World Happiness Report 2023

Mia Gmiza, Gabrijela Perković, Matija Roginić, Erika Tomakić

2023-12-06

Uvod

S obzirom na turbolentna događanja u svijetu i razne faktore koji se isprepleću u našoj svakodnevici, postavlja se pitanje koliko je sretan prosječan čovjek i što to utječe na sreću. Mreža rješenja za održivi razvoj Ujedinjenih naroda napravila je istraživanje pod imenom World Happiness Report u kojem mjeri razinu sreće, stopu konzumacije alkohola, stopu kriminala itd. po svim zemljama svijeta. U ovom smo projektu odlučili proučiti dostupne podatke kako bismo odredili je li svjetska populacija generalno bila sretnija u 2023. nego li u 2022. godini, kako koreliraju razni faktori sa stopom konzumacije alkohola te smo analizirali kvalitetu zdravstvene skrbi po regijama svijeta. Kako bismo donijeli zaključke na temelju dobivenih podataka problemu smo pristupili koristeći razne statističke metode.

Deskriptivna analiza

Učitavanje podataka.

```
opis_var = read.csv("datasets/opis_varijabli.csv")
WHR_22 = read.csv("datasets/WHR_2022.csv")
WHR_22 = head(WHR_22, -1) # preskacem zadnji red jer je "xx"
WHR_23 = read.csv("datasets/WHR_2023.csv")
```

Podatci za 2022. godinu sastoje se od 146 država i dvije varijable. Podatci za 2023. godinu sastoje se od 137 država i 15 varijabli.

```
cat("Varijable za 2022. godinu:\n")
```

```
## Varijable za 2022. godinu:
```

```
names(WHR_22)
```

```
## [1] "Country"         "Happiness.score"
```

```
cat("Varijable za 2023. godinu:\n")
```

```
## Varijable za 2023. godinu:
```

```
names(WHR_23)
```

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

```
## [10] "Alcohol.consumption.Both.Sexes..L.year."
## [11] "Alcohol.consumption.Male..L.year."
## [12] "Alcohol.consumption.Female..L.year."
## [13] "Crime.rate.Crime.Index"
## [14] "Healthcare.Legatum.Prosperity.Index.Health.Score"
## [15] "Gini.Coefficient...World.Bank"
```

```
any(is.na(WHR_22))
```

```
## [1] FALSE
```

```
cat("U podatcima za 2022. godinu nema nedostajućih vrijednosti.\n")
```

```
## U podatcima za 2022. godinu nema nedostajućih vrijednosti.
```

```
any(is.na(WHR_23))
```

```
## [1] TRUE
```

```
cat("U podatcima za 2023. godinu ima nedostajućih vrijednosti.\n")
```

```
## U podatcima za 2023. godinu ima nedostajućih vrijednosti.
```

```
for (col_name in names(WHR_23)) {
  if (sum(is.na(WHR_23[,col_name])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu ',col_name, ': ', sum(is.na(WHR_23[,col_name])),'
  }
}
```

```
## Ukupno nedostajućih vrijednosti za varijablu  Healthy.life.expectancy :  1
## Ukupno nedostajućih vrijednosti za varijablu  Alcohol.consumption.Both.Sexes..L.year. :  6
## Ukupno nedostajućih vrijednosti za varijablu  Alcohol.consumption.Male..L.year. :  6
## Ukupno nedostajućih vrijednosti za varijablu  Alcohol.consumption.Female..L.year. :  6
## Ukupno nedostajućih vrijednosti za varijablu  Crime.rate.Crime.Index :  24
## Ukupno nedostajućih vrijednosti za varijablu  Healthcare.Legatum.Prosperity.Index.Health.Score :  2
## Ukupno nedostajućih vrijednosti za varijablu  Gini.Coefficient...World.Bank :  10
```

Sljedeće što možemo napraviti je izračunati korelaciju između varijabli. Možemo izabrati i ispisati korelaciju između svakog para varijabli, ali takav ispis bi bio nepraktičan, a nije nam ni potreban. Stoga ćemo ispisati samo korelaciju svih varijabli s varijablom koja prikazuju indeks sreće u pojedinoj državi.

```
my_data <- WHR_23[, c(3,4,5,6,7,8,9,10,11,12,13,14,15)]
matrix = round(cor(my_data, use = "complete.obs"),2)
corrs <- matrix[, 1]
names <- colnames(matrix)
var = names[1]
df <- data.frame(Variable = colnames(matrix)[-1], Correlation = corrs[-1])
last12 <- tail(df,12)
cat(sprintf("%s %.2f\n", last12$Variable, last12$Correlation))
```

```
## GDP.per.capita 0.72
##  Social.support 0.80
##  Healthy.life.expectancy 0.71
##  Freedom.to.make.life.choices 0.60
##  Generosity 0.09
##  Perceptions.of.corruption -0.54
##  Alcohol.consumption.Both.Sexes..L.year. 0.54
##  Alcohol.consumption.Male..L.year. 0.51
##  Alcohol.consumption.Female..L.year. 0.60
```

```
##  Crime.rate.Crime.Index -0.38
##  Healthcare.Legatum.Prosperity.Index.Health.Score 0.74
##  Gini.Coefficient...World.Bank -0.32
```

1. Je li razina sreće u publikaciji za 2023. veća ili manja u usporedbi s istraživanjem provedenim godinu ranije?

S obzirom na dostupne podatke zanima nas postaju li ljudi sretniji ili nesretniji. Kako bismo odgovorili na to pitanje usporedit ćemo razine sreće u publikaciji iz 2022. i 2023. godine. Budući da nisu dostupni podaci za sve zemlje u obje godine, možemo uzeti presjek zajedničkih država. To nas ostavlja s podatcima za 133 države.

Fokusirajmo se na globalnu sliku razina sreće, odnosno po svim zemljama

Postavimo nul-hipotezu (H0) koja tvrdi da je razina sreće u 2023. godini manja ili jednaka onoj izmjerenoj u prethodnoj godini. Alternativna hipoteza (H1) tvrdi kako je izmjerena razina sreće u 2023. godini veća nego li ona izmjerena u prethodnoj godini.

Kako bismo testirali točnost nul-hipoteze koristit ćemo t-test. Međutim, prije nego što počnemo s testiranjem, moramo utvrditi da su podaci normalni, odnosno da pripadaju normalnoj razdiobi. Provest ćemo Shapiro-Wilk test pri čemu biramo da vrijedi $\alpha = .05$. Ako izračunata p vrijednost bude veća od zadanog α, podaci pripadaju normalnoj razdiobi.Također možemo pogledati i Q-Q graf podataka kako bismo potvrdili normalnost. U tom grafu razmatramo hoće li naši podaci formirati pravac uz potencijalna manja odstupanja. Ako to bude slučaj, podatke možemo smatrati normalnima.

