Módulo 18- Análise de Cluster

Tutorial SPSS – Análise dos Resultados

Método Hierárquico e Não-Hierárquico

Situação Problema

Apresentamos novamente a situação problema com o objetivo do leitor se localizar melhor. Caso tenha em mente a situação problema considerada, sugere-se ir para o próximo item – Análise de Resultados.

Uma varejista de roupas e acessórios femininos voltados para a classe A e B iniciará um programa de relacionamento com os clientes, oferecendo atendimento personalizado, promoções específicas e facilidades para cada grupo de clientes. É, portanto, necessário segmentar a clientela em grupos distintos, agrupando clientes com perfil semelhante, para que se possa direcionar a oferta.

A empresa elaborou um questionário com afirmações sobre compras e selecionou uma amostra, ao acaso, de 20 clientes para respondê-los. As questões são mensuradas através de uma nota de 1 a 5, onde 1 representa a nota mais baixa (discordância total) e a nota 5 representa a maior nota (concordância total), conforme é mostrado a seguir:

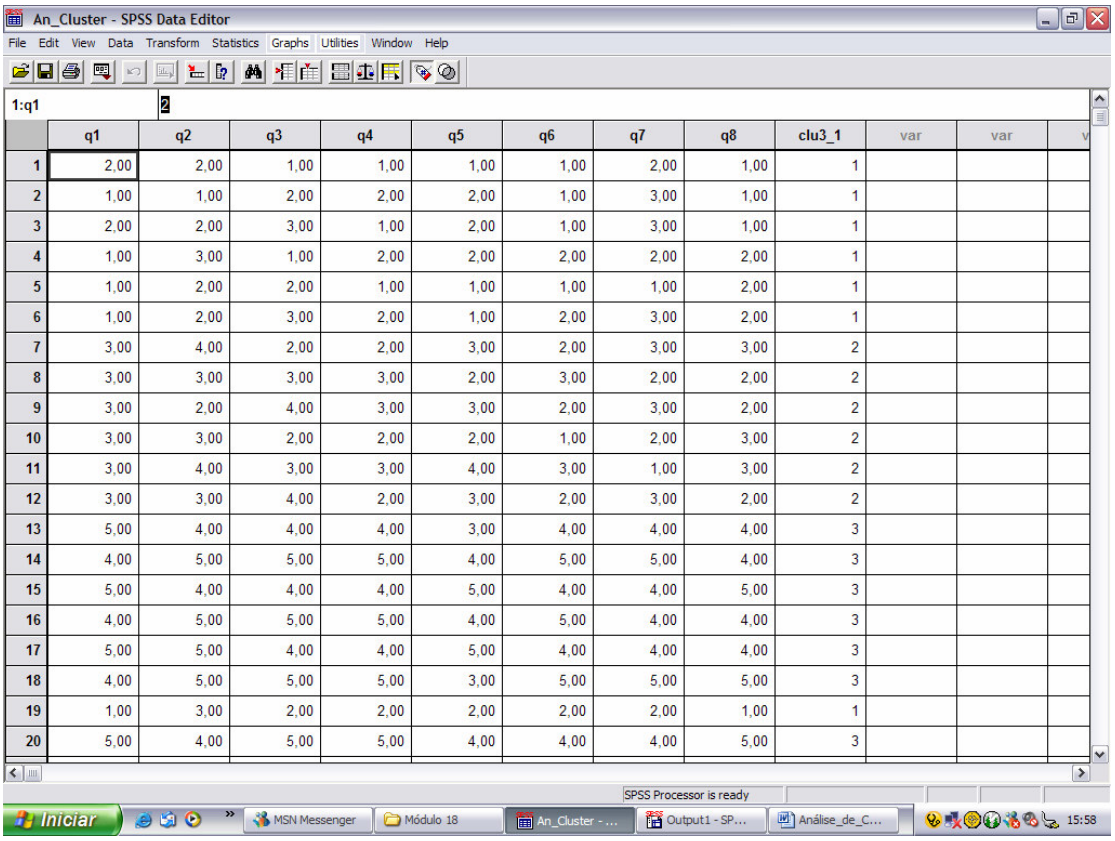
1) Só compro roupas quando realmente preciso	1	2	3	4	5
2) Uso as roupas que estão na moda	1	2	3	4	5
3) Compro roupas extravagantes mesmo não vá utilizá-las	1	2	3	4	5
4) Quando compro roupas costumo escolher também outras peças e acessórios para combinar	1	2	3	4	5
5) Costumo comprar mais roupas "curinga", fáceis de combinar	1	2	3	4	5
6) Quando gosto não me importo com o preço da peça	1	2	3	4	5
7) Procuro comprar sempre peças exclusivas	1	2	3	4	5
8) Só compro roupas se elas forem de marcas famosas	1	2	3	4	5

Tabela 1: Instrumento de pesquisa.

Para realizar o agrupamento devemos utilizar a ferramenta análise de clusters, conhecida também como análise de conglomerados. Conforme o texto teórico deste módulo, existem basicamente dois grandes grupos de métodos de clusterização: métodos hierárquicos e métodos não- hierárquicos. A forma de gerar tabelas, bem como de analisar os resultados possuem algumas diferenças. Assim, este tutorial foi dividido em quatro partes: geração de tabelas para o método hierárquico, análise dos resultados para o método hierárquico, geração de tabelas para o método não- hierárquico, análise dos resultados para o método não- hierárquico.

Método Hierárquico

Após ter finalizado o processo de obtenção das tabelas (tutorial anterior), observe que a planilha inicial ganhou uma nova variável "clu3_1", que mostra o número do cluster que o caso foi agrupado. Observe a figura 11. Percebemos que os casos 1, 2, 3, 4, 5, 6 e 19 pertencem ao cluster 1; os casos 7, 8, 9, 10, 11, 12 pertencem ao cluster 2 e os casos 13, 14, 15, 16, 17, 18 e 20 pertencem ao cluster 3.



