Statystyczne Metody Wspomagania Decyzji – Zestaw zadań nr 1 – Rozwiązania z opisem

Zadanie 1.2.

a) Wczytanie danych z pliku .CSV do ramki danych o nazwie auta przy użyciu funkcji read.csv2:

```
# 1.sposób – korzystając z linku do pliku z danymi:
auta <- read.csv2('http://www.ibspan.waw.pl/~pgrzeg/stat_lab/samochody.csv')
# 2.sposób – po zapisaniu pliku samochody.csv w katalogu roboczym:
getwd() # katalog roboczy
auta <- read.csv2('samochody.csv')
```

b) Sprawdzenie typów zmiennych.

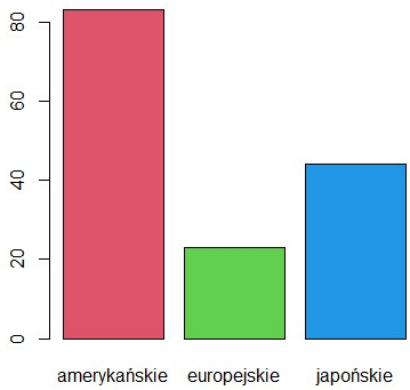
```
summary(auta)
sapply(auta, typeof)
```

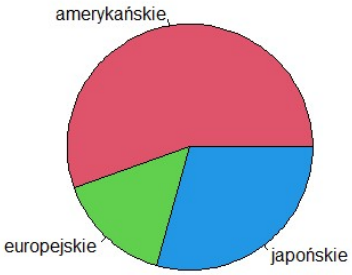
Podsumowanie zbioru danych – funkcja summary(). Można zauważyć, że zmienna producent została rozpoznana jako wektor liczbowy, tymczasem jest zmienną jakościową, opisującą kod producenta. Dokonujemy zatem **zmiany typu z wektora liczbowego na czynnik** odpowiadający zmiennej jakościowej – funkcja factor().

```
auta$producent <- factor(auta$producent)
levels(auta$producent) <- c('amerykańskie', 'europejskie', 'japońskie')
```

Można zliczyć auta poszczególnych producentów – funkcja table() oraz policzyć udziały procentowe – funkcja prop.test(). Następnie przedstawić je na wykresie kołowym lub słupkowym.

```
> table(auta$producent) # liczebności

amerykańskie  europejskie    japońskie
          83           23           44
> prop.table(table(auta$producent)) # odsetki

amerykańskie  europejskie    japońskie
   0.5533333    0.1533333    0.2933333
> round(100* prop.table(table(auta$producent)),1) # procenty

amerykańskie  europejskie    japońskie
        55.3         15.3         29.3

> barplot(table(auta$producent), col=2:4) # wykres słupkowy
> pie(table(auta$producent), col=2:4) # wykres kołowy
```





c) Usunięcie obserwacji z NA

```
auta <- na.omit(auta)
```

d) Utworzenie zmiennej zp opisującej zużycie paliwa w [l/100km] na podstawie mpg:

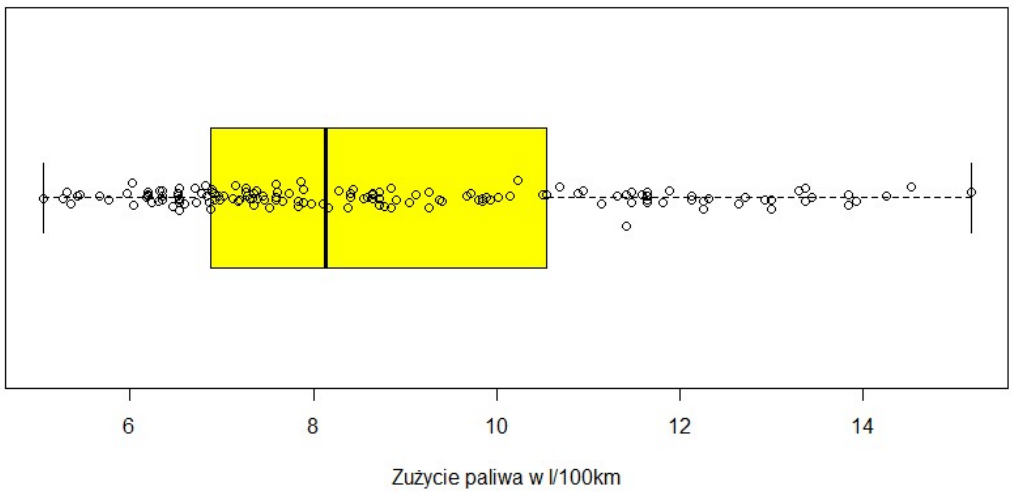
```
# mpg [mil]     - 1 [galon]
# mpg*1.609 [km] - 3.785 [l]
# 100 [km]      - zp [l]
# zp = (100*3.785)/(mpg*1.609)

auta$zp <-(100*3.785)/(auta$mpg*1.609) # zużycie paliwa w l/100km
```

