Moodle Tasks

Задача 1

От данните survey на пакета MASS определете средното $\overline{X}_n$ и стандартното отклонение S_n за височината на студентите.

```
> library(MASS)
> mean(survey$Height, na.rm = TRUE)
[1] 172.3809
> sd(survey$Height, na.rm = TRUE)
[1] 9.847528
```

Направете отделни изчисления за мъжете и за жените.

```
> mean(survey[survey$Sex == "Male", "Height"], na.rm = TRUE)
[1] 178.826
> mean(survey[survey$Sex == "Female", "Height"], na.rm = TRUE)
[1] 165.6867
> sd(survey[survey$Sex == "Male", "Height"], na.rm = TRUE)
[1] 8.380252
> sd(survey[survey$Sex == "Female", "Height"], na.rm = TRUE)
[1] 6.151777
```

a) Каква част от студентите попадат в интевалите:
$(\overline{X}_n - S_n, \overline{X}_n + S_n)$;

$$x_i \in (\overline{X}_n - S_n < x_i < \overline{X}_n + S_n), -S_n < x_i - \overline{X}_n < S_n,$$

$$-1 < \frac{x_i - \overline{X}_n}{S_n} < 1, \left| \frac{x_i - \overline{X}_n}{S_n} \right| < 1$$

```
> height.clean <- survey$Height[!is.na(survey$Height)]
> height.standardized <- abs(height.clean - mean(height.clean)) / sd(height.clean)
> sum(height.standardized < 1) / length(height.clean)
[1] 0.6842105
```

b) $(\overline{X}_n - 2S_n, \overline{X}_n + 2S_n)$;

$$x_i \in (\overline{X}_n - 2S_n < x_i < \overline{X}_n + 2S_n), \ -2 < \frac{x_i - \overline{X}_n}{S_n} < 2, \left| \frac{x_i - \overline{X}_n}{S_n} \right| < 2$$

```
> sum(height.standardized < 2) / length(height.clean)
[1] 0.9665072
```

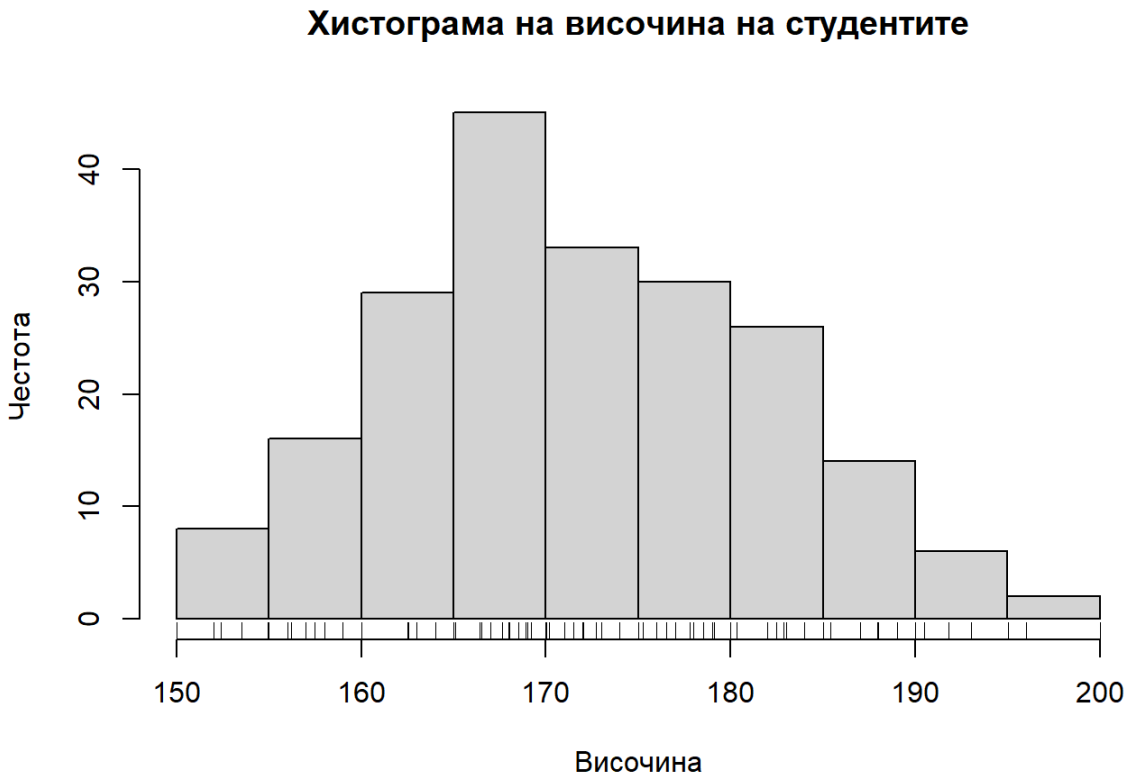
c) $(\overline{X}_n - 3S_n, \overline{X}_n + 3S_n)$;

$$x_i \in (\overline{X}_n - 3S_n < x_i < \overline{X}_n + 3S_n), \ -3 < \frac{x_i - \overline{X}_n}{S_n} < 3, \left| \frac{x_i - \overline{X}_n}{S_n} \right| < 3$$

```
> sum(height.standardized < 3) / length(height.clean)
[1] 1
```

Направете хистограма за височината на студентите.