	q1	q2	q3	q4	q5	q6	q7	q8	clu3_1	var	var
1	2,00	2,00	1,00	1,00	1,00	1,00	2,00	1,00	1		
2	1,00	1,00	2,00	2,00	2,00	1,00	3,00	1,00	1		
3	2,00	2,00	3,00	1,00	2,00	1,00	3,00	1,00	1		
4	1,00	3,00	1,00	2,00	2,00	2,00	2,00	2,00	1		
5	1,00	2,00	2,00	1,00	1,00	1,00	1,00	2,00	1		
6	1,00	2,00	3,00	2,00	1,00	2,00	3,00	2,00	1		
7	3,00	4,00	2,00	2,00	3,00	2,00	3,00	3,00	2		
8	3,00	3,00	3,00	3,00	2,00	3,00	2,00	2,00	2		
9	3,00	2,00	4,00	3,00	3,00	2,00	3,00	2,00	2		
10	3,00	3,00	2,00	2,00	2,00	1,00	2,00	3,00	2		
11	3,00	4,00	3,00	3,00	4,00	3,00	1,00	3,00	2		
12	3,00	3,00	4,00	2,00	3,00	2,00	3,00	2,00	2		
13	5,00	4,00	4,00	4,00	3,00	4,00	4,00	4,00	3		
14	4,00	5,00	5,00	5,00	4,00	5,00	5,00	4,00	3		
15	5,00	4,00	4,00	4,00	5,00	4,00	4,00	5,00	3		
16	4,00	5,00	5,00	5,00	4,00	5,00	4,00	4,00	3		
17	5,00	5,00	4,00	4,00	5,00	4,00	4,00	4,00	3		
18	4,00	5,00	5,00	5,00	3,00	5,00	5,00	5,00	3		
19	1,00	3,00	2,00	2,00	2,00	2,00	2,00	1,00	1		
20	5,00	4,00	5,00	5,00	4,00	4,00	4,00	5,00	3		

Figura 11: Planilha com a variável gerada (clu3_1).

Parte 2- Análise dos resultados

A tabela 2 mostra os casos considerados na análise, tanto os válidos quanto os casos faltantes.

Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
20	100,0%	0	,0%	20	100,0%

a. Squared Euclidean Distance used

Tabela 2: Número de casos considerados.

Neste problema obtivemos as respostas de todas as clientes entrevistadas, portanto, os 20 casos foram válidos.

Vertical Icicle

Number of clusters	Case																																							
	18:Case 18		16:Case 16		14:Case 14		17:Case 17		15:Case 15		20:Case 20		13:Case 13		12:Case 12		9:Case 9		8:Case 8		11:Case 11		10:Case 10		7:Case 7		6:Case 6		3:Case 3		2:Case 2		19:Case 19		4:Case 4		5:Case 5		1:Case 1
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X
5	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X
6	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X
7	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X		X		X	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X
8	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X		X		X	X	X		X	X	X	X	X		X	X	X		X	X	X	X
9	X	X	X	X	X		X	X	X		X	X	X		X	X	X	X	X		X		X	X	X		X	X	X	X	X		X	X	X		X	X	X	X
10	X	X	X	X	X		X	X	X		X	X	X		X	X	X		X		X		X	X	X		X	X	X	X	X		X	X	X		X	X	X	X
11	X	X	X	X	X		X	X	X		X	X	X		X	X	X		X		X		X	X	X		X		X	X	X		X	X	X		X	X	X	X
12	X	X	X	X	X		X	X	X		X	X	X		X	X	X		X		X		X		X		X		X	X	X		X	X	X		X	X	X	X
13	X	X	X	X	X		X	X	X		X	X	X		X	X	X		X		X		X		X		X		X	X	X		X	X	X		X		X	X
14	X	X	X	X	X		X	X	X		X		X		X	X	X		X		X		X		X		X		X	X	X		X	X	X		X		X	X
15	X	X	X	X	X		X	X	X		X		X		X	X	X		X		X		X		X		X		X		X		X	X	X		X		X	X
16	X		X	X	X		X	X	X		X		X		X	X	X		X		X		X		X		X		X		X		X	X	X		X		X	X
17	X		X	X	X		X		X		X		X		X	X	X		X		X		X		X		X		X		X		X	X	X		X		X	X
18	X		X	X	X		X		X		X		X		X		X		X		X		X		X		X		X		X		X	X	X		X		X	X
19	X		X	X	X		X		X		X		X		X		X		X		X		X		X		X		X		X		X		X		X		X	X

Tabela 3: Matriz “vertical icicle”.

A tabela 3 mostra uma representação de como foram feitos os agrupamentos. Devemos lê-la de baixo para cima. As colunas representam os casos agrupados, portanto 20 casos (colunas com escrito “18:case 18”). Inicialmente vemos que os casos 16 e 14 foram agrupados, sendo então, considerados 19 clusters, que podem ser vistos na linha 19. Em seguida, os dois os casos mais próximos (de acordo com a medida de distância selecionada) são

agrupados, que são cliente 4 e cliente 19, formando 18 clusters (linha 18), depois a cliente 12 e a cliente 9 são agrupadas, formando 17 clusters e assim sucessivamente até obtermos 3 clusters (observe a terceira linha de cima para baixo).

O cluster 3 possui os casos: 13, 14, 15, 16, 17, 18, 20; o cluster 2 possui: 7, 8, 9, 10, 11, 12; o cluster 1 possui: 1, 2, 3, 4, 5, 6, 19.

