

SAP projekt: Analiza specifikacija automobila

Fran Lubina, Zvonimir Stracenski, Luka Varga, Karlo Vešligaj

2026-01-23

Cilj i opis projekta

U okviru ovog projekta naglasak će biti na statističko zaključivanje vezano uz prethodne marketinške kampanje poduzeća, što je bitan korak u planiranju budućih kampanja. Tradicionalne masovne kampanje, poput reklamnih panoa, tipično imaju vrlo nisku uspješnost koja prema provedenim studijama često iznosi i ispod 1%. S druge strane, ciljani marketing često pokazuje značajno veću učinkovitost zbog usmjerenosti na kupce koji su skloniji kupnji određenih proizvoda i usluga.

U okviru ovog projekta naglasak će biti na statističko zaključivanje vezano uz specifikacije automobila, što je bitan korak u planiranju što objektivnijih odluka o modelu koji odgovara svim zahtjevima kupca. Automobilska industrija kontinuirano evoluirala, s naglaskom na ekološku održivost, sigurnost i tehnološku inovaciju. Razumijevanje odnosa između tehničkih specifikacija automobila i njihovih performansi ključno je kako za kupce tako i za proizvođače. Statističkom analizom podataka o automobilima moguće je identificirati ključne faktore koji utječu na cijenu, potrošnju goriva, i ukupne performanse vozila.

Inicijalni pregled i obrada podataka

```
# Učitavanje dataseta
my_cars <- read.csv("car_specifications.csv")
```

Podatci se sastoje od specifikacija automobila za 205 različitih modela od 22 proizvođača. Skup sadrži tehničke karakteristike i tržišne varijable s ukupno 26 atributa prikupljenih iz autoindustrije. Podaci uključuju dimenzije automobila (duljina, širina, visina), međuosovinskog razmak, obujam i snagu motora, vrstu pogonskog goriva, cijenu, broj vrata, potrošnju goriva u gradu i na autocesti, tip pogona (prednji, stražnji, 4WD) i druge relevantne specifikacije:

```
dim(my_cars)
```

```
## [1] 201 26
```

```
names(my_cars)
```

```
## [1] "make"           "aspiration"      "num.of.doors"
## [4] "body.style"     "drive.wheels"    "engine.location"
## [7] "wheel.base"     "length"          "width"
## [10] "height"         "curb.weight"     "engine.type"
## [13] "num.of.cylinders" "engine.size"     "fuel.system"
```

```
## [16] "bore"           "stroke"           "compression.ratio"
## [19] "horsepower"     "peak.rpm"         "price"
## [22] "city.L.100km"   "highway.L.100km"  "fuel"
## [25] "country"        "continent"
```

Prikažimo prvih nekoliko redaka:

```
head(my_cars)
```

```
##      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
```

Analiza podataka

Podatke zatim analiziramo i interpretiramo pomoću statističkih metoda vodeći se prethodno definiranim pitanjima.

Pitanje 1: Razlikuje li se snaga motora između automobila s turbopunjačem i atmosferskim motorima?

Najprije izračunamo mjere centralne tendencije za snagu motora ovisno o tipu usisavanja zraka.

```
summary.result1 <- my_cars %>%
  group_by(aspiration) %>%
  summarise(
    count = n(),
    mean_horsepower = mean(horsepower, na.rm = TRUE),
    median_horsepower = median(horsepower, na.rm = TRUE),
    sd_horsepower = sd(horsepower, na.rm = TRUE),
    min_horsepower = min(horsepower, na.rm = TRUE),
    max_horsepower = max(horsepower, na.rm = TRUE),
    q25 = quantile(horsepower, 0.25, na.rm = TRUE),
    q75 = quantile(horsepower, 0.75, na.rm = TRUE)
  )

summary.result1
```

```
## # A tibble: 2 x 9
##   aspiration count mean_horsepower median_horsepower sd_horsepower
##   <chr>      <int>      <dbl>          <dbl>          <dbl>
## 1 std         165        99.0            88            37.5
## 2 turbo        36       123.           120.           31.1
## # i 4 more variables: min_horsepower <int>, max_horsepower <int>, q25 <dbl>,
## #   q75 <dbl>
```

Postoje indikacije da bi motori s turbopunjačem trebali imati veću snagu od atmosferskih motora.

Ovakvo ispitivanje možemo provesti t-testom.

Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo dva uzoraka za motore koji se nalaze u različitim automobilima, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka koju provjeravamo histogramom.

```
clean_horsepower_std <- na.omit(my_cars[my_cars$aspiration == "std", ]$horsepower)
clean_horsepower_turbo <- na.omit(my_cars[my_cars$aspiration == "turbo", ]$horsepower)

xrange <- range(c(clean_horsepower_std, clean_horsepower_turbo))

ymax <- max(
  hist(clean_horsepower_std, plot = FALSE)$counts,
  hist(clean_horsepower_turbo, plot = FALSE)$counts
)

par(mfrow = c(1, 2))

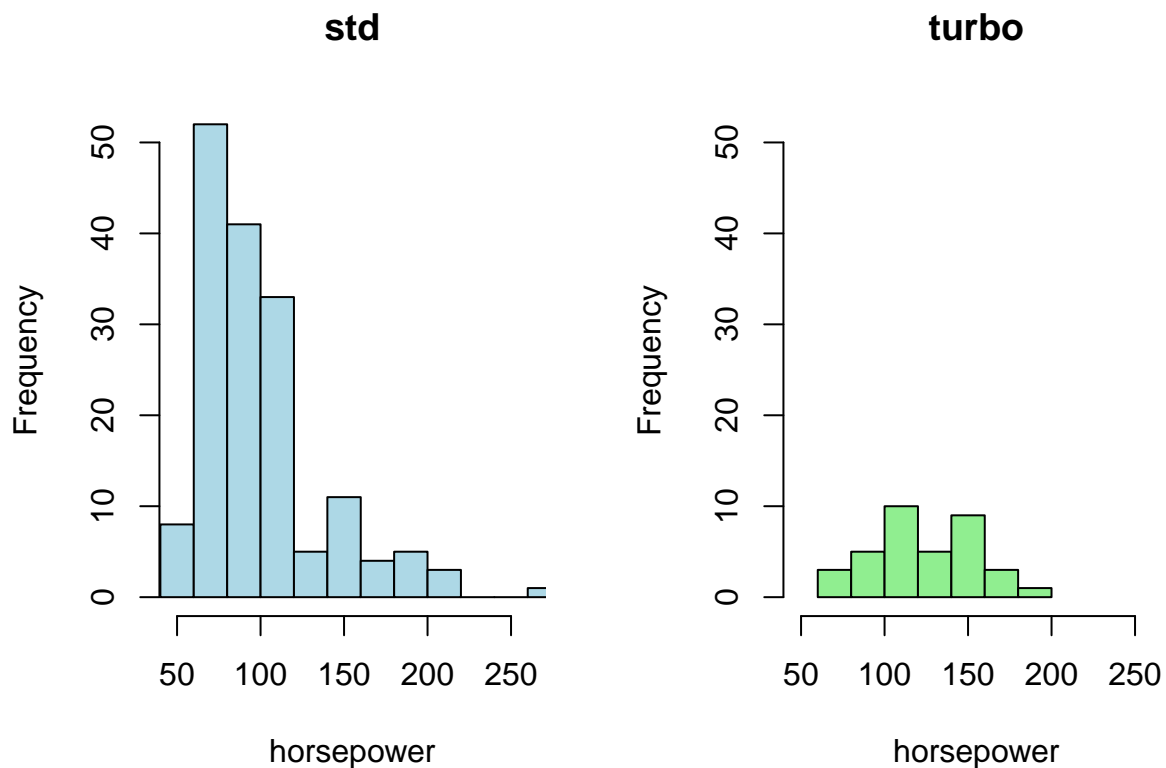
hist(clean_horsepower_std,
  main = "std",
  xlab = "horsepower",
  ylab = "Frequency",
  col = "lightblue",
```

```

border = "black",
xlim = xrange,
ylim = c(0, ymax))

hist(clean_horsepower_turbo,
main = "turbo",
xlab = "horsepower",
ylab = "Frequency",
col = "lightgreen",
border = "black",
xlim = xrange,
ylim = c(0, ymax))

```



Zbog prisutnosti ekstremnih vrijednosti i narušene pretpostavke o normalnosti distribucije (uočene vizualnim pregledom histograma), umjesto t-testa korišten je Mann-Whitney-Wilcoxonov test kao robusnija neparametrijska alternativa za usporedbu dviju nezavisnih skupina.

Iz gornjih histograma možemo zaključiti da podatci u dvije navedene grupe nisu normalno distribuirani.

S obzirom da podatci nisu normalno distribuirani, primijenit ćemo Mann-Whitney-Wilcoxonov test.

Mann-Whitney-Wilcoxonov test

Hipoteze:

H_0 : Ne postoji značajna razlika u snazi motora, odnosno snaga turbo motora je manja ili jednaka snazi atmosferskih motora

H_1 : Snaga motora automobila s turbopunjačem značajno je veća od snage automobila s atmosferskim motorom ($Mdn_{turbo} > Mdn_{std}$)

```
wilcox.test(clean_horsepower_turbo,
            clean_horsepower_std,
            alternative = "greater",
            conf.int = TRUE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  clean_horsepower_turbo and clean_horsepower_std
## W = 4306.5, p-value = 5.652e-06
## alternative hypothesis: true location shift is greater than 0
## 95 percent confidence interval:
##  19.00006      Inf
## sample estimates:
## difference in location
##                28.99995
```

Zaključak

Mann-Whitney-Wilcoxonovim testom utvrđeno je da automobili s turbopunjačem imaju statistički značajno veću snagu motora u usporedbi s automobilima s atmosferskim motorom ($p < 0.001$). Na temelju dobivene p-vrijednosti, odbacujemo nultu hipotezu (H_0) u korist alternativne (H_1).