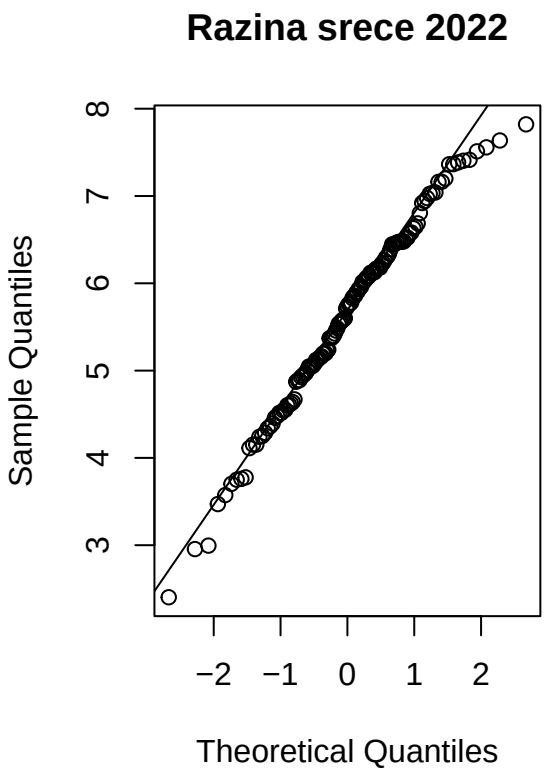
```
shapiro.test(ladder_score_2022)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ladder_score_2022
## W = 0.98742, p-value = 0.2656
```

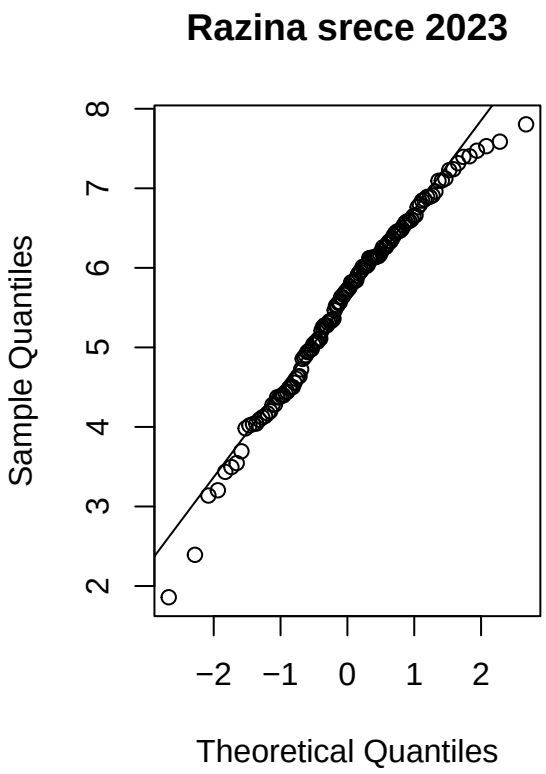
```
#data:  ladder_score_2022
#W = 0.98742, p-value = 0.2656

#Q-Q graf
par(mfrow=c(1,2))
qqnorm(ladder_score_2022, main='Razina sreće 2022')
qqline(ladder_score_2022)

shapiro.test(ladder_score_2023)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ladder_score_2023
## W = 0.98047, p-value = 0.0529
```

```
#data:  ladder_score_2023
#W = 0.98047, p-value = 0.0529

par(mfrow=c(1,2))
```




```
qqnorm(ladder_score_2023, main='Razina sreće 2023')
qqline(ladder_score_2023)
```



Rezultati pokazuju da su podaci iz 2022. godine normalno distribuirano s obzirom na Q-Q graf koji formira pravac te vrijednost p = 0.2656 > α. Slično pokazuju i podaci za 2023. godinu, iako je p vrijednost puno bliža granici normalnosti. Moguće je napraviti transformaciju podataka tako budu normalnije distribuirani, ali nije nužno jer podaci nisu upali u kritično područje.

Napokon možemo provesti t-test nad podacima kako bismo utvrdili točnost početne hipoteze. Neka vrijedi da je razine značajnosti $\alpha = 0.05$. S obzirom da proučavamo iste testne grupe po istim kategorijama, samo u nekom vremenskom razmaku, testne su skupine zavisne, odnosno uparene Stoga biramo upareni t-test .

```
rezultat_t_testa <- t.test(ladder_score_2023, ladder_score_2022, paired = TRUE, alternative = "greater")
print(rezultat_t_testa)
```

```
##
##  Paired t-test
##
## data:  ladder_score_2023 and ladder_score_2022
## t = -2.4778, df = 132, p-value = 0.9928
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  -0.08561081         Inf
## sample estimates:
## mean difference
##     -0.05130901
```

```
#Paired t-test

#data:  ladder_score_2023 and ladder_score_2022
#t = -2.4778, df = 132, p-value = 0.9928
```

```
#alternative hypothesis: true mean difference is greater than 0
#95 percent confidence interval:
# -0.08561081         Inf
#sample estimates:
#mean difference
    -0.05130901
```

```
## [1] -0.05130901
```

Rezultati testa ukazuju da ne odbacujemo nultu hipotezu (H0) na temelju p-vrijednosti koja je iznimno visoka. P-vrijednost izražava vjerovatnost dobivanja rezultata sličnih ili ekstremnijih od onih koje biste dobili ako je nulta hipoteza istinita. U našem slučaju p-vrijednost iznosi 0.9928 te je veća od standardne razine značajnosti od 0.05. Dakle, nemamo dovoljno statističkih dokaza da podržimo tvrdnju da je razina sreće u 2023. godini veća od razine sreće u 2022. godini.

2. Možemo li temeljem drugih dostupnih varijabli predvidjeti konzumaciju alkohola po zemljama?

S obzirom na to da pitanje traži predikciju konzumacije alkohola po zemljama, alat koji ćemo koristiti za to je linearna regresija. Nezavisne varijable će u tom slučaju biti sve one koje nisu vezane za alkohol, a varijabla koju predviđamo će biti konzumacija alkohola za oba spola. Uz to, iz skupa nezavisnih varijabli nećemo koristiti varijable koje pokazuju konzumaciju alkohola posebno za muškarce i žene. Iz toga slijedi da ćemo imati 12 nezavisnih varijabli i jednu nezavisnu.

```
my_data <- WHR_23[, c(3,4,5,6,7,8,9,10,13,14,15)]
head(my_data$Gini.Coefficient...World.Bank, 10)
```

```
##  [1] 27.3 28.2 26.1 39.0 28.1 30.0 27.6 33.1 35.4   NA
```

Problem na koji nalazimo su nedostajuće vrijednosti u našem skupu podataka. Na primjer, za značajku Gini Coefficient na desetoj poziciji u tablici imamo NA. Ono što ćemo napraviti je zamjena nedostajućih vrijednosti među nezavisnim varijablama prosječnom vrijednošću za tu regiju. Naravno, to nije jedini način na koji se možemo nositi s nedostajućim vrijednostima. Na primjer, možemo u potpunosti izbaciti te zapise, ali u tom slučaju gubimo previše podataka pa nam to nije opcija. Sljedeća opcija bi nam bila zamjena s ukupnim prosjekom, no procijenili smo da je zamjena prosjekom regije ipak točnija. Dakle, sad ćemo napraviti zamjenu nedostajućih podataka.

```
if(!require(dplyr)) install.packages("dplyr",repos = "http://cran.us.r-project.org")
```

```
## 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
```

```
library(dplyr)
my_data <- WHR_23 %>%
  group_by(Regional.indicator) %>%
  mutate(
    Ladder.score = ifelse(is.na(Ladder.score),
```

```
                  mean(Ladder.score, na.rm = TRUE), Ladder.score),
    GDP.per.capita = ifelse(is.na(GDP.per.capita),
                  mean(GDP.per.capita, na.rm = TRUE), GDP.per.capita),
    Social.support = ifelse(is.na(Social.support),
                  mean(Social.support, na.rm = TRUE), Social.support),
    Healthy.life.expectancy = ifelse(is.na(Healthy.life.expectancy), mean(Healthy.life.expectancy, na.rm
    Freedom.to.make.life.choices = ifelse(is.na(Freedom.to.make.life.choices), mean(Freedom.to.make.life
    Generosity = ifelse(is.na(Generosity), mean(Generosity, na.rm = TRUE), Generosity),
    Perceptions.of.corruption = ifelse(is.na(Perceptions.of.corruption), mean(Perceptions.of.corruption
    Crime.rate.Crime.Index = ifelse(is.na(Crime.rate.Crime.Index), mean(Crime.rate.Crime.Index, na.rm =
    Healthcare.Legatum.Prosperity.Index.Health.Score = ifelse(is.na(Healthcare.Legatum.Prosperity.Index
    Gini.Coefficient...World.Bank = ifelse(is.na(Gini.Coefficient...World.Bank), mean(Gini.Coefficient.
  ) %>%
  ungroup()

my_data <- my_data[, c(3,4,5,6,7,8,9,10,13,14,15)]
head(my_data$Gini.Coefficient...World.Bank, 10)
```

```
##  [1] 27.30000 28.20000 26.10000 39.00000 28.10000 30.00000 27.60000 33.10000
##  [9] 35.40000 36.36667
```

Sad vidimo da na desetom mjestu za značajku Gini Coefficient više nije NA, nego je ta nedostajuća vrijednost zamijenjena s prosjekom za regiju.

