**Fundação Getúlio Vargas (FGV)**

**MBA Business Analytics e Big Data**

**Ana Paula Gonçalves dos Santos**

**Marcelo K. Costa**

**Trabalho apresentado para disciplina Métodos Matriciais e Análise de Clusters, como requisito para avaliação (Banco de Dados Cereals\_2). Foi utilizado setseed=23.**

**Professor Abraham Laredo Sicsu**

**Brasília (DF), 10 de fevereiro de 2020.**

**Sumário**

[1. Introdução 3](#_Toc32160629)

[2. Objetivo do estudo 3](#_Toc32160630)

[3. Análise e tratamento dos dados 3](#_Toc32160631)

[3.1 Análise univariada 3](#_Toc32160632)

[3.1.1 Missing values 4](#_Toc32160633)

[3.1.2 Valores Negativos 4](#_Toc32160634)

[3.1.3 Outliers 4](#_Toc32160635)

[3.1.4 Medidas descritivas 5](#_Toc32160636)

[4. Análise da correlação entre as variáveis quantitativas 6](#_Toc32160637)

[5. Cluster Analysis 7](#_Toc32160638)

[5.1 Considerando todas as variáveis como drivers 7](#_Toc32160639)

[5.2 Considerando apenas as variáveis quantitativas como drivers 11](#_Toc32160640)

[6. Conclusões 14](#_Toc32160641)

# Introdução

O presente trabalho é requisito de avaliação da disciplina Métodos Matriciais e Análise de Clusters e compreende o tratamento dos dados e análise por clusters do *data frame* “Cereals\_2”, disponibilizado pelo professor.

Será utilizada a Linguagem R para execução das análises apresentadas a seguir.

O código fonte do trabalho está disponível no endereço https://github.com/anapaulagsantos/clusteranalysis.

# Objetivo do estudo

O objetivo deste estudo é definir clusters dos tipos de cereais matinais com características nutricionais semelhantes.

# Análise e tratamento dos dados

O *data frame* Cereals\_2 fornece informações nutricionais de cereais matinais, sendo composto por 77 linhas e 12 colunas.

As colunas compreendem as seguintes variáveis, cujas descrições foram traduzidas para o Português.

|  |  |  |
| --- | --- | --- |
| **Variável** | **Descrição da Variável** | **Tipo de Variável** |
| **name** | nome do cereal | qualitativa |
| **mfr** | fabricante do cereal, onde A = American Home Food Products; G = General Mills; K = Kelloggs; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina | qualitativa |
| **type** | frio ou quente | qualitativa |
| **calories** | calorias por porção | quantitativa |
| **protein** | quantidade de proteínas em gramas | quantitativa |
| **fat** | quantidade de gorduras em gramas | quantitativa |
| **sodium** | quantidade de sódio em gramas | quantitativa |
| **fiber** | quantidade de fibras em gramas | quantitativa |
| **carbo** | quantidade de carboidratos em gramas | quantitativa |
| **sugars** | quantidade de açúcares em gramas | quantitativa |
| **potass** | quantidade de potássio em miligramas | quantitativa |
| **rating** | classificação dos cereais | quantitativa |

## 3.1 Análise univariada

A análise univariada compreende a análise de cada variável isoladamente. Para tanto, será avaliada, por variável, a existência de outliers e *missing values*, além das medidas descritivas e gráficos.

### 3.1.1 Missing values

Missing values são valores ausentes nas observações de determinada variável.

Não foram encontrados casos de missing value na base de dados fornecida.

### 3.1.2 Valores Negativos

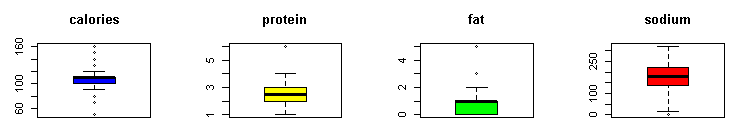
Foram identificados quatro (4) valores negativos em três (3) registros, nas variáveis “carbo”, “sugar” e “potass”, conforme tabela abaixo:

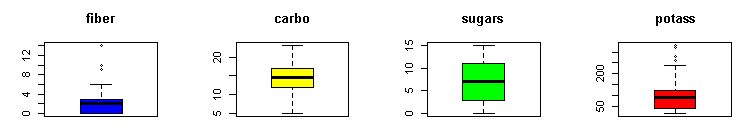
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **name** | **mfr** | **type** | **calories** | **protein** | **fat** | **sodium** | **fiber** | **carbo** | **sugars** | **potass** | **rating** |
| 1 | Almond\_Delight | R | C | 110 | 2 | 2 | 200 | 1.0 | 14 | 8 | -1 | 3.438.484 |
| 2 | Cream\_of\_Wheat\_(Quick) | N | H | 100 | 3 | 0 | 80 | 1.0 | 21 | 0 | -1 | 6.453.382 |
| 3 | Quaker\_Oatmeal | Q | H | 100 | 5 | 2 | 0 | 2.7 | -1 | -1 | 110 | 5.082.839 |

Como trata-se de valores nutricionais dos cereais, não é possível que haja valores negativos. Dessa forma, essas três linhas foram excluídas da análise.

