Trabalho

Grupo B

2022-09-10

Sobre o trabalho - O que precisa entregar

- Utilizar o R Markdown para documentar o código com saída html
- Fazer análise exploratória das variáveis com medidas de resumo e gráficos.
- Fazer uma análise de cluster.

Nós escolhemos a base Iris: https://archive.ics.uci.edu/ml/datasets/Iris

O nosso conjunto de dados consiste em 50 amostras de cada uma das três espécies de flores Iris .

Informações dos atributos

1. sepal length in cm - **Comprimento da sépala em cm**
2. sepal width in cm - **Largura da sépala em cm**
3. petal length in cm - **Comprimento da pétala em cm**
4. petal width in cm - **Largura da pétala em cm**
5. class - **espécies estudadas**
    1. **Iris Setosa**
    2. **Iris Versicolour**
    3. **Iris Virginica**

Quatro características (variáveis) foram medidas de cada amostra, são elas o comprimento e a largura da sépala e da pétala, em centímetros.

As espécies alvo do nosso estudo:

Criação das variável necessária e da lista de colunas

Aqui nós criamos a variável necessária para rodar nossa análise além disso também criamos uma lista para alterar o nome padrão das colunas.

```
urlDataSet <- 'http://archive.ics.uci.edu/ml/machine-learning-
databases/iris/iris.data'

colName <- c("sepala_comprimento", "sepala_largura", "petala_comprimento",
"petala_largura","especies")
```

Download da base

Baixamos a base e alteramos o nome das colunas

```
irisDataBase <- read.csv(url(urlDataSet), header = FALSE, col.names =
colName)
```

Verificação do nome das colunas alteradas:

```
irisDataBase %>% colnames()

## [1] "sepala_comprimento" "sepala_largura"     "petala_comprimento"
## [4] "petala_largura"     "especies"
```

Primeiras análises

Aqui nós realizaremos as primeiras análises da nossa base de dados. Executamos as funções:

- summary - Função para realizar uma análise estatística resumida
- head - Exibir os primeiros resultados
- str - Função para exibir a estrutura dos nossos dados
- Realizamos também uma validação para verificar se existe valores do tipo NA e a proporção com que ele existe para cada valor da base

```
irisDataBase %>% head()

##   sepala_comprimento sepala_largura petala_comprimento petala_largura
## 1                5.1            3.5                1.4            0.2
## 2                4.9            3.0                1.4            0.2
## 3                4.7            3.2                1.3            0.2
## 4                4.6            3.1                1.5            0.2
## 5                5.0            3.6                1.4            0.2
## 6                5.4            3.9                1.7            0.4
##      especies
## 1 Iris-setosa
## 2 Iris-setosa
## 3 Iris-setosa
## 4 Iris-setosa
## 5 Iris-setosa
## 6 Iris-setosa

irisDataBase %>% str()

## 'data.frame':    150 obs. of  5 variables:
##  $ sepala_comprimento: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
```

```
##  $ sepala_largura    : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ petala_comprimento: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ petala_largura    : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ especies          : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa"
"Iris-setosa" ...
```

```
irisDataBase %>% summary()
```

```
##  sepala_comprimento sepala_largura  petala_comprimento petala_largura
##  Min.   :4.300      Min.   :2.000   Min.   :1.000      Min.   :0.100
##  1st Qu.:5.100      1st Qu.:2.800   1st Qu.:1.600      1st Qu.:0.300
##  Median :5.800      Median :3.000   Median :4.350      Median :1.300
##  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
##    especies
##  Length:150
##  Class :character
##  Mode  :character
##
##
##
```

```
Count <- sum(is.na(irisDataBase))
CalcProportion <- irisDataBase %>% nrow() / Count

Proportion <- ifelse(is.infinite(CalcProportion), 0, CalcProportion)

data.frame(Index = colnames(irisDataBase), Count, Proportion)
```

```
##                Index Count Proportion
## 1 sepala_comprimento     0          0
## 2     sepala_largura     0          0
## 3 petala_comprimento     0          0
## 4     petala_largura     0          0
## 5           especies     0          0
```

Verificando a dimensionalidade dos dados

Para verificar a dimensionalidade nós utilizamos a função dim(), a função então retorna que:

- **150 linhas** de observações e **5 colunas** de variáveis

```
irisDataBase %>% dim()
```

```
## [1] 150   5
```

Verificação do desvio padrão

A seguir nós realizamos uma análise do desvio padrão das variáveis:

- sepala_comprimento

- sepala_largura
- petala_comprimento
- petala_largura

Antes de exibir os desvios padrões nós criamos uma função para evitar um pouco a duplicação de chamadas.

