Moodle Tasks

Разгледайте данните survey от пакета MASS

```
> install.packages("MASS")
> library(MASS)
> names(survey)
 [1] "Sex"    "Wr.Hnd" "NW.Hnd" "W.Hnd"  "Fold"   "Pulse"  "Clap"   "Exer"  
 [9] "Smoke"  "Height" "M.I"    "Age"   
> head(survey)
     Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height      M.I
1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00   Metric
2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80 Imperial
3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA     <NA>
4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00   Metric
5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00   Metric
6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72 Imperial
     Age
1 18.250
2 17.583
3 16.917
4 20.333
5 23.667
6 21.000
> summary(survey)
     Sex          Wr.Hnd          NW.Hnd        W.Hnd          Fold    
 Female:118   Min.   :13.00   Min.   :12.50   Left : 18   L on R : 99  
 Male  :118   1st Qu.:17.50   1st Qu.:17.50   Right:218   Neither: 18  
 NA's  :  1   Median :18.50   Median :18.50   NA's :  1   R on L :120  
              Mean   :18.67   Mean   :18.58                            
              3rd Qu.:19.80   3rd Qu.:19.73                            
              Max.   :23.20   Max.   :23.50                            
              NA's   :1       NA's   :1                                
     Pulse             Clap       Exer       Smoke         Height     
 Min.   : 35.00   Left   : 39   Freq:115   Heavy: 11   Min.   :150.0  
 1st Qu.: 66.00   Neither: 50   None: 24   Never:189   1st Qu.:165.0  
 Median : 72.50   Right  :147   Some: 98   Occas: 19   Median :171.0  
 Mean   : 74.15   NA's   :  1              Regul: 17   Mean   :172.4  
 3rd Qu.: 80.00                            NA's :  1   3rd Qu.:180.0  
 Max.   :104.00                                        Max.   :200.0  
 NA's   :45                                            NA's   :28     
       M.I           Age       
 Imperial: 68   Min.   :16.75  
 Metric  :141   1st Qu.:17.67  
 NA's    : 28   Median :18.58  
                Mean   :20.37  
                3rd Qu.:20.17  
                Max.   :73.00  
```

Задача 1

Въз основа на данните пресметнете вероятностите:

a) Случайно избран човек да се окаже редовен пушач

$$\mathbb{P}(\text{Smoke='Regul'}) = \frac{\text{\# of regularly smoking students}}{\text{\# of all students}} \approx 0.07$$

```
> prob.regul = sum(survey$Smoke == "Regul", na.rm = TRUE) / length(survey$Smoke)
> prob.regul
[1] 0.07172996
```

b) Случайно избран човек да се окаже редовно пушещ мъж

$$\mathbb{P}(\text{Smoke='Regul'} \cap \text{Sex='Male'}) \approx 0.05$$

```
> prob.regulANDsmoke = sum(survey$Smoke == "Regul" &
+       survey$Sex == "Male", na.rm = TRUE) / length(survey$Smoke)
> prob.regulANDsmoke
[1] 0.05063291
```

c) Случайно избран мъж да се окаже редовен пушач

$$\mathbb{P}(\text{Smoke='Regul'}|\text{Sex='Male'}) = \frac{\mathbb{P}(\text{Smoke='Regul'} \cap \text{Sex='Male'})}{\mathbb{P}(\text{Sex='Male'})} \approx 0.10$$

```
> prob.male = sum(survey$Sex == "Male", na.rm = TRUE)/length(survey$Sex)
> prob.regulANDsmoke/prob.male
[1] 0.1016949
```

d) Случайно избран редовен пушач да се окаже мъж

$$\mathbb{P}(\text{Sex='Male'}||\text{Smoke='Regul'}) = \frac{\mathbb{P}(\text{Sex='Male'} \cap \text{Smoke='Regul'})}{\mathbb{P}(\text{Smoke='Regul'})} \approx 0.71$$

```
> prob.regulANDsmoke/prob.regul
[1] 0.7058824
```

Задача 2

Представете графично данните за тютюнопушенето на студентите

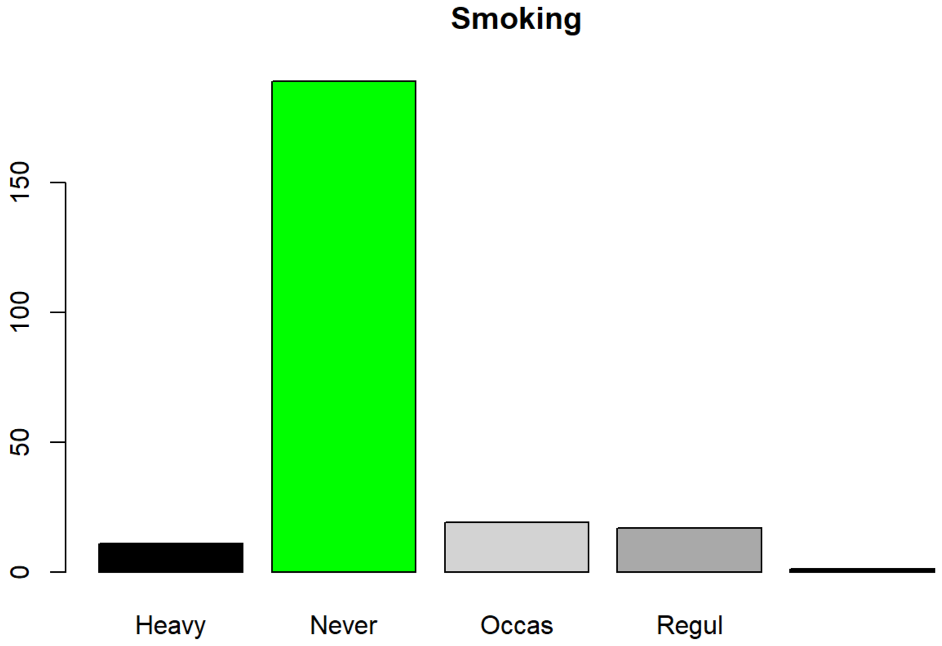
```
> summary(survey$Smoke)
Heavy Never Occas Regul  NA's
   11   189    19    17     1
```

Smoke е качественна променлива

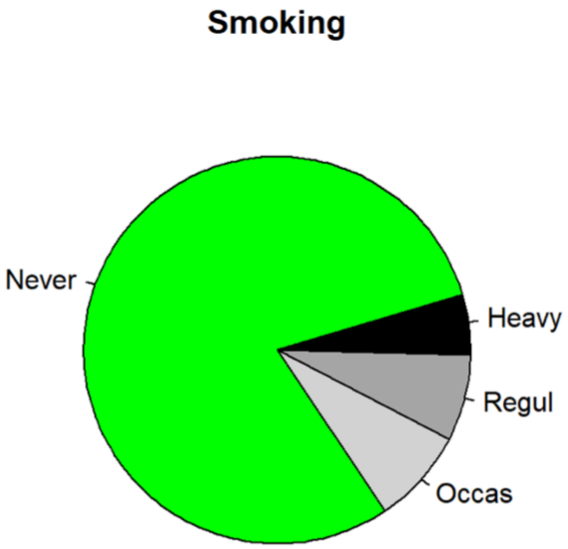
```
> table(survey$Smoke, useNA = "ifany")
Heavy Never Occas Regul  <NA>
```

```
  11   189    19    17     1
```

```
> barplot(table(survey$Smoke, useNA = "ifany"),
+         main = "Smoking",
+         col = c("black", "green", "lightgrey", "darkgrey"))
```



```
> pie(table(survey$Smoke, useNA = "ifany"),
+     main = "Smoking",
+     col = c("black", "green", "lightgrey", "darkgrey"))
```



Представете също така тютюнопушенето на студентите в зависимост от пола.