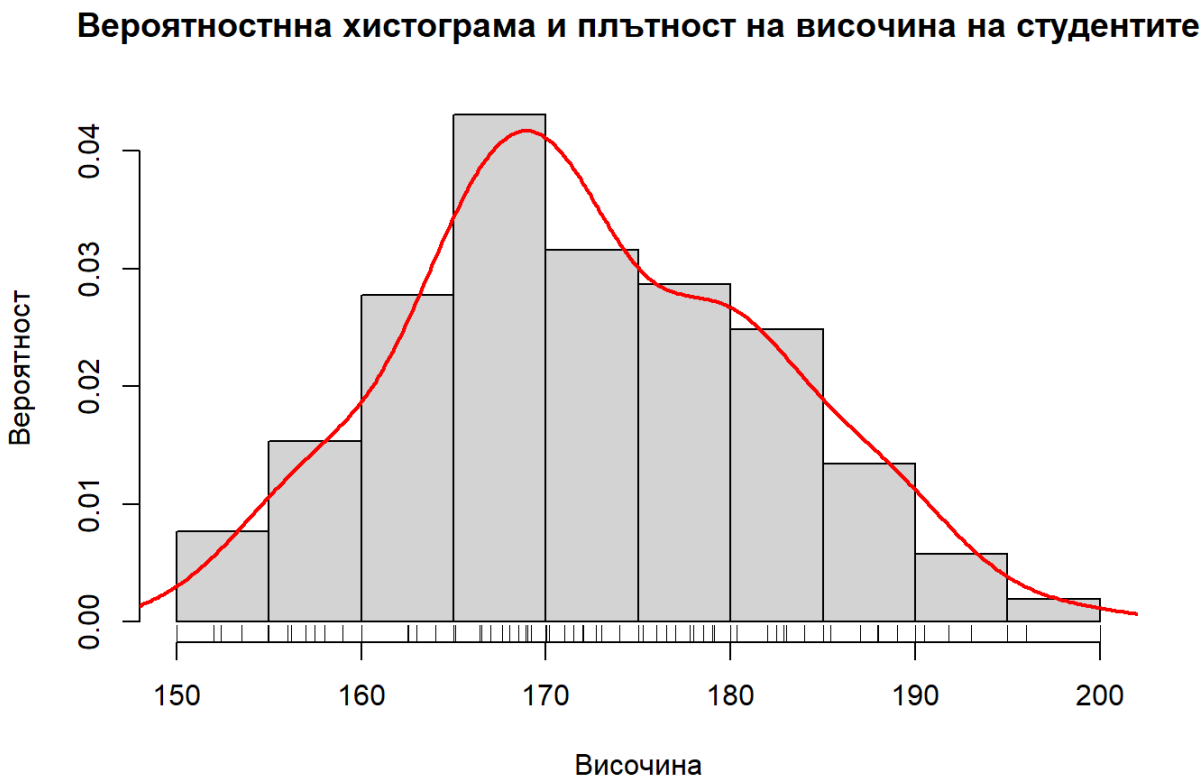
```
> hist(survey$Height,
+      main = "Хистограма на височина на студентите",
+      xlab = "Височина",
+      ylab = "Честота")
> rug(jitter(survey$Height))
```



Добавете полигона и плътността.

```
> h <-hist(survey$Height,
+          main = "Хистограма и полигон на височина на студентите",
+          xlab = "Височина",
+          ylab = "Честота")
> lines(x = c(min(h$breaks), h$mids, max(h$breaks)),
+       y = c(0, h$counts, 0),
+       type = "l",
+       lwd = 2)
```

