Учебен проект

Практикум Статистика и Емпирични Методи
2019-2020 уч. година

Изготвил: Александра Йовкова
ф.н. 62229 , III курс

За целта на този проект, беше създадена анкета, разпространена и попълнена от 61 души към днешна дата.
Анкетата се намира на следния линк: https://docs.google.com/forms/d/e/1FAIpQLSdHLHixXZhvFrdP6cqyf7ZyNw8lY-ul26sq_J7_GaVXH03eAw/viewform
Събраните данни бяха използвани за съставяне на графики и базов анализ на взаимоотношението между отделни фактори.

1. Структура на данните

```
$ Timestamp      : Factor w/ 61 levels "2019/12/27 5:14:14 сл.об. Гринуич+2",..: 1 2 3 4 5 6 7 8 9 10 ...
$ Gender         : Factor w/ 2 levels "Жена","Мъж": 1 2 2 2 1 1 2 2 2 1 ...
$ Age            : int  21 21 22 19 45 16 21 22 22 18 ...
$ Activity.level : int  2 3 7 8 4 4 7 6 4 3 ...
$ Sleep.hours    : Factor w/ 4 levels "10 до 12 часа (но по-малко от 12)",..: 3 3 3 1 2 3 4 4 3 2 ...
$ Take.pills     : logical  TRUE FALSE FALSE FALSE FALSE FALSE ...
$ Stress.level   : int  7 5 3 1 9 4 7 5 1 10 ...
$ Caffeine.drinks.per.day: int  2 2 2 0 3 0 0 1 0 0 …
```

data.csv файлът съдържа записи с горните 8 променливи за всеки от тях. Timestamp колоната няма да ни интересува в рамките на този проект. За по-ясна представа, нека видим първите 5 записа от този файл:

```
> head(data, n=5)
                        Timestamp Gender Age Activity.level                    Sleep.hours Take.pills Stress.level Caffeine.drinks.per.day
1 2019/12/27 5:14:14 сл.об. Гринуич+2   Жена  21              2    6 до 8 часа (но по-малко от 8)       TRUE            7                       2
2 2019/12/27 5:22:01 сл.об. Гринуич+2    Мъж  21              3    6 до 8 часа (но по-малко от 8)      FALSE            5                       2
3 2019/12/27 5:22:56 сл.об. Гринуич+2    Мъж  22              7    6 до 8 часа (но по-малко от 8)      FALSE            3                       2
4 2019/12/27 5:23:02 сл.об. Гринуич+2    Мъж  19              8 10 до 12 часа (но по-малко от 12)      FALSE            1                       0
5 2019/12/27 5:29:17 сл.об. Гринуич+2   Жена  45              4    4 до 6 часа (но по-малко от 6)      FALSE            9                       3
>
```

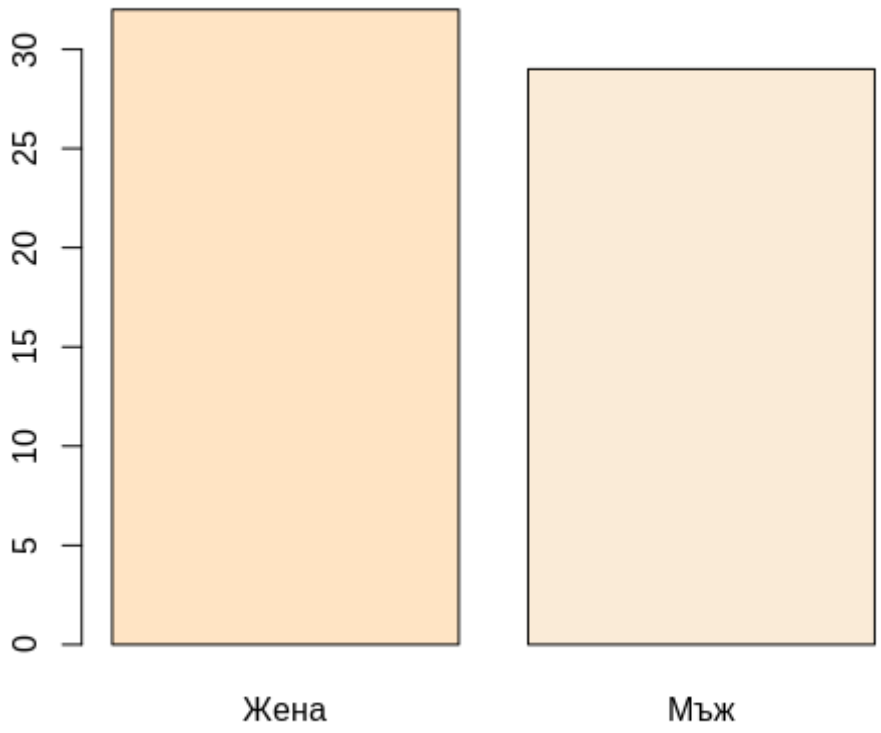
2. Базова информация и статистика за данните

```
> summary(data)
                          Timestamp   Gender        Age       Activity.level                          Sleep.hours Take.pills       Stress.level
 2019/12/27 5:14:14 сл.об. Гринуич+2: 1   Жена:32   Min.   :16.0   Min.   : 0.000   10 до 12 часа (но по-малко от 12): 1   Mode :logical   Min.   : 0.000
 2019/12/27 5:22:01 сл.об. Гринуич+2: 1   Мъж :29   1st Qu.:21.0   1st Qu.: 4.000   4 до 6 часа (но по-малко от 6)   :17   FALSE:52        1st Qu.: 3.000
 2019/12/27 5:22:56 сл.об. Гринуич+2: 1             Median :21.0   Median : 6.000   6 до 8 часа (но по-малко от 8)   :31   TRUE :9         Median : 6.000
 2019/12/27 5:23:02 сл.об. Гринуич+2: 1             Mean   :23.9   Mean   : 5.525   8 до 10 часа (но по-малко от 10) :12                   Mean   : 5.738
 2019/12/27 5:29:17 сл.об. Гринуич+2: 1             3rd Qu.:23.0   3rd Qu.: 7.000                                                         3rd Qu.: 8.000
 2019/12/27 5:32:52 сл.об. Гринуич+2: 1             Max.   :55.0   Max.   :10.000                                                         Max.   :10.000
 (Other)                            :55
 Caffeine.drinks.per.day
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 2.000
 Mean   : 1.787
 3rd Qu.: 2.000
 Max.   :10.000
```

