Regression Analysis on Car Specifications

Luka Babić, Dominik Barukčić, Andrija Merlin, Ivan Skukan

2024-01-11

Uvod, motivacija i opis problema

U procesu kupovine novog automobila korisno je razmotriti njihove specifikacije kako bi se donijela ˇsto objektivnija odluka o modelu koji odgovara svim zahtjevima kupca. U tu su svrhu prikupljeni detaljni podatci o modelima 22 proizvoda ˇca automobila razli ˇcitih cjenovnih kategorija.

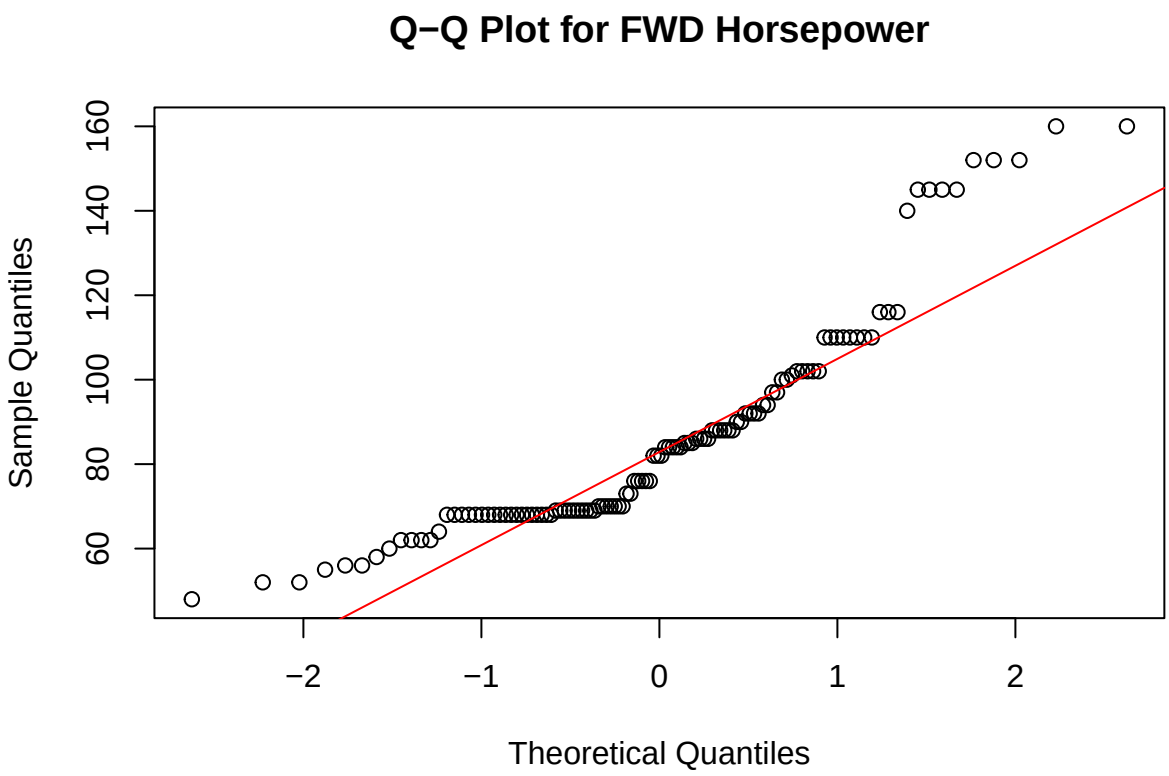
Pitanje 1 - Snaga automobila ovisno o pogonu

Je li snaga automobila s prednjim pogonom ve´ca od automobila s drugim vrstama pogona?

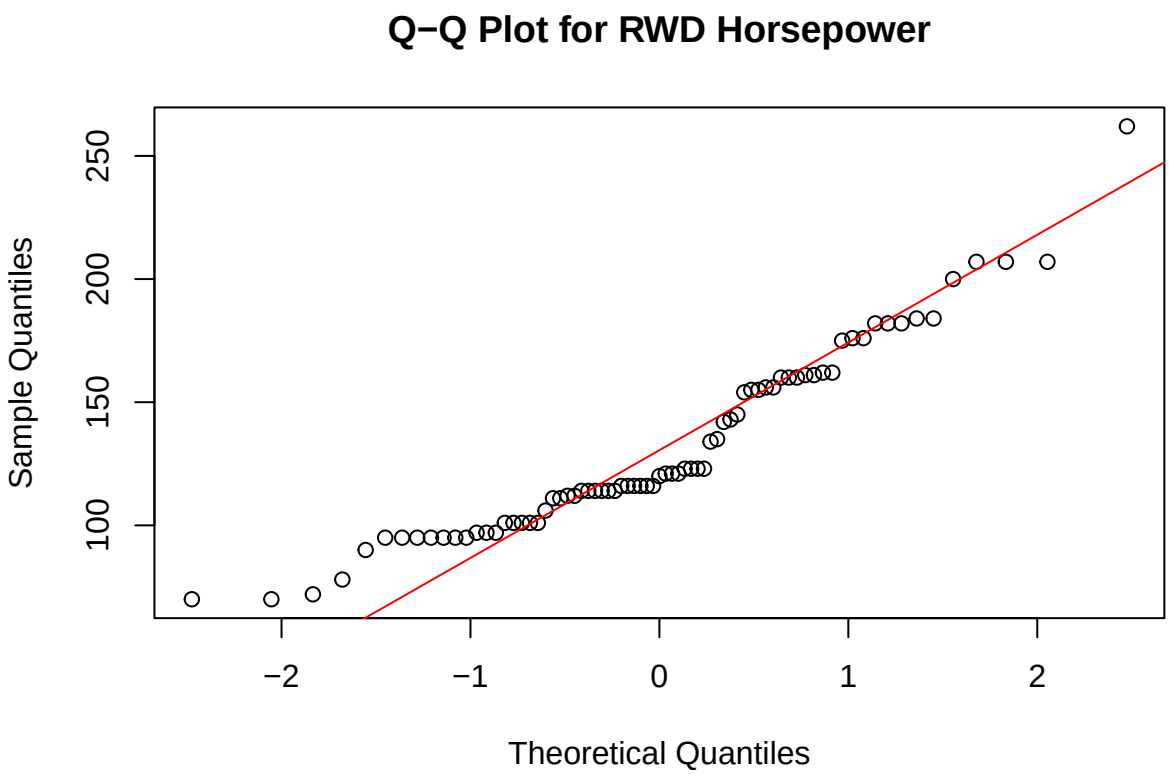
```
# Loading data
path <- "car_specifications.csv"
cardata <- read.csv(path)
```

Provjera normalnosti podataka

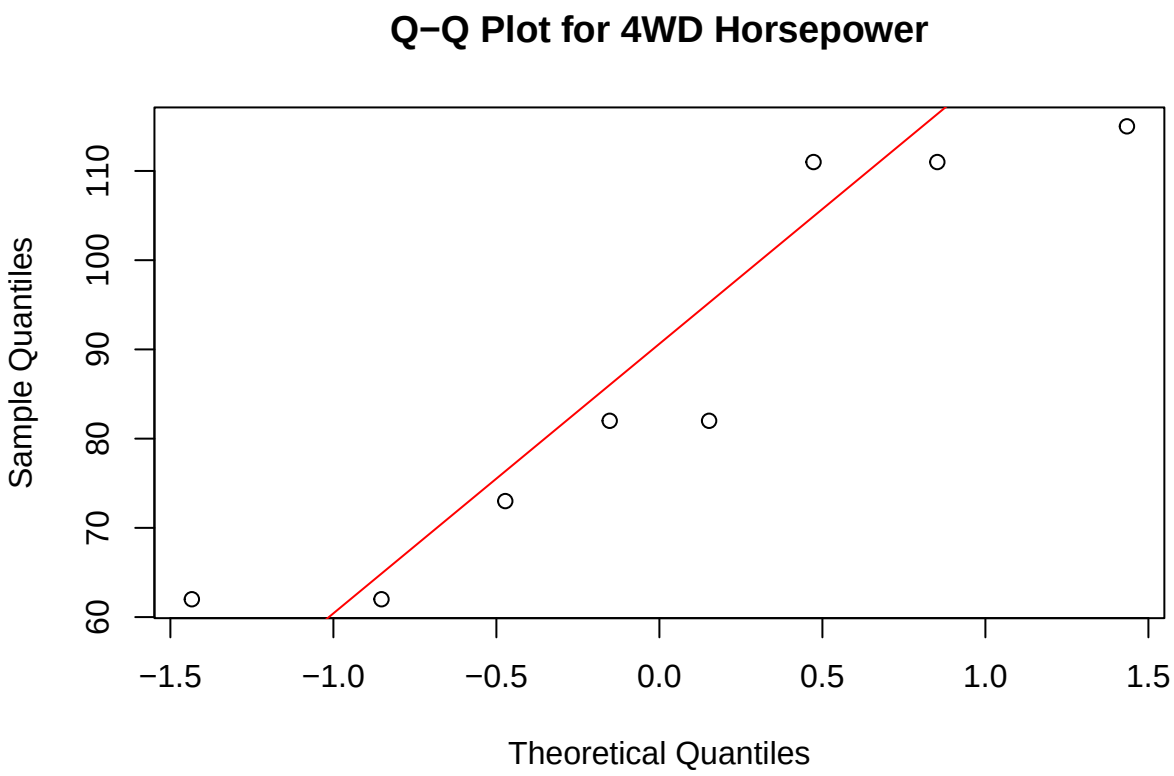
```
# QQ-plots to check the normality of 'horsepower' distribution for different drive types
# QQ-plot for 'horsepower' for front-wheel drive (fwd)

qqnorm(cardata$horsepower[cardata$drive.wheels == 'fwd'], main = "Q-Q Plot for FWD Horsepower")
qqline(cardata$horsepower[cardata$drive.wheels == 'fwd'], col = "red")
```




```r
# QQ-plot for 'horsepower' for rear-wheel drive (rwd)

qqnorm(cardata$horsepower[cardata$drive.wheels == 'rwd'], main = "Q-Q Plot for RWD Horsepower")
qqline(cardata$horsepower[cardata$drive.wheels == 'rwd'], col = "red")
```




```r
# QQ-plot for 'horsepower' for four-wheel drive (4wd)

qqnorm(cardata$horsepower[cardata$drive.wheels == '4wd'], main = "Q-Q Plot for 4WD Horsepower")
qqline(cardata$horsepower[cardata$drive.wheels == '4wd'], col = "red")
```



Kolmogorov-Smirnov test normalnosti

```
# Load package for Lilliefors (Kolmogorov-Smirnov) normality test
require(nortest)
```

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

```
# Normality tests for 'horsepower' across different 'drive.wheels' categories
lillie.test(cardata$horsepower)
lillie.test(cardata$horsepower[cardata$drive.wheels == 'fwd'])
lillie.test(cardata$horsepower[cardata$drive.wheels == 'rwd'])
lillie.test(cardata$horsepower[cardata$drive.wheels == '4wd'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cardata$horsepower
## D = 0.12738, p-value = 2.407e-08
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cardata$horsepower[cardata$drive.wheels == "fwd"]
```

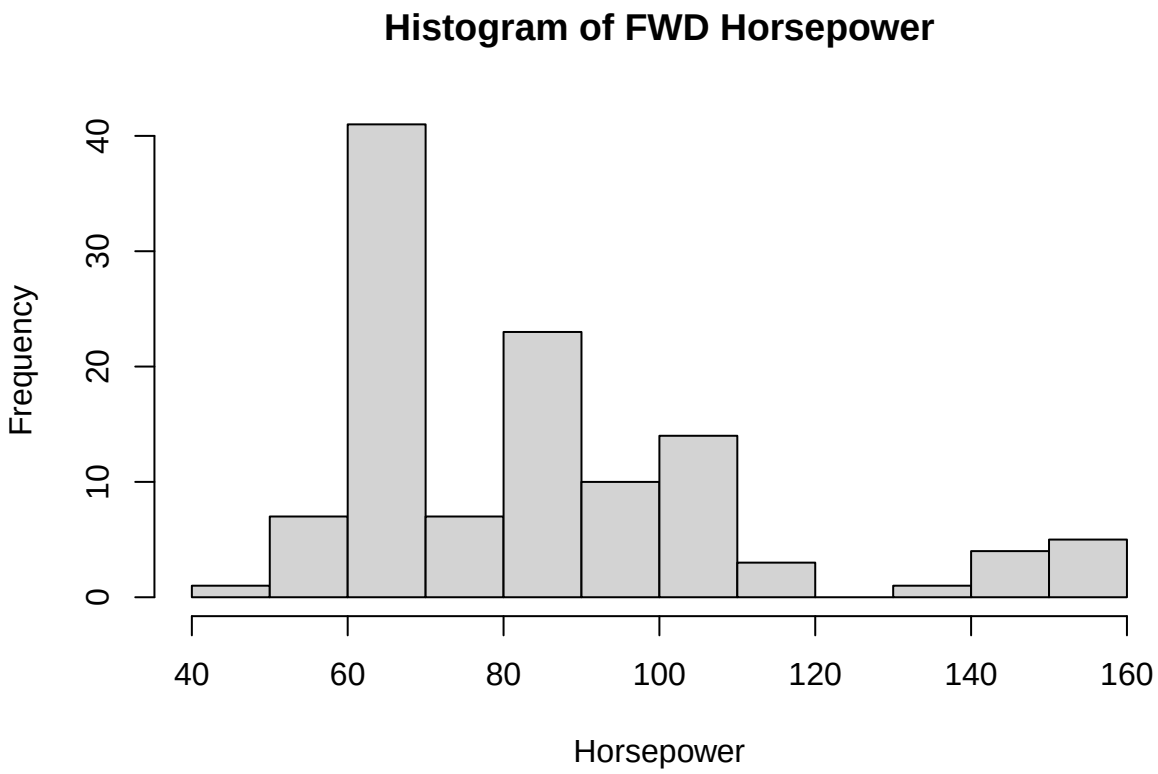
```
## D = 0.16274, p-value = 5.212e-08
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cardata$horsepower[cardata$drive.wheels == "rwd"]
## D = 0.19002, p-value = 4.296e-07
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  cardata$horsepower[cardata$drive.wheels == "4wd"]
## D = 0.23329, p-value = 0.2249
```

Kruskal-Wallisov test normalnosti

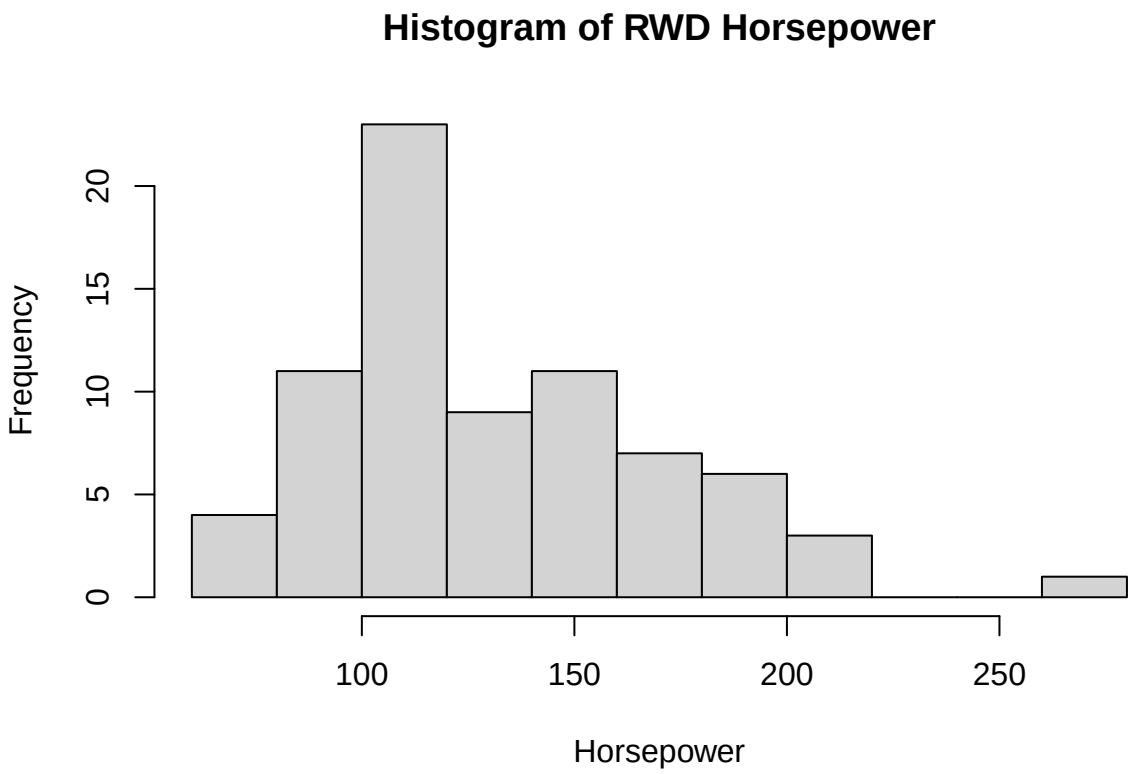
```
# Kruskal-Wallis test to check for differences in 'horsepower' among the 3 drive wheel categories
kruskal.test(horsepower ~ drive.wheels, data = cardata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  horsepower by drive.wheels
## Kruskal-Wallis chi-squared = 77.441, df = 2, p-value < 2.2e-16
```