Proximity Matrix

	Squared Euclidean Distance																			
Case	1:Case 1	2:Case 2	3:Case 3	4:Case 4	5:Case 5	6:Case 6	7:Case 7	8:Case 8	9:Case 9	10:Case 10	11:Case 11	12:Case 12	13:Case 13	14:Case 14	15:Case 15	16:Case 16	17:Case 17	18:Case 18	19:Case 19	20:Case 20
1:Case 1		3,594	3,671	3,274	2,278	4,970	10,083	8,489	11,942	4,906	17,979	11,108	30,922	49,711	42,305	46,057	42,099	50,033	3,336	45,275
2:Case 2	3,594		2,213	4,981	5,261	2,880	11,086	8,758	6,956	7,325	18,454	8,474	28,019	42,315	36,800	40,123	37,941	43,938	3,898	38,921
3:Case 3	3,671	2,213		5,644	5,121	2,601	7,400	6,280	4,681	4,986	14,629	3,847	22,199	35,936	30,980	33,744	30,773	37,559	3,416	32,962
4:Case 4	3,274	4,981	5,644		3,603	4,344	4,994	5,120	9,564	3,412	9,638	8,388	22,766	36,028	30,525	32,374	29,993	36,629	1,083	33,791
5:Case 5	2,278	5,261	5,121	3,603		4,471	11,563	8,385	12,154	4,925	15,391	11,320	32,408	50,586	42,770	45,471	43,585	49,887	3,541	44,595
6:Case 6	4,970	2,880	2,601	4,344	4,471		8,237	4,886	5,534	5,467	14,816	5,705	19,969	32,129	30,331	29,936	31,146	31,429	3,138	30,006
7:Case 7	10,083	11,086	7,400	4,994	11,563	8,237		4,114	5,998	2,528	5,121	3,474	9,288	19,150	13,423	16,958	12,564	20,032	5,954	16,845
8:Case 8	8,489	8,758	6,280	5,120	8,385	4,886	4,114		3,100	3,476	4,518	2,929	9,694	20,572	17,453	16,918	16,921	21,174	3,914	17,425
9:Case 9	11,942	6,956	4,681	9,564	12,154	5,534	5,998	3,100		5,830	7,824	1,176	9,718	18,981	14,874	16,789	15,689	20,884	7,214	15,003
10:Case 10	4,906	7,325	4,986	3,412	4,925	5,467	2,528	3,476	5,830		6,973	4,654	15,168	30,086	21,905	26,432	22,394	29,666	4,372	24,027
11:Case 11	17,979	18,454	14,629	9,638	15,391	14,816	5,121	4,518	7,824	6,973		6,306	11,142	19,531	12,674	14,416	11,816	21,714	9,454	15,248
12:Case 12	11,108	8,474	3,847	8,388	11,320	5,705	3,474	2,929	1,176	4,654	6,306		9,205	18,126	14,361	15,933	13,829	20,029	6,037	15,495
13:Case 13	30,922	28,019	22,199	22,766	32,408	19,969	9,288	9,694	9,718	15,168	11,142	9,205		4,067	3,113	3,336	3,276	3,927	22,458	2,236
14:Case 14	49,711	42,315	35,936	36,028	50,586	32,129	19,150	20,572	18,981	30,086	19,531	18,126	4,067		4,578	,731	3,393	1,161	34,575	2,852
15:Case 15	42,305	36,800	30,980	30,525	42,770	30,331	13,423	17,453	14,874	21,905	12,674	14,361	3,113	4,578		3,847	1,185	6,019	31,239	1,726
16:Case 16	46,057	40,123	33,744	32,374	45,471	29,936	16,958	16,918	16,789	26,432	14,416	15,933	3,336	,731	3,847		2,662	1,892	30,922	2,121
17:Case 17	42,099	37,941	30,773	29,993	43,585	31,146	12,564	16,921	15,689	22,394	11,816	13,829	3,276	3,393	1,185	2,662		5,856	29,685	2,910
18:Case 18	50,033	43,938	37,559	36,629	49,887	31,429	20,032	21,174	20,884	29,666	21,714	20,029	3,927	1,161	6,019	1,892	5,856		36,199	2,992
19:Case 19	3,336	3,898	3,416	1,083	3,541	3,138	5,954	3,914	7,214	4,372	9,454	6,037	22,458	34,575	31,239	30,922	29,685	36,199		33,360
20:Case 20	45,275	38,921	32,962	33,791	44,595	30,006	16,845	17,425	15,003	24,027	15,248	15,495	2,236	2,852	1,726	2,121	2,910	2,992	33,360	

This is a dissimilarity matrix

Tabela 4: Matriz de proximidades.

A tabela 4 mostra a matriz de proximidades ou similaridade. Indica a distância entre os casos. Como é uma matriz simétrica podemos observar somente uma das diagonais. Por exemplo, a distância entre o caso 1 e o caso 2 é de 3,594; a distância entre 2 e 3 é 2,213.
O software apresenta apenas a primeira tabela do cálculo das distâncias. Conforme ocorrem os agrupamentos, novas distâncias são calculadas com base nessa matriz. Observe também que a menor distância da tabela é entre os casos 14 e 16 (0,365) que foram agrupados primeiramente.

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	14	16	,365	0	0	5
2	4	19	,907	0	0	13
3	9	12	1,495	0	0	11
4	15	17	2,087	0	0	12
5	14	18	2,983	1	0	17
6	2	3	4,090	0	0	10
7	13	20	5,208	0	0	12
8	1	5	6,347	0	0	13
9	7	10	7,611	0	0	14
10	2	6	9,069	6	0	16
11	8	9	10,883	0	3	15
12	13	15	12,784	7	4	17
13	1	4	15,382	8	2	16
14	7	11	18,992	9	0	15
15	7	8	23,053	14	11	18
16	1	2	27,255	13	10	18
17	13	14	31,508	12	5	19
18	1	7	47,523	16	15	19
19	1	13	152,000	18	17	0

Tabela 5: Matriz de aglomerações.

A tabela 5 também indica como os agrupamentos foram feitos de acordo com o método escolhido. É outra maneira de indicar o procedimento. Inicialmente o caso 14 e 16 foram agrupados, com distância considerada era de 0,365. Note que tanto o caso 14 quanto o 16 não tinham sido agrupados em nenhum cluster anteriormente, por isso aparece o número 0 nas colunas “cluster 1” e “cluster 2” (em “stage cluster first appears”).

Note que na coluna “next stage” aparece o número 5, isto porque, na linha 5 o caso 14 volta a aparecer, quando é agrupado com o caso 18. Observe que nessa mesma linha, na coluna “Cluster 1” aparece o número 1, pois indica a linha onde o caso 14 já tinha sido agrupado anteriormente, com o caso 16.

Cluster Membership

Case	3 Clusters
1:Case 1	1
2:Case 2	1
3:Case 3	1
4:Case 4	1
5:Case 5	1
6:Case 6	1
7:Case 7	2
8:Case 8	2
9:Case 9	2
10:Case 10	2
11:Case 11	2
12:Case 12	2
13:Case 13	3
14:Case 14	3
15:Case 15	3
16:Case 16	3
17:Case 17	3
18:Case 18	3
19:Case 19	1
20:Case 20	3

Tabela 6: Resultados da aglomeração.

A tabela 6 mostra o resultado final da aglomeração, isto é, em quais clusters está cada caso. Por exemplo, o caso 1 está no cluster 1; o caso 8 está no cluster 2; o caso 20 está no cluster 3 e assim por diante.

