Efeitos do álcool nos estudos

Grupo 04:
Maria Eduarda Garcia
Mirela Mei

Prof.ª Ana Amélia

Dataset

DESCRIÇÃO

Acompanha o desempenho estudantil na educação secundária em duas escolas portuguesas.
Consideram-se matérias de matemática e português.
Na presente análise, estudaremos apenas o primeiro caso.

VARIÁVEIS

Os atributos incluem as notas dos alunos, dispositivos demográficos, sociais e relacionados à escola e foram coletados com o uso de relatórios e questionários escolares.

DADOS COLETADOS

- ORIGEM ??
- 395 linhas e 33 colunas
- Sem missing data

Dataset

VARIÁVEIS

- Escola em que estuda
- Gênero
- Idade
- Área em que reside
- Tamanho da família
- Estado de coabitação dos pais
- Nível de escolaridade da mãe
- Nível de escolaridade do pai
- Trabalho do pai
- Trabalho da mãe
- Razão pela qual escola foi escolhida
- Quem possui a guarda
- Tempo da escola para casa
- Tempo de estudo semanal
- Repetências
- Suporte escolar extra
- Suporte familiar
- Classes extra pagas
- Atividades extracurriculares
- Frequentouescola de enfermaria
- Pretende cursar ensino superior
- Acesso à internet
- Participa de relação romântica
- Qualidade da relação familiar
- Tempo livre após a escola
- Frequência com que sai com amigos
- Consumo de álcool diário
- Consumo de álcool aos finais de semana
- Estado de saúde
- Número de faltas
- Nota do primeiro período
- Nota do segundo período
- Nota final

Análise de Cluster

DESCRIÇÃO

Uma técnica estatística (exploratória e não-inferencial) usada para classificar elementos em grupos, com base nas similaridades e diferenças das características que estes elementos possuem. Os grupos são determinados de forma a obter-se homogeneidade dentro dos grupos e heterogeneidade entre eles.
Para definir a semelhança ou diferença entre os elementos é usada uma função de distância, que precisa ser definida considerando o contexto do problema em questão.

Podemos dividir a análise de cluster em dois grandes tipos de métodos: hierárquicos e não hierárquicos.

Análise de Cluster

MEDIDAS DE SIMILARIDADE

É necessário pré-especificar a medida de similaridade a ser utilizada no agrupamento, pois existem várias medidas de similaridades diferentes e cada uma delas produz um determinado tipo de agrupamento.

As mais usadas são:

- Distância Euclidiana
- Distância de Manhattan
- Distância de Correlação de Pearson
- Distância de Correlação de Eisen
- Distância de Correlação de Spearman
- Distância de Correlação de Kendal

Análise de Cluster

MÉTODO DE AGRUPAMENTO HIERÁRQUICO

- Uma série de sucessivos agrupamentos ou sucessivas divisões de elementos
- São subdivididos em métodos aglomerativos ou divisivos
- Representados por um diagrama bi-dimensional chamado de dendrograma ou diagrama de árvore
- Métodos de ligação, método de centróide, método de Ward

MÉTODO DE AGRUPAMENTO NÃO-HIERÁRQUICO

- A quantidade de grupos (k) deve ser pré-estabelecida
- Não há propriedade de hierarquia, ou seja, grupos unidos num determinado passo podem se separar em passos posteriores
- A partição dos dados se dá respeitando duas premissas: a coesão interna e o isolamento dos grupos
- K-Means

Análise de Cluster

EXEMPLO

Foi utilizado um dataset do repositório UCI contendo transações reais efetuadas entre 01/12/2010 e 09/12/2011 de um varejo online com sede alocada no Reino Unido, o qual contém um mix variado de produtos.

Primeiramente, filtra-se apenas os dados relevantes para análise, nesse caso, como as observações (produtos) se associam (agrupamento) e quão heterogêneos esses grupos são entre si em função do preço unitário e quantidade vendida.

Para a criação dos clusters, o cálculo de distância entre as variáveis de cada observação se dá pela distância euclidiana.

Análise de Cluster

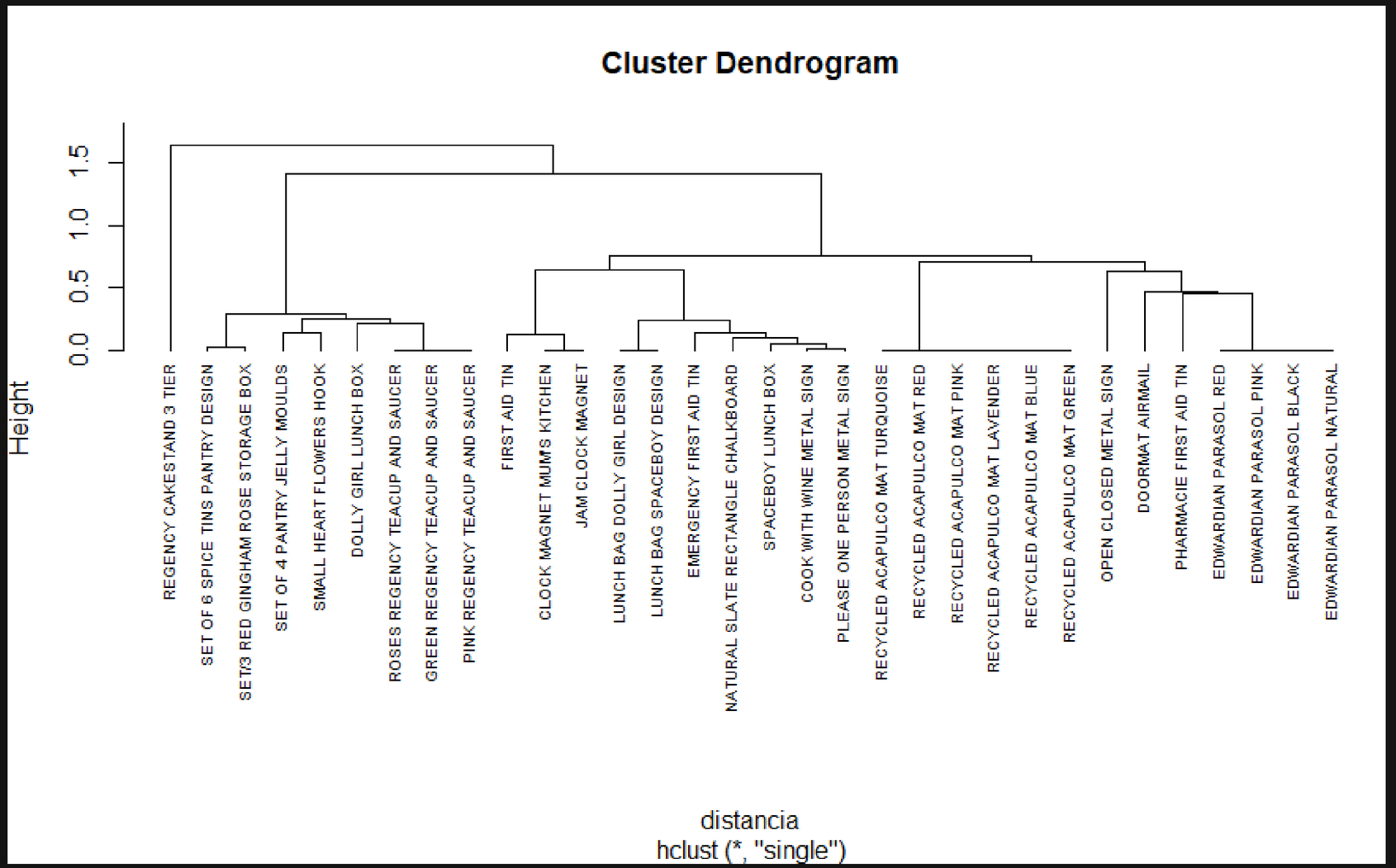
EXEMPLO

À princípio foi feito o clustering no modelo hierárquico e para o resultado desejado o método aplicado foi o de vizinho mais próximo (single linkage).