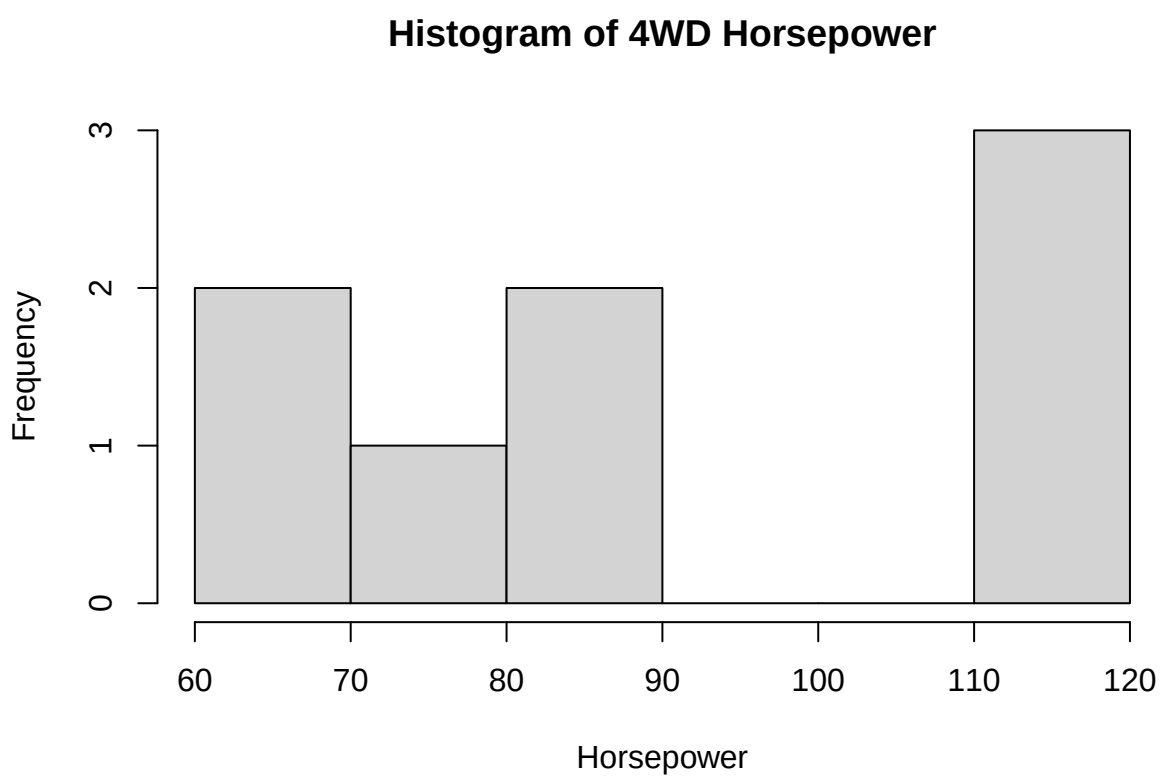
```
# Histograms of 'horsepower' for each 'drive.wheels' category

hist(cardata$horsepower[cardata$drive.wheels=='fwd'], main="Histogram of FWD Horsepower", xlab="Horsepo
```




```
hist(cardata$horsepower[cardata$drive.wheels=='rwd'], main="Histogram of RWD Horsepower", xlab="Horsepo
```




```
hist(cardata$horsepower[cardata$drive.wheels=='4wd'], main="Histogram of 4WD Horsepower", xlab="Horsepo
```



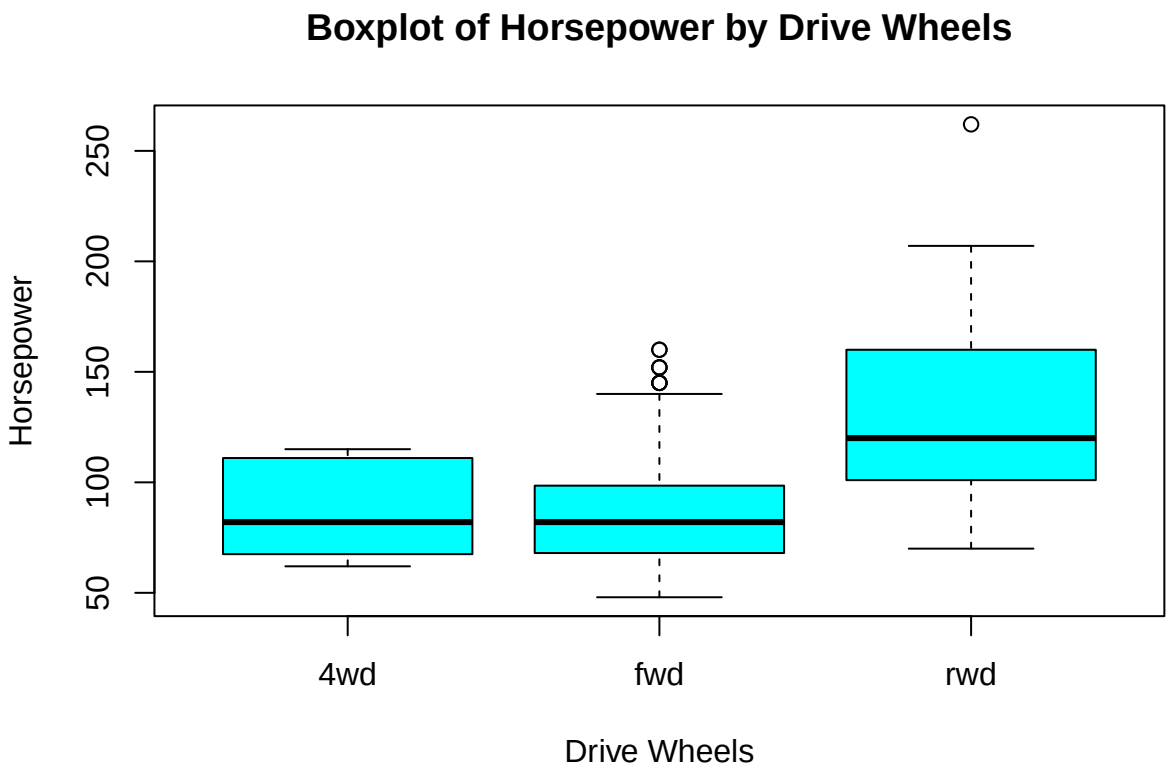

```
# Bartlett's test for homogeneity of variances across different 'drive.wheels' categories
bartlett.test(cardata$horsepower ~ cardata$drive.wheels)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  cardata$horsepower by cardata$drive.wheels
## Bartlett's K-squared = 16.339, df = 2, p-value = 0.0002832
```

```
# Variance calculations for 'horsepower' in each 'drive.wheels' category
var(cardata$horsepower[cardata$drive.wheels == 'fwd'])
var(cardata$horsepower[cardata$drive.wheels == 'rwd'])
var(cardata$horsepower[cardata$drive.wheels == '4wd'])
```

```
## [1] NA
## [1] 1441.206
## [1] 490.2143
```

```
# Boxplot showing distribution of 'horsepower' across 'drive.wheels' categories
boxplot(cardata$horsepower ~ cardata$drive.wheels, main="Boxplot of Horsepower by Drive Wheels", xlab="
```




```r
# ANOVA to check if mean 'horsepower' differs significantly across 'drive.wheels' categories
a = aov(cardata$horsepower ~ cardata$drive.wheels)
summary(a)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## cardata$drive.wheels   2  96017   48009   51.36 <2e-16 ***
## Residuals            196 183221     935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

```r
# Fit a linear model to understand the relationship between horsepower and drive wheel categories
model = lm(horsepower ~ drive.wheels, data = cardata)
summary(model)

# ANOVA test on the linear model, test overall significance of the model
anova(model)
```

```
##
## Call:
## lm(formula = horsepower ~ drive.wheels, data = cardata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61.64 -18.25 -10.25  15.75 130.36
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        87.25      10.81   8.071 6.86e-14 ***
## drive.wheelsfwd    -1.00      11.18  -0.089  0.92880
## drive.wheelsrwd    44.39      11.37   3.904  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.57 on 196 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3439, Adjusted R-squared:  0.3372
## F-statistic: 51.36 on 2 and 196 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: horsepower
##               Df Sum Sq Mean Sq F value    Pr(>F)
## drive.wheels   2  96017   48009  51.357 < 2.2e-16 ***
## Residuals    196 183221     935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# T-test for 'horsepower' between fwd and rwd
t.test(cardata$horsepower[cardata$drive.wheels == 'fwd'],
       cardata$horsepower[cardata$drive.wheels == 'rwd'])

# T-test for 'horsepower' between rwd and 4wd
t.test(cardata$horsepower[cardata$drive.wheels == 'rwd'],
       cardata$horsepower[cardata$drive.wheels == '4wd'])

# T-test for 'horsepower' between fwd and 4wd
t.test(cardata$horsepower[cardata$drive.wheels == 'fwd'],
       cardata$horsepower[cardata$drive.wheels == '4wd'])
```

```
##
##  Welch Two Sample t-test
##
## data:  cardata$horsepower[cardata$drive.wheels == "fwd"] and cardata$horsepower[cardata$drive.wheels
## t = -9.1332, df = 116.17, p-value = 2.518e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -55.23315 -35.54685
## sample estimates:
## mean of x mean of y
##     86.25    131.64
##
##
##  Welch Two Sample t-test
##
## data:  cardata$horsepower[cardata$drive.wheels == "rwd"] and cardata$horsepower[cardata$drive.wheels
## t = 4.9477, df = 11.967, p-value = 0.0003404
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  24.83626 63.94374
## sample estimates:
## mean of x mean of y
##    131.64     87.25
##
##
##  Welch Two Sample t-test
##
## data:  cardata$horsepower[cardata$drive.wheels == "fwd"] and cardata$horsepower[cardata$drive.wheels
## t = -0.12239, df = 8.3046, p-value = 0.9055
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.72192  17.72192
## sample estimates:
## mean of x mean of y
##     86.25     87.25
```

```
# Installing and Loading ggplot2 package
install.packages("ggplot2")
```

```
## Warning: package 'ggplot2' is in use and will not be installed
```

```
library(ggplot2)

# Density plot
density_plot <- ggplot(cardata, aes(x = horsepower, fill = drive.wheels)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Horsepower by Drive Wheels",
       x = "Horsepower",
       y = "Density") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")

# Scatter plot
scatter_plot <- ggplot(cardata, aes(x = engine.size, y = horsepower)) +
  geom_point() +
  labs(title = "Scatter Plot of Engine Size vs Horsepower",
       x = "Engine Size",
       y = "Horsepower") +
  theme_minimal()
```

Pitanje 2 - Razlike u potrošnji automobila prema regiji

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
```

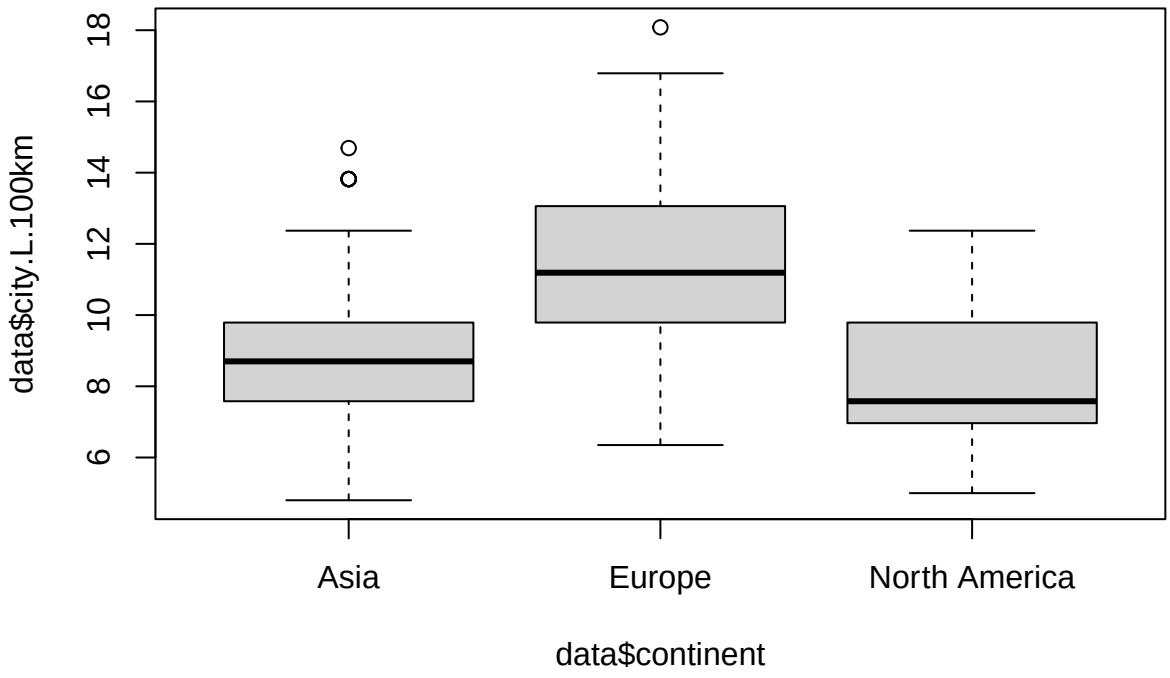
Nama su relevantni stupci 'city.L.100km', 'highway.L.100km' i 'continent'

Usporedba sredina

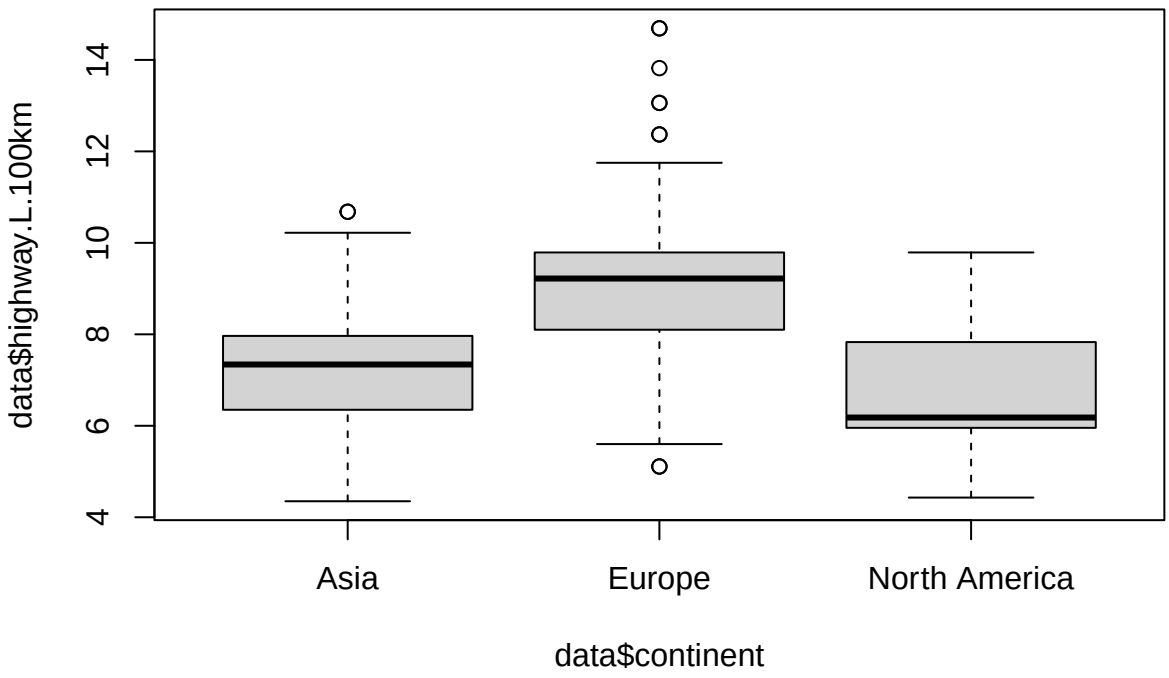
Provjera aritmetičkih sredina Prije nego krenemo sa ANOVA-om, možemo prvo usporediti aritmetičke sredine podataka. Ovo nije dovoljno da radimo bilo kakve konkretne zaključke, ali nam daje uvid u što bi možda očekivali. Također ćemo izračunati varijancu i standardnu devijaciju da imamo bolju ideju o izgledu 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)
```

```r
cat("\n")
print(sprintf("Aritmetička sredina potrošnje goriva u gradu: %f",overallMeanCity))
print(sprintf("Aritmetička sredina potrošnje goriva na autocestama: %f",overallMeanHighway))

print(sprintf("Sveukupna varijanca za gradove: %f",overallVarCity))
print(sprintf("Standardna devijacija: %f",sqrt(overallVarCity)))

print(sprintf("Sveukupna varijanca za autoceste: %f",overallVarHighway))
print(sprintf("Standardna devijacija: %f",sqrt(overallVarHighway)))

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




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




```
##
## [1] "Aritmetička sredina potrošnje goriva u gradu: 9.943582"
## [1] "Aritmetička sredina potrošnje goriva na autocestama: 8.043433"
## [1] "Sveukupna varijanca za gradove: 6.429238"
## [1] "Standardna devijacija: 2.535594"
## [1] "Sveukupna varijanca za autoceste: 3.390360"
## [1] "Standardna devijacija: 1.841293"
```