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * *
* * * *

 Dendrogram using Ward Method

                        Rescaled Distance Cluster Combine

   C A S E          0         5        10        15        20        25
 Label     Num  +---------+---------+---------+---------+---------+

 Case 14    14   -+
 Case 16    16   -+
 Case 18    18   -+-------------------------------------------------+
 Case 15    15   -+                                                 I
 Case 17    17   -+                                                 I
 Case 13    13   -+                                                 I
 Case 20    20   -+                                                 I
 Case 9      9   -+                                                 I
 Case 12    12   -+                                                 I
 Case 8      8   -+-----+                                           I
 Case 7      7   -+     I                                           I
 Case 10    10   -+     I                                           I
 Case 11    11   -+     +-----------------------------------------+
 Case 2      2   -+     I
 Case 3      3   -+     I
 Case 6      6   -+-----+
 Case 4      4   -+
 Case 19    19   -+
 Case 1      1   -+
 Case 5      5   -+
```

Gráfico 1: Dendrograma.

O gráfico 1 é o dendrograma. A escala vertical (situada à esquerda do gráfico) indica o nível de similaridade entre os casos, a escala horizontal mostra os casos fora de ordem, numa ordem que facilite a visualização dos agrupamentos. Os coeficientes de distância são reescalonados para valores entre 0 e 25.

Vamos entender os resultados da análise?

As saídas do software nos mostraram como os agrupamentos foram realizados e em quais clusters está cada caso. Mas o que isto quer dizer?
Observe o gráfico de pizza abaixo:

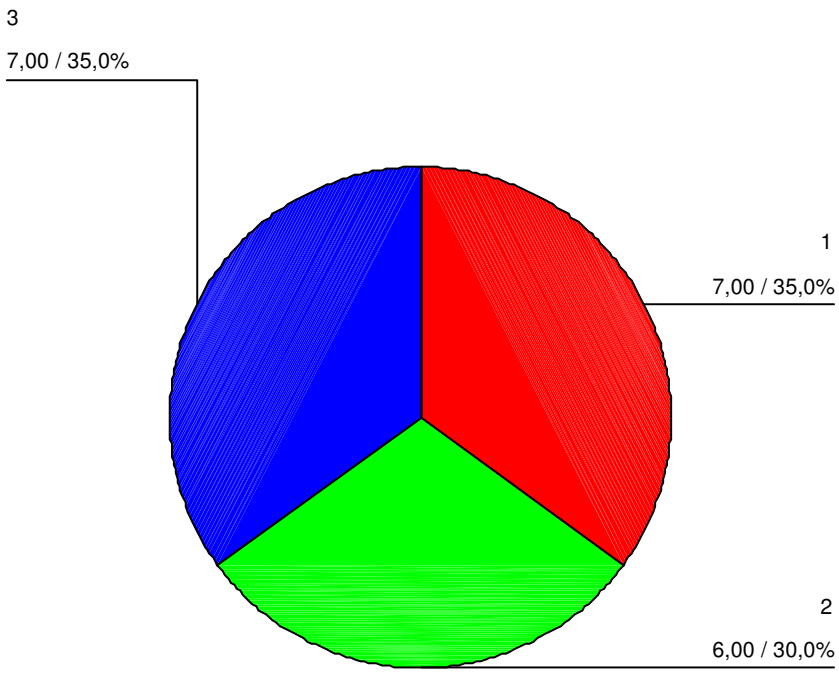


Gráfico 2: Porcentagem de casos em cada cluster.

Temos 35% dos casos no cluster 1; 30% no cluster 2 e 35% no cluster 3, os grupos possuem mais ou menos o mesmo tamanho.

Mas quais são as características de cada cluster?

Para relembrar as afirmações do questionário são as seguintes:

1) Só compro roupas quando realmente preciso	1	2	3	4	5
2) Uso as roupas que estão na moda	1	2	3	4	5
3) Compro roupas extravagantes mesmo não vá utilizá-las	1	2	3	4	5
4) Quando compro roupas costumo escolher também outras peças e acessórios para combinar	1	2	3	4	5
5) Costumo comprar mais roupas "curinga", fáceis de combinar	1	2	3	4	5
6) Quando gosto não me importo com o preço da peça	1	2	3	4	5
7) Procuro comprar sempre peças exclusivas	1	2	3	4	5
8) Só compro roupas se elas forem de marcas famosas	1	2	3	4	5

No tutorial "introdução ao SPSS" você aprendeu a calcular as estatísticas descritivas (statistics, summarize, descriptives), observe as tabelas descritivas abaixo:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q1	7	1,00	2,00	1,2857	,4880
Q2	7	1,00	3,00	2,1429	,6901
Q3	7	1,00	3,00	2,0000	,8165
Q4	7	1,00	2,00	1,5714	,5345
Q5	7	1,00	2,00	1,5714	,5345
Q6	7	1,00	2,00	1,4286	,5345
Q7	7	1,00	3,00	2,2857	,7559
Q8	7	1,00	2,00	1,4286	,5345
Valid N (listwise)	7				

Tabela 7: Estatísticas descritivas para o cluster 1.

A tabela 7 mostra as estatísticas para o cluster 1. Observamos que as clientes deste cluster atribuíram notas baixas para todas as questões, portanto, podemos classificá-las como clientes comedidas, que compram peças básicas e somente quando necessário.

