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

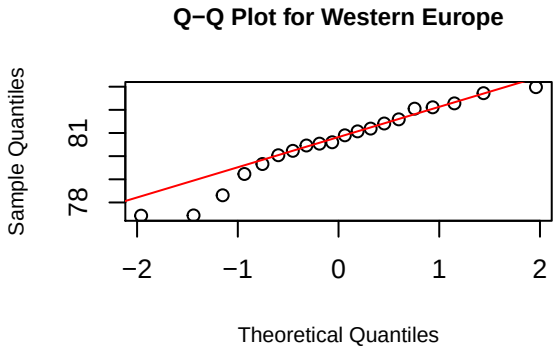
```
##
##
## Southeast Asia :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.21607, p-value = 0.7185
## alternative hypothesis: two-sided
##
##
## East Asia :
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_region
## D = 0.35303, p-value = 0.3569
## alternative 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
```

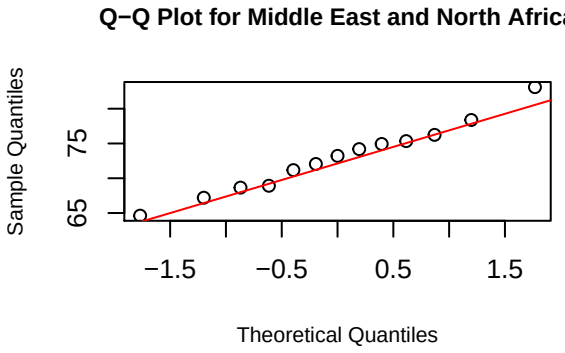
Za svaku regiju dobivamo veliku p-vrijednost, što znači da ne možemo odbaciti hipotezu H0 i zaključujemo da su podatci normalno distribuirani.

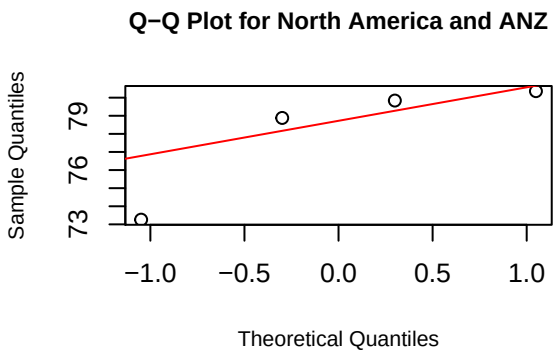
P-vrijednost ovisi o veličini uzorka pa ćemo se koristiti i grafičkom provjerom. Veći uzorak rezultira manjom p-vrijednošću.

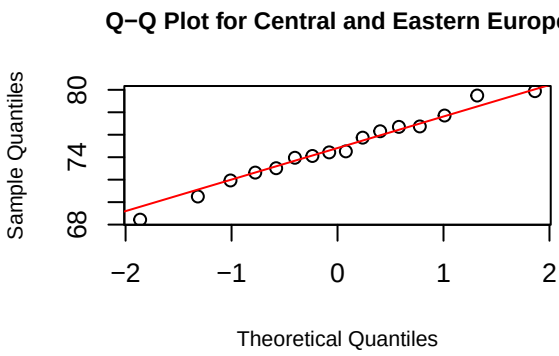
```
library(ggplot2)

par(mfrow = c(2, 2))  # 2 rows, 2 columns
```

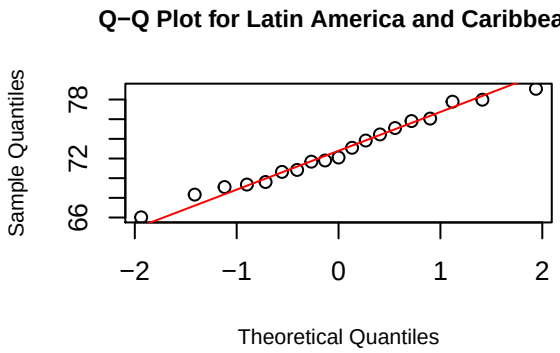
```r
for (region in regions) {
  data_region <- data3$Health.Score[data3$Region == region]
  qqnorm(data_region, main = paste("Q-Q Plot for", region), cex.main=0.9, cex.lab=0.8)
  qqline(data_region, col = "red")
}
```

