

Usporedba gradske potrošnje između kontinenata

Pitanje 2: Postoji li statistički značajna razlika u gradskoj potrošnji automobila između različitih kontinenata proizvođača?

```
grouped <- group_by(my_cars, continent)

summary.result1 <- summarise(
  grouped,
  count = n(),
  mean_city = mean(city.L.100km, na.rm = TRUE),
  median_city = median(city.L.100km, na.rm = TRUE),
  sd_city = sd(city.L.100km, na.rm = TRUE),
  min_city = min(city.L.100km, na.rm = TRUE),
  max_city = max(city.L.100km, na.rm = TRUE),
  q25 = quantile(city.L.100km, 0.25, na.rm = TRUE),
  q75 = quantile(city.L.100km, 0.75, na.rm = TRUE)
)
summary.result1
```

```
## # A tibble: 3 x 9
##   continent    count mean_city median_city sd_city min_city max_city   q25   q75
##   <chr>         <int>    <dbl>     <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 Asia           107     9.12       8.7     2.04     4.8    14.7  7.58  9.79
## 2 Europe          74    11.5       11.2     2.48     6.35    18.1  9.79 12.9
## 3 North Ameri~    20     8.46       7.58     2.19     5      12.4  7.27  9.79
```

Tablica prikazuje mjere centralne tendencije gradske potrošnje automobila grupiranih po kontinentu proizvođača. Može se uočiti kako se aritmetička sredina i medijan gradske potrošnje razlikuju među kontinentima.

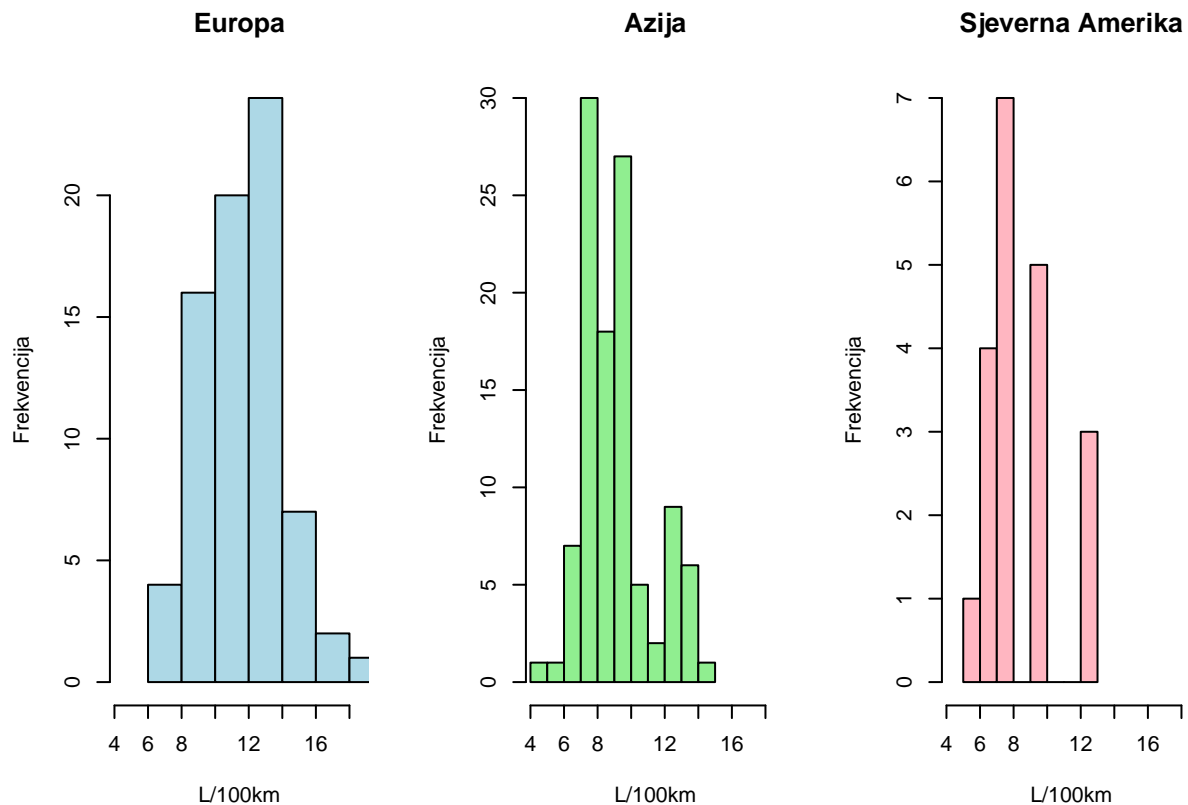
```
xrange <- range(my_cars$city.L.100km) + c(-0.5, 0.5)
ymax <- max(
  hist(my_cars[my_cars$continent == "Europe", ]$city.L.100km, plot = FALSE)$counts,
  hist(my_cars[my_cars$continent == "Asia", ]$city.L.100km, plot = FALSE)$counts,
  hist(my_cars[my_cars$continent == "North America", ]$city.L.100km, plot = FALSE)$counts
)

par(mfrow = c(1, 3))

hist(my_cars[my_cars$continent == "Europe", ]$city.L.100km,
  main = "Europa",
  xlab = "L/100km",
  ylab = "Frekvencija",
  col = "lightblue",
  border = "black",
  xlim = xrange)
```