Poderiamos ter realizado uma análise das variveis unicamente em um chamada, mas achamos mais didático analisar uma a uma.

```
fcStandardDeviation <- function(database, variable){
  if(variable %in% colnames(database)) {
    result <- database %>% dplyr::select(all_of(variable))

    result[,] %>% sd()
  } else {
    return(FALSE)
  }
}

desvio_sepala_comprimento = fcStandardDeviation(irisDataBase,
"sepala_comprimento")
desvio_sepala_largura = fcStandardDeviation(irisDataBase, "sepala_largura")
desvio_petala_comprimento = fcStandardDeviation(irisDataBase,
"petala_comprimento")
desvio_petala_largura = fcStandardDeviation(irisDataBase, "petala_largura")

resultadoDesvios <- data.frame(
  desvio_sepala_comprimento,
  desvio_sepala_largura,
  desvio_petala_comprimento,
  desvio_petala_largura
)

print(xtable(resultadoDesvios), type = "html")
```

desvio_sepala_comprimento

desvio_sepala_largura

desvio_petala_comprimento

desvio_petala_largura

1

0.83

0.43

1.76

0.76

O resultado exibe o nosso grau de dispersão dos nossos conjuntos de dados.

Análise de quantil

Iremos observar o quantil das variáveis:

- sepala_comprimento
- sepala_largura
- petala_comprimento
- petala_largura

Criamos também uma função para retornar os quantile, poderimos ter feito o resultado diretamente usando uma função apply:

```
apply(iris[,1:4], 2, quantile)
```

```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width
## 0%            4.3         2.0         1.00         0.1
## 25%           5.1         2.8         1.60         0.3
## 50%           5.8         3.0         4.35         1.3
## 75%           6.4         3.3         5.10         1.8
## 100%          7.9         4.4         6.90         2.5
```

Porém achamos mais didático deixar uma função e separar o valor em um DataFrame

```
fcQuantile <- function(database, variable){
  if(variable %in% colnames(database)) {
    result <- database %>% dplyr::select(all_of(variable))

    result[,] %>% quantile()
  } else {
    return(FALSE)
  }
}

quantile_sepala_comprimento = fcQuantile(irisDataBase, "sepala_comprimento")
quantile_sepala_largura = fcQuantile(irisDataBase, "sepala_largura")
quantile_petala_comprimento = fcQuantile(irisDataBase, "petala_comprimento")
quantile_petala_largura = fcQuantile(irisDataBase, "petala_largura")

resultadoQuantile <- data.frame(
  quantile_sepala_comprimento,
  quantile_sepala_largura,
  quantile_petala_comprimento,
  quantile_petala_largura
)

print(xtable(resultadoQuantile), type = "html")
```

quantile_sepala_comprimento

quantile_sepala_largura

quantile_petala_comprimento

quantile_petala_largura

0%

4.30

2.00

1.00

0.10

25%

5.10

2.80

1.60

0.30

50%

5.80

3.00

4.35

1.30

75%

6.40

3.30

5.10

1.80

100%

7.90

4.40

6.90

2.50

Agrupamento dos dados

Realizamos o agrupamento das espécies, para então realizar uma análise da média por variável agrupada:

```
irisDataBaseGroup <- irisDataBase %>% group_by(especies) %>%
  summarise(
    comprimento_medio_sepala = mean(sepala_comprimento, na.rm = TRUE),
    largura_media_sepala = mean(sepala_largura, na.rm = TRUE),
    comprimento_medio_petala = mean(petala_comprimento, na.rm = TRUE),
    largura_media_petala = mean(petala_largura, na.rm = TRUE),
)

print(xtable(irisDataBaseGroup), type = "html")
```

especies

comprimento_medio_sepala

largura_media_sepala

comprimento_medio_petala

largura_media_petala

1

Iris-setosa

5.01

3.42

1.46

0.24

2

Iris-versicolor

5.94

2.77

4.26

1.33

3

Iris-virginica

6.59

2.97

5.55

2.03

A partir do nosso agrupamento já somos capazes de determinar os tamanhos médios de sepala e petala de cada espécie.