Smoke и Sex са две качествени променливи. Можем да ги предсавим с честотната таблица по двете променливи

```
> table(survey$Sex, survey$Smoke)
        Heavy Never Occas Regul
```

```
 Female    5   99    9    5
 Male      6   89   10   12
```

Също така можем да ги представим с таблица на пропорциите по двете променливи.

За по ясна визуализация, нека закръглим числата до 2рия знак след десетичната запетая.

```
> options(digits = 1)
```

Видяхме, че можем да смятаме пропорциите с функцията prop.table спрямо всички наблюдения, спрямо редовете или спрямо колоните.
Пропорции спрямо редовете - тук пропорциите в реда се сумират до 1

```
> prop.table(table(survey$Sex, survey$Smoke), 1)
        Heavy Never Occas Regul
 Female  0.04  0.84  0.08  0.04
 Male    0.05  0.76  0.09  0.10
```

Пропорции спрямо колоните – тук пропорциите в колоните се сумират до 1

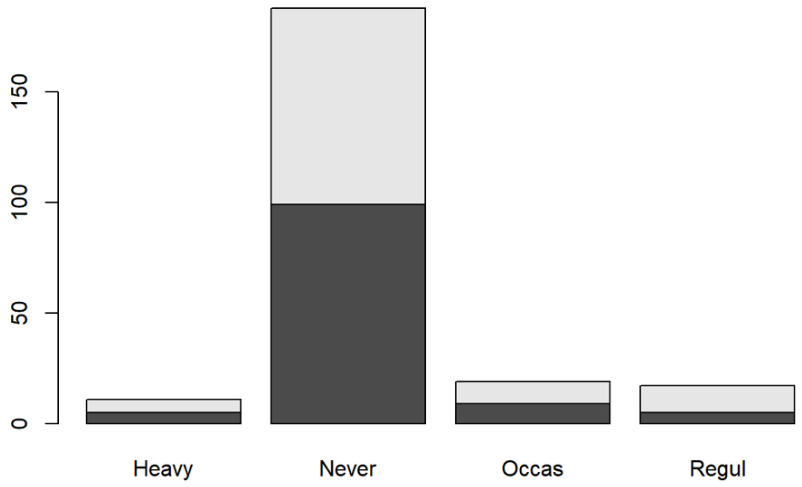
```
> prop.table(table(survey$Sex, survey$Smoke), 2)
        Heavy Never Occas Regul
 Female   0.5   0.5   0.5   0.3
 Male     0.5   0.5   0.5   0.7
```

Пропорциите спрямо всички наблюдения – тук всички пропорции в таблицата се сумират до 1

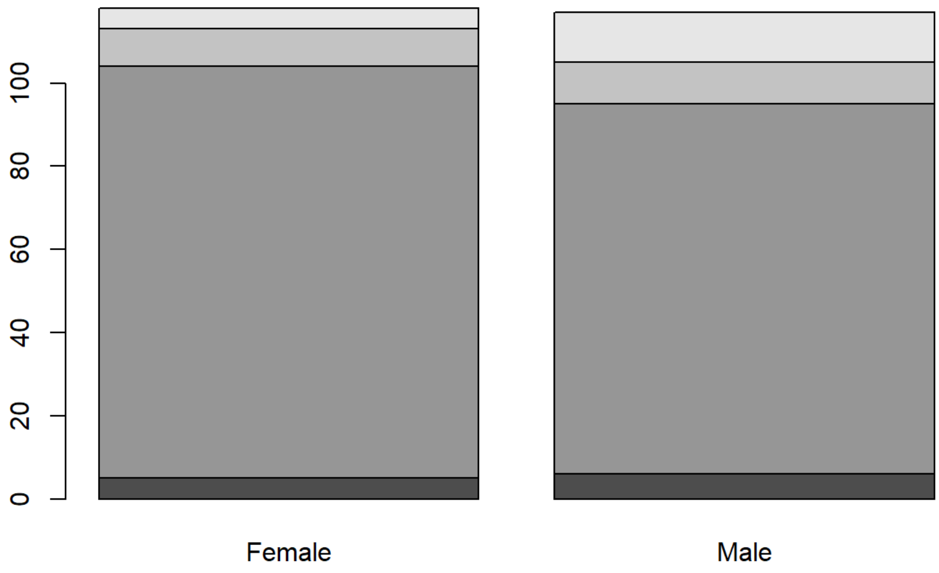
```
> prop.table(table(survey$Sex, survey$Smoke))
        Heavy Never Occas Regul
 Female  0.02  0.42  0.04  0.02
 Male    0.03  0.38  0.04  0.05
```

Също така можем да ги представим графично използвайки barplot

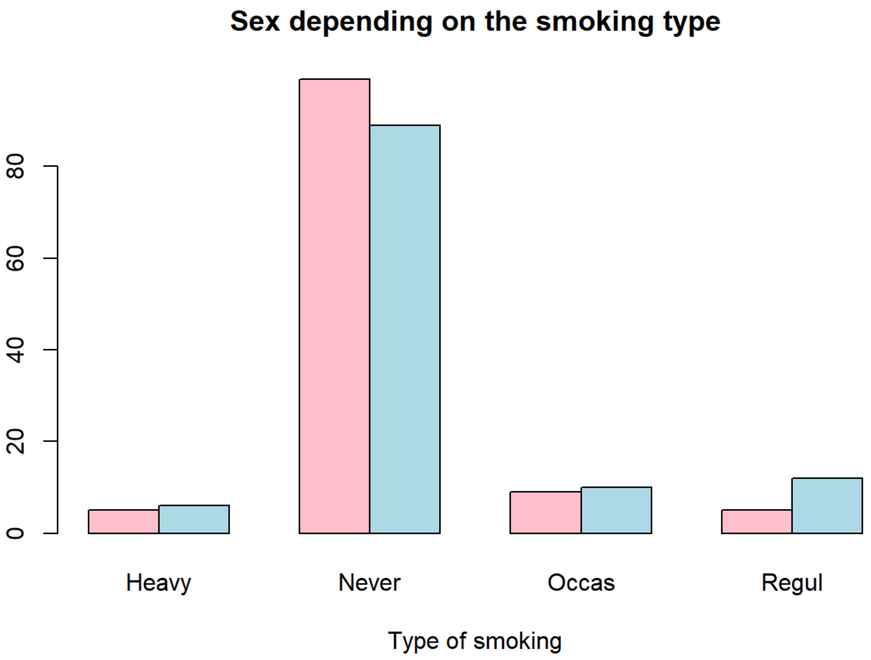
```
> barplot(table(survey$Sex, survey$Smoke))
```



