**FGV - FUNDAÇÃO GETÚLIO VARGAS  
MBA Executivo em Business Analytcs e Big Data**

**ALUNOS**

Guilherme Nakano Nalin  
Paulo Gustavo Maciel Lopes  
Patrícia Yoshie Kishi Bueno

**Projeto Análise de Agrupamentos**

Trabalho apresentado para obtenção de nota da disciplina Métodos Matriciais e Análise de Cluster  
Professor Abrahan Laredo Sicsu

São Paulo/SP  
Unidade Paulista  
Janeiro de 2021

**1. INTRODUÇÃO**

Foi disponibilizado a base de dados “*MINIMARKET*” para este Projeto de Análise de Agrupamentos.

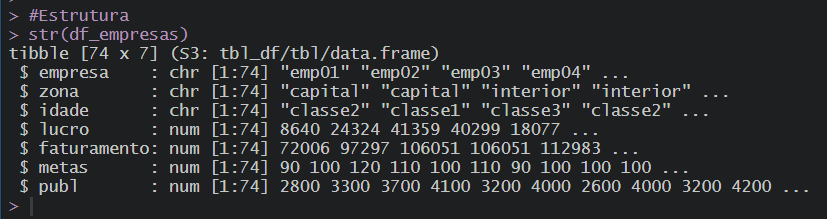
A base contém dados sobre empresas da capital e do interior com classe de idade e valores de faturamento, lucro, metas e público. Composta por 74 observações e 7 variáveis, sendo consideradas drivers: “zona”, “idade”, “lucro”, “faturamento”, “metas” e “publ” para a realização desse estudo.

A variável “idade” está classificada em:

* Classe1 até 5 anos;
* Classe2 6 a 10 anos;
* Classe3 11 anos ou mais.

**2. ANÁLISE DOS DRIVERS**

Analisamos a estrutura dos dados do *dataset* para identificação dos tipos das variáveis e possíveis transformações:

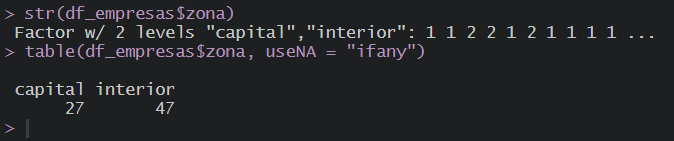


**2.1. Variável “zona”**

Do tipo *character* transformada em *factor* para ser classificada como qualitativa nominal.

Analisado a inexistência de *missing values* ou *outliers* a tratar.

Verificado um maior número de observações 63,5% para o “interior”.

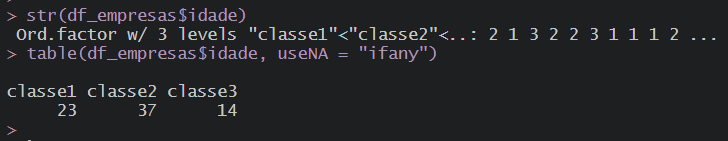


**2.2. Variável “idade”**

Do tipo *character* transformada em *ordered factor* para ser classificada como qualitativa ordinal.

Verificado a inexistência de *missing values* ou *outliers* a tratar.

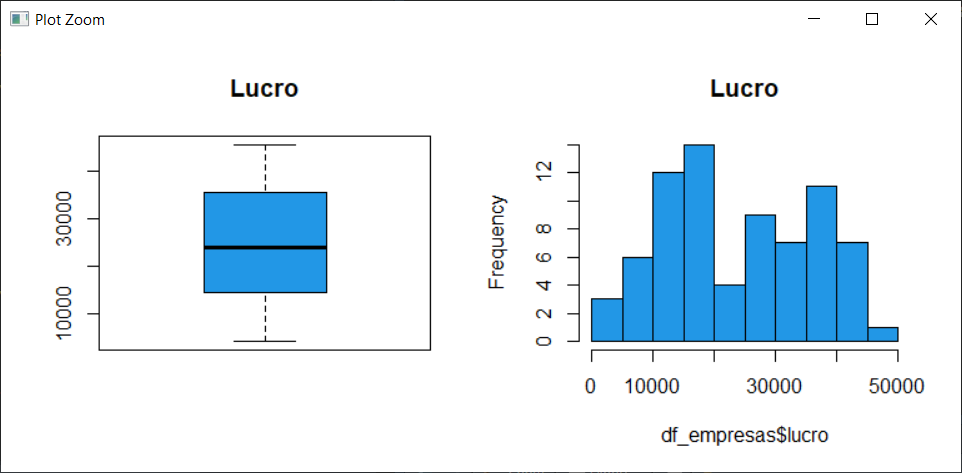
Analisado um maior número de observações 50% para a idade de “classe2”.