Przedstawienie rozkładu zmiennej zp na histogramie i wykresie łodygowo-liściowym:

```
#e)
hist(auta$zp) # histogram
hist(auta$zp, col='yellow', xlab='zużycie paliwa w l/100km', ylim=c(0,40),
    labels=TRUE, ylab='Liczebności', main = '')

# ta sama zmienna, a różne histogramy
par(mfrow=c(2,2)) # podział okna graficznego na 4 części
hist(auta$zp, breaks=3, col=2, xlim=c(0,20))
hist(auta$zp, breaks=5, col=3, xlim=c(0,20))
hist(auta$zp, breaks=11, col=4, xlim=c(0,20))
hist(auta$zp, breaks=20, col=5, xlim=c(0,20))
par(mfrow=c(1,1)) # powrót do jednego okna graficznego

#f)
stem(auta$zp) # wykres łodygowo-liściowy
```

h) Narysowanie wykresu skrzynkowego.

```
boxplot(auta$zp) # wykres skrzynkowy (pudełkowy, ramkowy)
boxplot(auta$zp, horizontal = TRUE, col='yellow',
        xlab='Zużycie paliwa w l/100km')
points(auta$zp, rnorm(length(auta$zp), 1, 0.02))
```



g) Wyznaczenie wskaźników położenia, rozproszenia i kształtu.

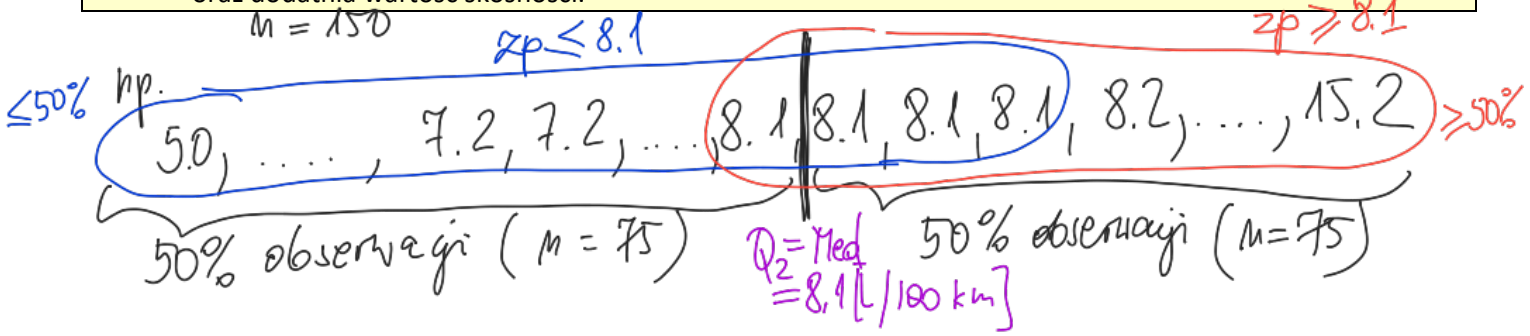
```
> mean(auta$zp) #średnia,
[1] 8.782034
> median(auta$zp) #mediana,
[1] 8.139865
> quantile(auta$zp) #kwartyle,
       0%       25%       50%       75%      100%
 5.048053  6.883384  8.139865 10.537073 15.176728
> quantile(auta$zp, c(0.1, 0.9)) #10. i 90. percentyl,
      10%       90%
 6.190507 12.349301
> min(auta$zp); max(auta$zp) #wartości ekstremalne,
[1] 5.048053
[1] 15.17673
> range(auta$zp)
[1]  5.048053 15.176728
> var(auta$zp) #wariancja,
[1] 5.942592
> sd(auta$zp) #odchylenie standardowe,
[1] 2.437743
> diff(range(auta$zp)) #rozstęp,
[1] 10.12867
> IQR(auta$zp) #rozstęp międzykwartylowy,
[1] 3.653689
> library(e1071)
> skewness(auta$zp) #współczynnik asymetrii,
[1] 0.6849622
> kurtosis(auta$zp) #kurtoza,
[1] -0.5968756
```

uwaga: jeśli pakiet e1071 nie został wcześniej zainstalowany, należy to zrobić wykonując polecenie: install.packages("e1071")

```
> mean(auta$zp)/sd(auta$zp) #współczynnik zmienności
[1] 3.602526
```

