Курсов проет по R

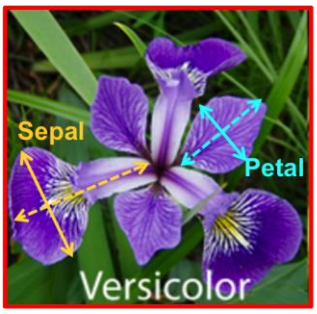
на Илкнур Бахтияр Мустафа, 3курс, спец. СИ, група 4
фн.62137

1. Въведение в “Изследователския анализ на данни”

Exploratory Data Analysis (EDA) представлява процесът на извършване на първоначални проучвания на данни, така че да се открият модели, трендове,аномалии, да се тестват хипотези и да се проверят предположения с помощта на обобщена статистика и графични изображения.

2. Въведение в Iris dataset

В текущия проект ще проучим един известен dataset - за цветето Iris с EDA методологията. Именно, причината да го изберем е, че позволява добро разпределяне на данните и следователно добро графично изобразяване на изследователския процес. Iris dataset съдържа 4 числови и 1 категорийна променливи. Съществуват 3 категории цветя (versicolor, virginica, setosa), като за всяка категория има по 50 наблюдения с измерване дължините и широчините на листата(sepal length/width, petal length/width).Именно равният брой изледвания за всяка категория показва доброто разпределяне на данните във всяка категория.



3. Общ вид на данните

Нека прочетем файлът с данните:

```
> data = read.csv("iris_data.csv")
```

Както описахме по-горе имаме 150 наблюдения с 5 променливи.

Проучваме цялостната структура на файла:

```
> ncol(data) # number of features/variables
[1] 5
```

Имаме 5 променливи.

```
> colnames(data) # name of the columns
[1] "sepal.length" "sepal.width"  "petal.length"
[4] "petal.width"  "iris"
```

Имената на променливите са показани отгоре.

```
> nrow(data) # number of observations
[1] 150
```

Има 150 наблюдения.

Нека разгледаме първите и последните 15 реда на файла:

```
> head(data, n= 15)
   sepal.length sepal.width petal.length
1           5.1         3.5          1.4
2           4.9         3.0          1.4
3           4.7         3.2          1.3
4           4.6         3.1          1.5
5           5.0         3.6          1.4
6           5.4         3.9          1.7
7           4.6         3.4          1.4
8           5.0         3.4          1.5
9           4.4         2.9          1.4
10          4.9         3.1          1.5
11          5.4         3.7          1.5
12          4.8         3.4          1.6
13          4.8         3.0          1.4
14          4.3         3.0          1.1
15          5.8         4.0          1.2
   petal.width        iris
1          0.2 Iris-setosa
2          0.2 Iris-setosa
3          0.2 Iris-setosa
4          0.2 Iris-setosa
5          0.2 Iris-setosa
6          0.4 Iris-setosa
7          0.3 Iris-setosa
8          0.2 Iris-setosa
9          0.2 Iris-setosa
10         0.1 Iris-setosa
11         0.2 Iris-setosa
12         0.2 Iris-setosa
13         0.1 Iris-setosa
14         0.1 Iris-setosa
15         0.2 Iris-setosa
```

```
> tail(data, n= 15)
    sepal.length sepal.width petal.length petal.width
136          7.7         3.0          6.1         2.3
137          6.3         3.4          5.6         2.4
138          6.4         3.1          5.5         1.8
139          6.0         3.0          4.8         1.8
140          6.9         3.1          5.4         2.1
141          6.7         3.1          5.6         2.4
142          6.9         3.1          5.1         2.3
143          5.8         2.7          5.1         1.9
144          6.8         3.2          5.9         2.3
145          6.7         3.3          5.7         2.5
146          6.7         3.0          5.2         2.3
147          6.3         2.5          5.0         1.9
148          6.5         3.0          5.2         2.0
149          6.2         3.4          5.4         2.3
150          5.9         3.0          5.1         1.8
              iris
136 Iris-virginica
137 Iris-virginica
138 Iris-virginica
139 Iris-virginica
140 Iris-virginica
141 Iris-virginica
142 Iris-virginica
143 Iris-virginica
144 Iris-virginica
145 Iris-virginica
146 Iris-virginica
147 Iris-virginica
148 Iris-virginica
149 Iris-virginica
150 Iris-virginica
```

4. Структура на променливите

```
> str(data)
'data.frame':   150 obs. of  5 variables:
 $ sepal.length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ sepal.width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ petal.length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ petal.width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ iris        : Factor w/ 3 levels "Iris-setosa",..: 1 1 1 1 1 1 1 1 1 1 ...
>
```

В заключение с тези резултати виждаме, че разполагаме с 4 числови - sepal.length, sepal.width, petal.length и petal.width, както и една категорийна променлива с 3 нива - iris.

5. Цялостна статистика на данните

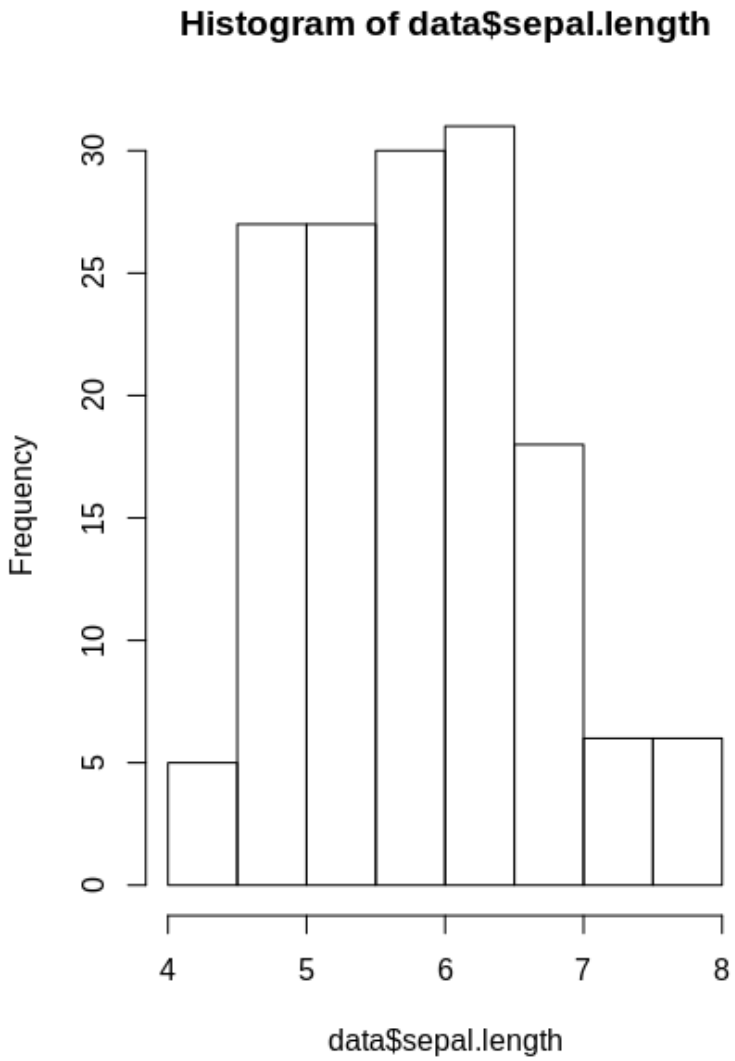
```
> summary(data)
  sepal.length    sepal.width     petal.length    petal.width                 iris
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Iris-setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Iris-versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   Iris-virginica :50
 Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

Изходът от горната команда показва минимумът, максимумът, средната стойност и модата за всяка една от четирите числови променливи, както първи и трети квартил.

6. Графично представяне на данните

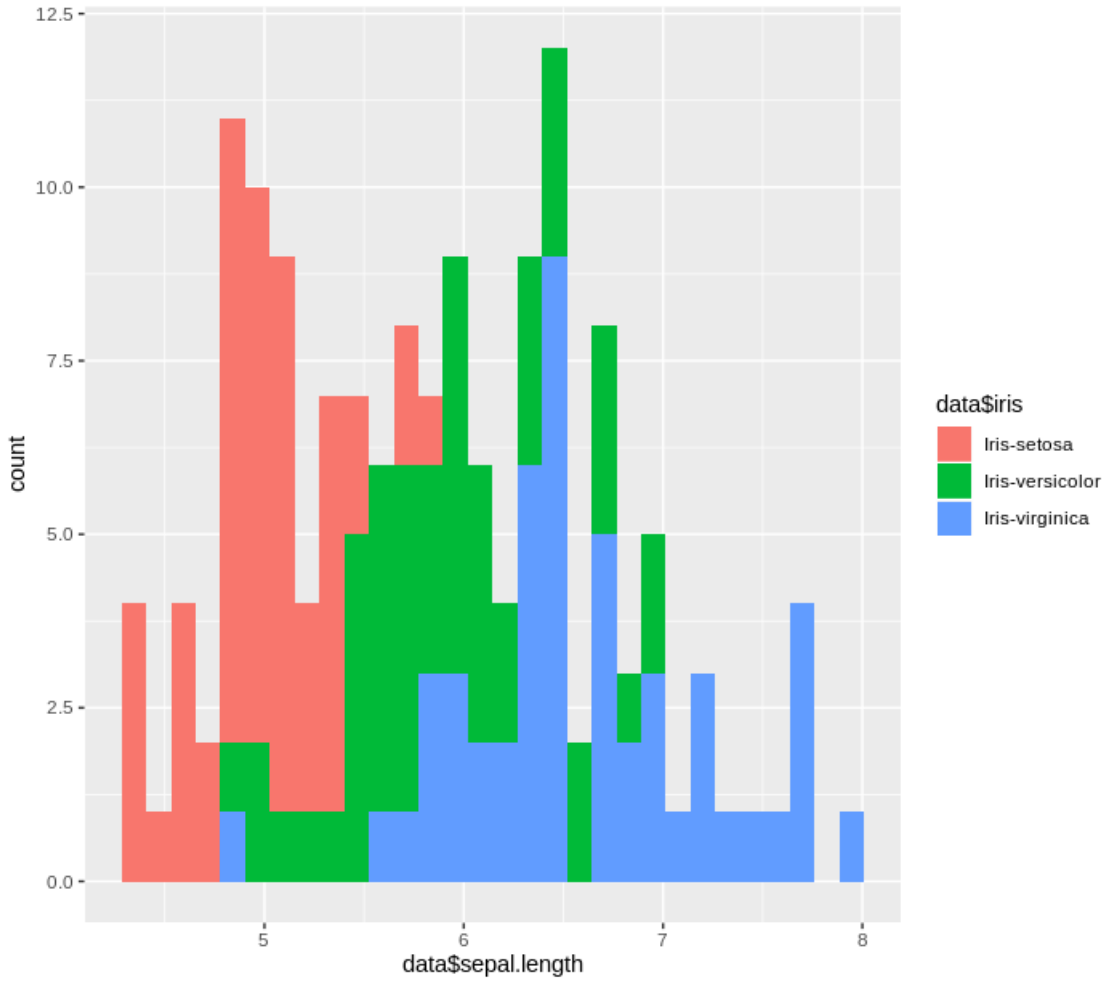
Ще разгледаме хистограма за всяка числова променлива, както и разпределението на всяка категория на цветето iris по разглежданата числова променлива:

```
> hist(data$sepal.length)
```



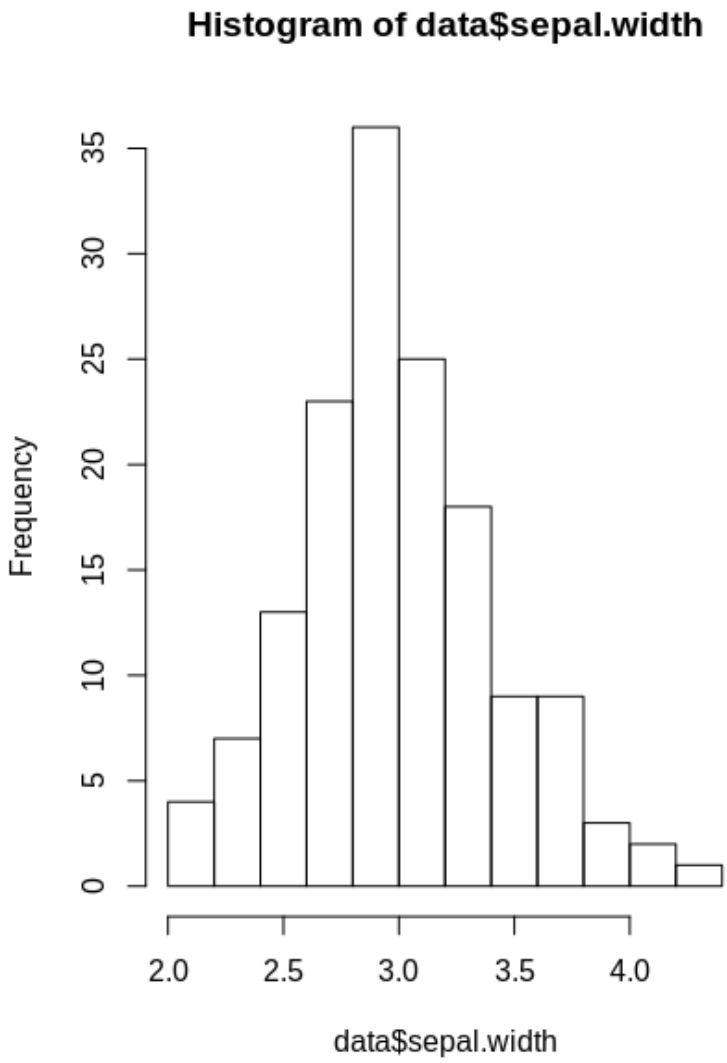
Коментар: От горните резултати имаме дистрибуция, която прилича на нормална , но с дълга дясна опашка(long right tail).

