Potrošnja goriva po regijama

Ivan Skukan

#Pitanje **Postoje li razlike u potrošnji automobila prema regiji kojoj pripada proizvođač?**

#Uvod Da odgovorimo na ovo pitanje, moramo analizirati podatke potrošnje goriva na 3 kontinenta. Zbog činjenice da imamo više od 2 regije, analiza varijance (ANOVA) će biti naš odabir modeliranja umjesto t-testa, ali prije toga ćemo morati testirati uvjete ANOVA-e.

Učitavanje podataka:

```
path <- "car_specifications.csv"
data <- read.csv(path)

data$continent = as.factor(data$continent)
data$country = as.factor(data$country)
head(data)
```

```
##         make aspiration num.of.doors  body.style drive.wheels engine.location
## 1 Alfa Romeo        std          two convertible          rwd           front
## 2 Alfa Romeo        std          two convertible          rwd           front
## 3 Alfa Romeo        std          two   hatchback          rwd           front
## 4       Audi        std         four       sedan          fwd           front
## 5       Audi        std         four       sedan          4wd           front
## 6       Audi        std          two       sedan          fwd           front
##   wheel.base length width height curb.weight engine.type num.of.cylinders
## 1      225.0  428.8 162.8  124.0        1156        dohc             four
## 2      225.0  428.8 162.8  124.0        1156        dohc             four
## 3      240.0  434.8 166.4  133.1        1280        ohcv              six
## 4      253.5  448.6 168.1  137.9        1060         ohc             four
## 5      252.5  448.6 168.7  137.9        1281         ohc             five
## 6      253.5  450.3 168.4  134.9        1137         ohc             five
##   engine.size fuel.system bore stroke compression.ratio horsepower peak.rpm
## 1        2130        mpfi 8.81   6.81               9.0        111     5000
## 2        2130        mpfi 8.81   6.81               9.0        111     5000
## 3        2491        mpfi 6.81   8.81               9.0        154     5000
## 4        1786        mpfi 8.10   8.64              10.0        102     5500
## 5        2229        mpfi 8.10   8.64               8.0        115     5500
## 6        2229        mpfi 8.10   8.64               8.5        110     5500
##   price city.L.100km highway.L.100km   fuel country continent
## 1 13495        11.19            8.70 petrol   Italy    Europe
## 2 16500        11.19            8.70 petrol   Italy    Europe
## 3 16500        12.37            9.04 petrol   Italy    Europe
## 4 13950         9.79            7.83 petrol Germany    Europe
## 5 17450        13.06           10.68 petrol Germany    Europe
## 6 15250        12.37            9.40 petrol Germany    Europe
```

Nama su relevantni stupci 'city.L.100km', 'highway.L.100km' i 'continent'

#Usporedba sredina ***Provjera aritmetičkih sredina*** Prije nego krenemo sa ANOVA-om, možemo prvo

usporediti aritmetičke sredine podataka. Ovo nije dovoljno da radimo bilo kakve konkretne zaključke, ali nam daje uvid u što bi možda očekivali. Također ćemo izračunati varijancu i standardnu devijaciju da imamo bolju ideju o izgledu raspršenosti podataka. Provjerimo sredine za pojedine kontinente sa boxplotom i sveukupnu sredinu i varijance kroz ispis:

```
continents = unique(data$continent) #lista svih kontinenta

overallMeanCity = mean(data$city.L.100km)
overallMeanHighway = mean(data$highway.L.100km)
overallVarCity = var(data$city.L.100km)
overallVarHighway = var(data$highway.L.100km)

cat("\n")
```

```
print(sprintf("Aritmetička sredina potrošnje goriva u gradu: %f",overallMeanCity))
```

```
## [1] "Aritmetička sredina potrošnje goriva u gradu: 9.943582"
```

```
print(sprintf("Aritmetička sredina potrošnje goriva na autocestama: %f",overallMeanHighway))
```

```
## [1] "Aritmetička sredina potrošnje goriva na autocestama: 8.043433"
```

```
print(sprintf("Sveukupna varijanca za gradove: %f",overallVarCity))
```

```
## [1] "Sveukupna varijanca za gradove: 6.429238"
```

```
print(sprintf("Standardna devijacija: %f",sqrt(overallVarCity)))
```

```
## [1] "Standardna devijacija: 2.535594"
```

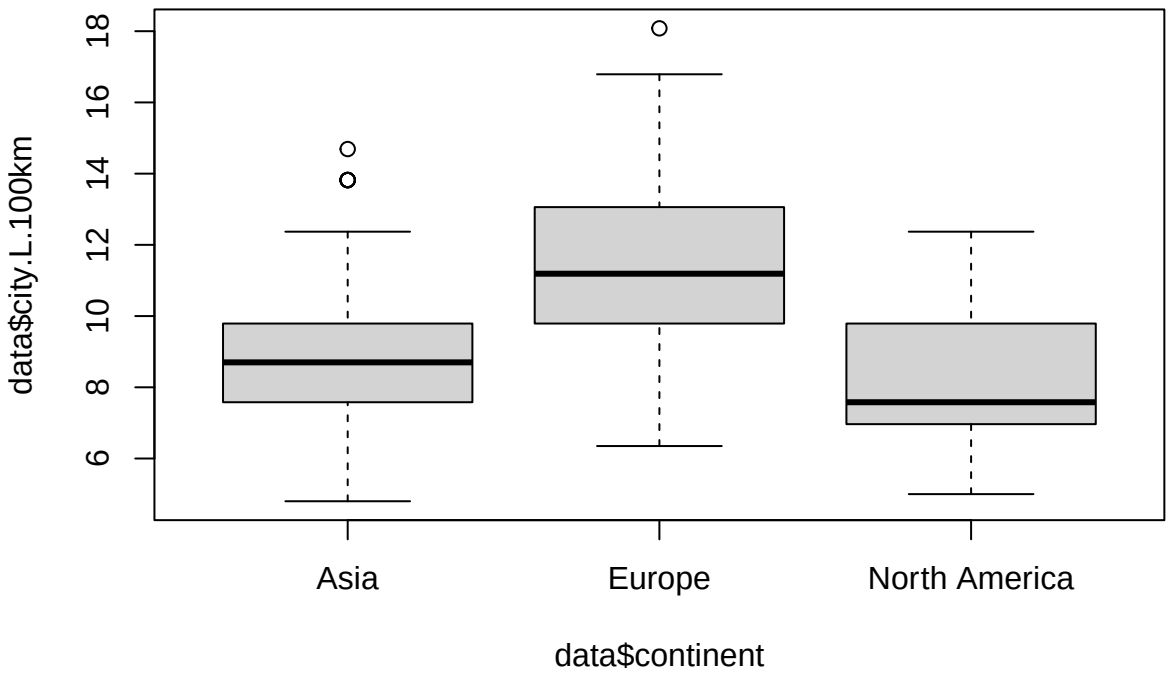
```
print(sprintf("Sveukupna varijanca za autoceste: %f",overallVarHighway))
```

```
## [1] "Sveukupna varijanca za autoceste: 3.390360"
```