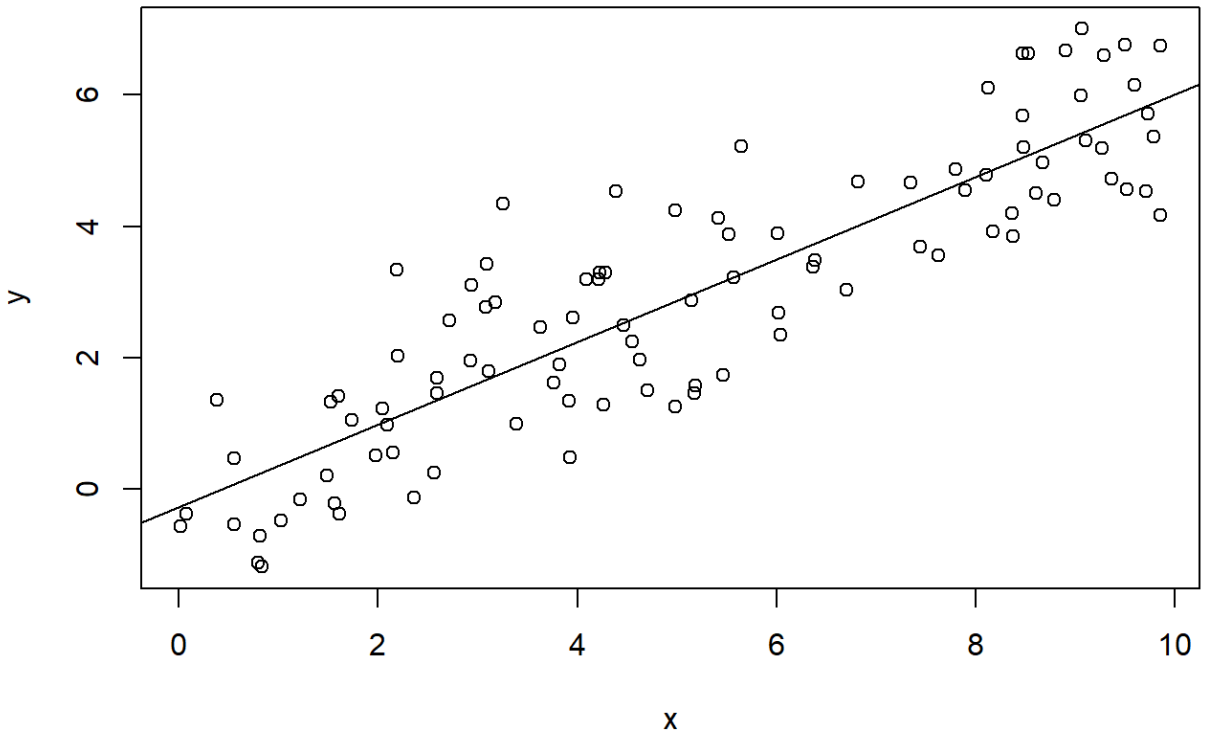

```
> hist(height.clean,
+      main = "Вероятностнна хистограма и плътност на височина на студентите",
+      xlab = "Височина",
+      ylab = "Вероятност",
+      probability = TRUE)
> rug(jitter(height.clean))
> lines(density(height.clean),
+       col='red',
+       lwd = 2)
```



Задача 2

Представете графично данните от файла Data.txt.

```
> data <- read.table("Data.txt", header = TRUE)
> head(data)
         x         y
1 5.1756426 1.4663523
2 8.3717110 3.8488988
3 0.3805756 1.3589665
4 4.3884145 4.5357171
5 1.9780963 0.5158695
6 0.5544184 0.4695551

> plot(data)
> abline(lm(data$y ~ data$x))
```




```
> cor(data$x, data$y)
[1] 0.8800885
```

Задача 3

Разгледайте данните anscombre.

```
> str(anscombe)
'data.frame':   11 obs. of  8 variables:
 $ x1: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x2: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x3: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x4: num  8 8 8 8 8 8 8 19 8 8 ...
 $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
 $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26 ...
 $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
 $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.91 ...
```