### 3.1.3 Outliers

Outliers são observações que se afastam da maioria dos dados da série avaliada. Para identificação de outliers foram gerados gráficos do tipo boxplot das variáveis quantitativas:





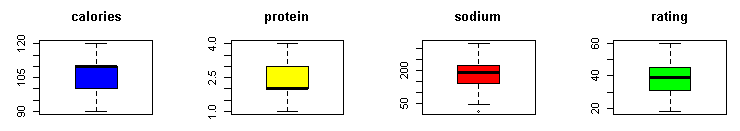


Observa-se outliers nas variáveis “calories”, “protein”, “fat”, “sodium”, “fiber”, “potass” e “rating”. Contudo, “fat”, “fiber” e “potass” foram consideradas assimétricas, não sendo realizado nenhum tratamento. Para as demais variáveis, optou-se pela exclusão dos outliers, totalizando 21 exclusões, conforme tabela e critérios definidos abaixo:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **name** | **mfr** | **type** | **calories** | **protein** | **fat** | **sodium** | **fiber** | **carbo** | **sugars** | **potass** | **rating** |
| 1 | 100%\_Bran | N | C | 70 | 4 | 1 | 130 | 10.0 | 5.0 | 6 | 280 | 6.840.297 |
| 2 | All-Bran | K | C | 70 | 4 | 1 | 260 | 9.0 | 7.0 | 5 | 320 | 5.942.551 |
| 3 | All-Bran\_with\_Extra\_Fiber | K | C | 50 | 4 | 0 | 140 | 14.0 | 8.0 | 0 | 330 | 9.370.491 |
| 4 | Basic\_4 | G | C | 130 | 3 | 2 | 210 | 2.0 | 18.0 | 8 | 100 | 3.703.856 |
| 5 | Cheerios | G | C | 110 | 6 | 2 | 290 | 2.0 | 17.0 | 1 | 105 | 5.076.500 |
| 6 | Frosted\_Mini-Wheats | K | C | 100 | 3 | 0 | 0 | 3.0 | 14.0 | 7 | 100 | 5.834.514 |
| 7 | Just\_Right\_Fruit\_&\_Nut | K | C | 140 | 3 | 1 | 170 | 2.0 | 20.0 | 9 | 95 | 3.647.151 |
| 8 | Maypo | A | H | 100 | 4 | 1 | 0 | 0.0 | 16.0 | 3 | 95 | 5.485.092 |
| 9 | Muesli\_Raisins,\_Dates,\_&\_Almonds | R | C | 150 | 4 | 3 | 95 | 3.0 | 16.0 | 11 | 170 | 3.713.686 |
| 10 | Muesli\_Raisins,\_Peaches,\_&\_Pecans | R | C | 150 | 4 | 3 | 150 | 3.0 | 16.0 | 11 | 170 | 3.413.976 |
| 11 | Mueslix\_Crispy\_Blend | K | C | 160 | 3 | 2 | 150 | 3.0 | 17.0 | 13 | 160 | 3.031.335 |
| 12 | Nutri-Grain\_Almond-Raisin | K | C | 140 | 3 | 2 | 220 | 3.0 | 21.0 | 7 | 130 | 4.069.232 |
| 13 | Oatmeal\_Raisin\_Crisp | G | C | 130 | 3 | 2 | 170 | 1.5 | 13.5 | 10 | 120 | 3.045.084 |
| 14 | Puffed\_Rice | Q | C | 50 | 1 | 0 | 0 | 0.0 | 13.0 | 0 | 15 | 6.075.611 |
| 15 | Puffed\_Wheat | Q | C | 50 | 2 | 0 | 0 | 1.0 | 10.0 | 0 | 50 | 6.300.565 |
| 16 | Raisin\_Squares | K | C | 90 | 2 | 0 | 0 | 2.0 | 15.0 | 6 | 110 | 5.533.314 |
| 17 | Shredded\_Wheat | N | C | 80 | 2 | 0 | 0 | 3.0 | 16.0 | 0 | 95 | 6.823.588 |
| 18 | Shredded\_Wheat\_'n'Bran | N | C | 90 | 3 | 0 | 0 | 4.0 | 19.0 | 0 | 140 | 7.447.295 |
| 19 | Shredded\_Wheat\_spoon\_size | N | C | 90 | 3 | 0 | 0 | 3.0 | 20.0 | 0 | 120 | 7.280.179 |
| 20 | Special\_K | K | C | 110 | 6 | 0 | 230 | 1.0 | 16.0 | 3 | 55 | 5.313.132 |
| 21 | Total\_Raisin\_Bran | G | C | 140 | 3 | 1 | 190 | 4.0 | 15.0 | 14 | 230 | 2.859.278 |

1. Calories: foram mantidas as observações maiores que 80 e menores que 130 calorias.
2. Protein: foram excluídas as observações iguais a 6.
3. Sodium: foram excluídas as observações iguais a 0.
4. Rating: foram excluídas as observações maiores ou iguais a 93,7.