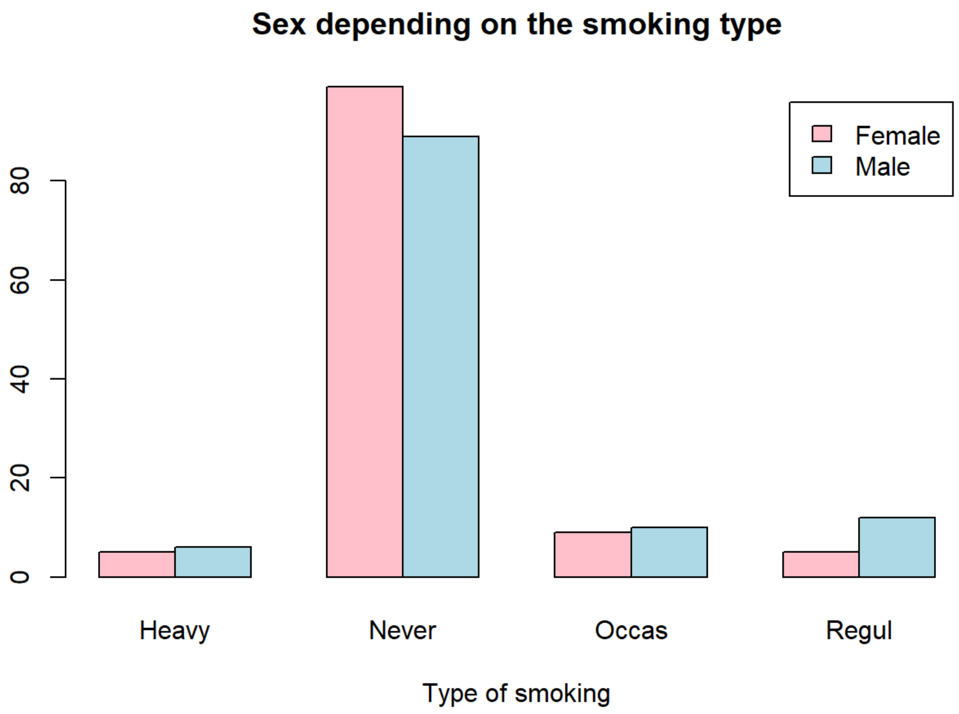
```
> barplot(table(survey$Smoke, survey$Sex))
```



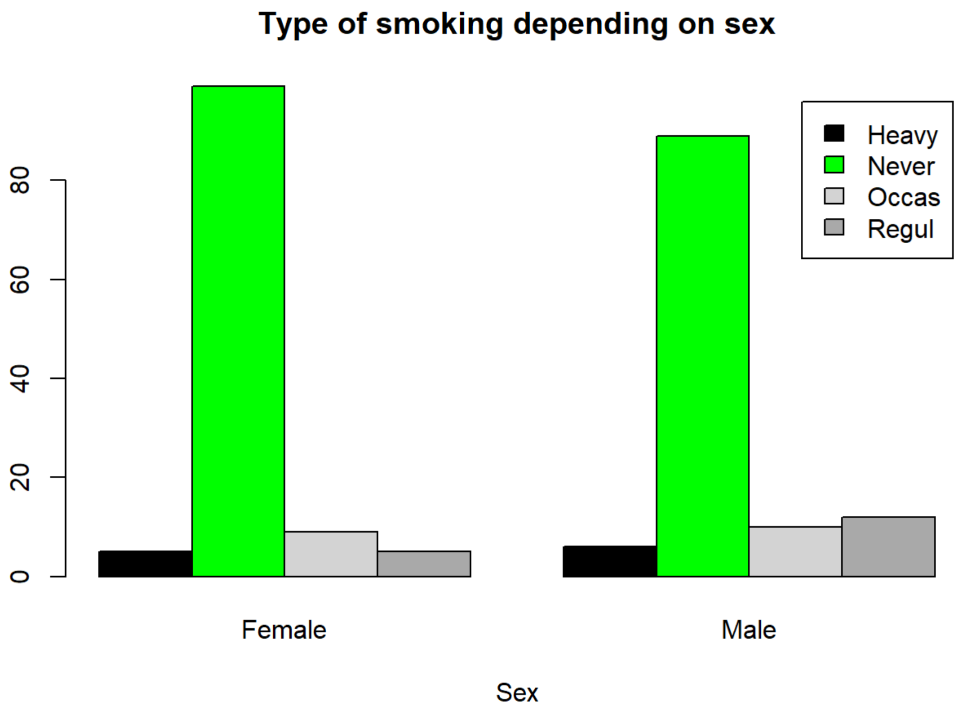
```
> barplot(table(survey$Sex, survey$Smoke),
+         main = "Sex depending on the smoking type",
+         names.arg = c("Heavy", "Never", "Occas", "Regul"),
+         beside = TRUE,
+         col = c("Pink", "lightblue"),
+         xlab = "Type of smoking")
```



```
> barplot(table(survey$Sex, survey$Smoke),
+         main = "Sex depending on the smoking type",
+         names.arg = c("Heavy", "Never", "Occas", "Regul"),
+         legend.text = TRUE,
+         beside = TRUE,
+         col = c("Pink", "lightblue"),
+         xlab = "Type of smoking")
```



```
> barplot(table(survey$Smoke, survey$Sex),
+       main = "Type of smoking depending on sex",
+       names.arg = c("Female", "Male"),
+       legend.text = TRUE,
+       beside = TRUE,
+       col = c("Black", "Green", "lightgrey", "darkgrey"),
+       xlab = "Sex")
```



Задача 3

Пресметнете оценки за средното, медианата, квартилите, стандартното отклонение и т.н. за височината на студентите

```
> min(survey$Height, na.rm = TRUE)
[1] 150
> max(survey$Height, na.rm = TRUE)
[1] 200
> mean(survey$Height, na.rm = TRUE)
[1] 172
> median(survey$Height, na.rm = TRUE)
[1] 171
> quantile(survey$Height, 0.25, na.rm = TRUE)
25%
165
> quantile(survey$Height, 0.75, na.rm = TRUE)
75%
180
> sd(survey$Height, na.rm = TRUE)
```

```
[1] 10
> var(survey$Height, na.rm = TRUE)
[1] 97
> diff(range(survey$Height, na.rm = TRUE))
[1] 50
> IQR(survey$Height, na.rm = TRUE)
[1] 15
> summary(survey$Height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    150     165     171     172     180     200      28 
> fivenum(survey$Height)
[1] 150 165 171 180 200
```