Embora o dendrograma gerado visualmente nos dê um norte quanto a quantidade de clusters ideal para utilizar, é fundamental que também seja aplicado o método de Elbow, que nada mais é do que a técnica que testa a variância dos dados em relação ao número de clusters.

É considerado um valor ideal de K (quantidade de grupos) quando o aumento no número de clusters (eixo X) não representa um valor significativo de ganho (eixo Y).

Análise de Cluster



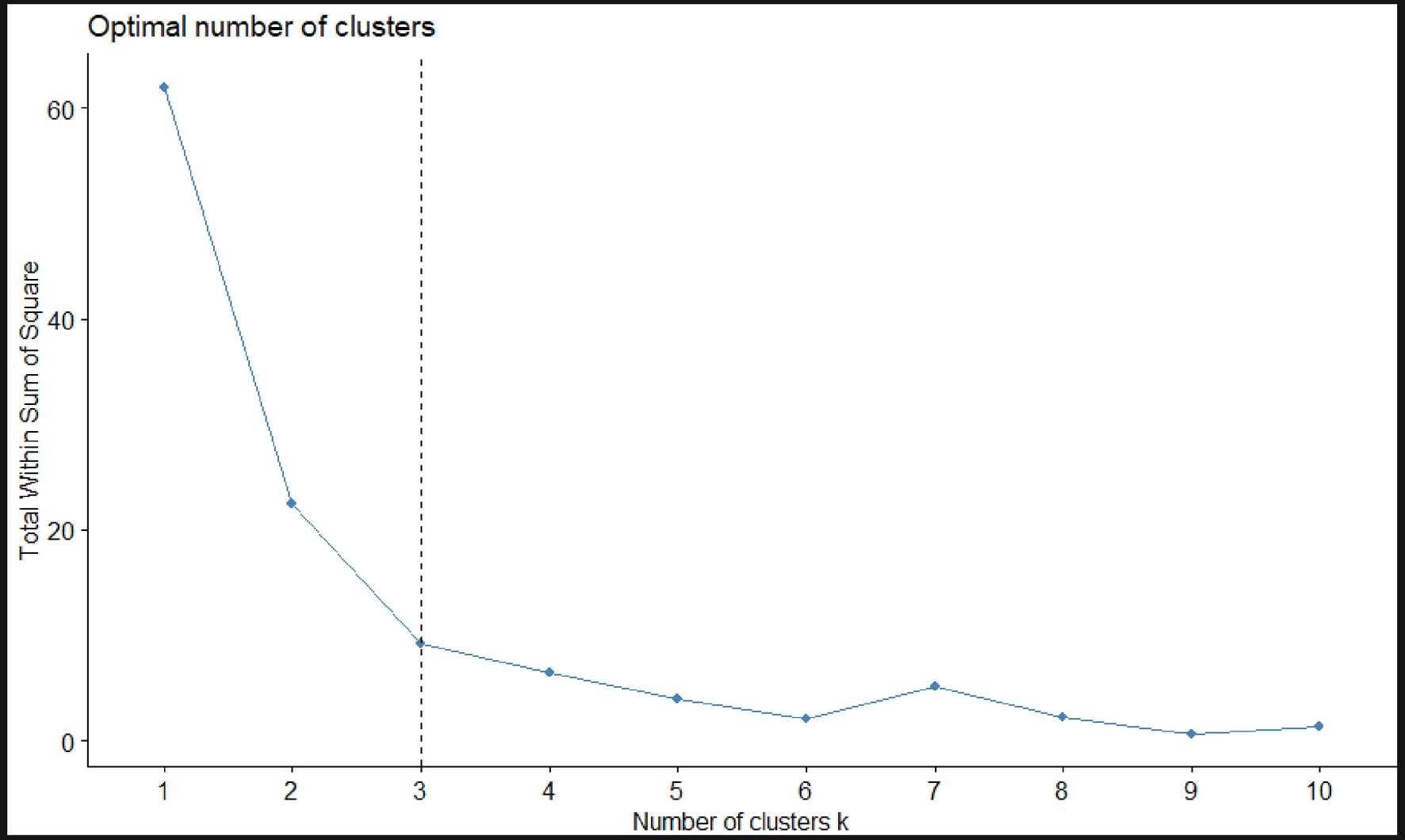
Análise de Cluster

EXEMPLO

No gráfico a seguir, plotou-se o somatório das variâncias dos dados em relação ao numero de clusters e como pode-se observar, a partir do terceiro cluster, as distâncias dos erros quadráticos praticamente se estabilizam. Sendo assim, considera-se o número total de 3 clusters para seguir com a análise.

Foram criados os 3 clusters agrupando:
Cluster 1 = 23 observações
Cluster 2 = 8 observações
Cluster 3 = 1 observação

Análise de Cluster



Análise de Cluster

EXEMPLO

Visualizando graficamente os grupos, podemos observar o quão próximas as observações estão representadas entre si dentro do mesmo cluster e o quão distantes estão entre grupos.

Análise de Cluster

ESCOLHA DAS VARIÁVEIS

As variáveis escolhidas para a analise foram:
G3 (nota final escolar), idade e número de faltas.

PERGUNTA DA ANÁLISE

Que grupos homogêneos de alunos emergem com base em resultados de nota escolar, idade e número de faltas?

Estatística Descritiva

G3 (NOTA FINAL ESCOLAR)

Tabelas de frequência:

```
> table(dados$G3)

 0  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
38  1  7 15  9 32 28 56 47 31 31 27 33 16  6 12  5  1
> prop.table(table(dados$G3))

          0           4           5           6           7           8
0.096202532 0.002531646 0.017721519 0.037974684 0.022784810 0.081012658
          9          10          11          12          13          14
0.070886076 0.141772152 0.118987342 0.078481013 0.078481013 0.068354430
         15          16          17          18          19          20
0.083544304 0.040506329 0.015189873 0.030379747 0.012658228 0.002531646
>
```