```
> ggplot(data,aes(x=data$sepal.length, fill=data$iris)) + geom_histogram()
```



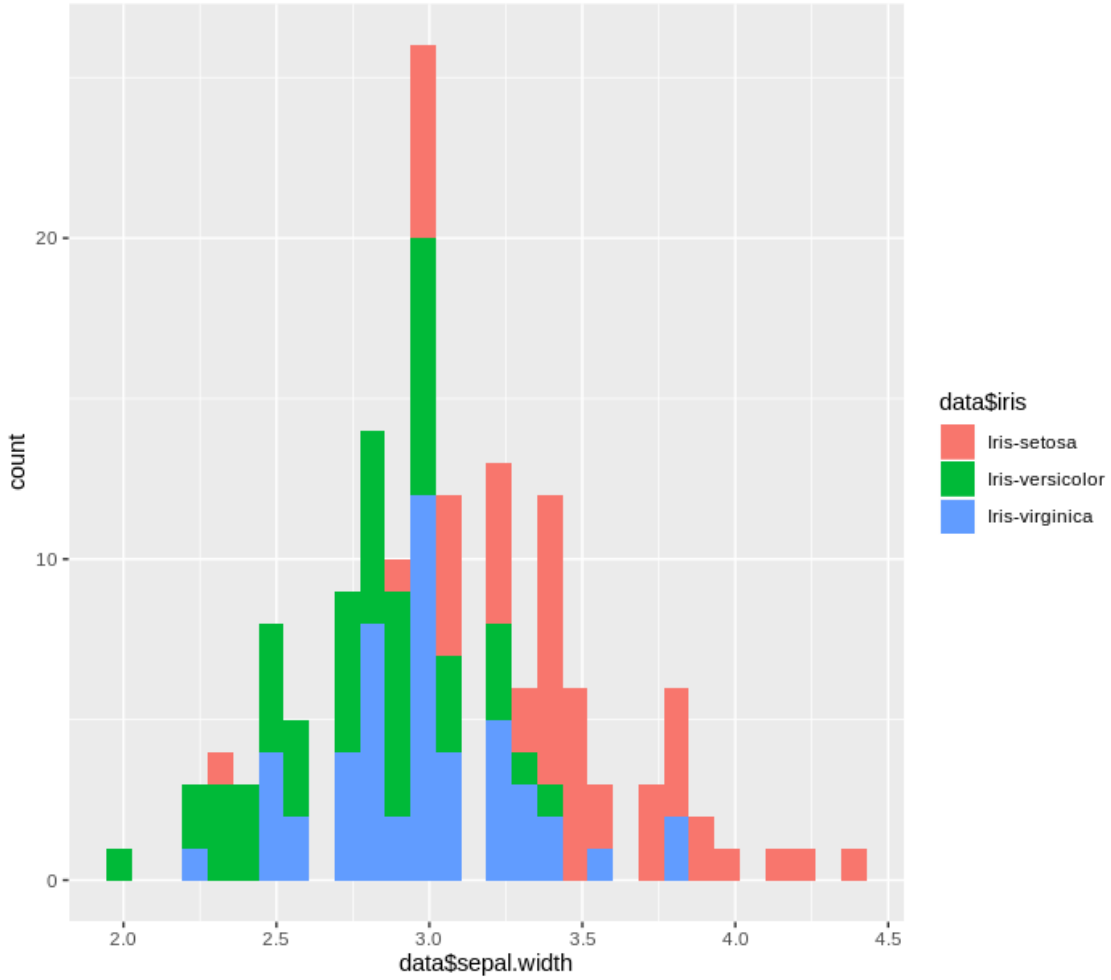
Коментар: Sepal.length не показва добро разпределение между категориите и следователно не се очаква добра корелация между двете променливи.

> hist(data$sepal.width)



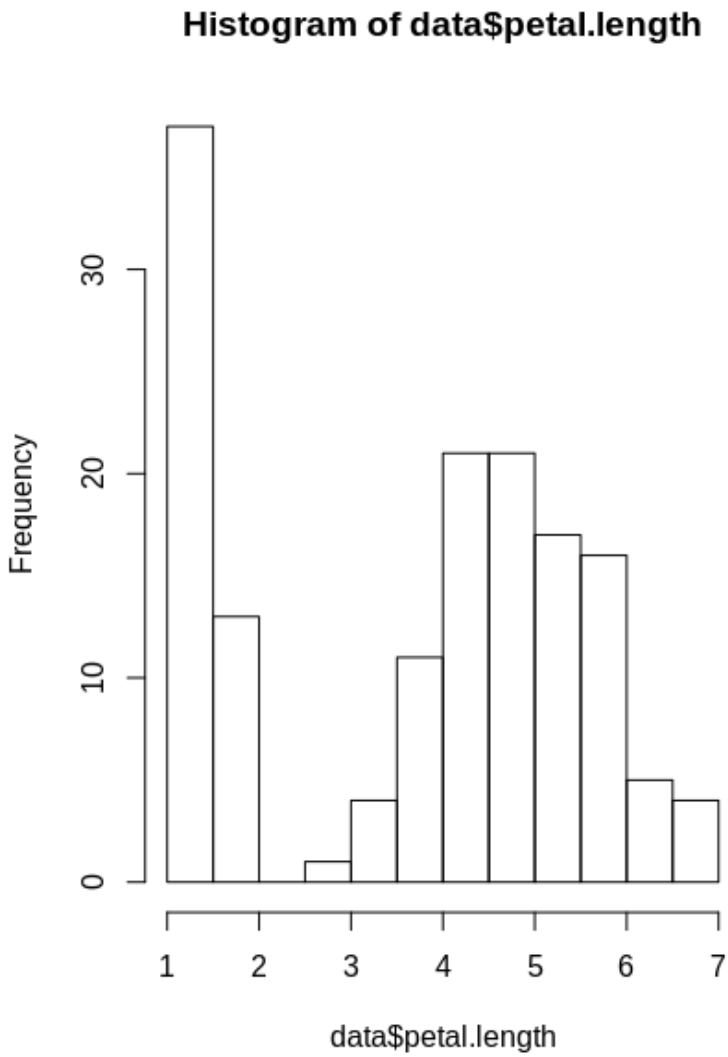
Коментар: Изглежда като нормална дистрибуция.

```
> ggplot(data,aes(x=data$sepal.width, fill=data$iris)) + geom_histogram()
```



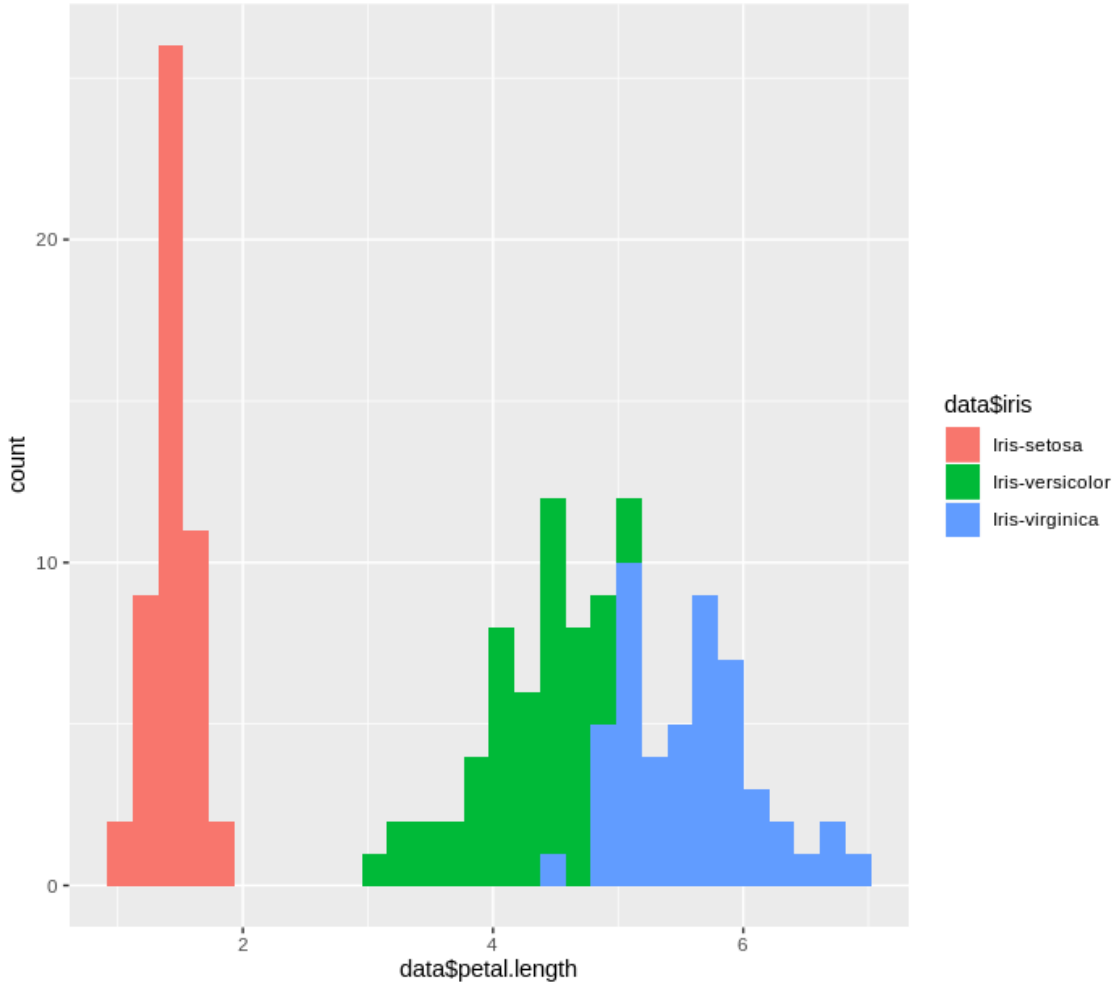
Коментар: Sepal.width също не показва добро разпределение между категориите - припокриващи се дистрибуции, следователно не се очаква добра корелация между двете променливи.