3. Графично представяне

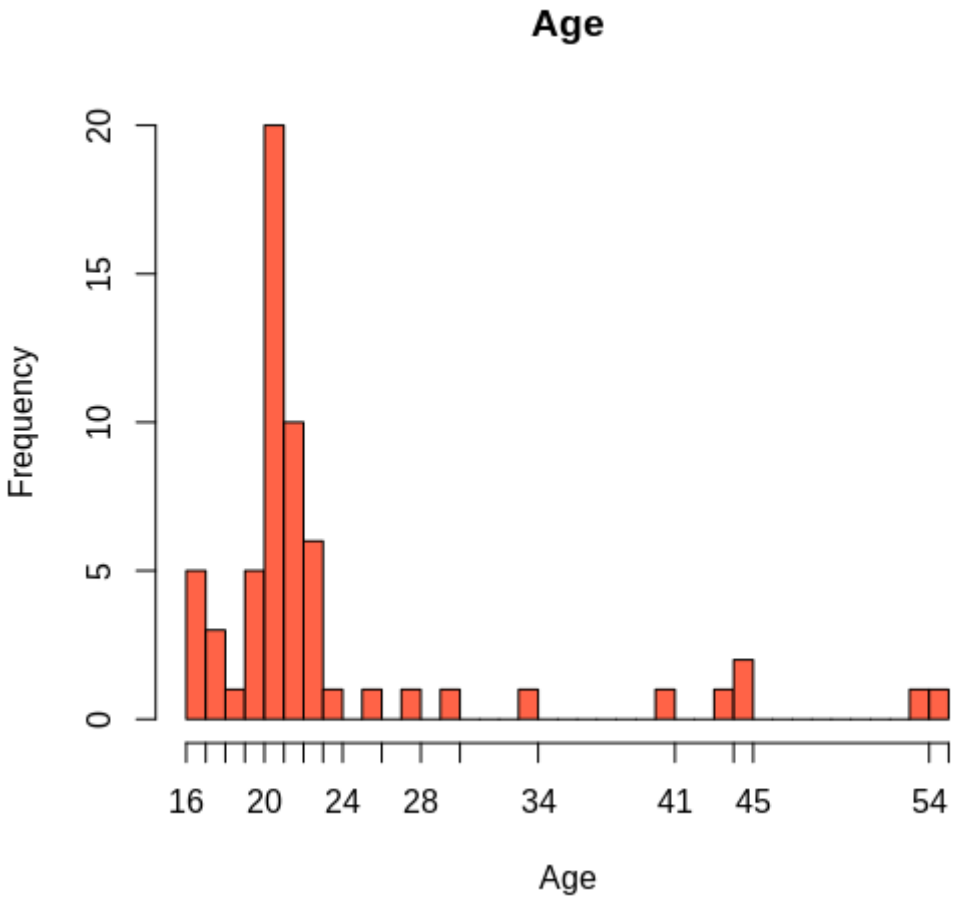
3.1. Самостоятелно графично представяне на всяка променлива