Estatística Descritiva

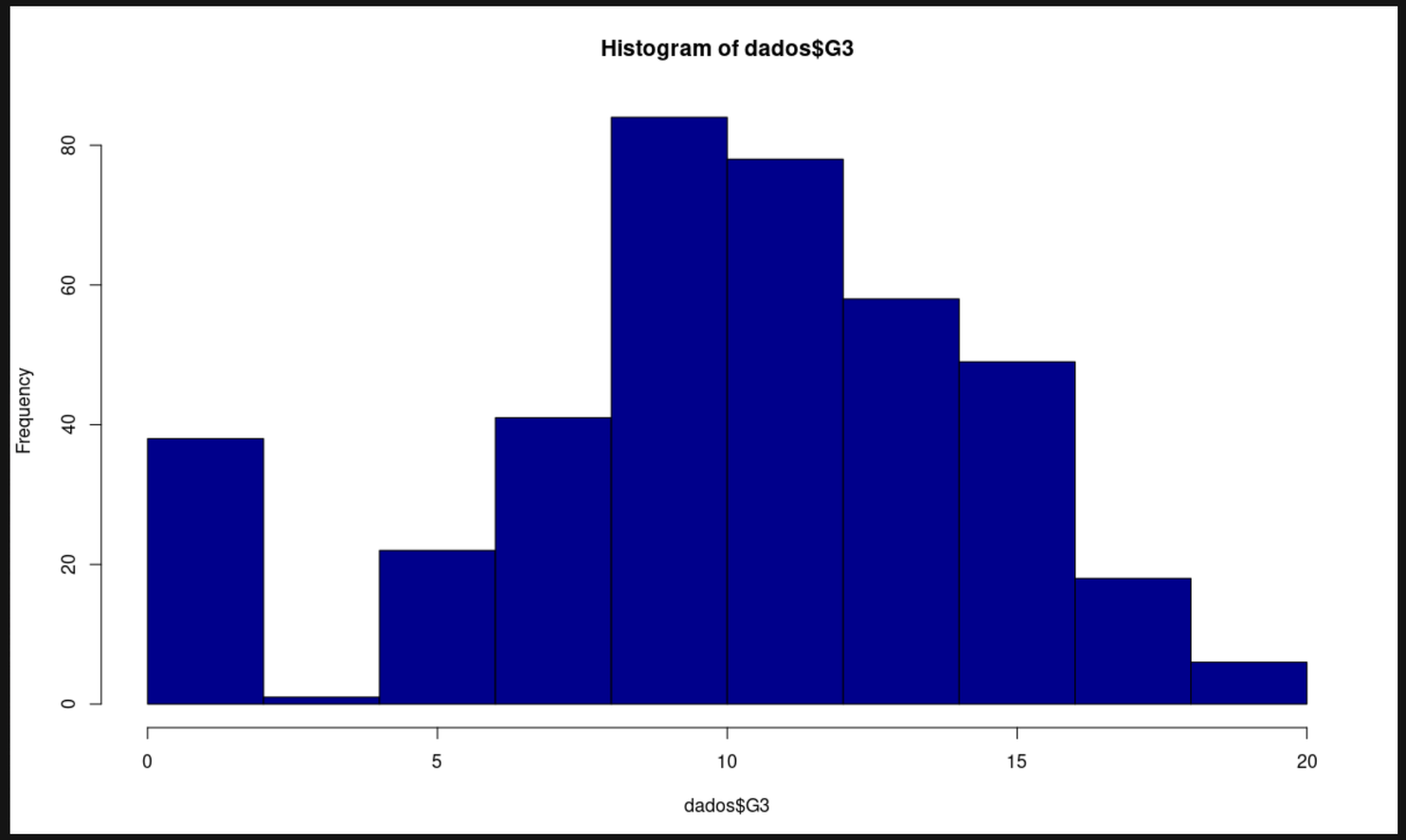
G3 (NOTA FINAL ESCOLAR)

Medidas de tendência central e de dispersão:

```
> summary(dados$G3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   11.00   10.42   14.00   20.00
> medianG3 <- median(dados$G3)
> print(medianG3)
[1] 11
> getmode <- function(dados) {
+   uniqv <- unique(dados)
+   uniqv[which.max(tabulate(match(dados, uniqv)))]
+ }
> resultG3 <- getmode(dados$G3)
> print(resultG3)
[1] 10
> describe(dados$G3)
   vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 395 10.42 4.58     11   10.84 4.45   0  20    20 -0.73     0.37 0.23
>
```

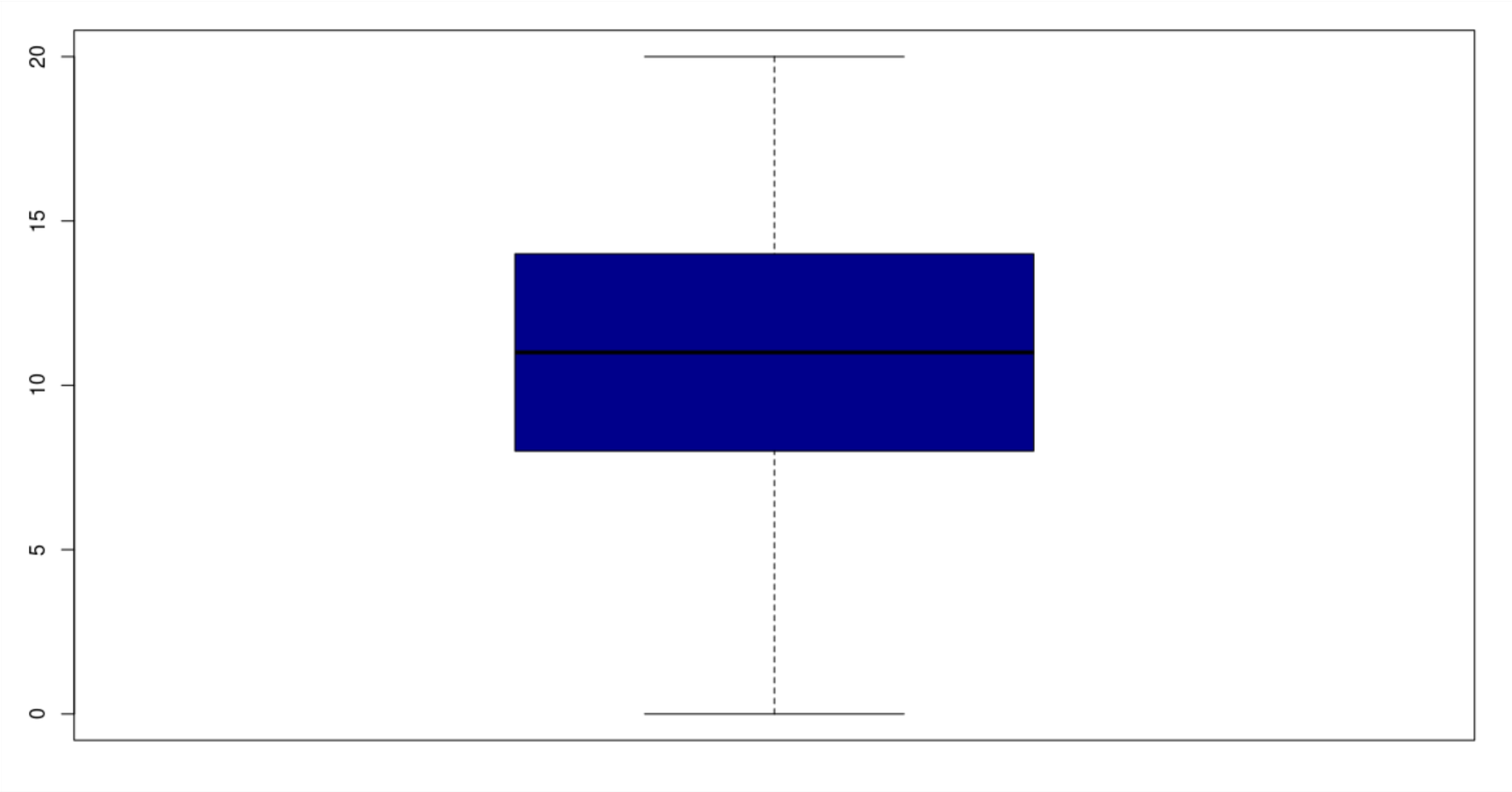
Estatística Descritiva

G3 (NOTA FINAL ESCOLAR)



