World Happiness Report 2023

Mia Gmiza, Gabrijela Perković, Matija Roginić, Erika Tomakić

2023-12-06

Uvod

*dodati opis naseg zadatka

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"
```

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

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

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

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

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

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

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

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

```r
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
```

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.

```r
data3 <- WHR_23[!is.na(WHR_23$Healthcare.Legatum.Prosperity.Index.Health.Score), ]

regions <- unique(data3$Regional.indicator)

ks_results <- list()

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

```
## Warning in ks.test(data_region, "pnorm", mean = mean(data_region), sd =
## sd(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 :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.11197, p-value = 0.9397
## alternative hypothesis: two-sided
##
##
## Middle East and North Africa :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.098219, p-value = 0.9982
## alternative hypothesis: two-sided
##
##
## North America and ANZ :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.34591, p-value = 0.6191
## alternative hypothesis: two-sided
##
##
## Central and Eastern Europe :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.092969, p-value = 0.9968
## alternative hypothesis: two-sided
##
##
## Latin America and Caribbean :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.10035, p-value = 0.9805
## alternative hypothesis: two-sided
##
##
## Southeast Asia :
```

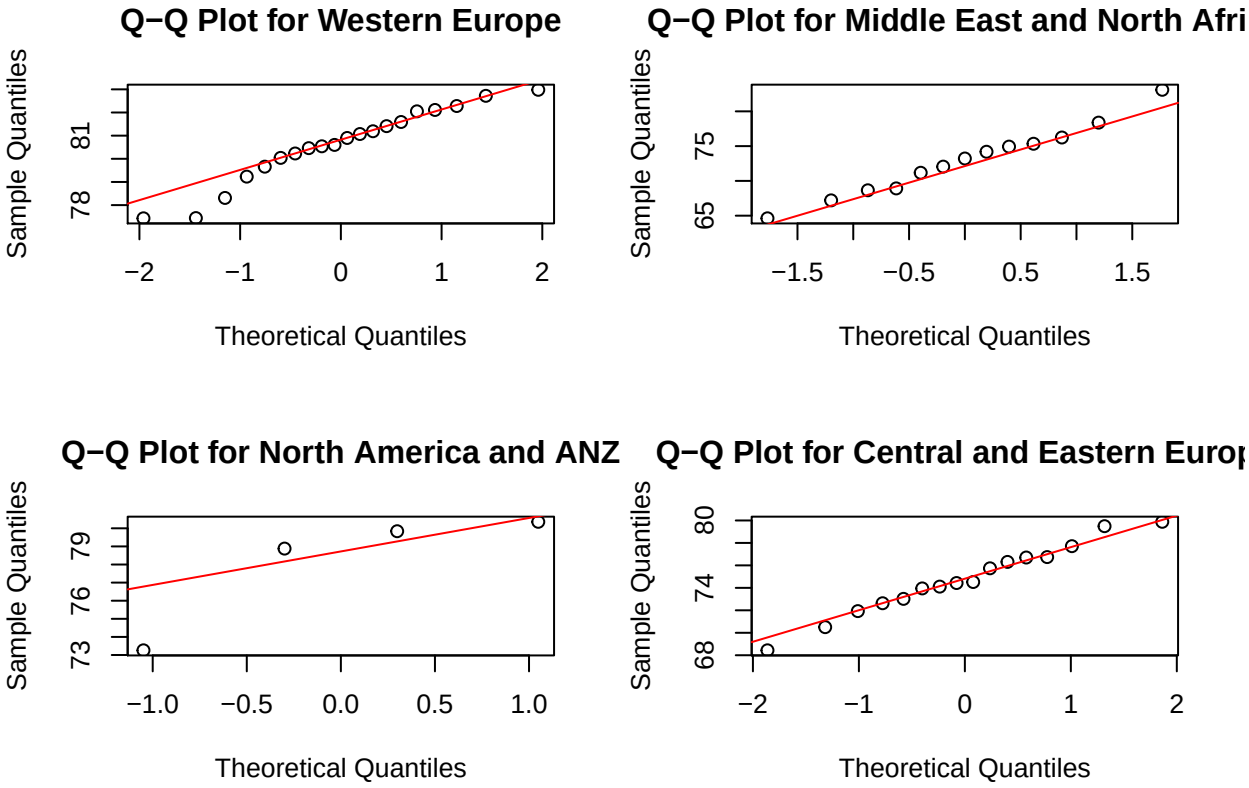
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.21607, p-value = 0.7185
## alternative hypothesis: two-sided
##
##
## East Asia :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.35303, p-value = 0.3569
## alternative hypothesis: two-sided
##
##
## Commonwealth of Independent States :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.19633, p-value = 0.8785
## alternative hypothesis: two-sided
##
##
## Sub-Saharan Africa :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.089175, p-value = 0.9346
## alternative hypothesis: two-sided
##
##
## South Asia :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.18957, p-value = 0.9535
## alternative hypothesis: two-sided
```

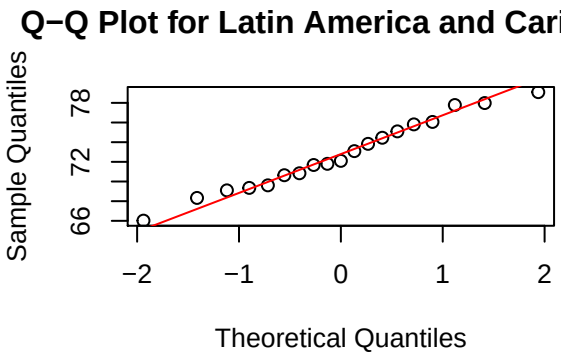
Za svaku regiju dobivamo veliku p-vrijednost, što znači da ne možemo odbaciti hipotezu H0 i zaključujemo da su podatci normalno distribuirani.

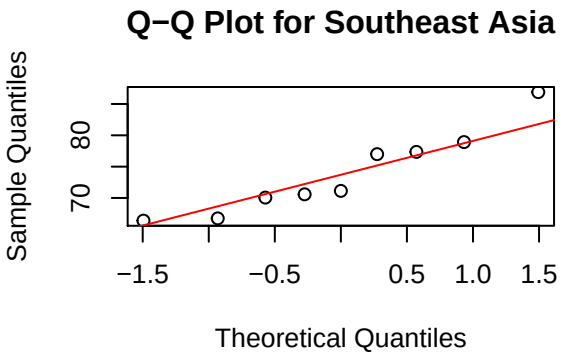
P-vrijednost ovisi o veličini uzorka pa ćemo se koristiti i grafičkom provjerom. Veći uzorak rezultira manjom p-vrijednošću.

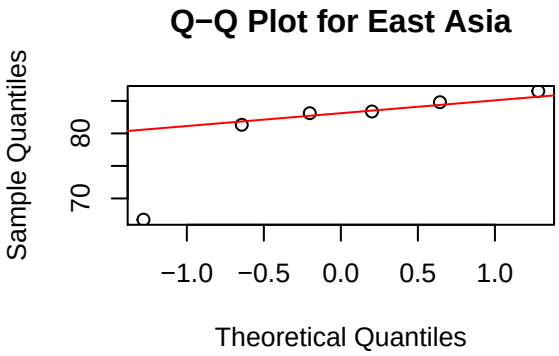
```
library(ggplot2)