Nakon što smo napravili zamjenu nedostajućih vrijednosti u nezavisnim varijablama, moramo riješiti taj problem i kod zavisne varijable što je u našem slučaju konzumacija alkohola za oba spola. Iz ranijeg ispisa (kod deskriptivne statistike) vidimo da kod te značajke imamo 6 nedostajućih zapisa. S obzirom na to da je cilj ovdje predvidjeti vrijednost konzumacije alkohola, nema smisla mijenjati te nedostajuće vrijednosti s prosjekom. Ono što ćemo ovdje napraviti je ignorirati tih 6 zapisa (6 država) te provesti linearnu regresiju na preostalim zapisima. To znači da ćemo regresiju raditi na temelju podataka iz 131 države umjesto početnih 137.

```
new <- my_data[!is.na(my_data$Alcohol.consumption.Both.Sexes..L.year.), ]
size <- dim(new)
cat("Broj redaka:", size[1], "\n")
```

```
## Broj redaka: 131
```

Sljedeći korak je provođenje linearne regresije.

```
if(!require(coefplot)) install.packages("coefplot",repos = "http://cran.us.r-project.org")
```

```
## Loading required package: coefplot
```

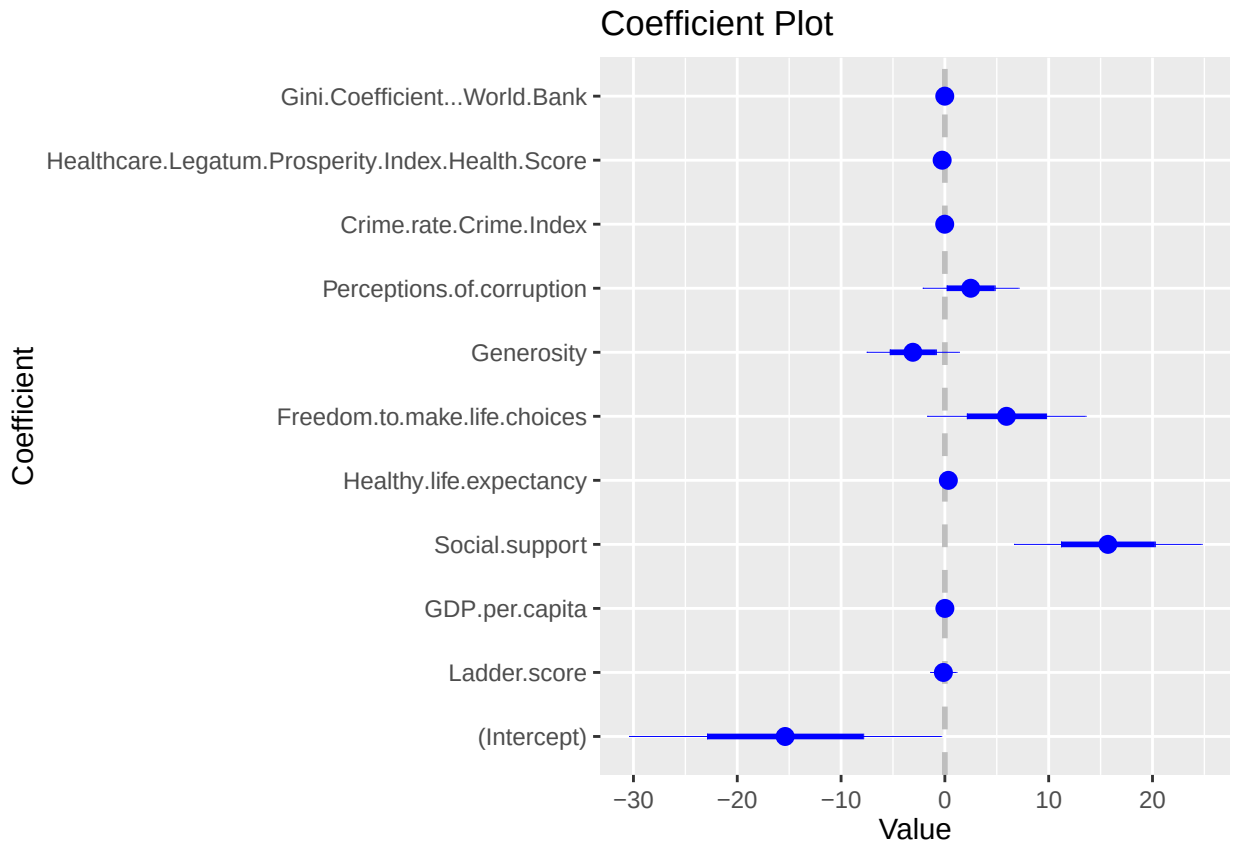
```
## Loading required package: ggplot2
```

```
library(coefplot)
model <- lm(Alcohol.consumption.Both.Sexes..L.year. ~ Ladder.score + GDP.per.capita + Social.support + H
            Perceptions.of.corruption + Crime.rate.Crime.Index +
            Healthcare.Legatum.Prosperity.Index.Health.Score +
            Gini.Coefficient...World.Bank,
          data = new)
summary(model)
```

```
##
## Call:
## lm(formula = Alcohol.consumption.Both.Sexes..L.year. ~ Ladder.score +
##     GDP.per.capita + Social.support + Healthy.life.expectancy +
```

```
##     Freedom.to.make.life.choices + Generosity + Perceptions.of.corruption +
##     Crime.rate.Crime.Index + Healthcare.Legatum.Prosperity.Index.Health.Score +
##     Gini.Coefficient...World.Bank, data = new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4664 -2.2199  0.3454  1.9004  8.9395
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                    -1.539e+01  7.518e+00  -2.047
## Ladder.score                                   -1.433e-01  6.304e-01  -0.227
## GDP.per.capita                                  5.767e-05  2.475e-05   2.330
## Social.support                                  1.571e+01  4.519e+00   3.475
## Healthy.life.expectancy                         3.353e-01  1.823e-01   1.839
## Freedom.to.make.life.choices                    5.935e+00  3.813e+00   1.556
## Generosity                                     -3.085e+00  2.219e+00  -1.390
## Perceptions.of.corruption                       2.490e+00  2.308e+00   1.079
## Crime.rate.Crime.Index                         -1.559e-02  3.165e-02  -0.493
## Healthcare.Legatum.Prosperity.Index.Health.Score -2.632e-01  1.044e-01  -2.521
## Gini.Coefficient...World.Bank                  -1.015e-02  5.157e-02  -0.197
##                                                Pr(>|t|)
## (Intercept)                                    0.042876 *
## Ladder.score                                   0.820540
## GDP.per.capita                                 0.021450 *
## Social.support                                 0.000711 ***
## Healthy.life.expectancy                        0.068374 .
## Freedom.to.make.life.choices                   0.122234
## Generosity                                     0.167111
## Perceptions.of.corruption                      0.282746
## Crime.rate.Crime.Index                         0.623261
## Healthcare.Legatum.Prosperity.Index.Health.Score 0.013025 *
## Gini.Coefficient...World.Bank                  0.844340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.239 on 120 degrees of freedom
## Multiple R-squared:  0.4304, Adjusted R-squared:  0.383
## F-statistic: 9.069 on 10 and 120 DF,  p-value: 5.119e-11
```

```
coefplot(model)
```



Sažetak linearnog modela kojeg nam je kao izlaz dao RStudio nam otkriva pojedinosti o koeficijentima uz regresore, statističkoj važnosti pojedinog regresora, te ostalim mjerama po kojima možemo vidjeti koliko se dobro model prilagođava podacima.

