Arthur Alves
Davidson Nilson
Izabel Barranco
Raquel Cristiane

Desempenho de estudantes em provas

Universidade de São Paulo
2021

Base de dados

Tema: Desempenho de alunos em provas

Descrição: Essa base de dados inclui pontuações em diferentes áreas de uma prova e alguns dados socioeconômicos dos alunos avaliados.

Variáveis de interesse: Nota em matemática, em leitura e em escrita (quantitativa discreta).

Informações: 1000 linhas e 8 variáveis

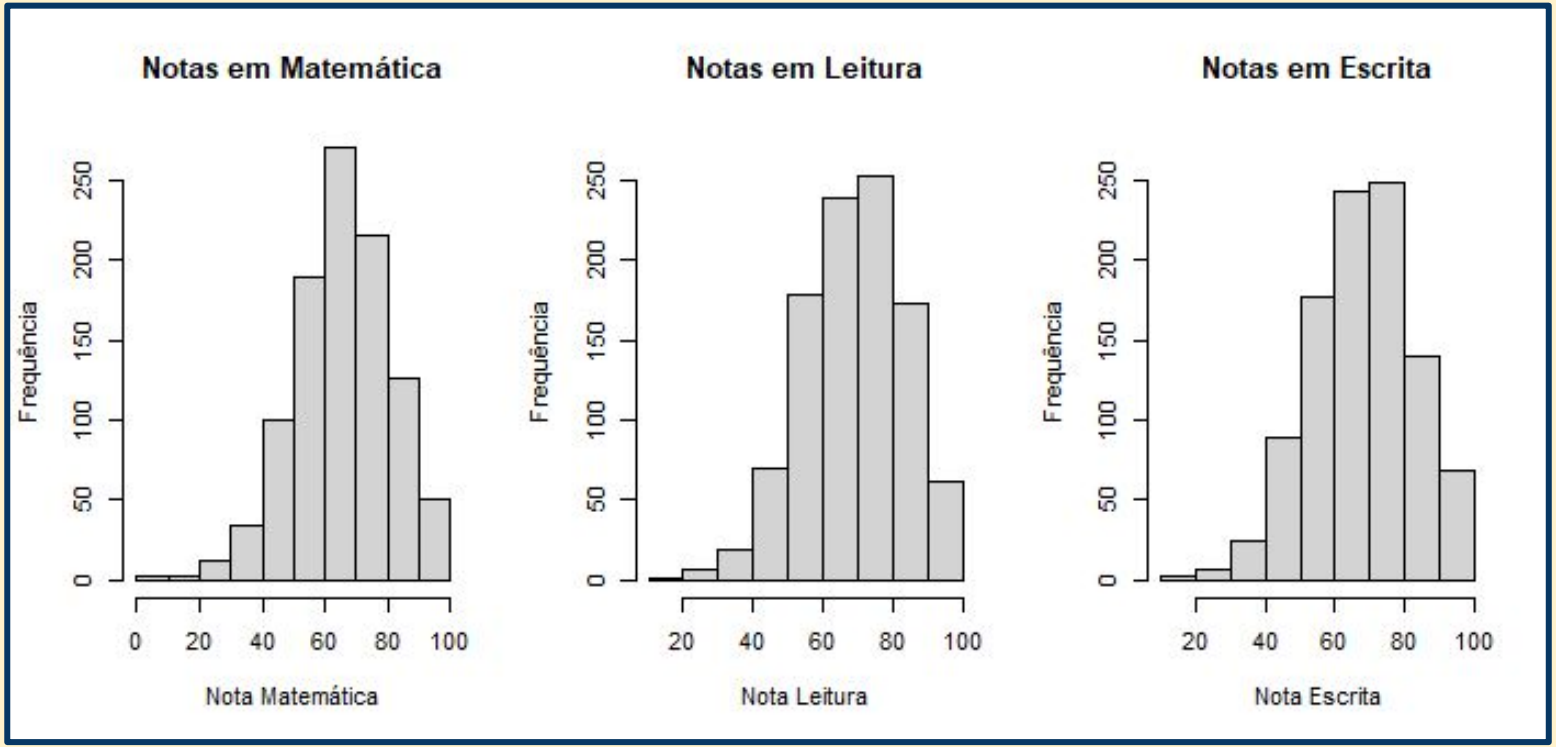
Variáveis

Medidas de dispersão

	Matemática	Leitura	Escrita
Min	0,00	17,00	10
1º quartil	57,00	59,00	57,75
Mediana	66,00	70,00	69,00
Média	66,09	69,17	68,05
3º quartil	77,00	79,00	79,00
Máx	100,00	100	100
Desvio padrão	15,16	14,60	15,20
Variância	229,92	213,17	230,91

Variáveis

Histogramas das variáveis



Análise de clusters

- Técnica exploratória e não-inferencial
- A classificação dos objetos em grupos (cluster) homogêneos;
- Similaridades e/ou distâncias entre os indivíduos;
- Medidas de distância: Distância Euclidiana, Distância de Manhattan, Distância de Correlação de Pearson, Distância de Correlação de Eisen, Distância de Correlação de Spearman e Distância de Correlação de Kendal.

Análise de clusters

K-Means (algoritmo Lloyd-Forgy):

- Número de grupos (k) é pré-especificado;
- Mais eficiente para uma grande quantidade de instâncias;
- Não Hierárquico

Análise de clusters

K-Means (algoritmo Lloyd-Forgy):

1. Gerar k centróides aleatoriamente;

2. Calcular distância entre todos os pontos e cada um dos centróides;

3. Cada registro será atribuído ao centróide (cluster) que tem a menor distância;

4. Recalcular centróides a partir dos pontos atribuídos;

5. Repetir algoritmo até não ocorrer alteração dos centróides.

Análise de clusters

Método hierárquico:

- Clusterização Aglomerativa;
- Clusterização Divisivas;
- Dendrogramas;

Análise de clusters

Método hierárquico:

- Clusterização Aglomerativa

  a) Cada objeto considerando um cluster individual;

  b) São formados pares de clusters com as menores distâncias entre si;

  c) Repete-se o procedimento em novos clusters maiores, chegando até um único grande cluster.

Análise de clusters

Método hierárquico:

- Distância Euclidiana
- Método Ward a soma de quadrados da distância entre os dois cluster.
- Diminuição da variância interna e aumento da variância externa.

Análise de clusters

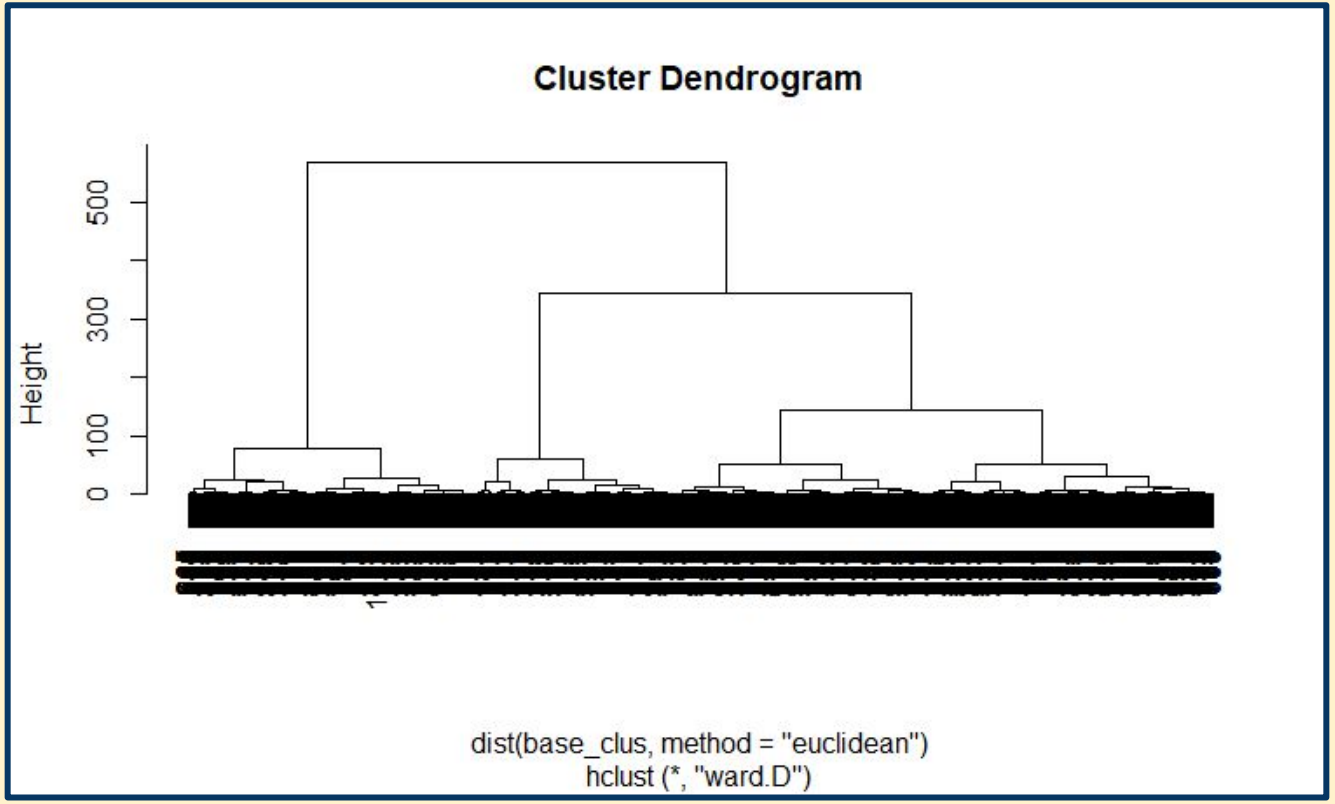
Método hierárquico (Método Ward):