```
> hist(data$petal.length)
```



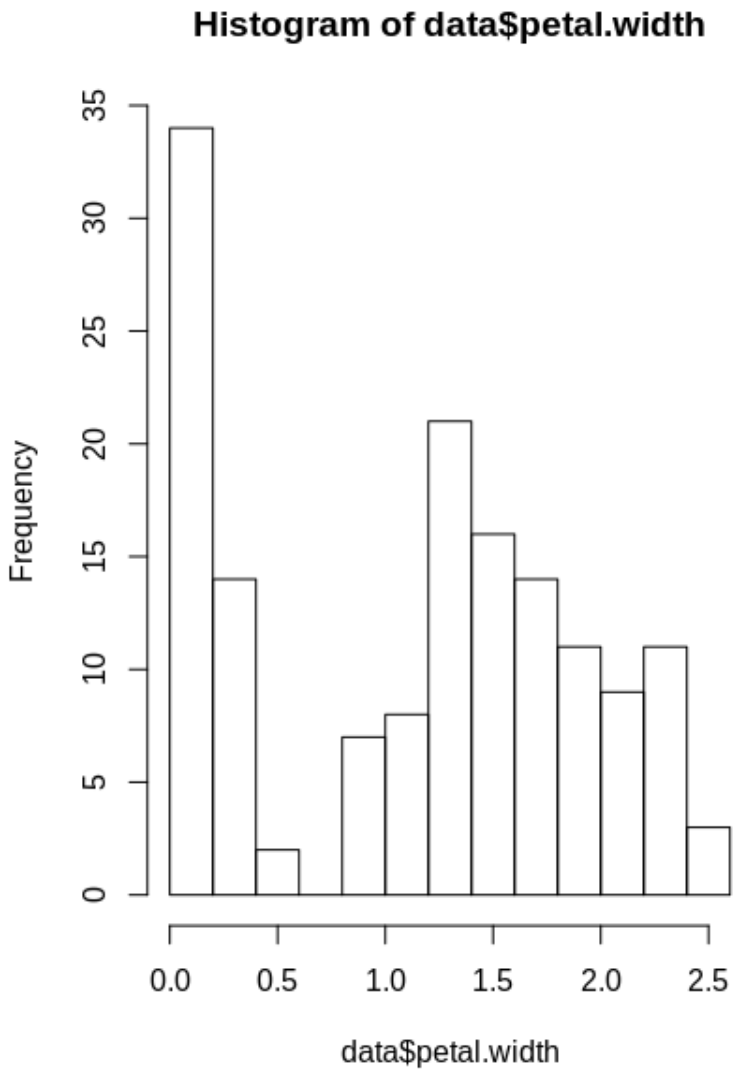
Коментар: Не е нормално разпределение, има два пика (бимодално разпределение)- единия в 1 и другия в 4.

```
> ggplot(data,aes(x=data$petal.length, fill=data$iris)) + geom_histogram()
```



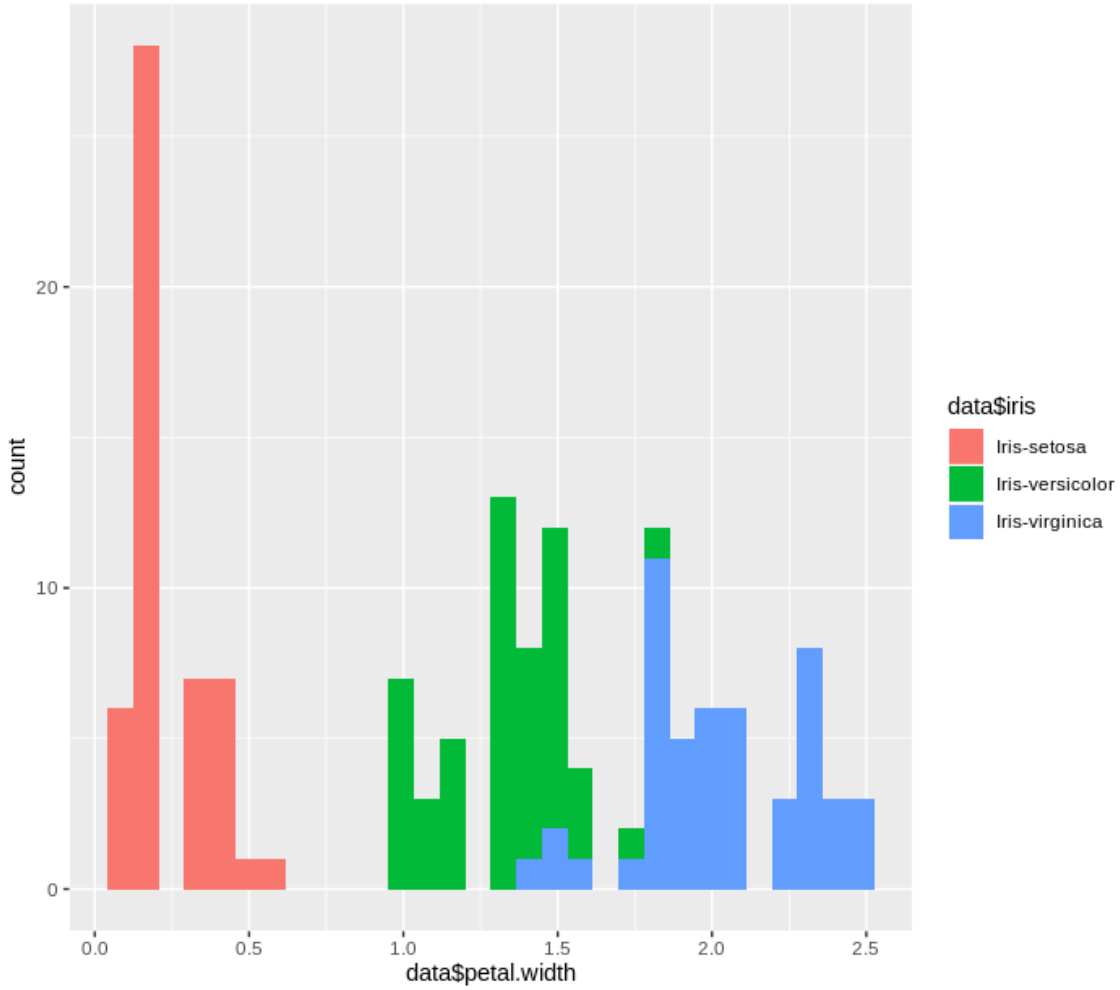
Коментар: Petal.length показва добро разпределение между категориите - не припокриващи се дистрибуции, следователно се очаква добра корелация между двете променливи.

```
> hist(data$petal.width)
```



Коментар: Подобно на горния резултат (бимодално разпределение) - пикове около 0.3 и 1.2 .

```
> ggplot(data,aes(x=data$petal.width, fill=data$iris)) + geom_histogram()
```

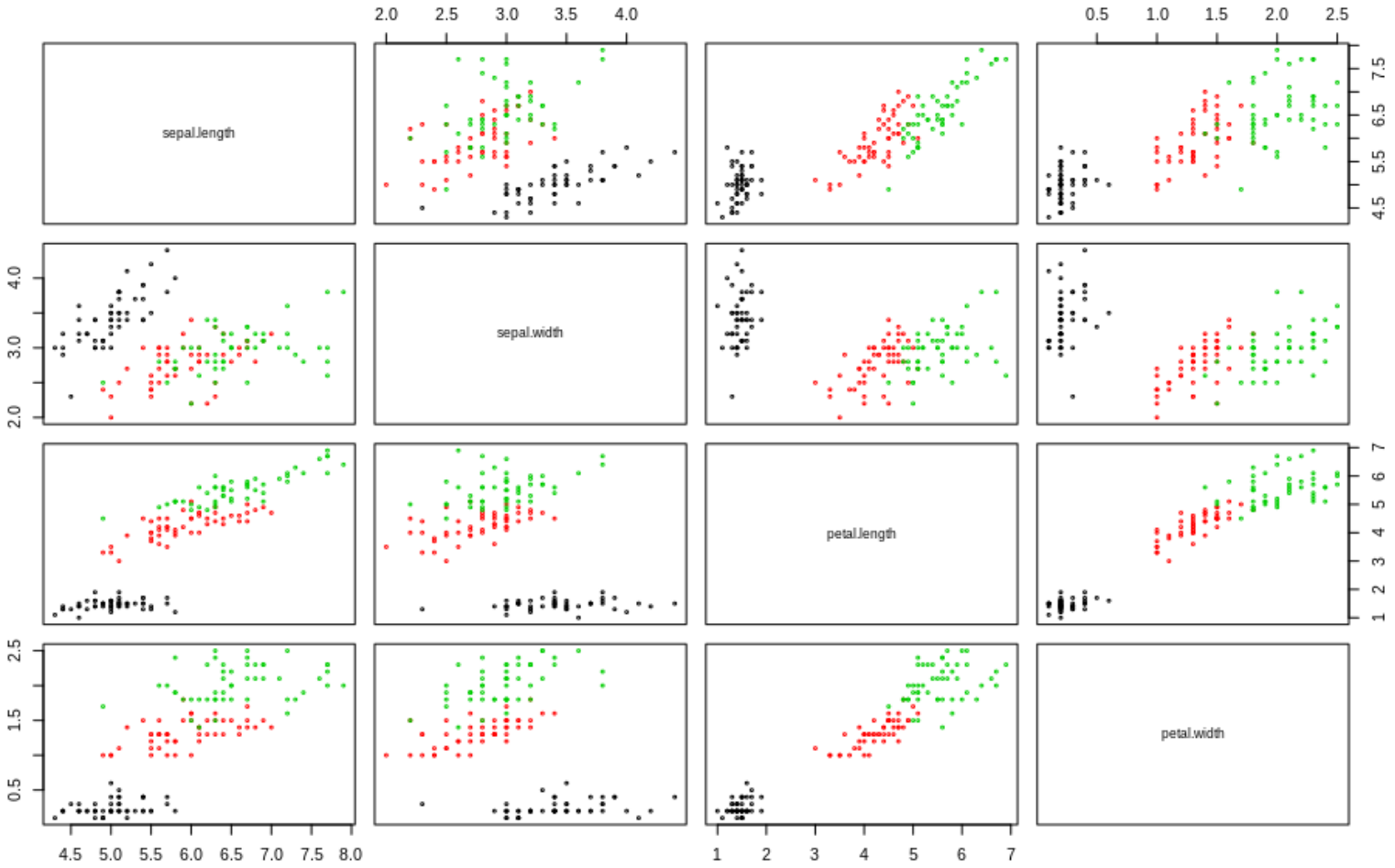


Коментар: Petal.width показва добро разпределение между категориите - не припокриващи се дистрибуции, следователно се очаква добра корелация между двете променлив

7. Сравнение по 2 числови и една категорийна променлива

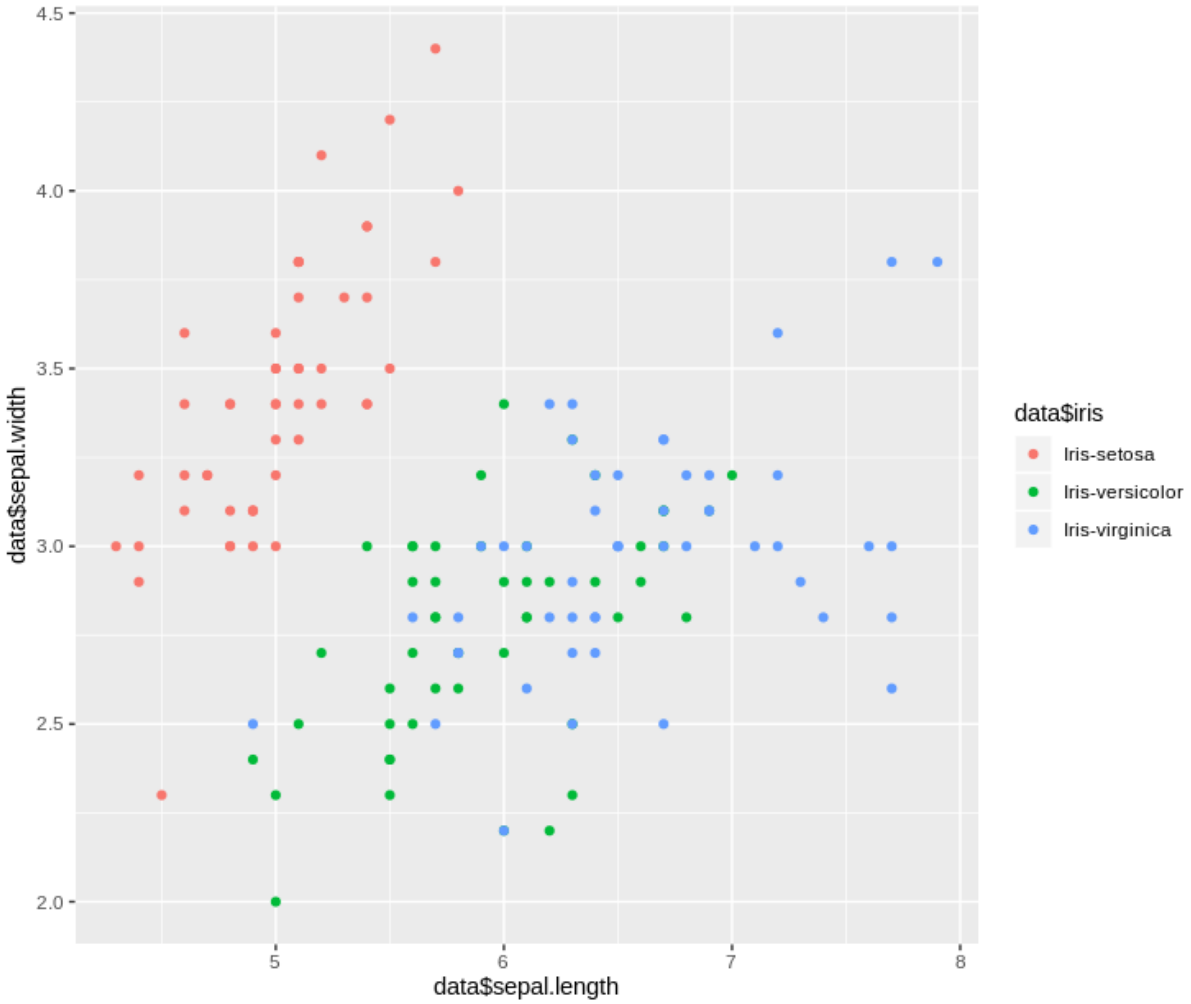
Визуализация на разпределението по всяка двойка числови променливи и взаимодействието на видовете категории.