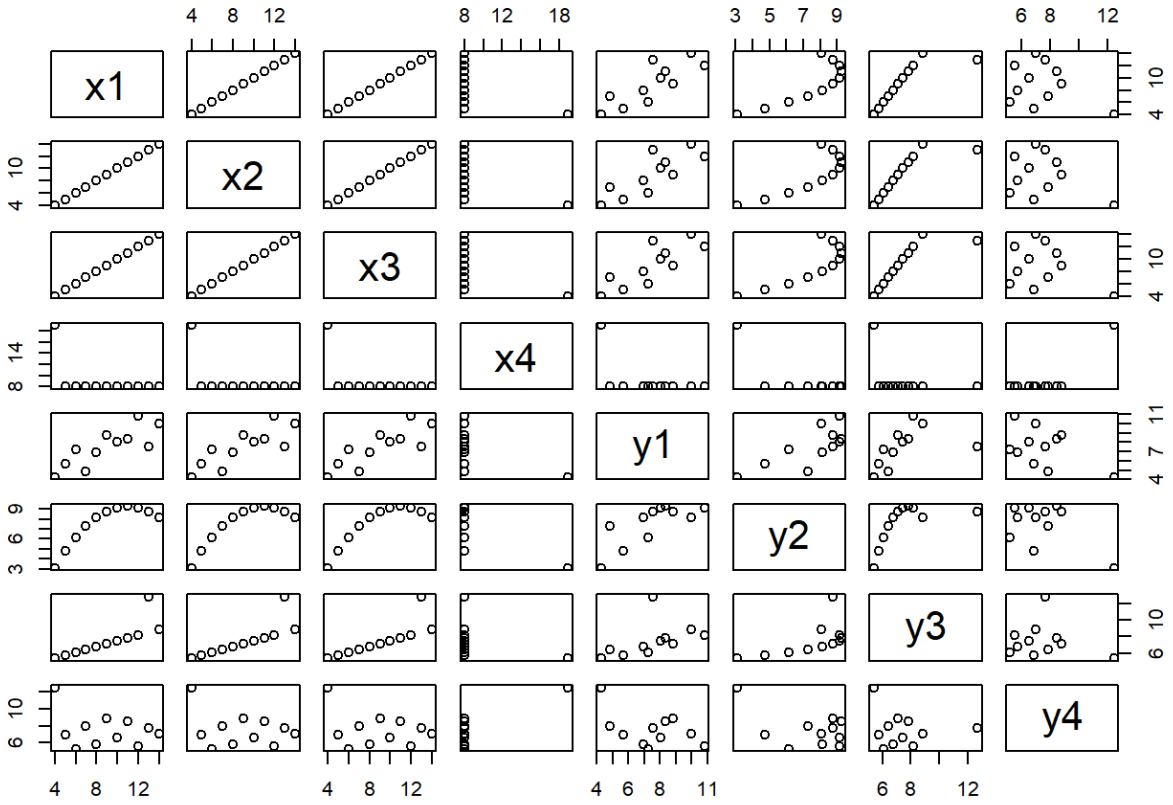
```
> summary(anscombe)
       x1              x2              x3              x4              y1
 Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8   Min.   : 4.260
 1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8   1st Qu.: 6.315
 Median : 9.0   Median : 9.0   Median : 9.0   Median : 8   Median : 7.580
 Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9   Mean   : 7.501
 3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8   3rd Qu.: 8.570
 Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19   Max.   :10.840
```

```
       y2              y3              y4
 Min.   :3.100   Min.   : 5.39   Min.   : 5.250
 1st Qu.:6.695   1st Qu.: 6.25   1st Qu.: 6.170
 Median :8.140   Median : 7.11   Median : 7.040
 Mean   :7.501   Mean   : 7.50   Mean   : 7.501
 3rd Qu.:8.950   3rd Qu.: 7.98   3rd Qu.: 8.190
 Max.   :9.260   Max.   :12.74   Max.   :12.500
```