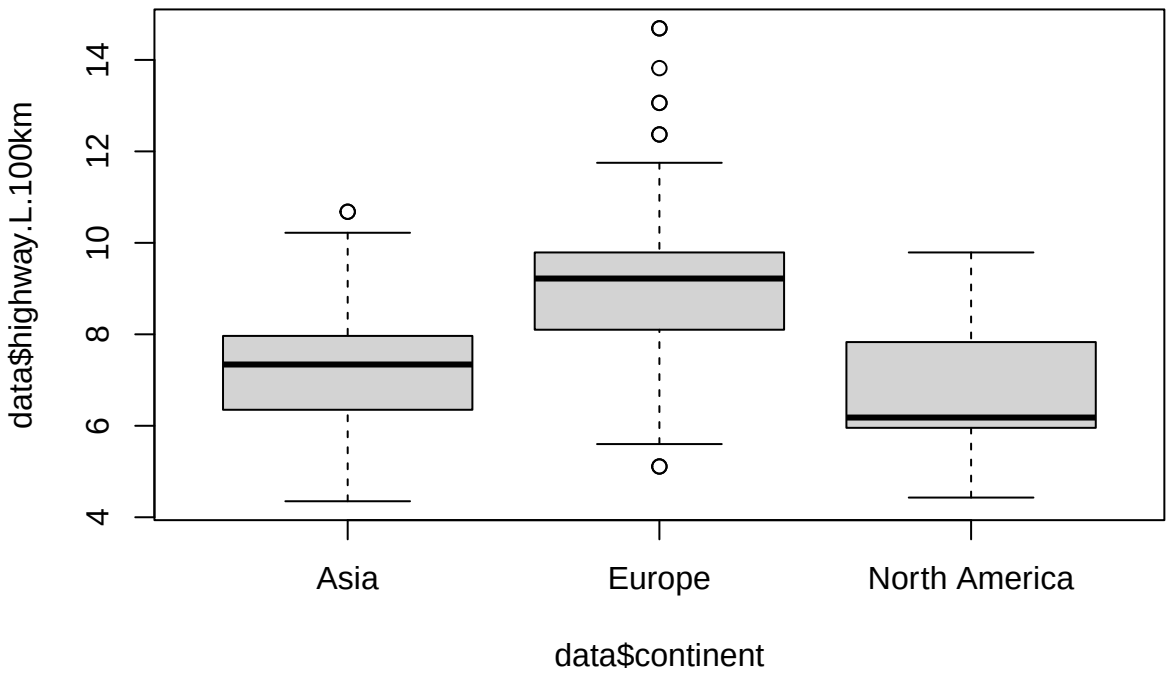
```
print(sprintf("Standardna devijacija: %f",sqrt(overallVarHighway)))
```

```
## [1] "Standardna devijacija: 1.841293"
```

```
boxplot(data$city.L.100km ~ data$continent)
```




```
boxplot(data$highway.L.100km ~ data$continent)
```



Već vidimo da je potrošnja u Europi u prosjeku veća nego druga dva kontinenta. Osim toga vidimo da je i disperzija nešto veća.

Uvjeti za ANOVA-u Podsjetimo se. Želimo testirati ima li značajno odstupanje u sredinama potrošnje goriva na 3 kontinenta i zato prirodno biramo ANOVA-u. Kako bi proveli ANOVA test na podacima, prvo moramo biti sigurni da dani podaci zadovoljavaju sljedeće uvjete: 1. Normalnost 2. Nezavisnost 3. Homogenost varijanci

#lillie **Testiranje normalnosti** Prvo ćemo proveti test normalnosti. Zanimaju nas tablice za potrošnju goriva u gradu i na autocesti za svaki kontinent. Koristit ćemo Lilliefors test koji se temelji na Kolmogorov-Smirnov testu i Q-Q plot za vizualizaciju. Postavljamo hipoteze:

$$H0 : Dani\ podaci\ za\ potrošnju\ goriva\ imaju\ normalnu\ distribucijuH1 : \neg H0$$

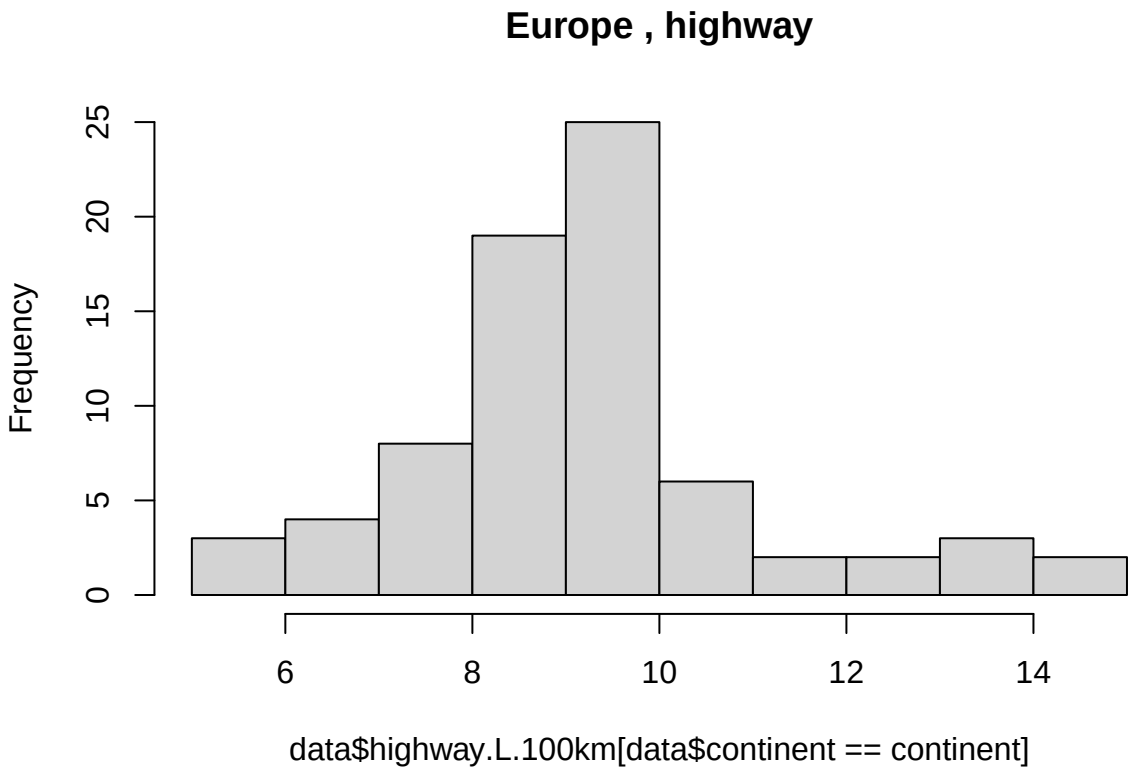
Provedimo test:

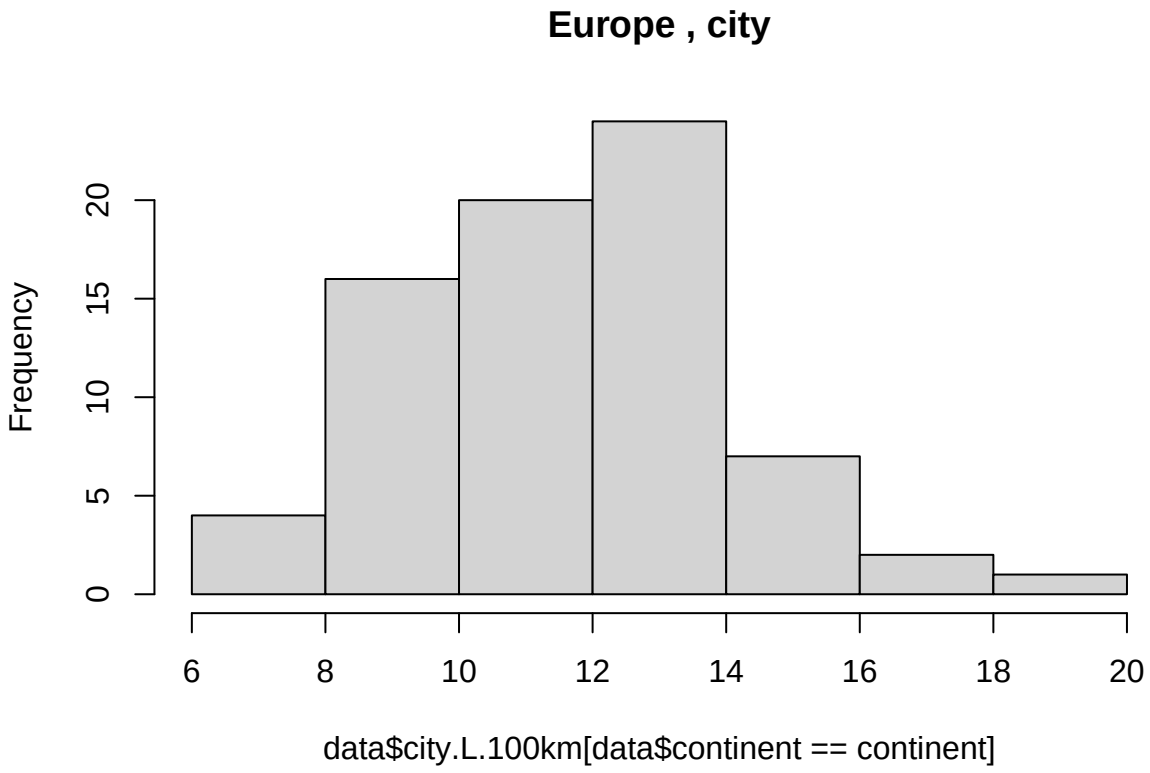
```
require(nortest) #potrebna biblioteka
```

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

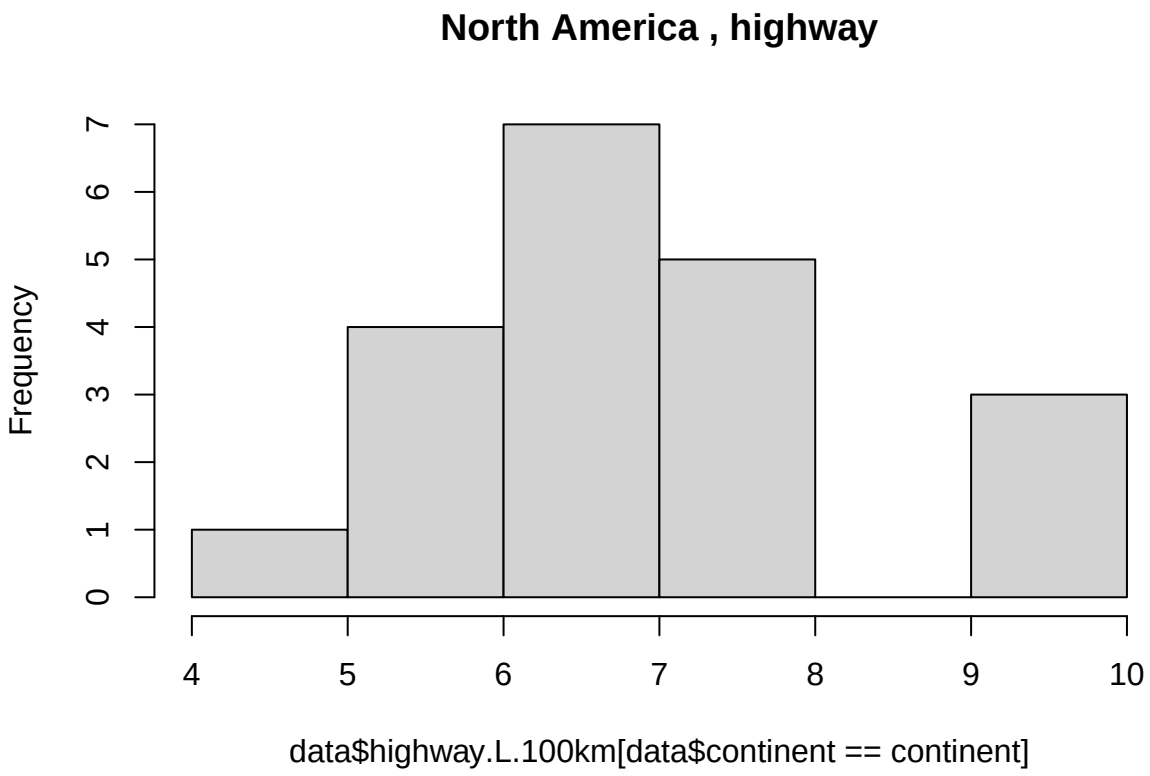
```
for (continent in continents) {
  print(paste("For continent:",continent))
  print(lillie.test(data$city.L.100km[data$continent == continent]))
  print(lillie.test(data$highway.L.100km[data$continent == continent]))
  titleHighway = paste(continent,", highway")
  titleCity = paste(continent,", city")
  hist(data$highway.L.100km[data$continent == continent],main=titleHighway)
  hist(data$city.L.100km[data$continent == continent],main=titleCity)
}
```

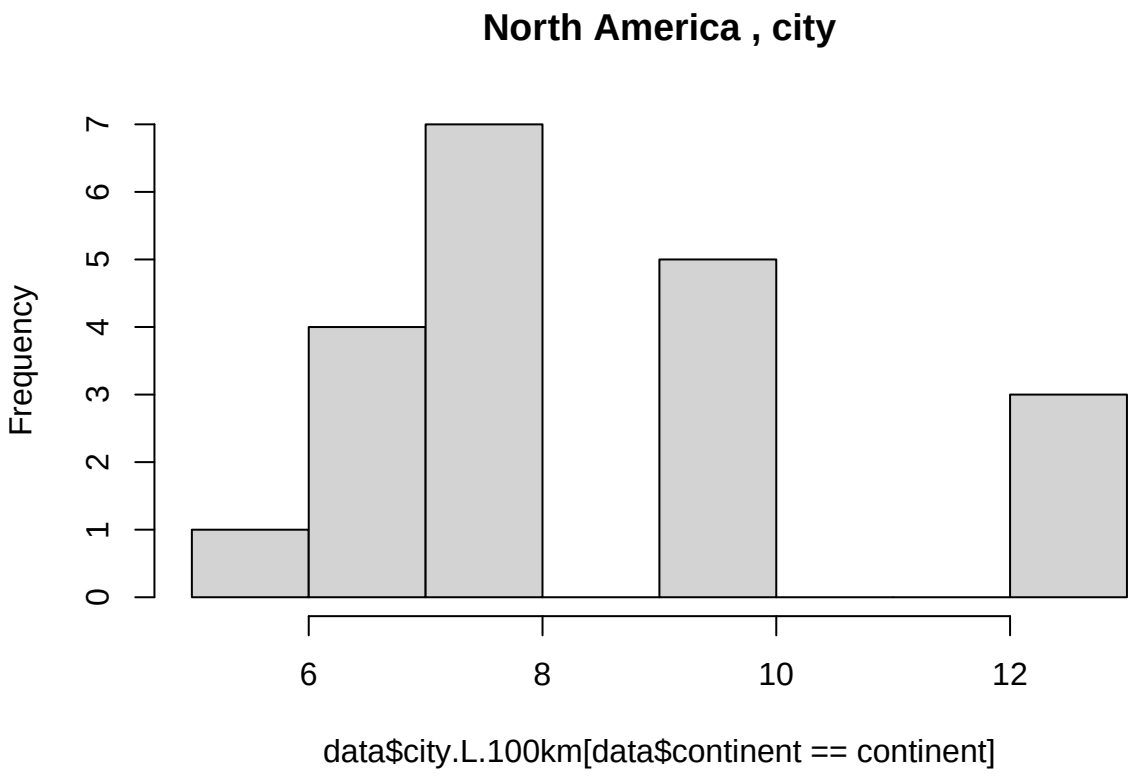
```
## [1] "For continent: Europe"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$city.L.100km[data$continent == continent]
## D = 0.11082, p-value = 0.02499
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$highway.L.100km[data$continent == continent]
## D = 0.17288, p-value = 9.739e-06
```