Estatística Descritiva

G3 (NOTA FINAL ESCOLAR) - BOXPLOT



Estatística Descritiva

IDADE

Tabelas de frequência:

```
> table(dados$age)

 15  16  17  18  19  20  21  22
 82 104  98  82  24   3   1   1
> prop.table(table(dados$age))

         15          16          17          18          19          20
0.207594937 0.263291139 0.248101266 0.207594937 0.060759494 0.007594937
         21          22
0.002531646 0.002531646
>
```

Estatística Descritiva

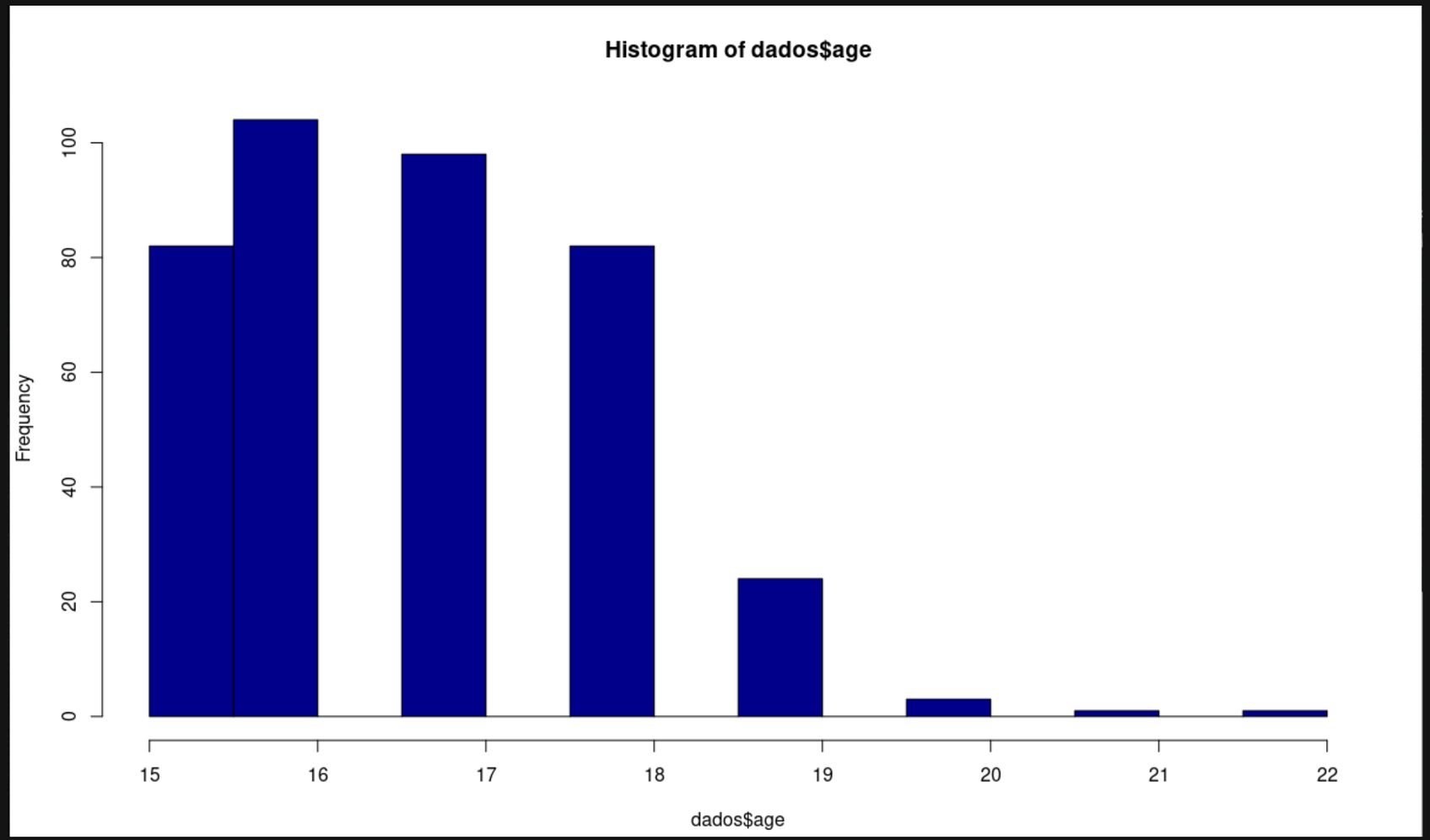
IDADE

Medidas de tendência central e de dispersão:

```
> summary(dados$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   15.0    16.0    17.0    16.7    18.0    22.0
> medianAge <- median(dados$age)
> print(medianAge)
[1] 17
> getmode <- function(dados) {
+   uniqv <- unique(dados)
+   uniqv[which.max(tabulate(match(dados, uniqv)))]
+ }
> resultAge <- getmode(dados$age)
> print(resultAge)
[1] 16
> describe(dados$age)
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 395 16.7 1.28     17   16.63 1.48  15  22     7 0.46    -0.03 0.06
>
```