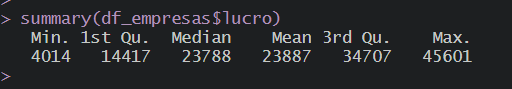
**2.3. Variável “lucro”**

Do tipo *numeric* e sem transformação, classificada como quantitativa ordinal.

Verificado a inexistência de *missing values* ou *outliers* a tratar, como pode ser observado nos gráficos abaixo:



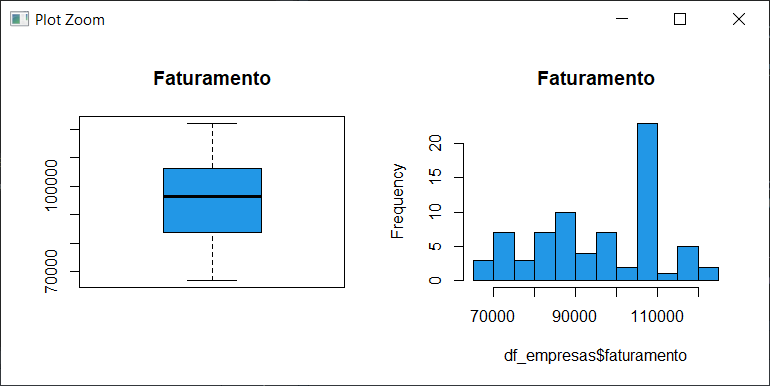
Analisando os dados observamos uma dispersão simétrica dos valores do “lucro” entre 14417 e 34707 com mediana de 23788.



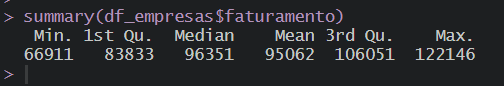
**2.4. Variável “faturamento”**

Do tipo *numeric* e sem transformação, classificada como quantitativa ordinal.

Verificado a inexistência de *missing values* ou *outliers* a tratar, como pode ser observado nos gráficos abaixo:



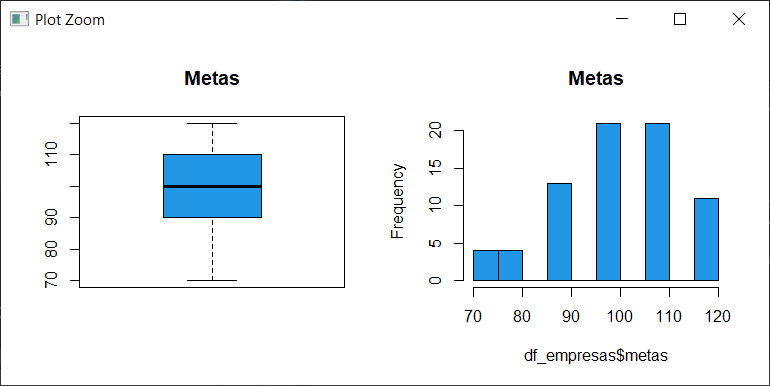
Analisando os dados observamos uma dispersão assimétrica à esquerda dos valores do “faturamento” entre 83833 e 106051 com mediana de 96351 e um pico elevado no valor 106051.



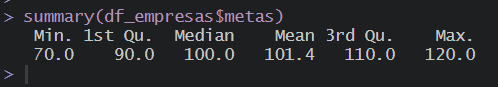
**2.5. Variável “metas”**

Do tipo *numeric* e sem transformação, classificada como quantitativa ordinal.