Após as exclusões, os gráficos dessas variáveis ficaram da seguinte forma:



### 3.1.4 Medidas descritivas

Após os tratamentos listados nos itens 3.1.2 e 3.1.3, restaram 53 observações (linhas) e 12 variáveis (colunas). Segue a sumarização dos dados por variável, extraída do software R.

name mfr type calories

100%\_Natural\_Bran : 1 G:18 C:53 Min. : 90.0

Apple\_Cinnamon\_Cheerios: 1 K:15 1st Qu.:100.0

Apple\_Jacks : 1 N: 1 Median :110.0

Bran\_Chex : 1 P: 9 Mean :107.9

Bran\_Flakes : 1 Q: 5 3rd Qu.:110.0

Cap'n'Crunch : 1 R: 5 Max. :120.0

(Other) :47

protein fat sodium fiber

Min. :1.000 Min. :0 Min. : 15.0 Min. :0.000

1st Qu.:2.000 1st Qu.:0 1st Qu.:140.0 1st Qu.:0.000

Median :2.000 Median :1 Median :190.0 Median :1.000

Mean :2.189 Mean :1 Mean :181.3 Mean :1.651

3rd Qu.:3.000 3rd Qu.:1 3rd Qu.:220.0 3rd Qu.:3.000

Max. :4.000 Max. :5 Max. :320.0 Max. :6.000

carbo sugars potass rating

Min. : 8.00 Min. : 2.000 Min. : 20.00 Min. :18.04

1st Qu.:12.00 1st Qu.: 3.000 1st Qu.: 35.00 1st Qu.:31.23

Median :14.00 Median : 8.000 Median : 60.00 Median :39.24

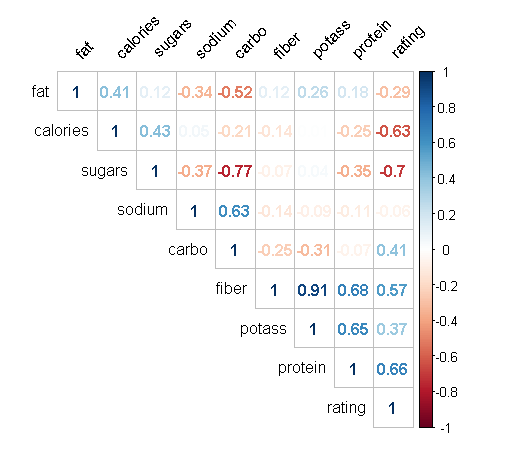
Mean :14.67 Mean : 7.774 Mean : 81.13 Mean :38.25

3rd Qu.:17.00 3rd Qu.:12.000 3rd Qu.:110.00 3rd Qu.:45.33

Max. :23.00 Max. :15.000 Max. :260.00 Max. :59.64

# 4. Análise da correlação entre as variáveis quantitativas

Para visualização das correlações foi utilizada a biblioteca “corrplot”. Para Interpretação da matriz abaixo, considere que quanto maior o círculo maior a correlação entre as variáveis. Além disso, quanto mais azul escuro, mais próxima a correlação fica de 1, que significa que além de forte a correlação é positiva. De mesmo modo, quanto mais próximo de vermelho escuro, mais próxima a correlação fica de -1, que significa que além de forte a correlação é negativa.



Observa-se alta correlação entre as variáveis fibra e potássio (0,91).

Considerando que a fibra é um componente nutricional muito valorizado pelos nutricionistas, optou-se por excluir a variável potássio.

# Cluster Analysis

O objetivo desta seção é utilizar técnicas para agrupamento dos cereais matinais em grupos com características nutricionais semelhantes.

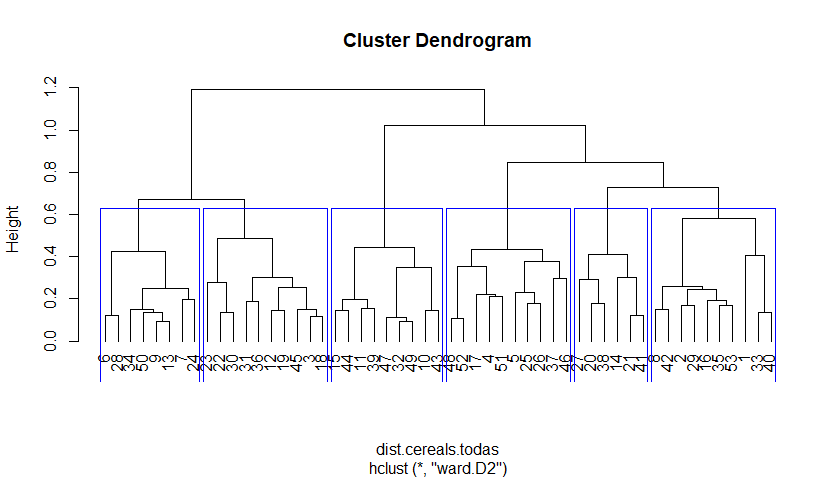
No primeiro experimento, serão utilizadas todas as variáveis e, no segundo, apenas as variáveis quantitativas.

## Considerando todas as variáveis como drivers

Como essa análise envolve variáveis qualitativas e quantitativas, foi utilizada a função “daisy” do pacote “Cluster” para cálculo da matriz de distâncias, não sendo necessário padronizar as variáveis porque a função faz isso por meio da métrica de Gower.

Para definição do número de cluster foram utilizadas as técnicas abaixo:

1. **Algoritmo hierárquico aglomerativo (função hclust):** para ter uma ideia visual do número de clusters. De acordo com o dendograma abaixo, verificam-se 6 clusters.



1. **Algorimo k-medoid (função pamk do pacote fpc):** ajuda a estabelecer o número de clusters a partir de técnicas matemáticas com ASW (para os casos de variáveis qualitativas). A critério da equipe, definiu-se que o algoritmo avaliaria a utilização de 2 a 6 clusters.

Conforme resultado abaixo, o algoritmo mostra como “ideal” a utilização de 4 clusters.