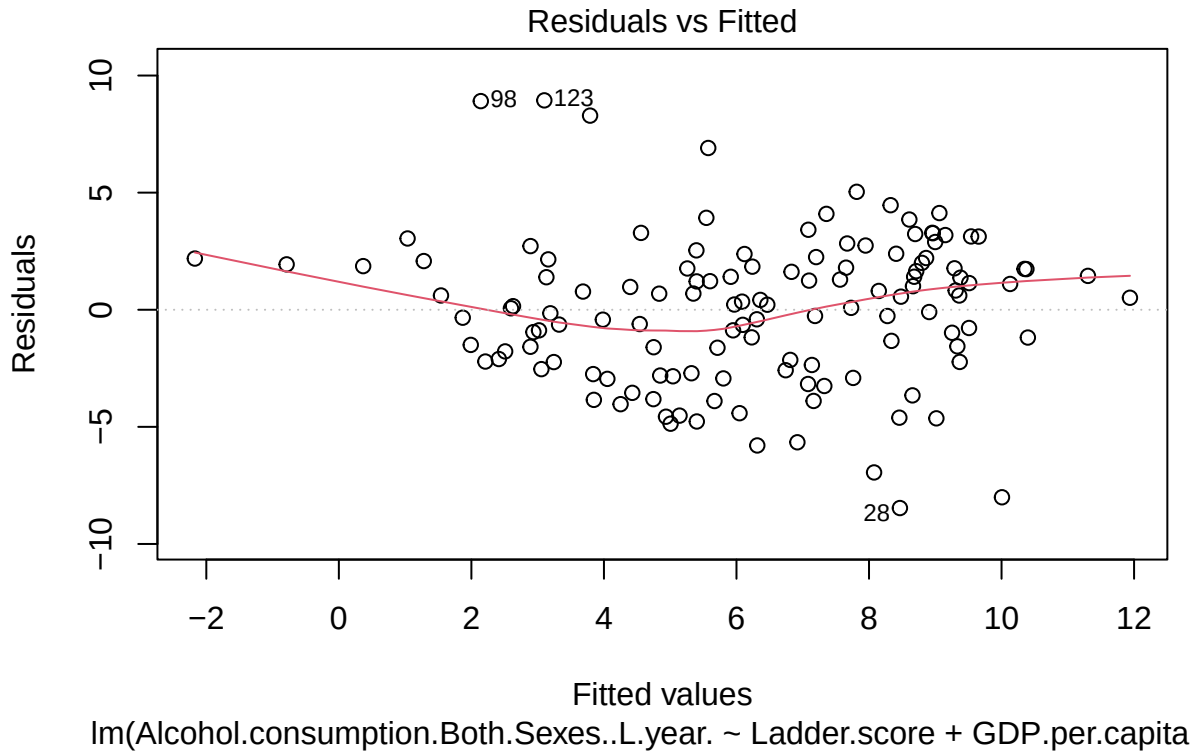
Prvi dio analize će je analiza koeficijenata uz regresore. Vrijednost koeficijenta nam govori o tome koliko promjena u vrijednosti regresora utječe na promjennu izlaza. T vrijednost za svaki regresor opisuje statističku važnost pojedinog regresora. S obzirom na oblik t-distribucije, možemo reći da regresor ima veći statistički značaj ako je njegova vrijednost po iznosu veća. Iz prikazanog sažetka, zaključujemo da regresor Social support ima najveći statistički značaj, te da je statistički značajan i pri p-vrijednosti 0.001, a zatim GDP per capita te Healthcare legatum prosperity index health score koji su statistički značajni pri p-vrijednosti 0.05. Naravno, bitno je gledati i predznak vrijednosti koeficijenta. Iz toga možemo zaključiti da npr. povećanje vrijednosti regresora GDP per capita rezultira povećanjem konzumacije alkohola u zemlji. Suprotno, povećanje vrijednosti regresora Healthcare legatum prosperity index health score rezultira smanjenjem konzumacije alkohola u zemlji.

Vrijednost standardne pogreške reziduala iznosi 3.239 pri 120 stupnjeva slobode. Ta vrijednost nam opisuje standardnu devijaciju reziduala u našem modelu te se izračunava na sljedeći način: $\sqrt{\sum \frac{(y-\hat{y})^2}{df}}$. U toj formuli, y je prava vrijednost varijable koju predviđamo, $\hat{y}$ je predviđena vrijednost, a df je stupanj slobode koji iznosi broj zapisa u našoj tablici - broj regresora (131-11) = 120. Naravno, cilj je imati što manju vrijednost standardne pogreške reziduala jer to znači da model uspješnije predviđa konzumaciju alkohola uz pomoć ostalih podataka.

Vrijednost R^2, odnosno koeficijenta determinacije u našem modelu iznosi 0.4304, a ona predstavlja proporciju varijance zavisne varijable koja se može opisati nezavisnim varijablama (regresorima) u modelu. Cilj je da ona bude što bliže 1 jer će u tom slučaju model biti uspješniji. No, u našem slučaju ta vrijednost nije visoka, štoviše, možemo zaključiti da se manje od pola varijance zavisne varijable može opisati nezavisnim varijablama. Odnosno, ako se sjetimo formule za izračunavanje $R^2 = 1 - \frac{SSE}{SST}$, vidimo da omjer SSE i SST ima vrijednost veću od otprilike 0.57, a mi bismo htjeli da bude bliže 1 jer bi tada prilagodba pravcu bila bolja.