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

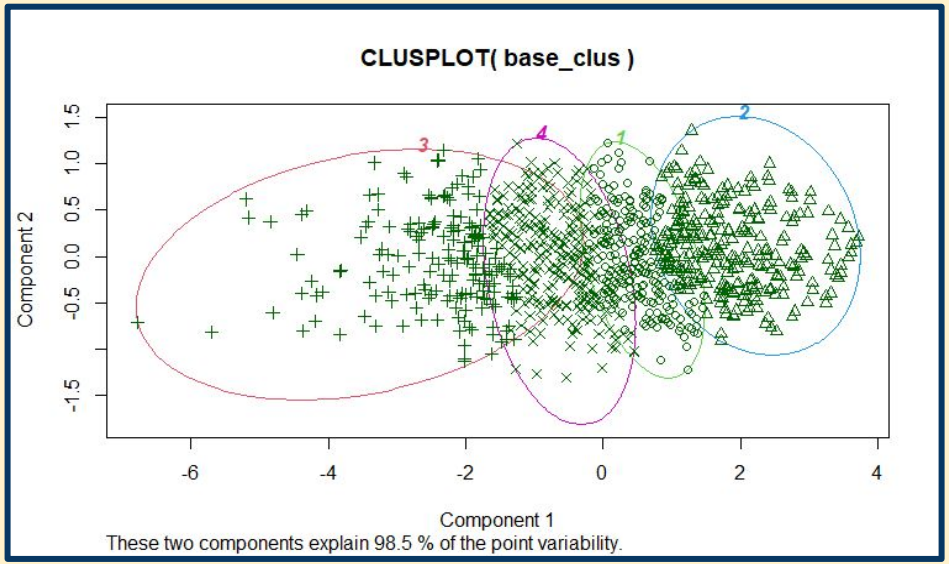
$$\sum_{i=1}^{n}(p_i - q_i)^2$$

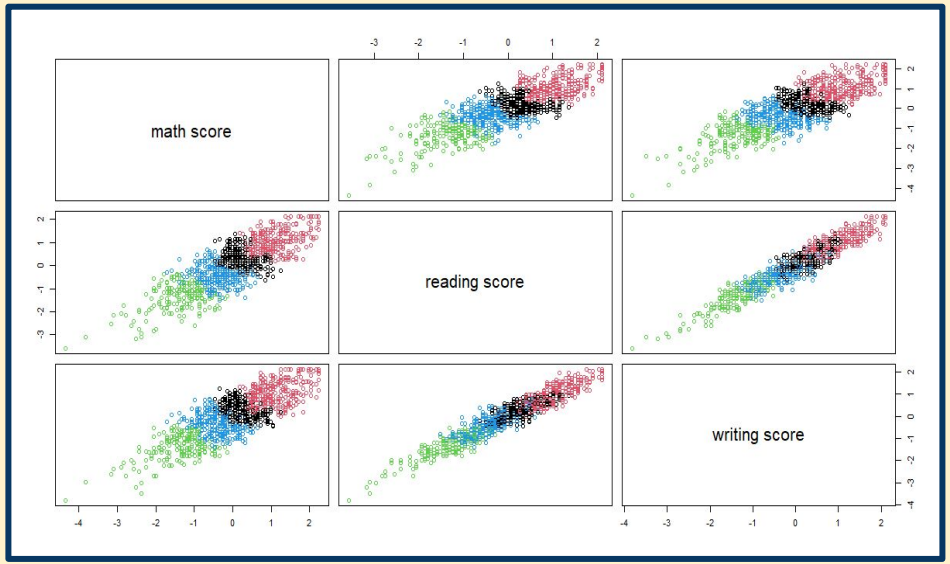
- pi: variável i da observação pertencente ao cluster p
- qi: variável i da observação pertencente ao cluster q

Análise de clusters - Hierárquico



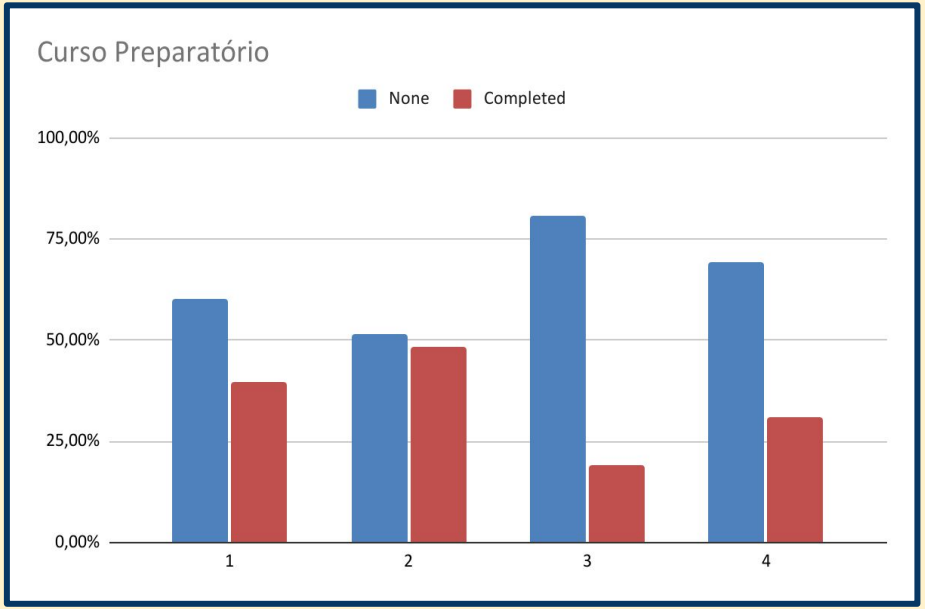
Análise de clusters - Hierárquico

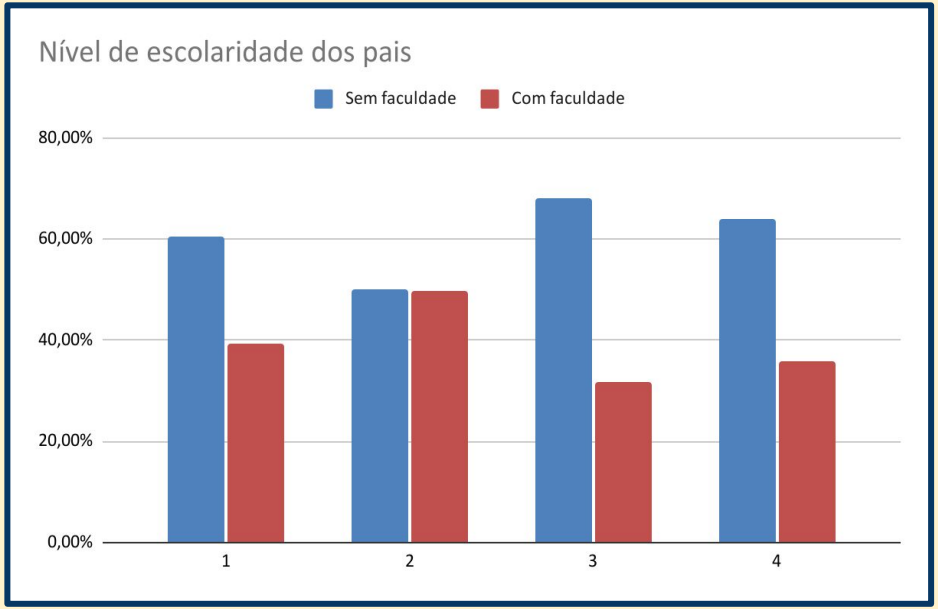




Análise de clusters - Hierárquico

Analisando os clusters com as demais variáveis do dataset.