Verificado a inexistência de *missing values* ou *outliers* a tratar, como pode ser observado nos gráficos abaixo:



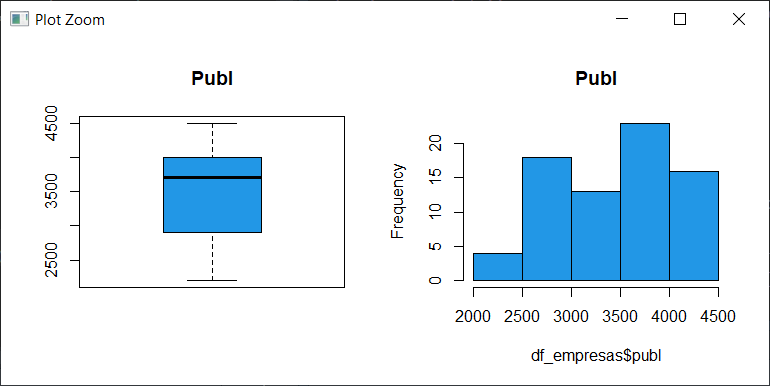
Analisando os dados observamos uma dispersão simétrica dos valores da “metas” entre 90 e 110 com mediana de 100.



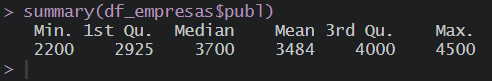
**2.. Variável “publ”**

Do tipo *numeric* e sem transformação, classificada como quantitativa ordinal.

Verificado a inexistência de *missing values* ou *outliers* a tratar, como pode ser observado nos gráficos abaixo:

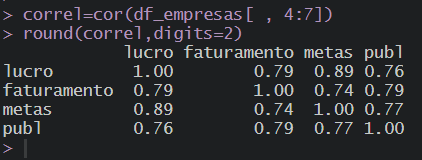


Analisando os dados observamos uma dispersão assimétrica à direita dos valores de “publ” entre 2925 e 4000 com mediana de 3700.



**3. MATRIZ DE CORRELAÇÃO**

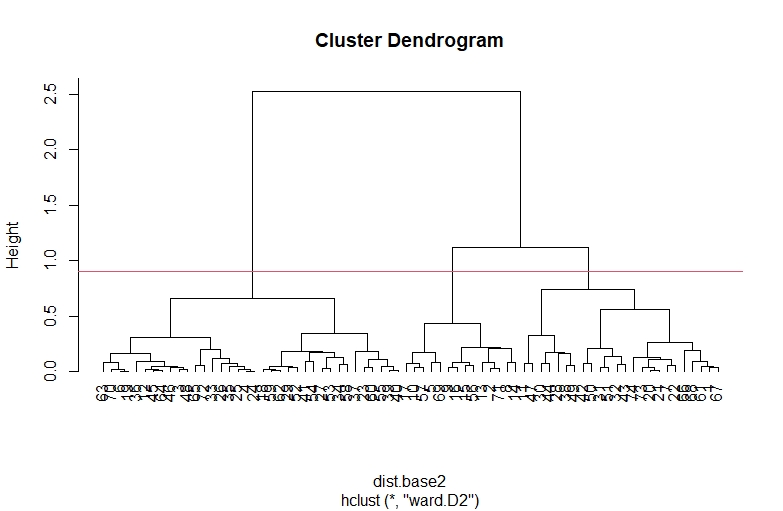
Observamos que os valores de correlação das variáveis são baixos não possuindo relação linear entre si, assim podemos seguir com todas as variáveis nas análises.



**4. MÉTODO HIERÁRQUICO AGLOMERATIVO**

Antes da execução do cálculo das distâncias entre todas as amostras do *dataset*, foi construído um vetor para atribuição de peso às variáveis. Para as variáveis *drivers* qualitativas, o peso atribuído foi 0.5, enquanto que para as variáveis quantitativas o peso atribuído foi 1. Esse passo tem o objetivo de evitar que a formação dos *clusters* sofra muita influência das variáveis qualitativas.