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

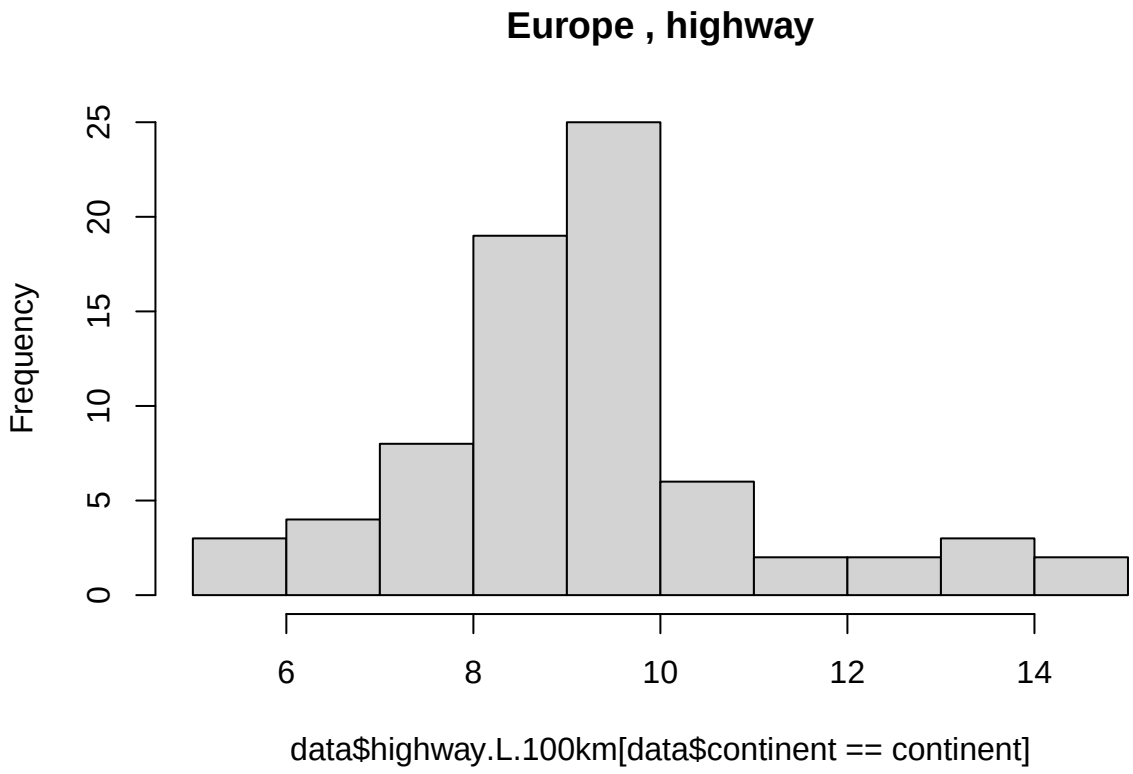
Lillie

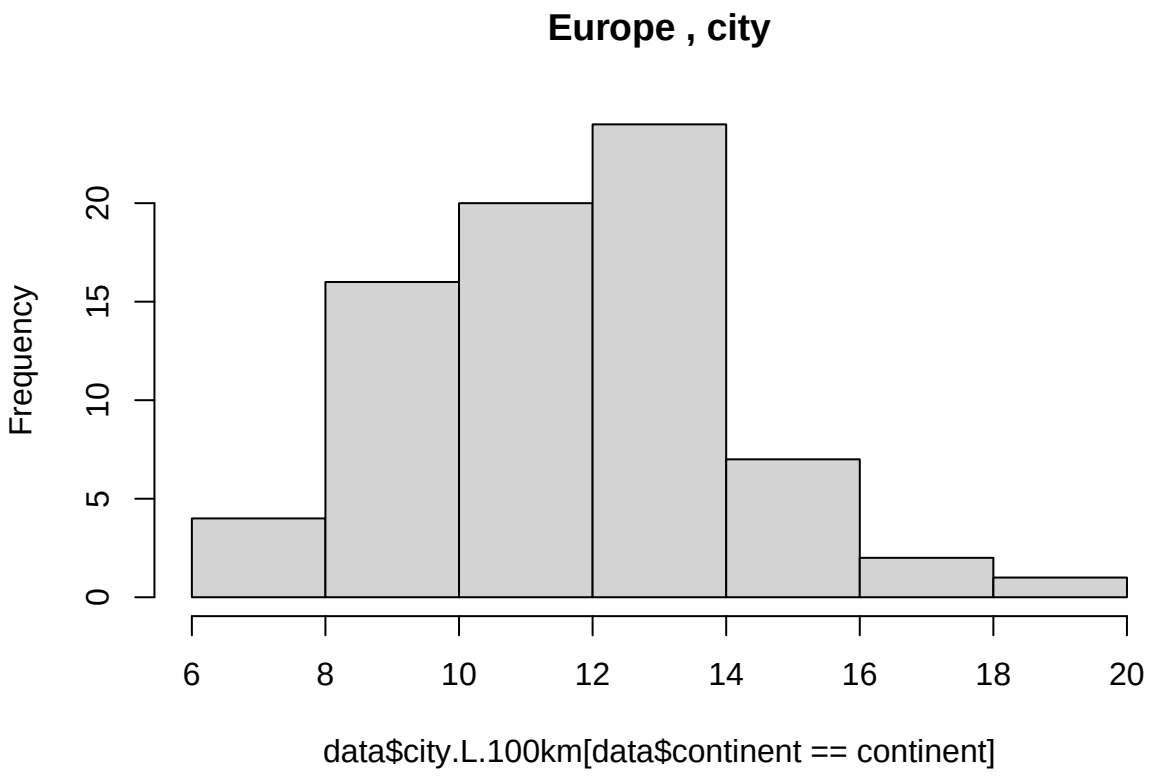
Testiranje normalnosti Prvo ćemo proveti test normalnosti. Zanimaju nas tablice za potrošnju goriva u gradu i na autocesti za svaki kontinent. Koristit ćemo Lilliefors test koji se temelji na Kolmogorov-Smirnov testu i Q-Q plot za vizualizaciju. Postavljamo hipoteze:

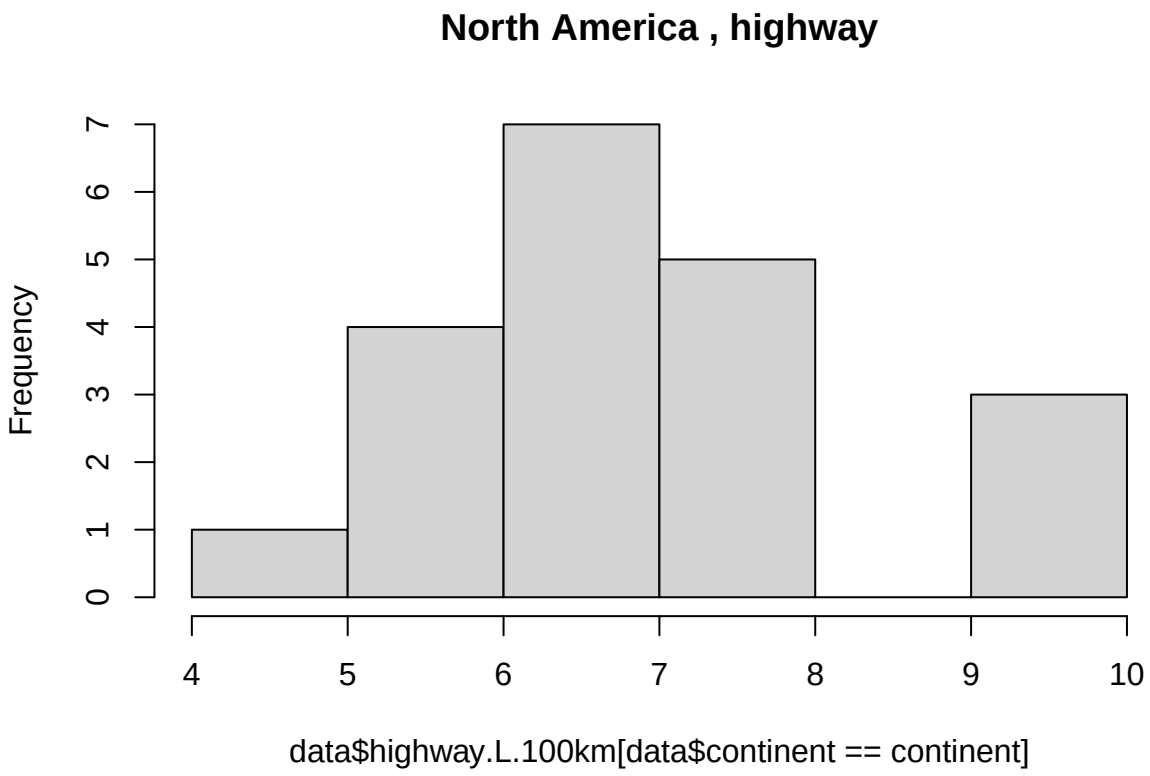
$$H0 : \textit{Dani podaci za potrošnju goriva imaju normalnu distribuciju} H1 : \neg H0$$

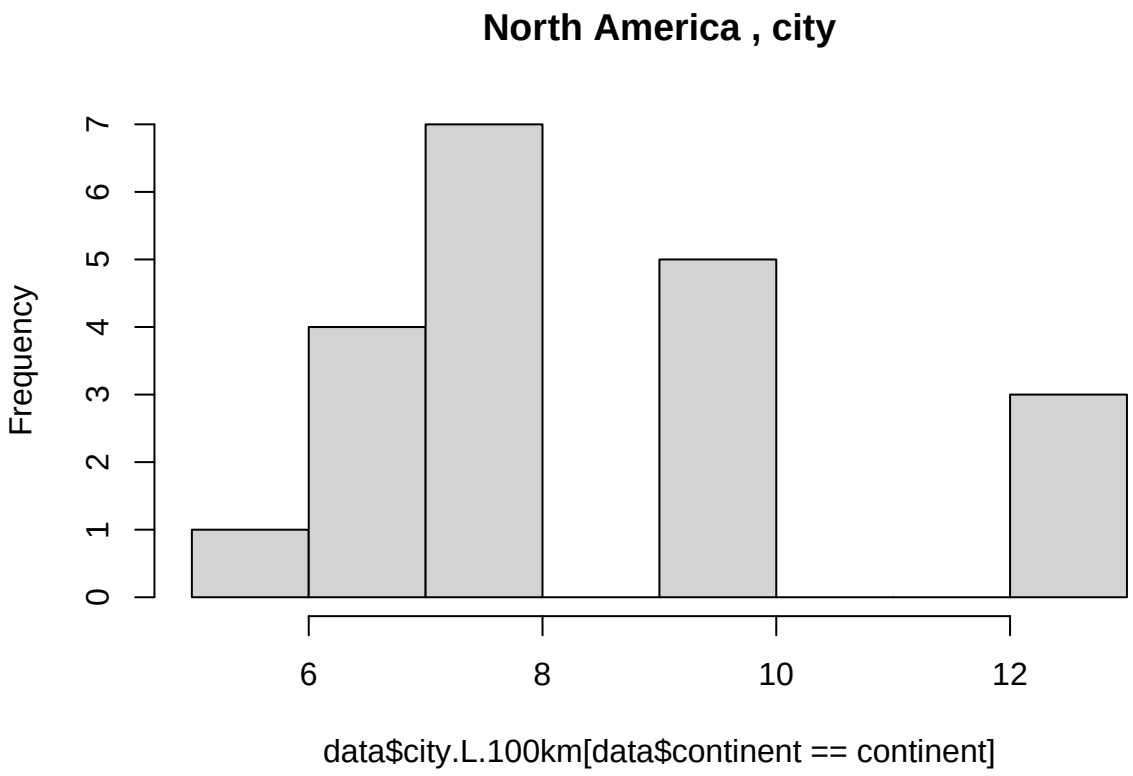
Provedimo test:

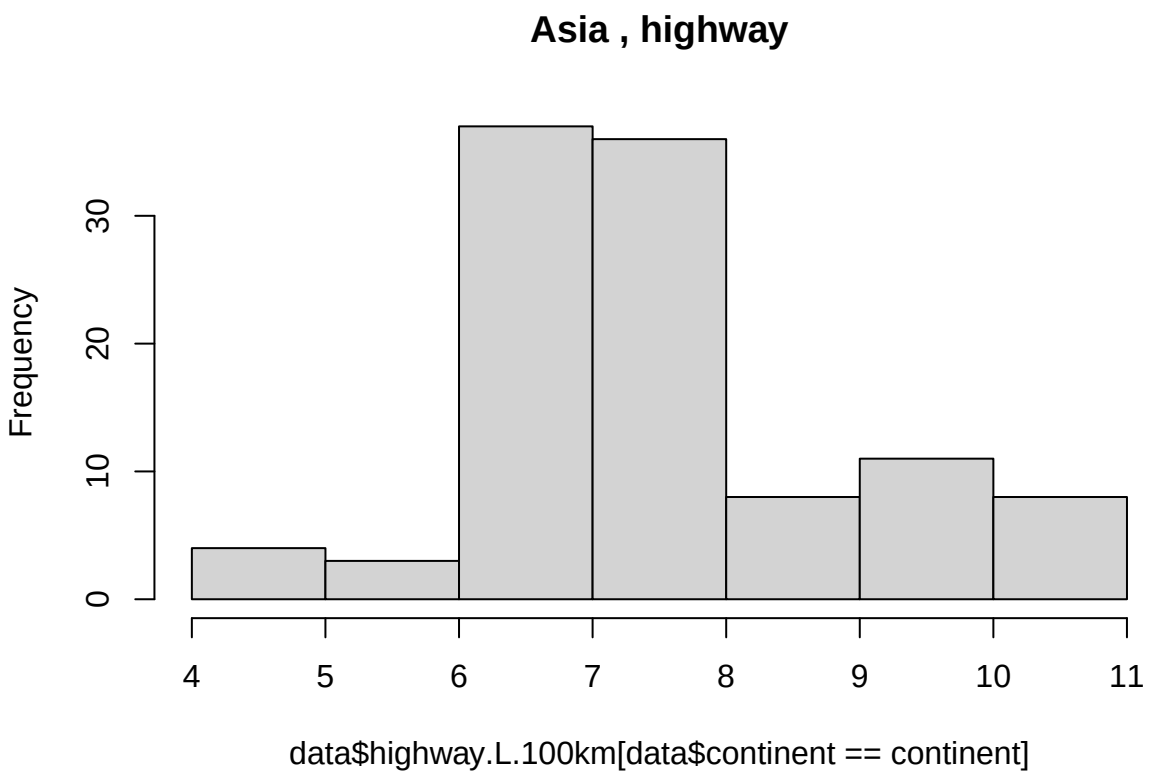
```r
require(nortest) #potrebna biblioteka

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)
}
```

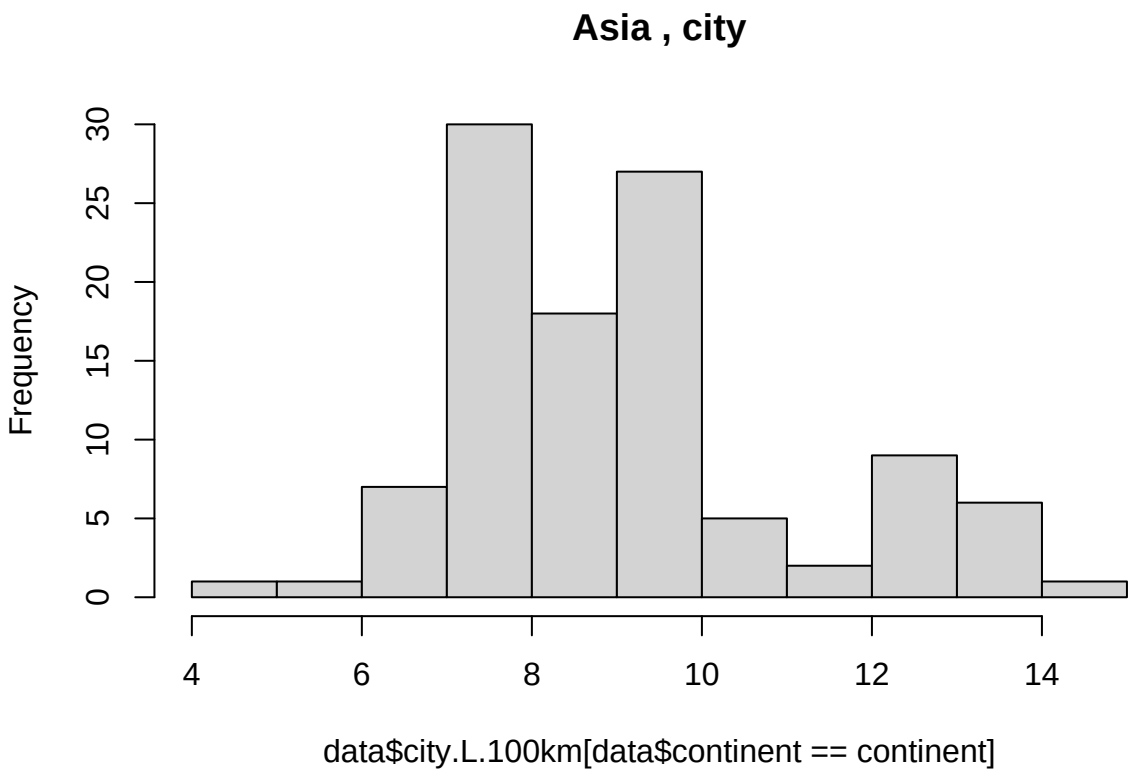



Europe , city
Frequency
0
5
10
15
20
6
8
10
12
14
16
18
20
data$city.L.100km[data$continent == continent]