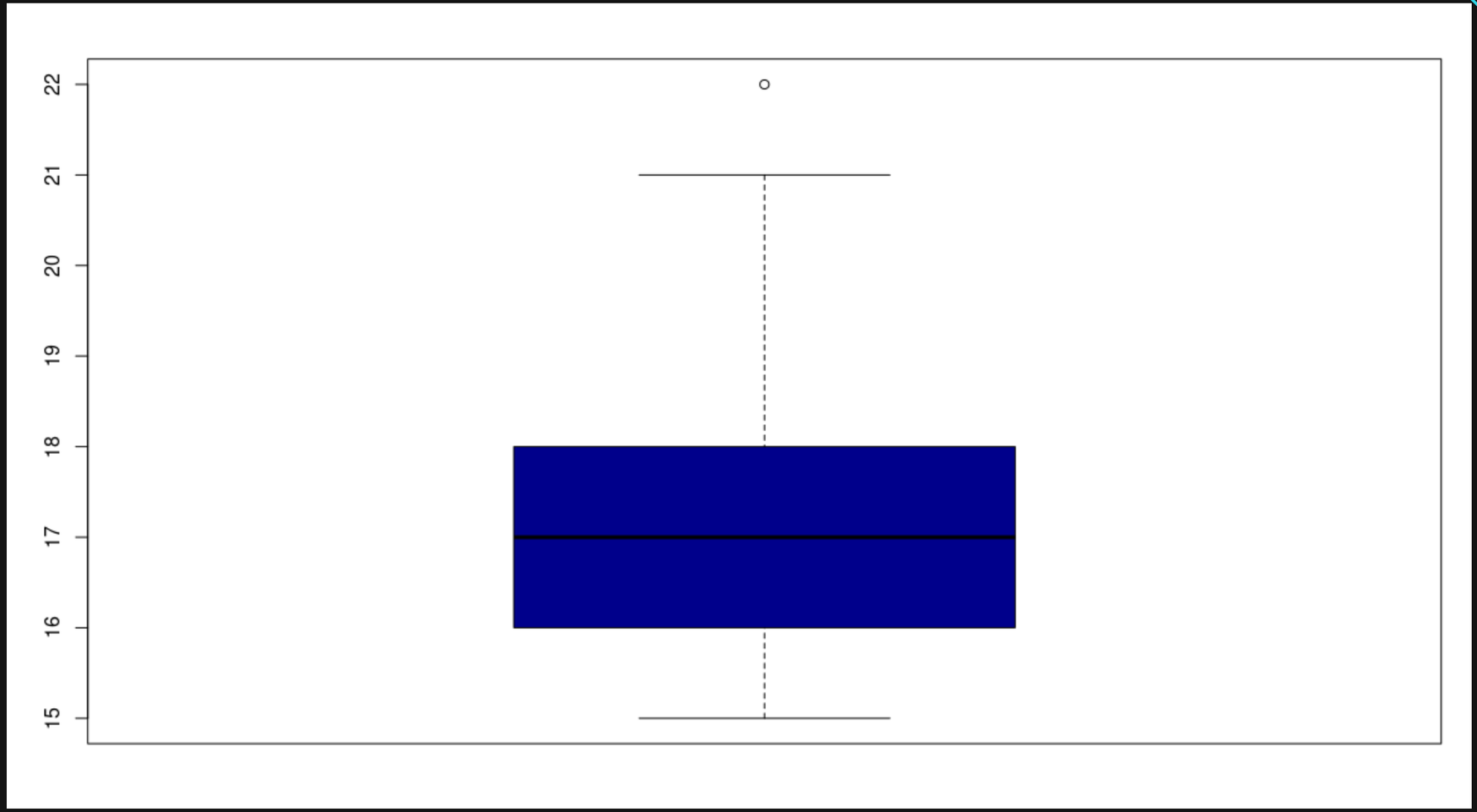
Estatística Descritiva

IDADE



Estatística Descritiva

IDADE - BOXPLOT



Estatística Descritiva

NÚMERO DE FALTAS

Tabelas de frequência:

```
> table(dados$absences)

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
115   3  65   8  53   5  31   7  22   3  17   3  12   3  12   3   7   1   5   1
 20  21  22  23  24  25  26  28  30  38  40  54  56  75
  4   1   3   1   1   1   1   1   1   1   1   1   1   1
> prop.table(table(dados$absences))

          0           1           2           3           4           5
0.291139241 0.007594937 0.164556962 0.020253165 0.134177215 0.012658228
          6           7           8           9          10          11
0.078481013 0.017721519 0.055696203 0.007594937 0.043037975 0.007594937
         12          13          14          15          16          17
0.030379747 0.007594937 0.030379747 0.007594937 0.017721519 0.002531646
         18          19          20          21          22          23
0.012658228 0.002531646 0.010126582 0.002531646 0.007594937 0.002531646
         24          25          26          28          30          38
0.002531646 0.002531646 0.002531646 0.002531646 0.002531646 0.002531646
         40          54          56          75
0.002531646 0.002531646 0.002531646 0.002531646
>
```

Estatística Descritiva

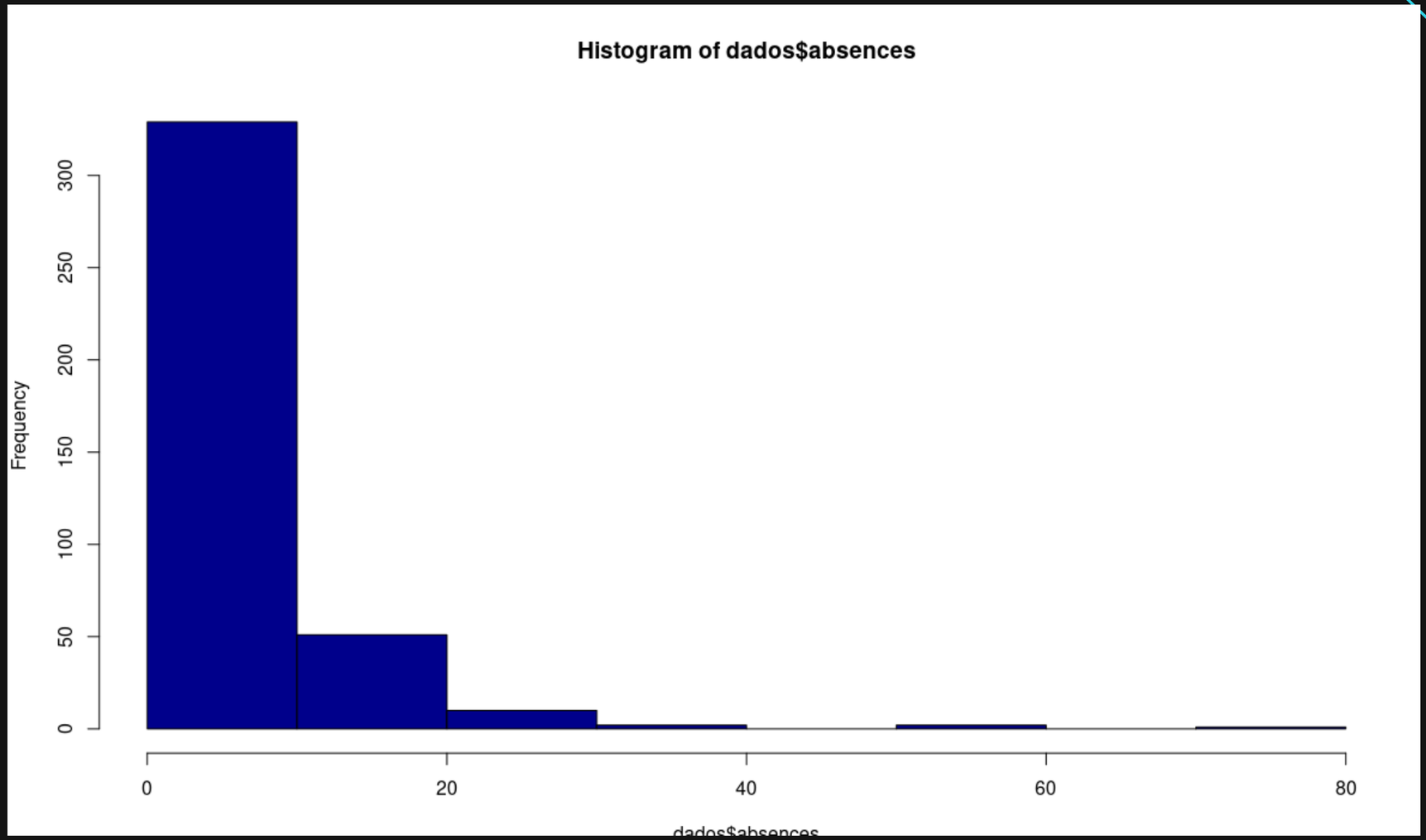
NÚMERO DE FALTAS

Medidas de tendência central e de dispersão:

```
> summary(dados$absences)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   4.000   5.709   8.000  75.000
> medianAbsences <- median(dados$absences)
> print(medianAbsences)
[1] 4
> getmode <- function(dados) {
+   uniqv <- unique(dados)
+   uniqv[which.max(tabulate(match(dados, uniqv)))]
+ }
> resultAbsences <- getmode(dados$absences)
> print(resultAbsences)
[1] 0
> describe(dados$absences)
   vars   n mean sd median trimmed  mad min max range skew kurtosis  se
X1    1 395 5.71  8      4    4.24 5.93   0  75    75 3.64    21.31 0.4
>
```