- Пол – категорийна променлива




```
> #gender
> plot(data$Gender, col=c('bisque', 'antiquewhite')) # distribution male/female
>
```

- Възраст – числова променлива

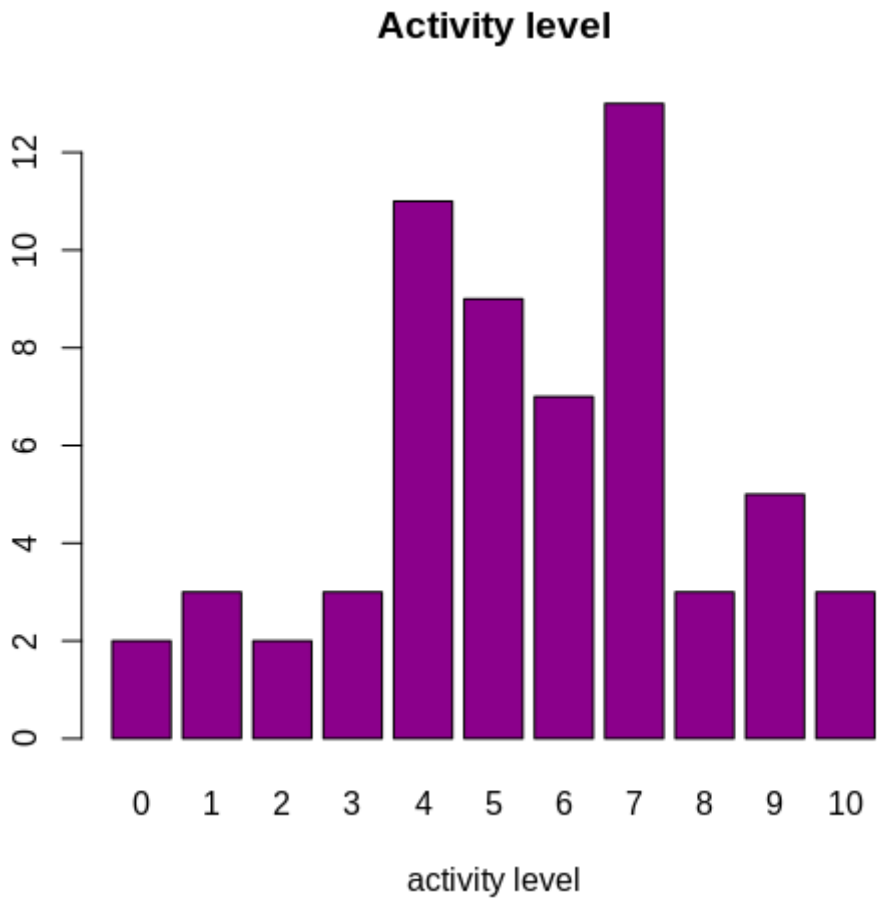



```
> #age
> histogram <- hist(data$Age, breaks=30, main="Age", xlab="Age", col="tomato", labels=levels(data$Age), xaxt='n')
> axis(1, at=unique(data$Age), labels=levels(data$Age))
>
```

- Ниво на стрес

```
>   #stress levels
>   barplot(table(data$Stress.level), main='Stress level', xlab='Stress level', col='mediumvioletred', space=0)
>
```