```
> pairs(data[,1:4], pch = 21,  cex = 0.4, col = data$iris)
```



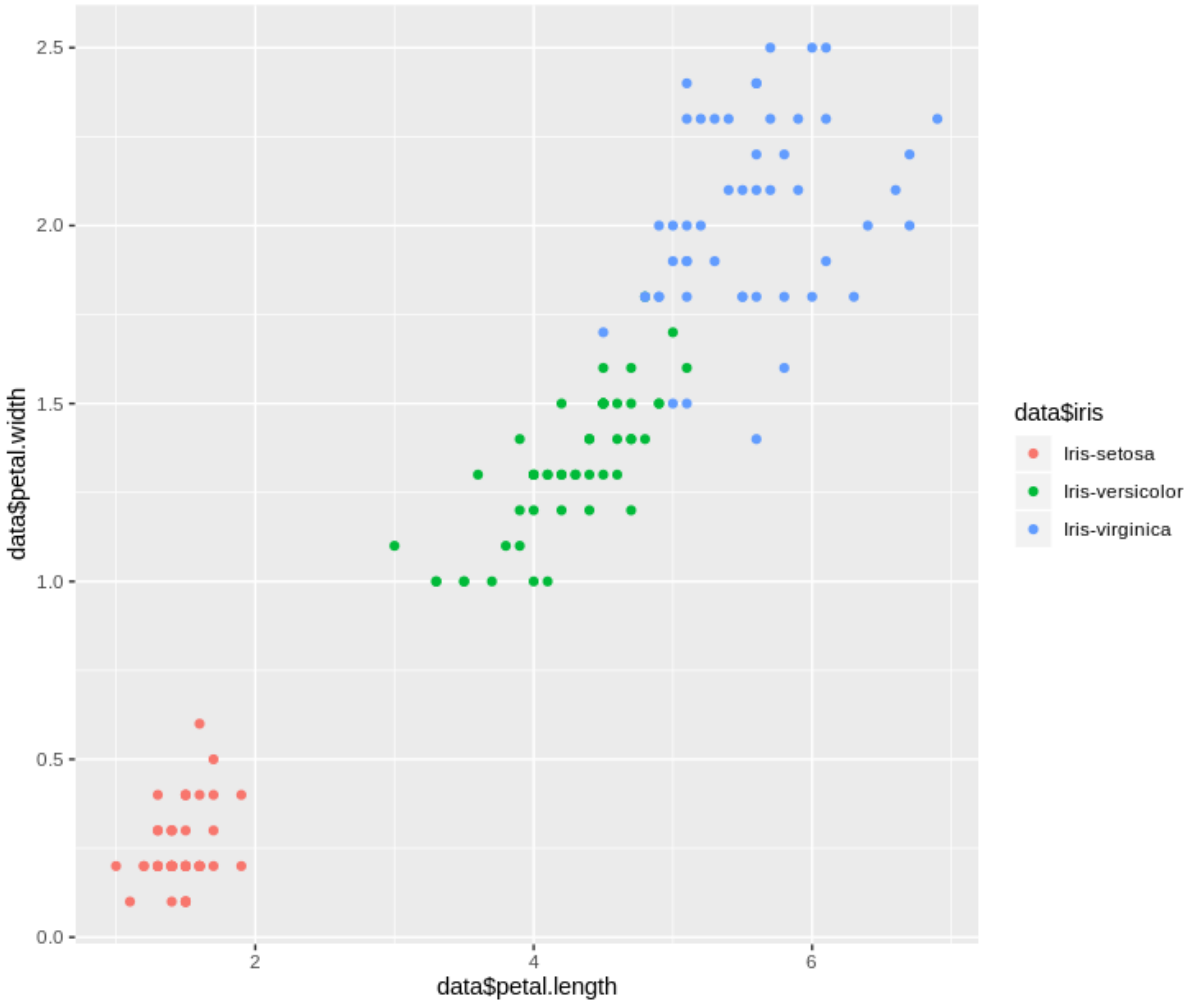
Ще разгледаме две от тези визуализации отблизо.

```
> ggplot(data,aes(data$sepal.length, data$sepal.width,colour = data$iris)) + geom_point()
```



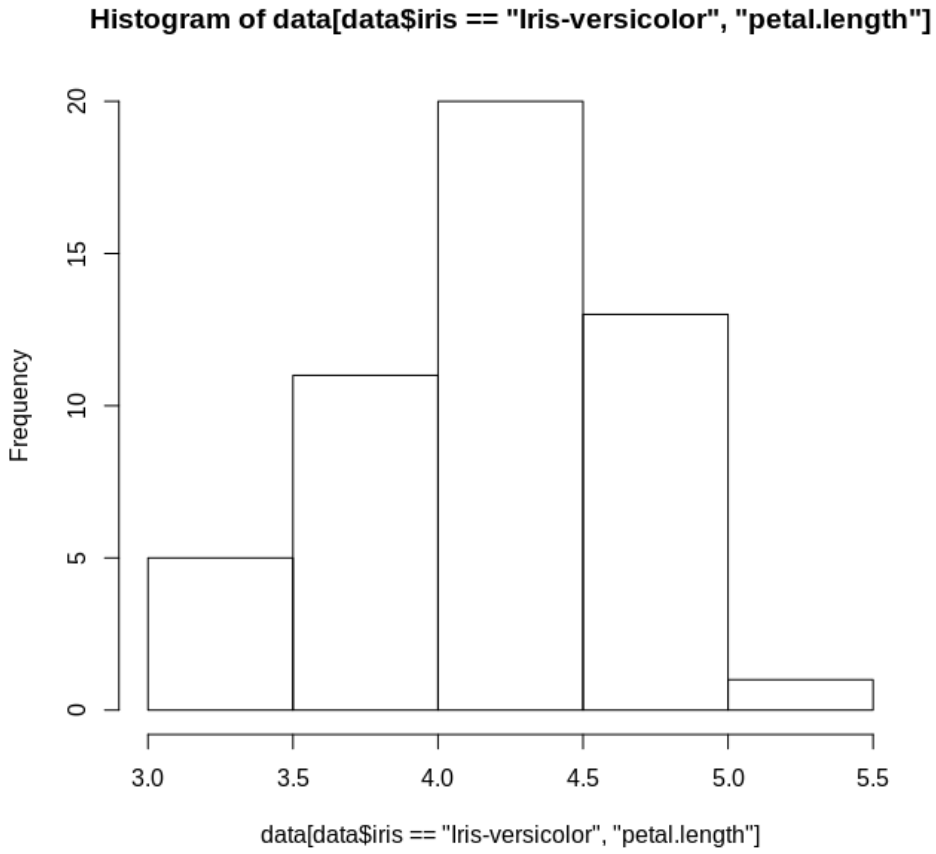
Коментар: Ако sepal.width е между 2.7 и 4.5 и sepal.length е между 4 и 6, тогава можем да категоризираме, че цветето е iris.setosa. Iris.versicolor и iris.virginica не са добре дефинирани и не можем да предскажем. Избираме Petal Length и Pelat Width, защото от предишния анализ се показва добра предварителна неприпокриваща се дистрибуция.

```
> ggplot(data,aes(data$petal.length, data$petal.width,colour = data$iris)) + geom_point()
```



Коментар: Получаваме добро разпределение за всяка категория, съответно можем да постигнем категоризиране по дадени petal.length и petal.width за всеки вид от цветето. Очевидно е, че при petal.length под 2 и petal.width под 0.5 можем да го категоризираме като iris.setosa, като iris.versicolor - за petal.length между 3 и 5, petal.width между 1.0 и ~ 1.75 и като iris.virginica при petal.length между 5 и 7 същевременно petal.width между 1.75 и 2.5.

8. Дистрибуция на сегментирана част от данните



Коментар: Разпределението е нормално.

9. Цялостна статистика по категории

Нека разгледаме средната стойност и вариацията на всяка числова променлива по дадена категория:

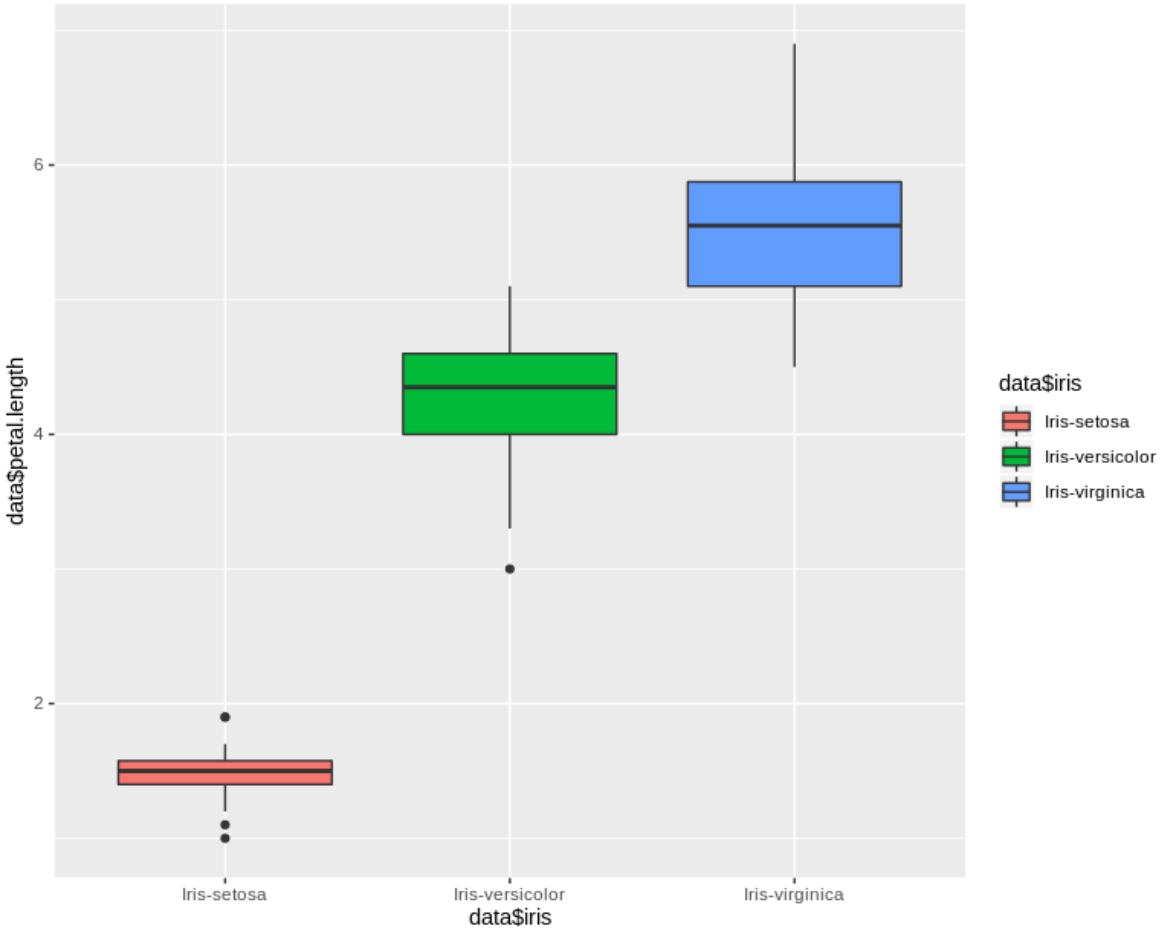
```
> aggregate(data[,1:4], list(data$iris), mean)
         Group.1 sepal.length sepal.width petal.length petal.width
1     Iris-setosa        5.006       3.418        1.464       0.244
2 Iris-versicolor        5.936       2.770        4.260       1.326
3  Iris-virginica        6.588       2.974        5.552       2.026
> aggregate(data[,1:4], list(data$iris), var)
         Group.1 sepal.length sepal.width petal.length petal.width
1     Iris-setosa    0.1242490  0.14517959   0.03010612  0.01149388
2 Iris-versicolor    0.2664327  0.09846939   0.22081633  0.03910612
3  Iris-virginica    0.4043429  0.10400408   0.30458776  0.07543265
>
```

Коментар: Забелязваме че, средната стойност на sepal.length и sepal.width за iris.versicolor и iris.virginica за близки, което не позволява лесно категоризиране на цветето както видяхме и по-горе. Докато средната стойност на petal.length и petal.width са различни/далечни, съответно постигаме лесно разпределяне и в 3-те класа.

10. Boxplot на данните

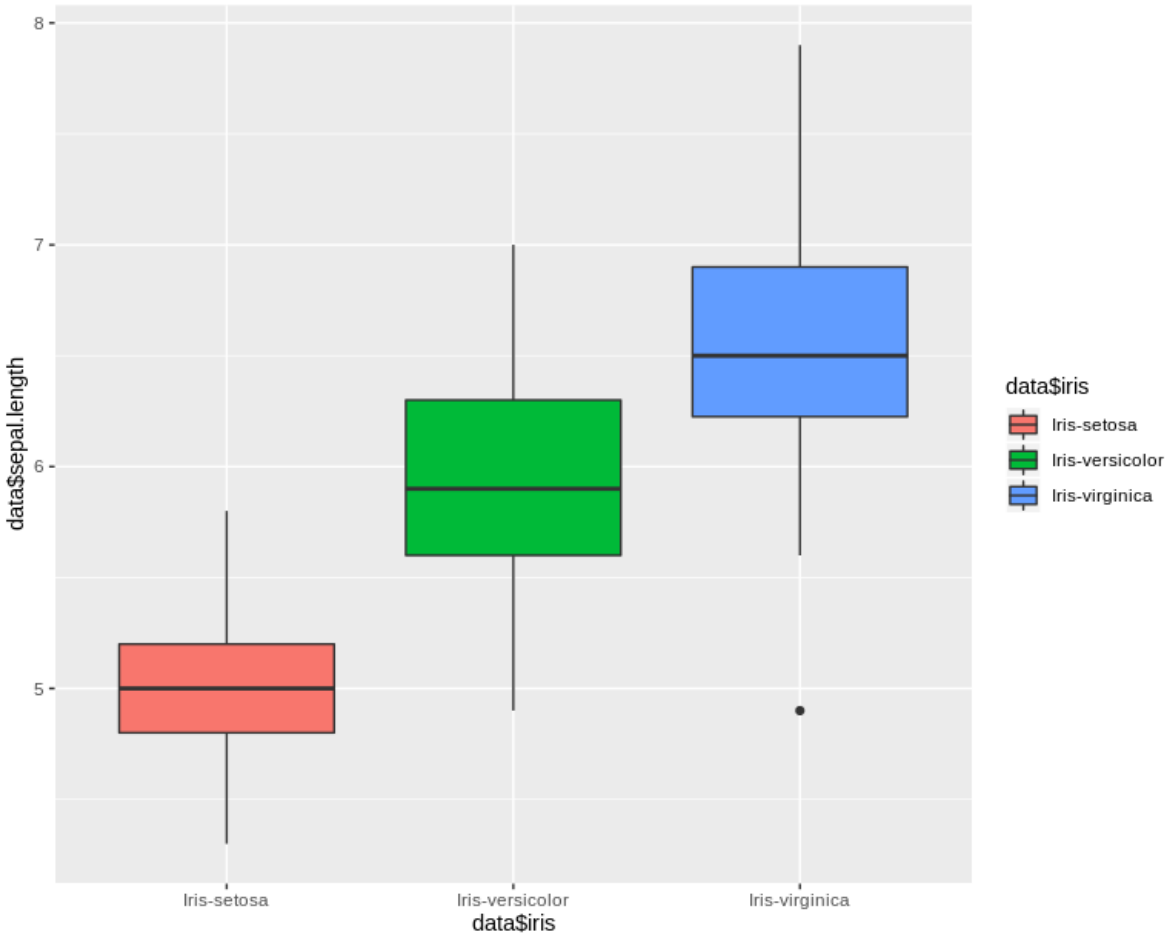
Ще покажем корелациите между 1 числова и 1 категорийна променливи .