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]


North America , city
Frequency
0
1
2
3
4
5
6
7
6
8
10
12
data$city.L.100km[data$continent == continent]


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

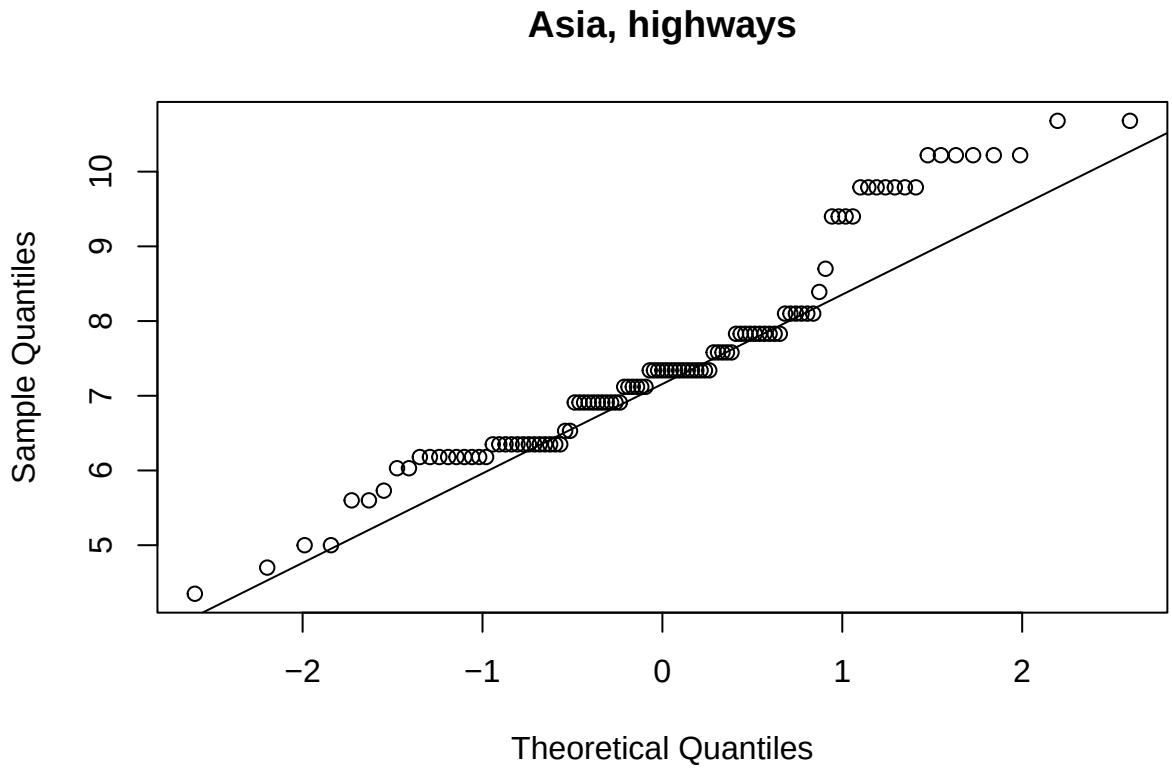



```
## [1] "For continent: Europe"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$city.L.100km[data$continent == continent]
## D = 0.11082, p-value = 0.02499
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$highway.L.100km[data$continent == continent]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

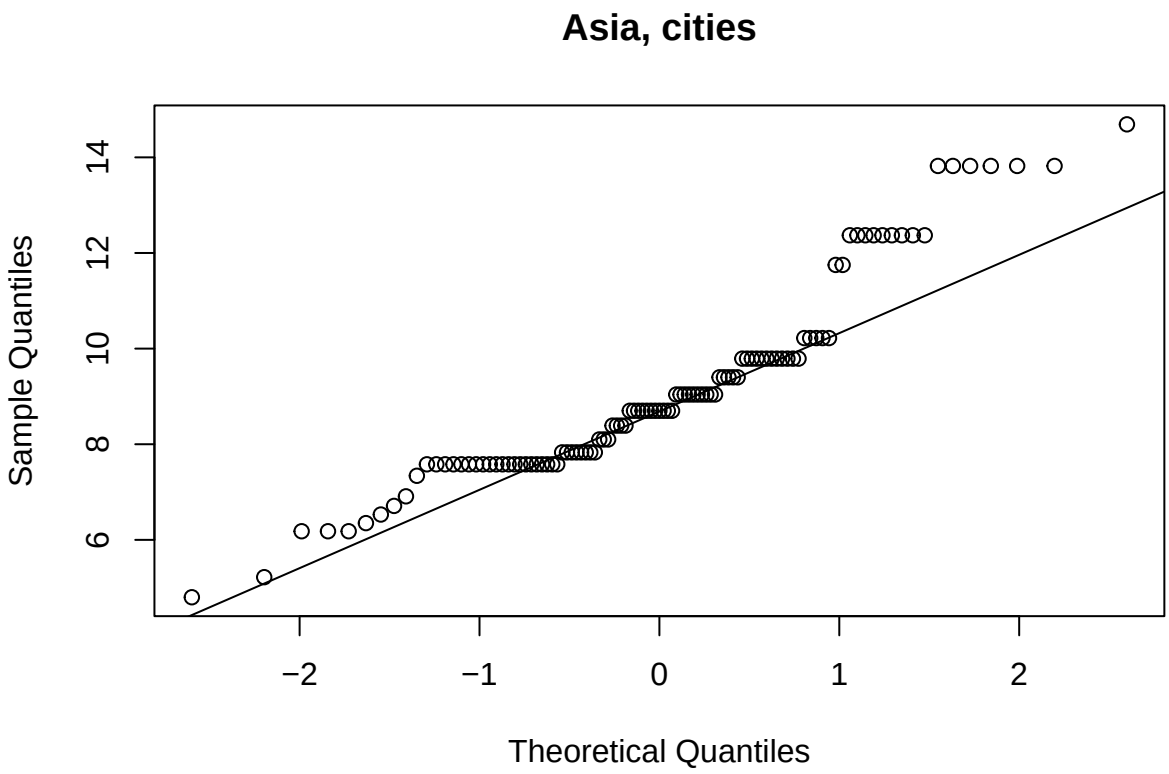
```
## [1] "For continent: Asia"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$city.L.100km[data$continent == continent]
## D = 0.15718, p-value = 7.516e-07
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$highway.L.100km[data$continent == continent]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

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:

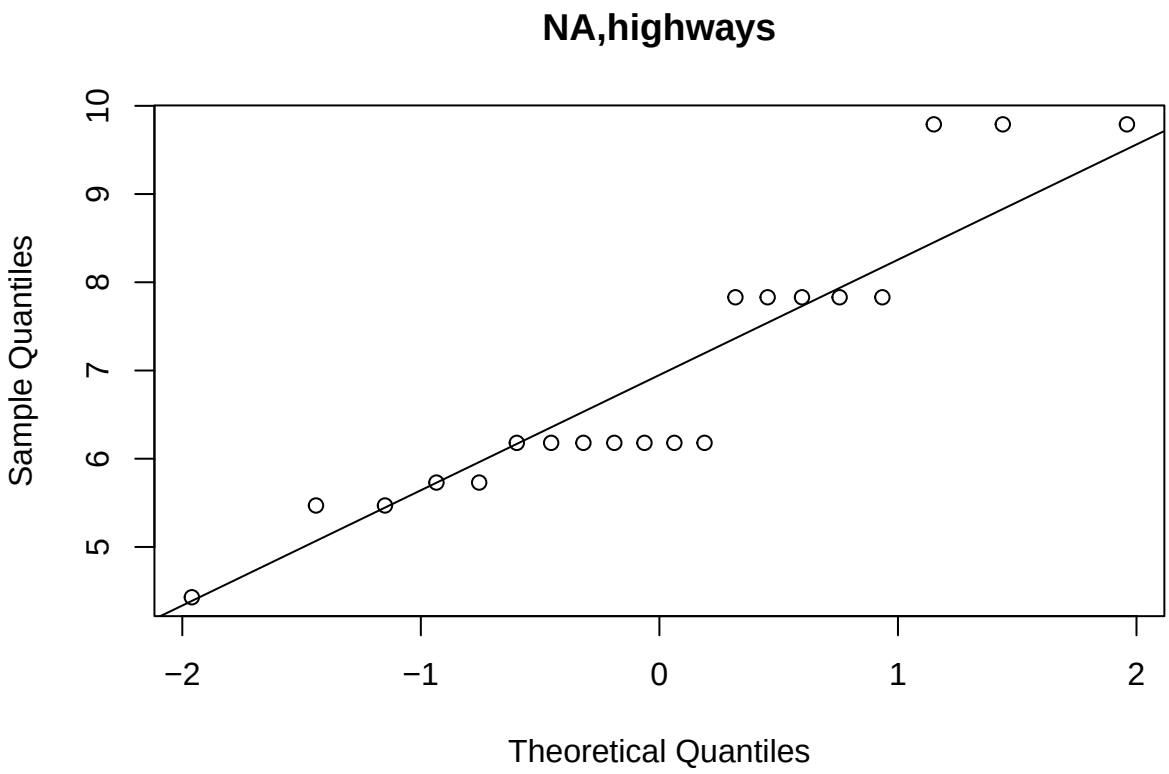
```
qqnorm(data$highway.L.100km[data$continent == "Asia"],main="Asia, highways")
qqline(data$highway.L.100km[data$continent == "Asia"])
```



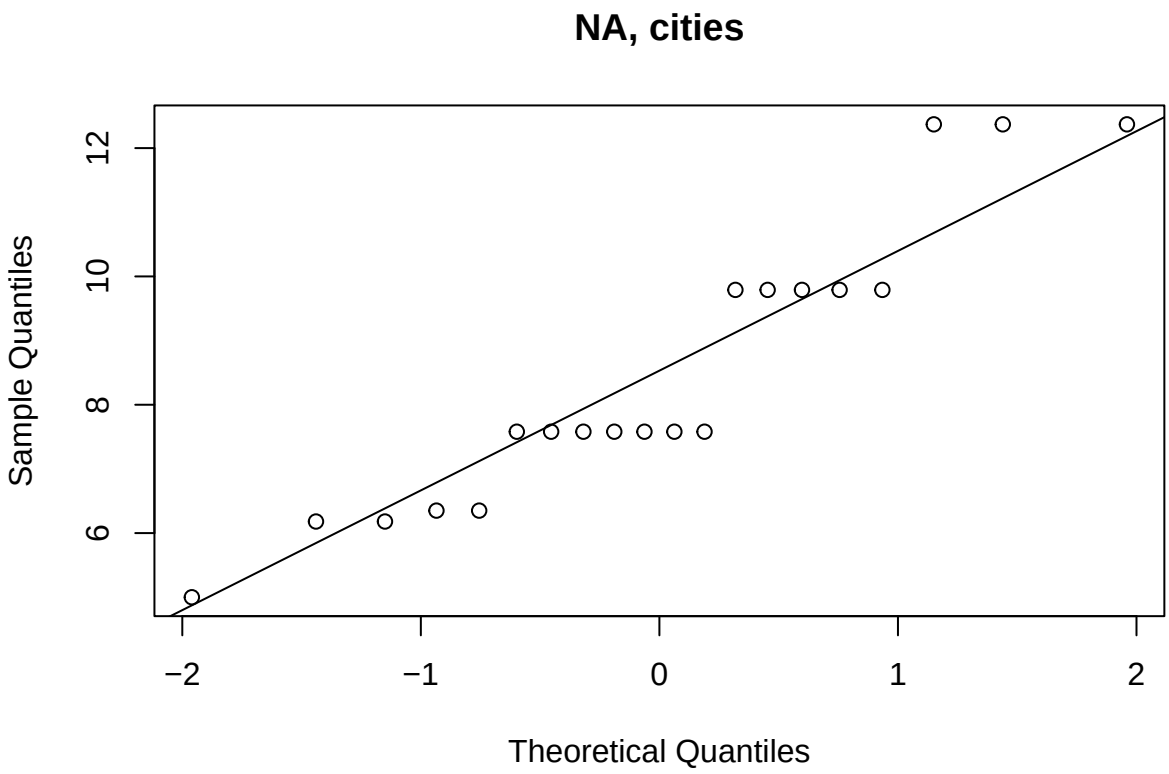

```
qqnorm(data$city.L.100km[data$continent == "Asia"],main = "Asia, cities")
qqline(data$city.L.100km[data$continent == "Asia"])
```



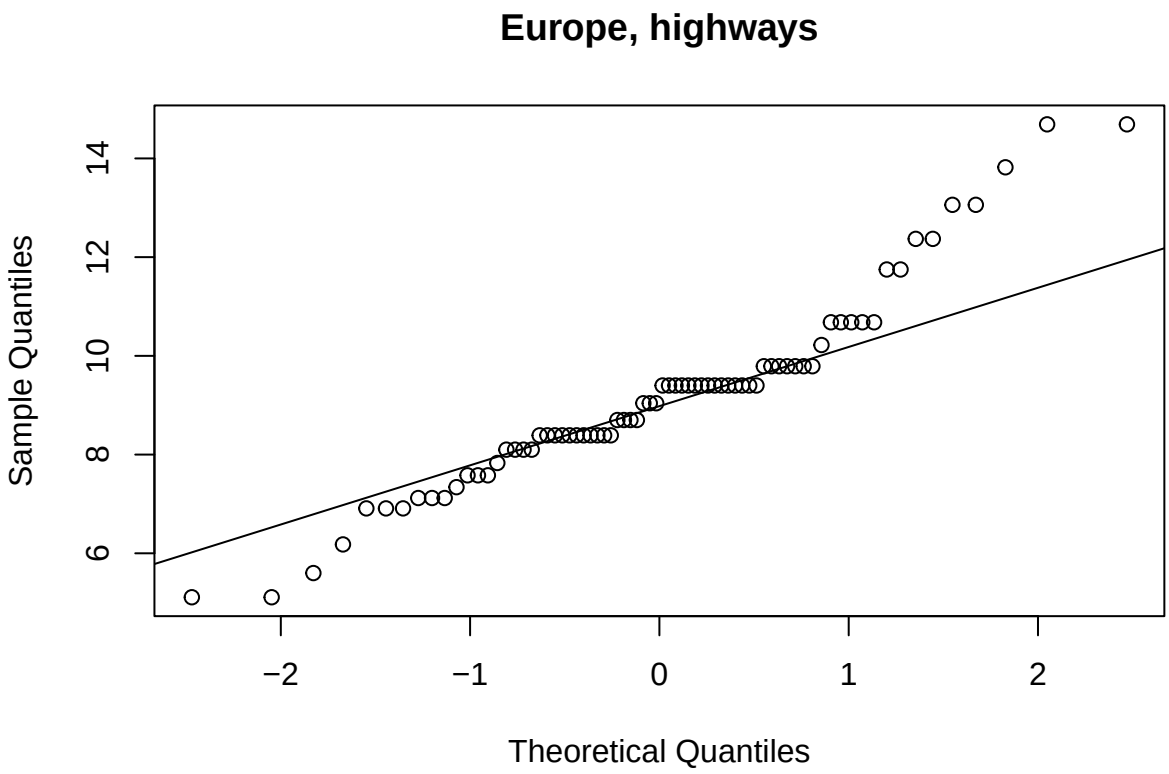

```r
qqnorm(data$highway.L.100km[data$continent == "North America"],main = "NA,highways")
qqline(data$highway.L.100km[data$continent == "North America"])
```



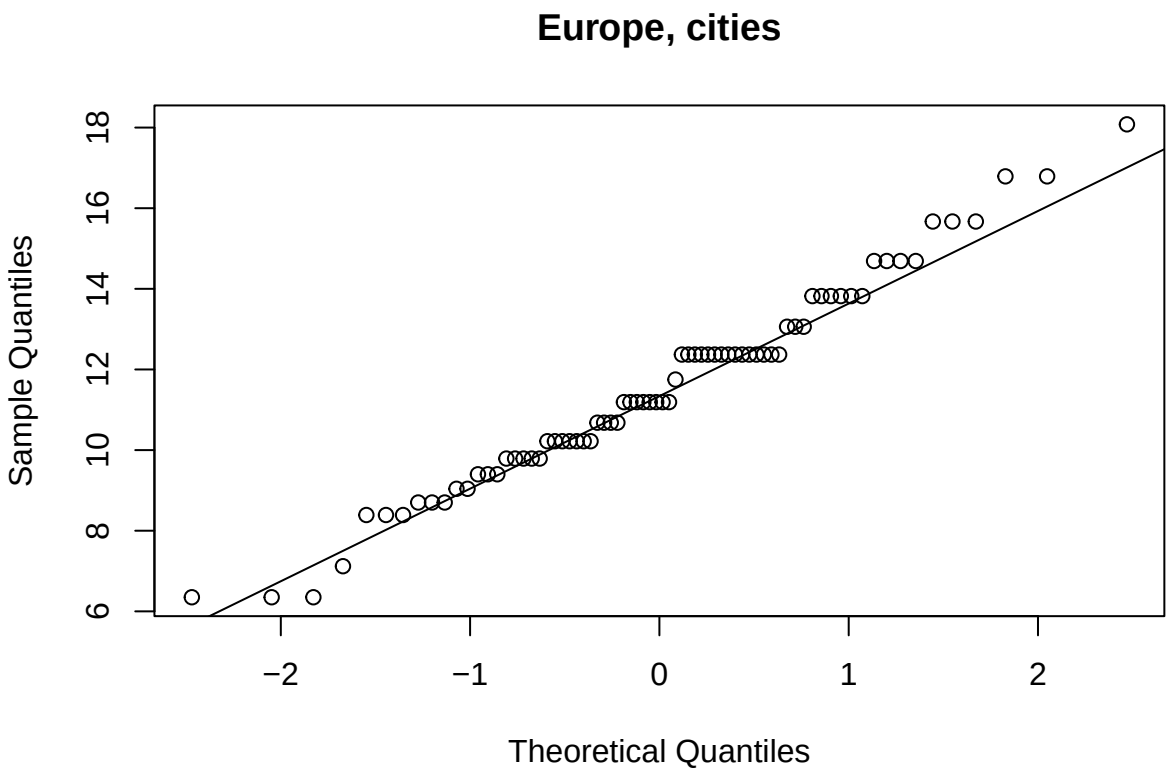

```
qqnorm(data$city.L.100km[data$continent == "North America"],main = "NA, cities")
qqline(data$city.L.100km[data$continent == "North America"])
```




```
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")
print(dunn)
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 = 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
## $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"
##
##   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.

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.