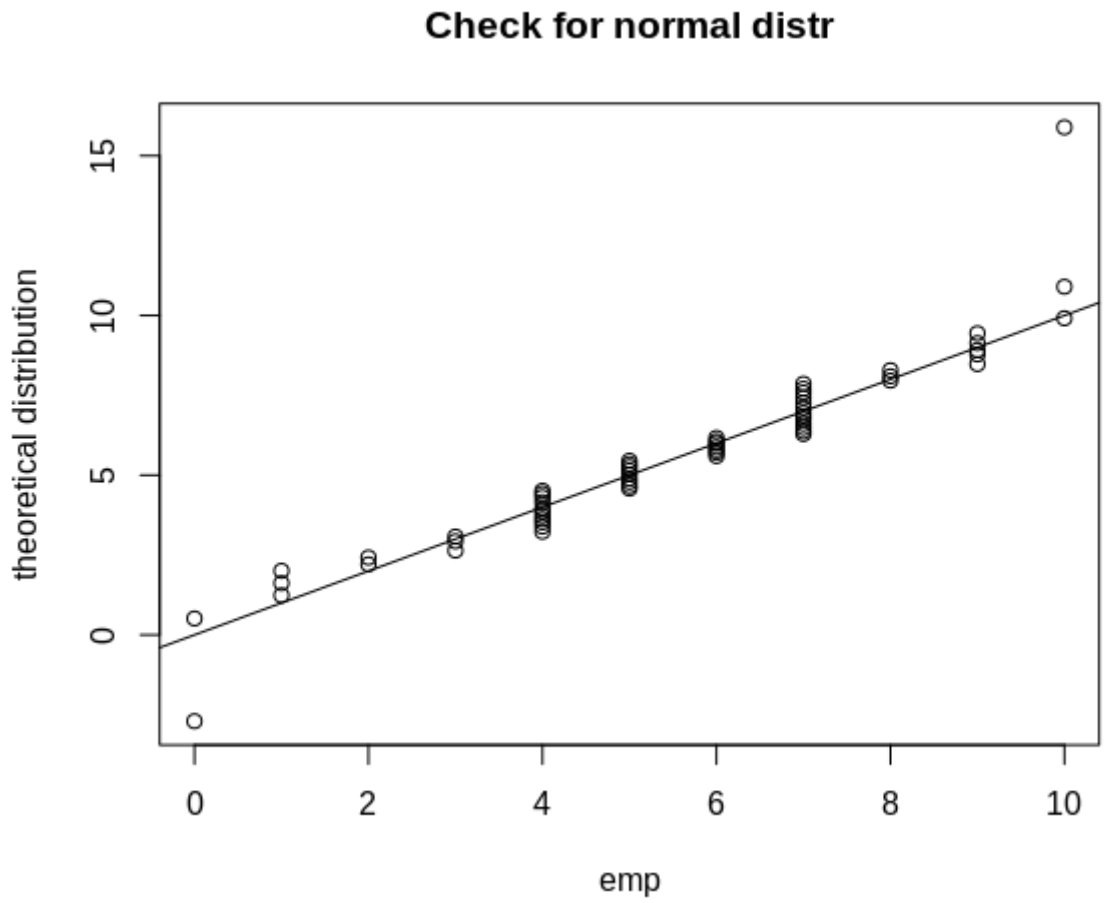
- Ниво на физическа активност



От графиката забелязваме, че разпределението е близко до нормалното. Правим проверка:

```
> emp <- data$Activity.level
> d <- rnorm(n = 10^3, mean = mean(emp), sd = sd(emp))
> qqplot(emp, d, ylab = "theoretical distribution", main = "Check for normal distr")
>
> abline(a = 0, b = 1)
>
```

И резултатът е следният:

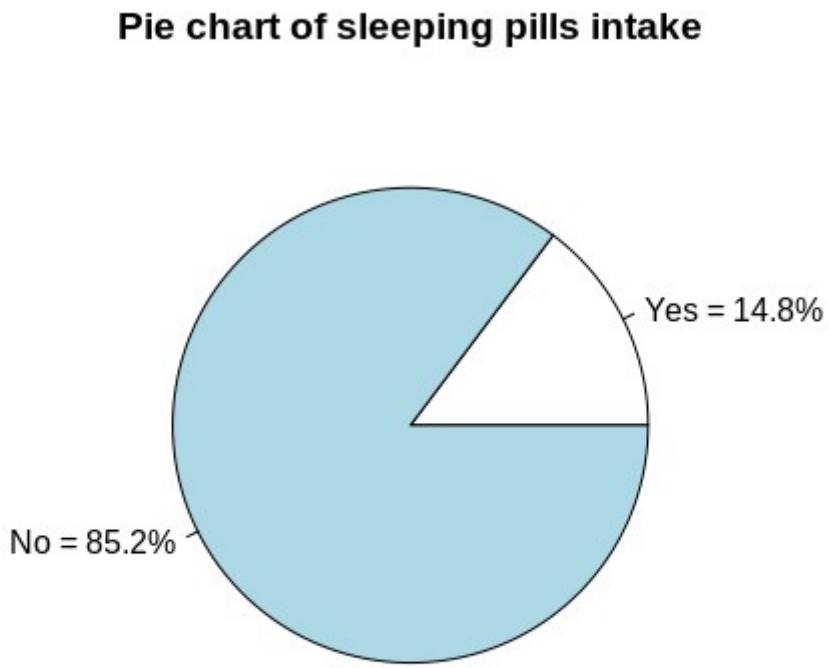


Друг начин за проверка е чрез `shapiro.test`:

```
> shapiro.test(data$Activity.level)

        Shapiro-Wilk normality test

data:  data$Activity.level
W = 0.96496, p-value = 0.07796

>
```