```
> ggplot(data, aes(x =data$iris, y = data$petal.length, fill = data$iris)) + geom_boxplot()
```



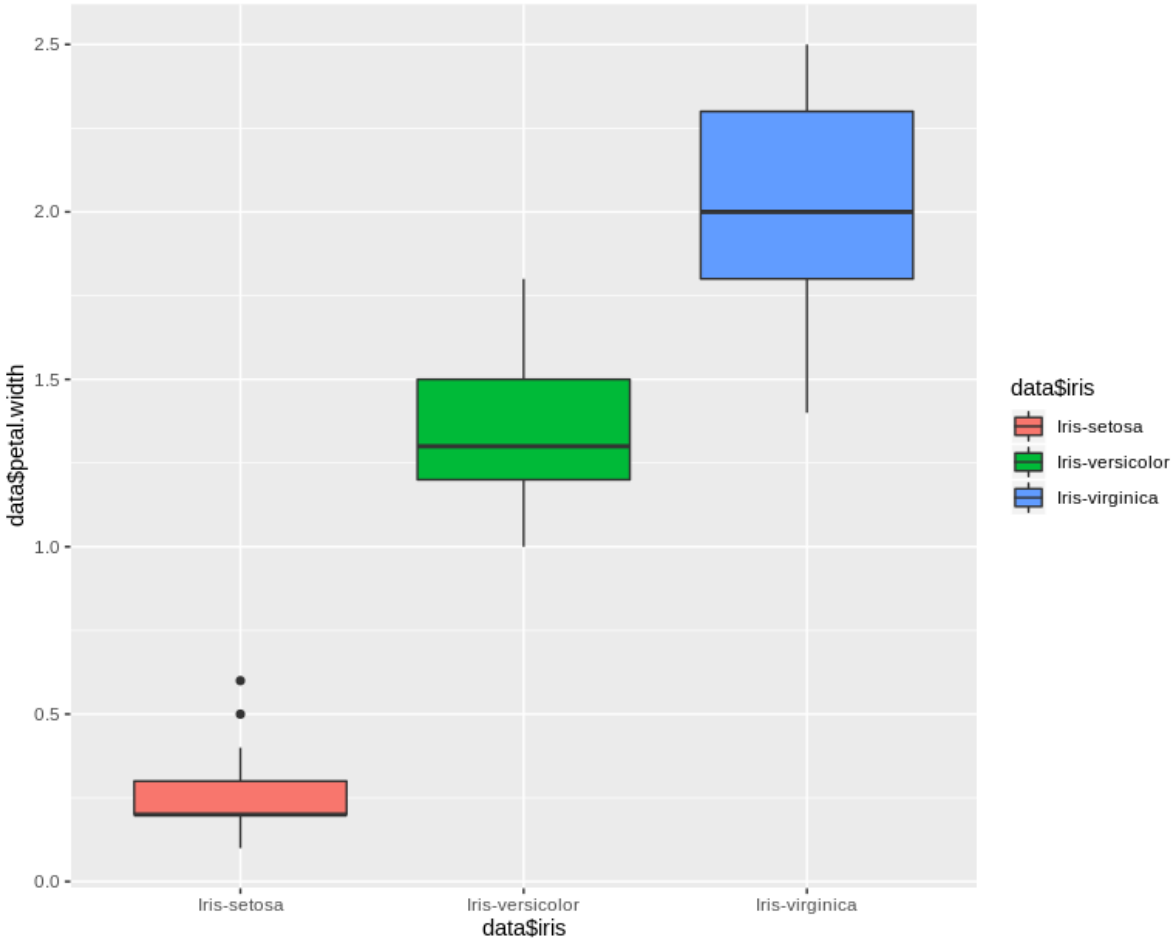
Коментар: Както се вижда от предният анализ, Petal Length има различни средни стойности за трите категории, както и вариацията за трите не се припокриват.

```
> ggplot(data, aes(x =data$iris, y = data$sepal.length, fill = data$iris)) + geom_boxplot()
```



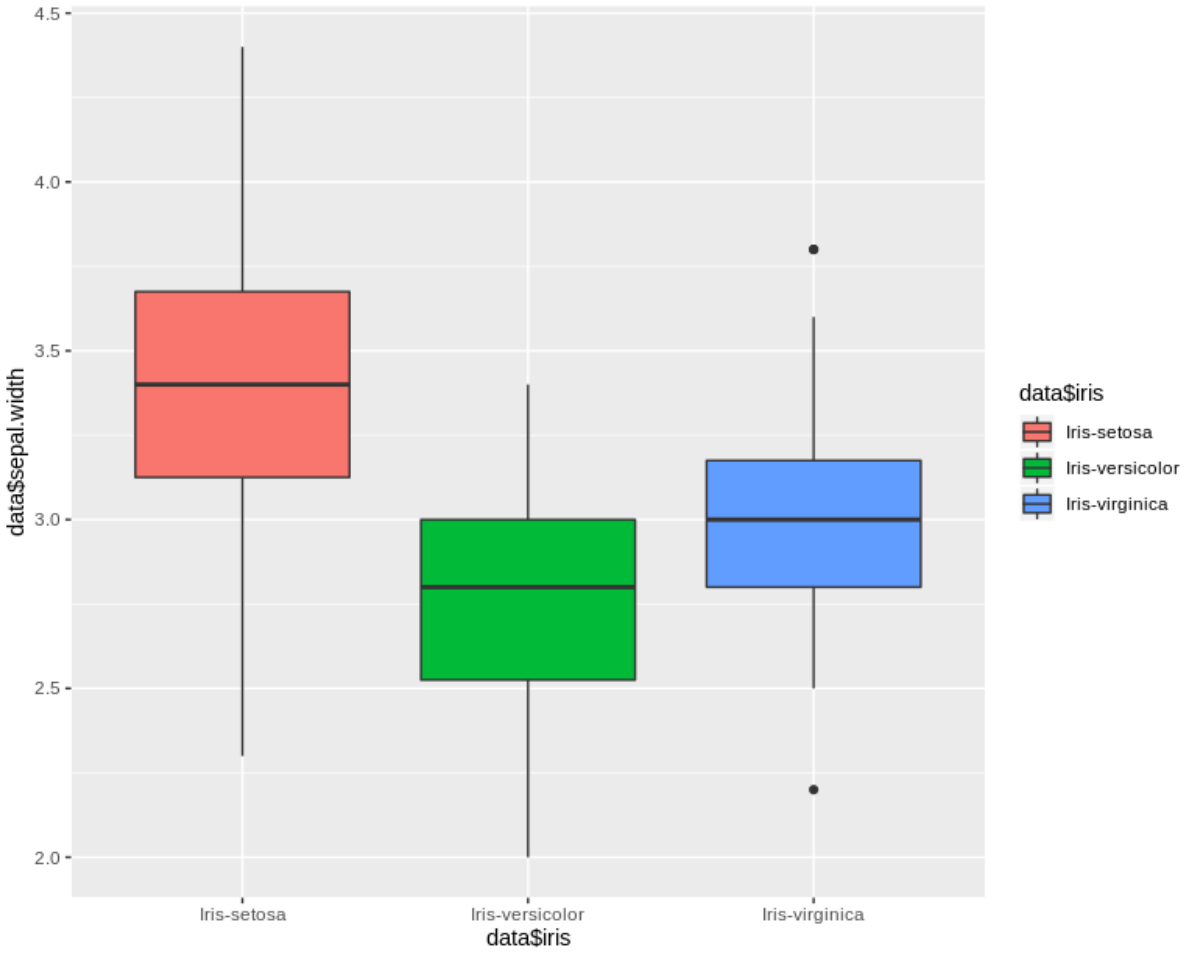
Коментар: Както се вижда от предният анализ, Sepal Length има различни средни стойности за трите категории, но вариациите се припокриват.

```
> ggplot(data, aes(x =data$iris, y = data$petal.width, fill = data$iris)) + geom_boxplot()
```



Коментар: Както се вижда от предният анализ, Petal Width има различни средни стойности за трите категории, както и вариацията за трите не се припокрива.