Agrupamento pelo desvios

Agora nós vamos agrupar as espécies com seus desvios padrões

```
irisDataBaseGroupDesv <- irisDataBase %>% group_by(especies) %>%
  summarise(
    desvio_comprimento_sepala = sd(sepala_comprimento, na.rm = TRUE),
    desvio_largura_sepala = sd(sepala_largura, na.rm = TRUE),
    desvio_comprimento_petala = sd(petala_comprimento, na.rm = TRUE),
    desvio_largura_petala = sd(petala_largura, na.rm = TRUE),
)

print(xtable(irisDataBaseGroupDesv), type = "html")
```

especies

desvio_comprimento_sepala

desvio_largura_sepala

desvio_comprimento_petala

desvio_largura_petala

1

Iris-setosa

0.35

0.38

0.17

0.11

2

Iris-versicolor

0.52

0.31

0.47

0.20

3

Iris-virginica

0.64

0.32

0.55

0.27

Categorização com base em quartis

Nós criaremos uma variável categórica com base no quartil a partir das variáveis.

```
quartils_comprimento_sepala<- cut(irisDataBase$sepala_comprimento,
breaks=quantile(irisDataBase$sepala_comprimento), include.lowes=T)

irisDataBaseQuartis <- irisDataBase

irisDataBaseQuartis$sepala_comprimento_quartil_grupo <-
quartils_comprimento_sepala

result <- aggregate(.~especies+sepala_comprimento_quartil_grupo,
irisDataBaseQuartis, mean)
```

Unificamos os dados e construimos uma tabela das contagens dos quartis

CrossTable dos quartis de comprimento de sepala

```
resultQuartisSepalaComprimento <- table(irisDataBaseQuartis$especies,
irisDataBaseQuartis$sepala_comprimento_quartil_grupo)

print(xtable(resultQuartisSepalaComprimento), type = "html")
```

[4.3,5.1]

(5.1,5.8]

(5.8,6.4]

(6.4,7.9]

Iris-setosa

36

14

0

0

Iris-versicolor

4

20

17

9

Iris-virginica

1

5

18

26

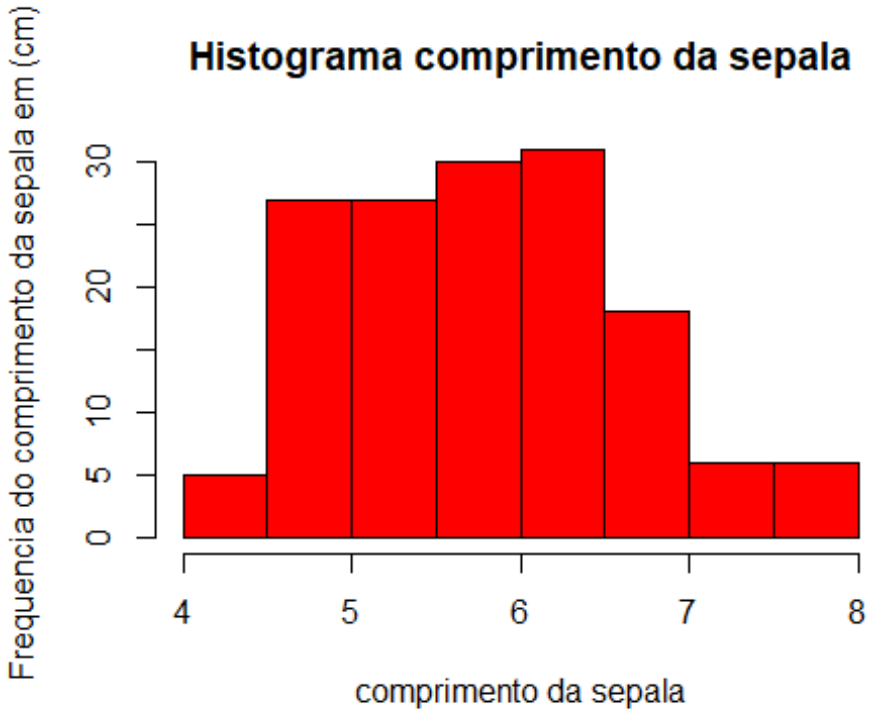
Histograma das variáveis

Nós agora iremos análiser o histograma das variáveis:

- sepala_comprimento
- sepala_largura
- petala_comprimento
- petala_largura

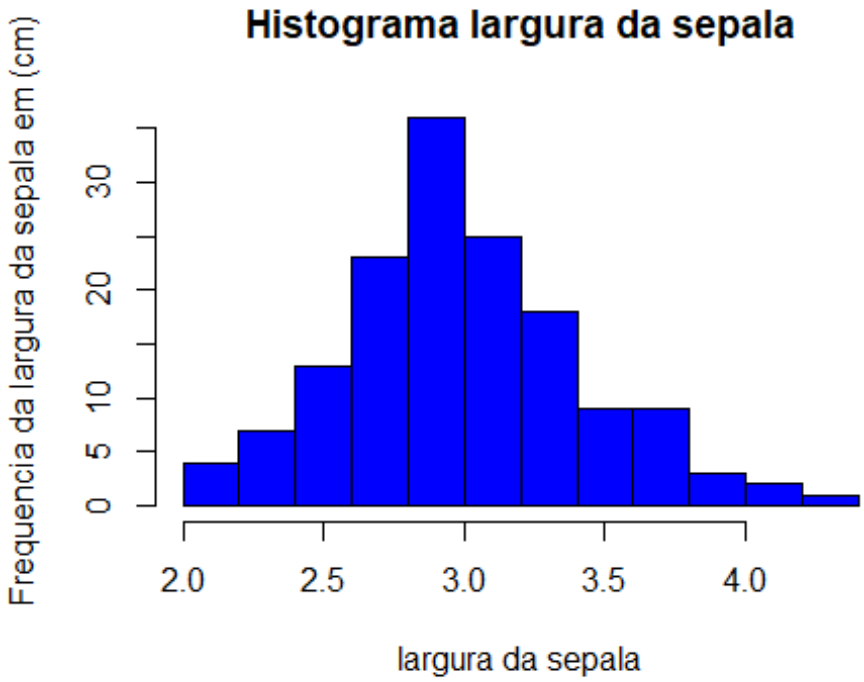
Histograma comprimento da sepala

```
hist(irisDataBase$sepala_comprimento, xlab = 'comprimento da sepala', ylab = 'Frequencia do comprimento da sepala em (cm)', main = 'Histograma comprimento da sepala', col = 'red')
```



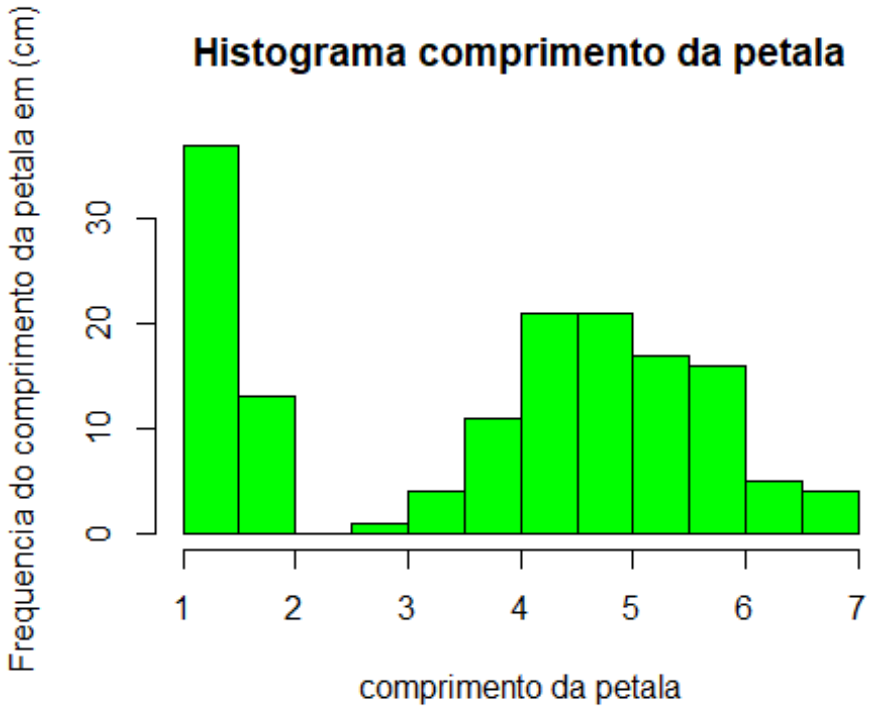
Histograma largura da sepala

```
hist(irisDataBase$sepala_largura, xlab = 'largura da sepala', ylab =
'Frequencia da largura da sepala em (cm)', main = 'Histograma largura da
sepala', col = 'blue')
```