```
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*

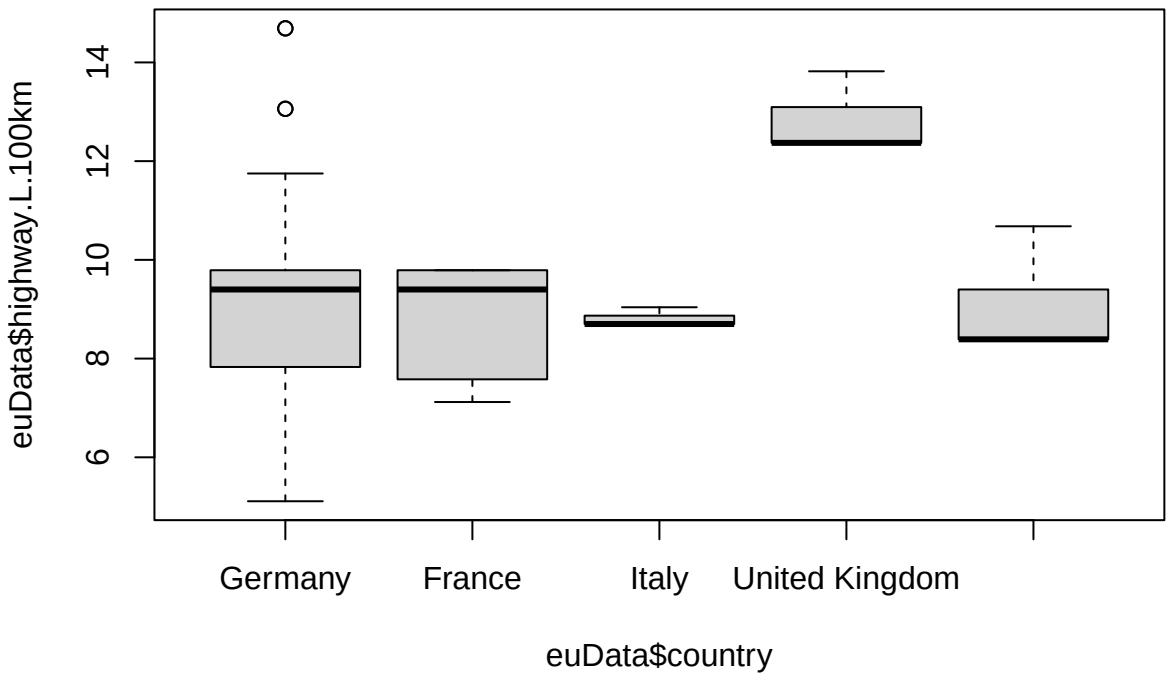
Provedimo testove:

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

print(euKruskalHighway)
print(euKruskalCity)
```

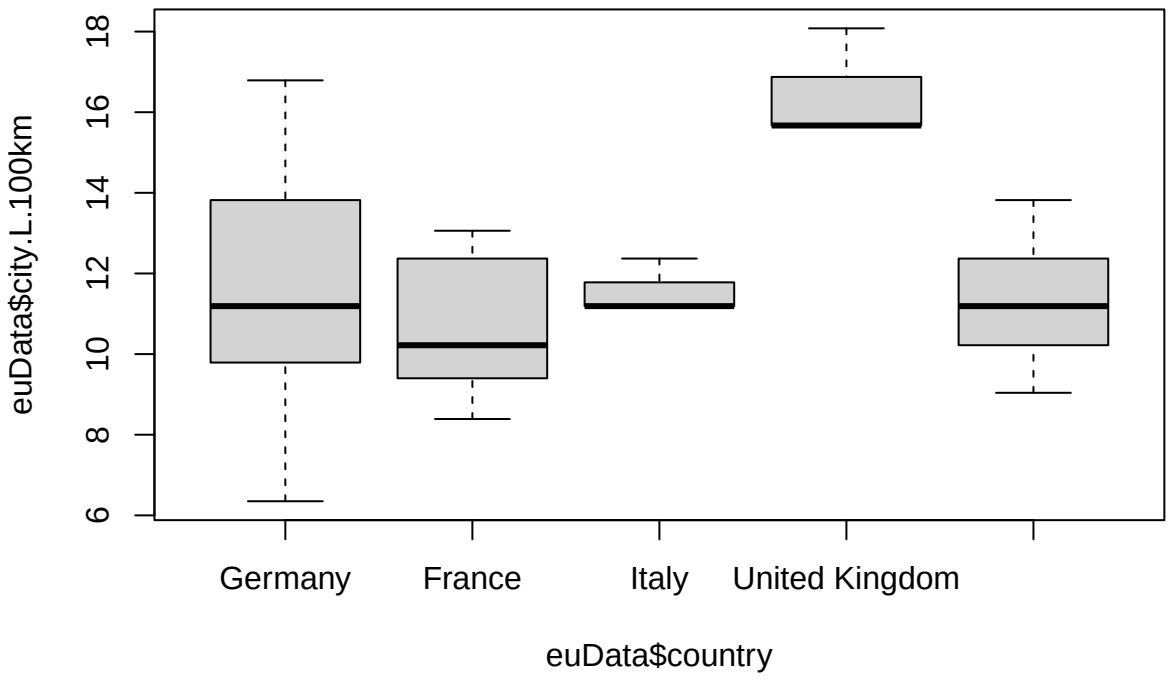
```
##
##  Kruskal-Wallis rank sum test
##
## data:  europeDataHighway
## Kruskal-Wallis chi-squared = 0.62478, df = 2, p-value = 0.7317
##
##
##  Kruskal-Wallis rank sum test
##
## data:  europeDataCity
## Kruskal-Wallis chi-squared = 2.151, df = 2, p-value = 0.3411
```

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)
```




```
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.

Pitanje 3 - Predvidanje cijene automobila

Mozemo li temeljem drugih dostupnih varijabli predvidjeti cijenu automobila? Koja varijabla pritom ima najznacajniji utjecaj?

Izrađeno je ukupno šest regresijskih modela, svaki kombinira različite karakteristike automobila kako bi se pronašla veza između cijene i tih karakteristika. Cilj je sastaviti najbolji model koji predviđa cijenu automobila.

Interpretacija plotova

Uz modele su priloženi tzv. Residual vs Fitted te Normal Q-Q grafovi.

Residual vs Fitted graf provjerava pretpostavku da ostaci (**reziduali**) imaju srednju vrijednost nula na svim razinama nezavisnih varijabli. Željeni izgled grafa je nasumična raspršenost tocaka, bez jasno vidljivog obrasca. Obrasci ukazuju na to da model nije adekvatno objasnio neki aspekt korištene strukture podataka. Nasumično raspršene točke oko horizontalne linije na nuli sugeriraju da su predviđanja modela nepristrana na svim razinama nezavisnih varijabli. To implicira da je model prikladan za podatke duž cijelog raspona predviđenih vrijednosti. Odsutnost obrazaca ili sustavnih struktura sugerira da su reziduali modela

raspoređeni nasumično, što podupire pretpostavku da je odnos između nezavisnih varijabli i zavisne varijable linearan.

Zakrivljeni oblik: Zakrivljen odnos u ostacima (kao oblik slova U ili obrnuto U) može sugerirati da postoji nelinearan odnos između nezavisnih varijabli i zavisne varijable koji nije uhvaćen modelom.

Raširenje/skupljanje (heteroscedastičnost): Ako se ostaci šire ili skupljaju kako se povećavaju ili smanjuju predviđene vrijednosti, to ukazuje na nejednaku varijancu (heteroscedastičnost), što krši jednu od pretpostavki linearne regresije.

Interpretacija ispisa

Ispis pruža rezultate analize linearne regresije, uključujući koeficijente modela, njihovu statističku značajnost i ukupno prilagođavanje modela. Razmotrimo ključne komponente:

Koeficijenti: Za svaki član u modelu dobivamo **procjenu veličine učinka** (Estimate), **standardnu pogrešku te procjene** (Std. Error) i **t-vrijednost**, koja je procjena podijeljena s njezinom standardnom pogreškom. **Stupac Pr(>|t|)** prikazuje p-vrijednost za t-test protiv nulte hipoteze da je koeficijent nula (nema učinka). Određeni članovi su statistički značajni prediktori cijene (Pr(>|t|) je manje od npr. 0,05), pri čemu ce određeni imati posebno malu p-vrijednost, što ukazuje na snažan odnos s cijenom. Određeni članovi neće biti statistički značajni na razini od npr. 0,05 (p-vrijednost je npr. 0,0641).

Oznake značajnosti: Zvjezdice označavaju razinu značajnosti, s više zvjezdica označava višu statističku značajnost.

Standardna pogreška ostataka: Ovo je procjena standardne devijacije ostataka (reziduala), što je otprilike prosječna udaljenost na kojoj se promatrane vrijednosti nalaze od regresijske linije.

Multiple R-kvadrat i Prilagođeni R-kvadrat: Multiple R-kvadrat od npr. 0,7661 ukazuje da se otprilike 76,61% varijabilnosti cijene može objasniti modelom. To je mjera dobrog prilagođavanja modela. Prilagođeni R-kvadrat je prilagođen broju prediktora u modelu i preciznija je mjera dobrog prilagođavanja. Za ovaj model može iznositi npr. 0,7625.

F-statistika i njezina p-vrijednost: F-statistika testira nultu hipotezu da su svi koeficijenti regresije jednaki nuli (tj. model nema objašnjavajuću snagu). Vrlo mala p-vrijednost ukazuje da je model statistički značajan i da je barem jedan od prediktora povezan s cijenom.

Regresijski modeli

Karakteristike motora

Sastavljena su tri regresijska modela, svaki od kojih tvori različitu kombinaciju karakteristika motora.

Model 1 Prvi model koristi varijable: konjske snage, veličinu motora, broj cilindara te sustav goriva kako bi predvidio cijenu automobila.

*KOD:**

```
# Učitavanje podataka
data <- read.csv("car_specifications.csv")

# Pretvaranje kategoričkih varijabli u faktore
data$make <- as.factor(data$make)
data$aspiration <- as.factor(data$aspiration)
data$num_of_doors <- as.factor(data$num.of.doors)
data$body_style <- as.factor(data$body.style)
```

```r
data$drive_wheels <- as.factor(data$drive.wheels)
data$engine_location <- as.factor(data$engine.location)
data$fuel <- as.factor(data$fuel)
data$country <- as.factor(data$country)
data$continent <- as.factor(data$continent)

# Prikazivanje razina engine_location
levels(data$engine_location)
```

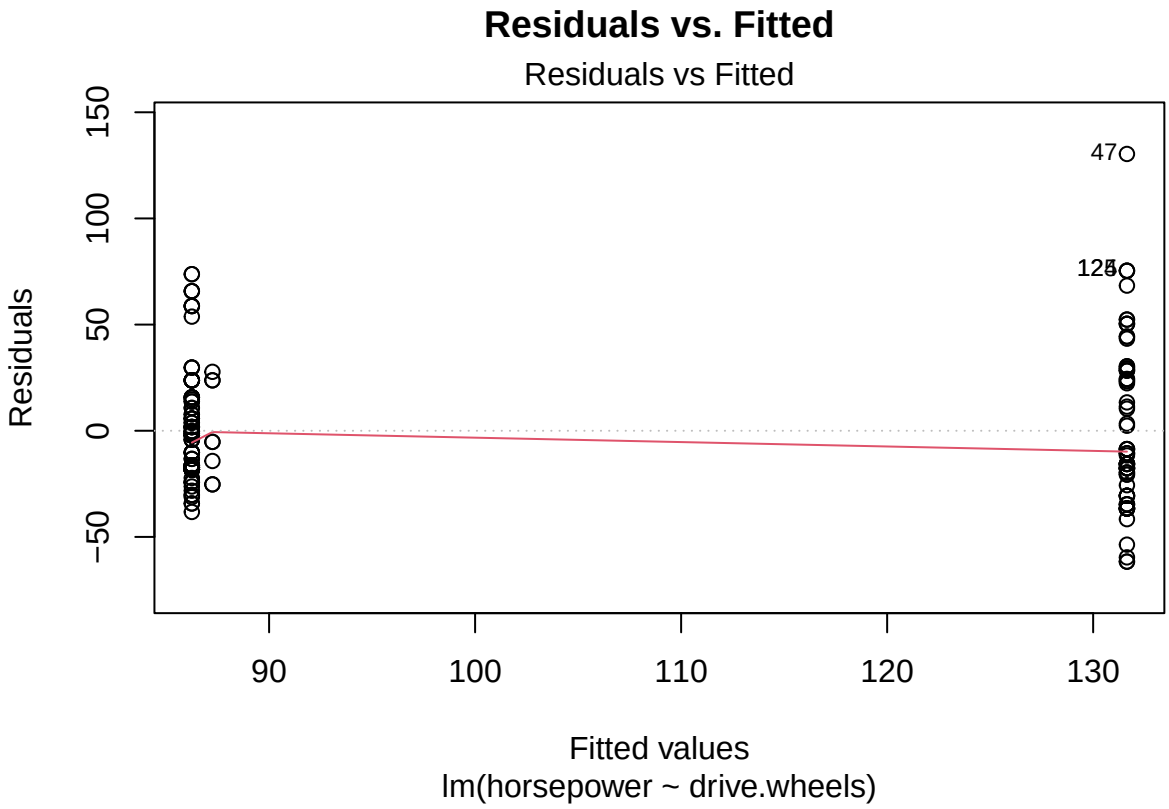
```
## [1] "front" "rear"
```

```r
# Izgradnja regresijskog modela
# Ovaj model sastoji se od razlicitih karakteristika motora
ec1_model <- lm(price ~ horsepower + engine.size + num.of.cylinders + fuel.system, data = data)

# Prikazivanje sažetka modela
summary(model)
```

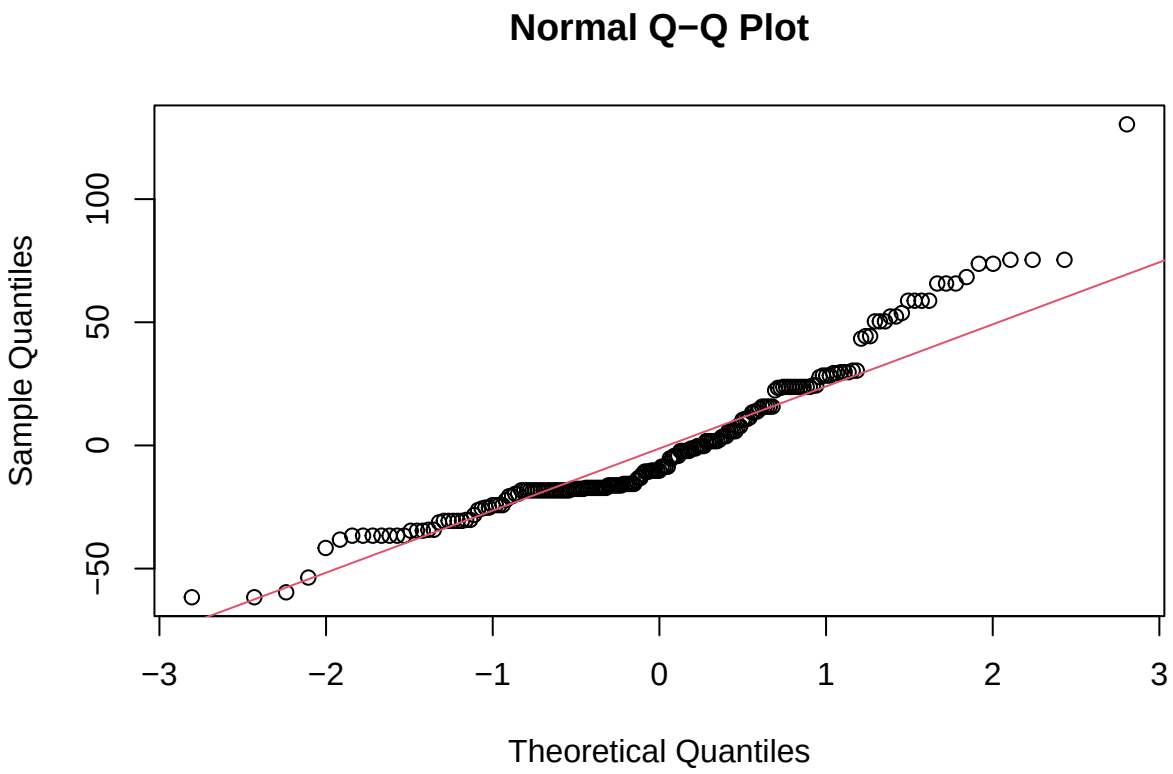
```
##
## Call:
## lm(formula = horsepower ~ drive.wheels, data = cardata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61.64 -18.25 -10.25  15.75 130.36
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        87.25      10.81   8.071 6.86e-14 ***
## drive.wheelsfwd    -1.00      11.18  -0.089  0.92880
## drive.wheelsrwd    44.39      11.37   3.904  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.57 on 196 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3439, Adjusted R-squared:  0.3372
## F-statistic: 51.36 on 2 and 196 DF,  p-value: < 2.2e-16
```

Promatrajući procijenjene doprinose varijable prediktoru, zaključujemo da konjske snage te veličina motora značajno doprinose određivanju cijene automobila. Također, 12 cilindara značajno utječe na cijenu automobila. R-squared metrike govore nam da sastavljeni model objasnjava 85.95%, odnosno, 84.80% varijabilnosti. P-vrijednost je vrlo niska, što na nam govori da je model statistički značajan.