Направете отделни изчисления за мъжете и за жените.

```
> summary(survey$Height[survey$Sex == "Male"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    155     173     180     179     185     200      13 
> sd(survey$Height[survey$Sex == "Male"], na.rm = TRUE)
[1] 8
> var(survey$Height[survey$Sex == "Male"], na.rm = TRUE)
[1] 70
> diff(range(survey$Height[survey$Sex == "Male"], na.rm = TRUE))
[1] 45
> IQR(survey$Height[survey$Sex == "Male"], na.rm = TRUE)
[1] 12
> fivenum(survey$Height[survey$Sex == "Male"])
[1] 155 173 180 185 200
> summary(survey$Height[survey$Sex == "Female"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    150     163     167     166     170     180      17 
> sd(survey$Height[survey$Sex == "Female"], na.rm = TRUE)
[1] 6
> var(survey$Height[survey$Sex == "Female"], na.rm = TRUE)
[1] 38
> diff(range(survey$Height[survey$Sex == "Female"], na.rm = TRUE))
[1] 30
> IQR(survey$Height[survey$Sex == "Female"], na.rm = TRUE)
[1] 7
> fivenum(survey$Height[survey$Sex == "Female"])
[1] 150 163 167 170 180
```

Каква част от студентите се различават от средната височина с неповече от едно 1 стандартно отклонение, т.е. наблюдението да попадне в интервала

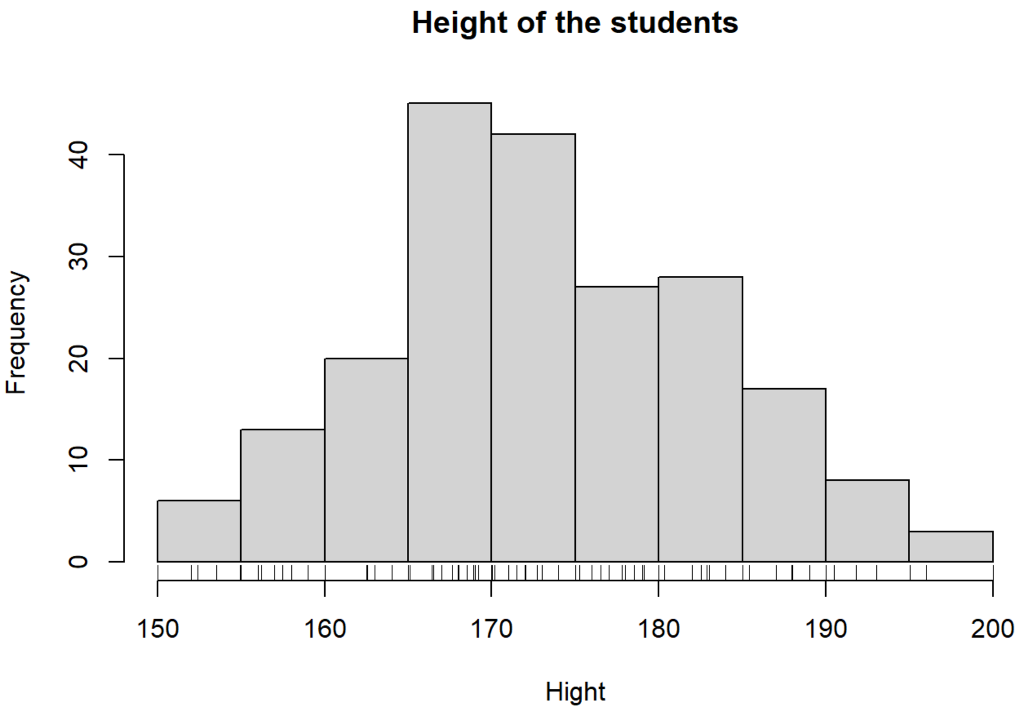
$$[\overline{X}-\sigma,\ \overline{X}+\sigma],\ \overline{X}-\sigma < x_i < \overline{X}+\sigma,\ -\sigma < x_i-\overline{X} < \sigma,\ \frac{|x_i-\overline{X}|}{\sigma} < 1.$$

```
> x.without.na <- survey$Height[!is.na(survey$Height)]
> x.mean <- mean(survey$Height, na.rm = TRUE)
> x.sd <- sd(survey$Height, na.rm = TRUE)
> sum(abs(x.without.na - x.mean) / x.sd < 1) / length(x.without.na)
[1] 0.7
```

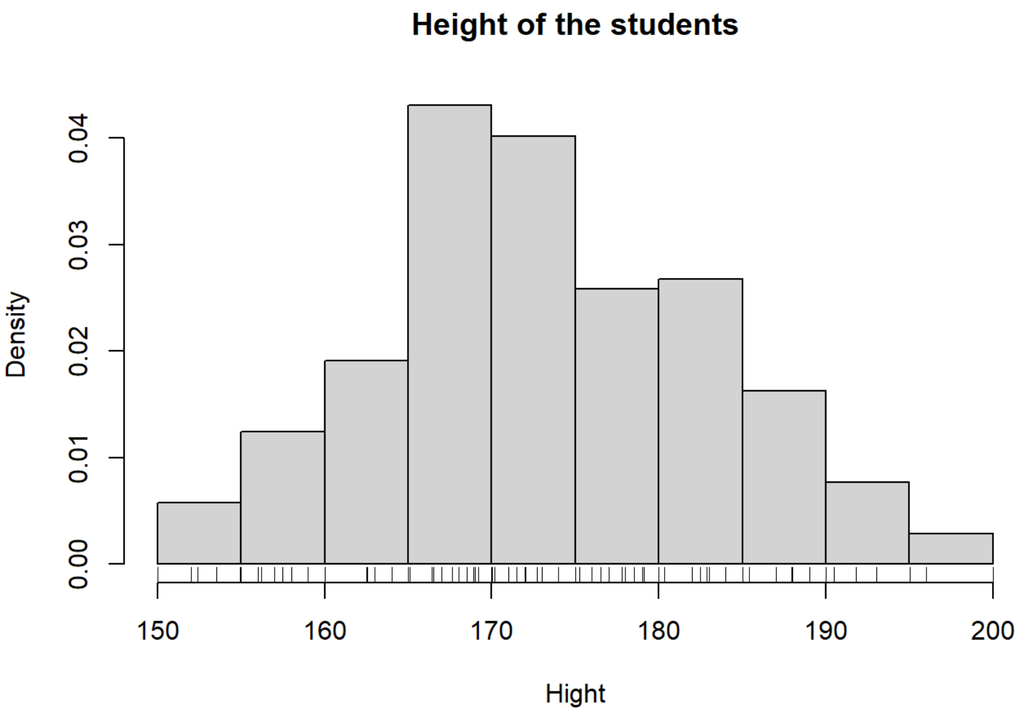
Задача 4

Представете графично височината на студентите
Височината на студентите е количествена непрекъсната променлива