Interpretacja wyników:

- Średnie zużycie paliwa wynosi 8,8 l/100km z odchyleniem standardowym 2,43 l/100km.
- Najmniejsze zużycie to 5l/100km, a największe 15.2l/100km.
- Dla co najmniej 50% aut zużycie paliwa jest nie większe niż 8.1 l/100km i jednocześnie dla co najmniej 50% aut zużycie paliwa jest nie mniejsze niż 8.1 l/100km.
- Dla co najmniej 25% samochodów zużycie wyniosło nie więcej niż 6,9 l/100km (Q1) i jednocześnie dla co najmniej 75% aut zużycie paliwa jest nie mniejsze niż 6,9 l/100km.
- Dla co najmniej 75% aut zużycie nie przekracza 10,4 i/100km (Q3) i jednocześnie dla co najmniej 25% aut zużycie paliwa jest nie mniejsze niż 10,4 l/100km.
- Dla co najmniej 10% aut spalanie jest nie mniejsze niż 12.3 l/100km (90.percentyl).
- Rozkład zużycia paliwa jest rozkładem prawostronnie skośnym – co jest widoczne na histogramie i wykresie skrzynkowym (wolniej opadające prawe ramię, dłuższy prawy wąs, mediana przesunięta w lewo) oraz dodatnia wartość skośności.



Zadanie 1.3.

a) Utworzenie zmiennej zp_kat opisującej kategorię zużycia paliwa.

```
auta$zp_kat[auta$zp <= 7] <- 'mało'
auta$zp_kat[auta$zp > 7 & auta$zp <= 10] <- 'średnio'
auta$zp_kat[auta$zp > 10] <- 'dużo'
```

b) Zliczenie kategorii (poziomów) czynnika zp_kat.

```
table(auta$zp_kat)
# dużo  mało średnio
#   42    44     64
```

← porządek słownikowy

Jeśli chcemy nadać inny porządek kategoriom czynnika, możemy wykonać polecenie

```
auta$zp_kat <- factor(auta$zp_kat, ordered = TRUE,  levels = c("mało", "średnio", "dużo"))
table(auta$zp_kat)
# mało średnio   dużo
#   44     64     42
```

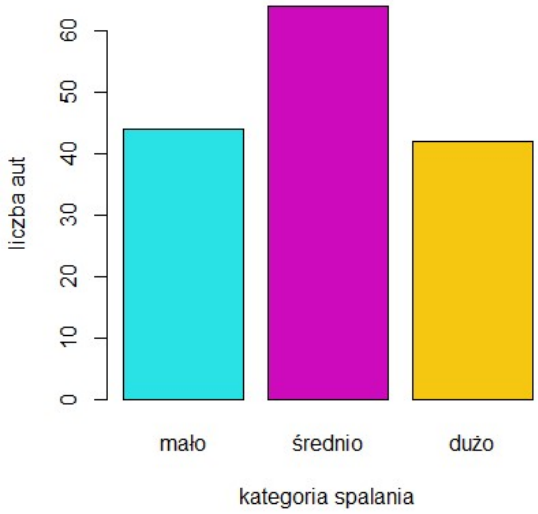
← nadany porządek

Obliczenie udziałów procentowych poszczególnych kategorii czynnika zp_kat.

```
prop.table(table(auta$zp_kat)) # odsetki
prop.table(table(auta$zp_kat))*100 # procenty
```

c) Narysowanie wykresu słupkowego i kołowego dla kategorii spalania.

```
barplot(table(auta$zp_kat),col=5:7,
        ylab='liczba aut', xlab = 'kategoria spalania') # wykres słupkowy
pie(table(auta$zp_kat),col=5:7) # wykres kołowy
```



Uwaga: można też utworzyć takie wykresy dla odsetków (udziałów procentowych).