Lembre-se que além de verificar a média, é importante analisar o quanto essa opinião é homogênea ou não para os respondentes. Em outras palavras, devemos verificar se existe uma grande variabilidade nos dados em relação à média. Para tanto, podemos utilizar o coeficiente de variação (CV= desvio-padrão/média), temos então:

Questão	Desvio-padrão	Média	Coeficiente de Variação (%)
Q1	0,4880	1,2857	38%
Q2	0,6901	2,1429	32%
Q3	0,8165	2,0000	41%
Q4	0,5345	1,5714	34%
Q5	0,5345	1,5714	34%
Q6	0,5345	1,4286	37%
Q7	0,7559	2,2857	33%
Q8	0,5345	1,4286	37%

Quadro 1: Coeficiente de variação cluster 1.

De maneira geral observa-se que este grupo é homogêneo, com uma variabilidade média nos dados.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q1	6	3,00	3,00	3,0000	,0000
Q2	6	2,00	4,00	3,1667	,7528
Q3	6	2,00	4,00	3,0000	,8944
Q4	6	2,00	3,00	2,5000	,5477
Q5	6	2,00	4,00	2,8333	,7528
Q6	6	1,00	3,00	2,1667	,7528
Q7	6	1,00	3,00	2,3333	,8165
Q8	6	2,00	3,00	2,5000	,5477
Valid N (listwise)	6				

Tabela 8: Estatísticas descritivas para o cluster 2.

A tabela 8 mostra as estatísticas para o cluster 2. Estas clientes deram notas por volta de 3 nas questões, assim, mostram indiferença, não adotam comportamento muito favorável nem muito contrário ao questionado. Temos para o cluster 2:

Questão	Desvio-padrão	Média	Coeficiente de Variação (%)
Q1	0,0000	3,0000	0%
Q2	0,7528	3,1667	24%
Q3	0,8944	3,0000	30%
Q4	0,5477	2,5000	22%
Q5	0,7528	2,8333	27%
Q6	0,7528	2,1667	35%
Q7	0,8165	2,3333	35%
Q8	0,5477	2,5000	22%

Quadro 2: Coeficientes de variação cluster 2.

De maneira geral observa-se que este grupo é homogêneo, com uma variabilidade média nos dados.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q1	7	4,00	5,00	4,5714	,5345
Q2	7	4,00	5,00	4,5714	,5345
Q3	7	4,00	5,00	4,5714	,5345
Q4	7	4,00	5,00	4,5714	,5345
Q5	7	3,00	5,00	4,0000	,8165
Q6	7	4,00	5,00	4,4286	,5345
Q7	7	4,00	5,00	4,2857	,4880
Q8	7	4,00	5,00	4,4286	,5345
Valid N (listwise)	7				

Tabela 9: Estatísticas descritivas para o cluster 3.

A tabela 9 mostra as estatísticas para o cluster 3. Estas clientes deram notas altas a todas as informações, podemos, então, classifica-las como preocupadas com a moda, que tentam sempre segui-la. Temos para o cluster 3:

Questão	Desvio-padrão	Média	Coeficiente de Variação (%)
Q1	0,5345	4,5714	12%
Q2	0,5345	4,5714	12%
Q3	0,5345	4,5714	12%
Q4	0,5345	4,5714	12%
Q5	0,8165	4,0000	20%
Q6	0,5345	4,4286	12%
Q7	0,4880	4,2857	11%
Q8	0,5345	4,4286	12%

Quadro 3: Coeficientes de variação cluster 3.

Neste cluster observamos a maior homogeneidade de opiniões entre os respondentes, pois existe pouca variabilidade dos dados em relação à média.

Observe o gráfico abaixo:

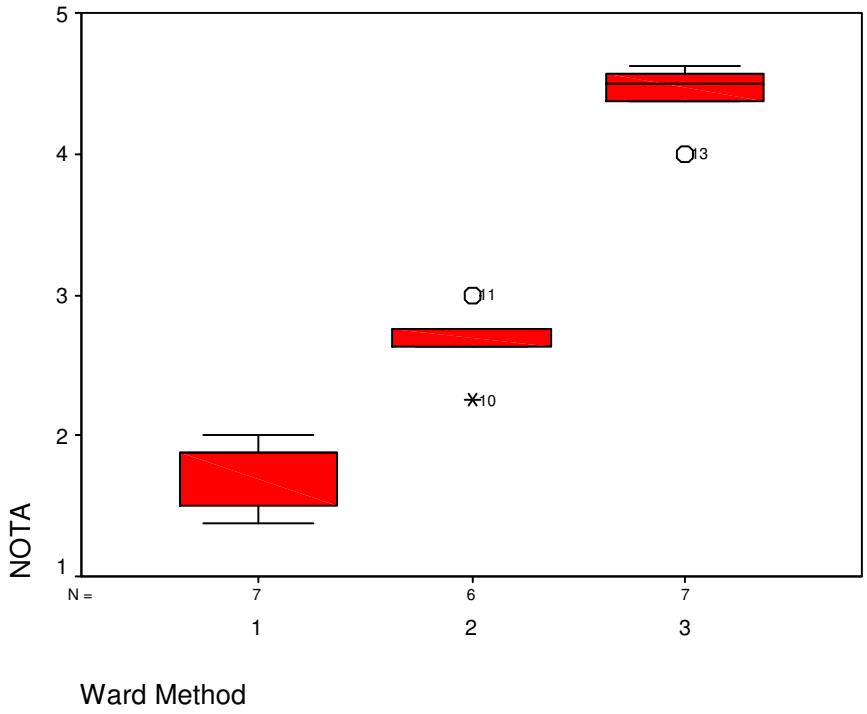


Gráfico 3: Box plot.

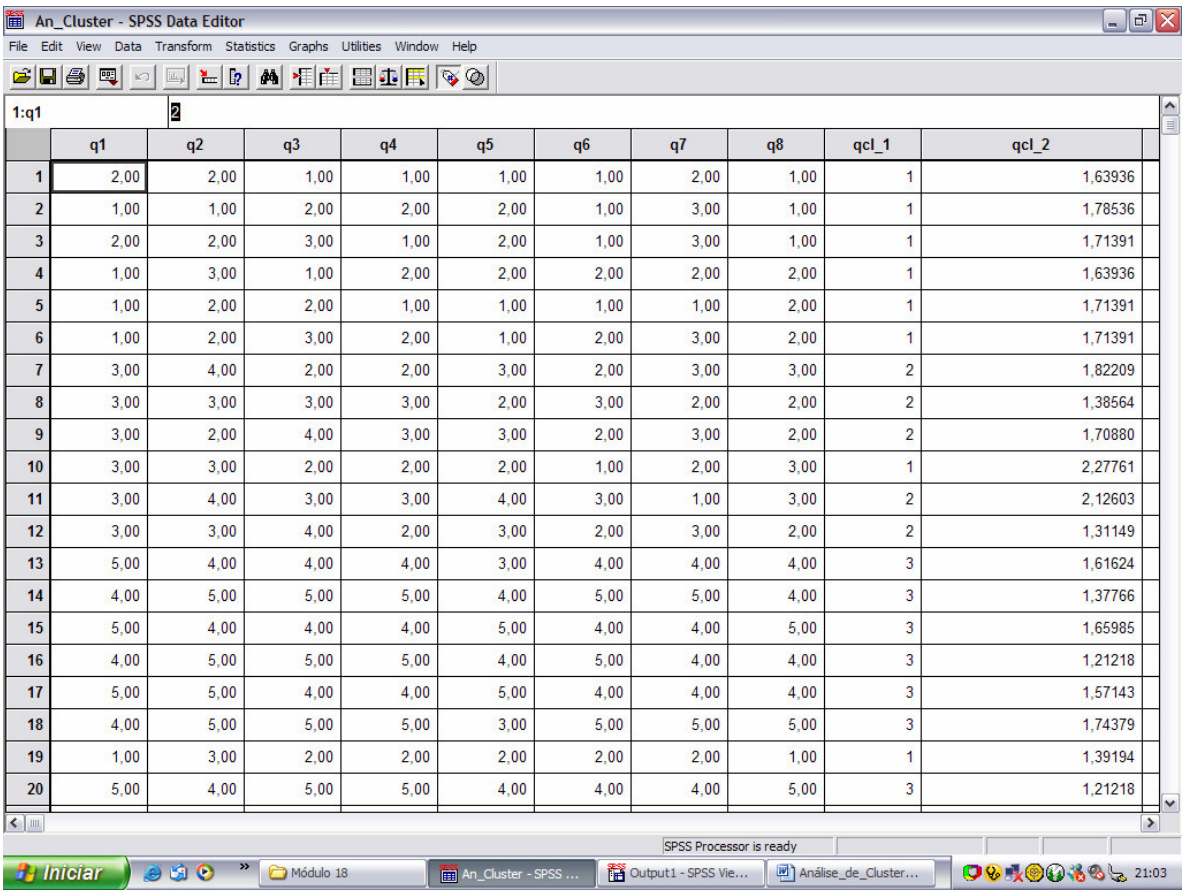
O box plot apresenta graficamente a relação dos clusters com as notas médias nas questões. Podemos observar que o cluster 1 atribui as notas mais baixas para as questões, o cluster 2 possui notas em torno de 3, ou seja, com uma tendência a indiferença e o cluster 3 possui as notas mais altas, mostrando concordância com as questões.

Atenção! Estas tabelas não são outputs da análise de cluster. São análises adicionais realizadas para entendermos melhor os resultados do agrupamento.

Método não hierárquico

Após ter finalizado o processo de obtenção das tabelas (tutorial anterior), observe que temos duas novas variáveis na planilha de dados. A variável qcl_1 mostra que os casos 1, 2, 3, 4, 5, 6, 10 e 19 pertencem ao cluster 1; os casos 7, 8, 9, 11, 12 pertencem ao cluster 2; os casos 13, 14, 15, 16, 17, 18, 20 pertencem ao cluster 3.