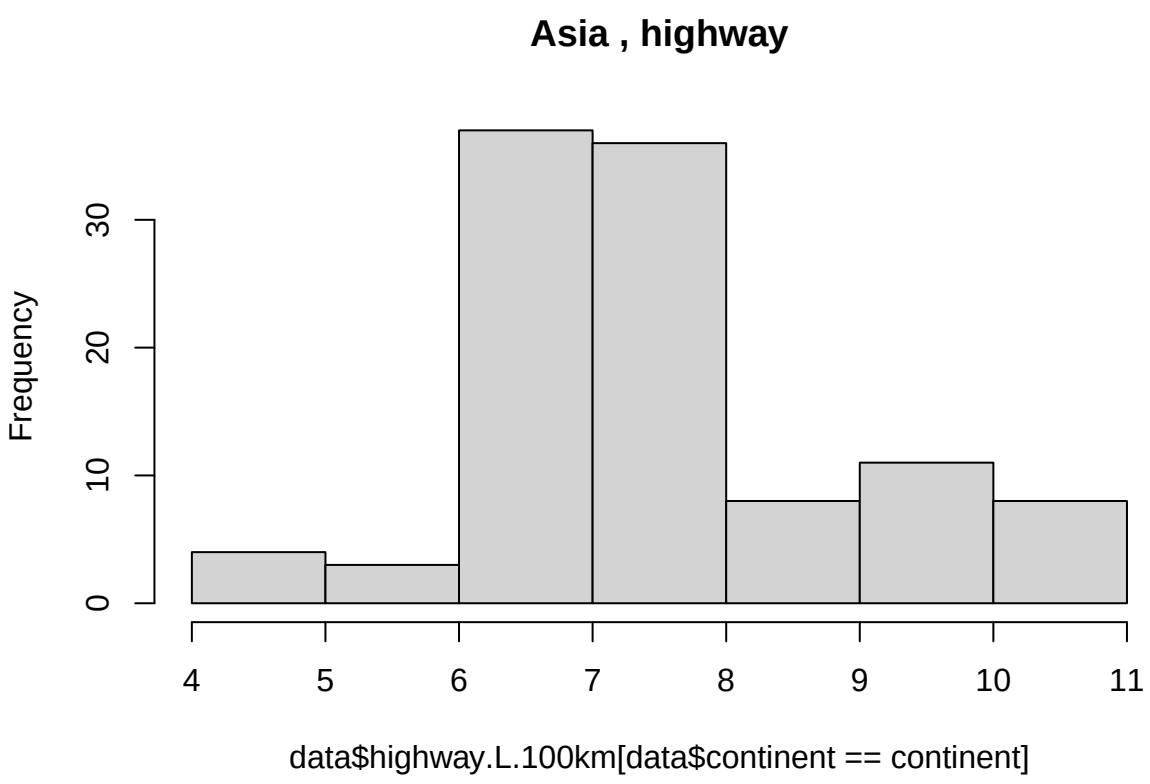


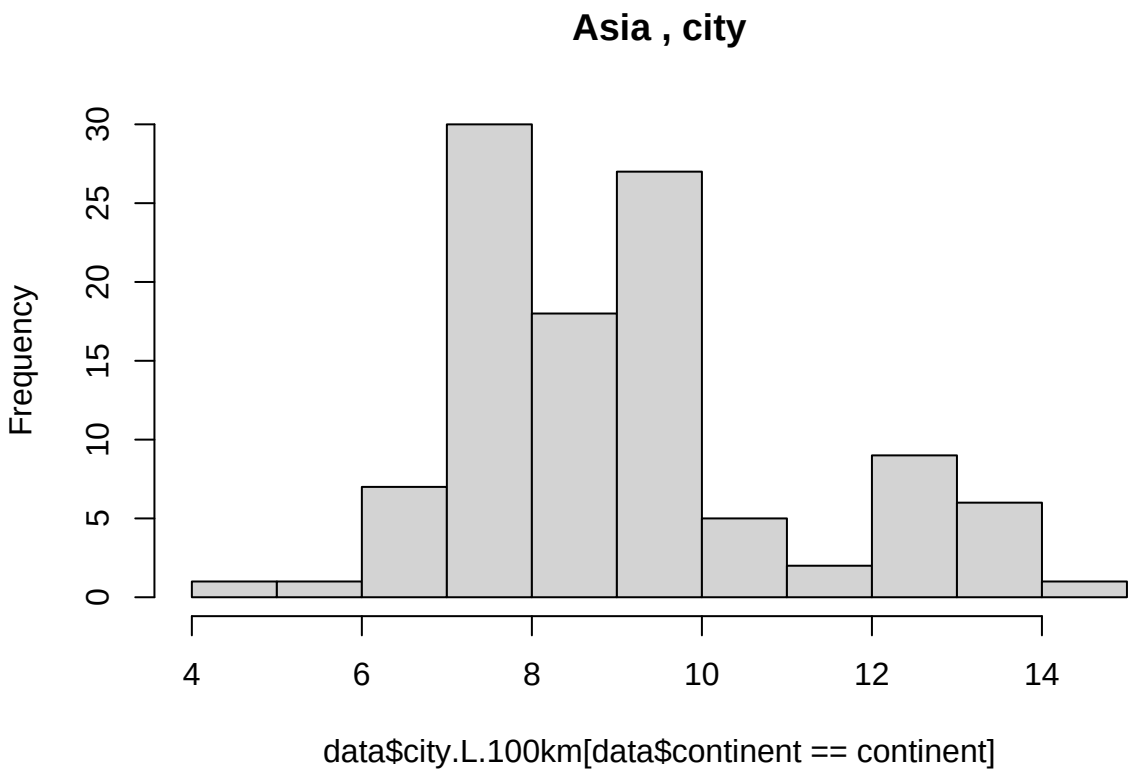

```
## [1] "For continent: North America"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$city.L.100km[data$continent == continent]
## D = 0.2557, p-value = 0.001348
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$highway.L.100km[data$continent == continent]
## D = 0.28537, p-value = 0.0001587
```


North America , highway
Frequency
0
1
2
3
4
5
6
7
4
5
6
7
8
9
10
data$highway.L.100km[data$continent == continent]