2 clusters 0.1741282

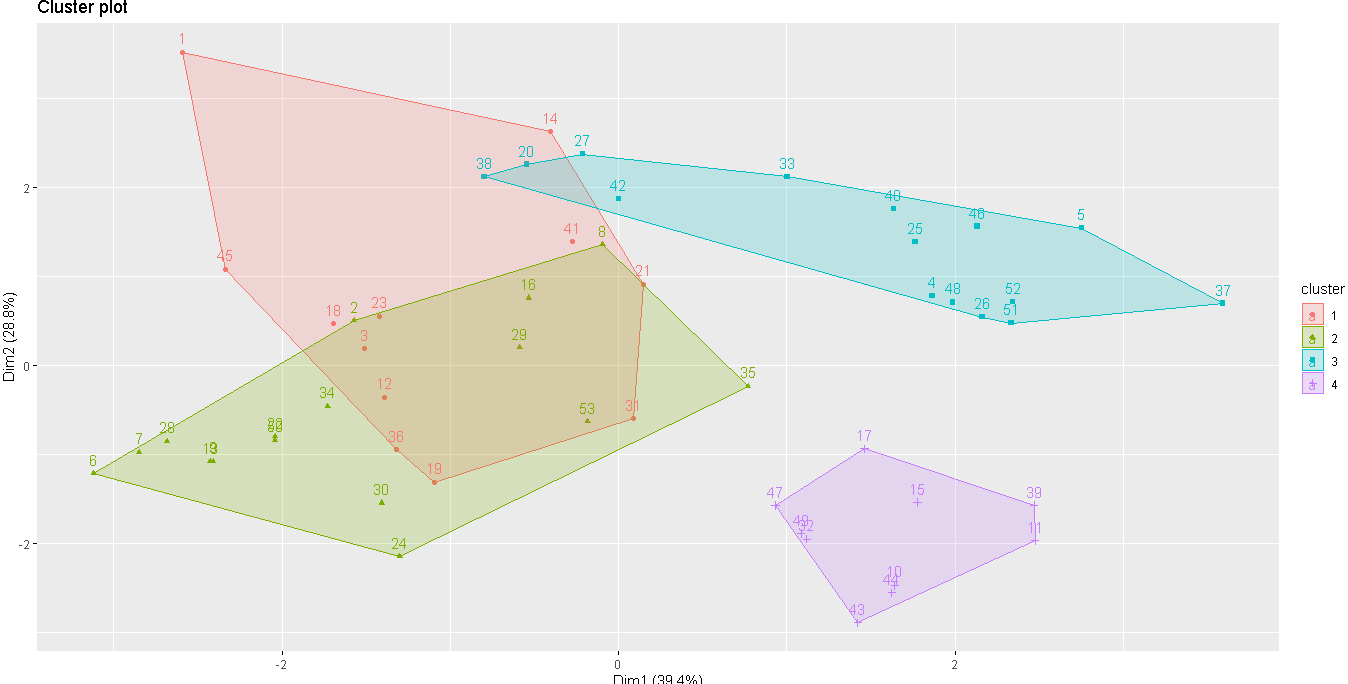
3 clusters 0.1297191

4 clusters 0.1909146

5 clusters 0.1569641

6 clusters 0.1759362

1. **Projeção dos 4 clusters**:



1. **Marcas e tipos em cada cluster:**

* Quanto às marcas:

**1 2 3 4**

American Home Food Products 0 0 0 0

General Mills 0 12 3 3

Kelloggs 10 0 1 4

Nabisco 0 0 1 0

Post 1 2 6 0

Quaker Oats 1 2 2 0

Ralston Purina 0 0 2 3

**Cluster 1:** formado prioritariamente por cereais da marca Kelloggs.

**Cluster 2:** formado prioritariamente por cereais da marcaGeneral Mills.

**Cluster 3:** sem marcas predominantes, porém com maior quantidade de cereais da marca Post.

**Cluster 4:** sem marcas predominantes.

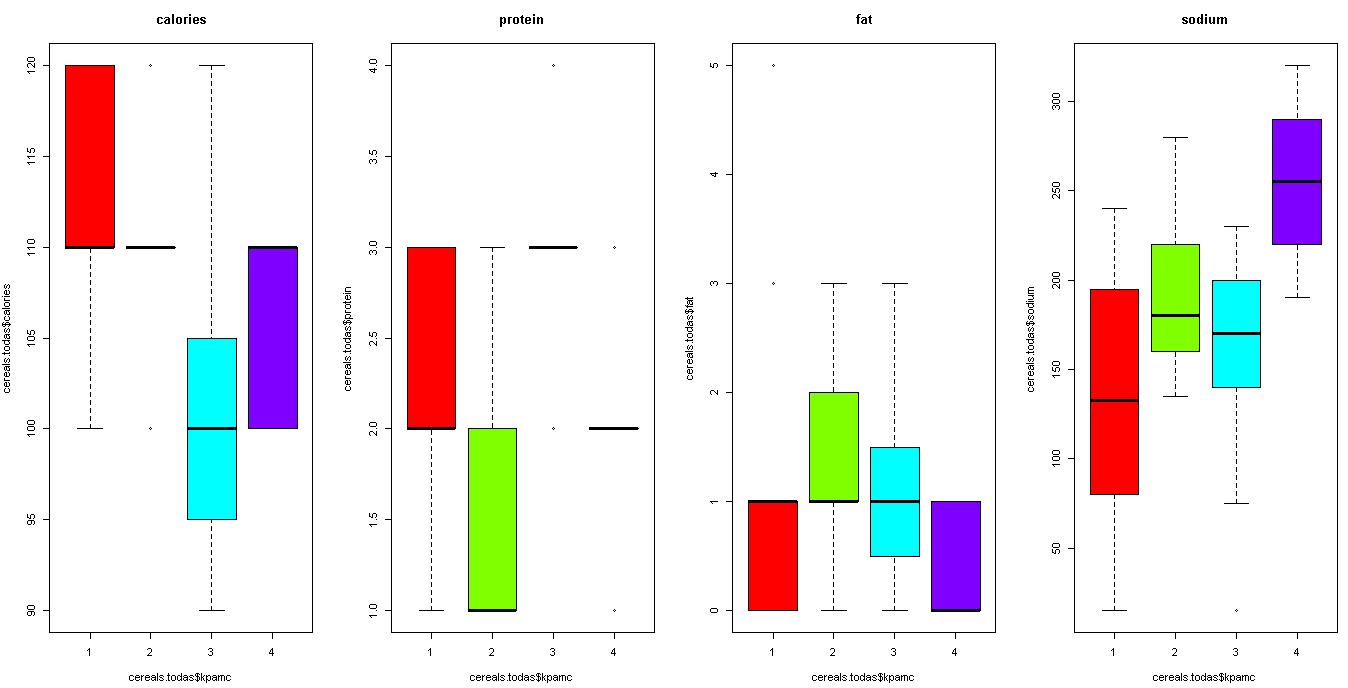
* Quanto ao tipo: todos são frios.

