Зад.1 От данните 'survey' на пакета 'MASS' определете средно $\overline{X_n}$ и стандартно отклонение S_n за височината на студентите. Направете отделни изчисления за мъжете и за жените. Каква част от студентите попадат в интервалите:

а) $(X_n - S_n, X_n + S_n)$;

б) $(X_n - 2S_n, X_n + 2S_n)$;

в) $(X_n - 3S_n, X_n + 3S_n)$?

```
> summary( Height )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  NA's
  150.0   165.0   171.0  172.4   180.0  200.0    28
```

Оценките могат да бъдат пресметнати и с отделни функции, например средното $\overline{X_n}$ и стандартното отклонение S_n.

```
> m = mean( Height, na.rm = T )
172.38
> s = sd( Height, na.rm = T )
9.84
```

Ще пресметнем само за мъжете.

```
> m.m = mean( Height[ Sex == "Male"], na.rm = T )
178.82
> s.m = sd( Height[ Sex == "Male"], na.rm = T )
8.38
```

Ще определеим студентите попадащи в интервала $(X_n - S_n, X_n + S_n)$.

а) $(\overline{X_n} - S_n, \overline{X_n} + S_n)$;

```
> p = sum( m - s < Height & Height < m + s , na.rm = T )
143
> 100 * p / sum( !is.na(Height))
68.4 %
```

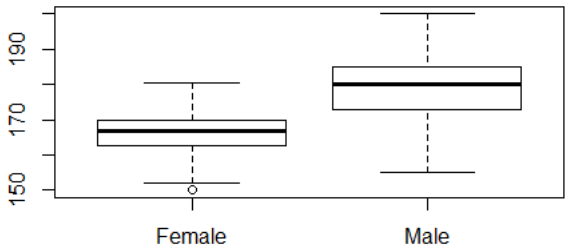
Алтернативен начин:

```
> ct = cut( Height, breaks = c(0, m-s, m+s, 300 ))
> table( ct )
 (0,163] (163,182] (182,300]
      28       143        38
> prop.table( table(ct) )
   (0,163]  (163,182]  (182,300]
 0.1339713 0.6842105 0.1818182
```

Аналогично се намират 96,6% за втория и 100% за третия интервал.

Зад.2 Представете графично височината на студентите. Постройте боксплот и хистограма, добавете полигона и плътността. Направете отделни графики за мъжете и за жените. Начертайте на една графика плътностите за ръстта на мъжете и жените.

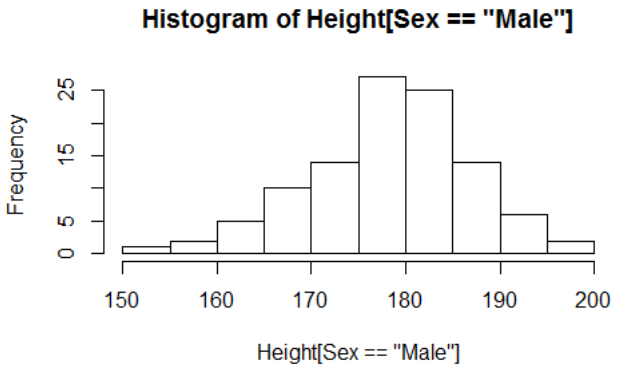
```
> boxplot(Height ~ Sex )
```



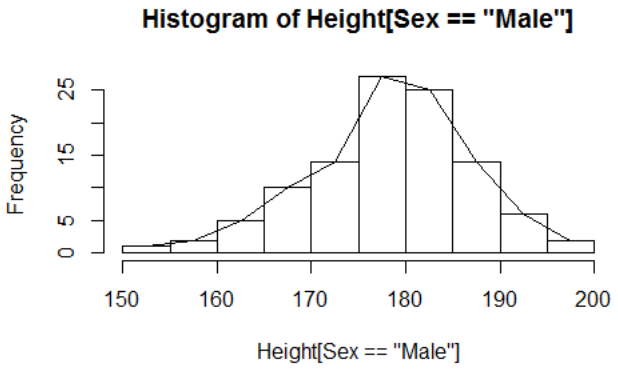
От боксплота се вижда, че жените са по-ниски от мъжете. А също и данните за мъжете са с по-голям размах, т.е. с по-голямо разсейване.