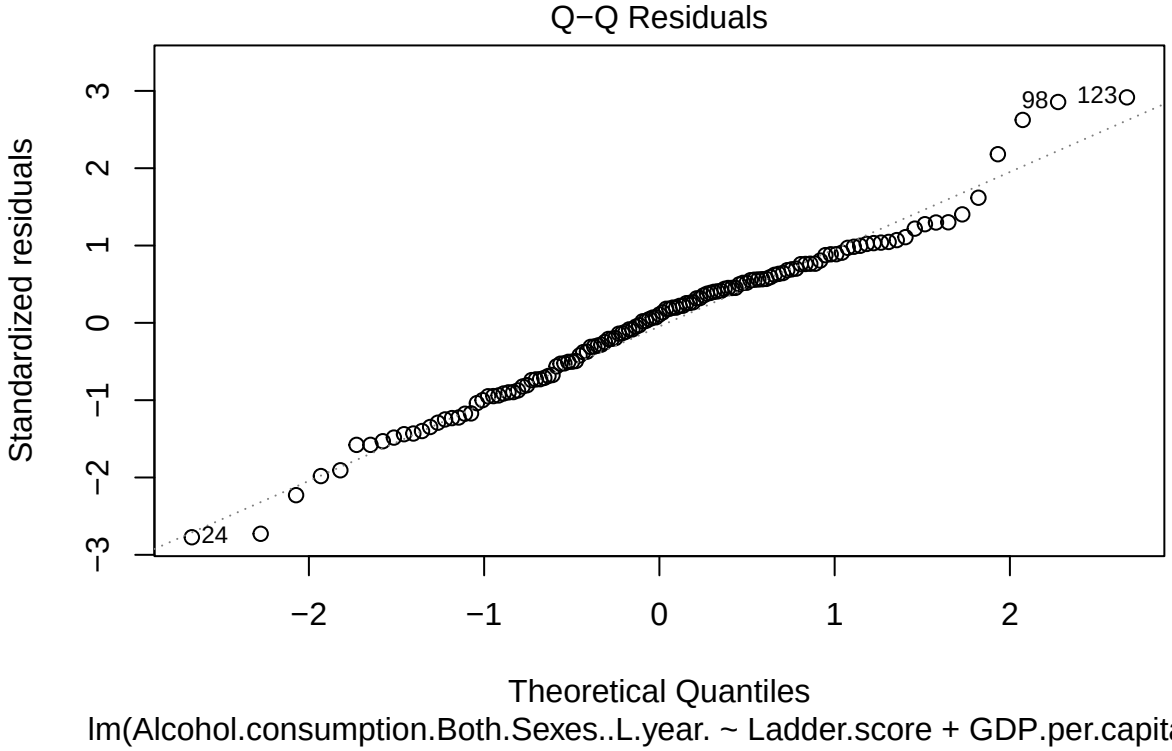
Prilagođeni R^2 je verzija mjere R^2 koja kažnjava velik broj parametara te nam daje točniju procjenu toga koliko je naš model prilagođen pravcu. Ta vrijednost u našem slučaju iznosi 0.383 što je još manje nego vrijednost R^2.
Vrijednost F statistike iznosi 9.069, a p-vrijednost je $5.119e^{-11}$ što je jako mala vrijednost. Interpretacija te vrijednosti je sljedeća: ona testira nultu hipotezu da su svi koeficijenti u modelu jednaki 0, odnosno da niti jedna od nezavisnih varijabli nije korisna za predikciju zavisne varijable. S obzirom da je p-vrijednost iznimno mala, definitvno možemo odbaciti nultu hipotezu te zaključiti da je barem jedna nezavisna varijabla korisna za predikciju.

```
plot(model, which = 1)
```



Ono što želimo ispitati grafom iznad je pretpostavka da reziduali imaju konstantnu varijancu za bilo koji ulaz u model. Da bi donijeli zaključak o toj pretpostavci, moramo pogledati graf iznad. Naime, ako su reziduali otprilike distribuirani u okolini osi apscisa bez nekog jasnog uzorka, možemo zaključiti da je naša pretpostavka točna. Iz grafa vidimo da u našem modelu to je slučaj, reziduali su distribuirani otprilike jednako iznad i ispod osi apscisa te ne možemo utvrditi neki specifičan uzorak među njima. Iz toga slijedi da je naša prvotna pretpostavka bila točna.

```
plot(model, which = 2)
```



Sljedeća pretpostavka je da su reziduali normalno distribuirani. Da bismo provjerili tu pretpostavku, poslužit ćemo se Q-Q grafom iznad. Za graf vrijedi, ako njegove točke otprilike formiraju pravac, pretpostavka o normalno distribuiranim rezidualima je točna. Q-Q graf za naš model otprilike formira pravac, naravno, moramo napomenuti da taj pravac nije savršen te da postoje točke koje odstupaju od pravca, posebno na krajnje lijevom i krajnje desnom dijelu spektra. No, točke formiraju uzorak koji je dovoljno blizu pravcu, te možemo zaključiti da je naša prvotna pretpostavka bila točna.\

Zaključak: Iz svih gore navedenih rezultata i analiza moramo donijeti konačni zaključak koji je ujedno i odgovor na pitanje postavljeno u podnaslovu: Možemo li temeljem drugih dostupnih varijabli predvidjeti konzumaciju alkohola po zemljama? Odgovor na pitanje nije jednostavan. Naravno da mi uvijek možemo dati podatke na ulaz modela i dobiti izlaz. No, s obzirom na to da je u našem slučaju vrijednost R^2 prilično niska, odgovor je da možemo predvidjeti, ali to predviđanje neće uvijek biti jako blizu stvarnom, odnosno da će to predviđanje imati određenu pogrešku koja nije zanemariva. Da bi uspješnije predviđali konzumaciju alkohola, vjerojatno bi bilo pametnije koristiti neki drugi matematički model, ili ući u domenu neuronskih mreža i strojnog učenja gdje bi se susreli s algoritmima koji bi ovaj zadatak odrađivali uspješnije, ali s time bi izašli iz područja ovog predmeta, te to ovdje nećemo raditi.

Predviđanje konzumacije alkohola za 6 zemalja s nedostajućim vrijednostima za tu značajku

Dodatak samoj analizi modela linearne regresije bit će predviđanje konzumacije alkohola pomoću našeg izračunatog modela linearne regresije. S obzirom na to da za 6 zemalja za koje ćemo raditi predikciju nemamo stvarne podatke, nećemo moći provjeriti koliko je naše predviđanje zbilja točno. Osim toga, iz analize iznad vidjeli smo da model linearne regresije ne opisuje zadane podatke idealno, te da ima nezanemarivu pogrešku, što također moramo uzeti u obzir za našu predikciju.
Prvi korak je utvrđivanje za koje države nema podataka o konzumaciji alkohola.

```
my_data <- WHR_23 %>%
group_by(Regional.indicator) %>%
mutate(
Ladder.score = ifelse(is.na(Ladder.score),
mean(Ladder.score, na.rm = TRUE), Ladder.score),
GDP.per.capita = ifelse(is.na(GDP.per.capita),
mean(GDP.per.capita, na.rm = TRUE), GDP.per.capita),
Social.support = ifelse(is.na(Social.support),
mean(Social.support, na.rm = TRUE), Social.support),
Healthy.life.expectancy = ifelse(is.na(Healthy.life.expectancy), mean(Healthy.life.expectancy, na.rm = 
Freedom.to.make.life.choices = ifelse(is.na(Freedom.to.make.life.choices), mean(Freedom.to.make.life.ch
Generosity = ifelse(is.na(Generosity), mean(Generosity, na.rm = TRUE), Generosity),
Perceptions.of.corruption = ifelse(is.na(Perceptions.of.corruption), mean(Perceptions.of.corruption, na
Crime.rate.Crime.Index = ifelse(is.na(Crime.rate.Crime.Index), mean(Crime.rate.Crime.Index, na.rm = TRU
Healthcare.Legatum.Prosperity.Index.Health.Score = ifelse(is.na(Healthcare.Legatum.Prosperity.Index.Hea
Gini.Coefficient...World.Bank = ifelse(is.na(Gini.Coefficient...World.Bank), mean(Gini.Coefficient...Wo
) %>%
ungroup()
my_data <- my_data[, c(1,3,4,5,6,7,8,9,10,13,14,15)]
t <- my_data[is.na(my_data$Alcohol.consumption.Both.Sexes..L.year.), ]
tmp <- t[,c(1)]
size <- dim(tmp)
cat("Broj redaka:", size[1], "\n")
```

```
## Broj redaka: 6
```

```
head(tmp, 6)
```

```
## # A tibble: 6 x 1
##   Country.name
##   <chr>
## 1 Czechia
## 2 Taiwan Province of China
## 3 Kosovo
## 4 Hong Kong S.A.R. of China
## 5 Congo (Brazzaville)
## 6 State of Palestine
```