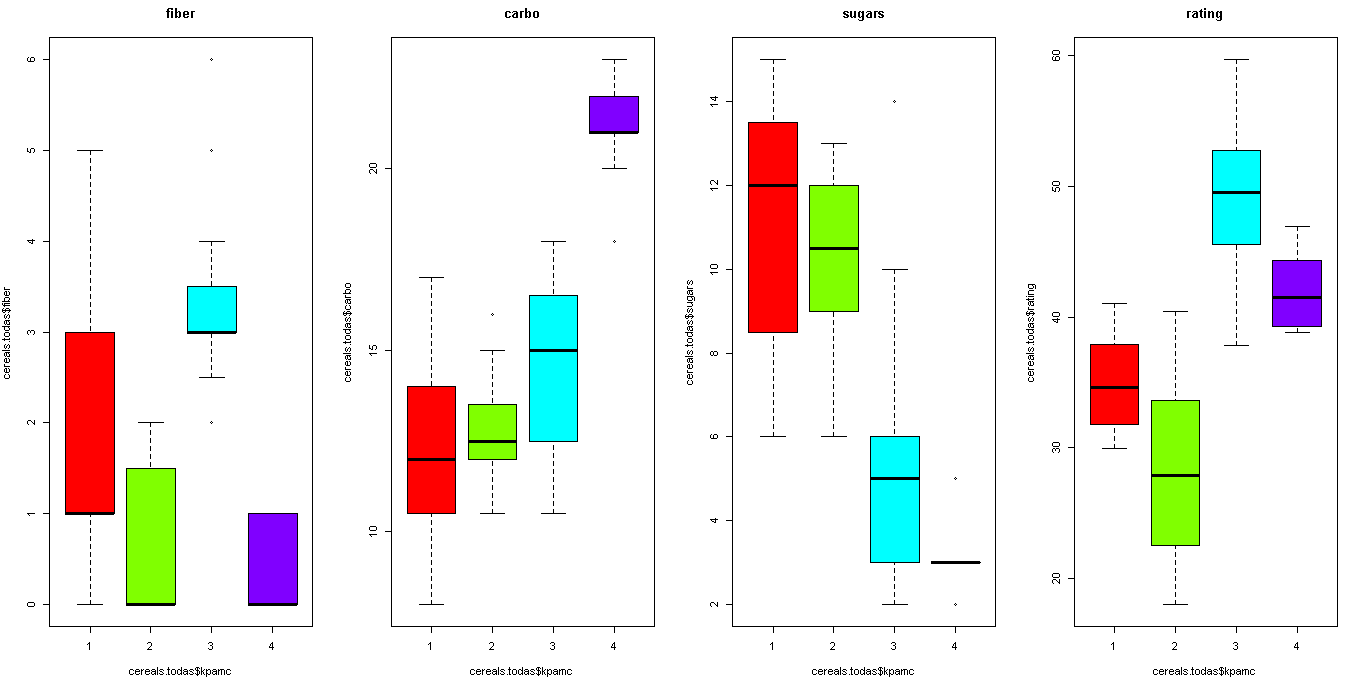
1 2 3 4

cold 12 16 15 10

hot 0 0 0 0

1. **Caracterização dos cluster:**





**Cluster 1 – Mais calóricos e com menos gorduras:**

* Destaca-se por conter os cereais com mais calorias e açúcares;
* Possui menos gorduras que os clusters 2 e 3;
* É mediano na quantidade de fibras e proteínas.
* Rating mediano.

**Cluster 2 – Poucas fibras e com mais gorduras:**

* Destaca-se por conter os cereais com mais gorduras e poucas fibras;
* Os cereais apresentam a mesma quantidade de calorias e proteínas;
* É o segundo cluster em quantidade de açúcares;
* Tem poucas fibras e proteínas;
* É o pior no rating.