```r
# Prikazivanje dijagnostike modela
# Residuals vs. Fitted plot
plot(model, which = 1, main = "Residuals vs. Fitted")
```




```r
# Normal Q-Q plot
qqnorm(resid(model))
qqline(resid(model), col = 2)
```



Model 2 Drugi model koristi varijable: konjske snage, lokaciju motora, veličinu motora te tip motora kako bi predvidio cijenu automobila.

KOD:

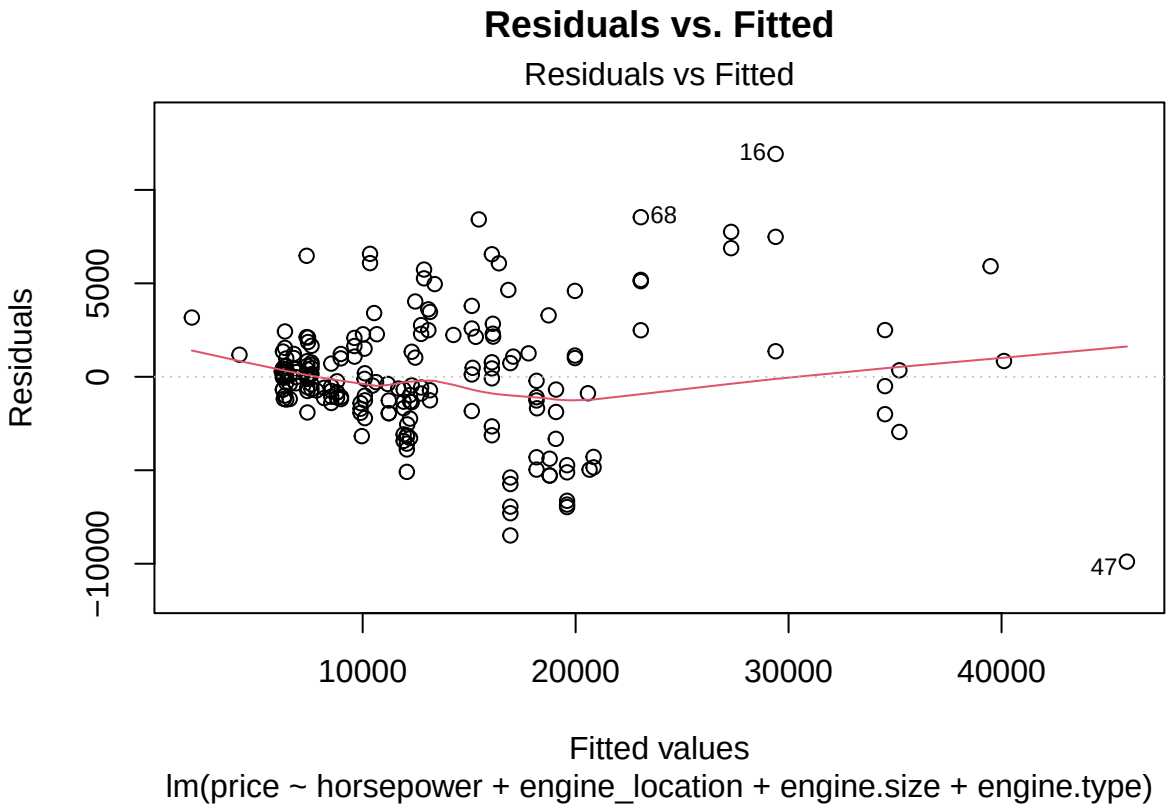
```
# Učitavanje podataka
data <- read.csv("car_specifications.csv")

# Pretvaranje kategoričkih varijabli u faktore
data$make <- as.factor(data$make)
data$aspiration <- as.factor(data$aspiration)
data$num_of_doors <- as.factor(data$num.of.doors)
data$body_style <- as.factor(data$body.style)
data$drive_wheels <- as.factor(data$drive.wheels)
data$engine_location <- as.factor(data$engine.location)
data$fuel <- as.factor(data$fuel)
data$country <- as.factor(data$country)
data$continent <- as.factor(data$continent)

# Prikazivanje razina engine_location
levels(data$engine_location)
```

```
## [1] "front" "rear"
```

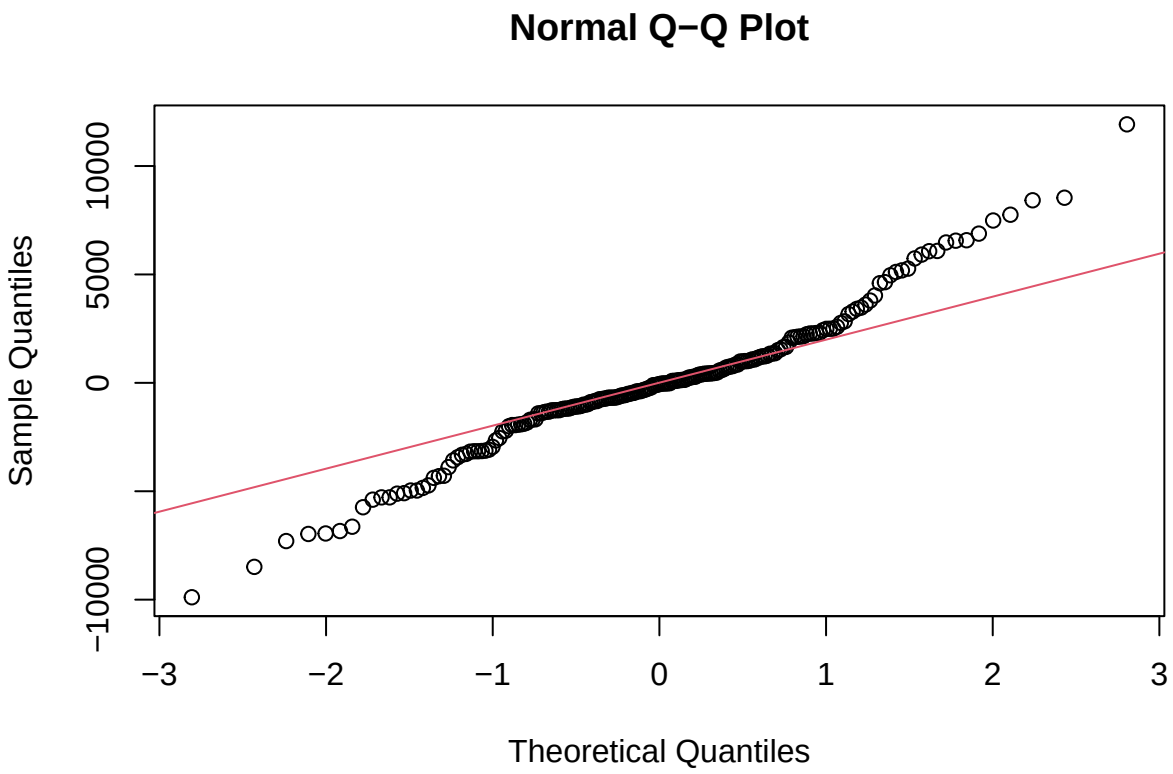
```r
# Izgradnja regresijskog modela
# Ovaj model kombinira različite karakteristike motora
model <- lm(price ~ horsepower + engine_location + engine.size  + engine.type , data = data)

# Prikazivanje sažetka modela
summary(model)
```

```
##
## Call:
## lm(formula = price ~ horsepower + engine_location + engine.size +
##     engine.type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9888.6 -1329.6   -58.6  1345.7 11922.5
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.229e+04  1.610e+03  -7.635 1.07e-12 ***
## horsepower           3.744e+01  1.291e+01   2.900  0.00417 **
## engine_locationrear  7.688e+03  2.362e+03   3.256  0.00134 **
## engine.size          9.675e+00  7.062e-01  13.700  < 2e-16 ***
## engine.typel         2.802e+03  1.419e+03   1.975  0.04971 *
## engine.typeohc       1.735e+03  1.084e+03   1.600  0.11128
## engine.typeohcf      6.262e+02  1.434e+03   0.437  0.66283
## engine.typeohcv     -3.312e+03  1.407e+03  -2.354  0.01962 *
## engine.typerotor     9.720e+03  2.022e+03   4.808 3.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3286 on 190 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8372, Adjusted R-squared:  0.8304
## F-statistic: 122.2 on 8 and 190 DF,  p-value: < 2.2e-16
```

```r
# Prikazivanje dijagnostike modela
# Residuals vs. Fitted plot
plot(model, which = 1, main = "Residuals vs. Fitted")
```




```r
# Normal Q-Q plot
qqnorm(resid(model))
qqline(resid(model), col = 2)
```



Promatrajući procijenjene doprinose varijabli prediktoru, zaključujemo da konjske snage, lokacija motora i veličina motora značajno doprinose određivanju cijene automobila. Tip motora "rotor" također ima značajan utjecaj na cijenu automobila. R-squared metrike govore nam da sastavljeni model objašnjava 83.72%, odnosno, 83.04% varijabilnosti. P-vrijednost je vrlo niska, što nam govori da je model statistički značajan.

Model 3 Treći model koristi varijable: konjske snage, veličinu motora, tip motora, broj cilindara, promjer cilindra, hod klipa, omjer kompresije, i najveću okretajnu brzinu motora kako bi predvidio cijenu automobila.

KOD:

```
# Učitavanje podataka
data <- read.csv("car_specifications.csv")

# Pretvaranje kategoričkih varijabli u faktore
data$make <- as.factor(data$make)
data$aspiration <- as.factor(data$aspiration)
data$num_of_doors <- as.factor(data$num.of.doors)
data$body_style <- as.factor(data$body.style)
data$drive_wheels <- as.factor(data$drive.wheels)
data$engine_location <- as.factor(data$engine.location)
data$fuel <- as.factor(data$fuel)
data$country <- as.factor(data$country)
data$continent <- as.factor(data$continent)

# Prikazivanje razina engine_location
levels(data$engine_location)
```

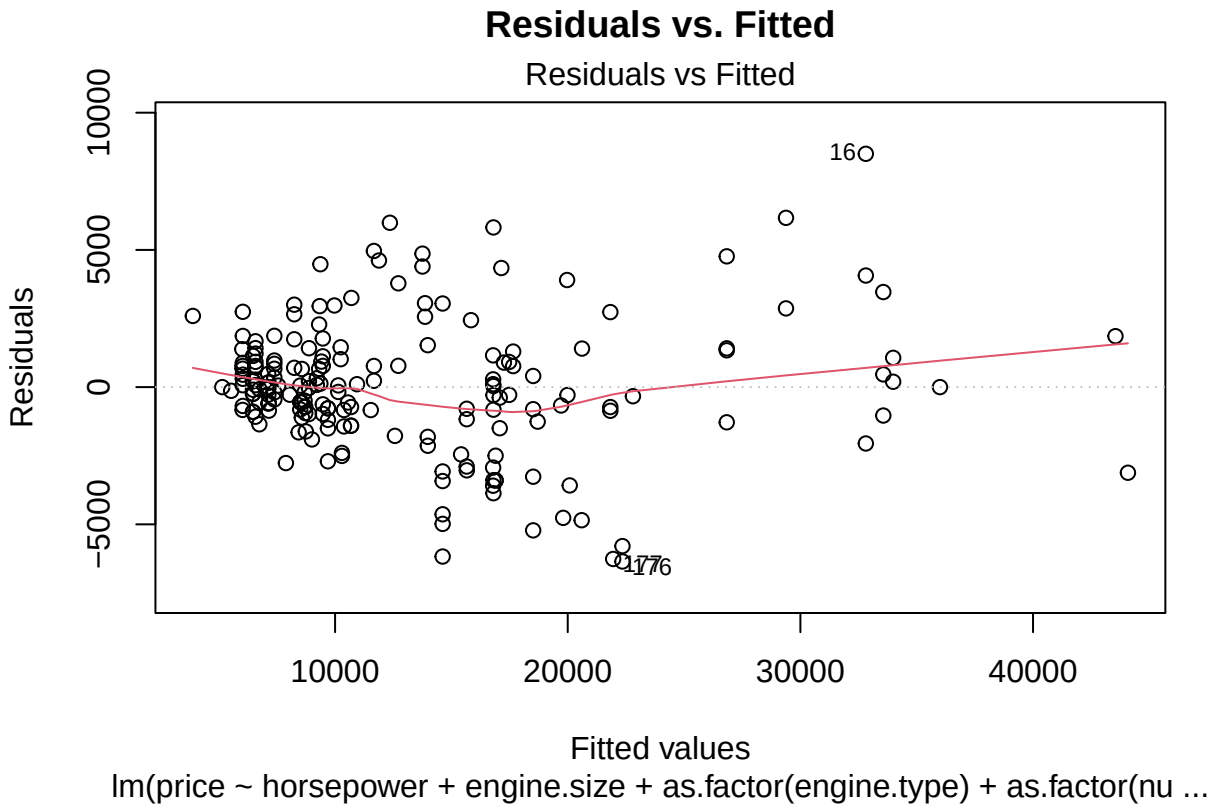
```
## [1] "front" "rear"
```

```r
# Izgradnja regresijskog modela
# Ovaj model kombinira različite karakteristike motora
model <- lm(price ~ horsepower + engine.size + as.factor(engine.type) +
             as.factor(num.of.cylinders) + bore + stroke + compression.ratio + peak.rpm,
            data = data)

# Prikazivanje sažetka modela
summary(model)
```

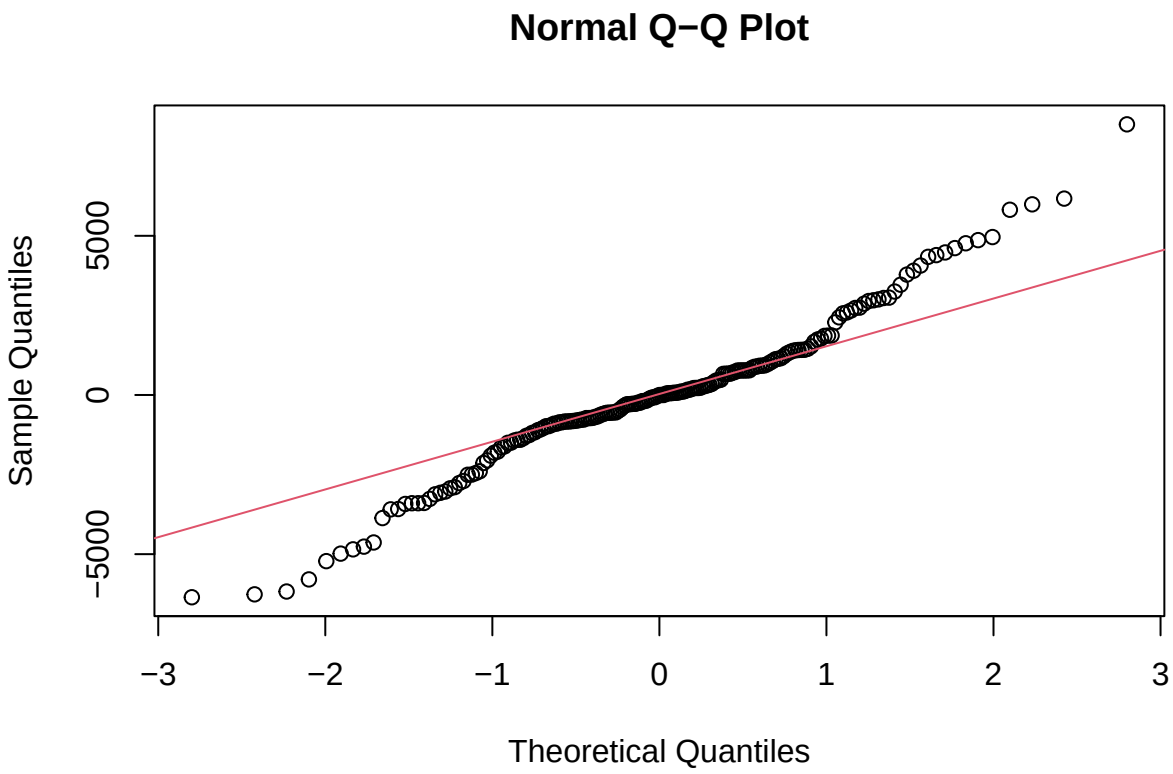
```
##
## Call:
## lm(formula = price ~ horsepower + engine.size + as.factor(engine.type) +
##     as.factor(num.of.cylinders) + bore + stroke + compression.ratio +
##     peak.rpm, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6353.2  -979.6    -2.6  1043.0  8501.4
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.588e+03  5.468e+03   0.839 0.402526
## horsepower                       7.290e+01  1.291e+01   5.645 6.38e-08 ***
## engine.size                      8.252e+00  1.267e+00   6.514 7.19e-10 ***
## as.factor(engine.type)l          4.349e+03  1.151e+03   3.779 0.000214 ***
## as.factor(engine.type)ohc        4.201e+03  8.395e+02   5.004 1.34e-06 ***
## as.factor(engine.type)ohcf       1.104e+03  1.145e+03   0.964 0.336350
## as.factor(engine.type)ohcv      -6.443e+03  1.195e+03  -5.391 2.19e-07 ***
## as.factor(num.of.cylinders)five -9.335e+03  2.770e+03  -3.370 0.000922 ***
## as.factor(num.of.cylinders)four -1.342e+04  3.132e+03  -4.285 2.98e-05 ***
## as.factor(num.of.cylinders)six  -9.762e+03  2.168e+03  -4.502 1.21e-05 ***
## as.factor(num.of.cylinders)three -9.967e+03  4.449e+03  -2.240 0.026303 *
## as.factor(num.of.cylinders)twelve -2.161e+04  3.179e+03  -6.797 1.53e-10 ***
## bore                             -1.051e+02  5.955e+02  -0.176 0.860114
## stroke                           -2.187e+03  3.364e+02  -6.500 7.72e-10 ***
## compression.ratio                 3.416e+02  5.432e+01   6.289 2.38e-09 ***
## peak.rpm                          1.725e+00  5.213e-01   3.310 0.001129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2451 on 179 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.9146, Adjusted R-squared:  0.9074
## F-statistic: 127.7 on 15 and 179 DF,  p-value: < 2.2e-16
```

Promatrajući procijenjene doprinose varijablama prediktorima, zaključujemo da konjske snage, veličina motora, tip motora, broj cilindara, promjer cilindra, hod klipa, omjer kompresije i najveća okretajna brzina motora značajno doprinose određivanju cijene automobila. R-squared metrike govore nam da sastavljeni model objašnjava 91.46%, odnosno, 90.74% varijabilnosti. P-vrijednost je vrlo niska, što nam govori da je model statistički značajan.