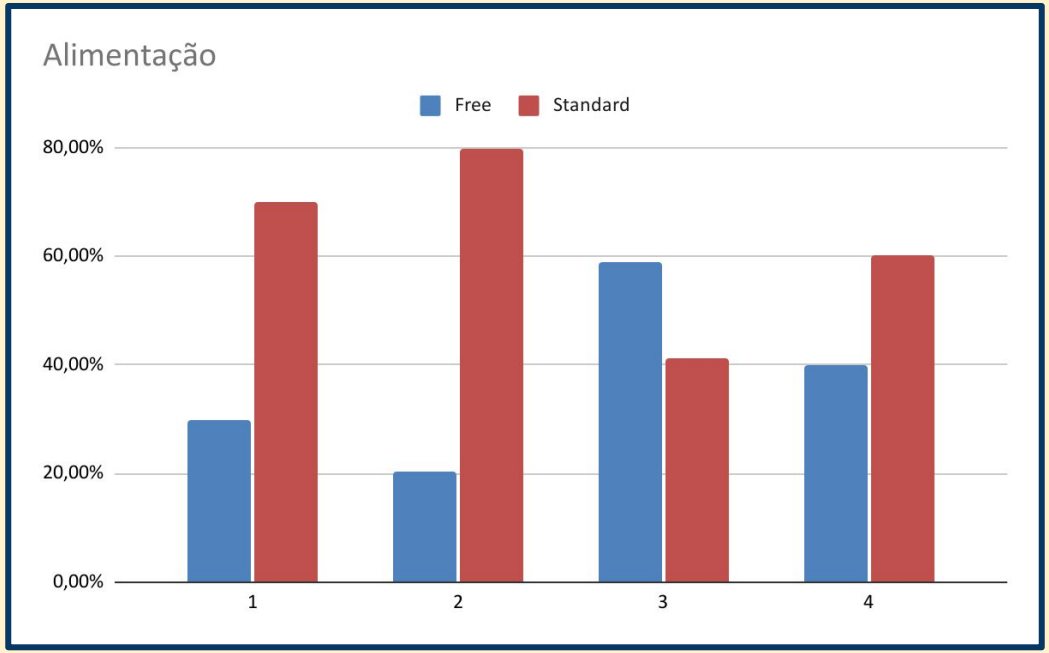




Cluster 1	Cluster 2	Cluster 3	Cluster 4
251	281	192	276

Análise de clusters - Hierárquico

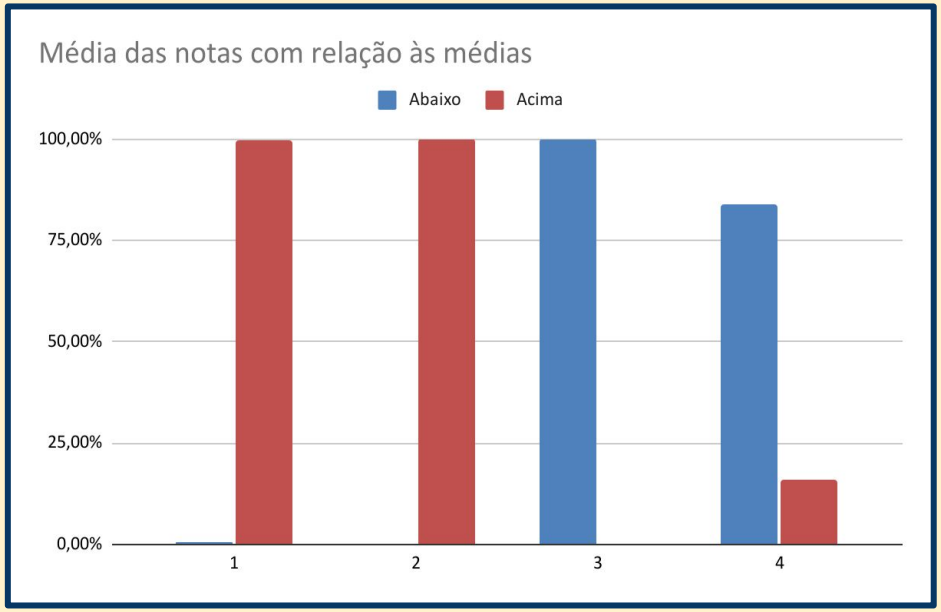
Analisando os clusters com as demais variáveis do dataset.

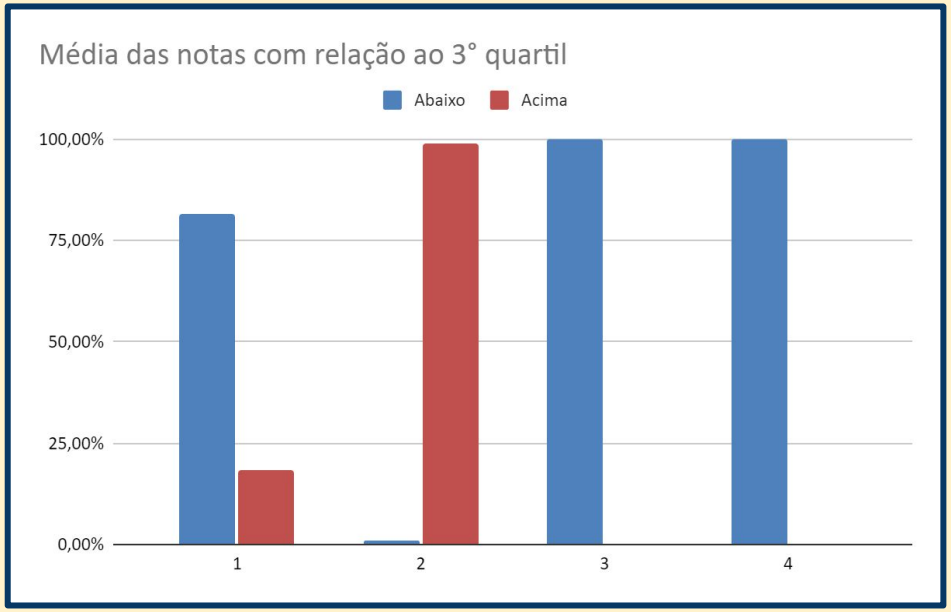


Cluster 1	Cluster 2	Cluster 3	Cluster 4
251	281	192	276

Análise de clusters - Hierárquico

Analisando os clusters com as demais variáveis do dataset.





