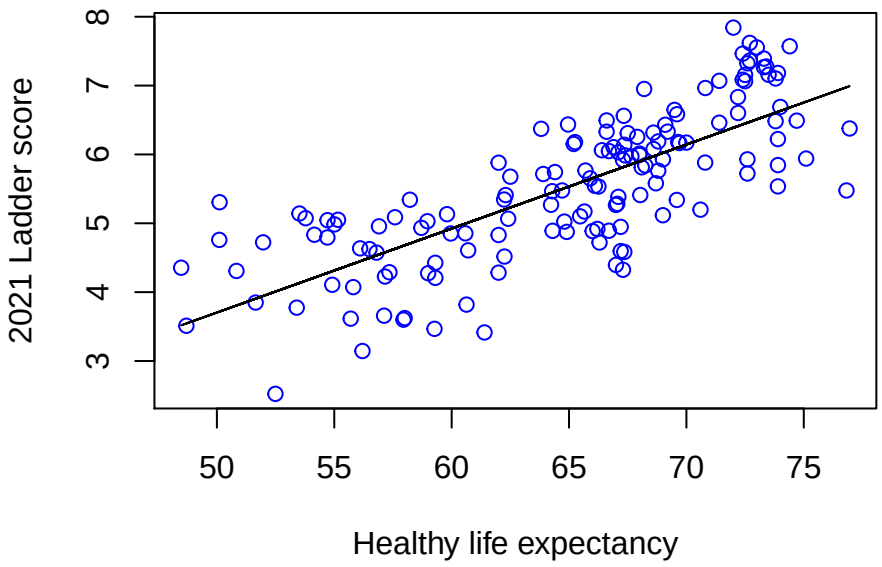
Ovisnost razine sreće o drugim varijablama u 2021. godini

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

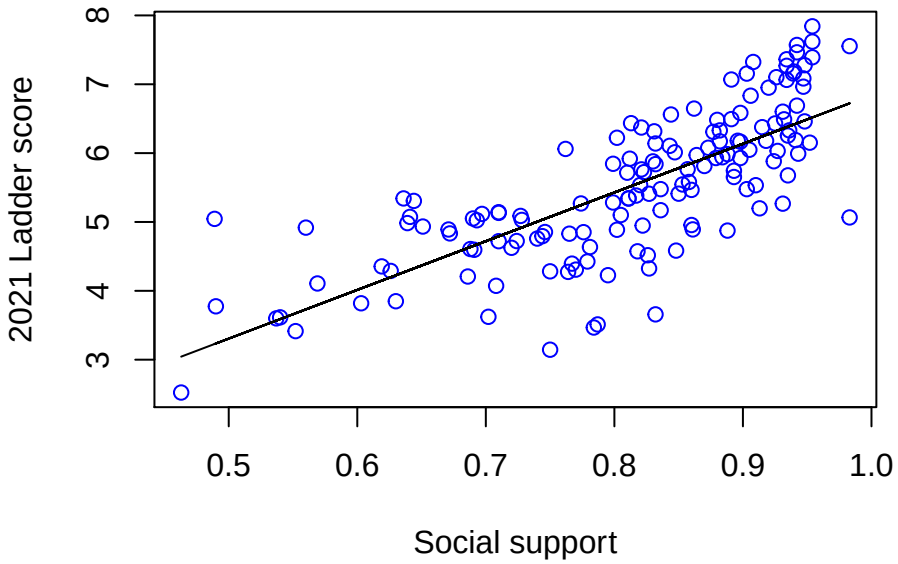
```
fitGDP = lm(whr2021$`Ladder score` ~ whr2021$`Logged GDP per capita`)
fitHealth = lm(whr2021$`Ladder score` ~ whr2021$`Healthy life expectancy`)
fitSocialSupport = lm(whr2021$`Ladder score` ~ whr2021$`Social support`)
fitFreedom = lm(whr2021$`Ladder score` ~ whr2021$`Freedom to make life choices`)
fitGenerosity = lm(whr2021$`Ladder score` ~ whr2021$Generosity)
fitCorruption = lm(whr2021$`Ladder score` ~ whr2021$`Perceptions of corruption`)
fitIncomeGini = lm(whr2021$`Ladder score` ~ whr2021$`Income Gini`)
fitWealthGini = lm(whr2021$`Ladder score` ~ whr2021$`Wealth Gini`)

plot(whr2021$`Logged GDP per capita`, whr2021$`Ladder score`,
     col="blue",
     xlab='Logged GDP per capita',
     ylab='2021 Ladder score')
lines(whr2021$`Logged GDP per capita`, fitGDP$fitted.values)
```




```
plot(whr2021$`Healthy life expectancy`, whr2021$`Ladder score`,
     col="blue",
     xlab='Healthy life expectancy',
     ylab='2021 Ladder score')
lines(whr2021$`Healthy life expectancy`, fitHealth$fitted.values)
```