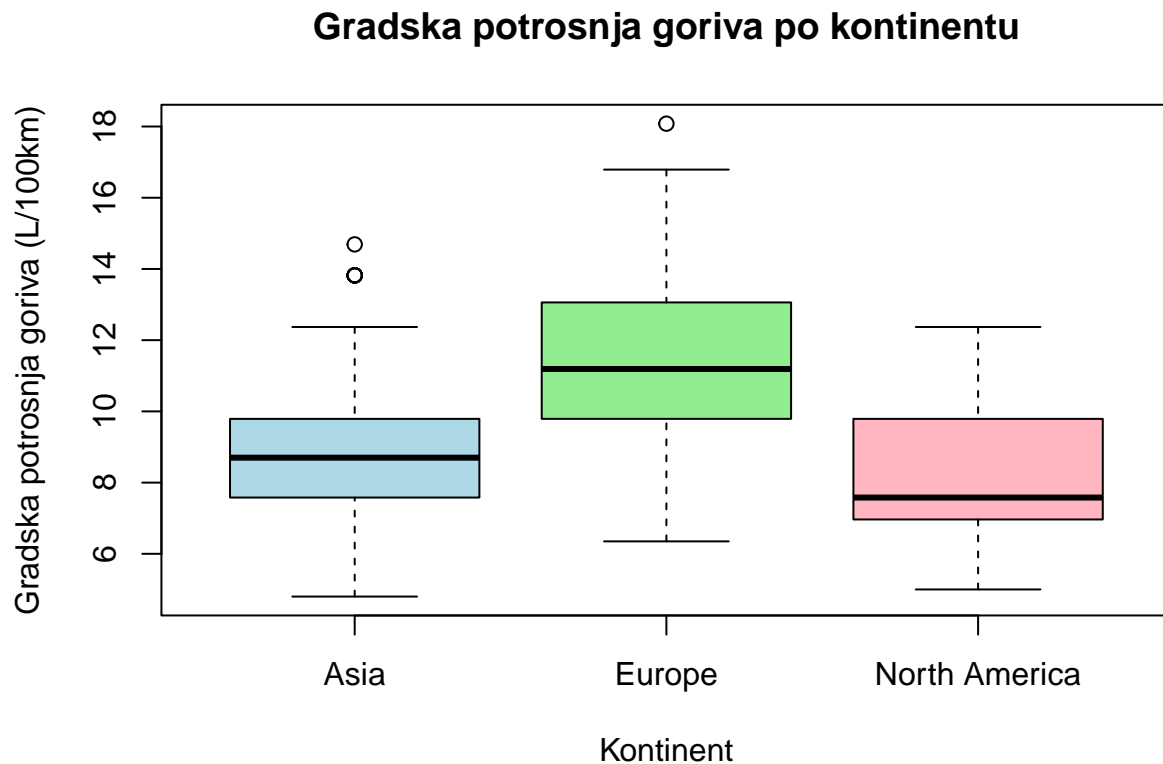
```
hist(my_cars[my_cars$continent == "Asia", ]$city.L.100km,
     main = "Azija",
     xlab = "L/100km",
     ylab = "Frekvencija",
     col = "lightgreen",
     border = "black",
     xlim = xrange)

hist(my_cars[my_cars$continent == "North America", ]$city.L.100km,
     main = "Sjeverna Amerika",
     xlab = "L/100km",
     ylab = "Frekvencija",
     col = "lightpink",
     border = "black",
     xlim = xrange)
```