Média Geral
67,77

Cluster 1	Cluster 2	Cluster 3	Cluster 4
251	281	192	276

3º quartil
78,33

Análise de clusters - Hierárquico

Características observáveis dos clusters

Cluster 1

Muitos com curso preparatório

Muitos pais com ensino superior

Muitos com alimentação de preço comum

Quase todos acima da média geral

Poucos com média acima do 3º quartil

Cluster 2

Maior quant. com curso preparatório

Maior com pais com ensino superior

Maior com alimentação de preço comum

Quase todos com média acima do 3º quartil

Cluster 3

Menor quant. com curso preparatório

Menor com pais com ensino superior

Maior com alimentação grátis ou preço reduzido

Todos abaixo da média geral

Cluster 4

Poucos com curso preparatório

Poucos pais com ensino superior

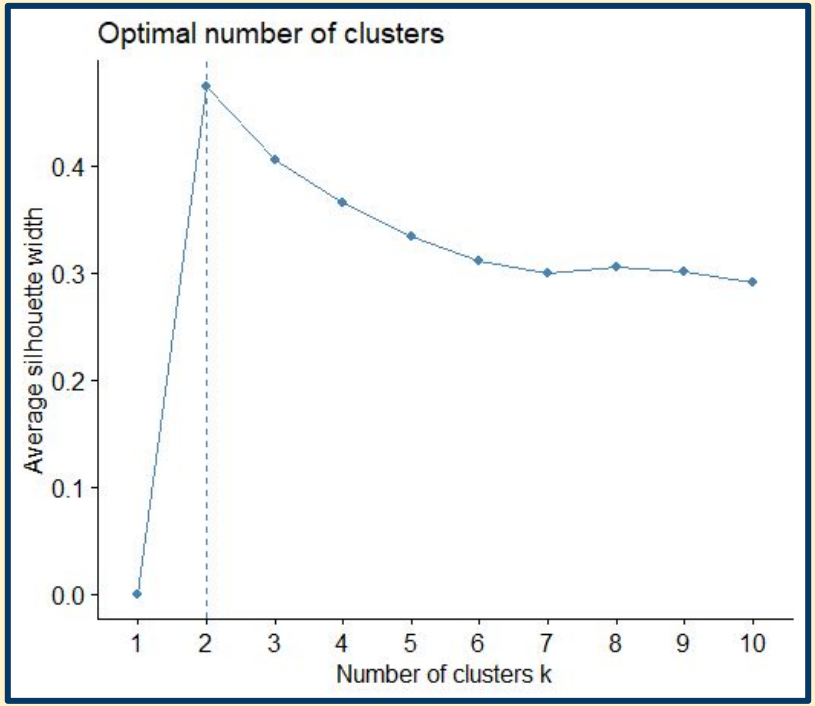
Muitos com alimentação grátis ou preço reduzido

Quase todos com média abaixo da média geral

Todos com média abaixo do 3º quartil

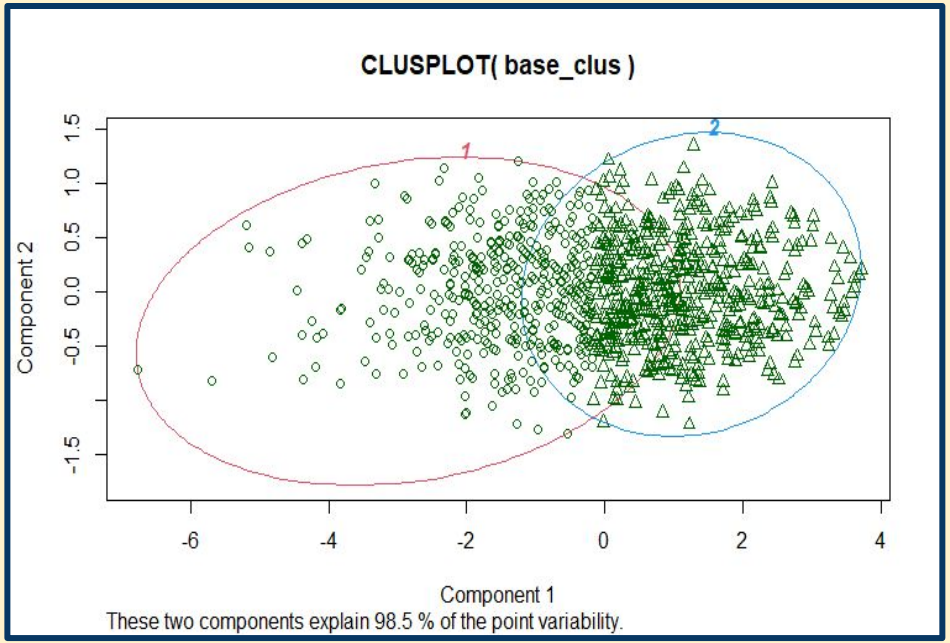
Análise de clusters - Não hierárquico

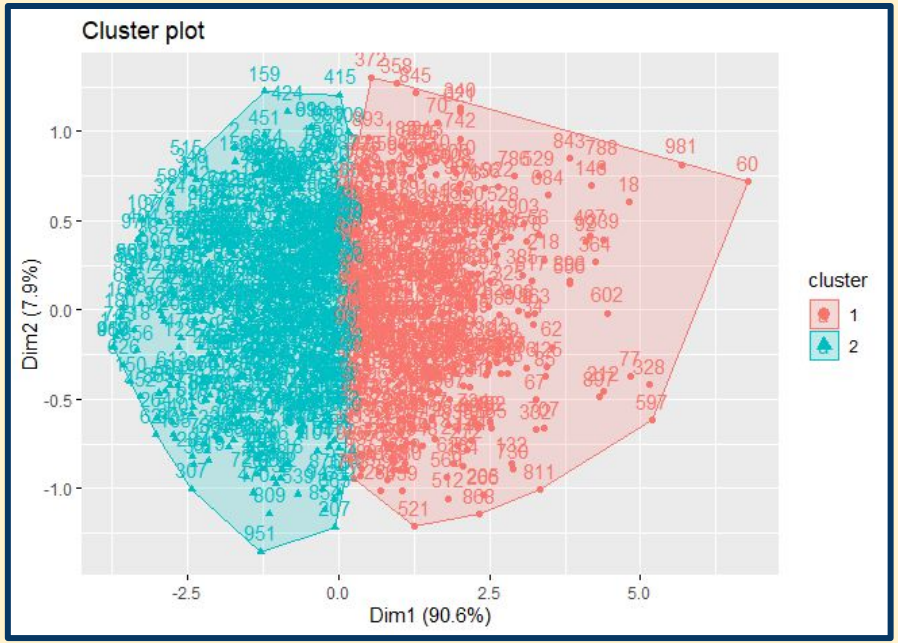
O algoritmo utilizado para essa análise foi o Kmeans, com distância euclidiana.



Análise de clusters - Não hierárquico

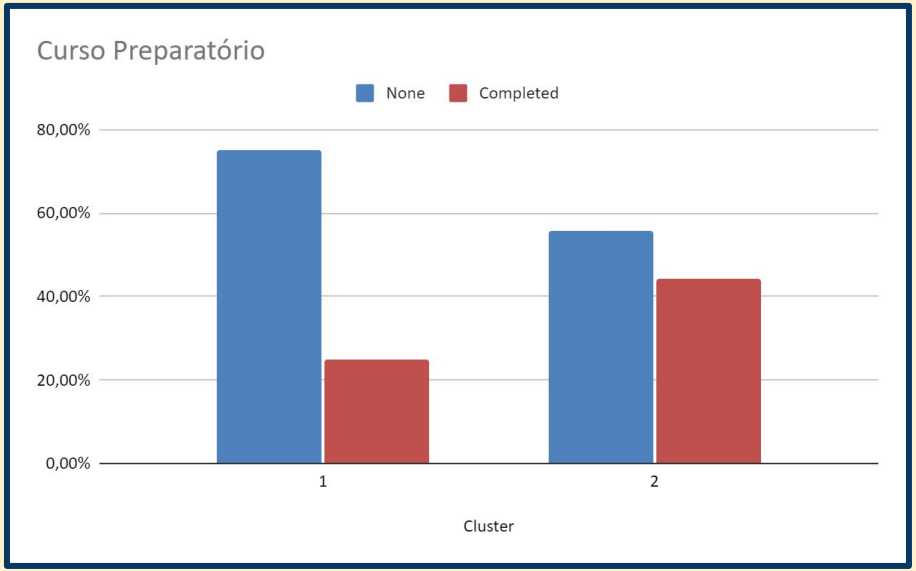
O algoritmo utilizado para essa análise foi o Kmeans, com distância euclidiana.