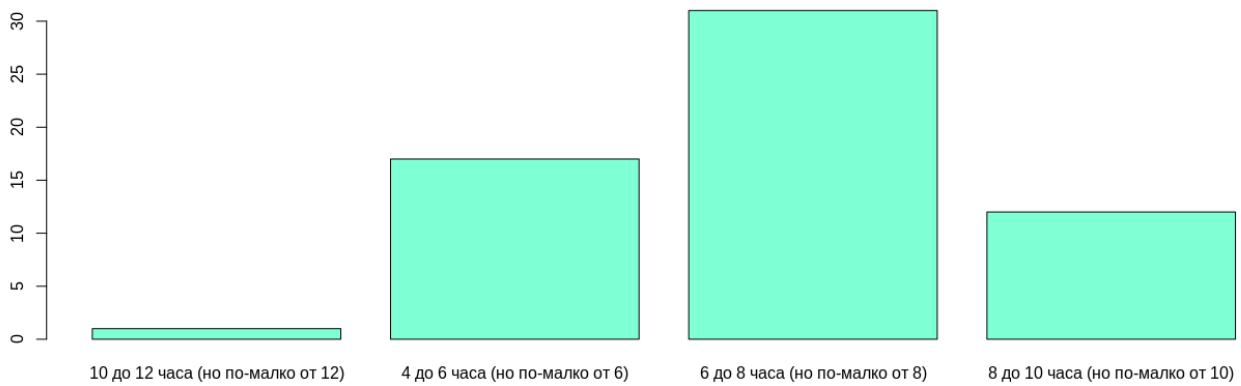
Полученият резултат за p-value е над 0.05, което показва, че величината е нормално разпределена.

- Прием на сънотворни



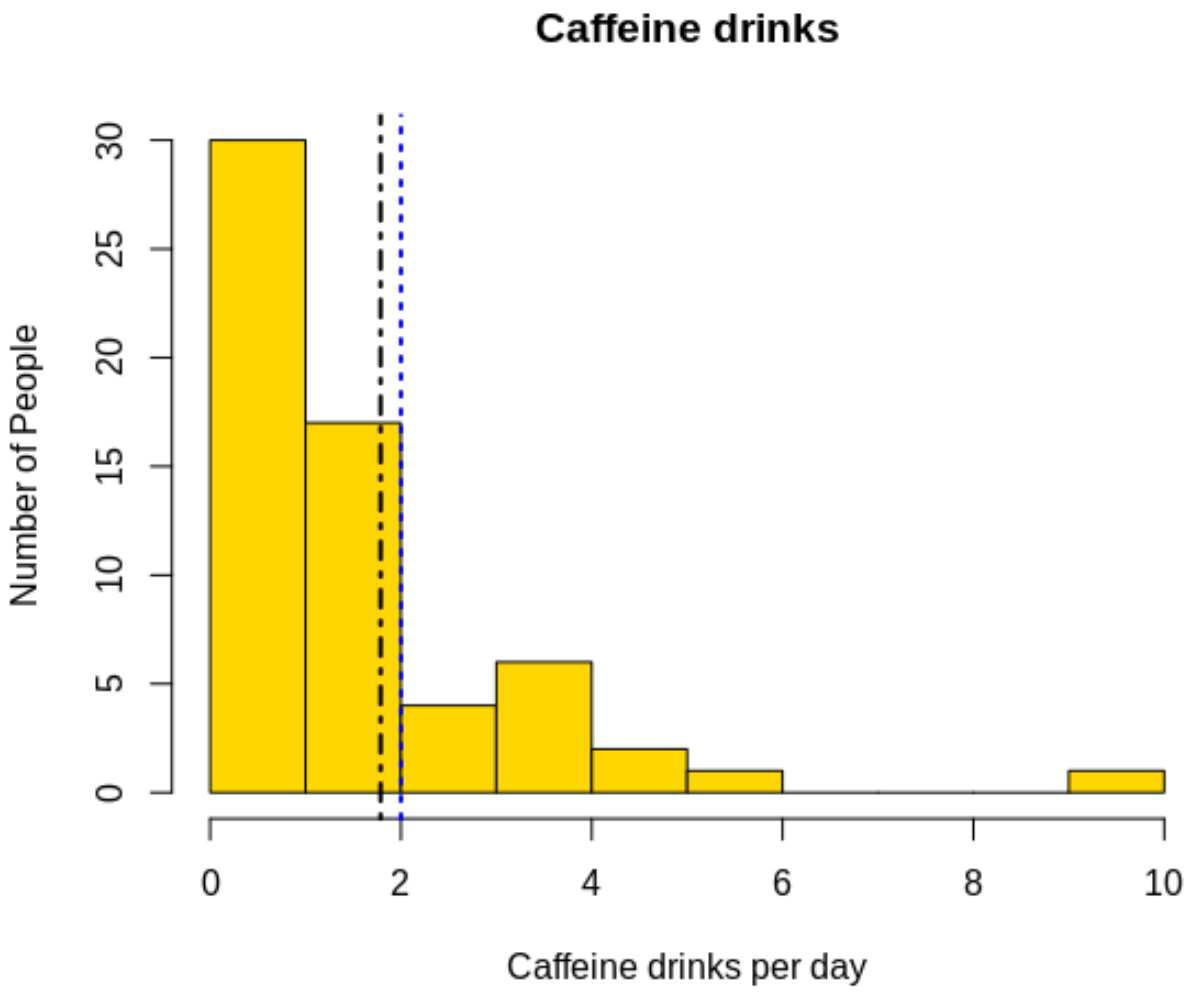

```
> #taking pills
> num_taking <- length(data$Take.pills[data$Take.pills==TRUE])
> num_not_taking <- length(data$Take.pills[data$Take.pills==FALSE])
> slices <- c(num_taking, num_not_taking)
> labels <- c("Yes", "No")
> pielabels <- sprintf("%s = %3.1f%s", labels, 100*slices/sum(slices), "%")
> pie(slices, labels = pielabels, main="Pie chart of sleeping pills intake")
>
```