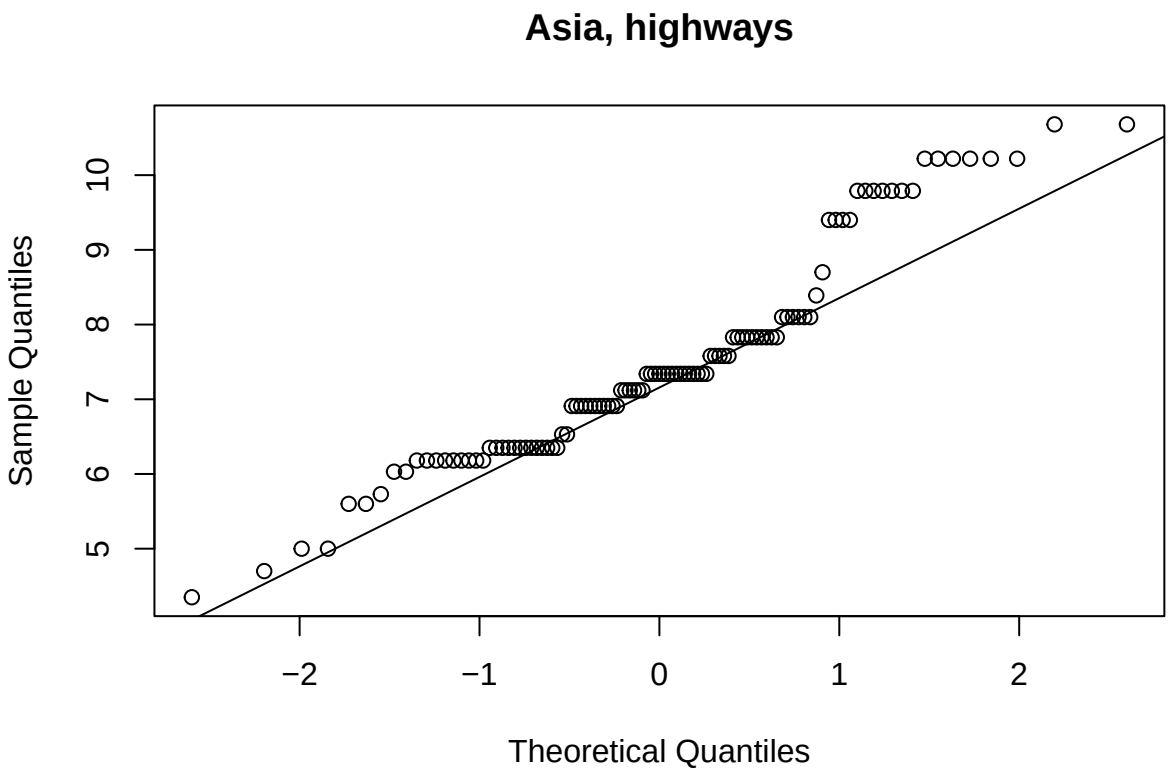

```
## [1] "For continent: Asia"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$city.L.100km[data$continent == continent]
## D = 0.15718, p-value = 7.516e-07
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$highway.L.100km[data$continent == continent]
## D = 0.14449, p-value = 9.635e-06
```


Asia , highway
Frequency
0
10
20
30
4
5
6
7
8
9
10
11
data$highway.L.100km[data$continent == continent]

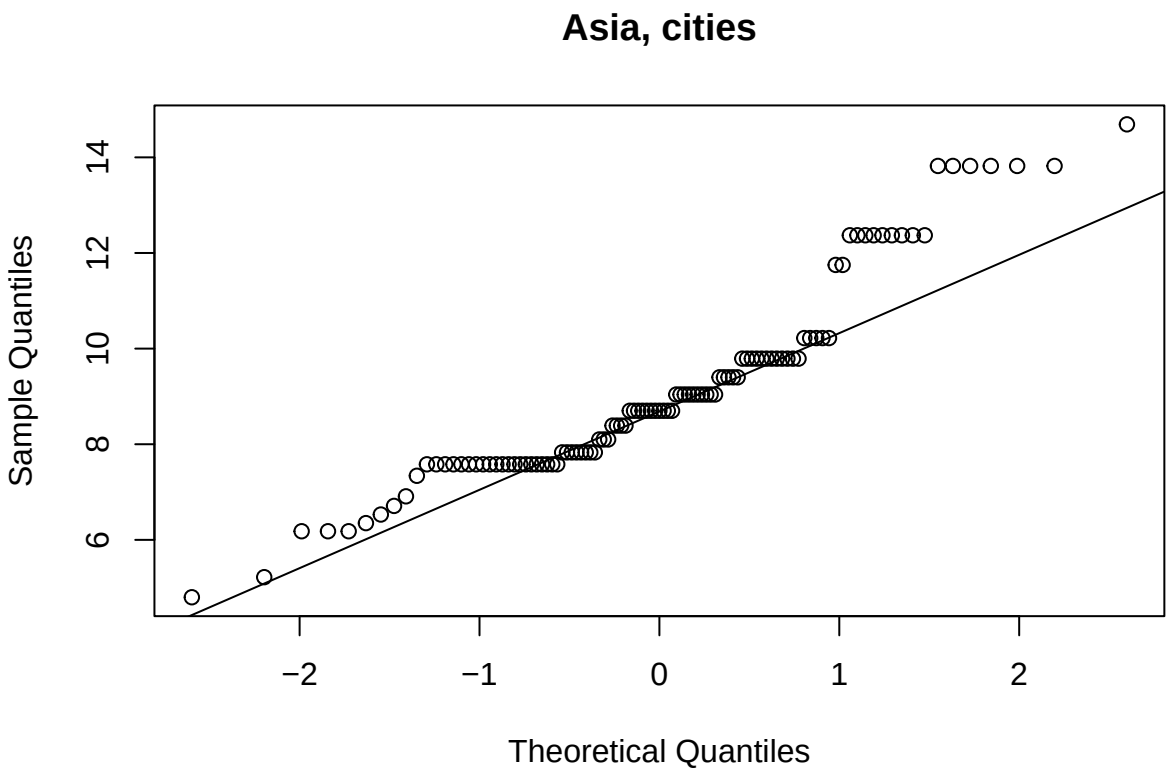


Dobili smo vrlo male p-vrijednosti. Sve su signifikantne na barem 0.05 razini, a većina je i ekstremnije od toga. Ovime možemo uvjereno odbaciti H0 i zaključiti kako podaci nisu normalno distribuirani. Osim toga, histogrami očito pokazuju kako podaci ne prate Gaussovu krivulju. Manjak normalnosti možemo vidjeti i na Q-Q plotu:

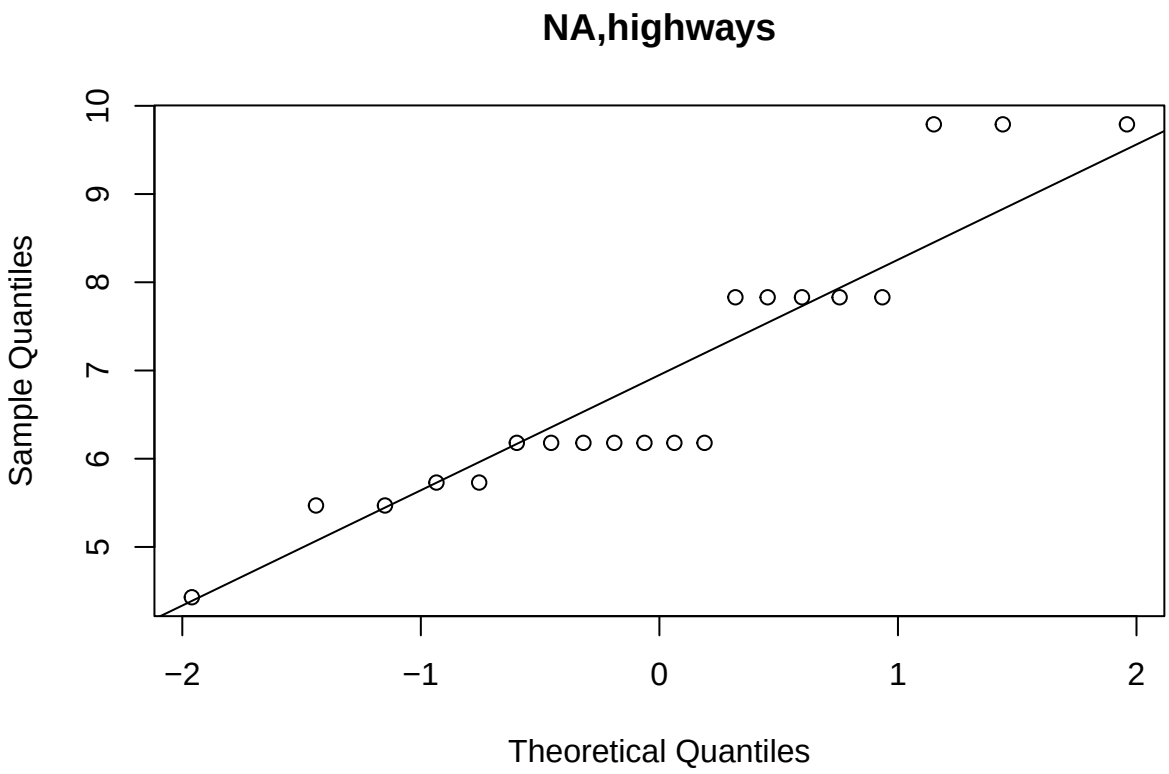
```
qqnorm(data$highway.L.100km[data$continent == "Asia"],main="Asia, highways")
qqline(data$highway.L.100km[data$continent == "Asia"])
```