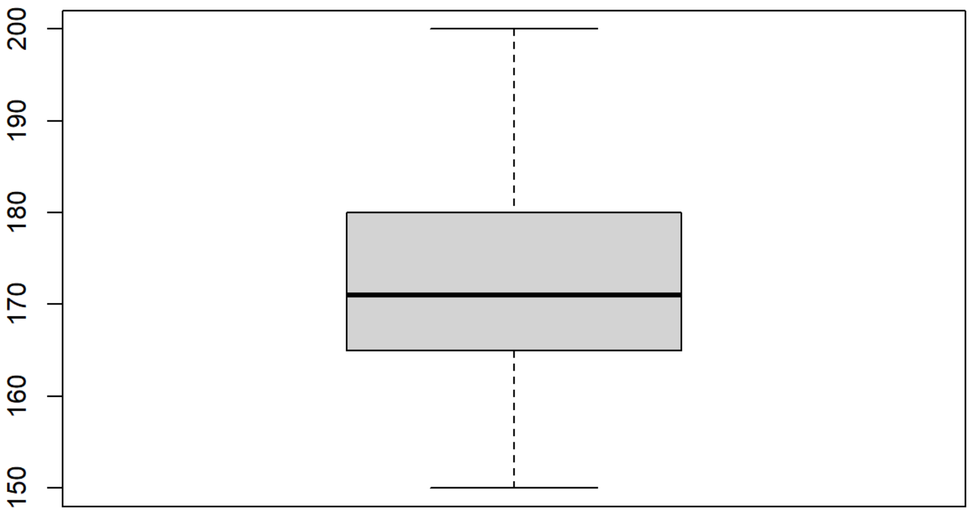
```
> hist(survey$Height,
+      right = FALSE,
+      main = "Height of the students",
+      xlab = "Hight")
> rug(jitter(survey$Height))
```



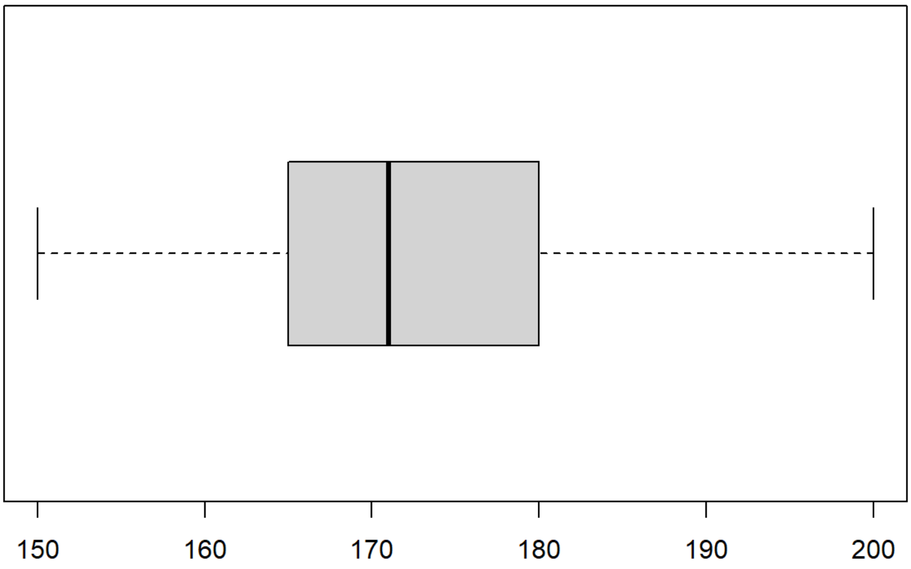
```
> hist(survey$Height,
+       right = FALSE,
+       main = "Height of the students",
+       xlab = "Hight",
+       probability = TRUE)
> rug(jitter(survey$Height))
```



```
> boxplot(survey$Height)
```

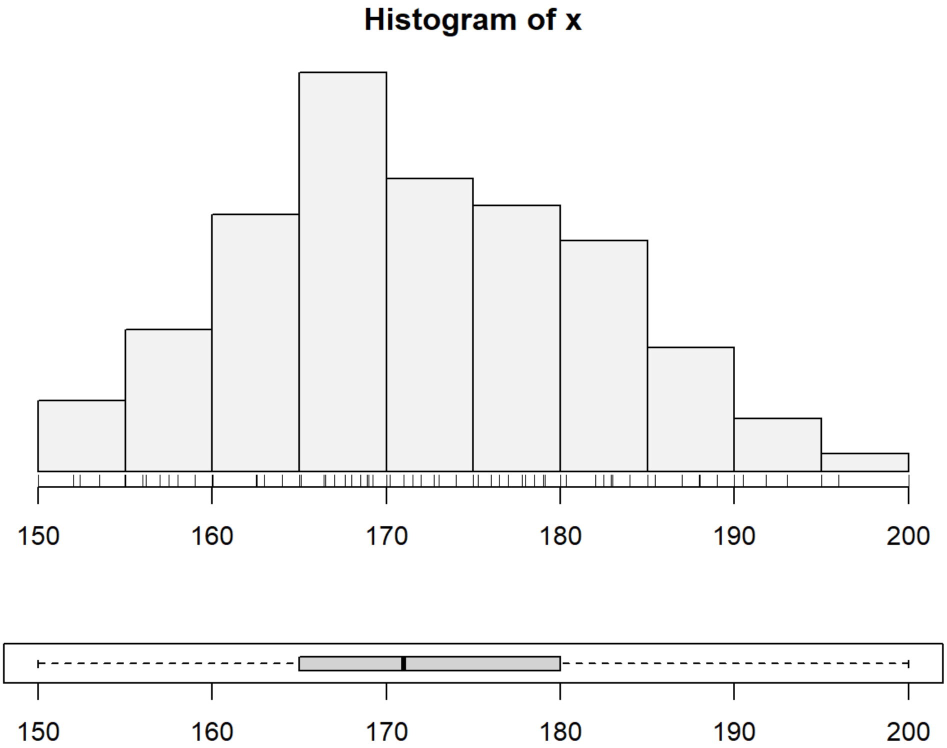


```
> boxplot(survey$Height, horizontal = TRUE)
```

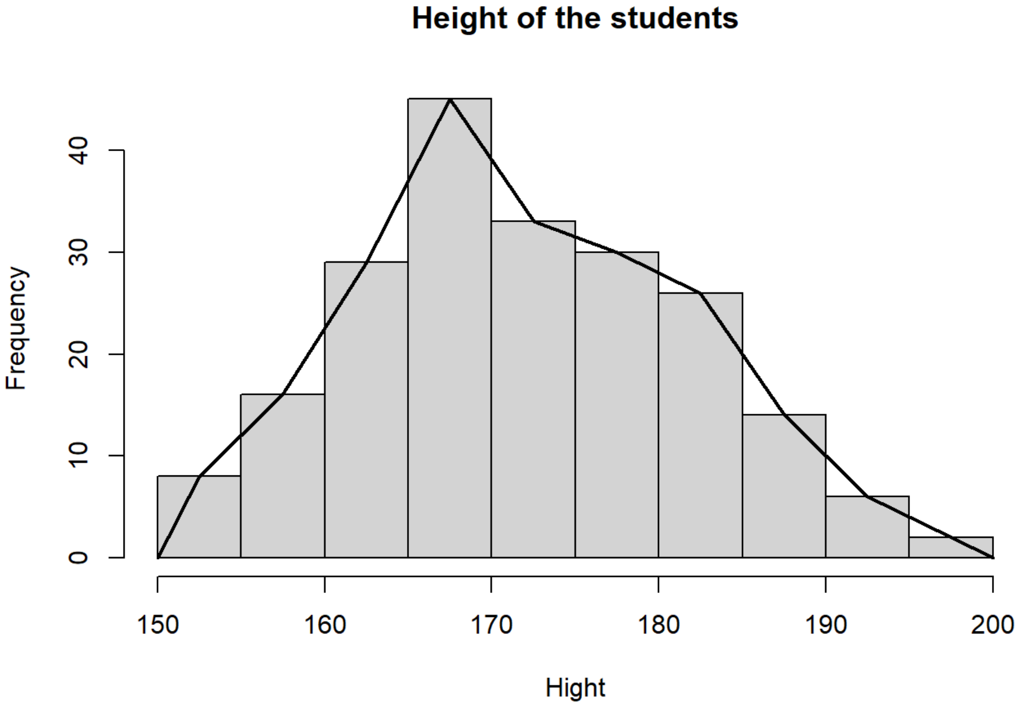


```
> library(UsingR)
Warning: package 'UsingR' was built under R version 4.0.3
Loading required package: HistData
Loading required package: Hmisc
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'
The following objects are masked from 'package:base':

    format.pval, units

Attaching package: 'UsingR'
The following object is masked from 'package:survival':

    cancer
```