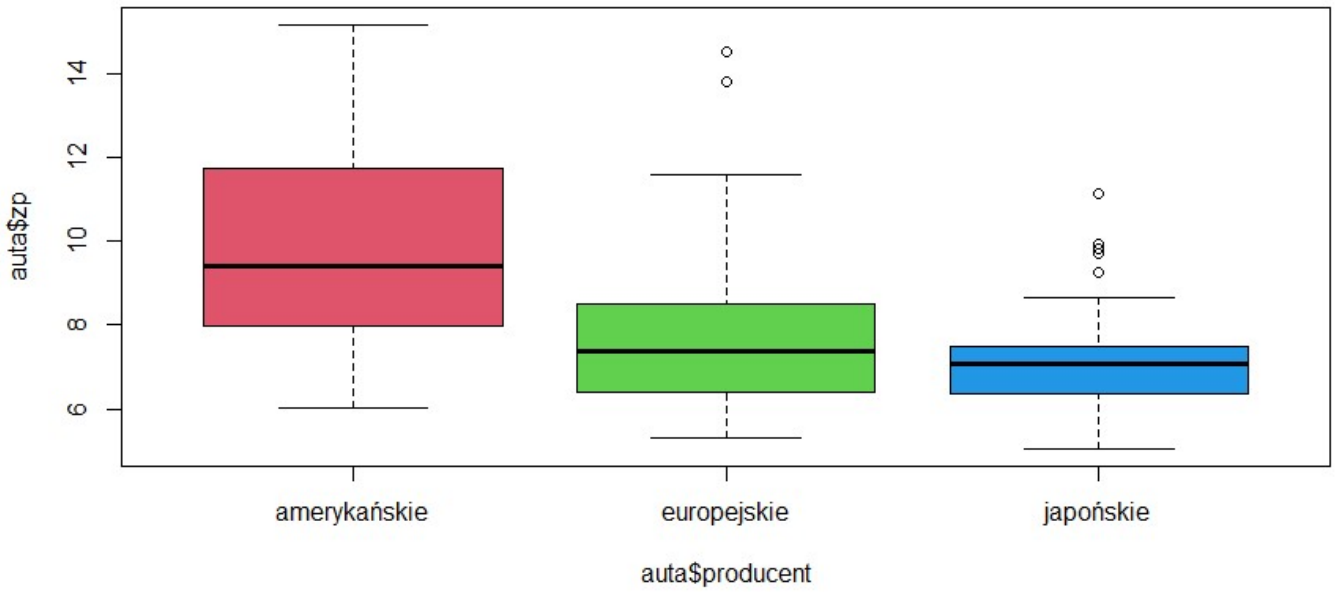
```
barplot(prop.table(table(auta$zp_kat))*100,col=5:7,
        ylab='odsetek (proporcja) aut', xlab = 'kategoria spalania') # wykres słupkowy
pie(prop.table(table(auta$zp_kat)),col=5:7) # wykres kołowy
```

Zadanie 1.4.

Funkcja tapply(), dla której pierwszym argumentem jest wektor liczbowy, drugim - wektor lub czynnik określający grupy, trzecim funkcja, która zostanie wyznaczona na wektorze liczbowym względem grup, np. średnia i odchylenie standardowe zużycia paliwa względem producenta:

```
tapply(auta$zp, auta$producent, mean)  # 9.847964   7.933492   7.214859
tapply(auta$zp, auta$producent, sd)    # 2.352059   2.564224   1.259573
boxplot(auta$zp~auta$producent, col=2:4)
```