- Часове сън



```
>
> #sleeping hours
> plot(data$Sleep.hours, col='aquamarine')
>
```

- Количество кофеинови напитки на ден

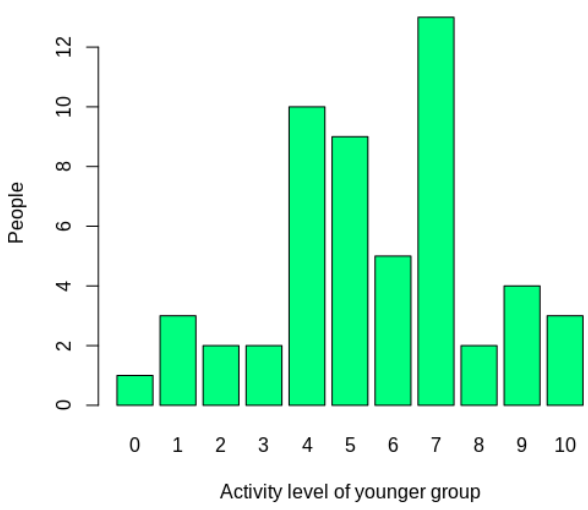


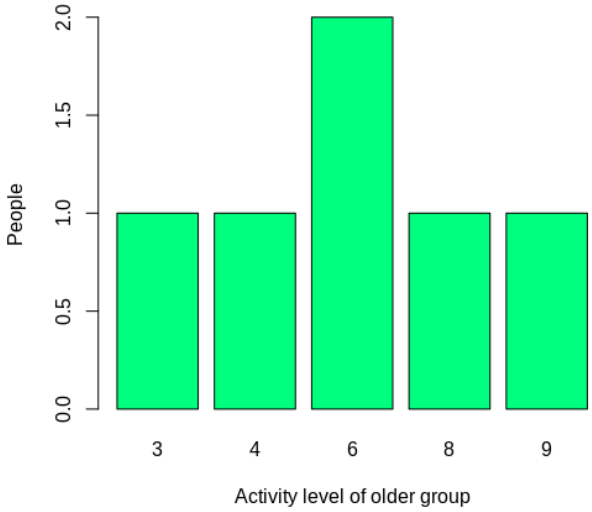
Като на горната графика със синя и черна линия са отбелязани съответно средната стойност и медианата.

```
>
> hist(data$Caffeine.drinks.per.day, main='Caffeine drinks', xlab='Caffeine drinks per day', ylab='Number of People', col='gold', breaks=10)
> abline(v = mean(data$Caffeine.drinks.per.day), lwd = 2, lty = 4)
> abline(v = median(data$Caffeine.drinks.per.day), lwd = 2, lty = 3, col = "blue")
>
```

3.2. Взаимодействие между променливите

- Нека анализираме физическата активност на хората в различни възрастови групи. За целта правим разбиване на анализираните обекти спрямо възрастта – под и над 30 години. Горе вече видяхме какво е разпределението на всички тях.