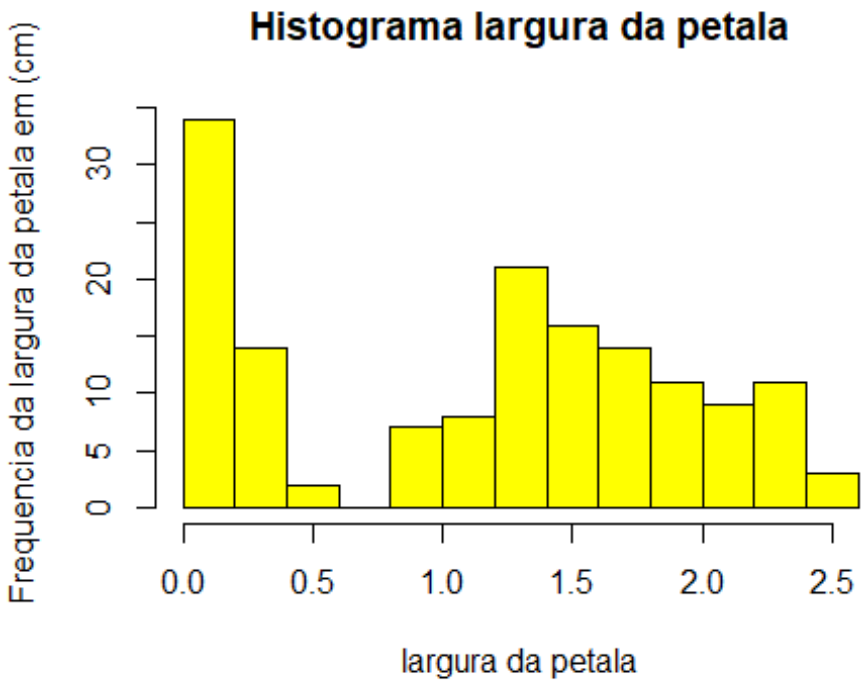
Histograma comprimento da petala

```
hist(irisDataBase$petala_comprimento, xlab = 'comprimento da petala', ylab = 'Frequencia do comprimento da petala em (cm)', main = 'Histograma comprimento da petala', col = 'green')
```



Histograma largura da petala

```
hist(irisDataBase$petala_largura, xlab = 'largura da petala', ylab =
'Frequencia da largura da petala em (cm)', main = 'Histograma largura da
petala', col = 'yellow')
```