```r
# Prikazivanje dijagnostike modela
# Residuals vs. Fitted plot
plot(model, which = 1, main = "Residuals vs. Fitted")
```




```r
# Normal Q-Q plot
qqnorm(resid(model))
qqline(resid(model), col = 2)
```



Marka automobila

Sastavljen je model koji koristi samo jednu varijablu - marku automobila. Model je vrlo jednostavan, no u kontekstu automobila smisleno je promatrati moć takvog modela.

KOD:

```
# Učitavanje podataka
data <- read.csv("car_specifications.csv")

data$make <- as.factor(data$make)
data$aspiration <- as.factor(data$aspiration)
data$num_of_doors <- as.factor(data$num.of.doors)
data$body_style <- as.factor(data$body.style)
data$drive_wheels <- as.factor(data$drive.wheels)
data$engine_location <- as.factor(data$engine.location)
data$fuel <- as.factor(data$fuel)
data$country <- as.factor(data$country)
data$continent <- as.factor(data$continent)
```

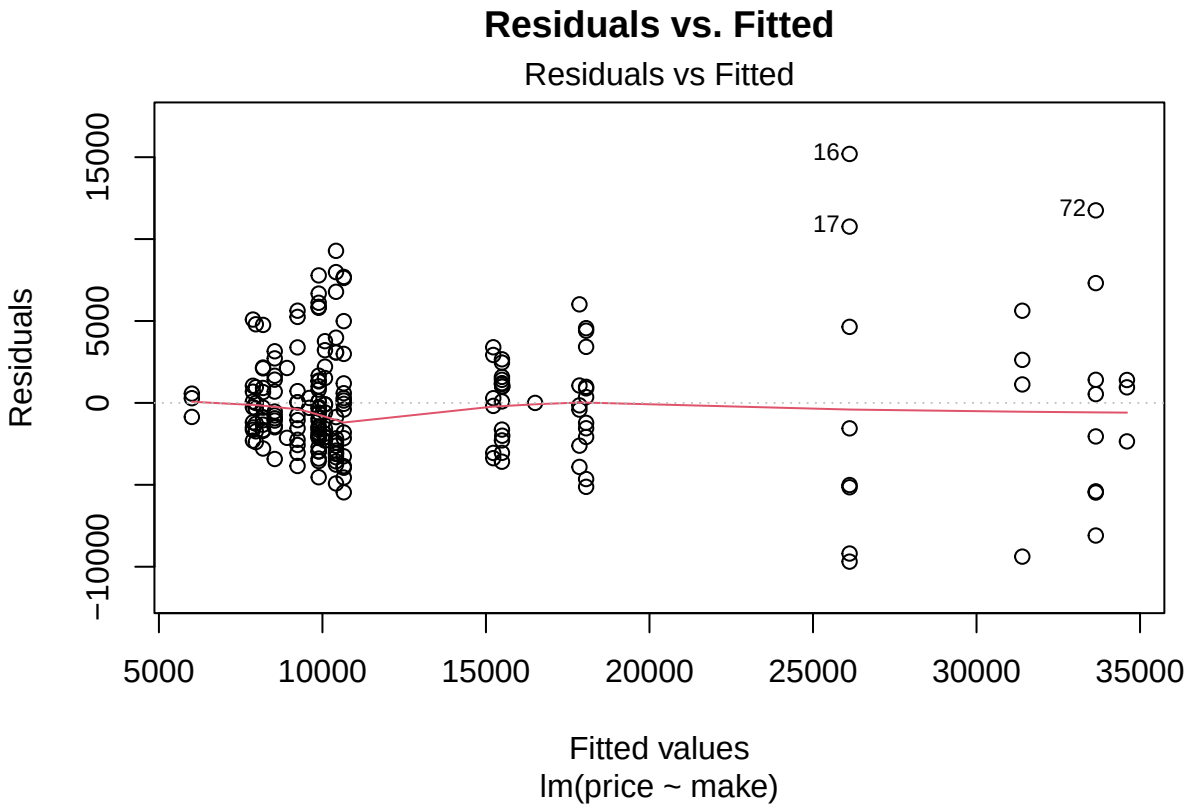
```
# sa levels printam sve kategorije ovih varijabli
levels(data$engine_location)
#levels(data$aspiration)

model <- lm(price ~ make, data = data)
```

```r
summary(model)
```

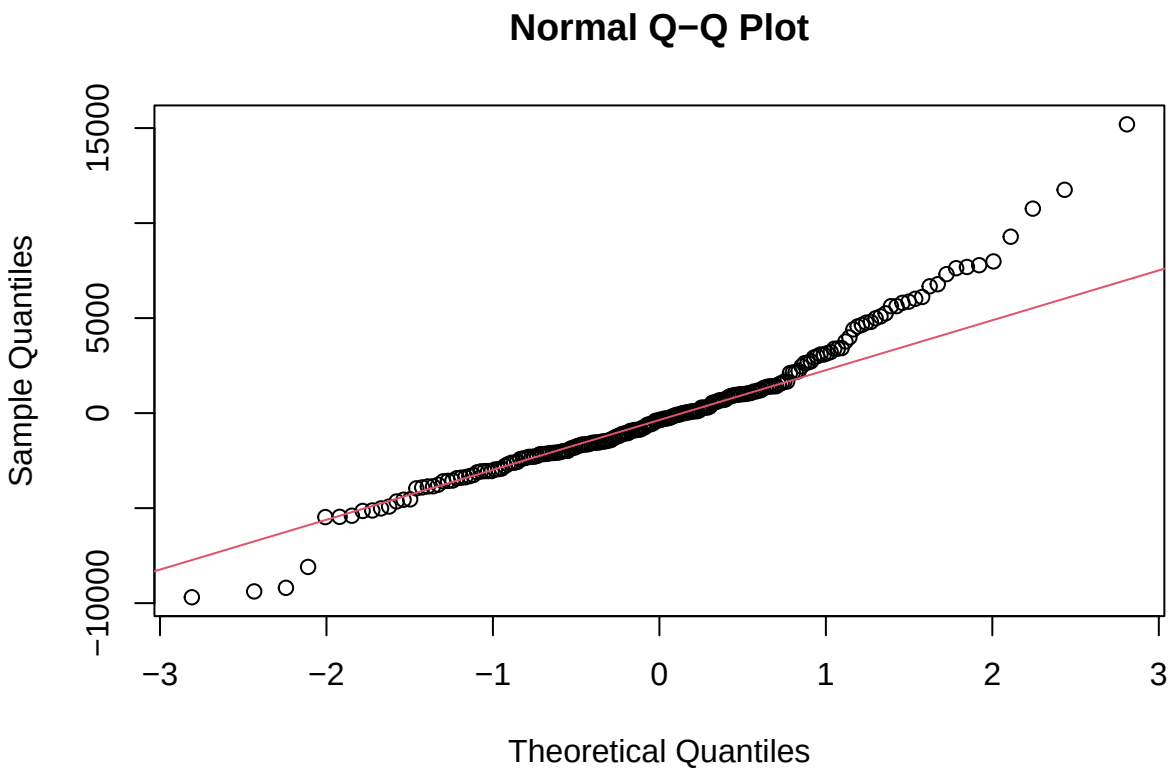
```
## [1] "front" "rear"
##
## Call:
## lm(formula = price ~ make, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9688.7 -2131.5  -354.4  1409.0 15196.2
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       15498.333   2191.253   7.073 3.30e-11 ***
## makeAudi           2360.833   2683.726   0.880  0.38021
## makeBMW           10620.417   2569.472   4.133 5.49e-05 ***
## makeChevrolet     -9491.333   3098.900  -3.063  0.00253 **
## makeDodge         -7622.889   2530.241  -3.013  0.00296 **
## makeHonda         -7313.641   2430.977  -3.009  0.00300 **
## makeIsuzu         -6581.833   3464.676  -1.900  0.05908 .
## makeJaguar        19101.667   3098.900   6.164 4.58e-09 ***
## makeMazda         -4845.451   2376.748  -2.039  0.04295 *
## makeMercedes-Benz 18148.667   2569.472   7.063 3.48e-11 ***
## makeMercury        1004.667   4382.507   0.229  0.81894
## makeMitsubishi    -6258.564   2430.977  -2.575  0.01085 *
## makeNissan        -5082.667   2366.824  -2.147  0.03310 *
## makePeugeot          -9.242   2472.067  -0.004  0.99702
## makePlymouth      -7534.905   2619.049  -2.877  0.00450 **
## makePorsche       15902.167   2898.756   5.486 1.39e-07 ***
## makeRenault       -5903.333   3464.676  -1.704  0.09014 .
## makeSaab           -275.000   2683.726  -0.102  0.91850
## makeSubaru        -6957.083   2449.896  -2.840  0.00504 **
## makeToyota        -5612.521   2291.668  -2.449  0.01528 *
## makeVolkswagen    -5420.833   2449.896  -2.213  0.02818 *
## makeVolvo          2564.848   2472.067   1.038  0.30089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3795 on 179 degrees of freedom
## Multiple R-squared:  0.7959, Adjusted R-squared:  0.7719
## F-statistic: 33.23 on 21 and 179 DF,  p-value: < 2.2e-16
```

Promatrajuci koeficijente zakljucujemo da sljedeće marke znacajno povećavaju cijenu automobila: BMW, Jaguar, Mercedes-Benz i Porsche. Marke koje smanjuju cijenu automobila (u usporedbi s referentnom markom) su: Honda, Mitsubishi, Subaru, Toyota, Volkswagen, Dodge i Chevrolet. R-squared metrika govori nam da model objašnjava 79.59% varijabilnosti u cijenama automobila, što je vrlo značajno za model sa samo jednom varijablom. F statistika ukazuje na značajnost modela, a p-vrijednost vrlo je niska.

```r
# Prikazivanje dijagnostike modela
# Residuals vs. Fitted plot
plot(model, which = 1, main = "Residuals vs. Fitted")
```




```r
# Normal Q-Q plot
qqnorm(resid(model))
qqline(resid(model), col = 2)
```



Metrike performansi

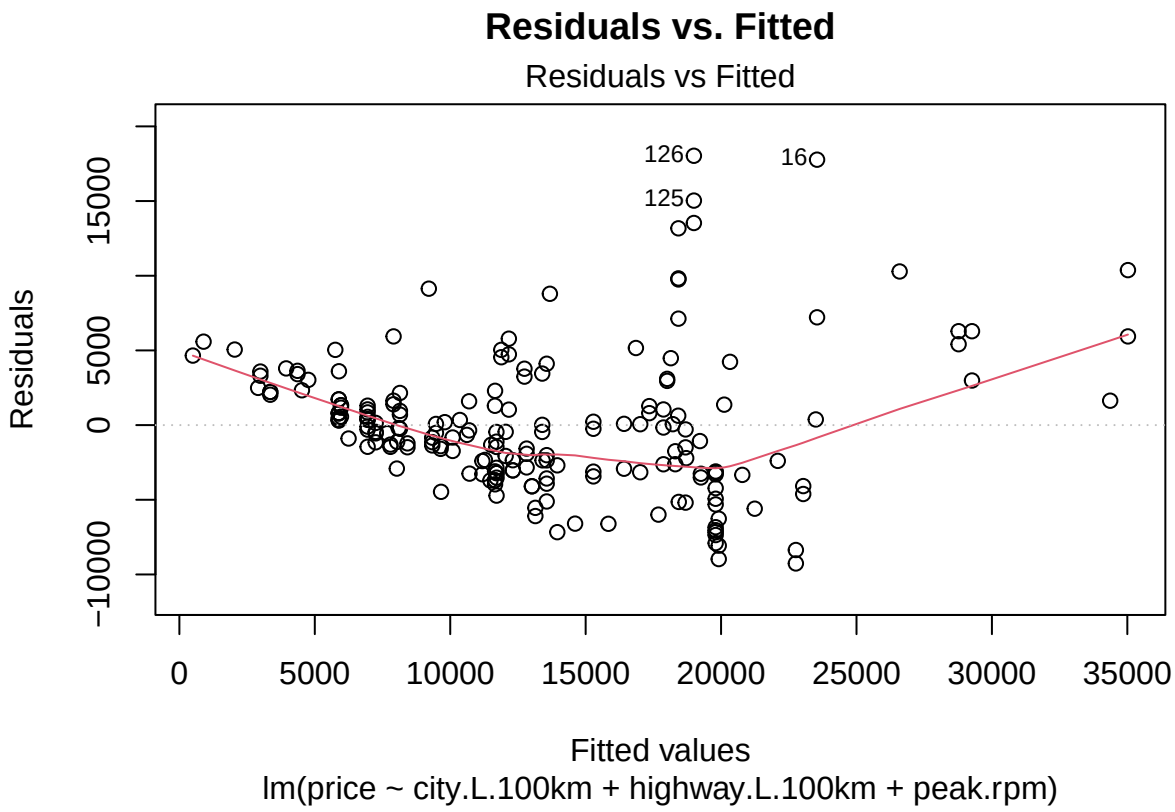
Regresijski model sastavljen je od sljedećih metrika: potrošnja goriva u gradskoj vožnji, potrošnja goriva na autocesti te maksimalan broj okretaja. Metrike su usko vezane uz performanse automobila.

KOD:

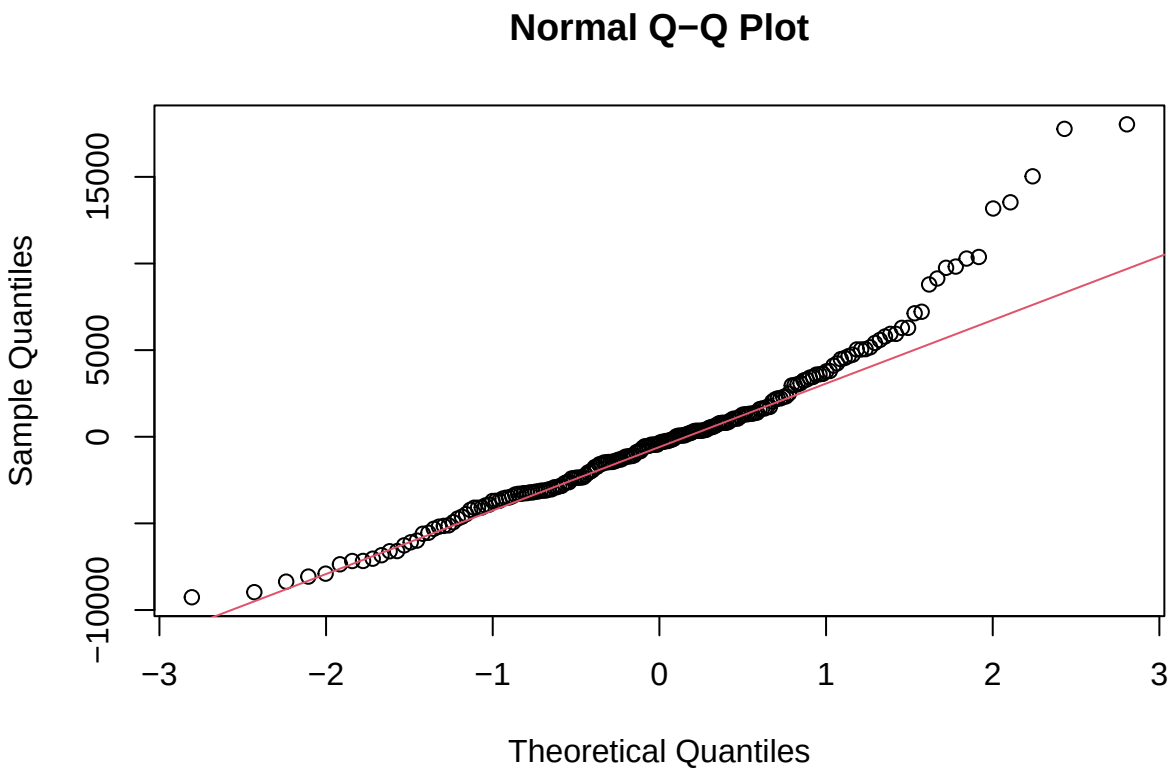
```
# Učitavanje podataka
data <- read.csv("car_specifications.csv")

data$make <- as.factor(data$make)
data$aspiration <- as.factor(data$aspiration)
data$num_of_doors <- as.factor(data$num.of.doors)
data$body_style <- as.factor(data$body.style)
data$drive_wheels <- as.factor(data$drive.wheels)
data$engine_location <- as.factor(data$engine.location)
data$fuel <- as.factor(data$fuel)
data$country <- as.factor(data$country)
data$continent <- as.factor(data$continent)
```