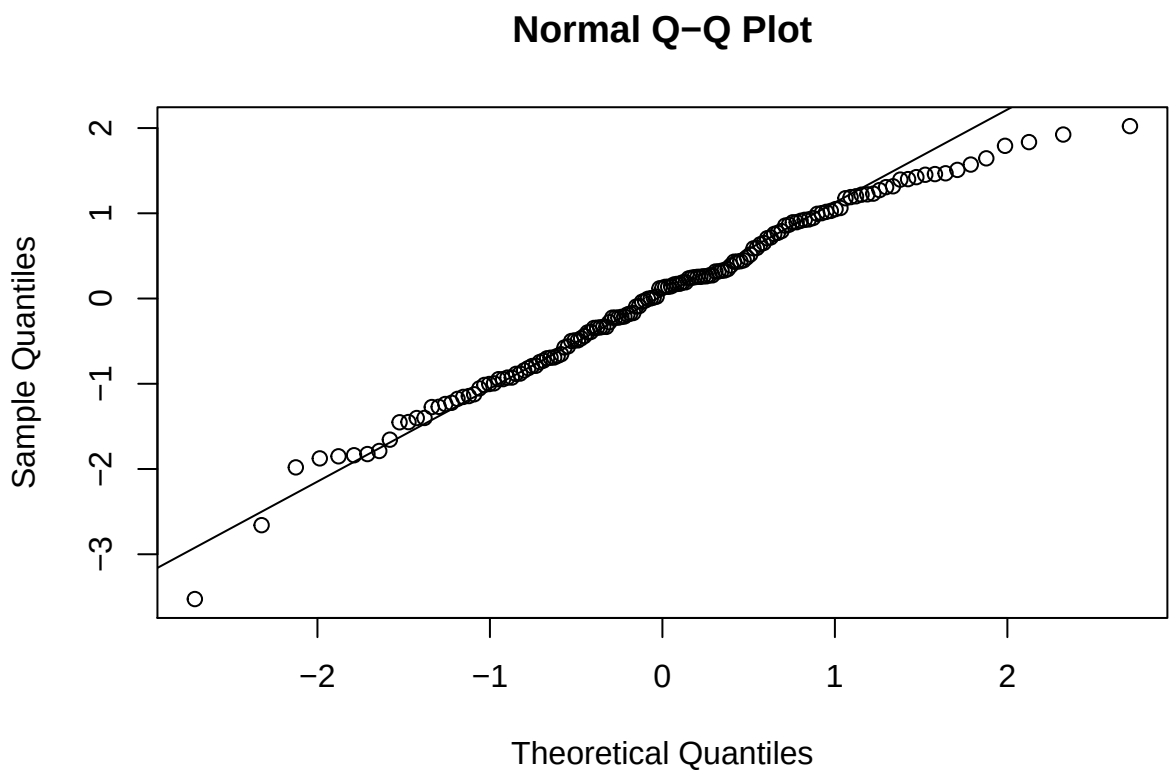



```
plot(whr2021$`Social support`, whr2021$`Ladder score`,
     col="blue",
     xlab='Social support',
     ylab='2021 Ladder score')
lines(whr2021$`Social support`, fitSocialSupport$fitted.values)
```



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. Normalnost reziduala ćemo provjeriti grafički pomoću kvantil-kvantil plota te statistički pomoću Lillieforsove inačice KS testa normalnosti.