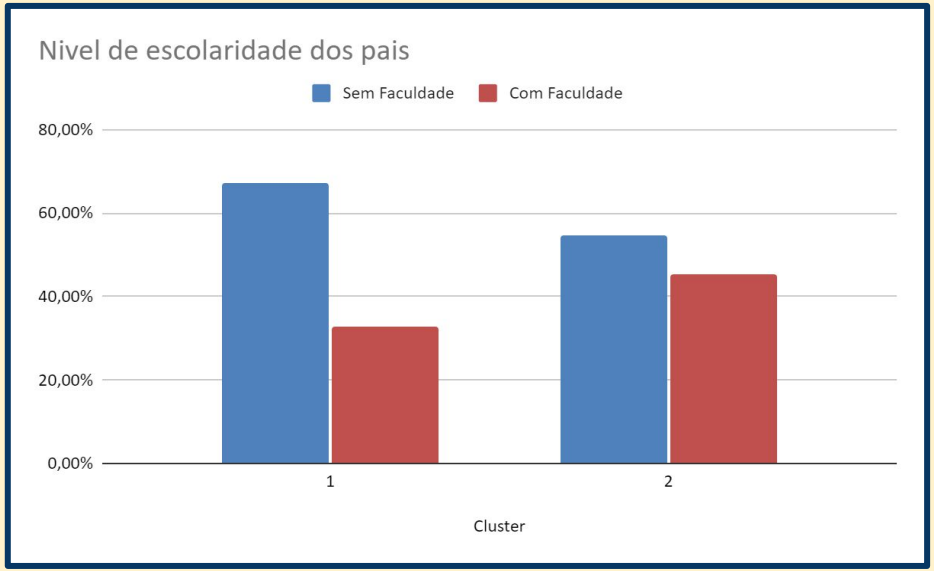




Análise de clusters - Não hierárquico

Analisando os clusters com as demais variáveis do dataset.

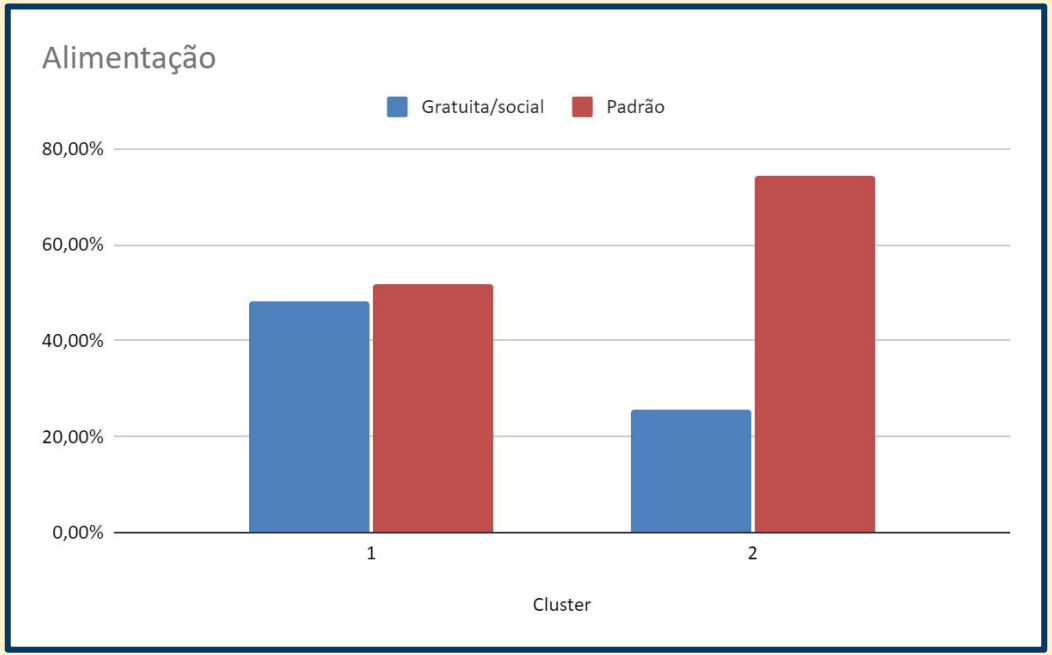




Cluster 1	Cluster 2
440	560

Análise de clusters - Não hierárquico

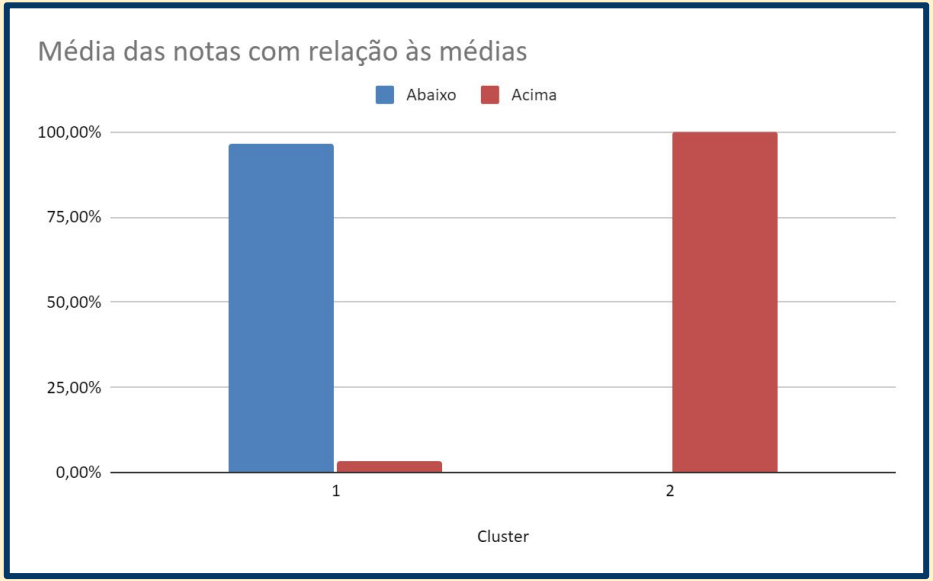
Analisando os clusters com as demais variáveis do dataset.

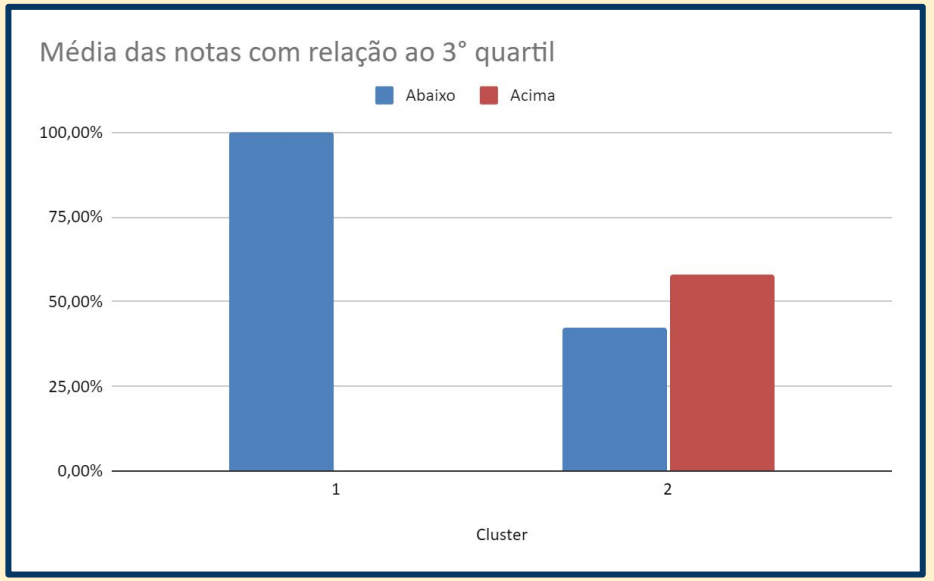


Cluster 1	Cluster 2
440	560

Análise de clusters - Não hierárquico

Analisando os clusters com as demais variáveis do dataset.