```
> ggplot(data, aes(x =data$iris, y = data$sepal.width, fill = data$iris)) + geom_boxplot()
```

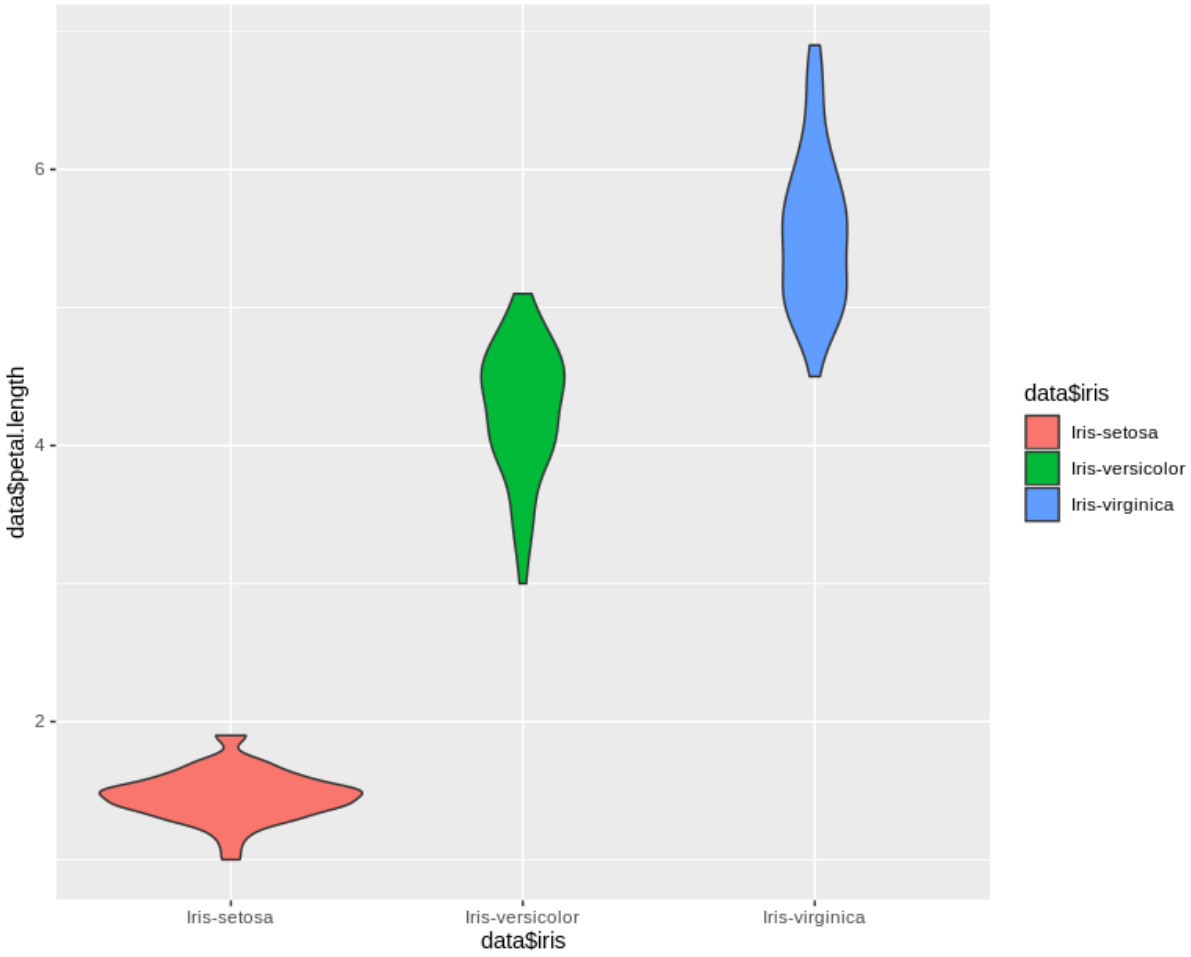


Коментар: Както се вижда от предният анализ, Sepal Width приближени средни стойности за трите категории, както и вариациите се припокриват.

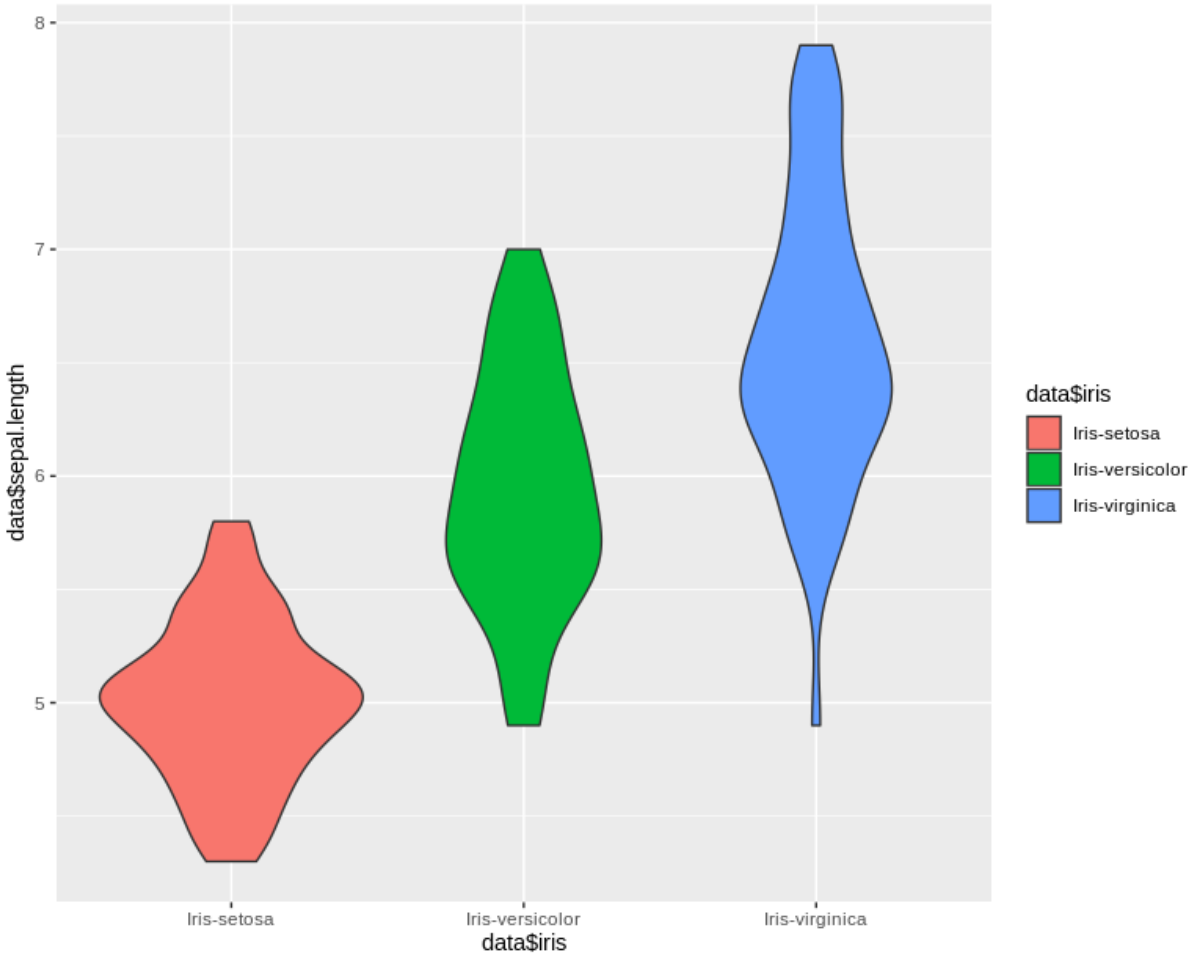
11. Violin plot на данните

Violin plots са подобни на boxplots, освен че показват също плътността на вероятността на ядрото на данните при различни стойности.

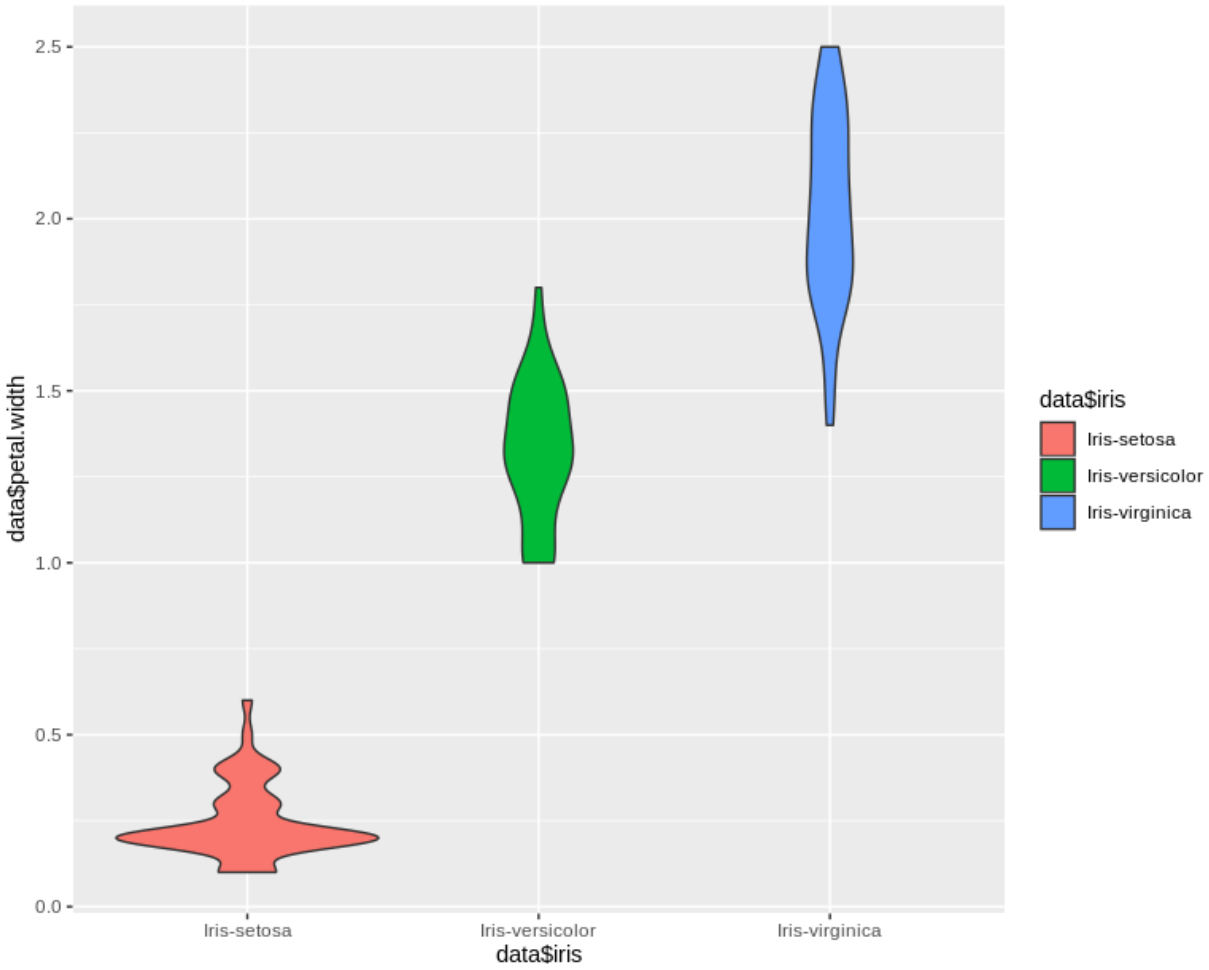
```
> ggplot(data, aes(x=data$iris, y = data$petal.length, fill = data$iris)) + geom_violin()
```



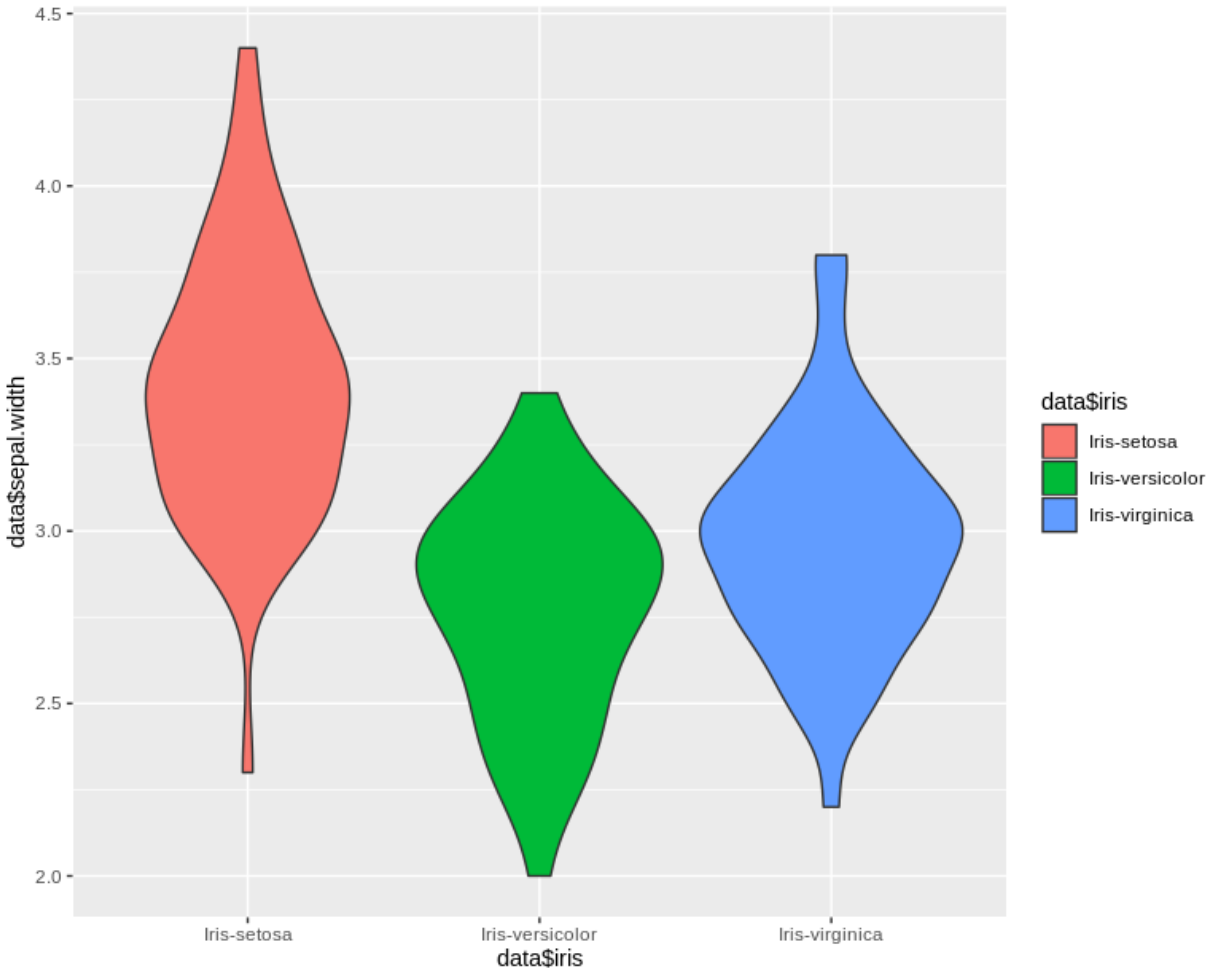