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)

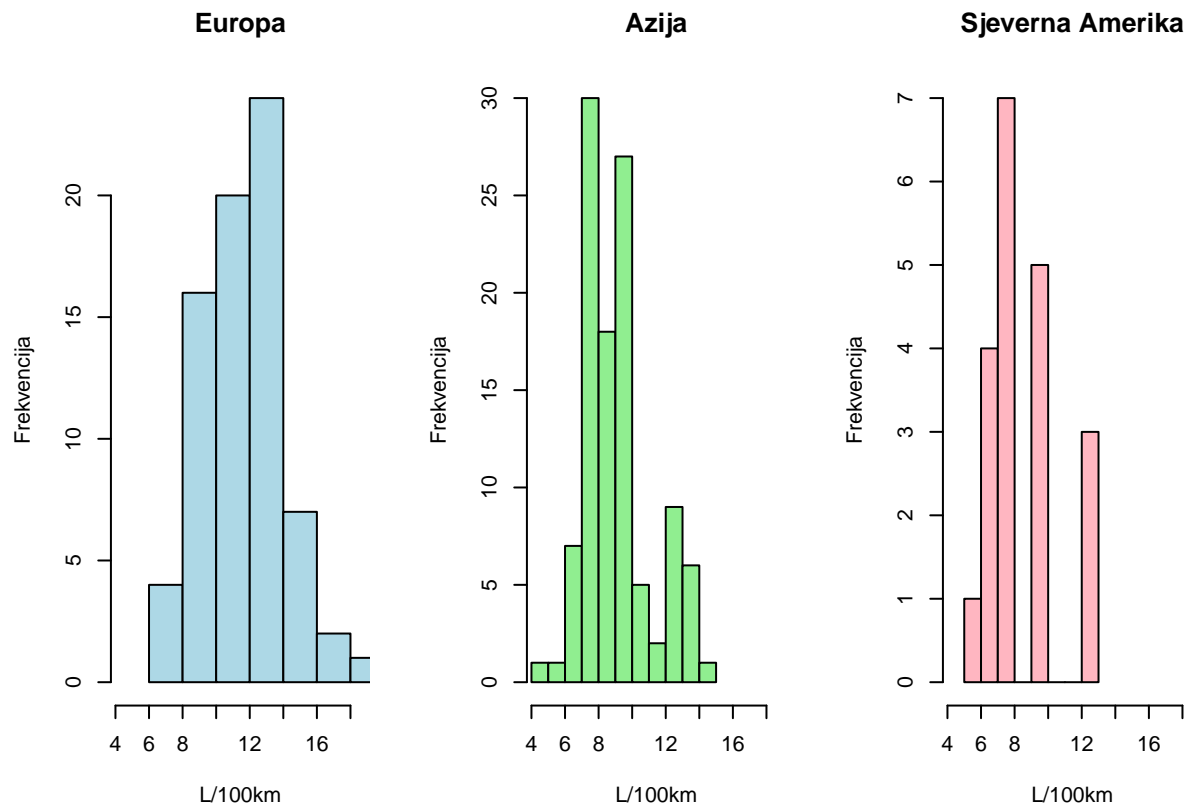
hist(my_cars[my_cars$continent == "Asia", ]$city.L.100km,
  main = "Azija",
  xlab = "L/100km",
  ylab = "Frekvencija",
  col = "lightgreen",
  border = "black",
```

```

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)