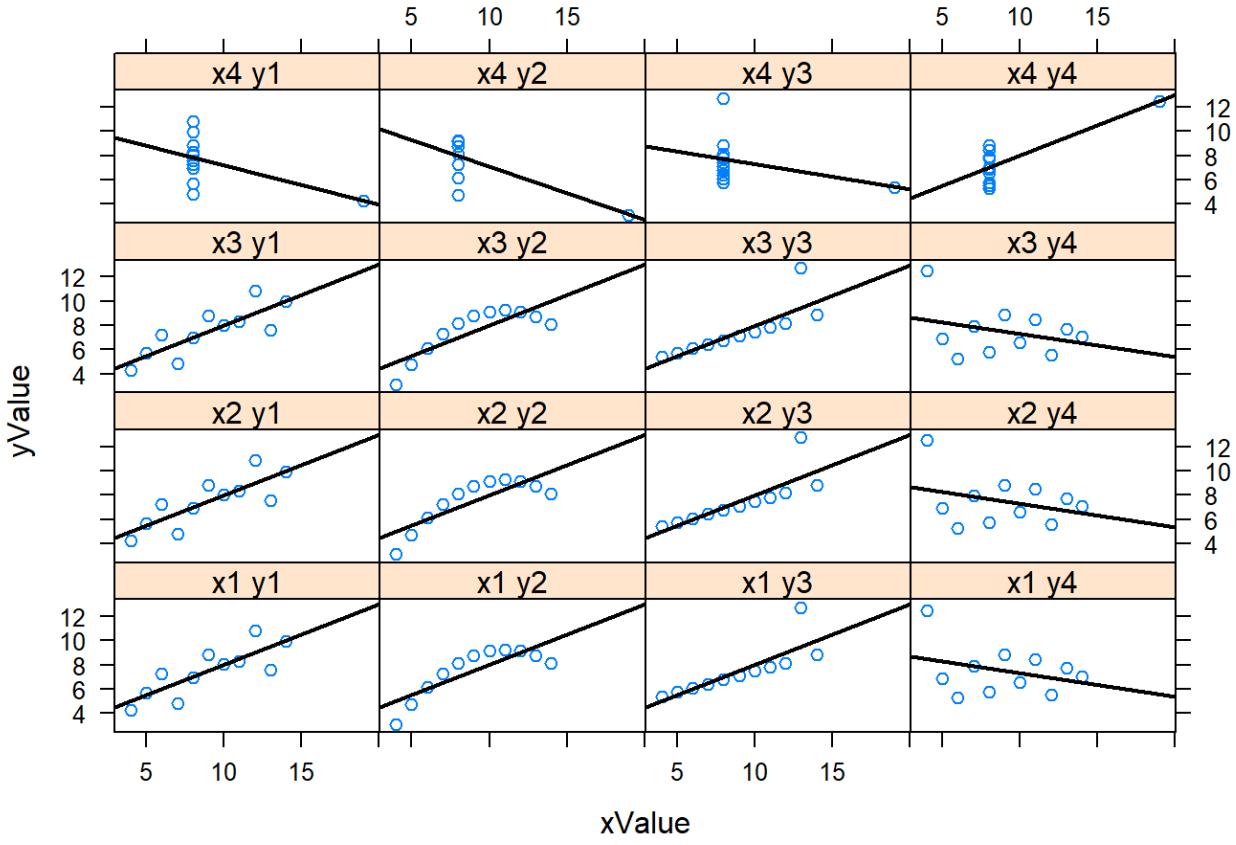
За всяка двойка $(x_i, y_i)_{i=1...4}$ пресметнете числовите характеристики, представете графично, отстранете outliers.

```
> cor(anscombe[, 1:4], anscombe[, 5:8])
          y1        y2        y3        y4
x1  0.8164205  0.8162365  0.8162867 -0.3140467
x2  0.8164205  0.8162365  0.8162867 -0.3140467
x3  0.8164205  0.8162365  0.8162867 -0.3140467
x4 -0.5290927 -0.7184365 -0.3446610  0.8165214
```