**Cluster 3 – Saudável:**

* Destaca-se por conter os cereais com menor quantidade de calorias, gorduras e açúcares e a maior quantidade de fibras;
* É mediano na quantidade de carboidratos;
* É o melhor avaliado no rating.

**Cluster 4 – Rico em sódio e baixo teor de fibras:**

* Destaca-se por conter os cereais com a maior quantidade de sódio e carboidratos e a menor quantidade de fibras;
* Possui baixa quantidade de gorduras;
* Mediano em quantidade de calorias;
* É o segundo melhor avaliado no rating.

1. **Comparação entre Algoritmo hierárquico aglomerativo x Algoritmo k-medoid**

Os resultados obtidos com as duas técnicas mostraram resultados bastantes divergentes, conforme descrito abaixo:

No caso do algoritmo hierárquico aglomerativo, a distribuição foi a seguinte:

**1 2 3 4**

16 18 10 9

Para o k-medoid, por sua vez, a distribuição foi a seguinte:

**1 2 3 4**

12 16 15 10

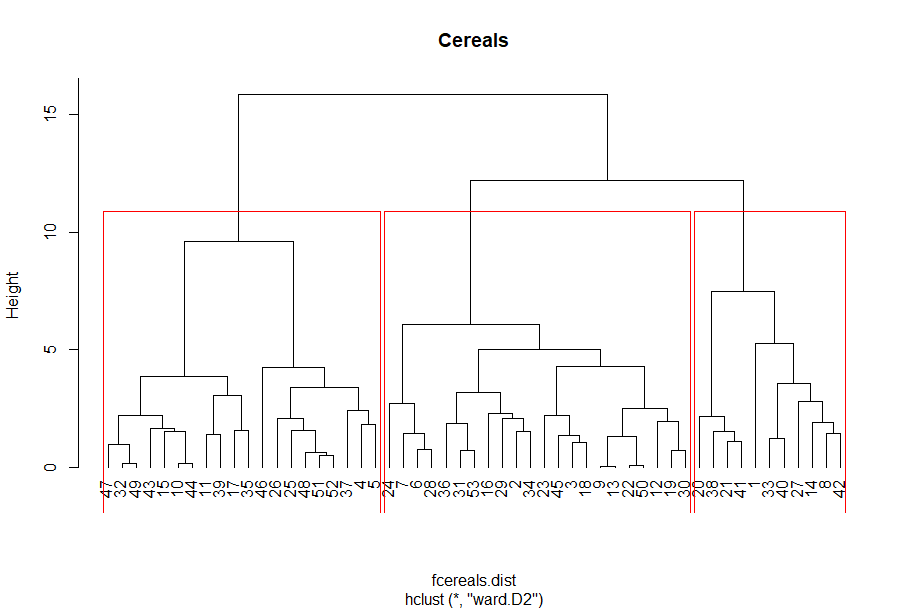
Adicionalmente, confirmando a divergência entre os agrupamentos realizados pelas duas técnicas, a função clust\_stats apontou o resultado de 0.3660209 (valor bem abaixo de 1, indicando pouca similaridade).

## 5.2 Considerando apenas as variáveis quantitativas como drivers

Como essa análise envolve apenas variáveis quantitativas, foi utilizada a função “dist” do pacote “Cluster” para cálculo da matriz de distâncias. Para padronização das variáveis utilizou-se a função “scale”.

Para definição do número de cluster foram utilizadas as técnicas abaixo:

1. **Algoritmo hierárquico aglomerativo (função hclust):** para ter uma ideia visual do número de clusters. De acordo com o dendograma abaixo, verificam-se 3 clusters.



1. **Algorimo NBclust:** ajuda a estabelecer o número de clusters a partir de técnicas matemáticas para os casos de variáveis quantitativas. A critério da equipe, definiu-se que o algoritmo avaliaria a utilização de 2 a 12 clusters.

Conforme resultado abaixo, o algoritmo mostra como “ideal” a utilização de 3 clusters:

\* Among all indices:

\* 1 proposed 2 as the best number of clusters

\* 6 proposed 3 as the best number of clusters

\* 5 proposed 4 as the best number of clusters

\* 6 proposed 5 as the best number of clusters

\* 1 proposed 6 as the best number of clusters

\* 1 proposed 10 as the best number of clusters

\* 3 proposed 12 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 3

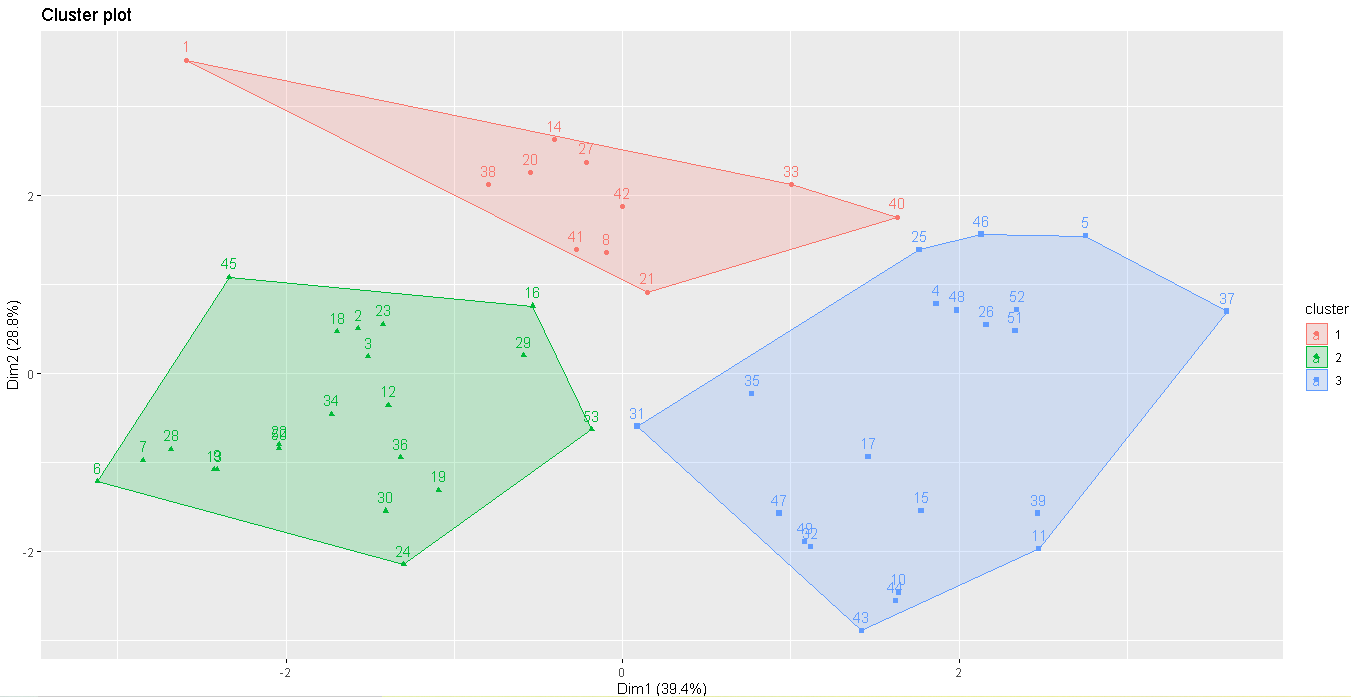
1. **Tamanho dos clusters:** utilizando a função K-means, foram gerados 2 clusters com 21 observações e 1 cluster com 11 observações.

> kmn <- kmeans(fcereals.scale,3,nstart=25)

> kmn$size

[1] 11 21 21

1. **Projeção dos 3 clusters**:



1. **Marcas e tipos em cada cluster:**

* Quanto às marcas:

**1 2 3**

General Mills 2 10 6

Kelloggs 3 6 6

Nabisco 0 0 1

Post 3 3 3

Quaker Oats 3 2 0

Ralston Purina 0 0 5

**Cluster 1:** sem marcas predominantes.

**Cluster 2:** formado prioritariamente por cereais da marcaGeneral Mills e Kelloggs.

**Cluster 3:** maior quantidade de cereais das marcas General Mills, Kelloggs e Ralston Purina.

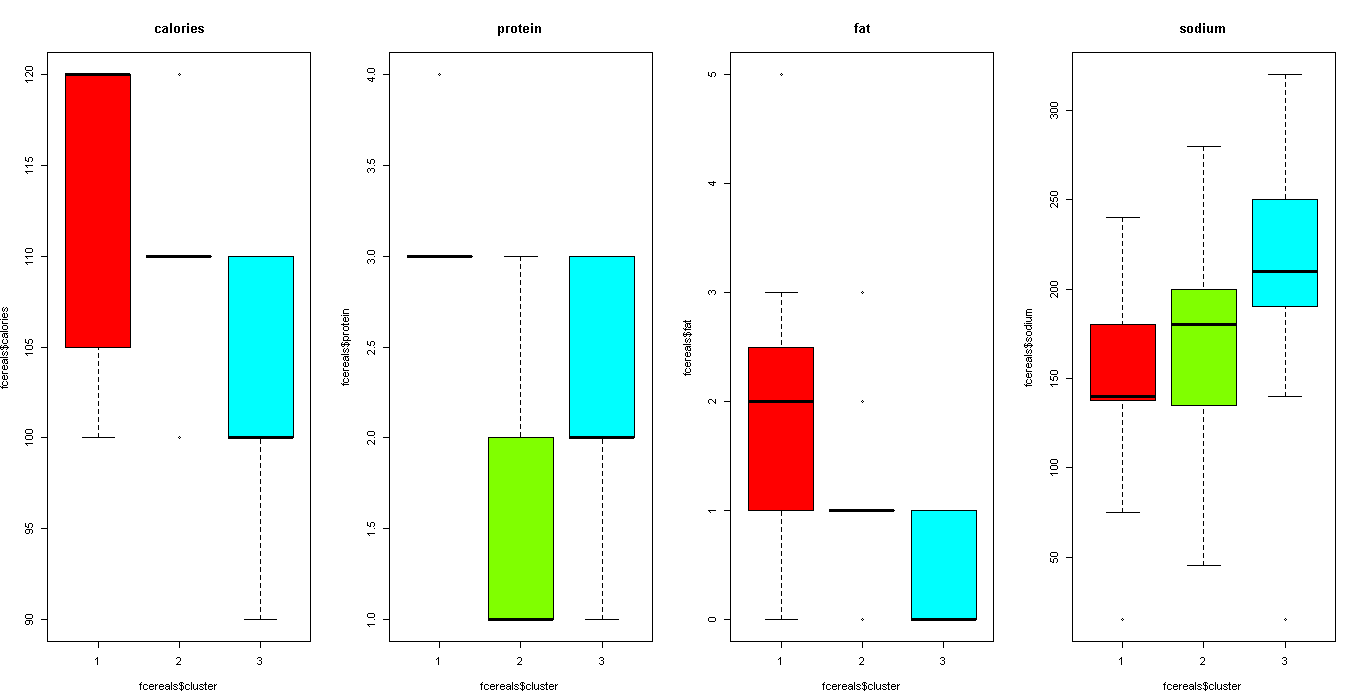
* Quanto ao tipo: todos são frios.