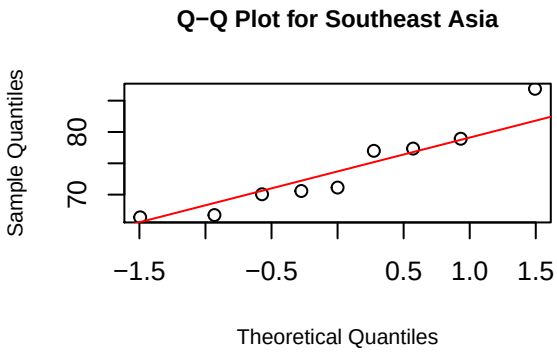


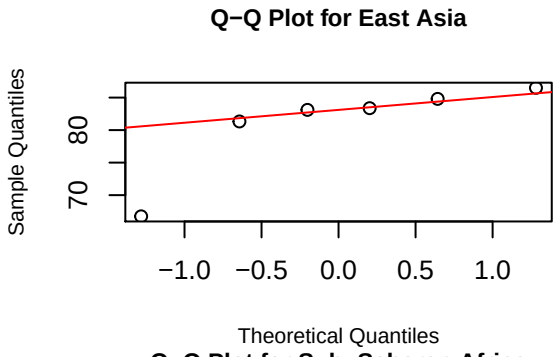


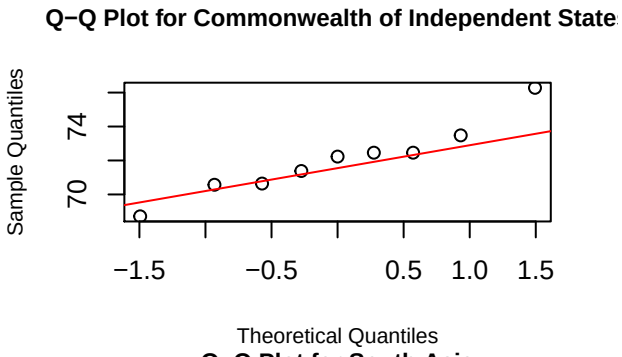


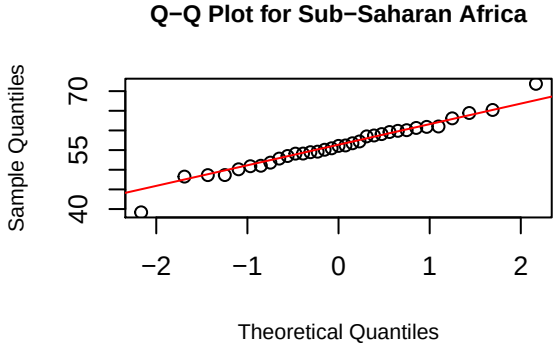


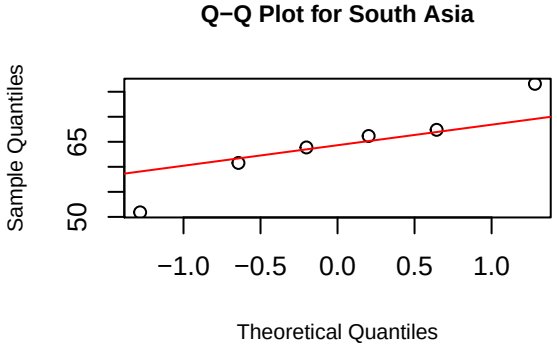












Općenito, što su podatci udaljeniji od crvene linije, to su manje normalno distribuirani. U našem slučaju se iz grafičkog prikaza može vidjeti da se uglavnom radi o normalnoj distribuciji, osim nekih stršećih vrijednosti.