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




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

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

Iz kvantil-kvantil 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".

```
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.

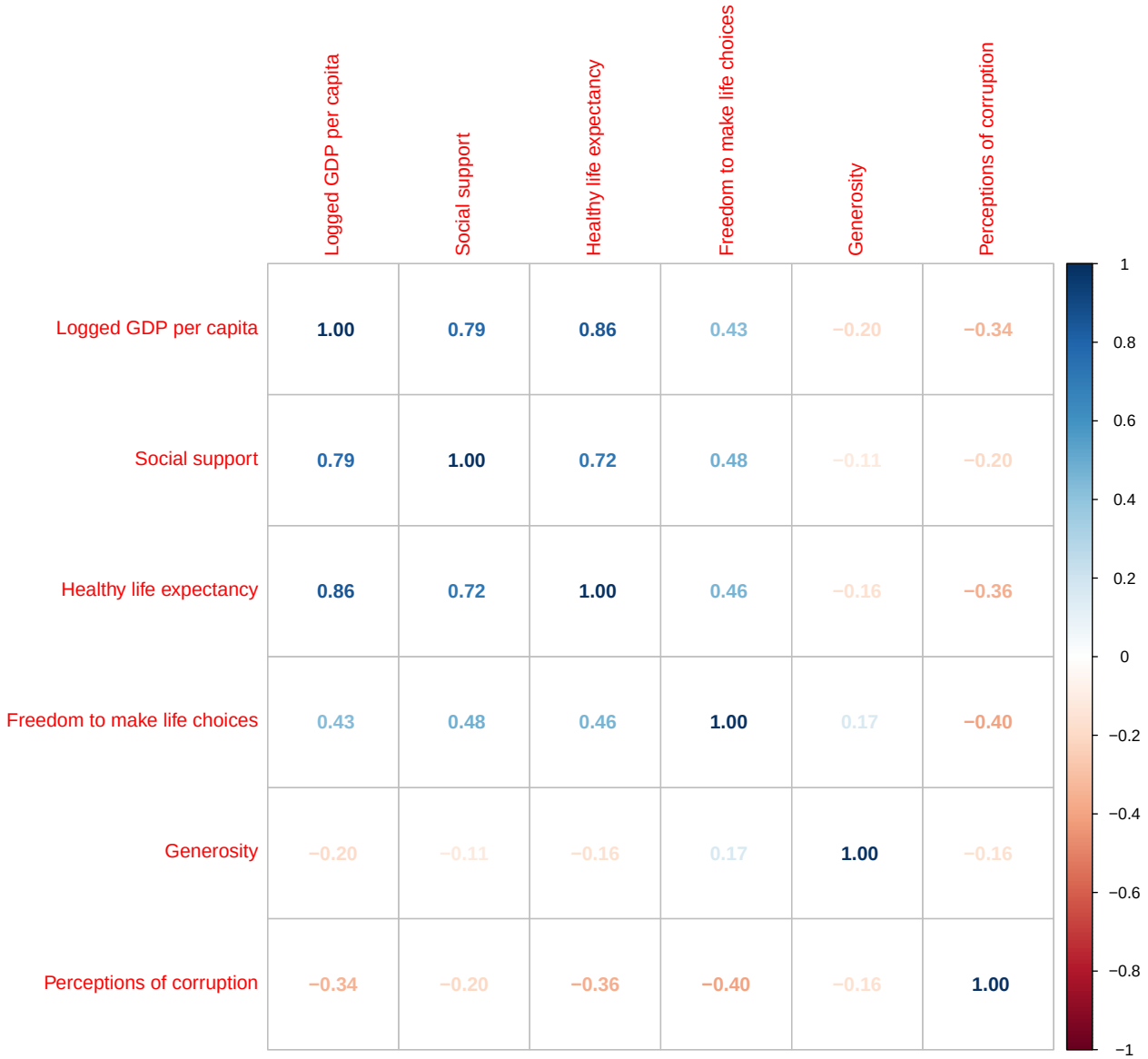
Analogno računamo za preostale varijable. Pogledajmo koliko iznose R-kvadrat i F-statistika za preostale jednostavne modele linearne regresije.

	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 kod jednostruke regresije 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.Ne provjeravamo korelaciju za Income Gini i Wealth Gini jer nedostaju podaci za pojedine države.

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

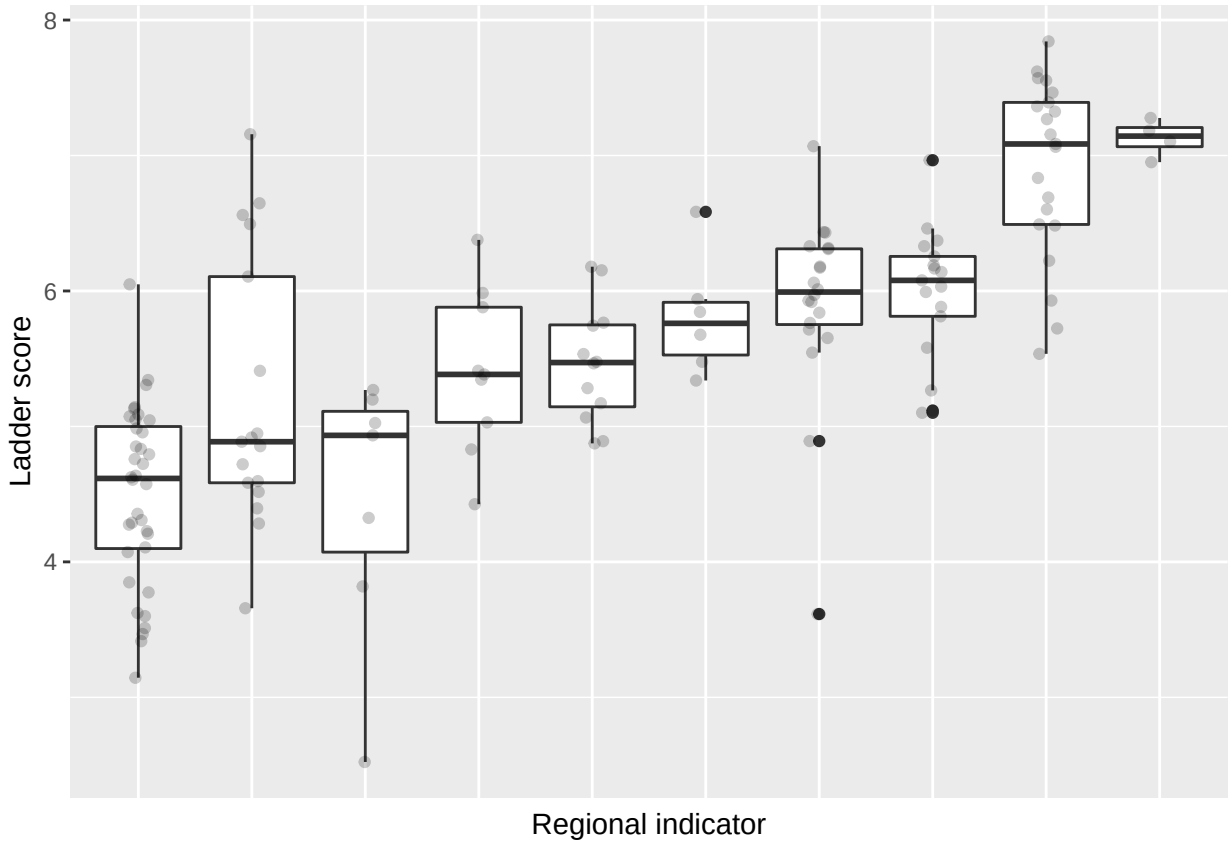


Varijabla "Logged GDP per capita" koja je u jednostavnoj linearnoj regresiji bila najznačajnija je jako korelirana s druge dvije najznačajnije ("Social support" i "Healthy life expectancy"). Zbog toga pri izgradnji modela višestruke linearne regresije možda nije potrebno koristiti sve 3 navedene varijable. Pokušajmo izgraditi model tako da R-kvadrat i F-statistika budu najveći.

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.