Sljedeći korak je napraviti predikcije za gore navedene države.

```
coefficients <- coef(model)
tmp <- t[,c(2,3,4,5,6,7,8,9,10,11,12)]
preds <- predict(model, tmp)
print(preds)
```

```
##        1        2        3        4        5        6
## 9.885382 8.061258 5.435370 8.958152 3.163369 6.401986
```

Sad smo dobili predikcije za svaku od 6 država. Vidimo da od tih 6 najveću konzumaciju alkohola po glavi stanovnika ima Češka, skoro 10 litara, a najmanju Kongo (Brazaville), nešto više od 3 litre. Također, vidimo da predikcija kaže da se u Palestini popije preko 6 litara alkohola godišnje po glavi stanovnika što je više nego na Kosovu na primjer. To ne bi trebao biti istinit podatak s obzirom na vjeru stanovnika Palestine. To nam pokazuje da predikcije koje smo napravili treba uzeti s rezervom, a daljnja analiza rezultata nema pretjeranog smisla s obzirom na to da nam stvarne vrijednosti nisu poznate.

3. Postoje li razlike u kvaliteti zdravstvene skrbi među različitim regijama?

Prvo je potrebno provjeriti jesu li podatci normalno distibuirani. To ćemo napraviti analitički i grafički. Analitičku provjeru čini Kolmogorov-Smirnov test. Postavljamo hipoteze:

H0: podatci su normalno distribuirani

H1: podatci nisu normalno distribuirani

Prisjetimo se, varijabla "Healthcare.Legatum.Prosperity.Index.Health.Score" ima dvije nedostajuće vrijednosti. Jedan od načina na koji se to može riješiti je da svedemo nedostajuće vrijednosti na srednju vrijednost te varijable za pripadnu regiju. U ovom slučaju ne možemo raditi takvu procjenu zato što pitanje kojim se bavimo ovisi o regijama. Nedostajuće vrijednosti su za Kosovo i Palestinu, države kojima to nije jedini nedostajući podatak. Prema tome, jednostavno ćemo te dvije države ukloniti iz daljnje procjene zdravstvene skrbi.

```
data3 <- WHR_23[!is.na(WHR_23$Healthcare.Legatum.Prosperity.Index.Health.Score), ]
names(data3)[names(data3) == "Healthcare.Legatum.Prosperity.Index.Health.Score"] <- "Health.Score"
names(data3)[names(data3) == "Regional.indicator"] <- "Region"
regions <- unique(data3$Region)

ks_results <- list()

# KS test se radi za svaku regiju
for (region in regions) {
  data_region <- data3$Health.Score[data3$Region == region]
  ks_result <- ks.test(data_region, "pnorm", mean = mean(data_region), sd = sd(data_region))
  ks_results[[region]] <- ks_result
}
```

```
## Warning in ks.test.default(data_region, "pnorm", mean = mean(data_region), :
## ties should not be present for the Kolmogorov-Smirnov test
```

```
for (k in names(ks_results)) {
  cat(k, ":\n")
  print(ks_results[[k]])
  cat("\n")
}
```

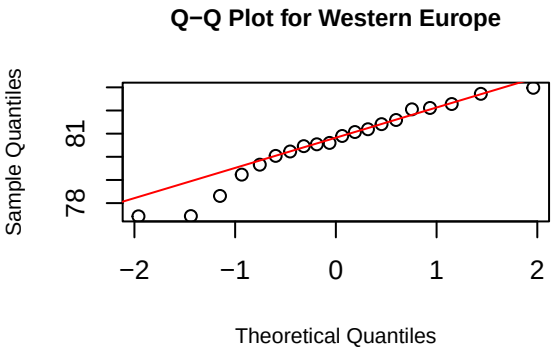
```
## Western Europe :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.11197, p-value = 0.9397
## alternative hypothesis: two-sided
##
##
## Middle East and North Africa :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.098219, p-value = 0.9982
## alternative hypothesis: two-sided
##
##
## North America and ANZ :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.34591, p-value = 0.6191
## alternative hypothesis: two-sided
##
##
## Central and Eastern Europe :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.092969, p-value = 0.9968
## alternative hypothesis: two-sided
##
##
## Latin America and Caribbean :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.10035, p-value = 0.9805
## alternative hypothesis: two-sided
##
##
## Southeast Asia :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.21607, p-value = 0.7185
## alternative hypothesis: two-sided
##
##
```

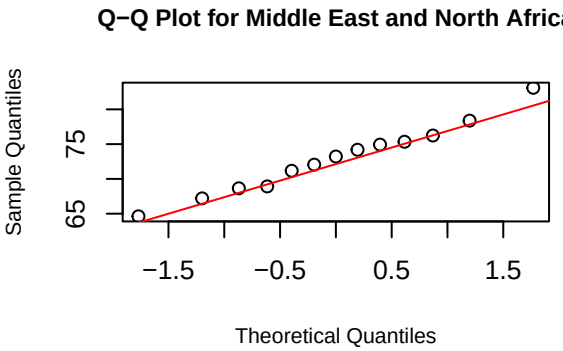
```
## East Asia :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.35303, p-value = 0.3569
## alternative hypothesis: two-sided
##
##
## Commonwealth of Independent States :
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.19633, p-value = 0.8785
## alternative hypothesis: two-sided
##
##
## Sub-Saharan Africa :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.089175, p-value = 0.9346
## alternative hypothesis: two-sided
##
##
## South Asia :
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.18957, p-value = 0.9535
## alternative hypothesis: two-sided
```

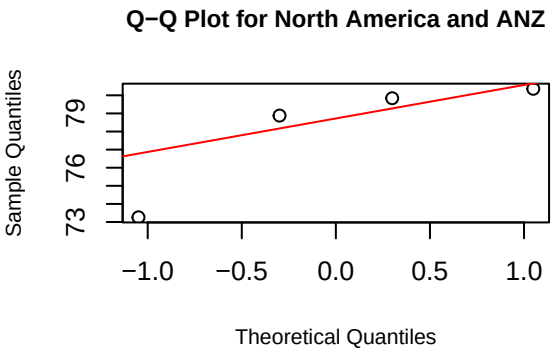
Za svaku regiju dobivamo veliku p-vrijednost, što znači da ne možemo odbaciti hipotezu H0 i zaključujemo da su podatci normalno distribuirani.

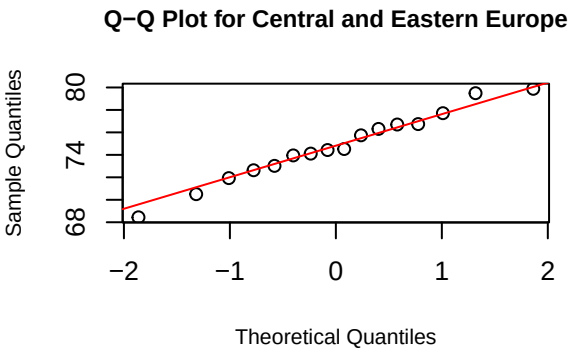
P-vrijednost ovisi o veličini uzorka pa ćemo se koristiti i grafičkom provjerom. Veći uzorak rezultira manjom p-vrijednošću.