Média Geral
67,77

Cluster 1	Cluster 2
440	560

3º quartil
78,33

Análise de clusters - Hierárquico

Características observáveis dos clusters

Cluster 1

Menor quant. com curso preparatório

Menor com pais com ensino superior

Maior com alimentação grátis ou preço reduzido

Quase todos abaixo da média geral

Todos com média abaixo do 3º quartil

Cluster 2

Maior quant. com curso preparatório

Maior com pais com ensino superior

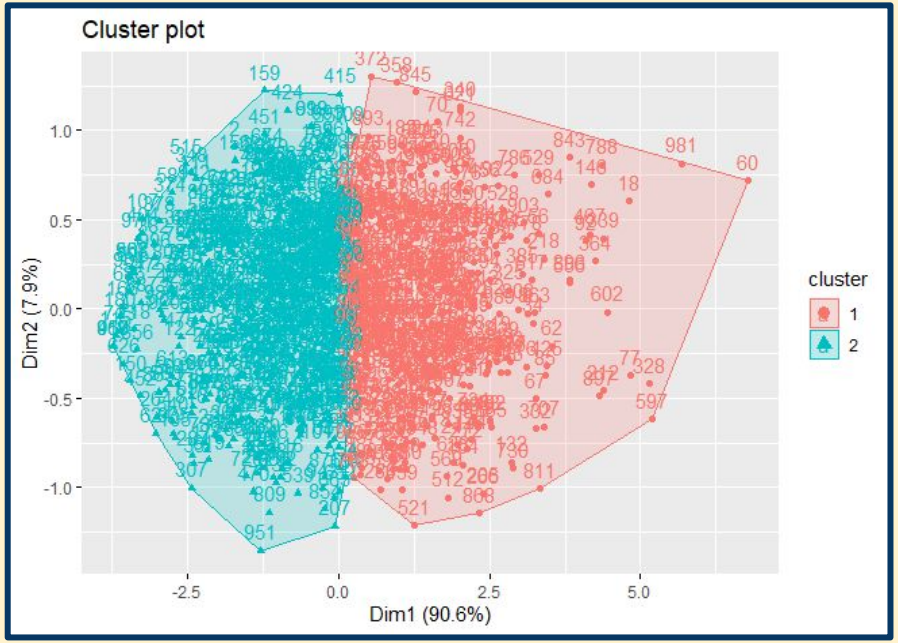
Maior com alimentação de preço comum

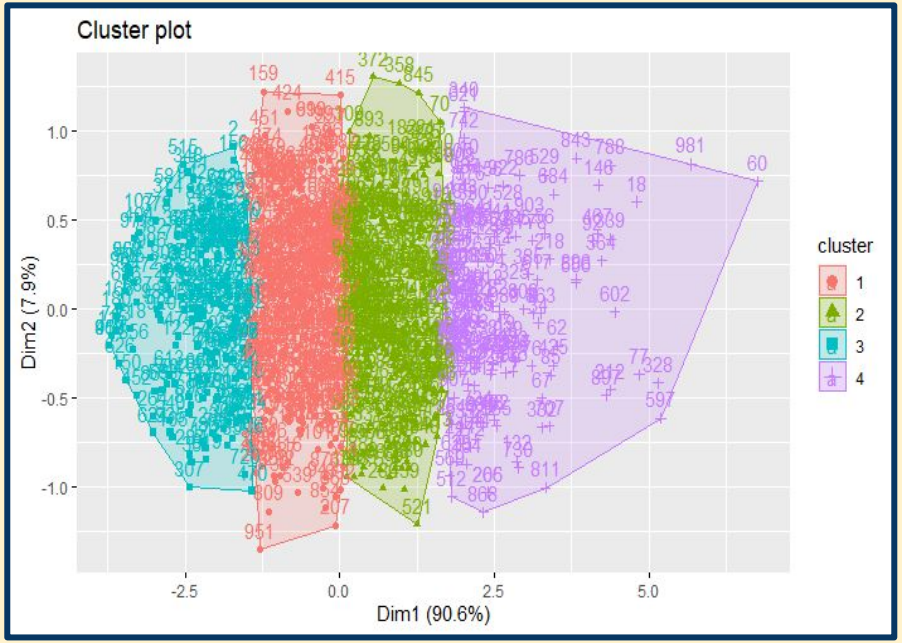
Todos com médias acima da média geral

Maioria com média acima do 3º quartil

Análise de clusters - Não hierárquico

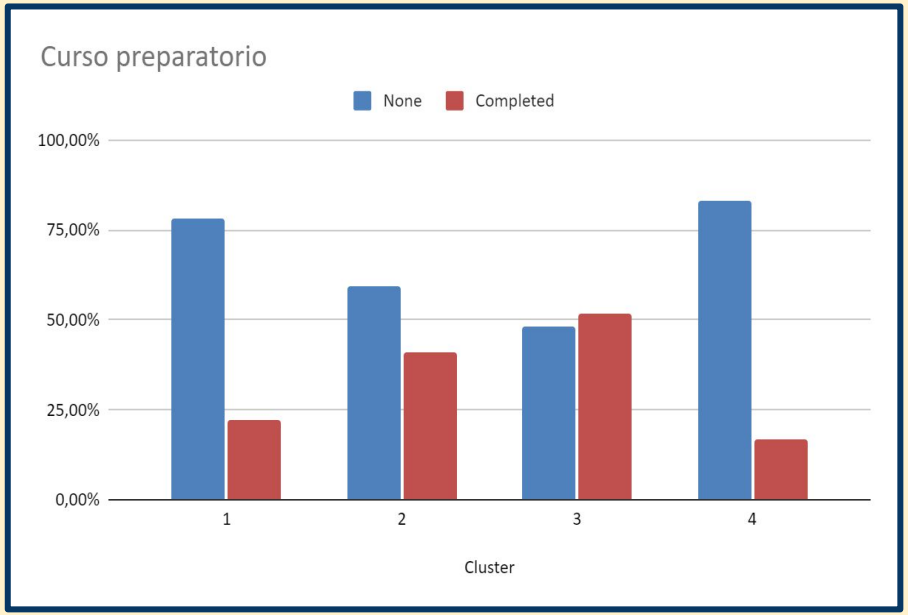
Kmeans com 2 e 4 clusters.

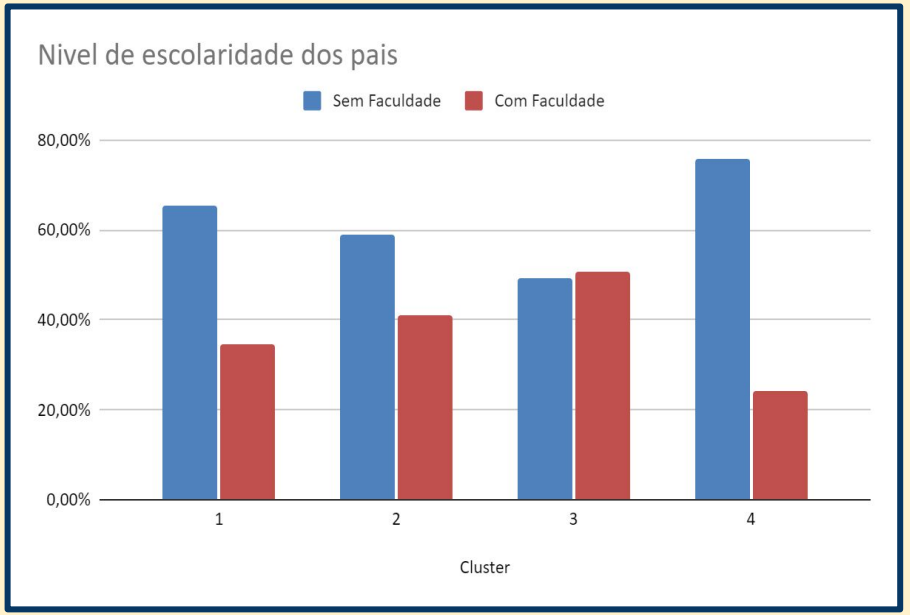




Análise de clusters - Não hierárquico

Analisando com 4 clusters com as demais variáveis do dataset.