```
whr2021 %>%
  ggplot(aes(x=fct_reorder(`Regional indicator`,`Ladder score`),y=`Ladder score`)) +
  geom_boxplot() +
  geom_jitter(width=0.1,alpha=0.2) +
  xlab("Regional indicator") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
```



Iz priloženog boxplota vidimo neki linearan trend, stoga ovu varijablu možemo iskoristiti u daljnjoj analizi.

```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 "Logged GDP per capita", "Social support", "Healthy life expectancy", "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.

Referencirajući se na tablicu koju smo dobili jednostavnom lin. regresijom, probat ćemo iskoristiti jedan ili dva od tri najznačajnija ("Logged GDP per capita", "Social support", "Healthy life expectancy"), te "Freedom to make life choices" i "Perception of corruption" koje koristimo u svim.

```
fitm1= lm(`Ladder score` ~ `Social support`+`Freedom to make life choices` +
            `Perceptions of corruption`, data = whr2021)
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 corruption`, data = whr2021)
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 corruption`, data = whr2021)
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 choices` + `Perceptions of corruption`, data = whr2021)
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` + `Perceptions of corruption`, data = whr2021)
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` + `Perceptions of corruption`, data = whr2021)
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`+`Freedom to make life choices` +
            `Perceptions of corruption`, data = whr2021)
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.

Nadalje, probajmo smanjiti broj parametara u modelu linearne regresije. Pokušat ćemo kombinirati zadnja dva parametra iz modela fit1.