```
> ggplot(data, aes(x =data$iris, y = data$sepal.length, fill = data$iris)) + geom_violin()
```




```
> ggplot(data, aes(x =data$iris, y = data$petal.width, fill = data$iris)) + geom_violin()
```

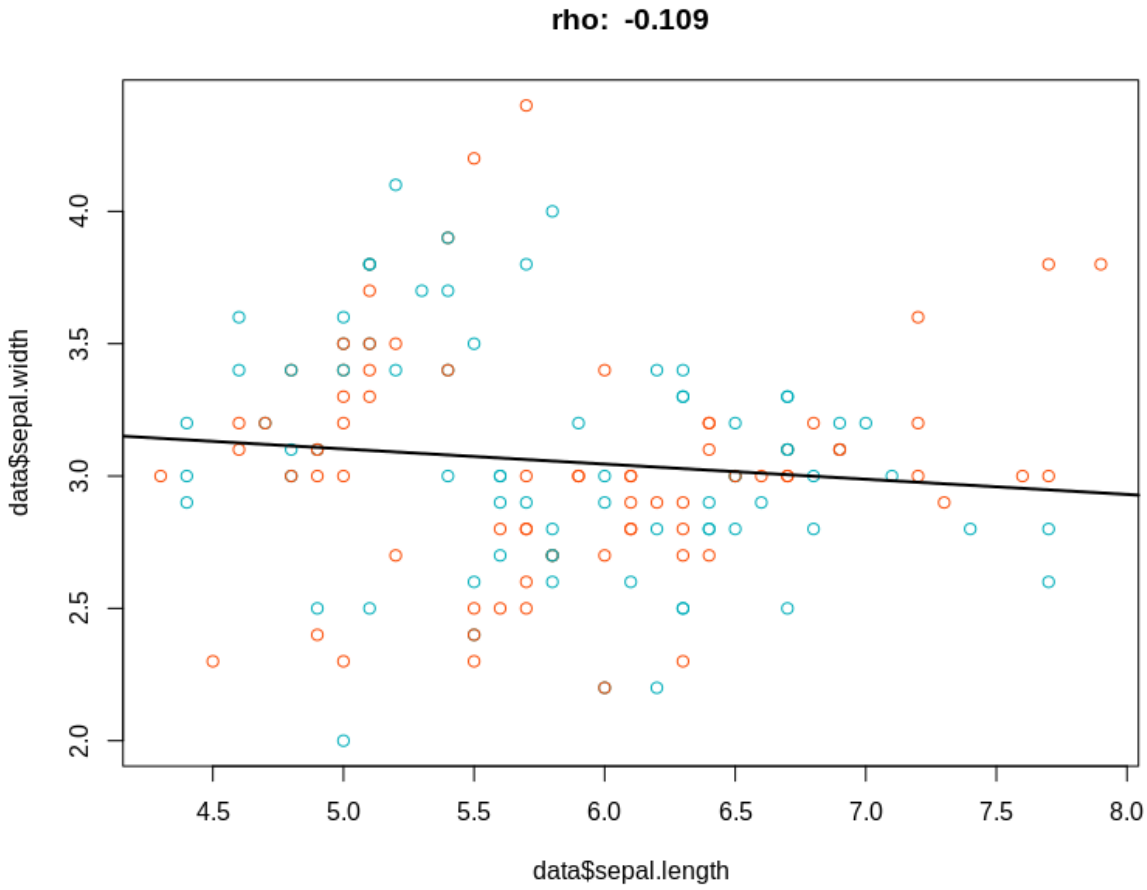



```
> ggplot(data, aes(x =data$iris, y = data$sepal.width, fill = data$iris)) + geom_violin()
```



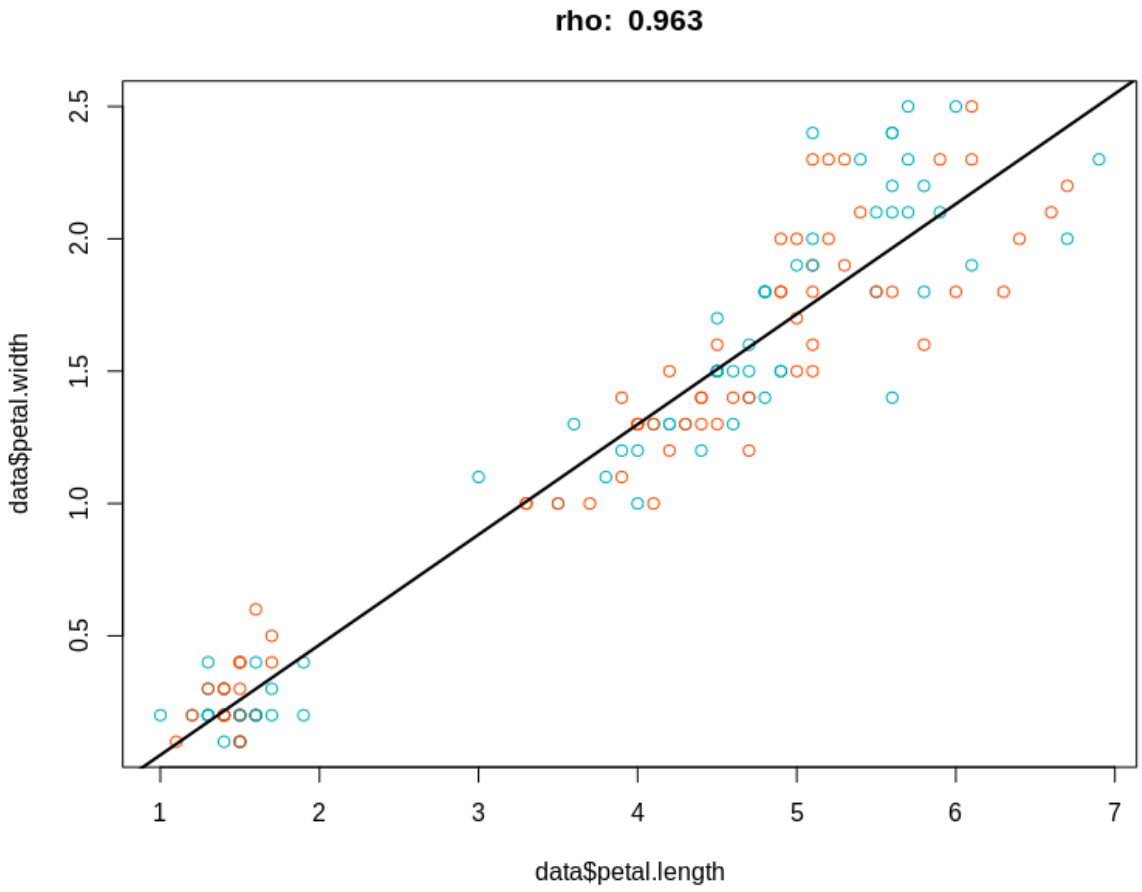
12. Корелационен анализ на числовите променливи

```
> rho1 <- round(cor(data$sepal.length,data$sepal.width),3)

>plot(data$sepal.length,data$sepal.width,col = cols, main=paste("rho: ", rho1))

> abline(lm(data$sepal.width ~ data$sepal.length), lwd=2)
```



Коментар: Както се вижда от горното представяне така и от коефициента на релация, съществува слаба корелация между sepal.length и sepal.width.

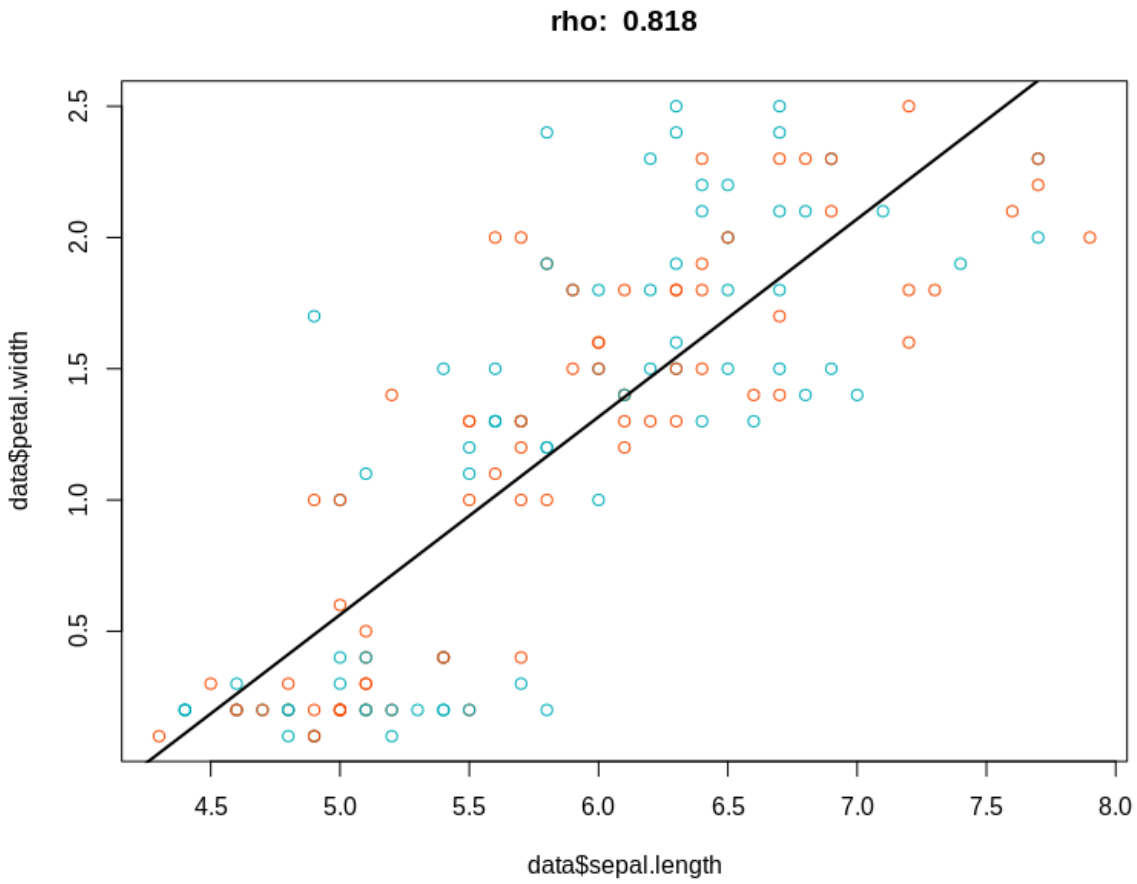
```
> rho2 <- round(cor(data$petal.length,data$petal.width),3)
> plot(data$petal.length,data$petal.width,col = cols, main=paste("rho: ", rho2))
> abline(lm(data$petal.width ~ data$petal.length), lwd=2)
```



Коментар: Коефициента на корелация е близък до 1, което ни приближава към детерминистична връзка между petal.length и petal.width,но в случая не е детерминистична, а по-скоро можем да твърдим че съществува много силна корелация между двете числови променливи.