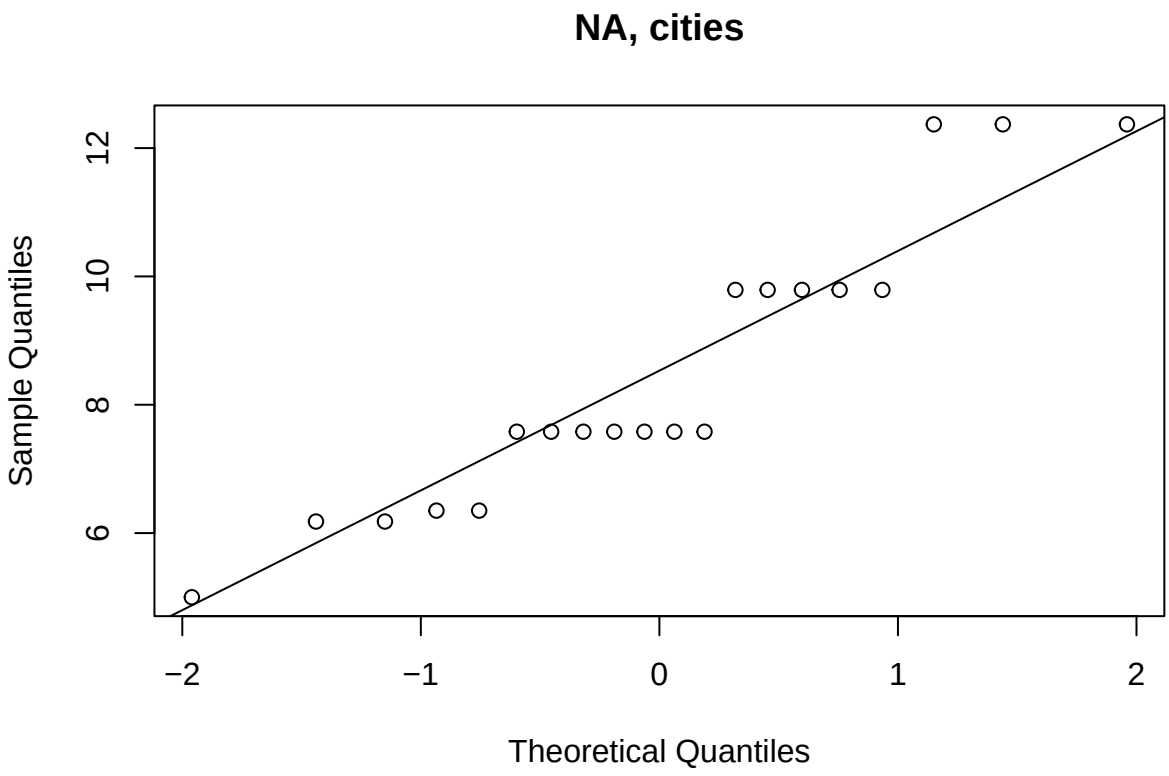

```
qqnorm(data$city.L.100km[data$continent == "Asia"],main = "Asia, cities")
qqline(data$city.L.100km[data$continent == "Asia"])
```



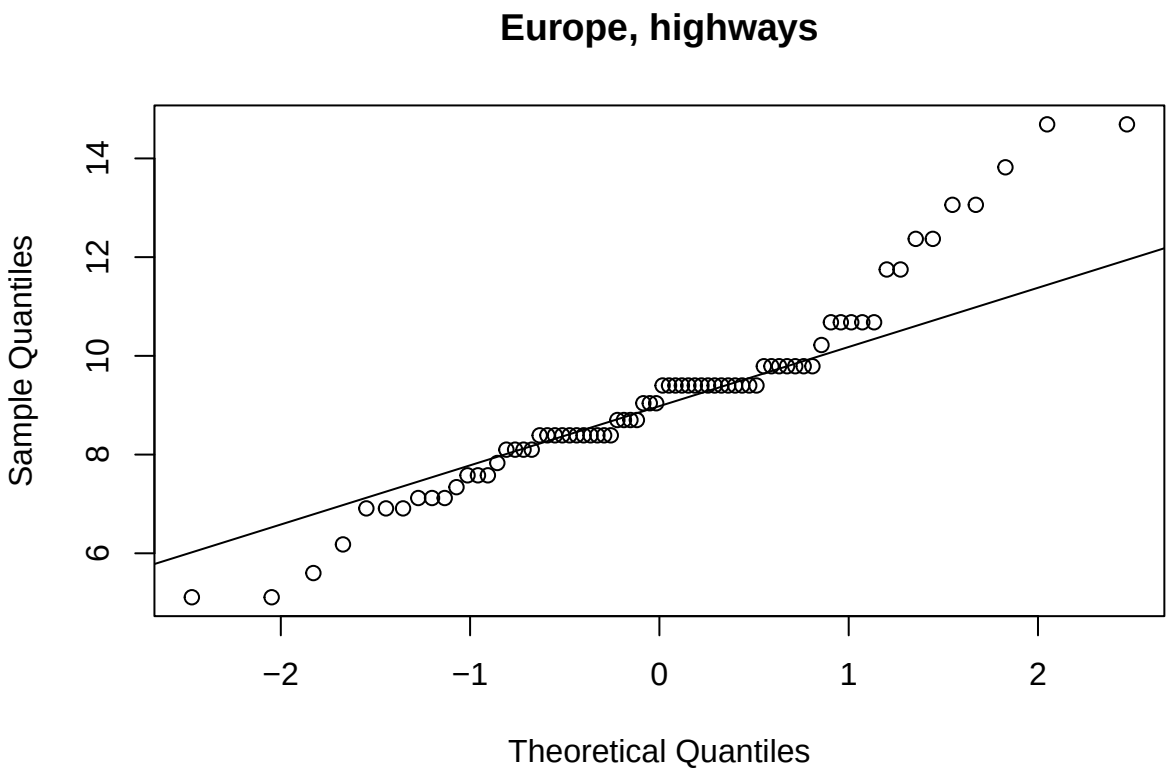

```
qqnorm(data$highway.L.100km[data$continent == "North America"],main = "NA,highways")
qqline(data$highway.L.100km[data$continent == "North America"])
```