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

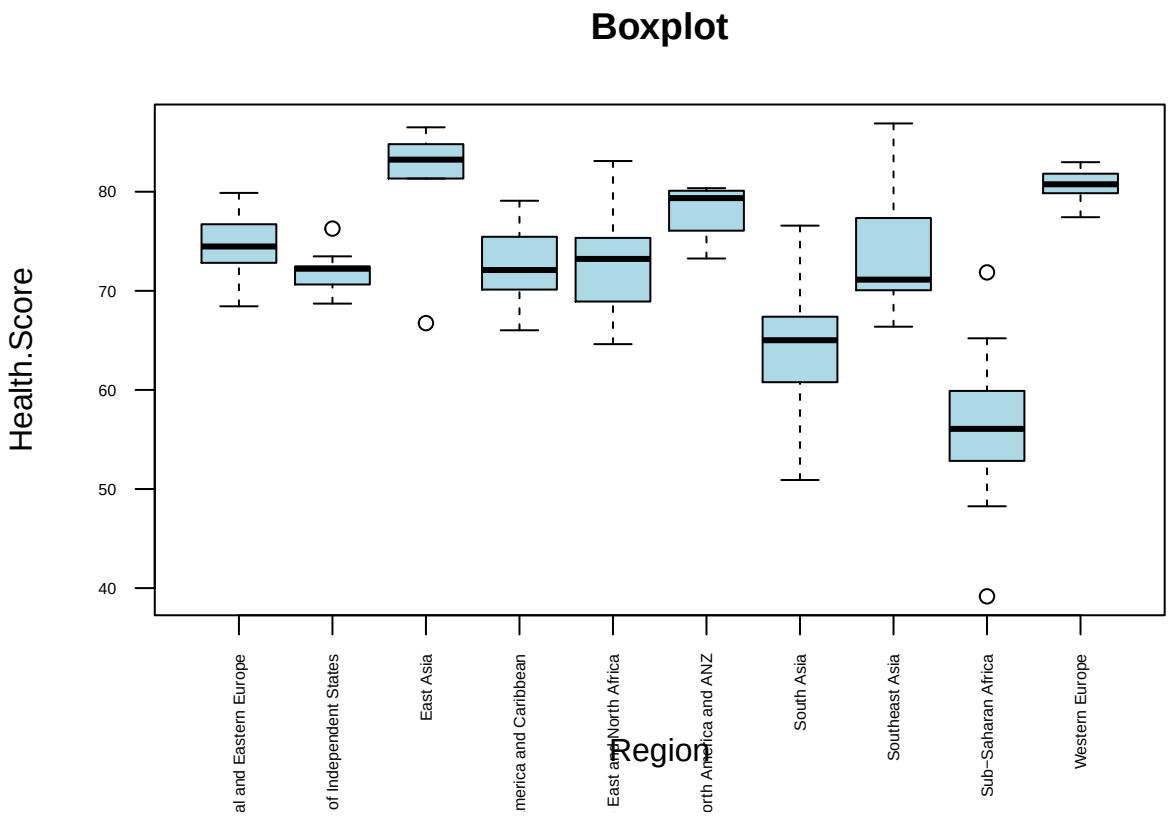
```
model <- aov(Health.Score ~ Region, data = data3)
summary(model)
```

```
##              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.

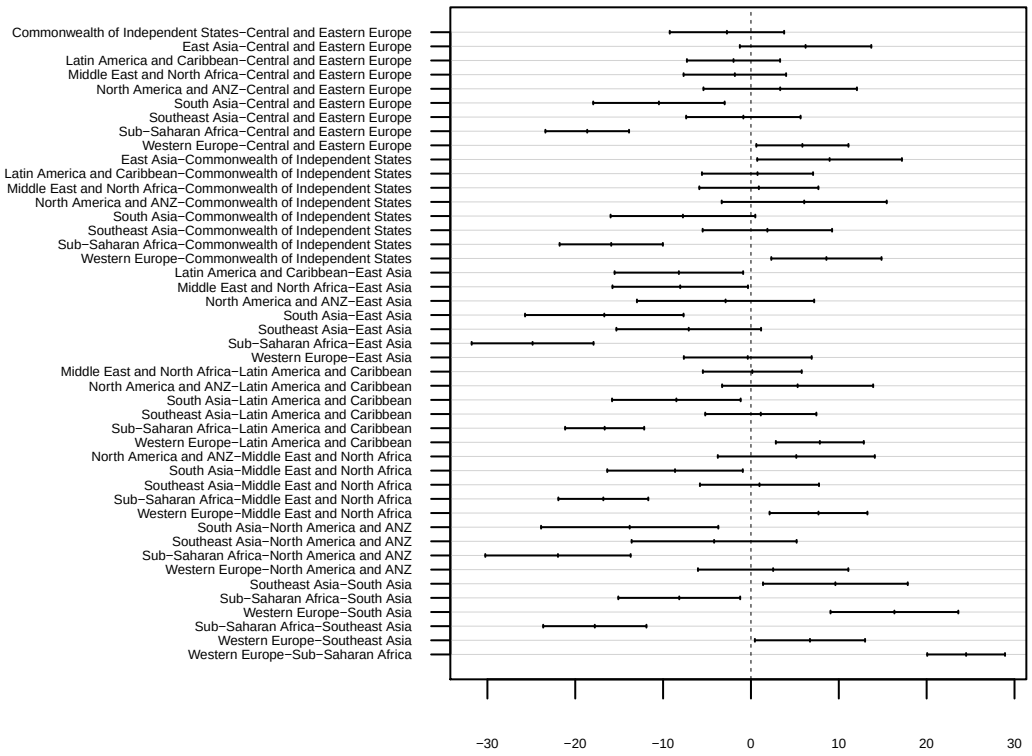
```
boxplot(Health.Score ~ Region, data = data3, col = "lightblue",
        main = "Boxplot", las = 2, cex.axis=0.5)
```



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.