Boxplot das variáveis

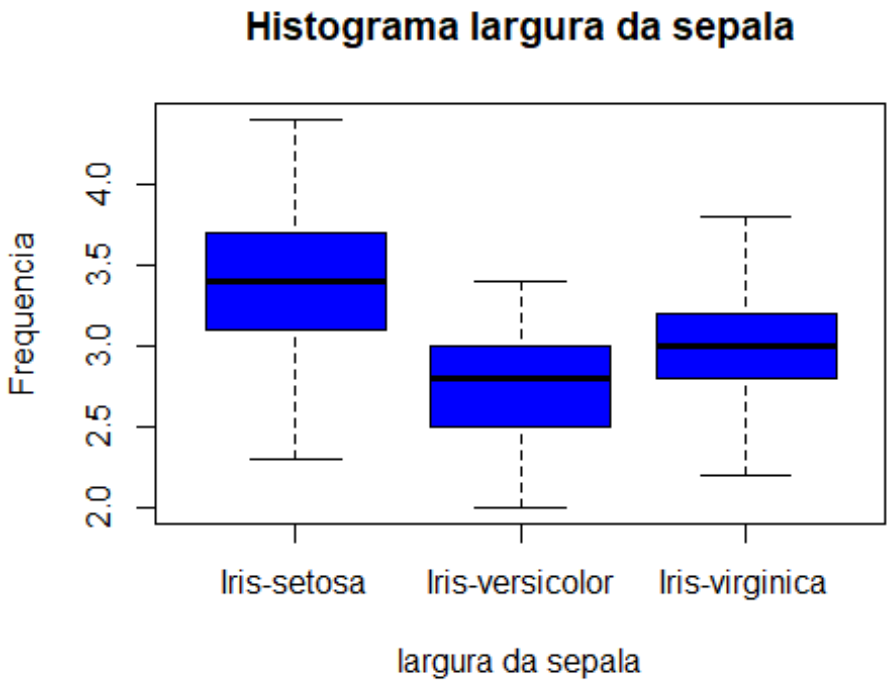
Nós agora iremos criar os boxplots das variáveis:

- sepala_comprimento
- sepala_largura
- petala_comprimento
- petala_largura

E verificaremos se existem outliers em nossos dados.

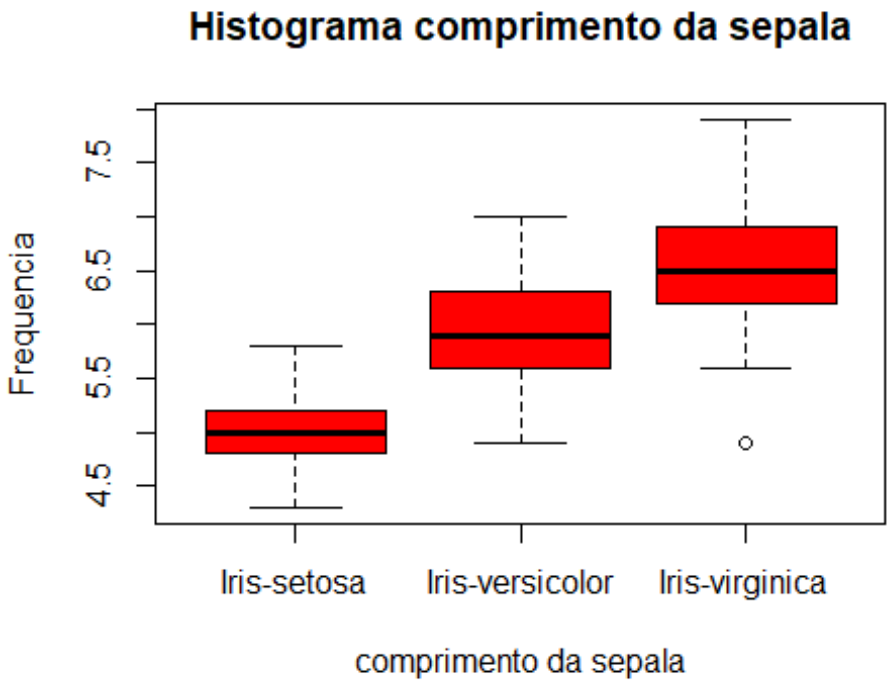
Boxplot largura Sepala

```
boxplot(sepala_largura ~ especies, data=irisDataBase, xlab = 'largura da sepala', ylab = 'Frequencia', main = 'Histograma largura da sepala', col = 'blue')
```