#if (!requireNamespace("car", quietly = TRUE)) {
#  install.packages("car")
#}
#library(car)
```

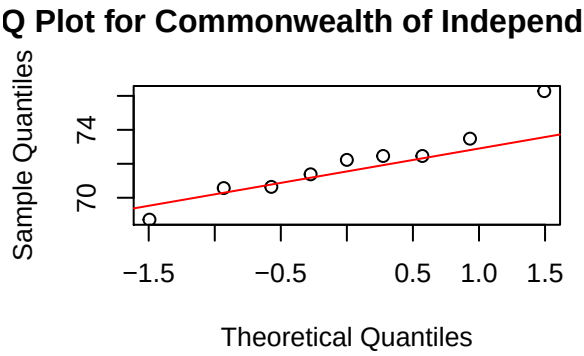
```r
par(mfrow = c(2, 2))  # 2 rows, 2 columns
for (region in regions) {
  data_region <- data3$Healthcare.Legatum.Prosperity.Index.Health.Score[data3$Regional.indicator == reg
  qqnorm(data_region, main = paste("Q-Q Plot for", region))
  qqline(data_region, col = "red")
}
```

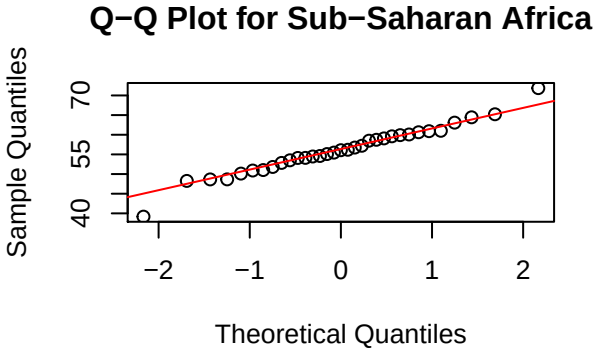


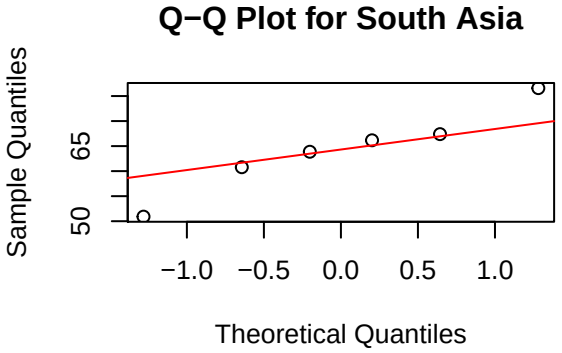












Općenito, što su podatci udaljeniji od crvene linije, to su manje normalno distribuirani. U našem slučaju se iz grafičkog prikaza može vidjeti da se uglavnom radi o normalnoj distribuciji, osim nekih stršećih vrijednosti.