Ova tri histograma prikazuju raspodjelu automobila po gradskoj potrošnji, pri čemu svaki histogram predstavlja jedan od kontinenata. Iz prikaza je jasno vidljiva razlika u potrošnji, osobito za Europu, gdje je gradska potrošnja, prema ovom uzorku, znatno viša u odnosu na Sjevernu Ameriku i Aziju.

```
boxplot(city.L.100km ~ continent, data = my_cars,
        main = "Gradska potrošnja goriva po kontinentu",
        xlab = "Kontinent",
        ylab = "Gradska potrošnja goriva (L/100km)",
        col = c("lightblue", "lightgreen", "lightpink"),
        family="Helvetica")
```



U boxplot dijagramima jasno je vidljiva značajna razlika u medijanima i ostalim kvartalima između kontinenta. Posebno je istaknuta razlika između Europe i preostalih dvaju kontinenta. Europa, prema boxplot dijagramu, ima značajno višu potrošnju. Donji kvartil Europe gotovo je veći od gornjeg kvartila preostala dva kontinenta, što znači da preko 70 % europskih automobila troši jednako ili više goriva od 25 % automobila s najvećom potrošnjom u Aziji i Sjevernoj Americi.

U uzorku automobila iz Azije i Europe postoji mali broj stršćih vrijednosti, što može utjecati na raspodjelu podataka, ali ne mijenja osnovni zaključak o razlikama među grupama.

Kako bismo izabrali kojim testom možemo testirati postoji li statistički značajna razlika u gradskoj potrošnji automobila između različitih kontinenata proizvođača potrebno je provjeriti normalnost podataka. To je učinjeno sljedećim Q-Q dijagramima.

```
par(mfrow = c(1, 3))

# Q-Q plot za Europu
qqnorm(my_cars$city.L.100km[my_cars$continent == "Europe"],
       main = "Q-Q Plot: Europa",
       xlab = "Teorijski kvantili",
```

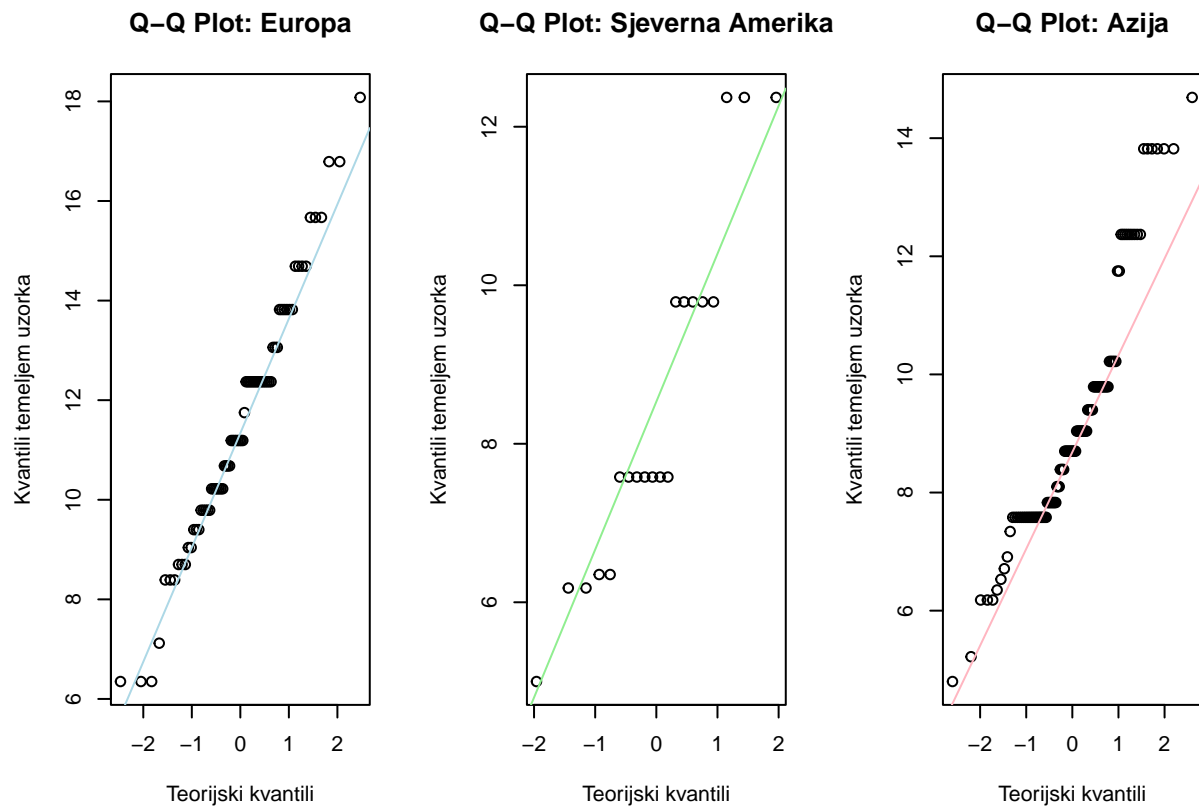
```

        ylab = "Kvantili temeljem uzorka")
qqline(my_cars$city.L.100km[my_cars$continent == "Europe"], col = "lightblue")

# Q-Q plot za Sjevernu Ameriku
qqnorm(my_cars$city.L.100km[my_cars$continent == "North America"],
        main = "Q-Q Plot: Sjeverna Amerika",
        xlab = "Teorijski kvantili",
        ylab = "Kvantili temeljem uzorka")
qqline(my_cars$city.L.100km[my_cars$continent == "North America"], col = "lightgreen")

# Q-Q plot za Aziju
qqnorm(my_cars$city.L.100km[my_cars$continent == "Asia"],
        main = "Q-Q Plot: Azija",
        xlab = "Teorijski kvantili",
        ylab = "Kvantili temeljem uzorka")
qqline(my_cars$city.L.100km[my_cars$continent == "Asia"], col = "lightpink")