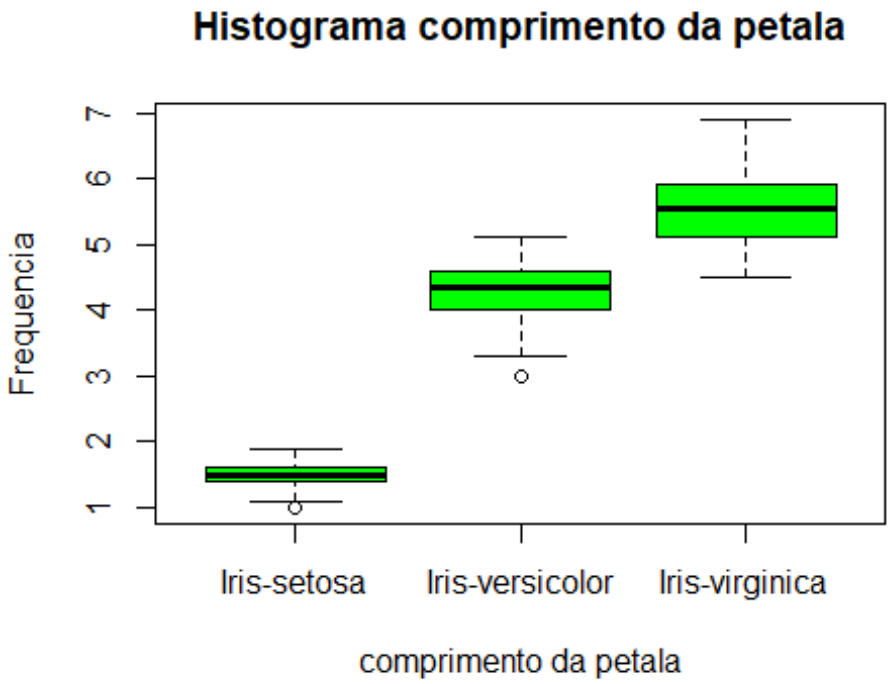
Boxplot comprimento Sepala

```
boxplot(sepala_comprimento ~ especies, data=irisDataBase, xlab = 'comprimento
da sepala', ylab = 'Frequencia', main = 'Histograma comprimento da sepala',
col = 'red')
```



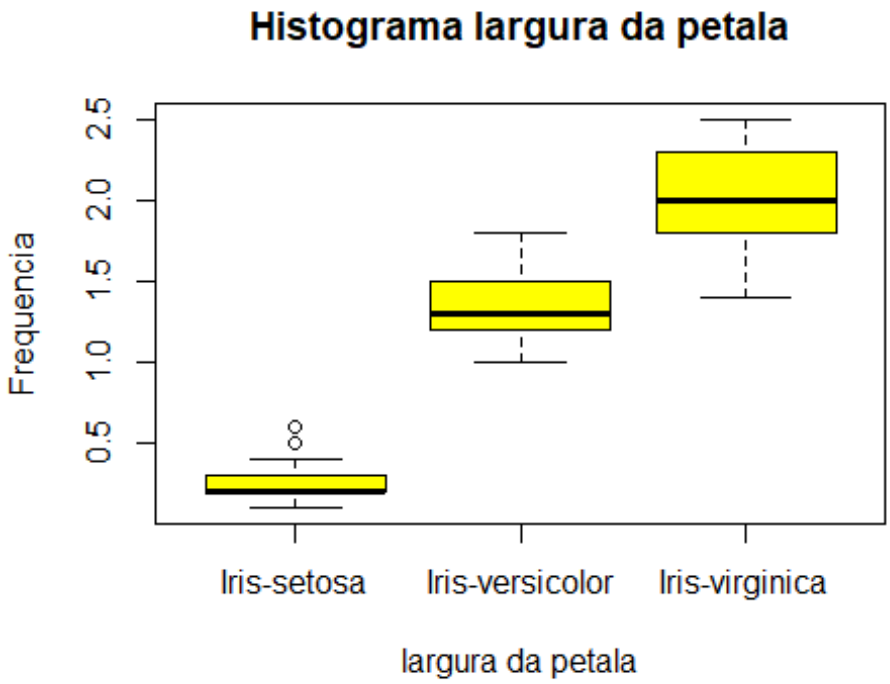
Boxplot comprimento da petala

```
boxplot(petala_comprimento ~ especies, data=irisDataBase, xlab = 'comprimento
da petala', ylab = 'Frequencia', main = 'Histograma comprimento da petala',
col = 'green')
```



Boxplot largura da petala

```
boxplot(petala_largura ~ especies, data=irisDataBase, xlab = 'largura da petala', ylab = 'Frequencia', main = 'Histograma largura da petala', col = 'yellow')
```


Histograma largura da petala
Frequencia
0.5
1.0
1.5
2.0
2.5
Iris-setosa
Iris-versicolor
Iris-virginica
largura da petala