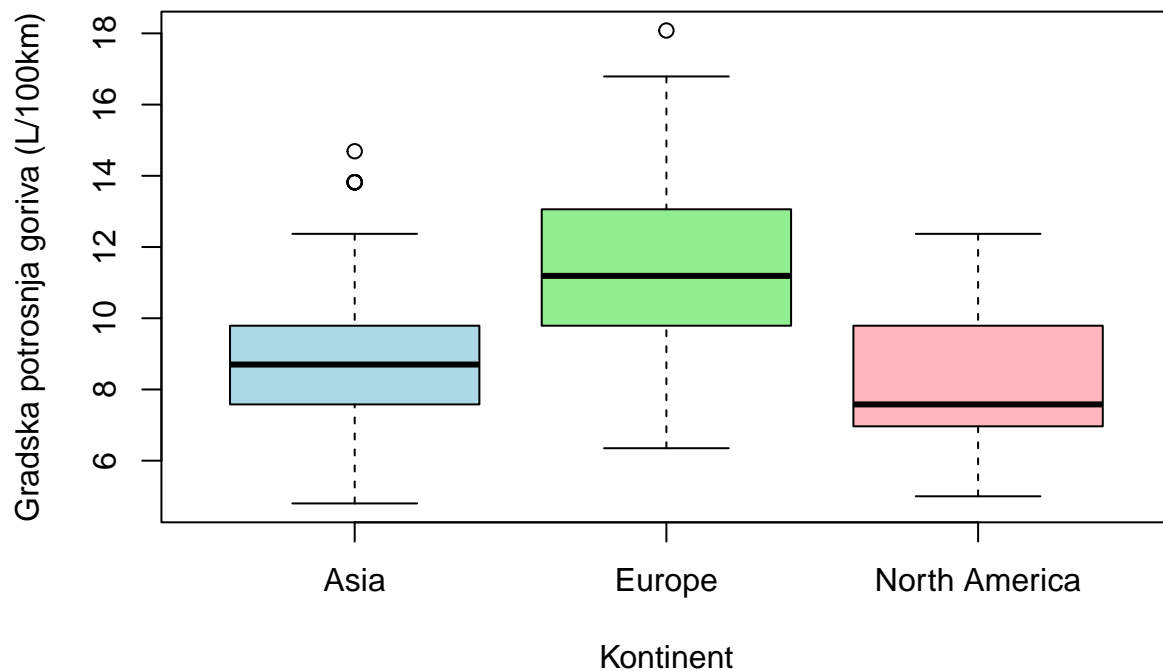
```



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")
```

Gradska potrošnja goriva po kontinentu



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 stršućih 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",
```



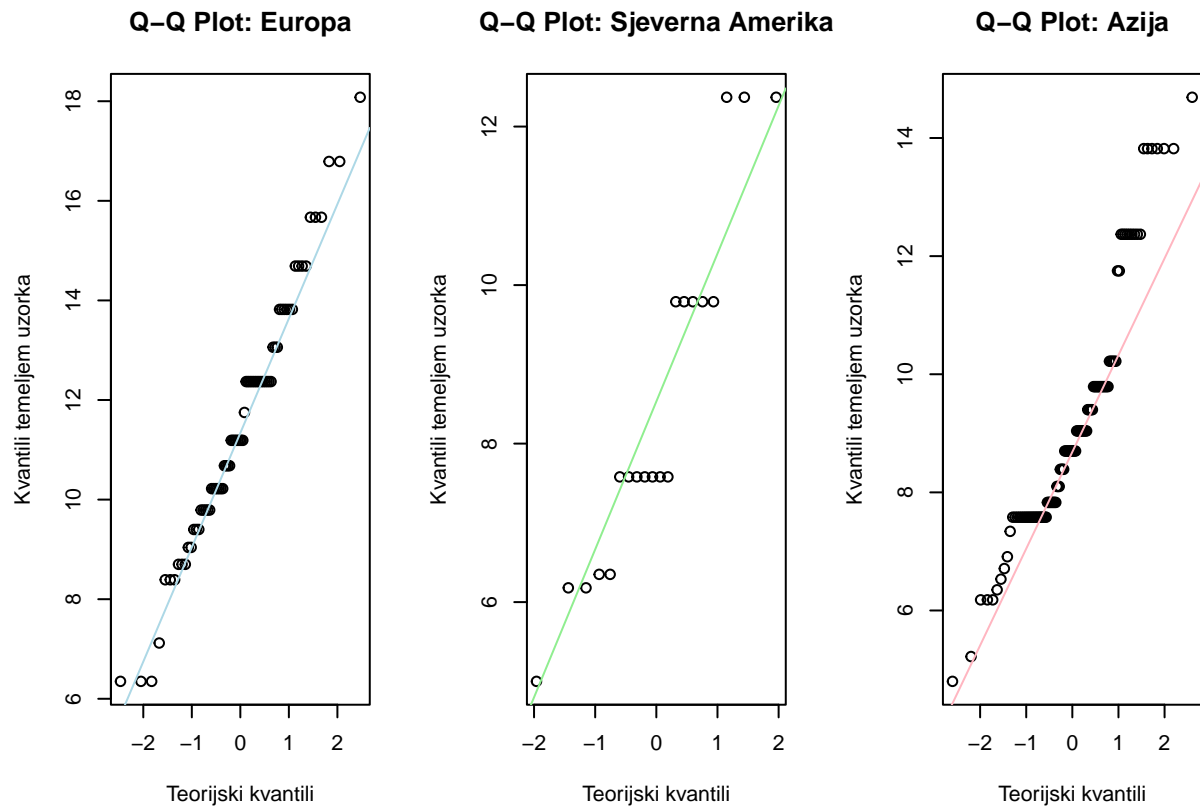
```

        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 proizvoljnu podjelu podataka. Logaritamska transformacija nije uspjela postići normalnost podataka za Aziju, što dodatno opravdava primjenu neparametarskog Kruskal-Wallisova testa.