A variável qcl_2 mostra distância de cada caso ao centro do seu cluster. Por exemplo, o caso 1, que pertence ao cluster 1 está a uma distância de 1,63936 do centro do cluster 1.



	q1	q2	q3	q4	q5	q6	q7	q8	qcl_1	qcl_2
1	2,00	2,00	1,00	1,00	1,00	1,00	2,00	1,00	1	1,63936
2	1,00	1,00	2,00	2,00	2,00	1,00	3,00	1,00	1	1,78536
3	2,00	2,00	3,00	1,00	2,00	1,00	3,00	1,00	1	1,71391
4	1,00	3,00	1,00	2,00	2,00	2,00	2,00	2,00	1	1,63936
5	1,00	2,00	2,00	1,00	1,00	1,00	1,00	2,00	1	1,71391
6	1,00	2,00	3,00	2,00	1,00	2,00	3,00	2,00	1	1,71391
7	3,00	4,00	2,00	2,00	3,00	2,00	3,00	3,00	2	1,82209
8	3,00	3,00	3,00	3,00	2,00	3,00	2,00	2,00	2	1,38564
9	3,00	2,00	4,00	3,00	3,00	2,00	3,00	2,00	2	1,70880
10	3,00	3,00	2,00	2,00	2,00	1,00	2,00	3,00	1	2,27761
11	3,00	4,00	3,00	3,00	4,00	3,00	1,00	3,00	2	2,12603
12	3,00	3,00	4,00	2,00	3,00	2,00	3,00	2,00	2	1,31149
13	5,00	4,00	4,00	4,00	3,00	4,00	4,00	4,00	3	1,61624
14	4,00	5,00	5,00	5,00	4,00	5,00	5,00	4,00	3	1,37766
15	5,00	4,00	4,00	4,00	5,00	4,00	4,00	5,00	3	1,65985
16	4,00	5,00	5,00	5,00	4,00	5,00	4,00	4,00	3	1,21218
17	5,00	5,00	4,00	4,00	5,00	4,00	4,00	4,00	3	1,57143
18	4,00	5,00	5,00	5,00	3,00	5,00	5,00	5,00	3	1,74379
19	1,00	3,00	2,00	2,00	2,00	2,00	2,00	1,00	1	1,39194
20	5,00	4,00	5,00	5,00	4,00	4,00	4,00	5,00	3	1,21218

Figura 9: Variáveis criadas pelo software.

Parte 2- Análise dos resultados

Initial Cluster Centers

	Cluster		
	1	2	3
Q1	1,00	3,00	4,00
Q2	2,00	4,00	5,00
Q3	2,00	3,00	5,00
Q4	1,00	3,00	5,00
Q5	1,00	4,00	3,00
Q6	1,00	3,00	5,00
Q7	1,00	1,00	5,00
Q8	2,00	3,00	5,00

Tabela 1: Centróides iniciais.

A tabela 1 mostra as sementes iniciais selecionadas pelo software para iniciar a análise. O cluster 1 possui o centro igual a 1, o cluster 2 possui centro igual a 3, o cluster 3 possui centro igual a 4 para a variável “q1” .

Cluster Membership

Case Number	Cluster	Distance
1	1	1,639
2	1	1,785
3	1	1,714
4	1	1,639
5	1	1,714
6	1	1,714
7	2	1,822
8	2	1,386
9	2	1,709
10	1	2,278
11	2	2,126
12	2	1,311
13	3	1,616
14	3	1,378
15	3	1,660
16	3	1,212
17	3	1,571
18	3	1,744
19	1	1,392
20	3	1,212

Tabela 2: Membros pertencentes a cada cluster e distância dos casos ao centróide.

A tabela 2 mostra quais casos pertencem aos clusters e a distância de cada caso ao centro do seu cluster.

Final Cluster Centers

	Cluster		
	1	2	3
Q1	1,50	3,00	4,57
Q2	2,25	3,20	4,57
Q3	2,00	3,20	4,57
Q4	1,63	2,60	4,57
Q5	1,63	3,00	4,00
Q6	1,38	2,40	4,43
Q7	2,25	2,40	4,29
Q8	1,63	2,40	4,43

Tabela 3: Centróides finais.

A tabela 3 mostra o centro final dos clusters para cada variável. Estes valores correspondem à média dos valores do caso que pertencem ao cluster. Por exemplo, para a variável 1 no cluster 1 vamos fazer a média aritmética dos valores dados pelas clientes à questão 1 (da planilha de dados):

$$Centro\ cluster1\ q1 = \frac{2+1+2+1+1+1+3+1}{8} = 1{,}50$$

Distances between Final Cluster Centers

Cluster	1	2	3
1		3,018	7,555
2	3,018		4,786
3	7,555	4,786	

Tabela 4: Distância entre os centróides dos clusters.

A tabela 4 mostra a distância entre os centros dos clusters. O centro do cluster 1 está a uma distância de 3,018 do centro do cluster 2 e está a uma distância de 7,555 do centro do cluster 3. O centro do cluster 2 está a uma distância de 4,786 do centro do cluster 3. Percebemos então que o cluster 1 e o 3 estão mais distantes e o cluster 1 e o 2 são os mais próximos, ou seja, o cluster 1 é mais diferente do cluster 3 e está mais parecido com o cluster 2.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Q1	17,618	2	,336	17	52,413	,000