wśród aut japońskich najmniejsze zróżnicowanie pod względem zużycia paliwa

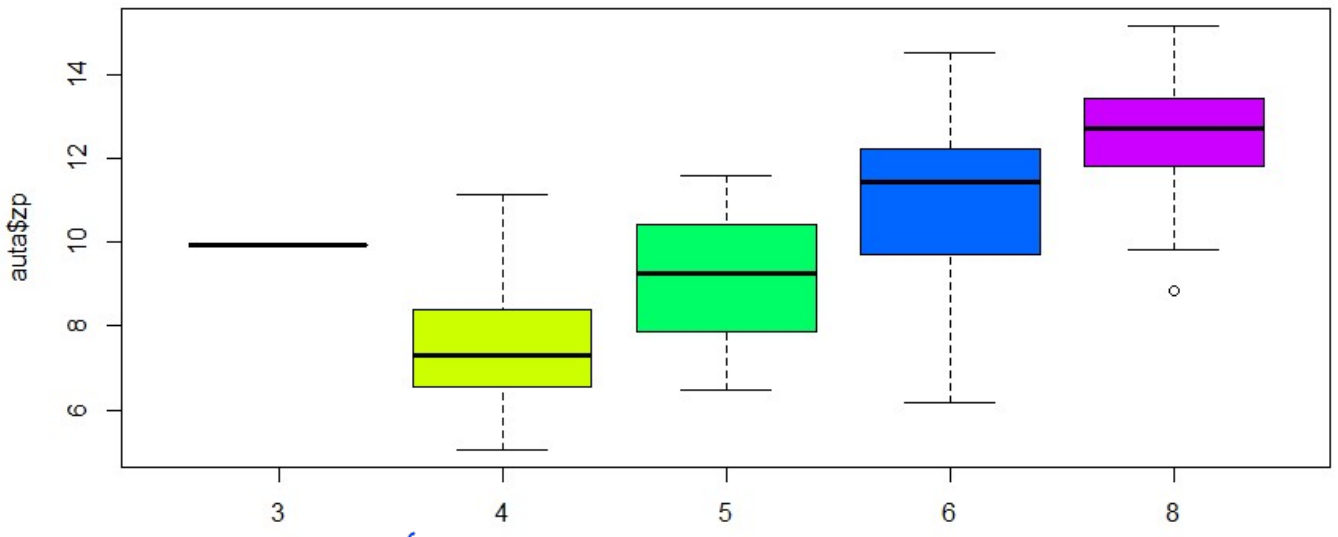


Interpretacja:
Największe zużycie paliwa obserwuje się dla aut amerykańskich – średnio 9,6 l/100km z odchyleniem standardowym 2,32 l/100km. Dla aut europejskich i japońskich – średnie zużycie jest zbliżone i wynosi 7.8 dla tych pierwszych i 7.2 dla tych drugich. Dla aut japońskich obserwuje się jednak najmniejszą zmienność – odchylenie standardowe zużycia paliwa dla nich jest najmniejsze i wynosi 1.26 l/100km. Dla europejskich jest podobne do odchylenia standardowego zużycia paliwa dla amerykańskich i wynosi 2.50 l/100km.
Ponadto łatwo zauważyć, że wykres skrzynkowy ma największą rozpiętość dla amerykańskich, mniejszą dla europejskich i najmniejszą dla japońskich. Wartość zużycia paliwa odpowiadająca Q1 dla aut amerykańskich, czyli 8.1 l/100km, jest maksymalną wartością dla niemalże wszystkich aut japońskich (wykres skrzynkowy dla japońskich (bez obserwacji odstających) „mieści się" niemalże pod Q1 dla amerykańskich). Z kolei dla samochodów europejskich wartość Q3=8.4 jest zbliżona do Q1=8.1 dla amerykańskich, co oznacza, że 75% aut europejskich zużywa maksymalnie tyle paliwa co 25% aut amerykańskich.

Uwaga: wykresy skrzynkowe bardzo dobrze sprawdzają się przy porównywaniu rozkładów dla grup.
Na podstawie samego wykresu nie możemy jeszcze stwierdzać, że np. auta amerykańskie mają istotnie większe spalanie niż japońskie. Dlatego że wizualizacja dotyczy wyłącznie tej próby danych (można przypuszczać, że dla innej próby danych, wykresy byłyby inne). *Dopiero testowanie hipotez statystycznych da nam możliwość wnioskowania o tym jak jest w ogóle, czyli np., że samochody amerykańskie mają istotnie większe spalanie niż japońskie.*

Zadanie 1.5.

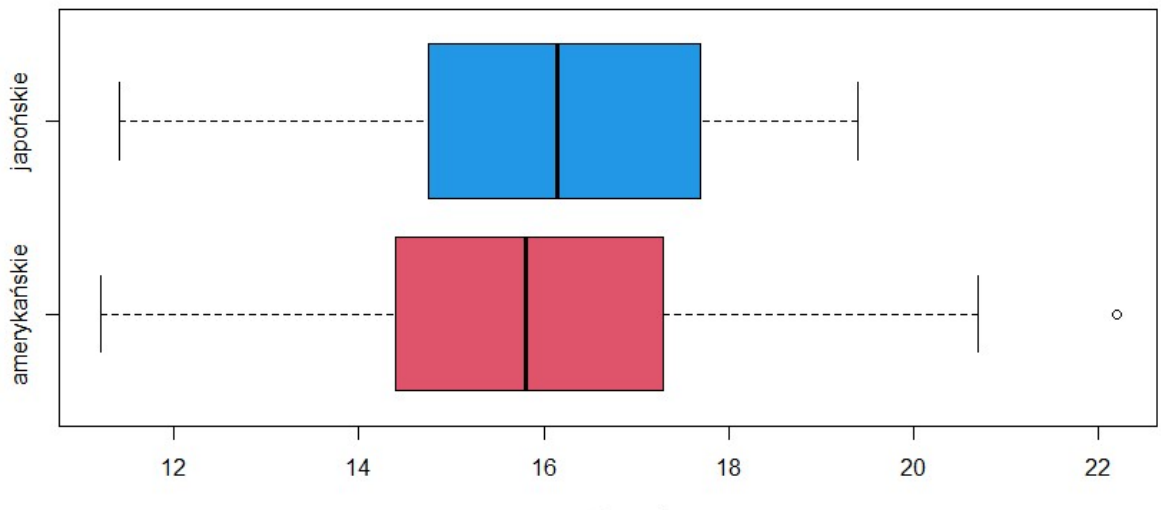
```
table(auta$cylindry)
tapply(auta$zp, auta$cylindry, summary)
boxplot(auta$zp~auta$cylindry, col=rainbow(5))
```