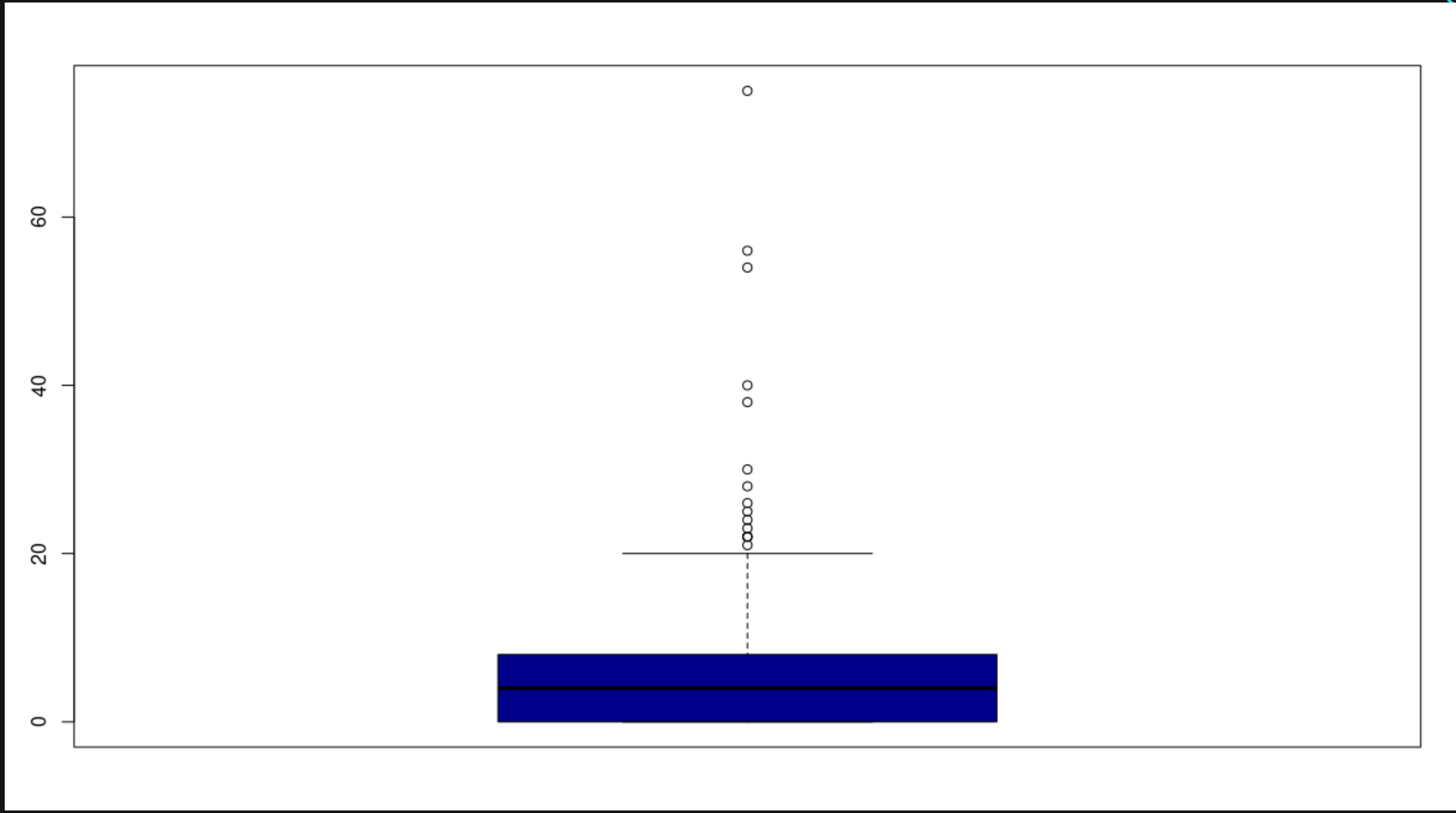
Estatística Descritiva

NÚMERO DE FALTAS



Estatística Descritiva

NÚMERO DE FALTAS - BOXPLOT



Análise de Cluster

MÉTODO DE AGRUPAMENTO HIERÁRQUICO

- Utilizamos a Distância Euclidiana e a Ligação Completa

Distância Euclidiana: essa é provavelmente a mais conhecida e usada medida de distância. Ela simplesmente é a distância geométrica no espaço multidimensional.
É calculada como:

$$d(\underline{x},\underline{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Análise de Cluster

DISTÂNCIA EUCLIDIANA

Nos comandos abaixo, calculamos a distância e geramos a matriz:

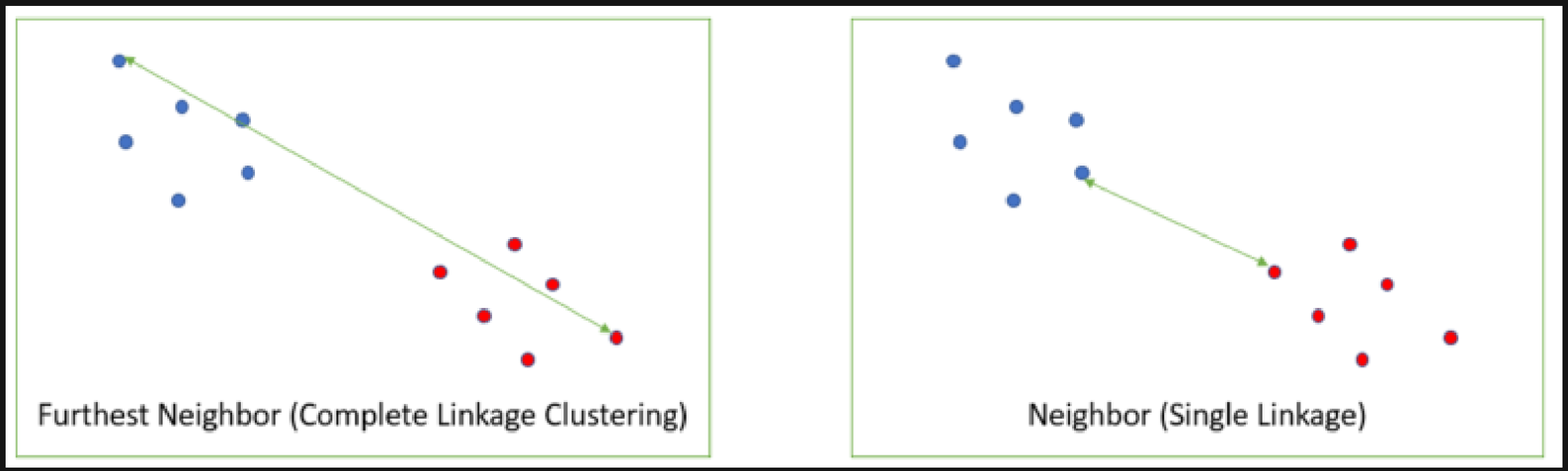
```
> ############ AGRUPAMENTO HIERÁRQUICO ############
> # gerando matriz de distância euclidiana
> df.dist = dist(df)
> df.dist
           1         2         3         4         5         6         7
2  0.8225538
3  2.5572204 1.9444585
           8         9        10        11        12        13        14
2
3
          15        16        17        18        19        20        21
2
3
          22        23        24        25        26        27        28
2
3
          29        30        31        32        33        34        35
2
3
          36        37        38        39        40        41        42
2
3
          43        44        45        46        47        48        49
2
3
          50        51        52        53        54        55        56
2
3
          57        58        59        60        61        62        63
2
3
          64        65        66        67        68        69        70
2
3
```

Análise de Cluster

MÉTODO DE AGRUPAMENTO HIERÁRQUICO

- Utilizamos a Distância Euclidiana e a Ligação Completa

Complete-Linkage: medida de similaridade entre dois clusters é definida pela maior distância de qualquer ponto do 1º cluster para qualquer ponto do 2º cluster (furthest neighbour).