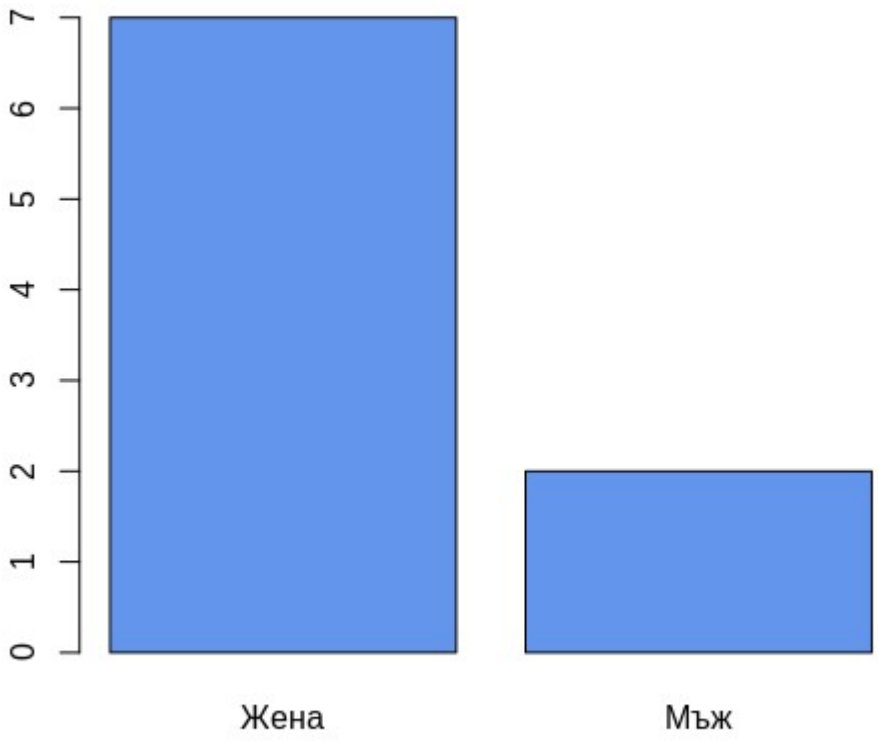

```
> #activity level of people under 30
> younger_group <- data[data$Age < 30, c("Activity.level")]
> barplot(table(younger_group), col='springgreen', xlab = "Activity level of younger group", ylab="People")
> print(mean(younger_group, na.rm = TRUE))
[1] 5.754717
>
> #activity level of people over 30
> older_group <- data[data$Age > 30, c("Activity.level")]
> barplot(table(older_group), col='springgreen',  xlab = "Activity level of older group", ylab="People")
> print(mean(older_group, na.rm = TRUE))
[1] 4.571429
>
```

Забелязва се спад на нивото на физическата активност в групата на хора над 30 години.

- Прием на сънотворни и пол

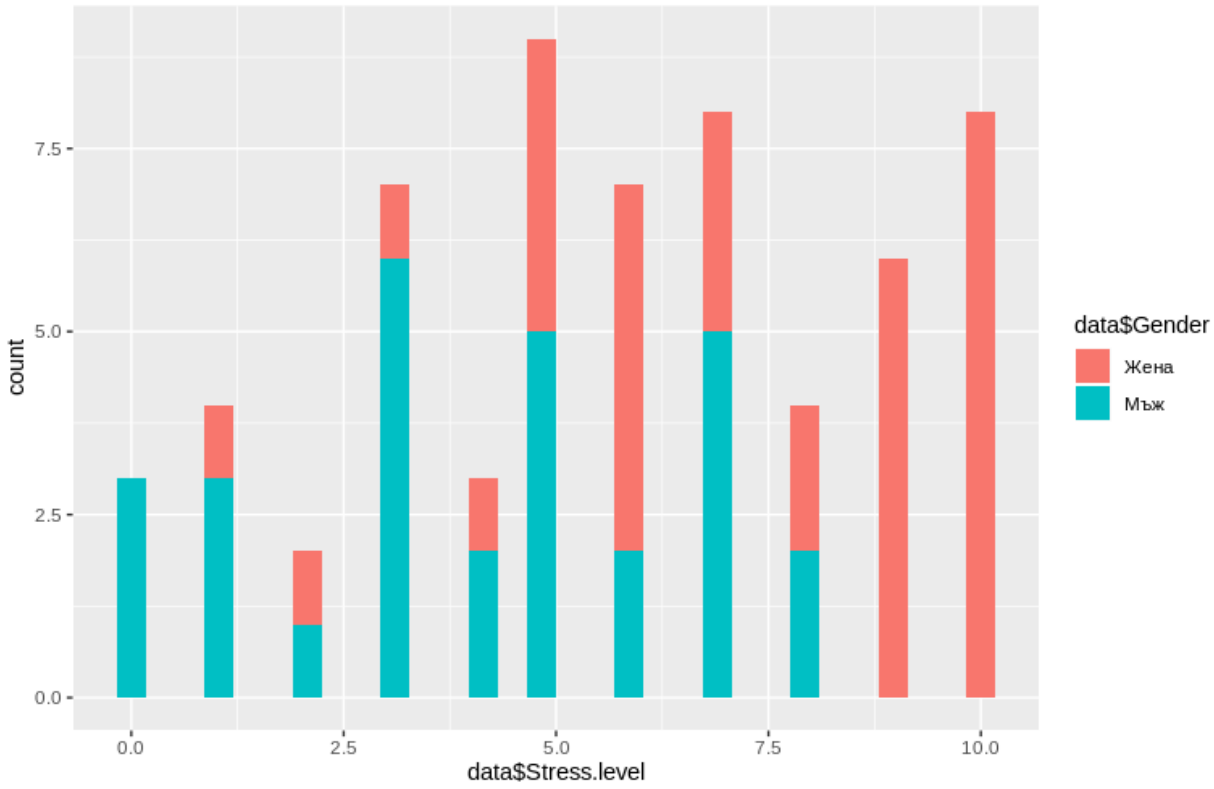
  Анализираме само хората, които са отбелязали, че приемат сънотворни. Разпределението за пола на тези хора е следното:

```
>
> barplot(table(data$Gender[data$Take.pills==TRUE]), col='cornflowerblue') # women are more likely to take pills
>
```