Można zaobserwować, że wraz ze wzrostem liczby cylindrów zwiększa się zużycie paliwa.

Zadanie 1.6.

```
przysp.A <- auta$przysp[auta$producent=='amerykańskie']
przysp.J <- auta$przysp[auta$producent=='japońskie']

boxplot(przysp.A, przysp.J, names = c('amerykańskie','japońskie'), horizontal = TRUE, col=c(2,4),
      xlab = 'przyspieszenie')
```



Rozkłady przyspieszenia aut amerykańskich i japońskich wydają się zbliżone.

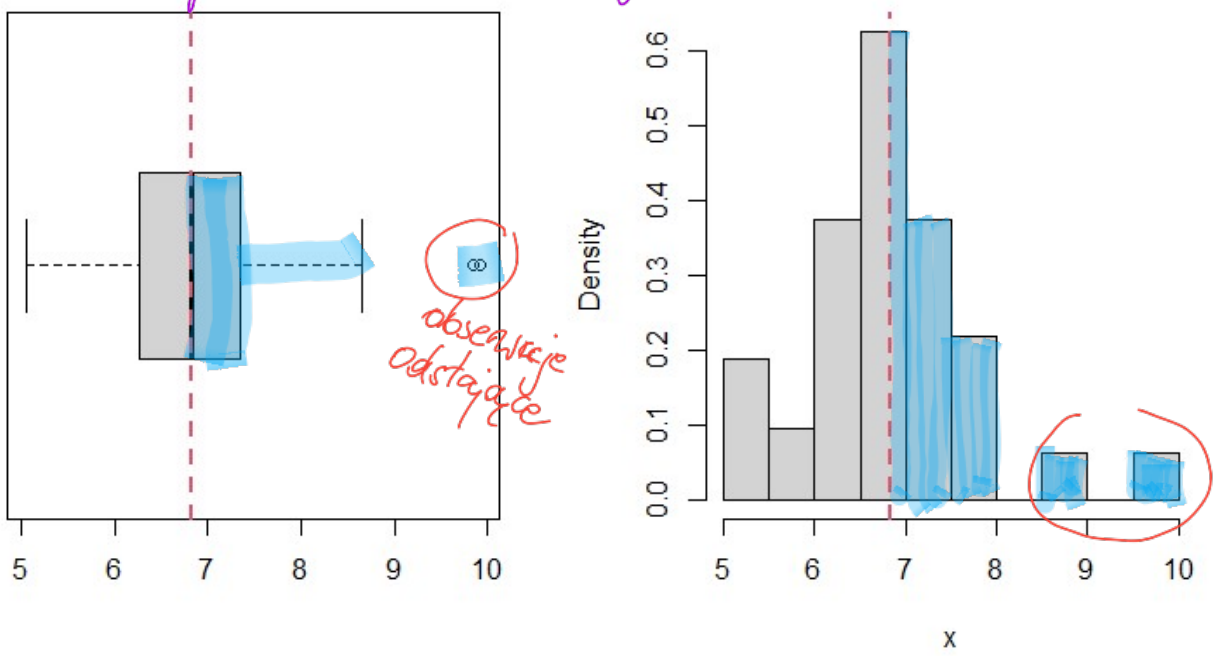
Zadanie 1.7.

```
> x <- auta$zp[auta$waga < 2500]
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.048   6.290   6.828   6.845   7.334   9.926
> var(x)
[1] 0.9097303
> sd(x)
[1] 0.9537978
> skewness(x)
[1] 0.8377937
```

← tworzę wektor x, bo dalej wygodnie jest odwoływać się tylko do wektora x

Średnie zużycie paliwa aut o wadze < 2500 wynosi 6,8 l/100 km z odchyleniem standardowym 0,95 l/100 km.