Ще направим хистограма за ръстта на мъжете, ще добавим и полигона.

```
> HM = Height[Sex == 'Male']
> h = hist( HM )
```



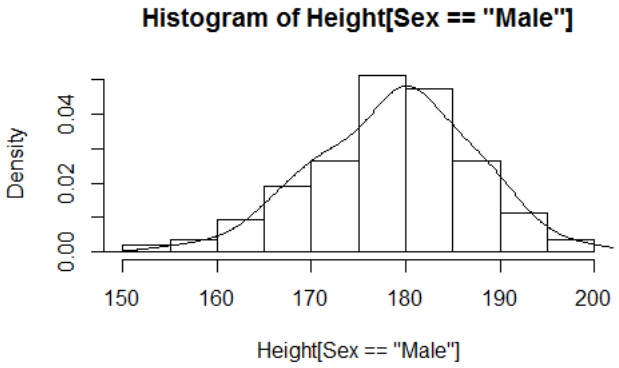

```
> lines(h$mids, h$counts )
```



Хистограмата показва, че болшинството мъже са с ръст около 180.

Ще добавим и плътността.

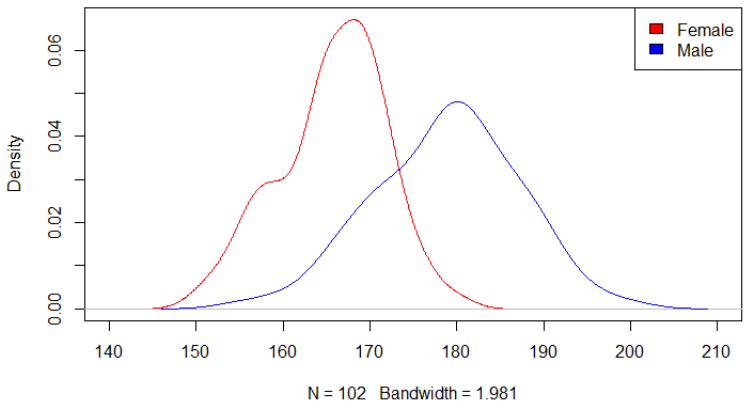
```
> hist( HM, probability = T)
> lines( density(HM, na.rm = T) )
```



Емперичната (пресметнатата по данните) плътност дава идея за истинската плътност на данните. В случая, разпределението е симетрично със средна стойност около 180, има вид на нормално.

Ще начертаем на една графика плътността на височината за мъжете и жените.

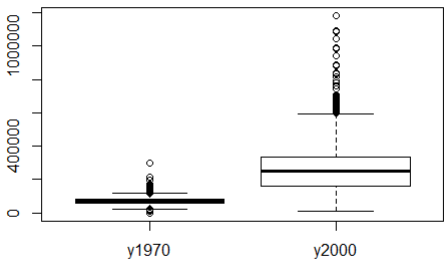
```
> HF = Height[Sex == 'Female']
> plot( density (HF, na.rm = T ), xlim = c(140, 210), col = 'red', main = " )
> lines( density (HM, na.rm = T ), col = 'blue')
> legend( 'topright', legend = c('Female', 'Male'), fill = c('red', 'blue'))
```



Отново можем да отчетем, жените са по-ниски от мъжете, освен това са и с по-малка вариация, т.е. стойностите са по-скупчени около средната стойност.

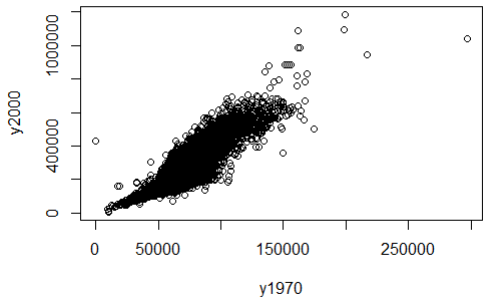
Зад.4 Разглеждаме таблицата 'homedata' от пакета 'UsingR'. Пред-ставете променливите графично - поотделно, както и заедно. Пресметнете корелацията.

```
> boxplot( homedata )
```



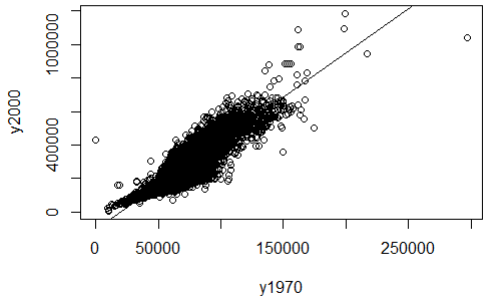
Вижда се, че в 2000 година къщите са значително по-скъпи от 1970. В 2000г. има и по-голяма вариация в цените.

```
> plot( homedata )
```



Изглежда съществува линейна връзка между цената в 1970 и в 2000г.

```
> l = lm( y2000 ~ y1970 )
> abline( l )
> cor( y1970, y2000)
      0.8962
```



Ще потърсим аутлайерите, т.е. наблюденията, които най-съществено се отличават от останалите в случая са най-далеч от правата.

```
> identify(y1970, y2000 )
      220  1064  2048
```