Pitanje 3: Možemo li predvidjeti cijenu automobila na temelju dimenzija (length, width), snage motora (horsepower), obujma motora (engine-size) i gradske potrošnje goriva?

Predviđanje cijene vozila

Istražujemo može li se cijena vozila predvidjeti na temelju snage motora (horsepower), veličine motora (engine.size), gradske potrošnje goriva (city.L.100km), dužine (length) i širine (width) vozila. Primjenjuje se višestruka linearna regresija jer je zavisna varijabla kontinuirana, a koristi se više kvantitativnih regresora.

Eksplorativna analiza

```
library(dplyr)
library(ggplot2)

cars_model <- my_cars %>%
  select(price, horsepower, engine.size, city.L.100km, length, width)

summary(cars_model)
```

```
##      price      horsepower      engine.size      city.L.100km
##  Min.   : 5118   Min.   : 48.0   Min.   :1000   Min.   : 4.800
## 1st Qu.: 7775   1st Qu.: 70.0   1st Qu.:1606   1st Qu.: 7.830
## Median :10295   Median : 95.0   Median :1966   Median : 9.790
## Mean   :13207   Mean   :103.4   Mean   :2079   Mean   : 9.944
## 3rd Qu.:16500   3rd Qu.:116.0   3rd Qu.:2311   3rd Qu.:12.370
## Max.   :45400   Max.   :262.0   Max.   :5342   Max.   :18.080
##
##           NA's      :2
##      length      width
##  Min.   :358.4   Min.   :153.2
## 1st Qu.:423.7   1st Qu.:162.8
## Median :439.9   Median :166.4
## Mean   :442.5   Mean   :167.4
## 3rd Qu.:466.1   3rd Qu.:169.2
## Max.   :528.6   Max.   :182.9
##
```

```
library(GGally)

ggpairs(cars_model,
  upper = list(continuous = wrap("cor", size = 4)),
  lower = list(continuous = wrap("points", alpha = 0.5, size = 1)),
  diag = list(continuous = wrap("barDiag", fill = "lightblue"))) +
  theme_minimal()
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value 'binwidth'.
```

```
## Warning: Removed 2 rows containing missing values
```

```

## Warning: Removed 2 rows containing missing values or values outside the scale range
## ('geom_point()').

## 'stat_bin()' using 'bins = 30'. Pick better value 'binwidth'.

## Warning: Removed 2 rows containing non-finite outside the scale range
## ('stat_bin()').

## Warning: Removed 2 rows containing missing values
## Removed 2 rows containing missing values
## Removed 2 rows containing missing values
## Removed 2 rows containing missing values

## Warning: Removed 2 rows containing missing values or values outside the scale range
## ('geom_point()').

## 'stat_bin()' using 'bins = 30'. Pick better value 'binwidth'.

## Warning: Removed 2 rows containing missing values or values outside the scale range
## ('geom_point()').

## 'stat_bin()' using 'bins = 30'. Pick better value 'binwidth'.

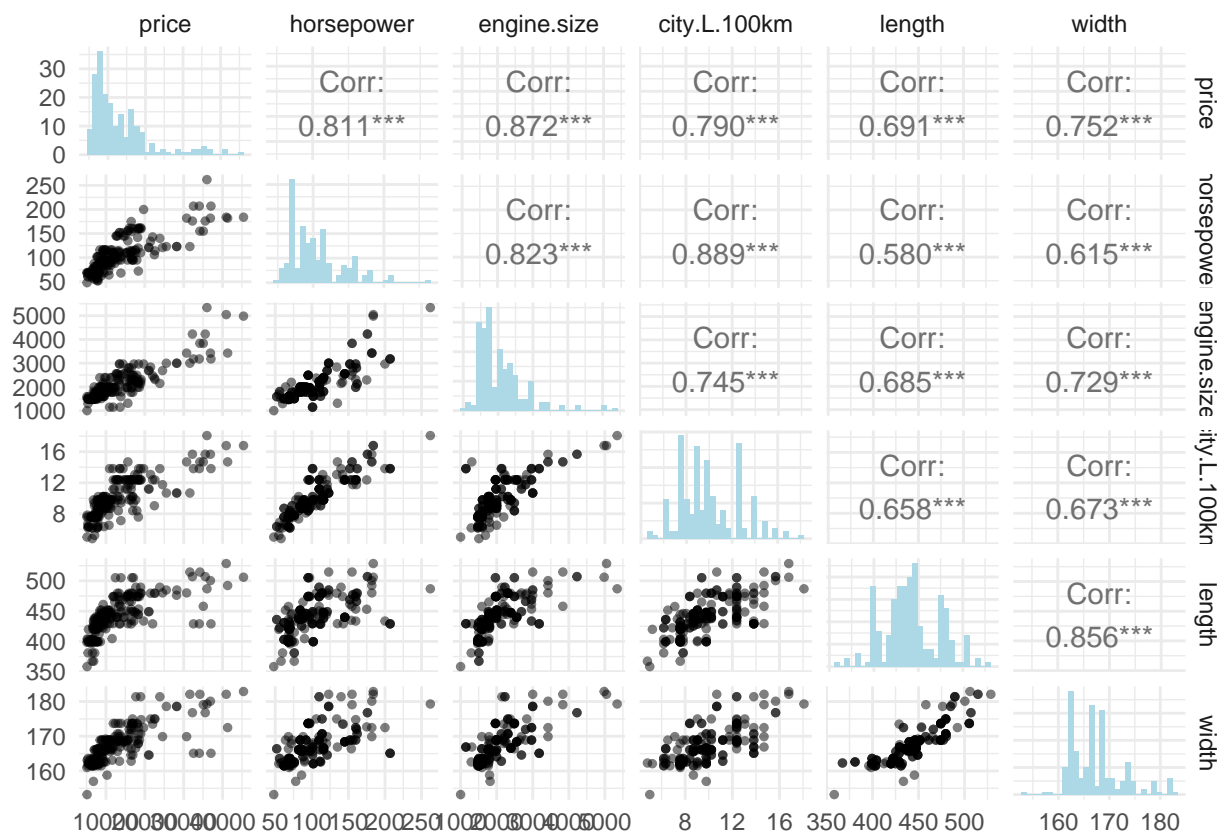
## Warning: Removed 2 rows containing missing values or values outside the scale range
## ('geom_point()').

## 'stat_bin()' using 'bins = 30'. Pick better value 'binwidth'.

## Warning: Removed 2 rows containing missing values or values outside the scale range
## ('geom_point()').

## 'stat_bin()' using 'bins = 30'. Pick better value 'binwidth'.