```
fit1= lm(`Ladder score` ~ `Logged GDP per capita` + `Social support`+
           `Freedom to make life choices` + `Perceptions of corruption`, data = whr2021)
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` , data = whr2021)
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 = whr2021)
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".

Pokušajmo sada u model uvrstiti varijablu "Regional indicator" koja je kategorijska.

```
fit_reg = lm(`Ladder score` ~ `Logged GDP per capita` + `Social support` +
               `Freedom to make life choices`+ `Regional indicator`, data=whr2021)
summary(fit_reg)
```

```
##
## Call:
## lm(formula = `Ladder score` ~ `Logged GDP per capita` + `Social support` +
##     `Freedom to make life choices` + `Regional indicator`, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93817 -0.23999  0.04016  0.31490  1.19057
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                      -0.65386    0.65883
## `Logged GDP per capita`                           0.27793    0.07616
## `Social support`                                  1.91892    0.64414
## `Freedom to make life choices`                    2.66749    0.46406
## `Regional indicator`Commonwealth of Independent States -0.34556    0.19514
## `Regional indicator`East Asia                     -0.10515    0.23957
## `Regional indicator`Latin America and Caribbean    0.12809    0.17338
## `Regional indicator`Middle East and North Africa  -0.25467    0.17849
## `Regional indicator`North America and ANZ          0.58936    0.28252
## `Regional indicator`South Asia                    -0.70788    0.24762
## `Regional indicator`Southeast Asia                -0.55599    0.22101
## `Regional indicator`Sub-Saharan Africa            -0.36212    0.19556
## `Regional indicator`Western Europe                 0.51539    0.16999
##                                                  t value Pr(>|t|)
## (Intercept)                                       -0.992 0.322735
## `Logged GDP per capita`                            3.649 0.000374 ***
## `Social support`                                   2.979 0.003425 **
## `Freedom to make life choices`                     5.748 5.67e-08 ***
## `Regional indicator`Commonwealth of Independent States  -1.771 0.078824 .
## `Regional indicator`East Asia                     -0.439 0.661428
## `Regional indicator`Latin America and Caribbean    0.739 0.461298
## `Regional indicator`Middle East and North Africa  -1.427 0.155915
## `Regional indicator`North America and ANZ          2.086 0.038841 *
## `Regional indicator`South Asia                    -2.859 0.004924 **
## `Regional indicator`Southeast Asia                -2.516 0.013043 *
## `Regional indicator`Sub-Saharan Africa            -1.852 0.066236 .
## `Regional indicator`Western Europe                 3.032 0.002911 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4988 on 136 degrees of freedom
## Multiple R-squared:  0.8017, Adjusted R-squared:  0.7842
## F-statistic: 45.83 on 12 and 136 DF,  p-value: < 2.2e-16
```

Vidimo da R sam stvara tzv. "dummy" varijable za kategorijsku varijablu "Regional indicator". Neke novo stvorene "dummy" varijable imaju veliku p-vrijednost što ukazuje da nisu značajne, no neke imaju malu p-vrijednost te se koeficijent determinacije povećao u usporedbi s ostalim modelima, ali se F-statistika smanjila.

Postoje li razlike u iskazanoj sreći među različitim regijama?

Zanima nas ima li značajnih razlika u sredinama razina sreće u različitim regijama. Prvo ćemo usporediti razine sreće u Srednjoj i Istočnoj Europi pomoću t-testa, a zatim razine sreće u tri svjetske regije pomoću ANOVA testa.