```



Q-Q dijagrami prikazuju značajna odstupanja od normalne distribucije za proizvođače iz Sjeverne Amerike i Azije. Zbog toga moramo koristiti test koji ne pretpostavlja normalnost podataka. Zato koristimo Kruskal-Wallisov test koji je neparametarska alternativa ANOVA testu.

H0: Ne postoji razlika u distribuciji gradske potrošnje goriva između Europe, Azije i Sjeverne Amerike.

H1: Postoji razlika u distribuciji gradske potrošnje goriva između Europe, Azije i Sjeverne Amerike.

Odabrana razina značajnosti: $\alpha=0.05$

```
my_cars$continent <- as.factor(my_cars$continent)
kruskal.test(city.L.100km ~ continent, data = my_cars)

##
##   Kruskal-Wallis rank sum test
##
## data:   city.L.100km by continent
## Kruskal-Wallis chi-squared = 49.079, df = 2, p-value = 2.201e-11
```

Zbog izrazito niske p-vrijednosti (< 0.05) Kruskal-Wallisovog testa zaključujemo da ova tri skupa podataka ne proističu iz iste distribucije, tj. da postoji razlika u gradskoj potrošnji između Azije, Europe i Sjeverne Amerike.

Kako bismo dodatno usporedili pojedine parove kontinenata provodimo Mann-Whitney-Wilcoxonov test.

Zbog višestrukih parnih usporedbi, primijenjena je Bonferronijeva korekcija kako bi se kontrolirala ukupna razina značajnosti (Alpha se dijeli s brojem usporedbi). Odabrana razina značajnosti za Bonferroni korekciju:

$$\alpha_{\text{Bonferroni}} = \frac{0.05}{3} \approx 0.0167$$

```
pairwise.wilcox.test(my_cars$city.L.100km,
                     my_cars$continent,
                     p.adjust.method = "bonferroni")

##
##   Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:   my_cars$city.L.100km and my_cars$continent
##
##               Asia      Europe
## Europe          1.9e-10 -
## North America 0.37      2.7e-05
##
## P value adjustment method: bonferroni
```

Zbog visoke p-vrijednosti pri usporedbi gradske potrošnje Azije i Sjeverne Amerike ne možemo odbaciti mogućnost da te dvije potrošnje proističu iz iste distribucije. Ostale kombinacije (Azije i Europa te Europa i Sjeverna Amerika) imaju izrazito nisku p-vrijednost, pa možemo zaključiti da proizlaze iz različitih distribucija.

Prije odabira konačnog testa, isprobana su još dva pristupa: hi-kvadrat test, pri čemu su podaci podijeljeni u tri skupine po potrošnji, te ANOVA test nakon logaritamske pretvorbe podataka. Kruskal-Wallisov test pokazuje se primjerenijim od hi-kvadrat testa kod kontinuiranih podataka, jer ne zahtijeva proizvoljna podjela podataka. Logaritamska transformacija nije uspjela postići normalnost podataka za Aziju, što dodatno opravdava primjenu neparametarskog Kruskal-Wallisova testa.