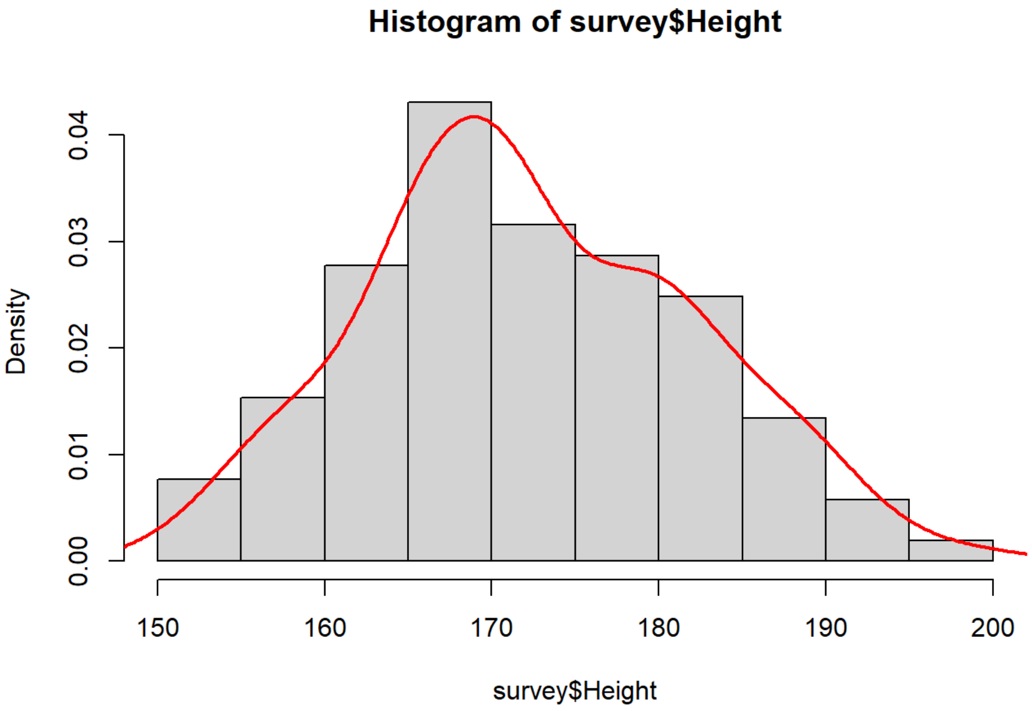
```
> simple.hist.and.boxplot(survey$Height)
```



```
> h <- hist(survey$Height, main = "Height of the students", xlab = "Hight")
> lines(x = c(min(h$breaks), h$mids, max(h$breaks)),
+       y = c(0, h$counts, 0),
+       type = "l",
+       lwd = 2)
```

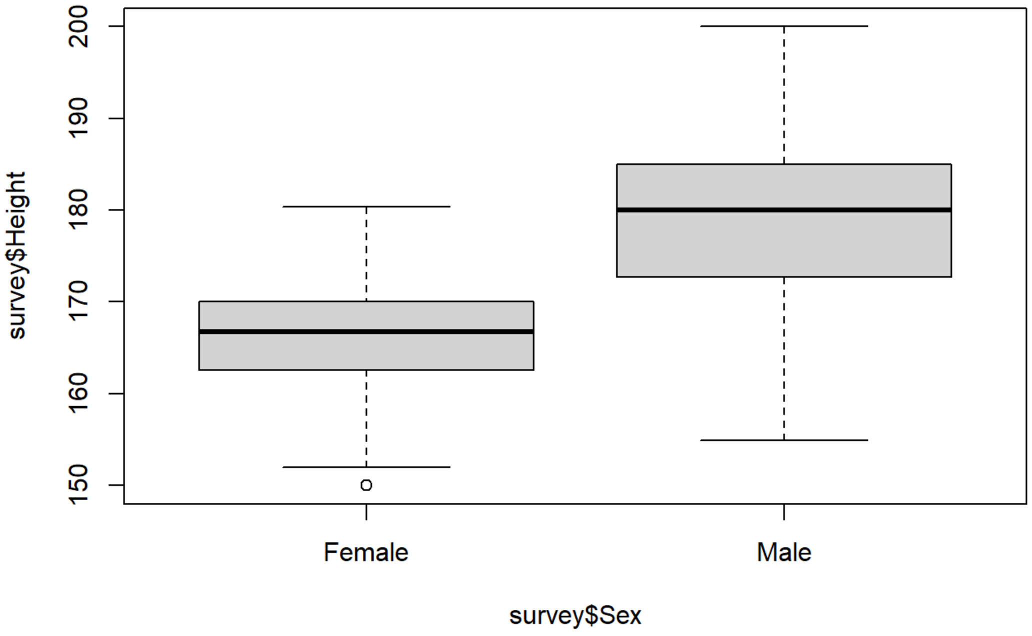


```
> hist(survey$Height, probability = TRUE)
> lines(density(survey$Height, na.rm = TRUE), lwd = 2, col = 'red')
```