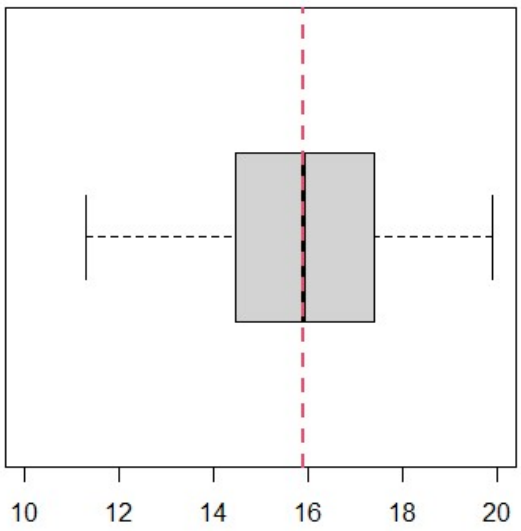
↓ bliskie 1 – rozkład prawostronnie skośny

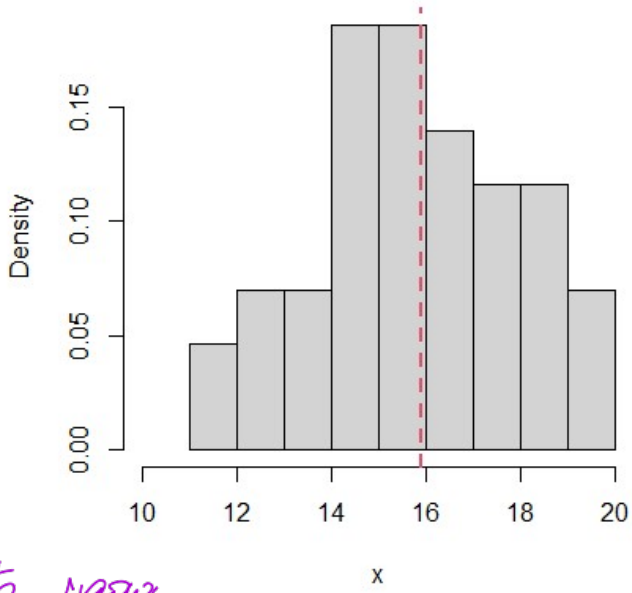


Zadanie 1.8.

```
> x <- auta$przysp[auta$waga > 2500 & auta$waga < 3000]
> par(mfrow=c(1,2))
> boxplot(x, horizontal = TRUE, ylim=c(10,20))
> abline(v=median(x), lty=2, col=2, lwd=2)
> hist(x, xlim=c(10,20),main="",probability = TRUE)
> abline(v=median(x), lty=2, col=2, lwd=2)
> quantile(x, 0.75)
 75%
17.4
> par(mfrow=c(1,1))
```

← co najmniej 75% aut o wadze (2500, 3000) ma przyspieszenie nie większe niż 17.4 (co najmniej 25% aut ma przyspieszenie nie mniejsze niż 17.4)