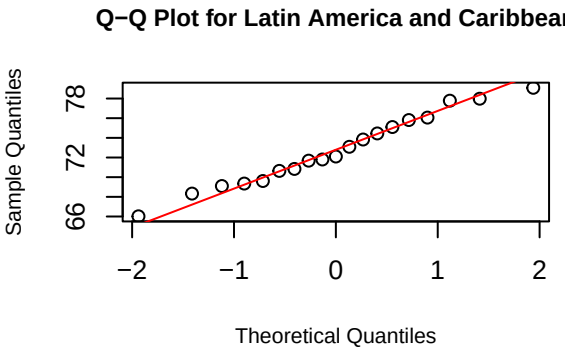
```r
library(ggplot2)

par(mfrow = c(2, 2))  # 2 rows, 2 columns
for (region in regions) {
  data_region <- data3$Health.Score[data3$Region == region]
  qqnorm(data_region, main = paste("Q-Q Plot for", region), cex.main=0.9, cex.lab=0.8)
  qqline(data_region, col = "red")
}
```

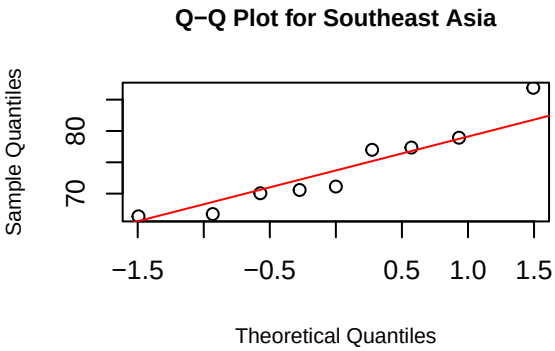

Q−Q Plot for Western Europe
Sample Quantiles
78
81
−2
−1
0
1
2
Theoretical Quantiles

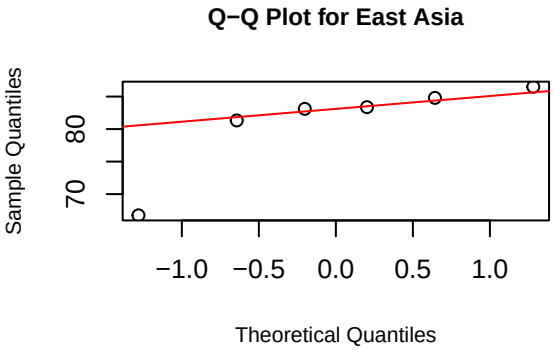

Q−Q Plot for Middle East and North Africa
Sample Quantiles
65
75
−1.5
−0.5
0.5
1.5
Theoretical Quantiles

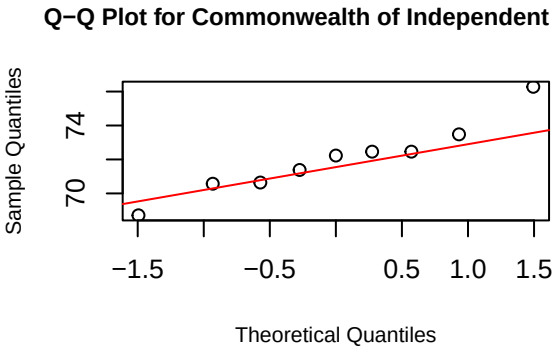

Q−Q Plot for North America and ANZ
Sample Quantiles
73
76
79
−1.0
−0.5
0.0
0.5
1.0
Theoretical Quantiles

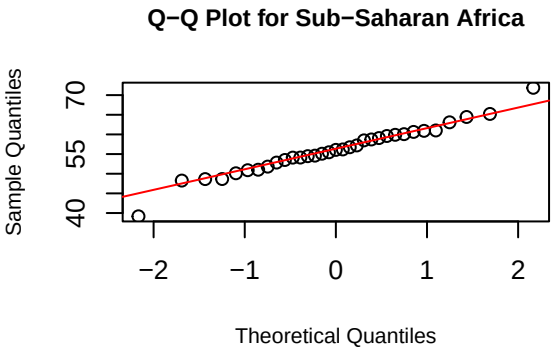

Q−Q Plot for Central and Eastern Europe
Sample Quantiles
68
74
80
−2
−1
0
1
2
Theoretical Quantiles

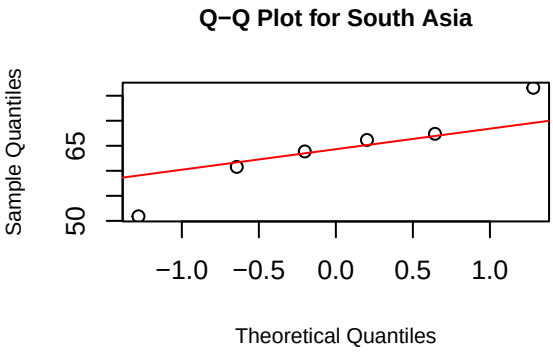

Q−Q Plot for Latin America and Caribbean
Sample Quantiles
66
72
78
−2
−1
0
1
2
Theoretical Quantiles


Q−Q Plot for Southeast Asia
Sample Quantiles
70
80
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles


Q−Q Plot for East Asia
Sample Quantiles
70
80
−1.0
−0.5
0.0
0.5
1.0
Theoretical Quantiles


Q−Q Plot for Commonwealth of Independent State
Sample Quantiles
70
74
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles





Općenito, što su podatci udaljeniji od crvene linije, to su manje normalno distribuirani. U našem slučaju se iz grafičkog prikaza može vidjeti da se uglavnom radi o normalnoj distribuciji, osim nekih stršećih vrijednosti.

Nakon što smo zaključili da su podaci normalno distribuirani, mogli smo provesti test koji će dati odgovor na postavljeno pitanje - ANOVA test, odnosno test analize varijance. To je statistički test koji procjenjuje tako što traži ima li značajne razlike u varijabilnosti srednjih vrijednosti. Ovaj test je pogodan jer imamo više od dvije regije za koje trebamo provjeriti razliku u zavisnoj varijabli - kvaliteti zdravstvene skrbi. U rezultatima testa gledamo p-vrijednost te F-statistiku.F- statistika zapravo nam govori kolika je razlika u srednjim vrijednostima regija. Dakle, ako je F-vrijednost niska, te razlike nisu statistički bitne, dok u obrnutom slučaju znači da postoji bitna razlika između skupina, odnosno, regija. Što se tiče p-vrijendosti, moramo gledati kako se odnosi s obzirom na našu odabranu razinu značajnosti. Ako je manja od značajnosti, odbacuje se nulta hipoteza, a to znači da postoje bitne razlike između nekih skupina.

Hipoteze:

H0: srednje vrijednosti regija su jednake

H1: barem neka srednja vrijednost regije se razlikuje od ostalih

```
model <- aov(Health.Score ~ Region, data = data3)
summary(model)
```