Начертайте графиките за височината на студентите спрямо пола

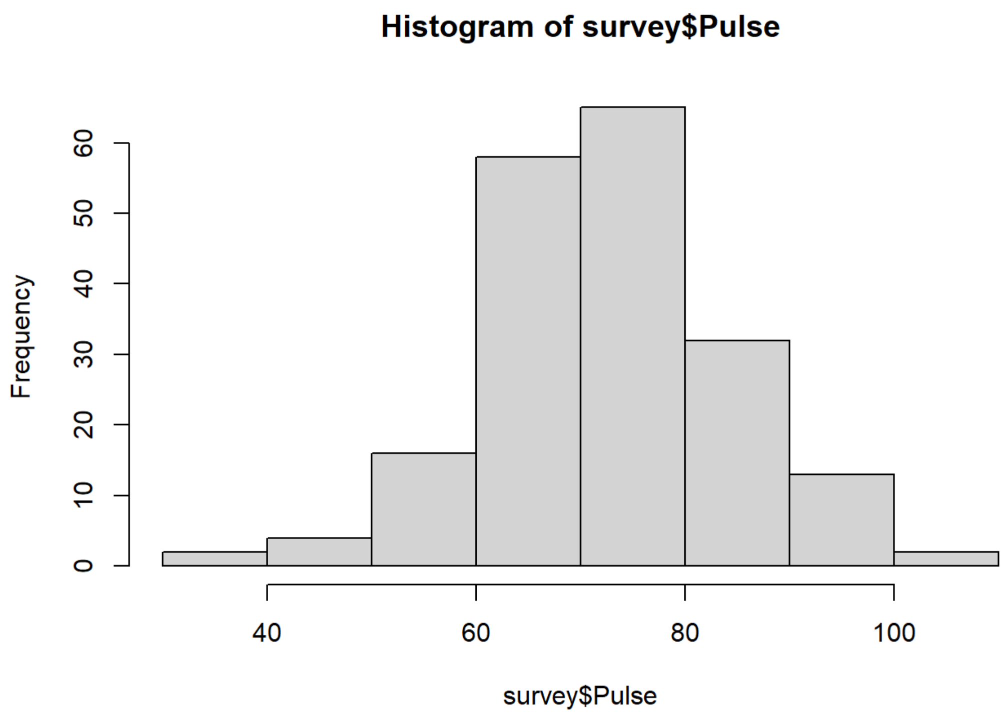
```
> boxplot(survey$Height~survey$Sex)
```



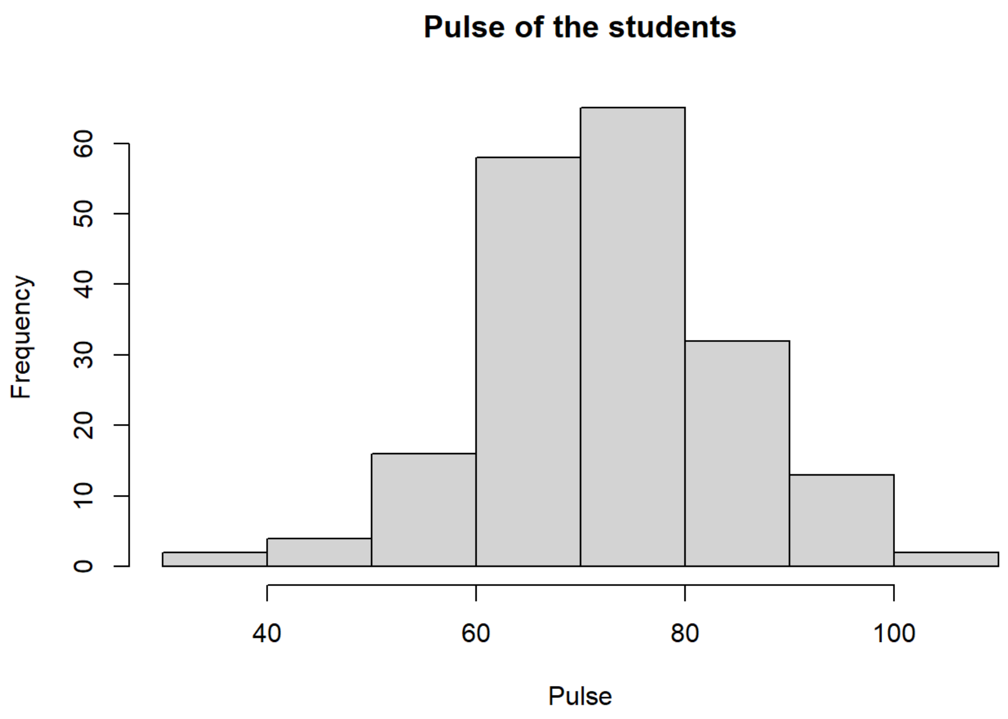
Задача 5

Направете хистограма за пулса на студентите

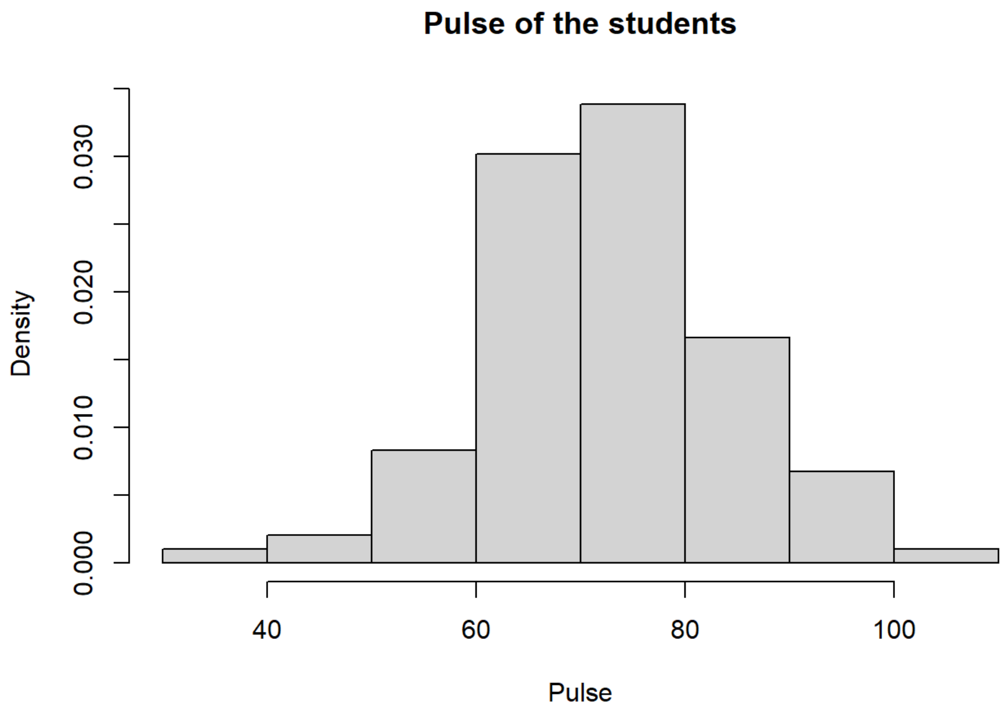
```
> hist(survey$Pulse)
```



```
> hist(survey$Pulse,
+       main = "Pulse of the students",
+       xlab = 'Pulse' )
```