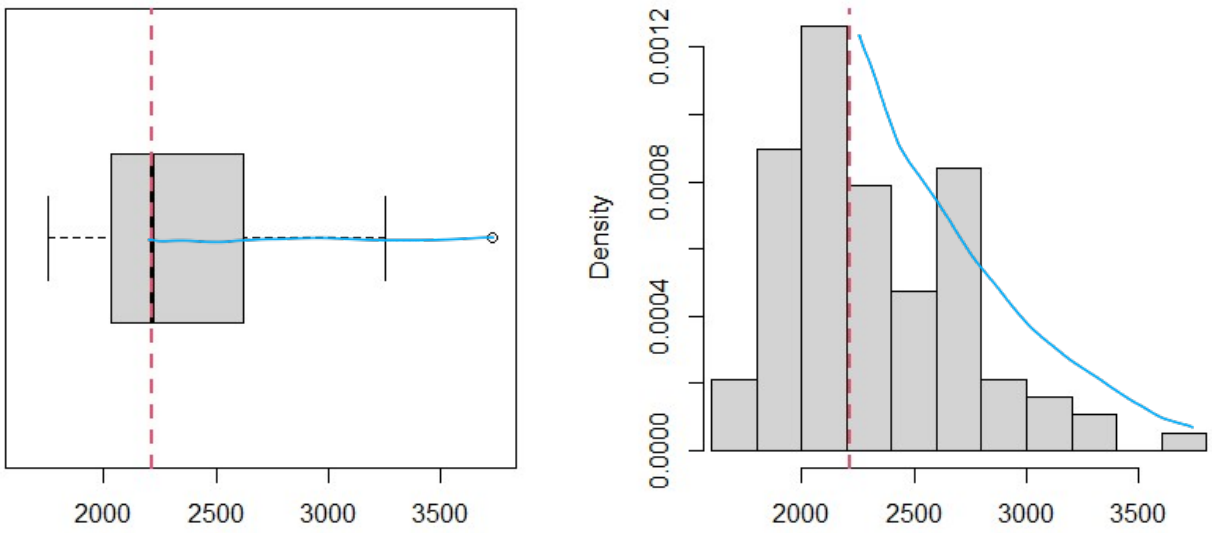


Rozkład przyspieszenia aut z tą wagą jest raczej symetryczny (średnia ≈ mediana, skośność blisko 0)
Najwięcej aut z przyspieszeniem 14–16.

Zadanie 1.9.

```
> x <- auta$waga[auta$mpg > 26]
> par(mfrow=c(1,2))
> boxplot(x, horizontal = TRUE, ylim=c(1650,3750))
> abline(v=median(x), lty=2, col=2, lwd=2)
> hist(x, main="",probability = TRUE, xlim=c(1650,3750))
> abline(v=median(x), lty=2, col=2, lwd=2)
> quantile(x, 0.95)
   95%
3058.5
> par(mfrow=c(1,1))
```

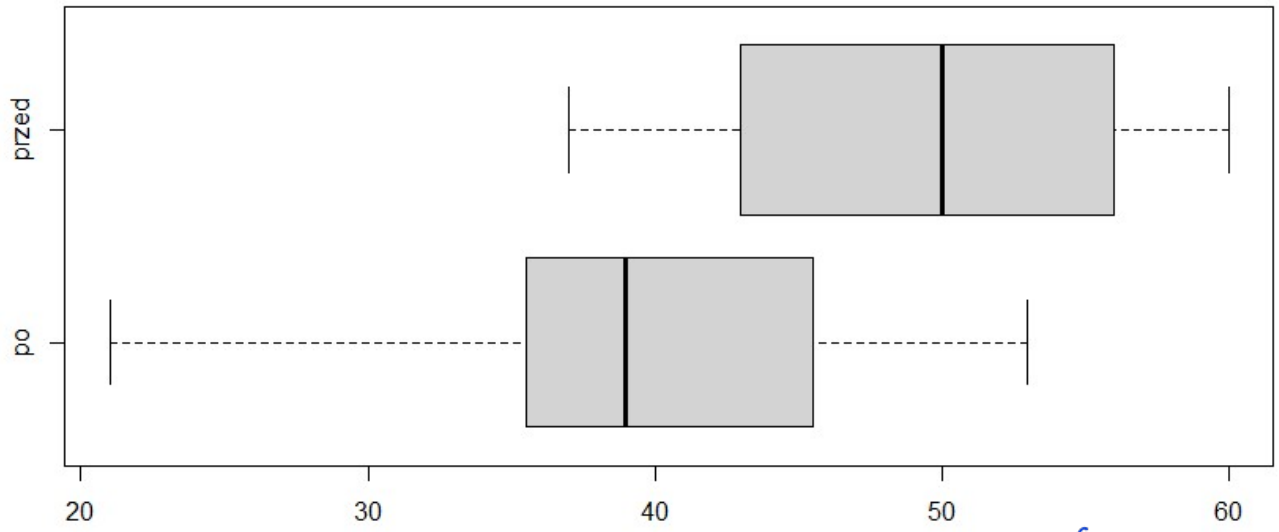
← co najmniej 95% aut z mpg > 26 ma wagę ≤ 3058.5



Rozkład wagi aut z mpg > 26 jest prawostronnie skośny (silna okośność, skośność blisko 1, średnia > mediana). Występuje obserwacja odstająca.

Zadanie 1.11.

```
przed <- c(56, 47, 49, 37, 38, 60, 50, 43, 43, 59, 50, 56, 54, 58)
po <- c(53, 21, 32, 49, 45, 38, 44, 33, 32, 43, 53, 46, 36, 48, 39, 35, 37, 36, 39, 45)

boxplot(przed, po)
boxplot(przed, po, names = c('przed','po'))
summary(przed)
summary(po)
```



Rozkłady liczby wiązań przed i po wydają się różne (mediany nawet daleko od siebie). Sprawdzenie, czy medians różnią się istotnie wymagałoby testowania odpowiedniej hipotezy statystycznej (Lab 4).

```
> summary(przed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     37      44      50      50      56      60
> summary(po)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   35.75   39.00   40.20   45.25   53.00
```