От тук е видимо, че жените са по-склонни да приемат подобен тип лекарства.

- Нива на стрес и пол

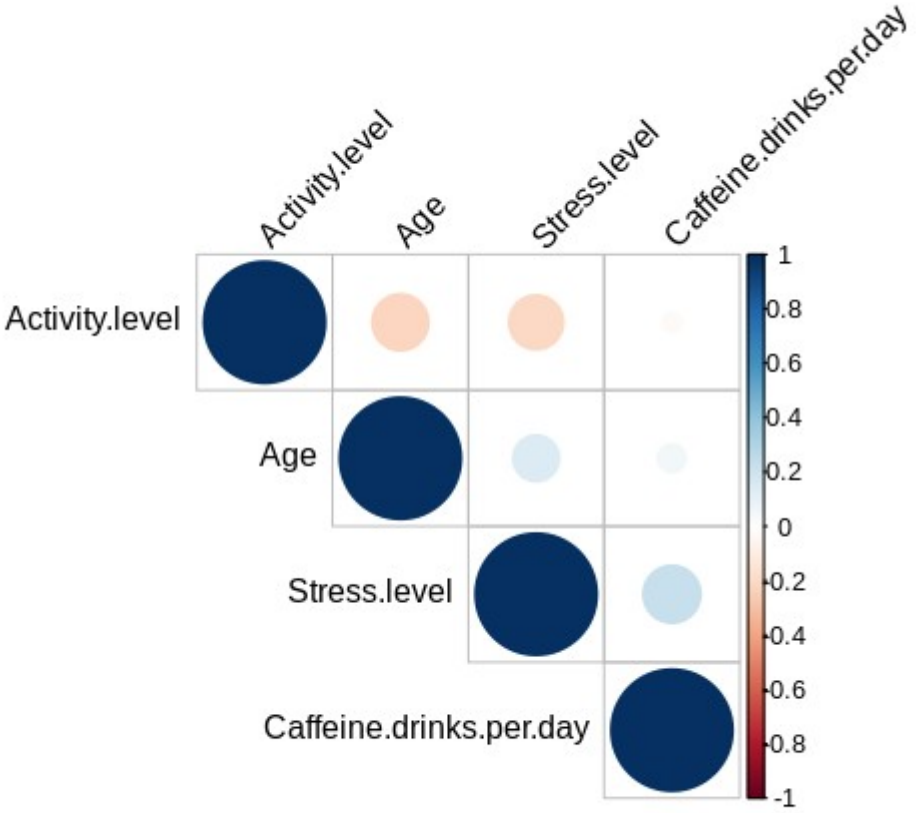



```
>
> ggplot(data,aes(x=data$Stress.level, fill=data$Gender)) + geom_histogram(bins=30) # women are with higher stress levels
>
```

Забелязва се тенденция жените да са с по-високи нива на стрес.

- Корелации между числовите променливи

```
> vector = data[, c(3,4,7, 8)]
> cor(vector)
                            Age Activity.level Stress.level Caffeine.drinks.per.day
Age                  1.00000000    -0.21992847    0.1487125             0.05499151
Activity.level      -0.21992847     1.00000000   -0.2050442            -0.02636357
Stress.level         0.14871245    -0.20504415    1.0000000             0.22762013
Caffeine.drinks.per.day  0.05499151    -0.02636357    0.2276201             1.00000000
> library(corrplot)
> corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
>
```



Очевидно корелацията между всеки две числови променливи е слаба.

Целият код за изграждане на горните графики(и още няколко, непоместени тук) се намира във файла project.R .
Допълнително е използвана библиотеката corrplot за по-добра визуализаия на корелационната графика.