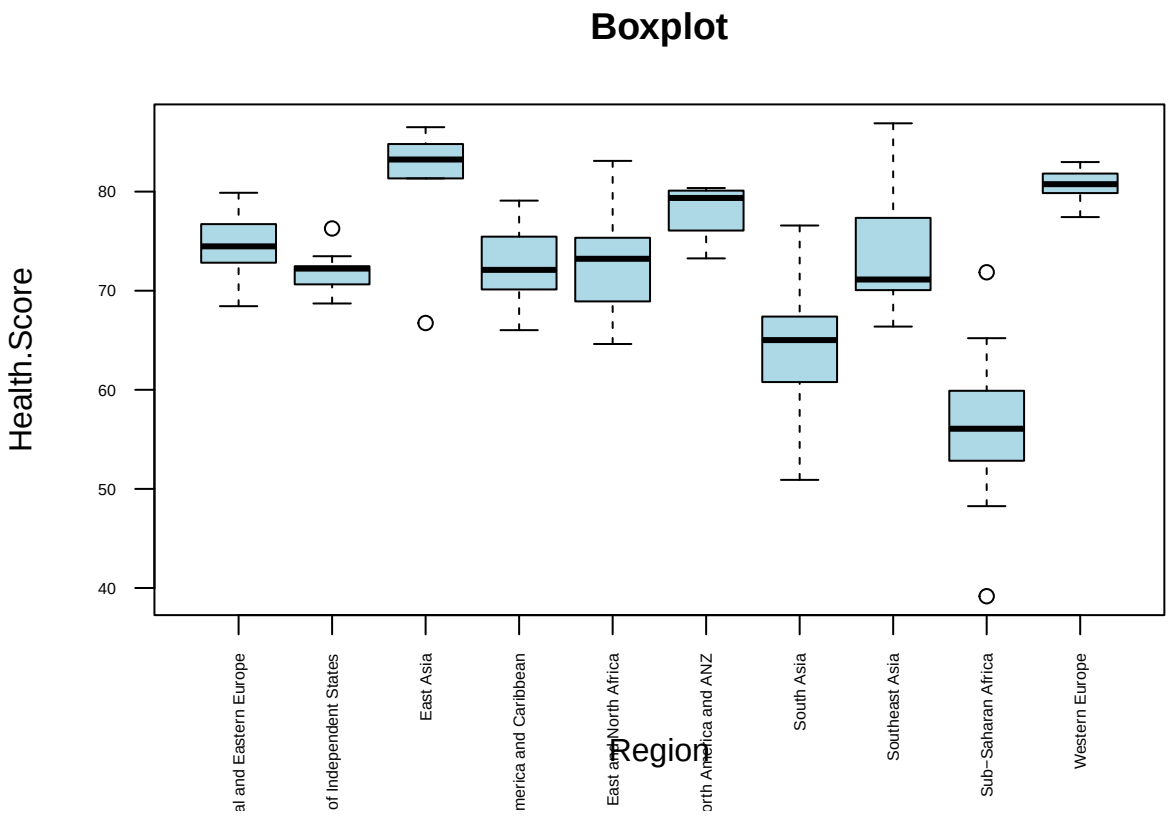
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Region        9  10572  1174.7   49.95 <2e-16 ***
## Residuals   125   2940    23.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dakle, p-vrijednost je $<$ 2e-16 što se može pročitati iz sažetka testa prikazanog gore (Pr(>F)), a to je manje

od naše razine značajnosti (0.05) te možemo odbaciti nultu hipotezu. Isto tako, vidimo da je F vrijednost dosta velika.

Kako bi bolje dočarali rezultate testa, odlučili smo se za vizualizaciju box plot dijagramom.

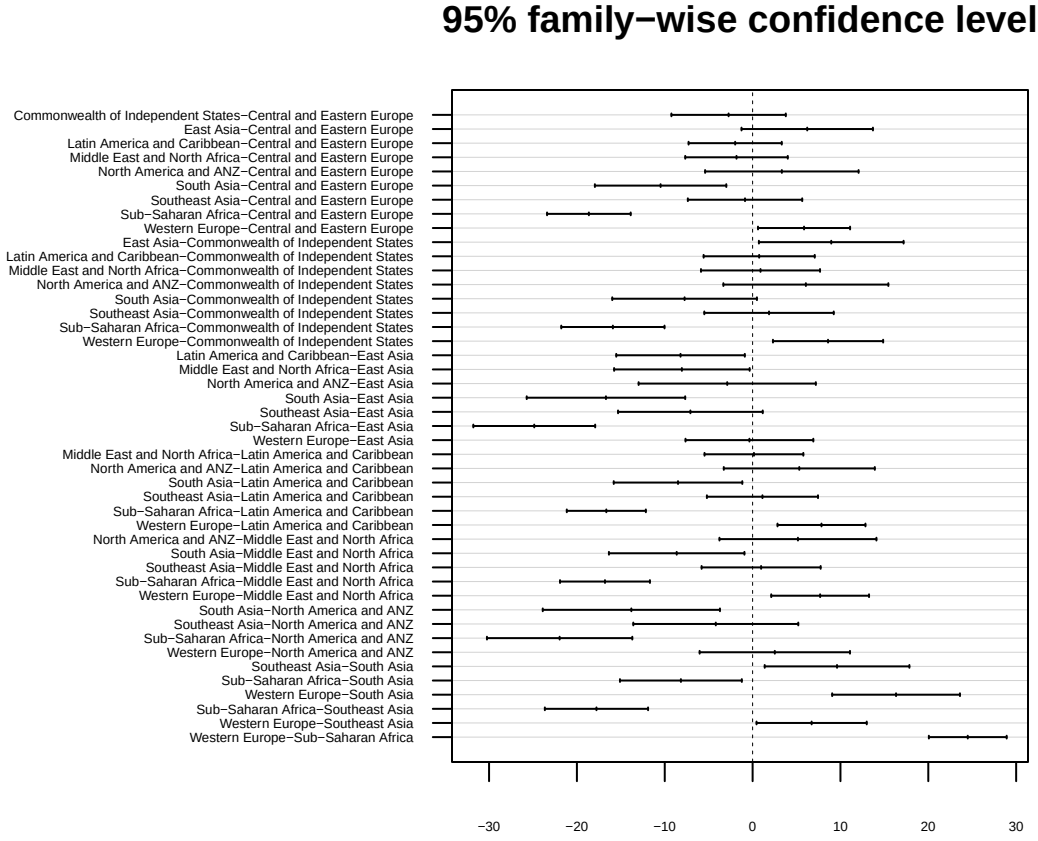
```
boxplot(Health.Score ~ Region, data = data3, col = "lightblue",
        main = "Boxplot", las = 2, cex.axis=0.5)
```



Ovaj graf dobro vizualizira razlike u zdravstvenoj skrbi prema regijama. Možemo primijetiti kako najslabiji indeks zdravstvene skrbi ima subsaharska Afrika dok najbolji indeks ima istočna Azija.

Sada kada znamo da razlike u zdravstvenoj skrbi između regija postoje, možemo napraviti TukeyHSD test koji će nam detaljnije objasniti te razlike.

```
tukey <- TukeyHSD(model)
tukey_table<-as.data.frame((tukey)[1])
write.csv(tukey_table, 'TK_data.csv')
with(par(mai=c(0.5,2.5,0.5,1)),{plot(tukey, las=1,cex.axis=0.4)})
```



Tukey HSD test nam daje vrijednosti "diff", koja pokazuje razliku u srednjoj vrijednosti između dvije promatrane grupe, "lwr" i "upr" koji predstavljaju granice intervala pouzdanosti, te p-vrijednost. Ako je p-vrijednost manja od 0.05, imamo par grupa koje se statistički značajno razlikuju.

Ako gledamo samo Europu, jednu grupu čini zapadni dio, dok drugu grupu čine centralni i istočni dio. Razlika srednjih vrijednosti iznosi 5.86, što se ne čini puno, ali se pokazalo statistički značajno s p-vrijednošću 0.016.

Unutar same Azije postoje velike razlike razvijenosti zdravstva, s istočnom Azijom koja je značajno bolja od južne Azije. U istočnoj Aziji indeks razvijenosti zdravstva iznosi više od 80, s iznimkom Mongolije čiji je indeks 66.74.

Najmanje zdravstveno razvijena pokazala se regija subsaharska Afrika čiji rezultati odskaču od svih ostalih regija. Najviše odstupa od istočne Azije s razlikom srednjih vrijednosti 24.86 i zapadne Europe s razlikom 24.49.

Najbolje zdravstveno razvijene regije su istočna Azija i zapadna Europa, te odmah uz njih Sjeverna Amerika, Australija i Novi Zeland.