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')
```

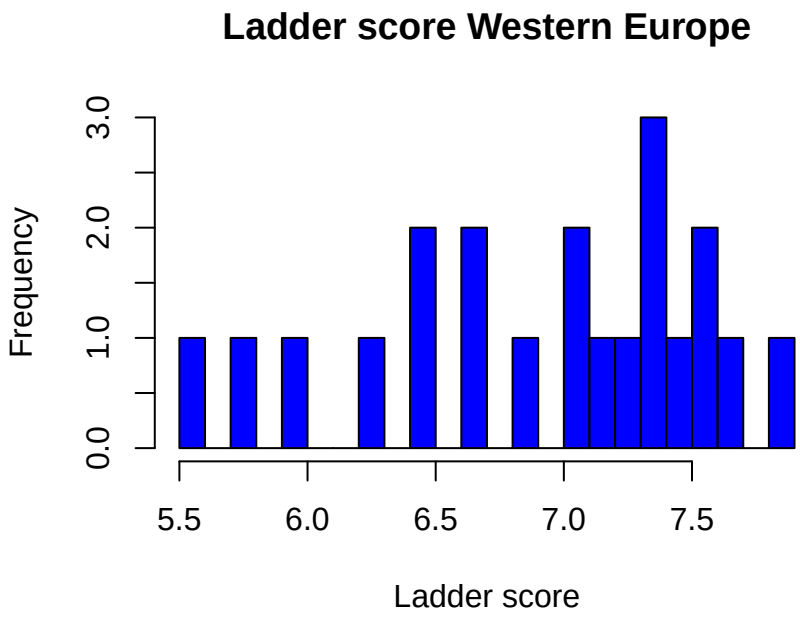
```
## Prosječan Ladder score zemalja iz Zapadne Europe  6.914905
```

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

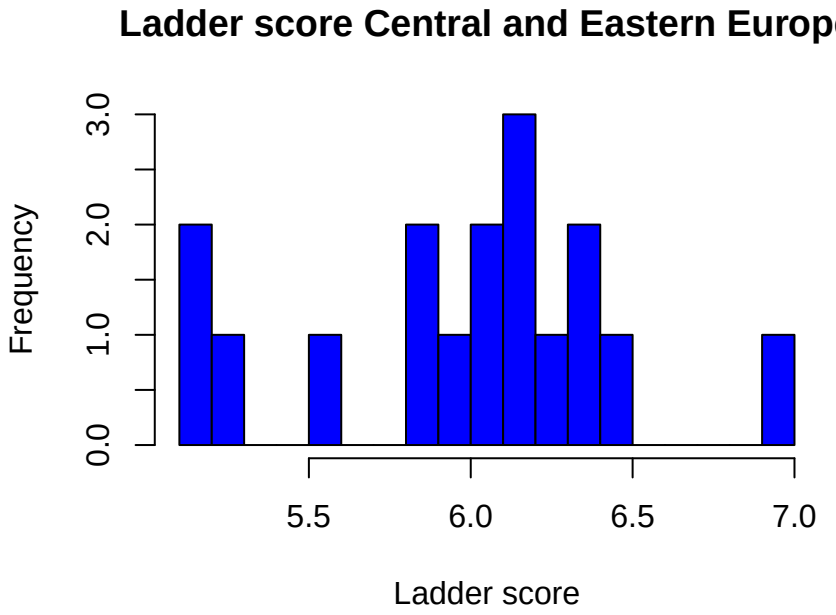
```
## Prosječan Ladder score zemalja iz Srednje i Istočne Europe 5.984765
```

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
         )
```

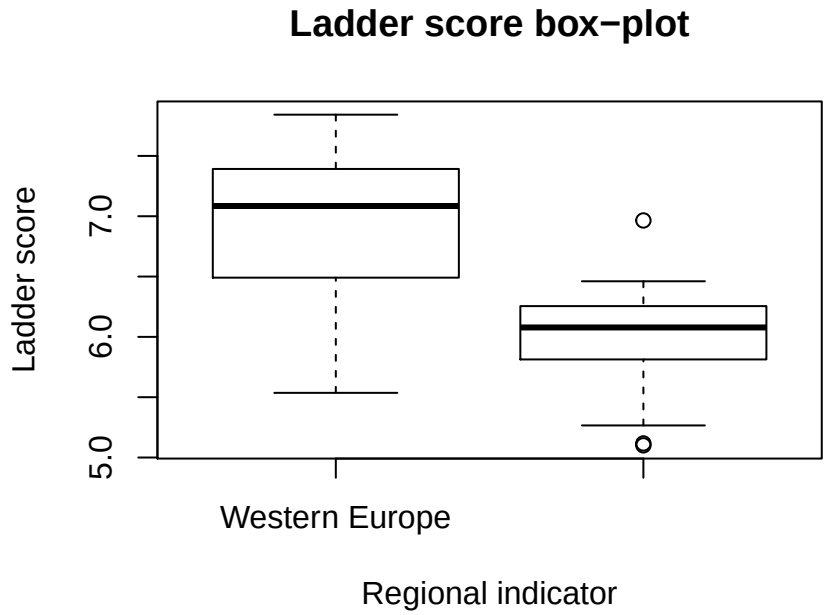


```
h = hist(central_eastern_europe$`Ladder score`,
         main="Ladder score Central and Eastern Europe",
         xlab="Ladder score",
         ylab='Frequency',
         col="blue",
         breaks = 20
         )
```