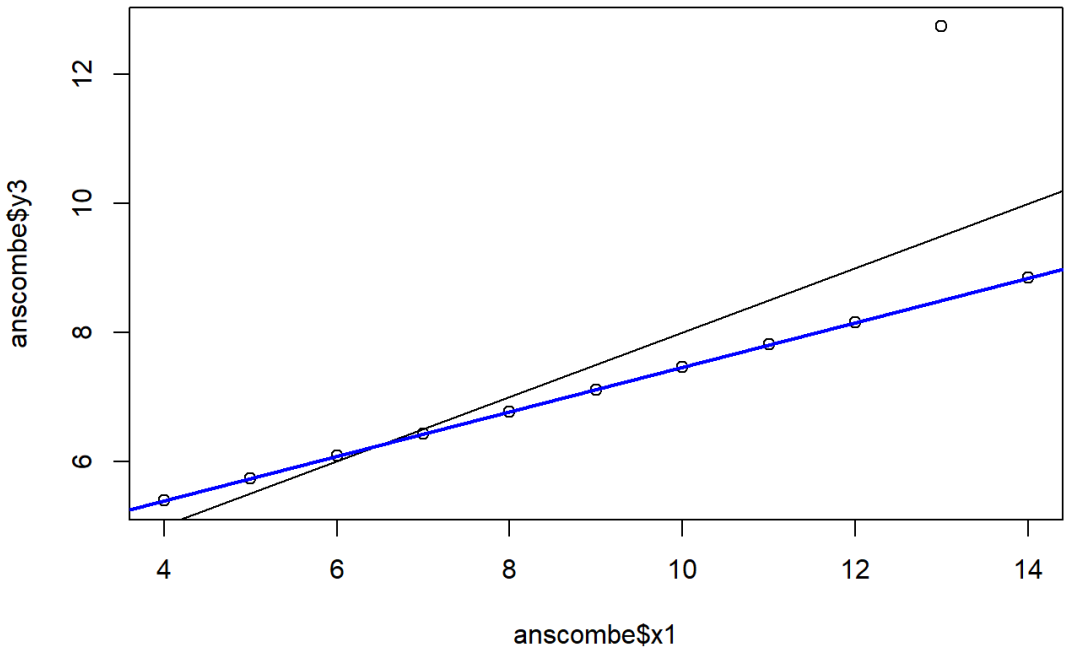
```
> pairs(anscombe)
```



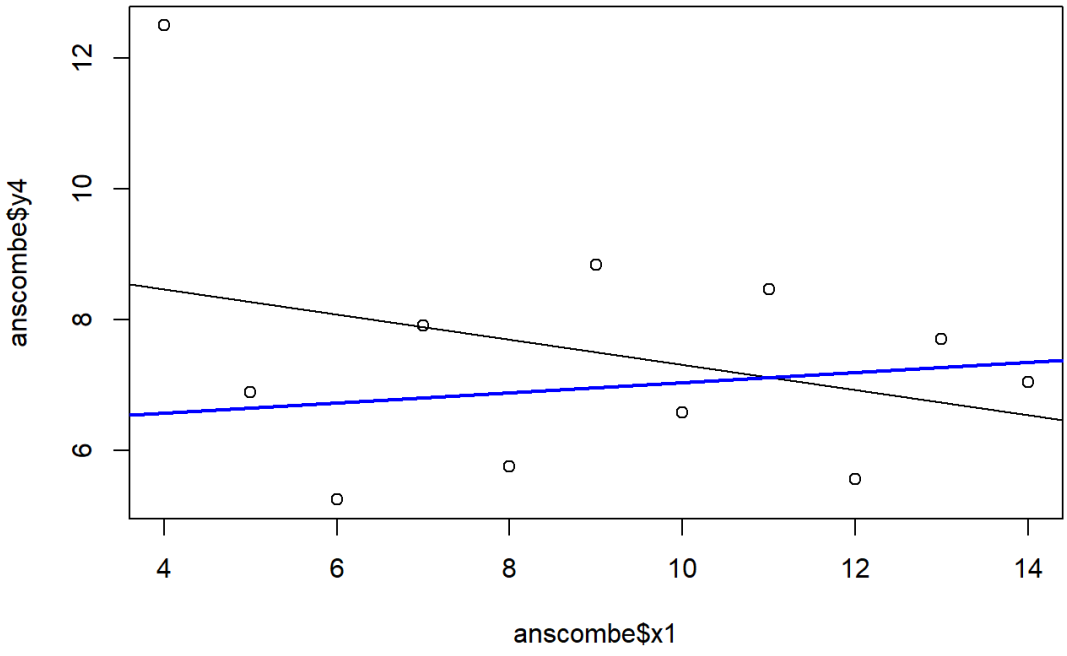
```
> library(lattice)
> x <- c(rep("x1", 11), rep("x2", 11), rep("x3", 11), rep("x4", 11))
> anscombre.stack <- data.frame(xValue = rep(c(anscombe$x1, anscombe$x2,
anscombe$x3, anscombe$x4), 4),
+                               yValue = c(rep(anscombe$y1, 4), rep(anscombe$y2, 4),
rep(anscombe$y3, 4), rep(anscombe$y4, 4)),
+                          category = c(paste(x, "y1"), paste(x, "y2"), paste(x, "y3"), paste(x,
"y4")))
> xyplot(yValue ~ xValue | category, data = anscombre.stack,
+        panel = function(x, y, ...){
+          panel.xyplot(x, y, ...)
+          fit <- lm(y ~ x)
+          panel.abline(fit, lwd = 2)
+        },
+        layout = c(4, 4))
```