```
> rho3 <- round(cor(data$sepal.length,data$petal.width),3)
> plot(data$sepal.length,data$petal.width,col = cols, main=paste("rho: ", rho3))
```

> abline(lm(data$petal.width ~ data$sepal.length), lwd=2)

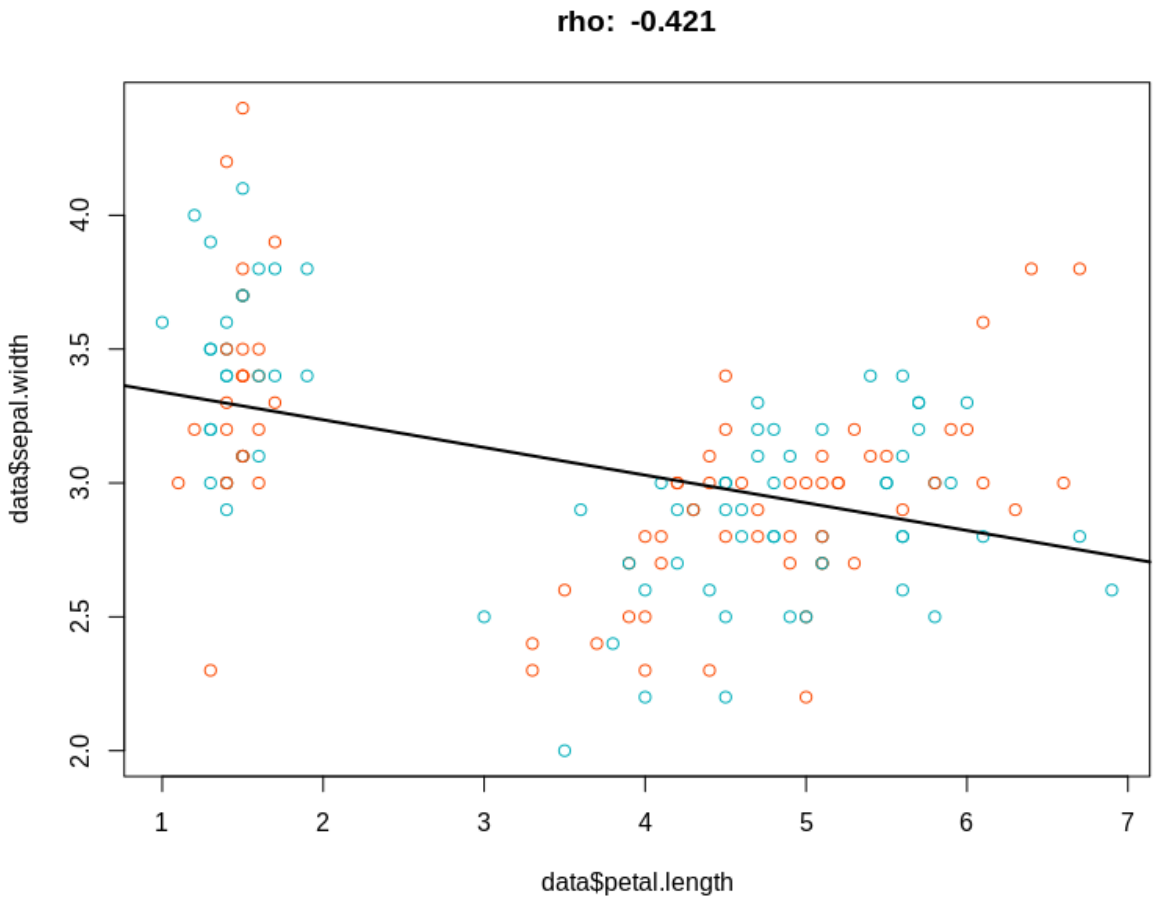


Коментар: Съществува силна корелация между променливите.

> rho4 <- round(cor(data$petal.length,data$sepal.width),3)

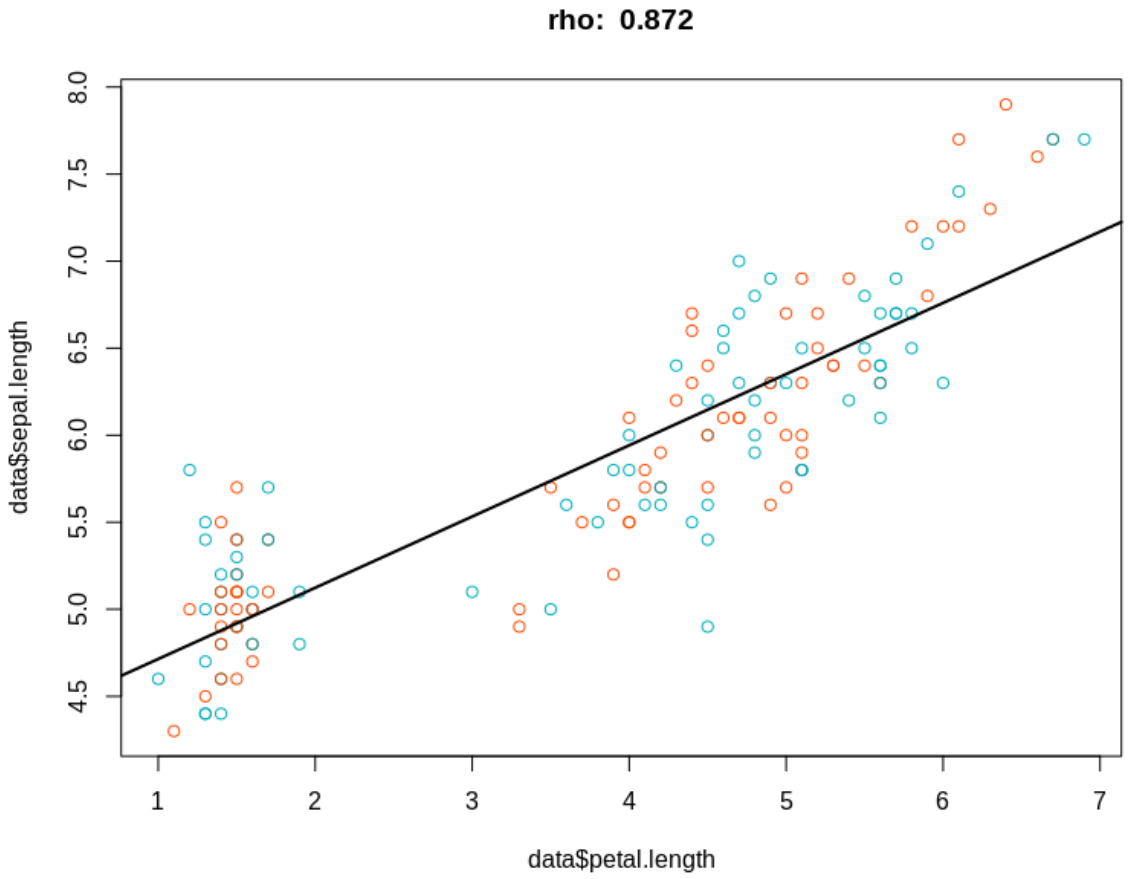
> plot(data$petal.length,data$sepal.width,col = cols, main=paste("rho: ", rho4))

> abline(lm(data$sepal.width ~ data$petal.length), lwd=2)



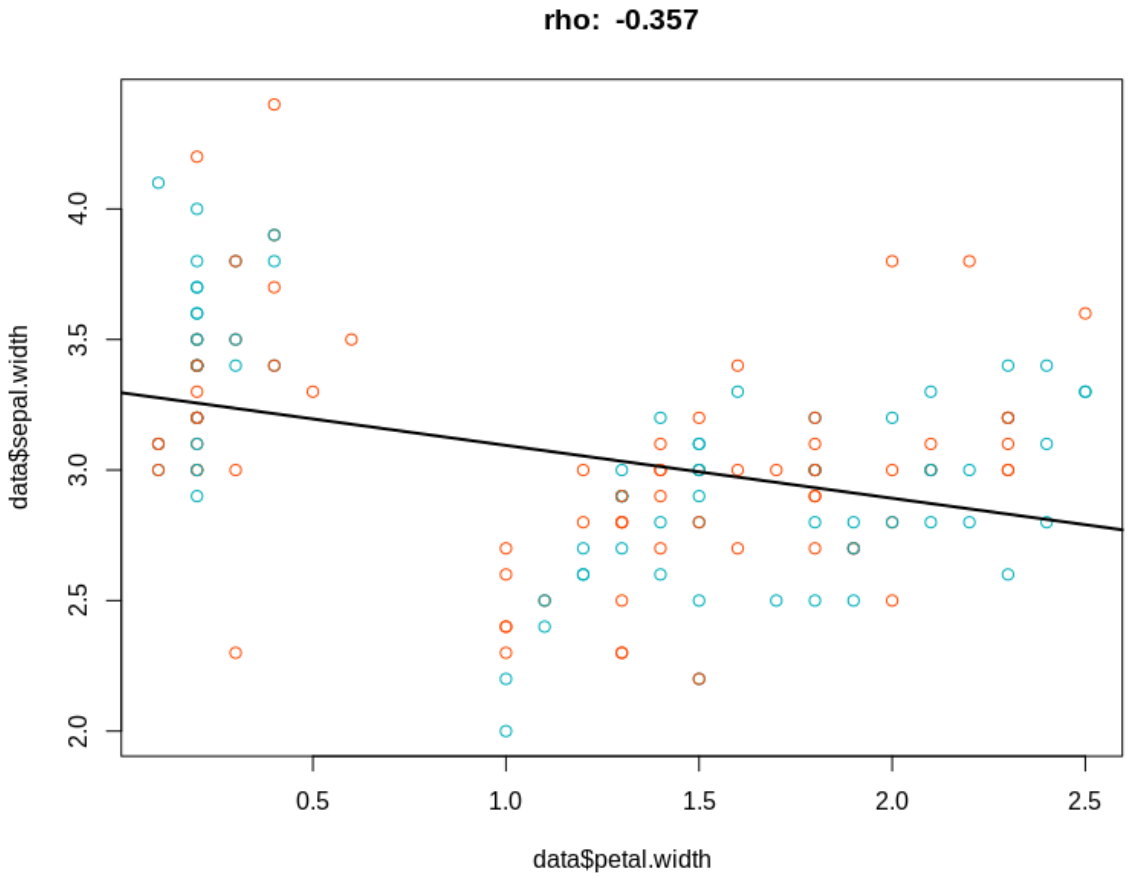
Коментар: Отново от графиката и регресията, както и от коефициента на корелация можем да твърдим, че съществува слаба корелация между променливите.

```
> rho5 <- round(cor(data$petal.length,data$sepal.length),3)
> plot(data$petal.length,data$sepal.length,col = cols, main=paste("rho: ", rho5))
> abline(lm(data$sepal.length ~ data$petal.length), lwd=2)
```



Коментар: Както по-горе, наблюдаваме силна корелация.

```
> rho6 <- round(cor(data$petal.width, data$sepal.width),3)
> plot(data$petal.width,data$sepal.width,col = cols, main=paste("rho: ", rho6))
> abline(lm(data$sepal.width ~ data$petal.width), lwd=2)
```



Коментар: Съществува слаба корелация между двете променливи.

Матрица на корелация:

```
> cor(data[,1:4])
             sepal.length sepal.width petal.length
sepal.length    1.0000000  -0.1093692    0.8717542
sepal.width    -0.1093692   1.0000000   -0.4205161
petal.length    0.8717542  -0.4205161    1.0000000
petal.width     0.8179536  -0.3565441    0.9627571
             petal.width
sepal.length   0.8179536
sepal.width   -0.3565441
petal.length   0.9627571
petal.width    1.0000000
>
```

13. Използвани библиотеки и други източници

Проектът изцяло е написан на R. Визуализацията на данните е постигната с помощта на библиотеките - ggplot2 и data.tables.

Източник на Iris dataset - http://mlr.cs.umass.edu/ml/.