```



Višestruka linearna regresija

```
model <- lm(
  price ~ horsepower + engine.size + city.L.100km + length + width,
  data = cars_model
)

summary(model)
```

```
##
## Call:
## lm(formula = price ~ horsepower + engine.size + city.L.100km +
##     length + width, data = cars_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8807.3 -1883.0   201.1  1375.1 14115.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.638e+04  1.160e+04  -4.859 2.44e-06 ***
## horsepower    2.348e+01  1.729e+01   1.358 0.17609
## engine.size    5.848e+00  7.322e-01   7.988 1.22e-13 ***
```

```
## city.L.100km 6.000e+02 2.289e+02 2.621 0.00946 **
## length      -3.442e+00 1.544e+01 -0.223 0.82381
## width       3.024e+02 9.533e+01 3.172 0.00176 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3399 on 193 degrees of freedom
## (2 observations deleted due to missingness)
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8185
## F-statistic: 179.6 on 5 and 193 DF, p-value: < 2.2e-16
```

```
model<-lm(
price~ engine.size+width,
data=cars_model
)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ engine.size + width, data = cars_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7971.4 -2094.5  -309.6  1451.6 15760.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.507e+04  1.108e+04  -5.874 1.78e-08 ***
## engine.size  8.085e+00  5.565e-01  14.530 < 2e-16 ***
## width       3.672e+02  7.105e+01   5.169 5.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3665 on 198 degrees of freedom
## Multiple R-squared:  0.7894, Adjusted R-squared:  0.7873
## F-statistic: 371.1 on 2 and 198 DF, p-value: < 2.2e-16
```

```
model<-lm(
price~ engine.size+city.L.100km+width,
data=cars_model
)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ engine.size + city.L.100km + width, data = cars_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8999.6 -1911.0   93.4  1349.0 14224.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -5.152e+04 1.061e+04 -4.855 2.45e-06 ***
## engine.size 6.414e+00 5.998e-01 10.693 < 2e-16 ***
## city.L.100km 8.276e+02 1.490e+02 5.553 9.02e-08 ***
## width 2.579e+02 6.910e+01 3.732 0.000248 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3417 on 197 degrees of freedom
## Multiple R-squared: 0.8179, Adjusted R-squared: 0.8151
## F-statistic: 294.9 on 3 and 197 DF, p-value: < 2.2e-16
```

```
model<-lm(
price~ horsepower+width,
data=cars_model
)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ horsepower + width, data = cars_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8499.7 -2402.9  -86.1  1691.4 15605.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.015e+05  1.046e+04  -9.704  <2e-16 ***
## horsepower   1.185e+02  9.394e+00  12.611  <2e-16 ***
## width        6.123e+02  6.586e+01   9.296  <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3913 on 196 degrees of freedom
## (2 observations deleted due to missingness)
## Multiple R-squared: 0.7619, Adjusted R-squared: 0.7595
## F-statistic: 313.6 on 2 and 196 DF, p-value: < 2.2e-16
```