Q2	10,093	2	,471	17	21,409	,000
Q3	12,343	2	,501	17	24,644	,000
Q4	16,505	2	,282	17	58,587	,000
Q5	10,662	2	,463	17	23,017	,000
Q6	17,705	2	,282	17	62,847	,000
Q7	8,936	2	,478	17	18,688	,000
Q8	15,205	2	,399	17	38,073	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Tabela 5: Tabela ANOVA.

A tabela 5 mostra os resultados da análise de variância. A média dos quadrados entre é representada na coluna "cluster- mean square" com (m-1) graus de liberdade, onde m=número de clusters; a média dos quadrados dentro é representada na coluna "error- mean square".

Não devemos interpretar os valores do nível de significância observado, pois esta interpretação é equivocada. O teste F é apenas descritivo. Verifica-se que o maior F calculado é o da questão 6 (62,847), sendo provavelmente esta a questão que mais diferencia os clusters, em seguida tem o maior valor F para a questão 4 (58,587) o que indica que esta questão pode ser a segunda que mais diferencia os clusters e assim por diante.

Number of Cases in each Cluster

Cluster	1	8,000
	2	5,000
	3	7,000
Valid		20,000
Missing		,000

Tabela 6: Número de casos em cada cluster.

A tabela 6 mostra o número de casos nos clusters. Temos 8 clientes no cluster 1, 5 clientes no cluster 2 e 7 clientes no cluster 3.

Observe que o resultado obtido é ligeiramente diferente do resultado com o método hierárquico de agrupamento. Por isso é tão importante possuir respaldo teórico ao escolher o tipo de método a ser utilizado.

Vamos entender os resultados da análise?

As saídas do software nos mostraram como os agrupamentos foram realizados e em quais clusters está cada caso. Mas o que isto quer dizer?

Observe o gráfico de pizza abaixo:

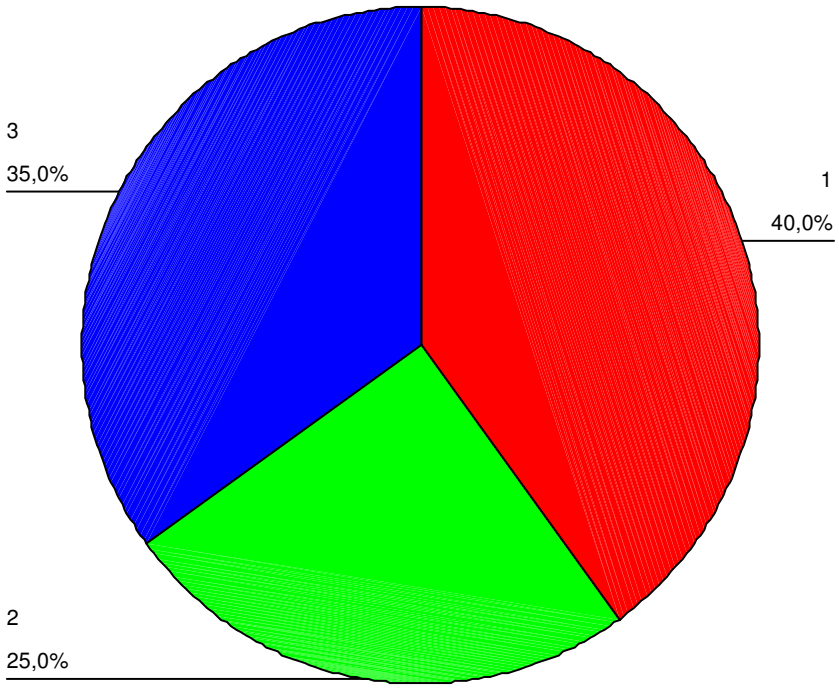


Gráfico 1: Porcentagem de casos em cada cluster.

Temos 40% dos casos no cluster 1; 25% no cluster 2 e 35% no cluster 3, ou seja, o maior grupo é o cluster 1 e o menor é o cluster 2.

Mas quais são as características de cada cluster?

Para relembrar as afirmações do questionário são as seguintes:

1) Só compro roupas quando realmente preciso	1	2	3	4	5
2) Uso as roupas que estão na moda	1	2	3	4	5
3) Compro roupas extravagantes mesmo não vá utilizá-las	1	2	3	4	5
4) Quando compro roupas costumo escolher também outras peças e acessórios para combinar	1	2	3	4	5
5) Costumo comprar mais roupas "curinga", fáceis de combinar	1	2	3	4	5
6) Quando gosto não me importo com o preço da peça	1	2	3	4	5
7) Procuro comprar sempre peças exclusivas	1	2	3	4	5
8) Só compro roupas se elas forem de marcas famosas	1	2	3	4	5

No tutorial "introdução ao SPSS" você aprendeu a calcular as estatísticas descritivas (statistics, summarize, descriptives), observe as tabelas descritivas abaixo:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q1	8	1,00	3,00	1,5000	,7559
Q2	8	1,00	3,00	2,2500	,7071
Q3	8	1,00	3,00	2,0000	,7559
Q4	8	1,00	2,00	1,6250	,5175
Q5	8	1,00	2,00	1,6250	,5175
Q6	8	1,00	2,00	1,3750	,5175
Q7	8	1,00	3,00	2,2500	,7071
Q8	8	1,00	3,00	1,6250	,7440
Distance of Case from its Classification Cluster Center	8	1,39194	2,27761	1,7344209	,2490767
Valid N (listwise)	8				

Tabela 7: Estatísticas descritivas para o cluster 1.

A tabela 7 mostra as estatísticas para o cluster 1. Observamos que as clientes deste cluster atribuíram notas baixas para todas as questões, portanto, podemos classificá-las como clientes comedidas, que compram peças básicas e somente quando necessário.