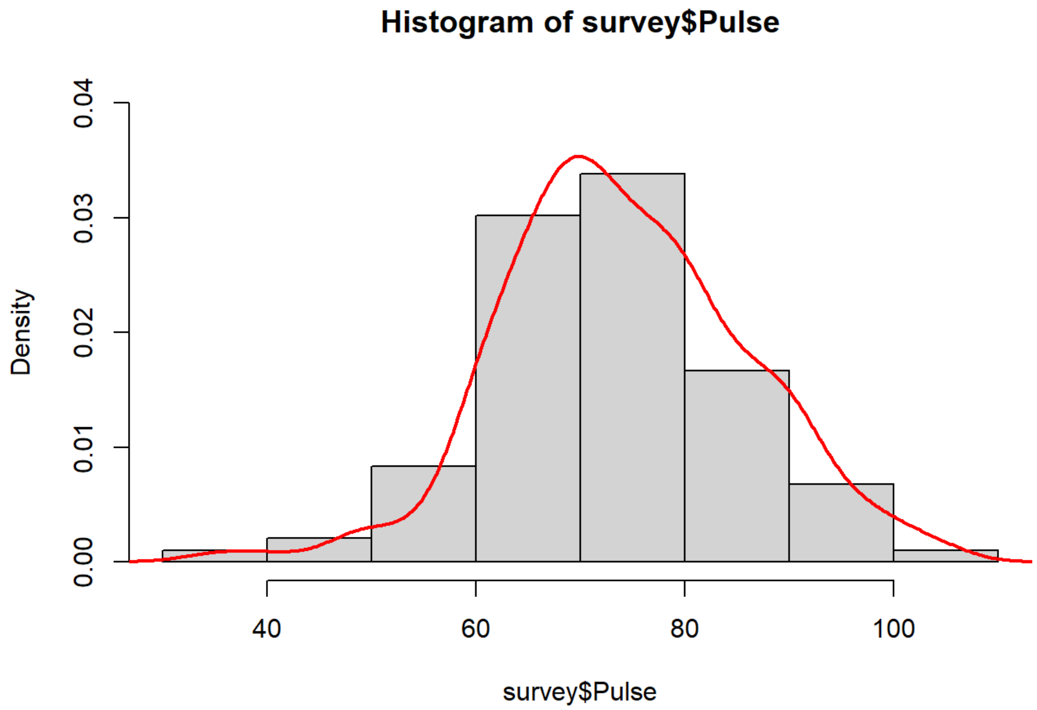


```
> hist(survey$Pulse,
+       main = "Pulse of the students",
+       xlab = 'Pulse',
+       probability = TRUE)
```



Добавете плътността.

```
> hist(survey$Pulse, probability = TRUE, ylim = c(0, 0.04))
> lines(density(survey$Pulse, na.rm = TRUE), lwd = 2, col = 'red')
```



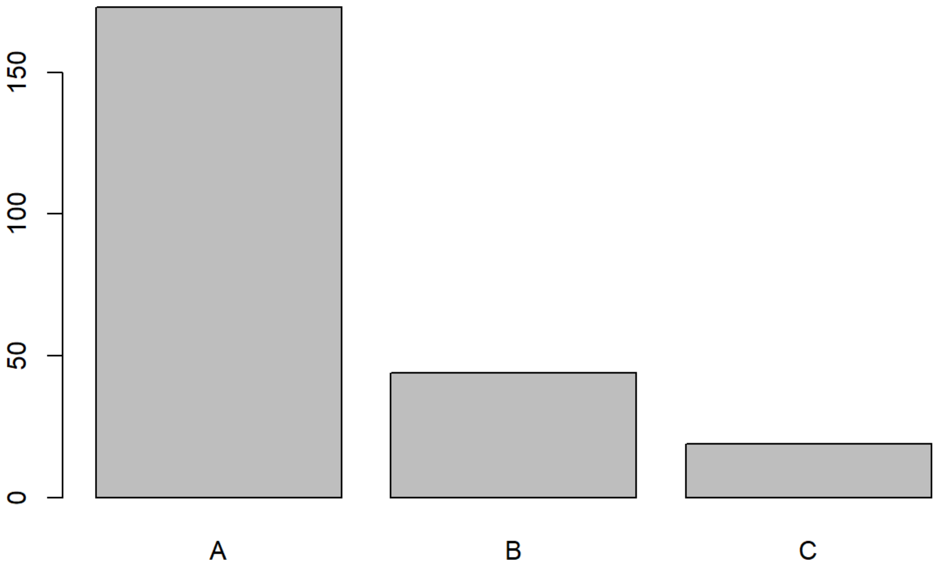
Задача 6

Разделете студентите според възрастта им на три групи: А - до 20г, В - от 20 до 25 и С – над 25.

```
> age.breaks <- c(min(survey$Age), 20, 25, max(survey$Age))
> survey$AgeCut <- cut(survey$Age, breaks = age.breaks, labels = c("A", "B", "C"))
> age.freq <- table(survey$AgeCut); age.freq
  A   B   C
173  44  19
```

Представете графично.

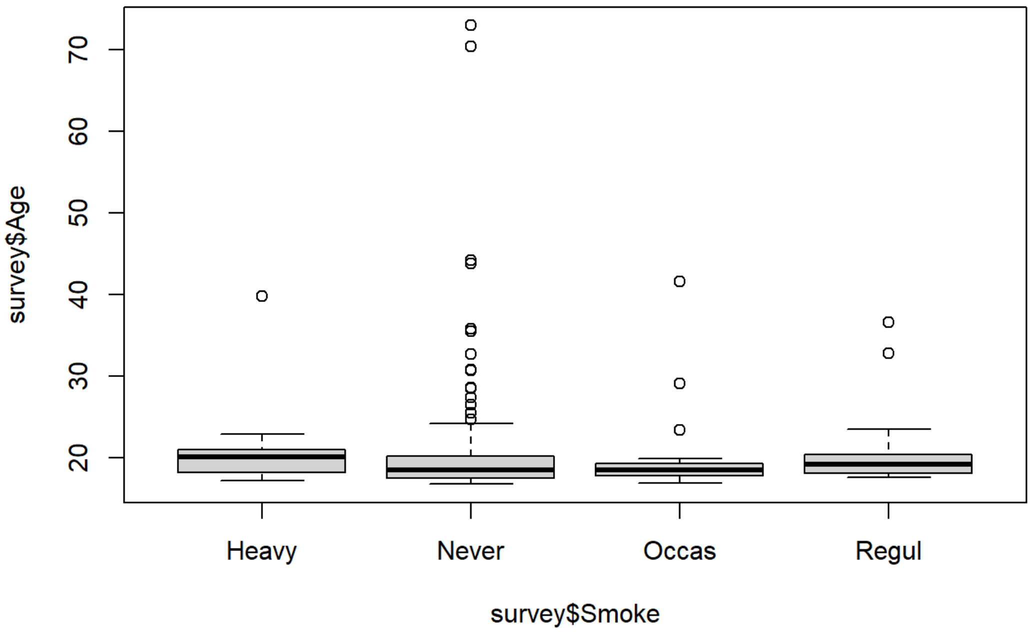
```
> barplot(age.freq)
```



Направете таблица за разпределението на пушачите в различните възрасти, представете графично.

```
> table(survey$AgeCut, survey$Smoke)
    Heavy Never Occas Regul
  A     5   139    16    12
  B     5    35     1     3
  C     1    14     2     2
```

```
> boxplot(survey$Age ~ survey$Smoke)
```


survey$Age
Heavy
Never
Occas
Regul
survey$Smoke