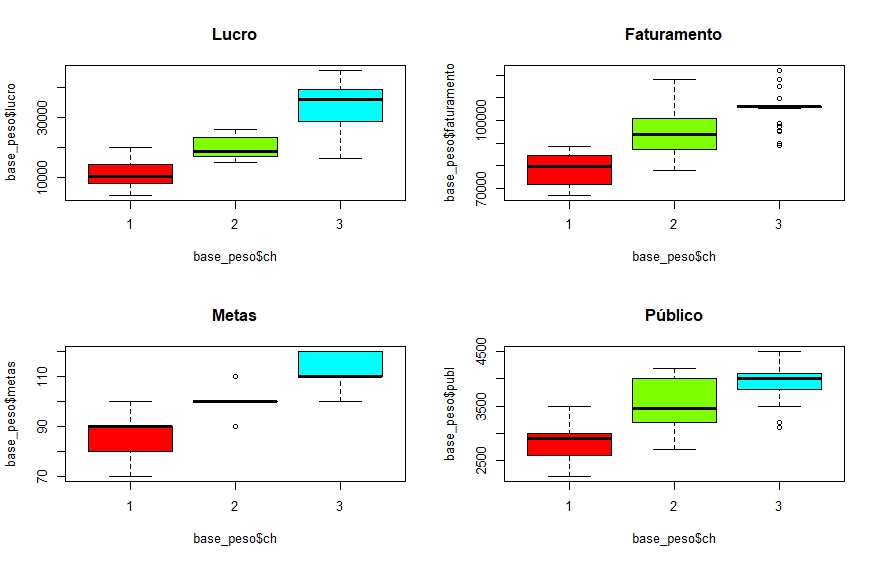
A partir do cálculo das distâncias das amostras, foi executado o algoritmo de criação de *clusters* com o método de *Ward*. Foi obtido o seguinte dendrograma:



Com base no dendrograma acima, foi escolhido a quebra em 3 *clusters* (conforme ponto de corte mostrado acima).

**4.1 Caracterização e comparação dos clusters**

Os *boxplots* a seguir ilustram os 3 *clusters* criados a partir das variáveis *drivers* quantitativas:

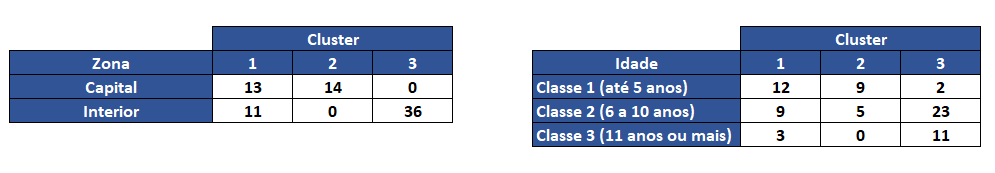


O *cluster* 1 ficou com as empresas que possuem as menores métricas: menor lucro, menor faturamento, menor meta e menor público.

O *cluster* 3, por outro lado, ficou com as empresas com as maiores métricas: maior lucro, maior faturamento, maior meta e maior público.

O *cluster* 2 é o meio termo, possui empresas cujas métricas são intermediárias, menores que no *cluster* 3 e maiores que no *cluster* 1. Há apenas uma pequena sobreposição dos valores de público com *cluster* 3.

Pelo ponto de vista das variáveis quantitativas, os *clusters* criados apresentam boa heterogeneidade. Agora, serão analisadas através das variáveis *drivers* qualitativas:



O *cluster* 1 contém empresas tanto na Capital como no Interior e, a maior parte delas (87,5%) possuem idade de até 10 anos. Este *cluster* é o que apresenta a maior quantidade de empresas jovens (até 5 anos).

O *cluster* 2 apresenta apenas empresas localizadas na Capital e é o único que não possui empresas com 11 anos ou mais de idade.

O *cluster* 3 possui a maior quantidade de empresas, todas localizadas no Interior. Praticamente todas as empresas (94,4%) possuem 6 anos de idade ou mais.