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 Europe"))
```

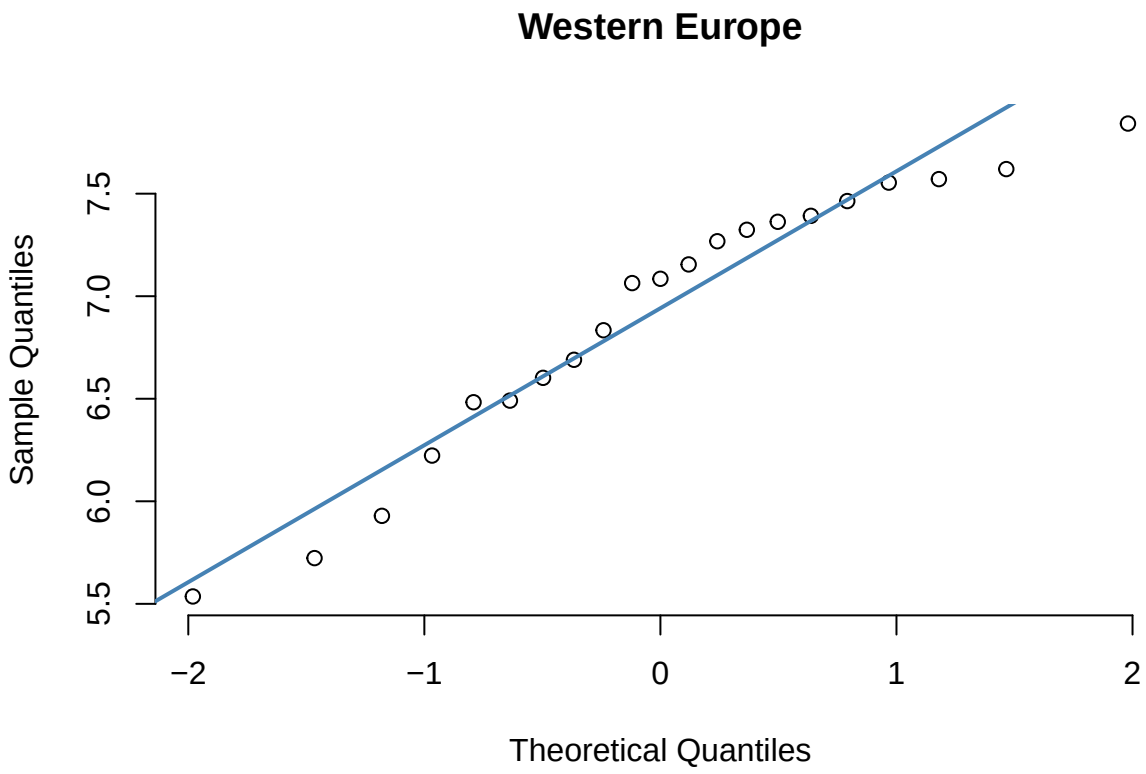


Postoje indikacije da bi ljudi iz zemalja Zapadne Europe trebali biti sretniji od ljudi iz zemalja Srednje i Istočne Europe.

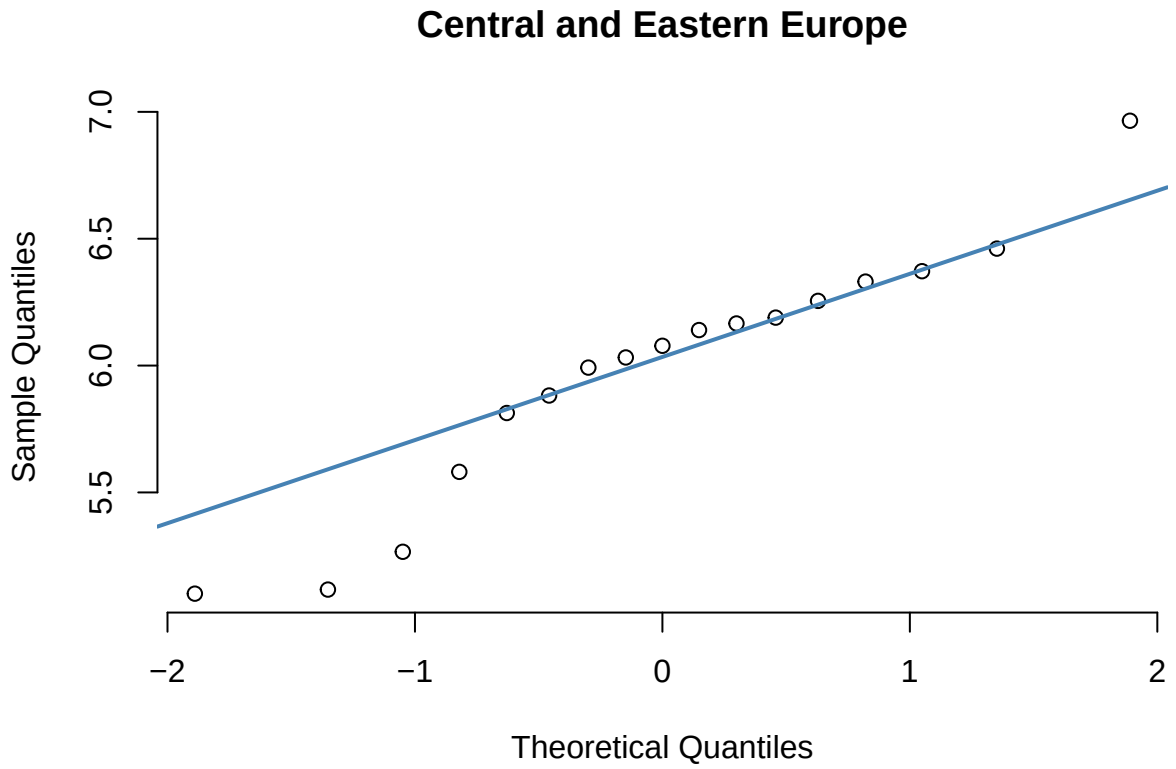
Postavimo sljedeće hipoteze:
H_0: Ladder score je jednak za Zapadnu i Srednju i Istočnu Europu
H_1: Ladder score je veći u Zapadnoj Europi od onog u Srednjoj i Istočnoj Europi

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 KS testa normalnosti (umjesto samog KS testa) 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.

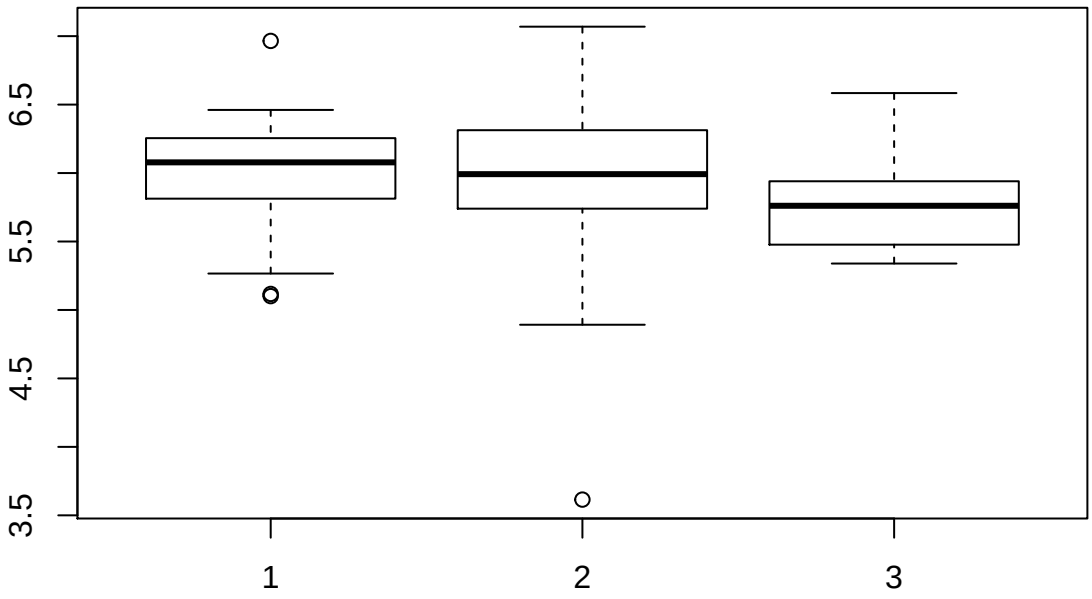
```
t.test(western_europe$`Ladder score`, central_eastern_europe$`Ladder score`,
       alt = "greater", var.equal = TRUE)
```

```
##
##  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.

Usporedba sredina razina sreće triju regija

```
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$`Regional indicator` == "Latin America and Caribbean"
                  | whr2021$`Regional indicator` == "East Asia", ]

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
```

```
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:

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

```
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
```

```
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.

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

```
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