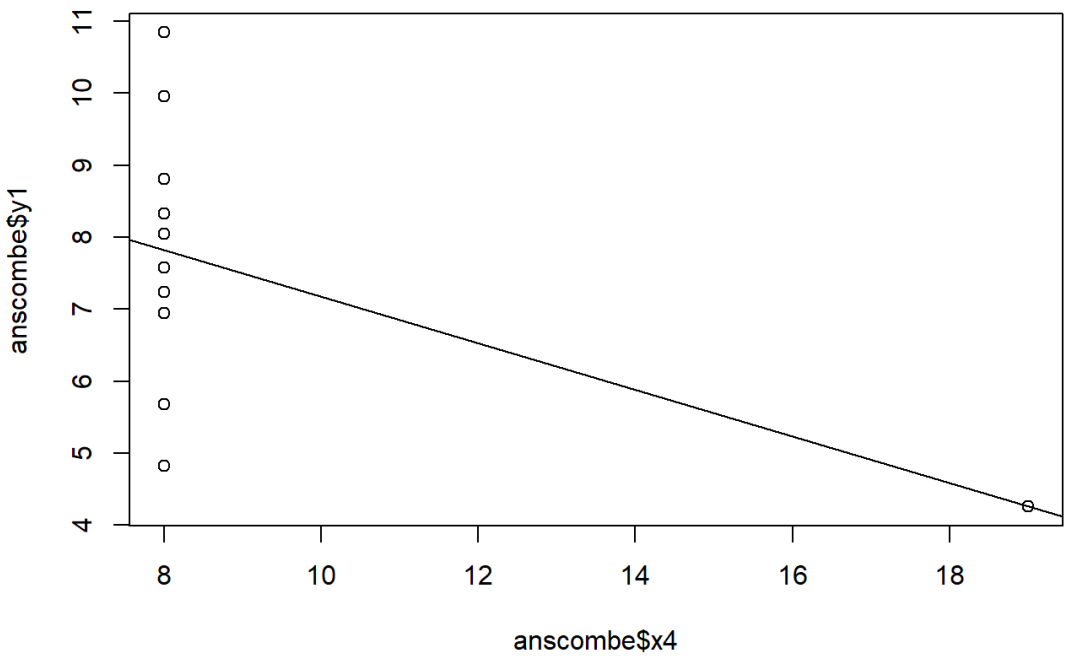

```
> plot(anscombe$y3 ~ anscombe$x1)
> abline(lm(anscombe$y3 ~ anscombe$x1))
> identify(anscombe$x1, anscombe$y3, n = 1)
integer(0)
> abline(lm(anscombe[-3, "y3"] ~ anscombe[-3, "x1"]), col = "Blue", lwd = 2)
```




```
> plot(anscombe$y4 ~ anscombe$x1)
> abline(lm(anscombe$y4 ~ anscombe$x1))
> identify(anscombe$x1, anscombe$y4, n = 1)
integer(0)
> abline(lm(anscombe[-8, "y4"] ~ anscombe[-8, "x1"]), col = "Blue", lwd = 2)
```



```
> plot(anscombe$y1 ~ anscombe$x4)
> abline(lm(anscombe$y1 ~ anscombe$x4))
> identify(anscombe$x4, anscombe$y1, n = 1)
integer(0)
> abline(lm(anscombe[-8, "y1"] ~ anscombe[-8, "x4"]), col = "Blue", lwd = 2)
```



Задача 4

Разгледайте данните titanic.

```
> titanic <- read.csv("../Data/titanic.csv")
> str(titanic)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...

> attach(titanic)
```

Има ли връзка между пола и шанса за оцеляване?

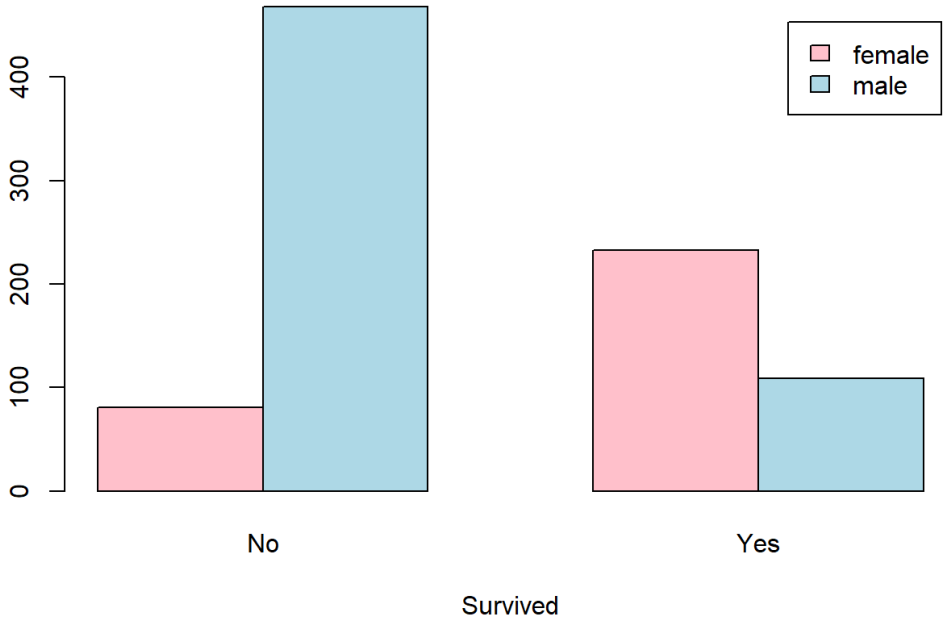
```
> table(Sex, Survived)
        Survived
Sex        0   1
  female  81 233
  male   468 109

> prop.table(table(Sex, Survived), 1)
        Survived
Sex              0         1
  female 0.2579618 0.7420382
  male   0.8110919 0.1889081
```

А между класата и оцеляването?