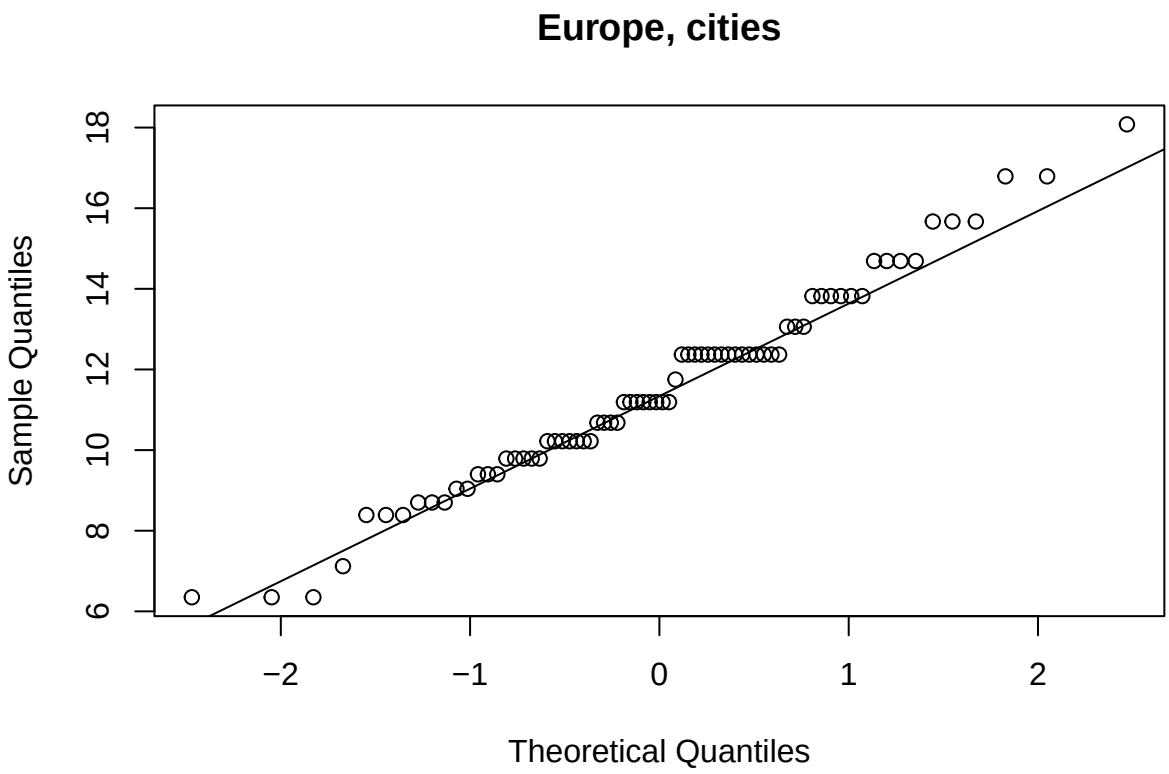

```
qqnorm(data$city.L.100km[data$continent == "North America"],main = "NA, cities")
qqline(data$city.L.100km[data$continent == "North America"])
```




```
qqnorm(data$highway.L.100km[data$continent == "Europe"],main = "Europe, highways")
qqline(data$highway.L.100km[data$continent == "Europe"])
```




```
qqnorm(data$city.L.100km[data$continent == "Europe"],main = "Europe, cities")
qqline(data$city.L.100km[data$continent == "Europe"])
```



Budući da nemamo zadovoljen uvjet normalnosti, ne možemo koristit ANOVA-u.

Kruskal-Wallis test Okrećemo se alternativi ANOVA testa, Kruskal-Wallis test. To je neparametarski test pa ne trebamo da nam podaci prate određenu distribuciju. Jedini uvjet je da je broj podataka barem 5. Hipoteze se postavljaju na isti način kao u ANOVA-i.

Hipoteze:

$H0:$ *Očekivana vrijednost potrošnje goriva po kontinentima je jednaka*$H1:$ *Barem jedna očekivana vrijednost se razlik*

Ovu istu hipotezu postavljamo za gradove i autoceste te zbog toga provodimo test dvaput.

Provedimo Kruskal-Wallis test:

```
filteredDataHighway = list(Asia_Highway = data$highway.L.100km[data$continent == "Asia"], NA_Highway = c
filteredDataCity = list(Asia_City = data$city.L.100km[data$continent == "Asia"], NA_City = data$city.L.1

kruskalHighway = kruskal.test(filteredDataHighway)
kruskalCity = kruskal.test(filteredDataCity)
print(kruskalHighway)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  filteredDataHighway
## Kruskal-Wallis chi-squared = 46.203, df = 2, p-value = 9.272e-11
```

Provjerimo prvo rezultat za potrošnju na autocestama. Dobili smo vrlo malu p-vrijednost, praktički je jednaka nuli, dakle bez sumnje odbacujemo H0 i zaključujemo da se barem jedna očekivana vrijednost razlikuje od ostalih.

Provjerimo sad rezultat za gradove:

```
print(kruskalCity)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  filteredDataCity
## Kruskal-Wallis chi-squared = 49.079, df = 2, p-value = 2.201e-11
```

Opet imamo p-vrijednost koja je efektivno jednaka nuli. Za potrošnju u gradovima također odbacujemo H0.

Na temelju boxplotova koje smo ranije vidjeli, najvjerojatnije su podaci za Europu zaslužni za odbijanje H0. Možemo koristiti Dunnov test da vidimo između kojih grupa su razlike značajne. Dunnov test ima istu hipotezu kao Kruskal-Wallis test.

```
require(dunn.test)
```

```
## Loading required package: dunn.test
```

```
dunn = dunn.test(data$highway.L.100km , g = data$continent,method="bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 46.2028, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       Asia     Europe
## ---------+----------------------
##   Europe |  -6.059794
##          |    0.0000*
##          |
## North Am |   1.440839   5.028198
##          |     0.2244    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

```
print(dunn)
```

```
## $chi2
## [1] 46.20277
##
## $Z
## [1] -6.059794  1.440840  5.028199
##
## $P
## [1] 6.814784e-10 7.481500e-02 2.475542e-07
##
## $P.adjusted
## [1] 2.044435e-09 2.244450e-01 7.426625e-07
##
## $comparisons
## [1] "Asia - Europe"          "Asia - North America"   "Europe - North America"
```

```r
dunn = dunn.test(data$city.L.100km, g = data$continent, method = "bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 49.0792, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |       Asia     Europe
## ---------+----------------------
##   Europe |  -6.382981
##          |    0.0000*
##          |
## North Am |   1.173223   4.963401
##          |     0.3611    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Zanima nas "p-adjusted" redak koji se također može vidjeti u tablici(drugi redak u svakoj čeliji). Za oba testa imamo isti zaključak Vidimo da je p-adjusted za North_America-Asia relativno velik i nije signifikantan na niti jednoj tipičnoj razini. Međutim, imamo vrlo male p-adjusted vrijednosti između Europe i bilo kojeg drugog kontinenta. Ovime možemo zaključiti da je odbijanje nul hipoteze Kruskal-Wallis testa bilo primarno zbog razine potrošnje u Europi. #pitaj za dunnov test!!!!

#subregionalno **Testiranje među regijama u Europi** Budući da Azija i Sjeverna Amerika imaju relativno slične potrošnje, ne zanima nas detaljnije subregionalno testiranje. Osim toga, ti kontinenti imaju samo po jednu državu u podacima, dakle ni ne možemo podjeliti na manje regije. Međutim, Europu, koja je imala poprilično veliku potrošnju goriva i mnogo država, možemo podijeliti na manje regije, no ne možemo testirati na pojedinim državama jer za neke nemamo dovoljno podataka. Opet ćemo napraviti Kruskal-Wallis test i, po potrebi, Dunnov test. Ovime možemo probati zaključiti ako se u nekim regijama više troši.