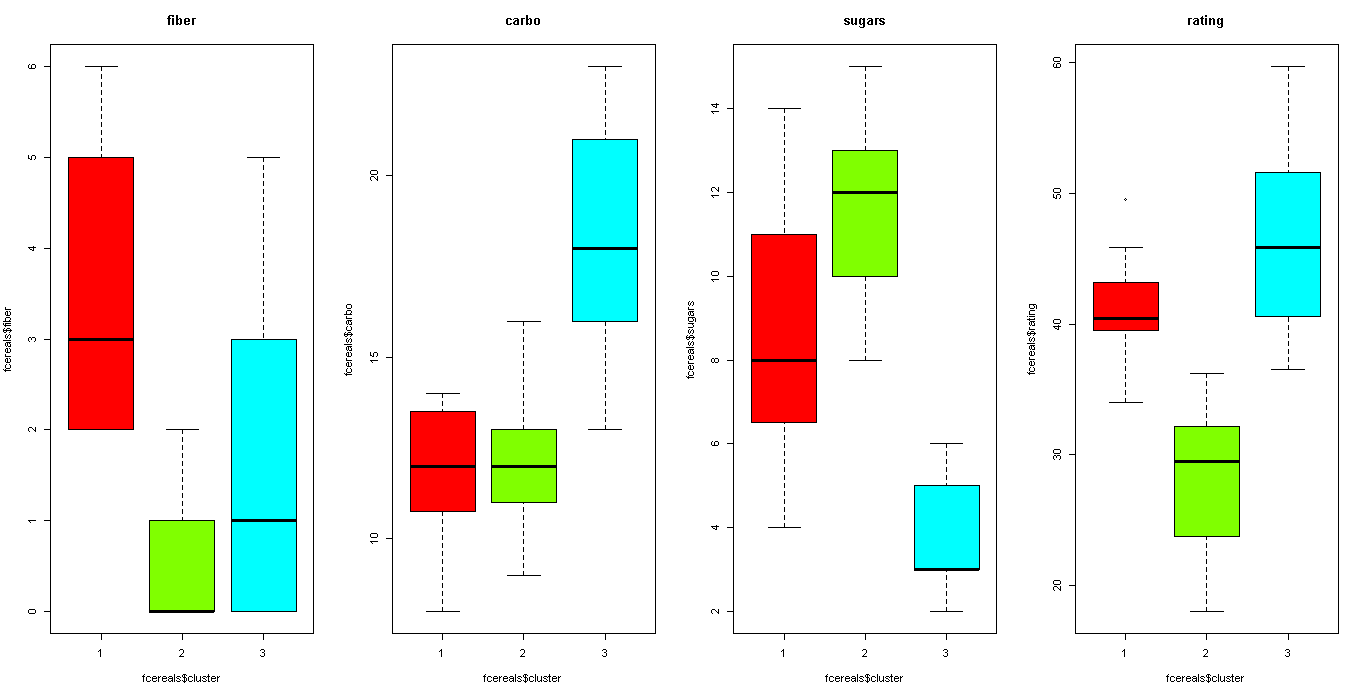
1 2 3

cold 11 21 21

hot 0 0 0

1. **Caracterização dos cluster:**





**Cluster 1 – o mais calórico:**

* Destaca-se por conter os cereais com mais calorias, gorduras e fibras;
* Os cereais apresentam a mesma quantidade de proteínas;
* É mediano com relação ao rating, açúcares e sódio.

**Cluster 2 – o mais “açucarado”:**

* Destaca-se por conter os cereais com a maior quantidade de açucares e menor quantidade de fibras e proteínas;
* Os cereais apresentam a mesma quantidade de gorduras;
* É semelhante ao cluster 1 em quantidade de carboidratos;
* Tem o pior rating.

**Cluster 3 – o melhor rating:**

* Destaca-se pelo melhor rating;
* Contém os cereais com maior quantidade de sódio e carboidratos e menor quantidade de açúcares e gorduras.

1. **Comparação entre Algoritmo hierárquico aglomerativo x Algoritmo k-means**

Os resultados obtidos com as duas técnicas mostraram resultados convergentes, conforme descrito abaixo:

No caso do algoritmo hierárquico aglomerativo, a distribuição foi a seguinte:

**1 2 3**

11 22 20

Para o k-means, por sua vez, a distribuição foi a seguinte:

**1 2 3**

11 21 21

Adicionalmente, confirmando a similaridade entre os agrupamentos realizados pelas duas técnicas, a função clust\_stats apontou o resultado de 0.9341728 (valor próximo de 1).

# Conclusões

Pela avaliação dos resultados obtidos com os dois experimentos, utilizando todas as variáveis do *dataframe* e apenas as quantitativas, conclui-se que os agrupamentos fazem mais sentido quando utilizamos somente variáveis quantitativas.

Além disso, a distribuição dos clusters quando somente as variáveis quantitativas são utilizadas fica mais definida. Já quando são utilizadas todas as variáveis há muita sobreposição entre os clusters.