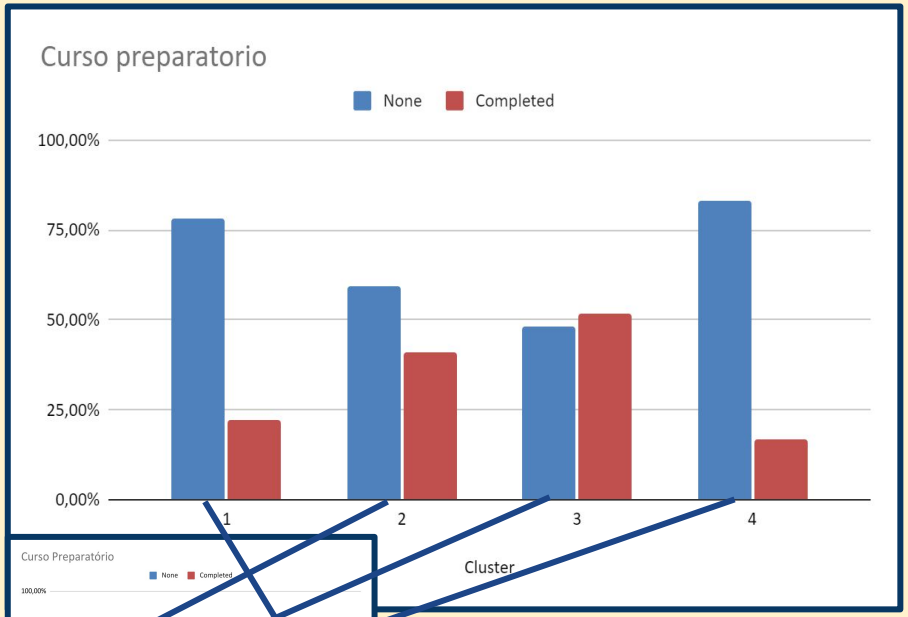


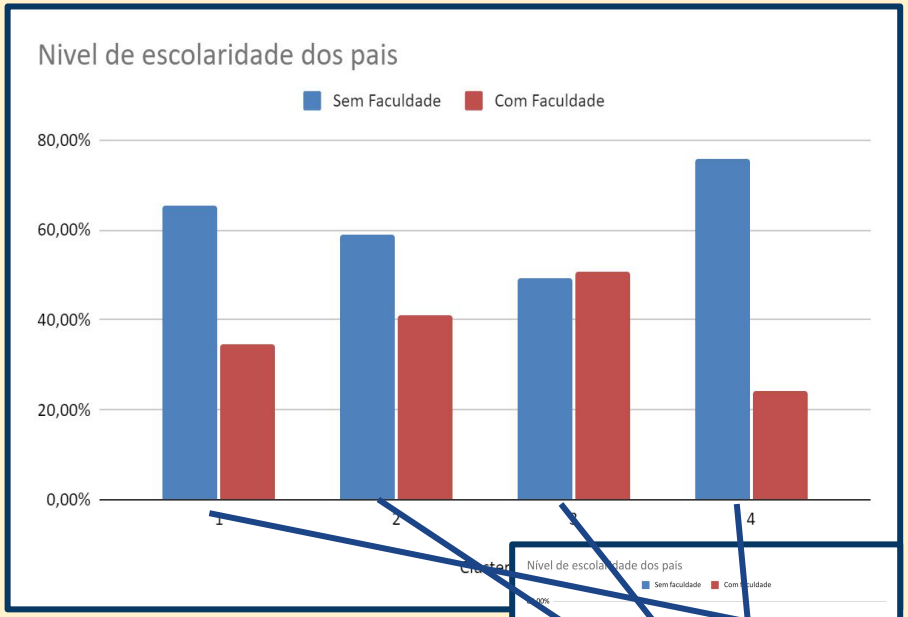


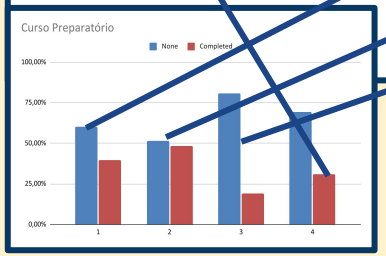
Cluster 1	Cluster 2	Cluster 3	Cluster 4
278	397	230	95

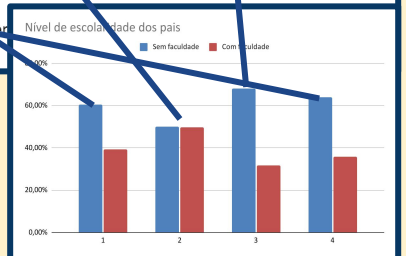
Análise de clusters - Não hierárquico

Comparação com clusters hierárquicos





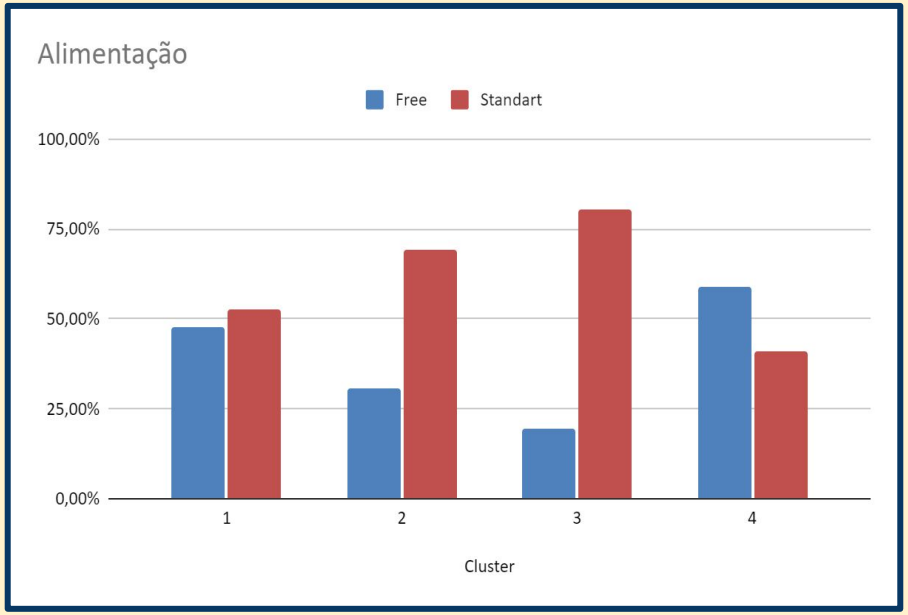




Cluster 1	Cluster 2	Cluster 3	Cluster 4
278	397	230	95

Análise de clusters - Não hierárquico

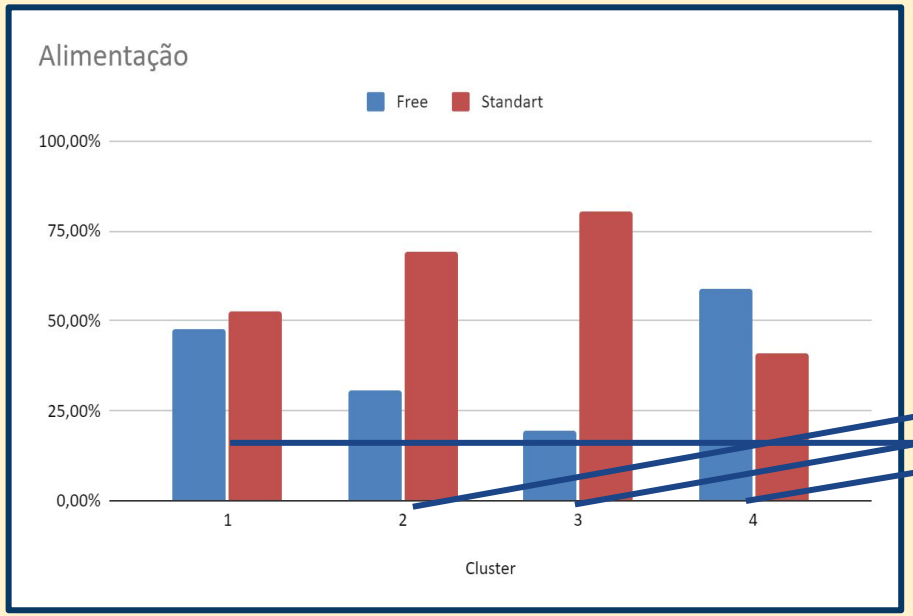
Analisando com 4 clusters com as demais variáveis do dataset.