Podjela na subregije Trebamo odrediti kako želimo grupirati države. Ciljat ćemo na podjelu koja otprilike dijeli Europu na zapadnu, sjevernu i središnju Europu. Budući da za Italiju i UK nemamo dovoljno podataka, grupirat ćemo ih sa Francuskom i Švedskom respektivno. Francuska i Italija će predstavljati zapadnu, UK i Švedska sjevernu, a Njemačka središnju Europu.

```r
westEuRegions = c("France","Italy")
northEuRegions = c("United Kingdom","Sweden")
centralEuRegions = c("Germany")

westEuData = subset(data, country %in% westEuRegions, select = c(highway.L.100km,city.L.100km,country))
northEuData = subset(data, country %in% northEuRegions, select = c(highway.L.100km,city.L.100km,country
centralEuData = subset(data, country %in% centralEuRegions, select = c(highway.L.100km,city.L.100km,cou

europeDataHighway = list(west = westEuData$highway.L.100km, north = northEuData$highway.L.100km, centra
europeDataCity = list(west = westEuData$city.L.100km, northEuData$city.L.100km, central = centralEuData
```

Hipoteze su na istu logiku:

$H0:$ *Očekivana vrijednost potrošnje goriva po podregijama je jednaka*$H1:$ *Barem jedna očekivana vrijednost se razliku*

Provedimo testove:

```
euKruskalHighway = kruskal.test(europeDataHighway)
euKruskalCity = kruskal.test(europeDataCity)

print(euKruskalHighway)
```

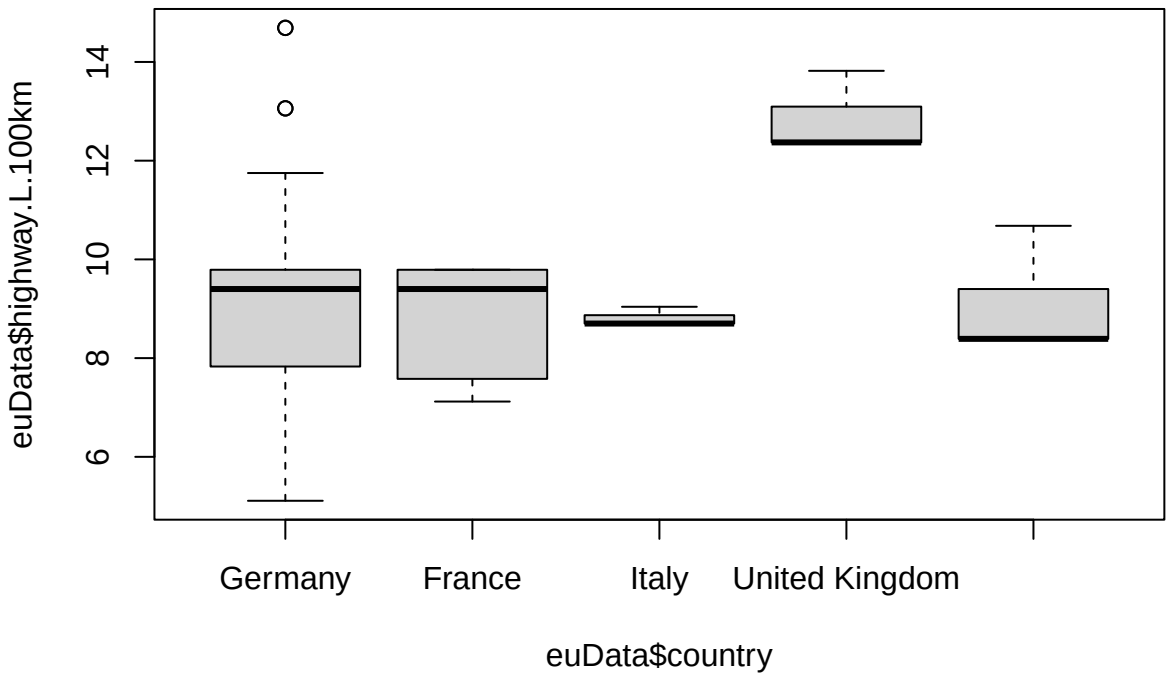
```
##
##  Kruskal-Wallis rank sum test
##
## data:  europeDataHighway
## Kruskal-Wallis chi-squared = 0.62478, df = 2, p-value = 0.7317
```

```
print(euKruskalCity)
```

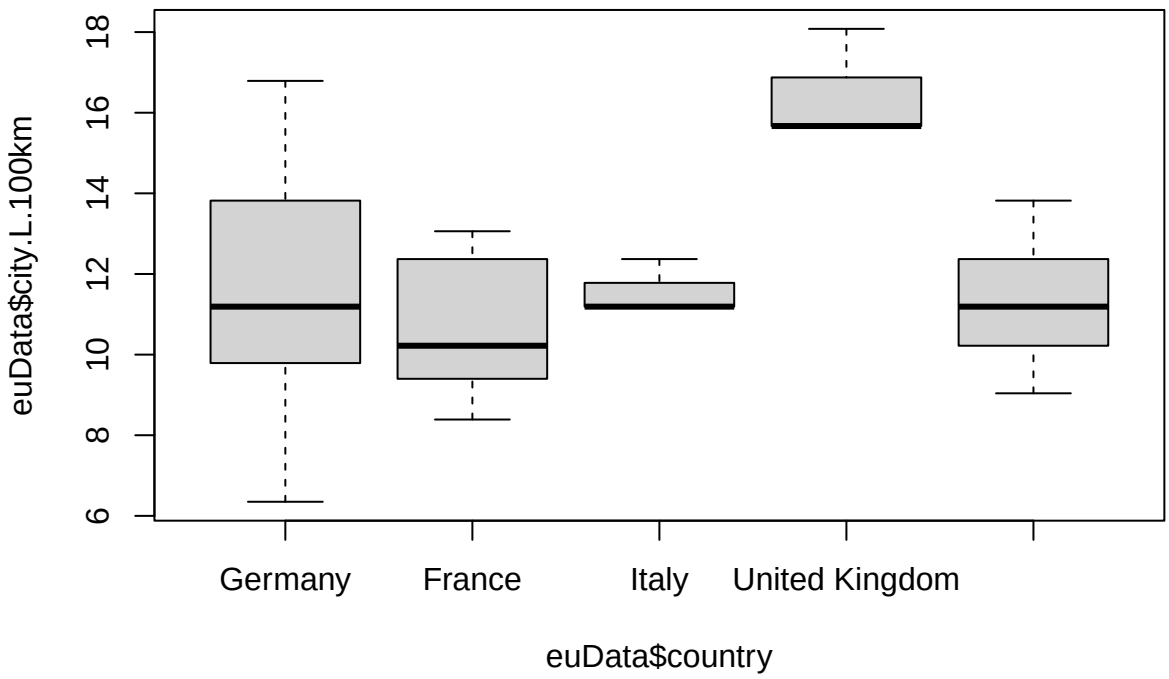
```
##
##  Kruskal-Wallis rank sum test
##
## data:  europeDataCity
## Kruskal-Wallis chi-squared = 2.151, df = 2, p-value = 0.3411
```

Velika p-vrijednost za oba testa nam govori da nema značajne razlike među ovim regijama Europe. Za kraj ćemo napraviti boxplot potrošnje goriva za svaku pojedinu europsku državu.

```
eu = c("Germany","France","Italy","United Kingdom","Sweden")
dataCopy = data
dataCopy$country = factor(data$country, levels=eu)
euData = subset(dataCopy, country %in% eu, select = c(highway.L.100km,city.L.100km,country))
boxplot(euData$highway.L.100km ~ euData$country)
```



```r
boxplot(euData$city.L.100km ~ euData$country)
```



Za neke države imamo jako malo podataka, ali možemo vidjeti da UK značajno odstupa od ostalih država. Vjerojatno jer su u datasetu uključeni Jaguari sa jačim motorima koji više troše. Zaključujemo da nisu određene europske države zaslužne za značajno odstupanje u prosječnoj potrošnji goriva, već vidimo da se generalno više troši, bilo u gradovima bilo na autocestama.