Análise de Cluster

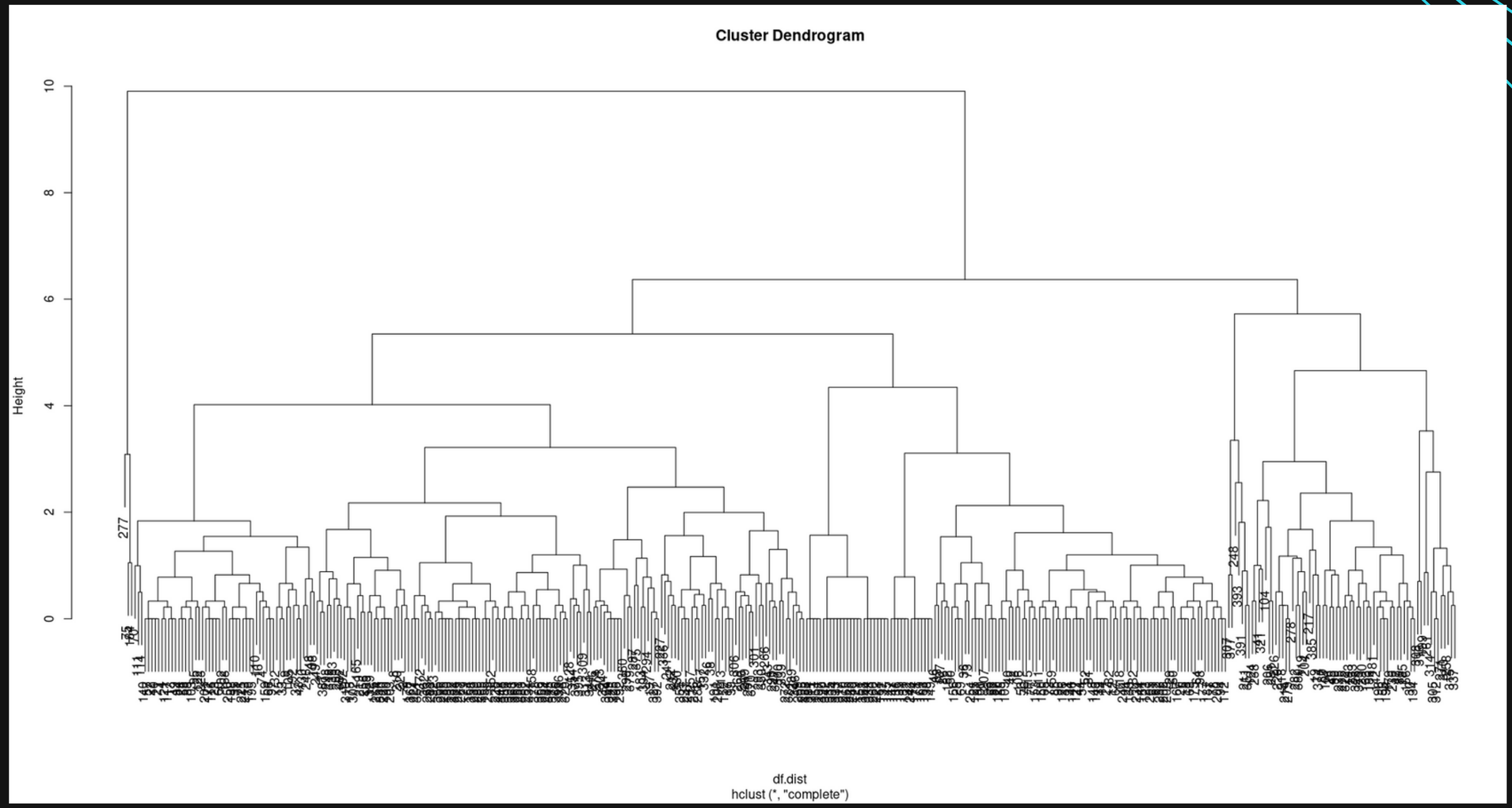
LIGAÇÃO COMPLETA

Nos comandos abaixo, utilizamos a função hclust, enviando como argumento a matriz calculada anteriormente. Essa função executa uma análise de cluster hierárquica usando um conjunto de diferenças para os objetos que estão sendo agrupados. Em cada estágio, as distâncias entre os agrupamentos são recalculadas pela fórmula de atualização de dissimilaridade de Lance-Williams de acordo com o método de agrupamento específico que está sendo usado.

```
# análise de clusters utilizando a ligação completa
df.hclust = hclust(df.dist)
plot(df.hclust)
```

Análise de Cluster

DENDROGRAMA



Análise de Cluster

Pelo dendrograma anterior, é possível observar que os clusters ficaram com muitas subdivisões. Portanto, para uma análise mais acurada, decidimos brecar a separação em apenas 5 clusters.

AGRUPAMENTO EM 5 CLUSTERS

Separando os registros em 5 clusters diferentes, e avaliando a quantidade de registros separados em cada um deles, tem-se:

```
> # selecionando apenas 5 grupos de clusters
> groups.5 = cutree(df.hclust, 5)
> table(groups.5)
groups.5
  1   2   3   4   5
199 125  61   3   7
```


199 registros no cluster 1, 125 no cluster 2, 61 no cluster 3, 3 no cluster 4 e 7 no cluster 5.

Análise de Cluster

Observando os registros individuais de cada cluster, tem-se:

- Cluster 1: Mais jovens (maioria entre 15 e 16 anos), pouquíssimas faltas (de 0 a 10, com muitos zeros e pouquíssimas acima de 6), muitas notas altas, indo de 10 a 20.
- Cluster 2: Majoritariamente mais jovens (de 15 a 16 anos), poucas faltas (de 0 a 10), muitas notas 0 e a maioria das notas mais baixas (com limite máximo 13)
- Cluster 3: Majoritariamente de idade mediana, entre 16 e 19 anos (maior quantidade entre 17 e 19), alto número de faltas (poucos registros abaixo de 10, vários acima deste, chegando em 28), notas medianas (de 5 a 18, mas poucas acima de 11)

Análise de Cluster

Observando os registros individuais de cada cluster, tem-se:

- Cluster 4: Majoritariamente idade mediana (de 16 a 17 anos), altíssimo número de faltas (variando entre 54 e 75), notas medianas (entre 8 e 11)
- Cluster 5: Idades mais altas (de 20 a 22 anos), poucas faltas (indo de 0 a 16, com pouquíssimas maiores que 10), notas medianas (de 5 a 15, com poucas acima de 10)

```
> medias <- aggregate(dados2, by=list(cluster=groups.5), mean)
> print(medias)
  cluster       age  absences        G3
1       1 16.92462  3.748744 12.788945
2       2 15.95200  2.184000  7.048000
3       3 17.06557 16.377049  9.622951
4       4 17.00000 61.666667  9.333333
5       5 20.14286  7.428571 10.428571
```


Médias dos valores das variáveis escolhidas a cada cluster, conforme análise descrita acima

Análise de Cluster

MÉTODO DE AGRUPAMENTO NÃO HIERÁRQUICO

- Utilizamos o algoritmo K-means

O algoritmo K-means atribui cada ponto de dados de entrada a um dos clusters minimizando a soma de quadrados dentro do cluster. Quando ele processa os dados de treinamento, o algoritmo K-means começa com um conjunto inicial de centroides escolhidos aleatoriamente.