```
# sa levels printam sve kategorije ovih varijabli
#levels(data$make)
#levels(data$aspiration)

model <- lm(price ~ city.L.100km + highway.L.100km + peak.rpm, data = data)
```

```r
summary(model)
# Residuals vs. Fitted plot
plot(model, which = 1, main = "Residuals vs. Fitted")
```




```r
# Normal Q-Q plot
qqnorm(resid(model))
qqline(resid(model), col = 2)
```




```
##
## Call:
## lm(formula = price ~ city.L.100km + highway.L.100km + peak.rpm,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9262.2 -3064.7  -351.8  1878.3 18031.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      325.1097  4089.8995   0.079 0.936724
## city.L.100km    1526.5314   482.9152   3.161 0.001823 **
## highway.L.100km 1451.6345   660.8091   2.197 0.029217 *
## peak.rpm          -2.7238     0.7266  -3.749 0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4604 on 195 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6721, Adjusted R-squared:  0.6671
## F-statistic: 133.3 on 3 and 195 DF,  p-value: < 2.2e-16
```

Iako model slabije objašnjava varijabilnost u podacima od predhodno korištenih modela, gledajući F-statistiku primjećujemo da je model i dalje statistički znacajan. Manje je kompleksan od nekih predhodno

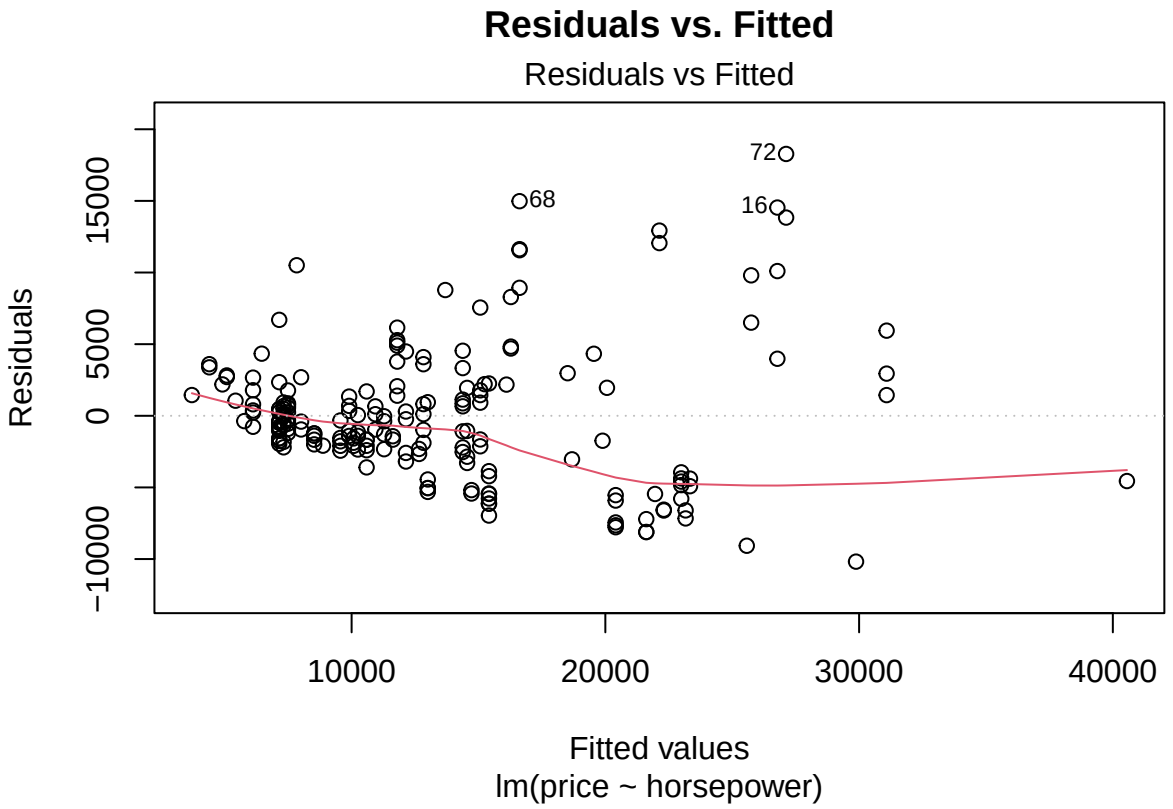
korištenih, sto ga čini interpretabilnijim. Varijabla koja je statistički najznačajnija je maksimalan broj okretaja motora. Model objašnjava 67.21% varijabilnosti podataka.

Konjske snage

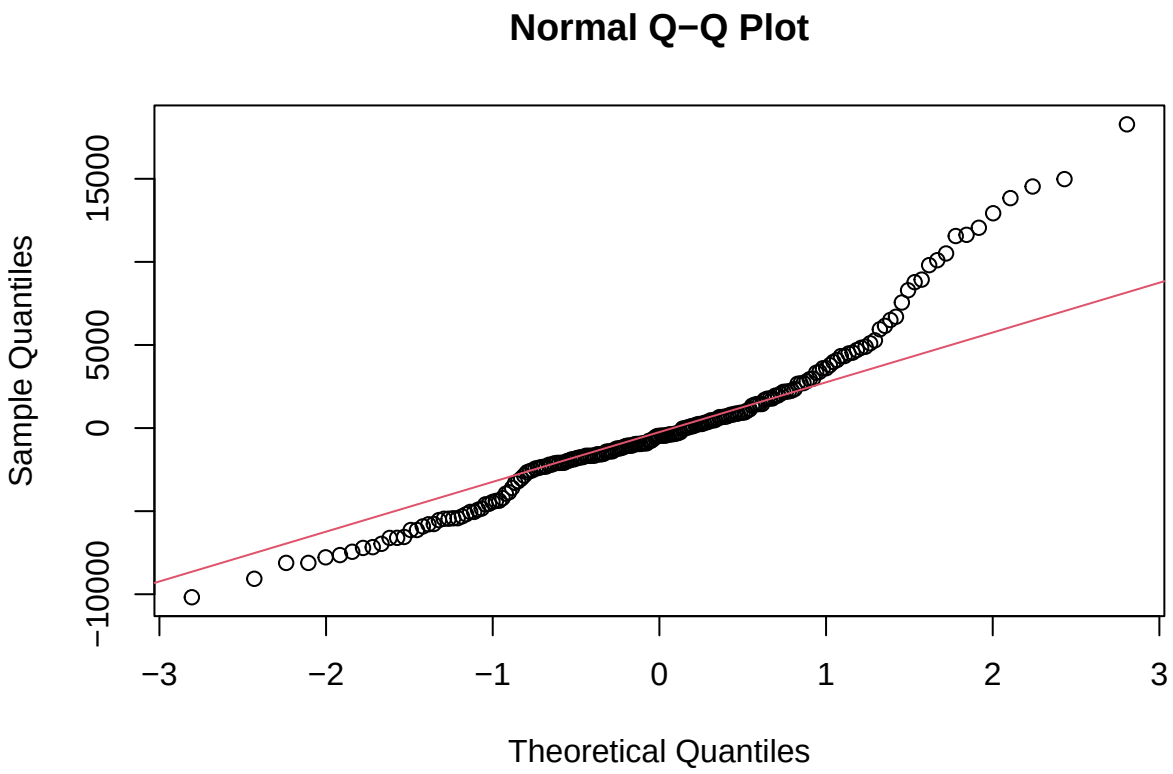
Model se sastoji od jedne varijable - konjskih snaga. Model je na prvi pogled prejednostavan, no u kontekstu automobila ima smisla promotriti ga (automobili sa više konjskih snaga motora u pravilu su skuplji).

KOD:

```
# Učitavanje podataka
data <- read.csv("car_specifications.csv")

data$make <- as.factor(data$make)
data$aspiration <- as.factor(data$aspiration)
data$num_of_doors <- as.factor(data$num.of.doors)
data$body_style <- as.factor(data$body.style)
data$drive_wheels <- as.factor(data$drive.wheels)
data$engine_location <- as.factor(data$engine.location)
data$fuel <- as.factor(data$fuel)
data$country <- as.factor(data$country)
data$continent <- as.factor(data$continent)

# sa levels printam sve kategorije ovih varijabli
#levels(data$make)
#levels(data$aspiration)

model <- lm(price ~ horsepower, data = data)

summary(model)
# Residuals vs. Fitted plot
plot(model, which = 1, main = "Residuals vs. Fitted")
```




```r
# Normal Q-Q plot
qqnorm(resid(model))
qqline(resid(model), col = 2)
```




```
##
## Call:
## lm(formula = price ~ horsepower, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10180.1  -2262.0   -471.1   1779.5  18276.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4562.175    974.995  -4.679 5.35e-06 ***
## horsepower    172.206      8.866  19.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4685 on 197 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.657,  Adjusted R-squared:  0.6552
## F-statistic: 377.3 on 1 and 197 DF,  p-value: < 2.2e-16
```

Model objašnjava 65.7% varijabilnosti u cijenama automobila, brojka je slična modelu koji koristi metrike performanse za predviđanje cijene koji smo promatrali maloprije. F statistika nam govori da je model statistički značajan (p vrijednost je vrlo mala). Iako model koristi samo jednu varijablu za predikciju cijene, multiple R-squared je relativno visok te pokazuje na dobru sposobnost objasnjavanja varijabilnosti u cijenama vozila.

Pitanje 4 - Analiza omjera kompresije izmedu atmosferskih motora i motora s turbopunjačem

Postoji li razlika u omjeru kompresije izmedu atmosferskih motora i motora s turbopunjačem?

1. Učitavanje potrebnih biblioteka i podataka

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

2. Filtriranje podataka

```
# Filtriram podatke prema tipu motora
atmosferski_motori <- podaci %>% filter(aspiration == "std")
turbopunjaci <- podaci %>% filter(aspiration == "turbo")
```

3. Deskriptivna statistika

```
# Izračun osnovne statističke mjere za omjer kompresije
summary(atmosferski_motori$compression.ratio)
summary(turbopunjaci$compression.ratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.50    8.70    9.00    9.59    9.40   23.00
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00    7.50    8.15   12.79   21.00   23.00
```

4. Testiranje hipoteza

```
# Testiranje razlike u omjeru kompresije između dvije grupe motora
t.test(atmosferski_motori$compression.ratio, turbopunjaci$compression.ratio)
```

```
##
##  Welch Two Sample t-test
##
## data:  atmosferski_motori$compression.ratio and turbopunjaci$compression.ratio
## t = -2.7559, df = 37.532, p-value = 0.008982
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.5585661 -0.8494743
## sample estimates:
## mean of x mean of y
##  9.590424 12.794444
```

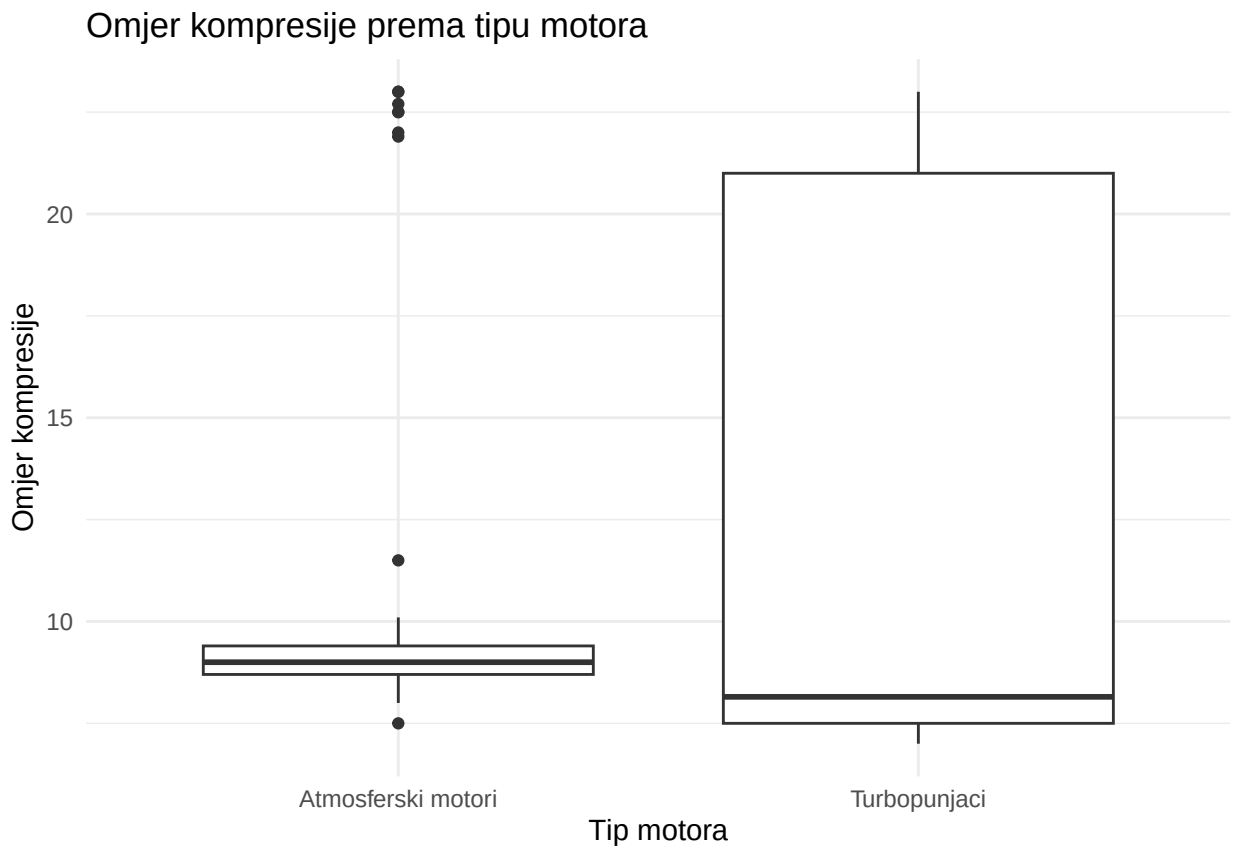
Ili, ako pretpostavke t-testa nisu zadovoljene, možemo koristiti Mann-Whitney U test:

```
wilcox.test(atmosferski_motori$compression.ratio, turbopunjaci$compression.ratio)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  atmosferski_motori$compression.ratio and turbopunjaci$compression.ratio
## W = 3517.5, p-value = 0.0813
## alternative hypothesis: true location shift is not equal to 0
```

5. Vizualizacija

```
# Vizualizacija razlika u omjeru kompresije između dvije grupe motora
ggplot() +
  geom_boxplot(data = rbind(atmosferski_motori %>% mutate(Type = "Atmosferski motori"),
                            turbopunjaci %>% mutate(Type = "Turbopunjaci")),
               aes(x = Type, y = compression.ratio)) +
  labs(title = "Omjer kompresije prema tipu motora",
       x = "Tip motora",
       y = "Omjer kompresije") +
  theme_minimal()
```



6. Zaključak:

Na temelju provedenog Wilcoxonovog testa rangova s korekcijom kontinuiteta, ne možemo odbaciti nultu hipotezu na razini značajnosti od 0,05. P-vrijednost iznosi 0,0813, što je veće od konvencionalnog praga značajnosti od 0,05.

To sugerira da nema statistički značajne razlike u omjeru kompresije između atmosferskih motora i motora s turbopunjačem na razini značajnosti od 0,05.

Međutim, važno je napomenuti da je p-vrijednost vrlo blizu pragu značajnosti od 0,05, što znači da postoji mala vjerojatnost (8,13%) da bismo dobili ovakav ili ekstremniji uzorak ako je nulta hipoteza istinita. U praksi, ovo bi moglo sugerirati potrebu za daljnjim istraživanjem ili povećanjem veličine uzorka kako bismo s većom sigurnošću potvrdili ove rezultate.

Stoga, na temelju dostupnih podataka, ne možemo potvrditi da postoji razlika u omjeru kompresije između atmosferskih motora i motora s turbopunjačem.