Jednostavna provjera povezanosti regresora

```
cor(cars_model, use = "complete.obs")
```

```
##           price horsepower engine.size city.L.100km  length  width
## price      1.0000000  0.8105331  0.8738902  0.7912969 0.6940555 0.7541672
## horsepower 0.8105331  1.0000000  0.8226605  0.8894963 0.5803396 0.6153860
## engine.size 0.8738902  0.8226605  1.0000000  0.7450069 0.6851805 0.7293491
## city.L.100km 0.7912969  0.8894963  0.7450069  1.0000000 0.6577905 0.6734888
## length     0.6940555  0.5803396  0.6851805  0.6577905 1.0000000 0.8564603
## width      0.7541672  0.6153860  0.7293491  0.6734888 0.8564603 1.0000000
```

Matrica korelacija koristi se kao jednostavna provjera međusobne povezanosti varijabli. Umjerene korelacije među regresorima ne predstavljaju problem, dok vrlo visoke korelacije mogu ukazivati na redundantne informacije.

Rezultati i interpretacija

Dobiveni koeficijent determinacije iznosi $R^2 \approx 0.82$, a prilagođeni $R^2 \approx 0.81$, što znači da model objašnjava oko 82% varijabilnosti cijene vozila. To ukazuje na snažnu linearnu povezanost između cijene i korištenih regresora.

Pojedinačni regresori pokazuju sljedeće: Veličina i snaga motora te potrošnja statistički značajno utječu na cijenu. Dimenzije također imaju značajan pozitivan utjecaj, što pokazuje da veći automobili imaju veću cijenu.

Snaga motora je nesignifikantna zbog snažne povezanosti s veličinom motora. Vidimo da je p-vrijednosti njegovog koeficijenta velika ($p \approx 0.18$). Dužina je također nesignifikatna, vjerojatno zbog redundatnosti sa drugom dimenzijom veličine - širinom. Vidimo veliku korelaciju između njih te veliku p-vrijednost njenog regresijskog koeficijenta.

Iako korelacijska matrica ukazuje na visku koreliranost između potrošnje i veličine, odnosno snage motor, vidimo da izbacivanjem potrošnje iz modela dobivamo lošiji koeficijent determinacije.

$R_{adj}^2 \approx R^2$ potvrđuje da je model prikladan i da dodavanje novih regresora opravdava svoje mjesto u modelu. Posebice u modelu koji ne uzima u obzir redundatne regresore.

Zaključak

Model objašnjava visok udio varijabilnosti cijene, što ukazuje na snažnu linearnu povezanost između cijene i tehničkih karakteristika automobila. Automobili veće snage, veće veličine i veće potrošnje goriva obično imaju veću cijenu, dok prilagođeni koeficijent determinacije ($R_{adj}^2 \approx R^2$) potvrđuje da je model statistički pouzdan i da nije rezultat pukog dodavanja regresora.

Pitanje 4: Postoji li veza između tipa pogona (prednji vs. stražnji) i tipa karoserije (sedan vs. hatchback)?

Cilj zadatka je utvrditi da li postoji povezanost između tipa pogona (prednji ili stražnji) i tipa karoserije (sedan ili hatchback). Koristimo χ^2 test za neovisnost. Prvo moramo učitati i obraditi naše podatke.

4.1. Unos podataka

Prvo obradimo tablicu tako da maknemo stupce koji nas ne zanimaju i ostavimo one koje trebamo (tip pogona i tip karoserije)

```
data2 = select(my_cars, c("body.style", "drive.wheels"))
```

Za svaki slučaj mićemo podatke gdje je jedna od varijabli nedefinirana.

```
data2 = na.omit(data2)
```

Zatim filtriramo naše podatke tako da ostanu samo oni tipovi pogona i karoserije koje mi koristimo (prednji i stražnji za pogon, hatchback i sedan za karoseriju).

```
#filtriraj za tip karoserije
data2 = filter(data2, is.element(body.style, c("hatchback", "sedan")))

#filtriraj za tip pogona
data2 = filter(data2, is.element(drive.wheels, c("fwd", "rwd")))
```