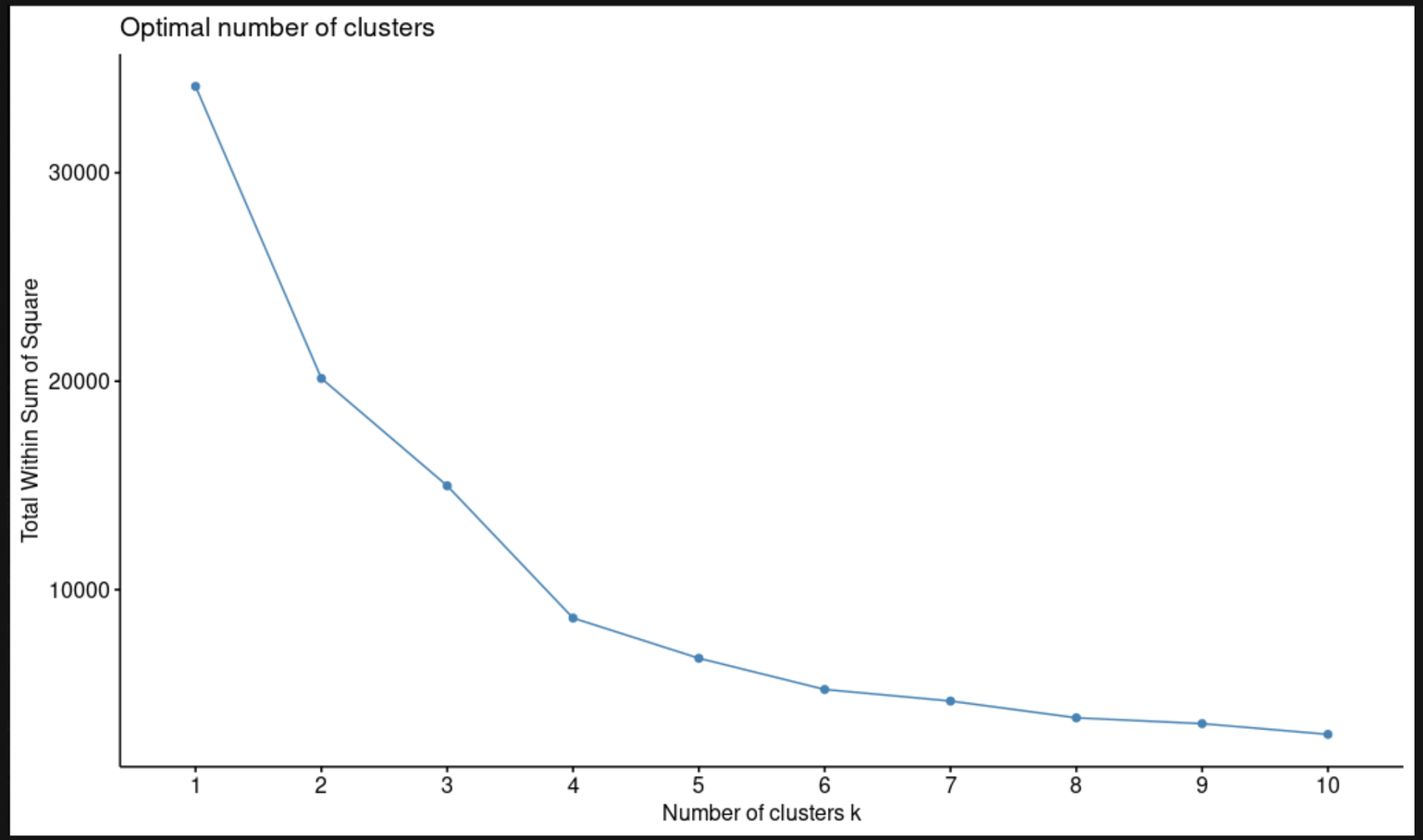
No código abaixo, utilizamos a biblioteca factoextra e o método para definir a melhor quantidade de clusters no agrupamento.

```
############ AGRUPAMENTO NÃO-HIERÁRQUICO ############
library(factoextra)

# método de soma de quadrados, para avaliar a quantidade de clusters
fviz_nbclust(dados2, kmeans, method = "wss")
```

Análise de Cluster

MÉTODO DE AGRUPAMENTO NÃO HIERÁRQUICO



Análise de Cluster

MÉTODO DE AGRUPAMENTO NÃO HIERÁRQUICO

Com base no gráfico da imagem anterior, é possível notar que o número de clusters se estabiliza em 5. A reta decai até o número 4 de clusters, e estabiliza no quinto, a partir do qual o número não altera de forma significativa na soma de quadrados interna (base do cálculo da distância euclidiana).

Gerando clusters com 5 grupos, tem-se:

```
> # cálculo do algoritmo k-means
> set.seed(123)
> km <- kmeans(df, 5, nstart = 25)
> km
K-means clustering with 5 clusters of sizes 5, 155, 39, 52, 144
```


5 registros no cluster 1, 155 no cluster 2, 39 no cluster 3, 52 no cluster 4 e 144 no cluster 5.

Análise de Cluster

Observando os registros individuais de cada cluster, tem-se:

- Cluster 1: Idades medianas (distribuição igual de 16 a 19 anos), altíssimo número de faltas (de 38 a 75), notas medianas, de 8 a 11.
- Cluster 2: Idades entre 16 e 17 anos, alto número de faltas (de 12 a 26), notas medianas, de 5 a 12
- Cluster 3: Majoritariamente de idade mediana, entre 16 e 19 anos (maior quantidade entre 17 e 19), baixo número de faltas (alto número de zeros, sem ultrapassar 10), notas medianas (de 8 a 15)

Análise de Cluster

MÉTODO DE AGRUPAMENTO NÃO HIERÁRQUICO

Observando os registros individuais de cada cluster, tem-se:

- Cluster 4: Majoritariamente idade mediana (de 16 a 17 anos), número de faltas mediano e notas medianas
- Cluster 5: Majoritariamente idade mediana (de 16 a 17 anos), número de faltas baixíssimo, com muitos zeros e notas medianas

```
> medias2 <- aggregate(dados2, by=list(cluster=km$cluster), mean)
> print(medias2)
  cluster      age    absences         G3
1       1 17.80000 52.60000000  9.4000000
2       2 15.51613  3.30322581 12.2709677
3       3 17.07692  0.05128205  0.1538462
4       4 17.01923 16.84615385  9.7115385
5       5 17.70833  4.18055556 11.4861111
```


Médias dos valores das variáveis escolhidas a cada cluster, conforme análise descrita acima

Conclusão

Ao comparar os clusters das duas análises, percebemos que:

- O cluster 4 da análise hierárquica equivale ao cluster 1 da análise não-hierárquica, pois são elementos agrupados por altíssimos números de faltas
- O cluster 2 da análise hierárquica corresponde ao 3 da não hierárquica, por ter os menores índices de faltas
- O cluster 5 da análise hierárquica equivale ao cluster 5 da não hierárquica, por registrar o mesmo intervalo de notas
- O cluster 1 da análise hierárquica contém os números medianos de faltas, assim como o cluster 2 da análise não hierárquica
- O cluster 3 da análise hierárquica equivale ao cluster 4 da não hierárquica, por possuírem a mesma média de número de faltas

Obrigada!