Lembre-se que além de verificar a média, é importante analisar o quanto essa opinião é homogênea ou não para os respondentes. Em outras palavras, devemos verificar se existe uma grande variabilidade nos dados em relação à média. Para tanto, podemos utilizar o coeficiente de variação (CV= desvio-padrão/média), temos então:

Questão	Desvio-padrão	Média	Coeficiente de Variação (%)
Q1	0,7559	1,5000	50%
Q2	0,7071	2,2500	31%
Q3	0,7559	2,0000	38%
Q4	0,5175	1,6250	32%
Q5	0,5175	1,6250	32%
Q6	0,5175	1,3750	38%
Q7	0,7071	2,2500	31%
Q8	0,7440	1,6250	46%

Quadro 1: Coeficiente de variação cluster 1.

De maneira geral observa-se que este grupo possui variabilidade relativamente alta, o que indica que as opiniões entre as clientes que compõem o cluster não são muito homogêneas.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q1	5	3,00	3,00	3,0000	,0000
Q2	5	2,00	4,00	3,2000	,8367
Q3	5	2,00	4,00	3,2000	,8367
Q4	5	2,00	3,00	2,6000	,5477
Q5	5	2,00	4,00	3,0000	,7071
Q6	5	2,00	3,00	2,4000	,5477
Q7	5	1,00	3,00	2,4000	,8944
Q8	5	2,00	3,00	2,4000	,5477
Distance of Case from its Classification Cluster Center	5	1,31149	2,12603	1,6708090	,3324103
Valid N (listwise)	5				

Tabela 8: Estatísticas descritivas para o cluster 2.

A tabela 8 mostra as estatísticas para o cluster 2. Estas clientes deram notas por volta de 2,5 nas questões, assim, mostram opinião tendendo à indiferença, não adotam comportamento muito favorável nem muito contrário ao questionado. Temos para o cluster 2:

Questão	Desvio-padrão	Média	Coeficiente de Variação (%)
Q1	0,0000	3,0000	0%
Q2	0,8367	3,2000	26%
Q3	0,8367	3,2000	26%
Q4	0,5477	2,6000	21%
Q5	0,7071	3,0000	24%
Q6	0,5477	2,4000	23%
Q7	0,8944	2,4000	37%
Q8	0,5477	2,4000	23%

Quadro 2: Coeficientes de variação cluster 2.

De maneira geral observa-se que este grupo é homogêneo, com uma variabilidade mediana nos dados, ou seja, não há muita variação entre as opiniões das clientes classificadas neste grupo.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q1	7	4,00	5,00	4,5714	,5345
Q2	7	4,00	5,00	4,5714	,5345
Q3	7	4,00	5,00	4,5714	,5345
Q4	7	4,00	5,00	4,5714	,5345
Q5	7	3,00	5,00	4,0000	,8165
Q6	7	4,00	5,00	4,4286	,5345
Q7	7	4,00	5,00	4,2857	,4880
Q8	7	4,00	5,00	4,4286	,5345
Distance of Case from its Classification Cluster Center	7	1,21218	1,74379	1,4847638	,2170331
Valid N (listwise)	7				

Tabela 9: Estatísticas descritivas para o cluster 3.

A tabela 9 mostra as estatísticas para o cluster 3. Estas clientes deram notas altas a todas as informações, podemos, então, classifica-las como preocupadas com a moda, que tentam sempre segui-la. Temos para o cluster 3:

Questão	Desvio-padrão	Média	Coeficiente de Variação (%)
Q1	0,5345	4,5714	12%
Q2	0,5345	4,5714	12%
Q3	0,5345	4,5714	12%
Q4	0,5345	4,5714	12%
Q5	0,8165	4,0000	20%
Q6	0,5345	4,4286	12%
Q7	0,4880	4,2857	11%
Q8	0,5345	4,4286	12%

Quadro 3: Coeficientes de variação cluster 3.

Neste cluster observamos a maior homogeneidade de opiniões entre os respondentes, pois existe pouca variabilidade dos dados em relação à média.

Observe o seguinte gráfico:

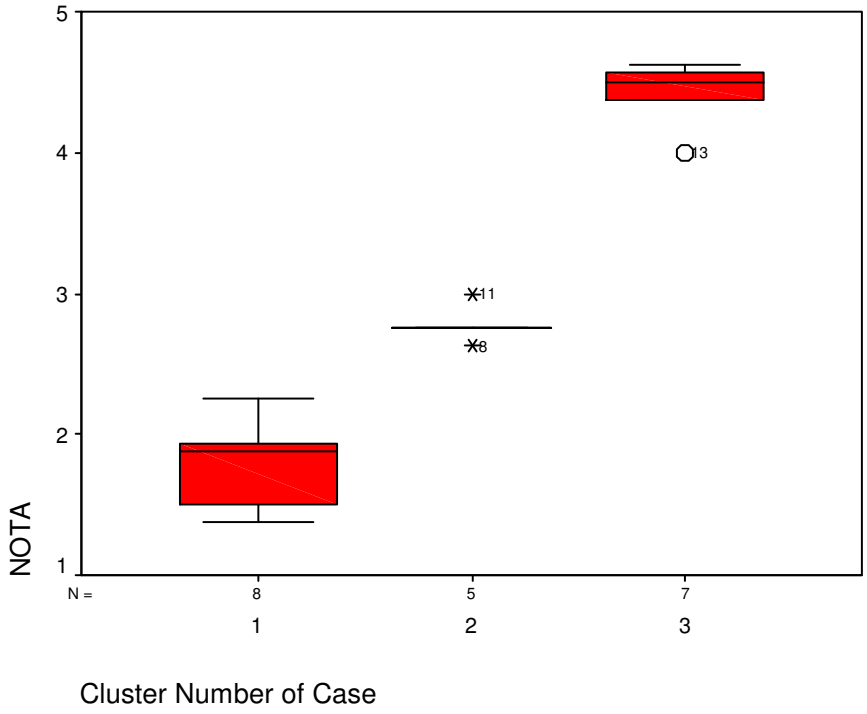


Gráfico 2: Box plot.

O box plot apresenta graficamente a relação dos clusters com as notas médias nas questões. Podemos observar que o cluster 1 atribui as notas mais baixas para as questões, o cluster 2 possui notas em torno de 3, ou seja, com uma tendência a indiferença e o cluster 3 possui as notas mais altas, mostrando concordância com as questões.

Atenção! Estas tabelas não são outputs da análise de cluster. São análises adicionais realizadas para entendermos melhor os resultados do agrupamento.