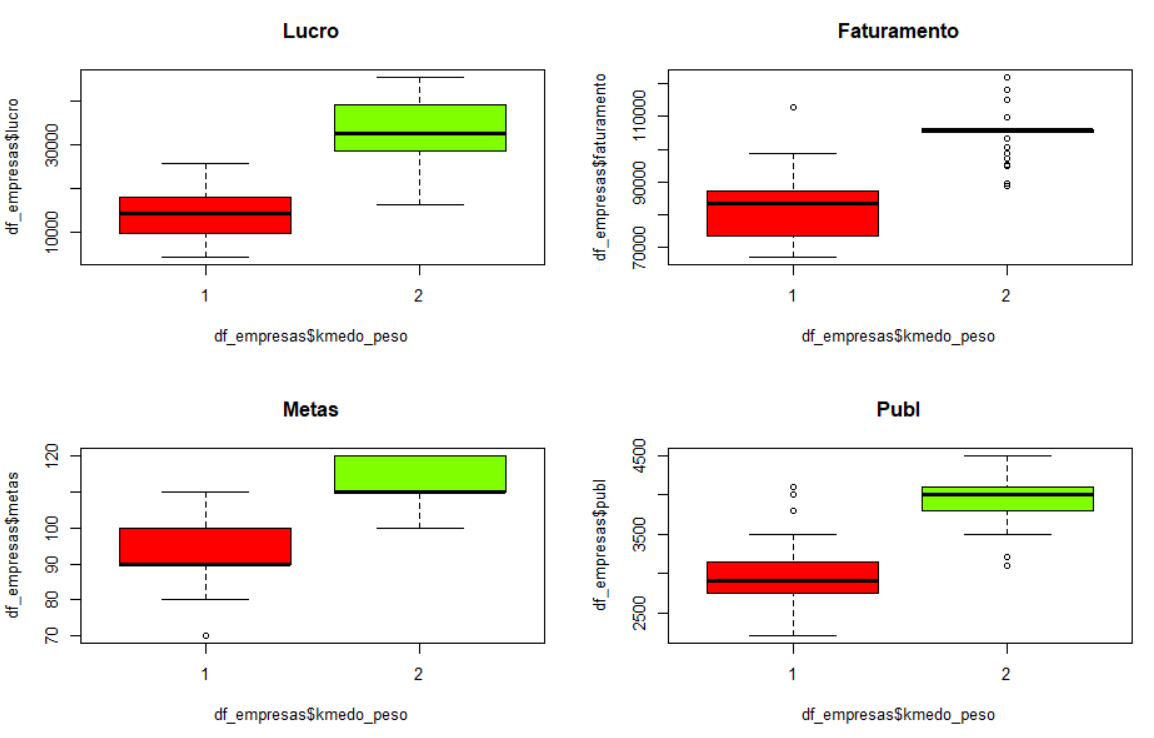
Através das variáveis *drivers* qualitativas, também foi possível encontrar boa distinção entre os *clusters* (há poucas similaridades).

**5. MÉTODO *K-MEDOIDS***

Para a realização do método *K-medoid* foi rodada a função *pamk* com K= 2:6, sendo usado como *set.seed* o ano de nascimento do mais velho (1987), tendo como resultado a sugestão de 2 *clusters* diferindo do escolhido no Método Hireráquico Aglomerativo através do dendrograma que foram 3 *clusters*.

**5.1 Caracterização e comparação entre os clusters obtidos**

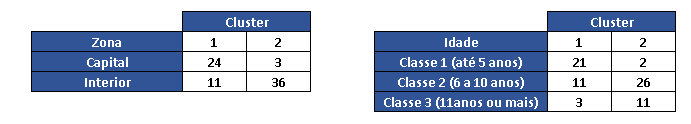
Os *boxplots* a seguir demonstram as distribuições dos *drivers* numéricos dos *clusters* obtidos a partir da função *K-medoid*:



O *cluster* 1foi formado por empresas que apresentaram o menor lucro, faturamento, metas e público.

O *cluster* 2 foi formado por empresas que apresentaram o maior lucro, faturamento, metas e público.

As tabelas a seguir mostram a distribuição dos drivers qualitativos entre os Clusters.



O *cluster* 1 apresentou predomínio de empresas mais novas (até 5 anos) e localizadas na Capital.

O *cluster* 2 apresentou predomínio de empresas de 6 ou mais anos e localizadas no Interior.

**6. RESULTADO**

O cruzamento dos resultados obtidos pelos dois tipos de métodos teve Índice de *Rand* de 0.6992659, indicando que a maioria dos pares de objetos ocorreram no mesmo grupo ou grupos distintos, em ambos os *clusters*.