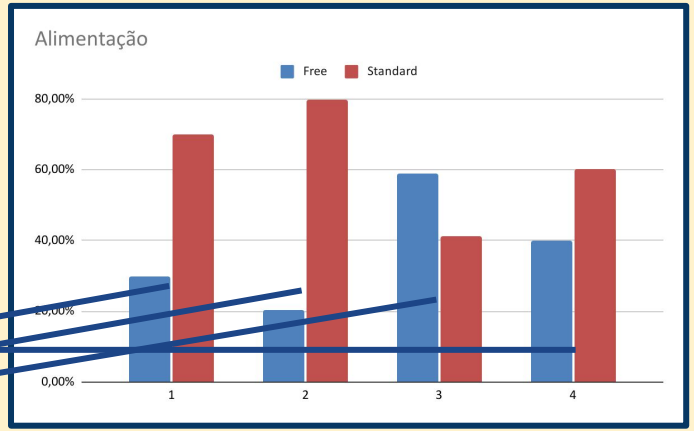


Cluster 1	Cluster 2	Cluster 3	Cluster 4
278	397	230	95

Análise de clusters - Não hierárquico

Comparação com clusters hierárquicos

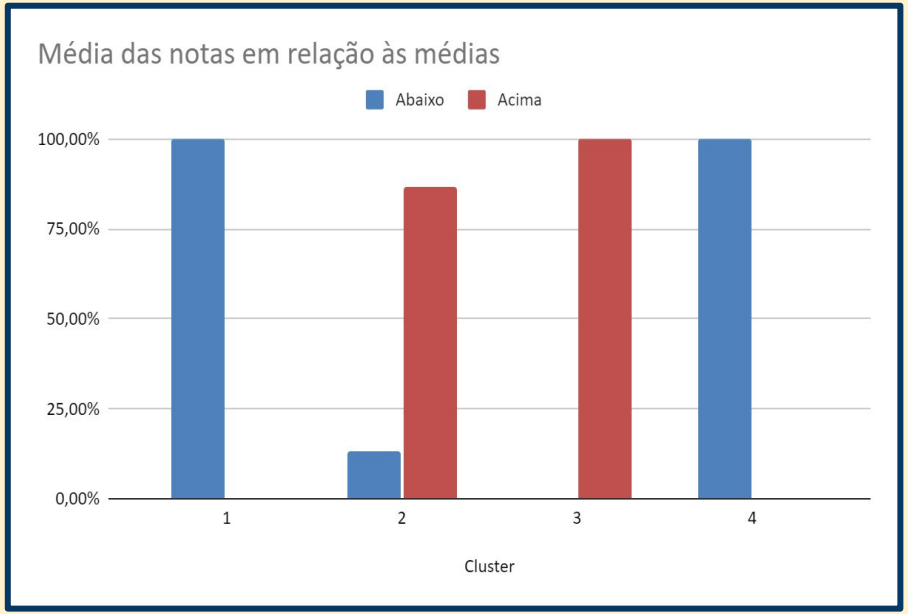


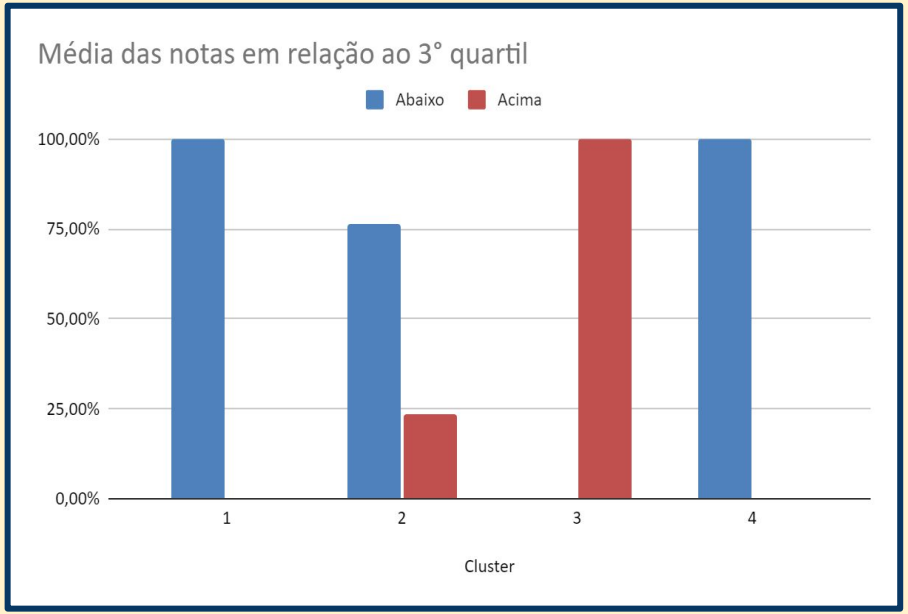


Cluster 1	Cluster 2	Cluster 3	Cluster 4
278	397	230	95

Análise de clusters - Não hierárquico

Analisando com 4 clusters com as demais variáveis do dataset.





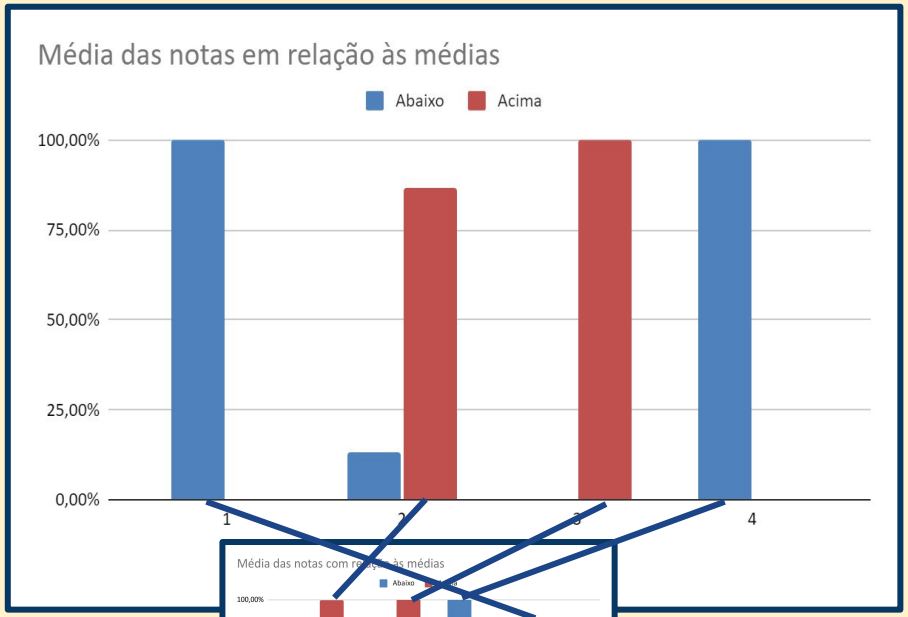
Média Geral
67,77

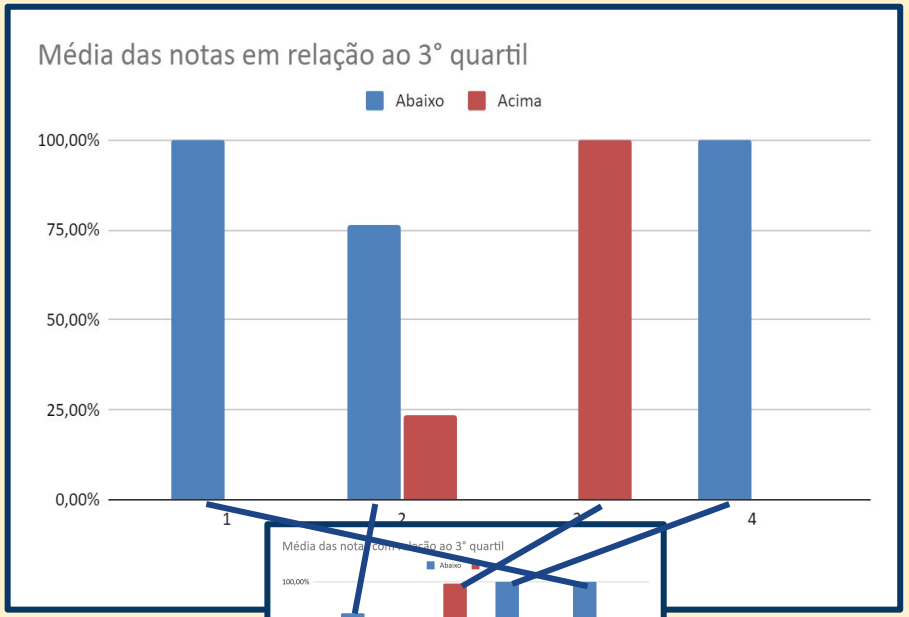
Cluster 1	Cluster 2	Cluster 3	Cluster 4
278	397	230	95

3º quartil
78,33

Análise de clusters - Não hierárquico

Comparação com clusters hierárquicos





Média G	er 1	Cluster 2	Cluster 3	Clu	º quartil
67,7	78	397	230		78,33

Análise de clusters - Conclusão

A partir das notas, os modelos de clusters dividiram os alunos entre grupos de notas semelhantes, onde pode-se observar que, não somente as médias dos alunos entre as provas mostra a diferença entre eles, mas também os fatores socioeconômicos.

Esses fatores demonstram que os alunos com melhores notas são aqueles que tiveram melhores condições para fazer a prova, como curso preparatório, pais com formação acadêmica e alimentação de preço integral.

Obrigado!