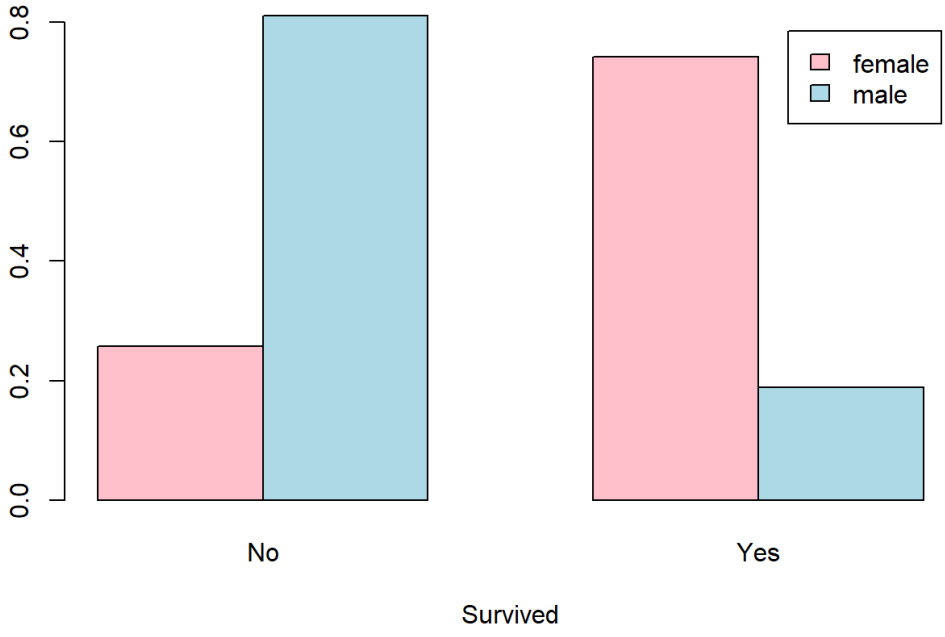
```
> table(Pclass, Survived)
      Survived
Pclass   0   1
     1  80 136
     2  97  87
     3 372 119

> prop.table(table(Pclass, Survived), 1)
      Survived
Pclass         0         1
     1 0.3703704 0.6296296
     2 0.5271739 0.4728261
     3 0.7576375 0.2423625
```

Направете подходящи графики.

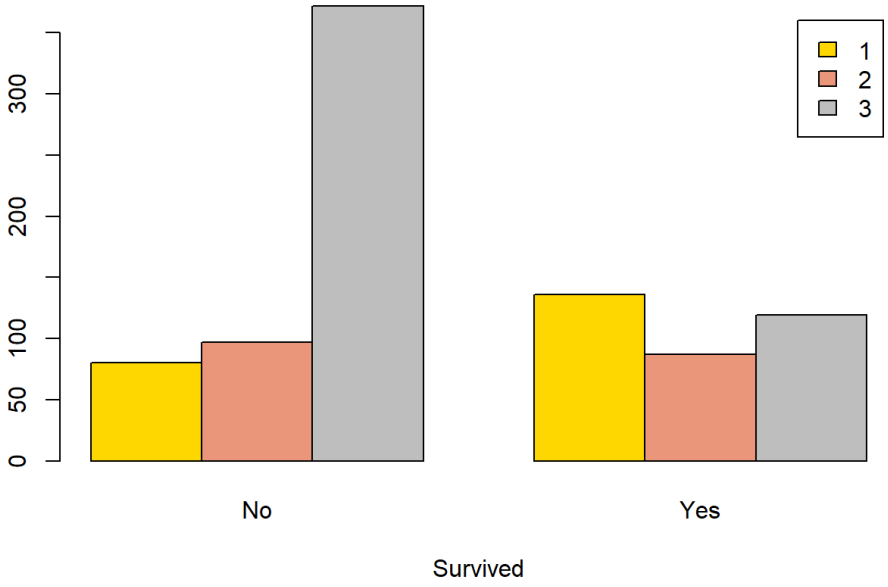
```
> barplot(table(Sex, Survived),
+         beside = TRUE,
+         xlab = "Survived",
+         names.arg = c("No", "Yes"),
+         col = c("pink", "lightblue"),
+         legend.text = TRUE)
```




```
> barplot(prop.table(table(Sex, Survived), 1),
+         beside = TRUE,
+         xlab = "Survived",
+         names.arg = c("No", "Yes"),
+         col = c("pink", "lightblue"),
+         legend.text = TRUE)
```



```
> barplot(table(Pclass, Survived),
+         beside = TRUE,
+         xlab = "Survived",
+         names.arg = c("No", "Yes"),
+         col = c("gold", "darksalmon", "gray"),
+         legend.text = TRUE)
```




```
> barplot(prop.table(table(Pclass, Survived), 1),
+         beside = TRUE,
+         xlab = "Survived",
+         names.arg = c("No", "Yes"),
+         col = c("gold", "darksalmon", "gray"),
+         legend.text = TRUE)
```