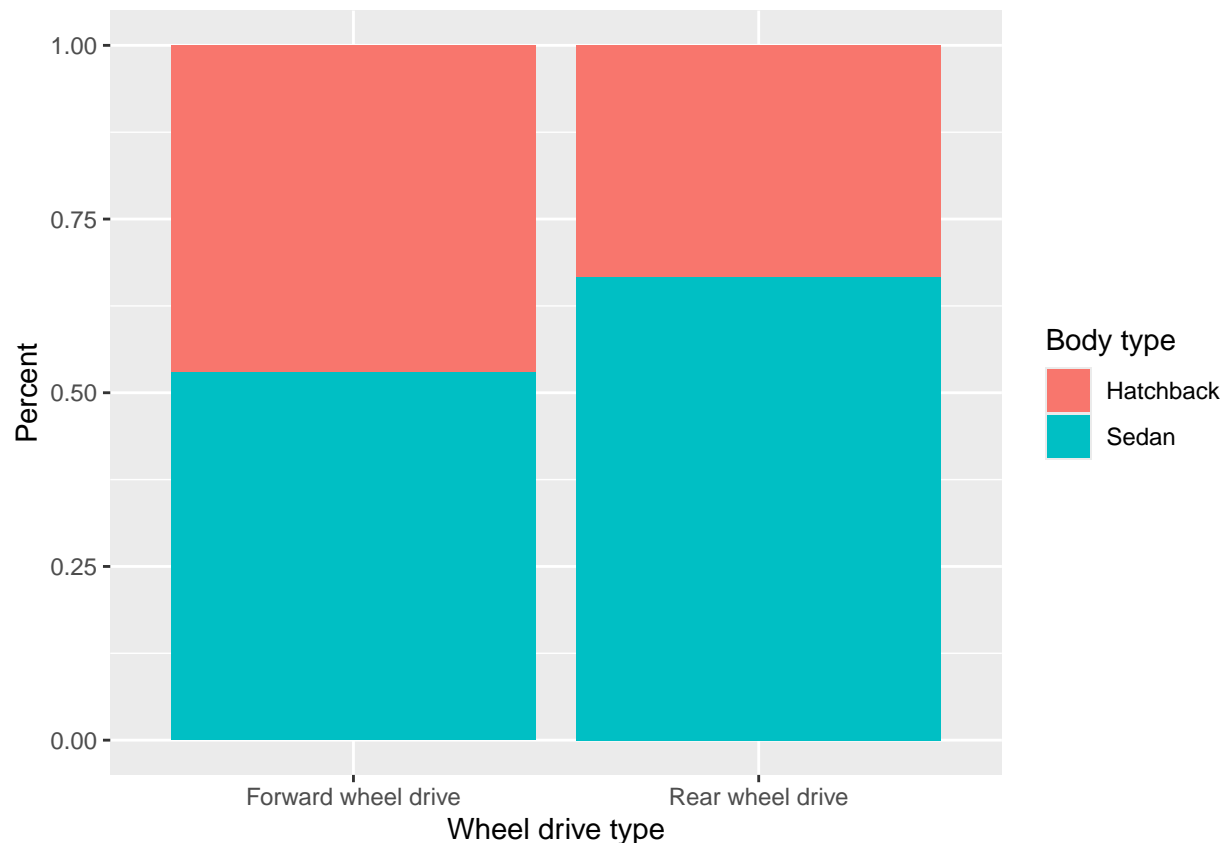
4.2. χ^2 test

Generiramo kontingencijsku tablicu da proučimo podatke i graf za vizualizaciju odnosa podataka međusobno.

```
tab = table(data2$body.style, data2$drive.wheels)
tab
```

```
##
##           fwd rwd
## hatchback  49  18
## sedan      55  36
```

```
#graf
library(ggplot2)
xx = c(rep("Forward wheel drive", 2), rep("Rear wheel drive", 2))
yy = rep(c("Hatchback", "Sedan"), 2)
tabpl = tab[1:4]
datapl = data.frame(tab)
ggplot(datanpl, aes(fill=yy, x=xx, y=tabpl)) +
  geom_bar(position = "fill", stat = "identity") +
  xlab("Wheel drive type") +
  ylab("Percent") +
  scale_fill_discrete(name = "Body type")
```



Vidimo da je jedina prava pretpostavka za χ^2 test (svaka observacija u tablici iznad 5) ispunjena. Druga pretpostavka, isključivost kategorija, implicitno ispunjujemo prema značenju kategorija (npr. auto ne može istovremeno biti hatchback i sedan). Graf nam ukazuje da je moguće da postoji neka povezanost između tipa karoserije i pogona.

Možemo postaviti našu hipotezu:

$$H_0 : o_i = e_i, i = \{1, \dots, k\}$$

$$H_1 : o_i \neq e_i, \exists i$$

$$\alpha = 0.05$$

gdje je k ukupan broj svih kategorija.

Zatim jednostavno napravimo χ^2 test.

```
result = chisq.test(tab)
result
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 2.2289, df = 1, p-value = 0.1355
```

Vidimo da je dobivena p-vrijednost $> \alpha$.

Zaključak

Prema rezultatu χ^2 testa, ne možemo odbaciti H_0 hipotezu, te nastavljamo pod prepostavkom da su tip pogona i tip karoserije uistinu nezavisni.

Dodatno

Također možemo vidjeti tablicu očekivanih vrijednosti i usporediti sa dobivenim.

Očekivane vrijednosti

```
##
##           fwd rwd
## hatchback  44  23
##   sedan    60  31
```

Dobivene vrijednosti

```
##
##           fwd rwd
## hatchback  49  18
##   sedan    55  36
```

Vidimo da su vrijednosti zapravo dosta različite (što spada sa relativno niskom p-vrijednosti), no ne dovoljno da bi kategorije bile statistički značajno povezane.