Análise de Agrupamentos

Caroline Vasconcelos

Recordes Nacionais de Corrida Feminina

Procedimentos rudimentares de exploração de dados são quase sempre de grande ajuda para entender a complexidade da natureza das relações multivariadas. Analisar um grande conjunto de dados a procura de agrupamentos que sejam naturais é uma importante técnica de análise exploratória e tais agrupamentos podem oferecer formas de acessar dimensionalidades, identificação de anomalias (outliers), sugerir hipóteses interessantes sobre a relação entre as variáveis, classificação não supervisionada que é útil quando não há conhecimento prévio sobre categorias, segmentação de mercado em aplicações de negócios, facilitando a personalização de estratégias de marketing.

Agrupamento, ou clusterização (clustering), é uma análise mais primitiva em que não são feitas suposições prévias sobre o número de grupos das variáveis ou sobre a estrutura dos mesmos, o agrupamento é feito através de similaridades ou dissimilaridades (distâncias). Em aplicações práticas da análise de clusters, o investigador conhece o problema o suficiente para saber se determinado agrupamento é "bom" ou "ruim".

A análise de agrupamentos deste estudo foi feito com dados do National Track Records for Women, inclui as seguintes variáveis:

x_1: 100m (s);

x_2: 200m (s);

x_3: 400m (s);

x_4: 800m (min);

x_5: 1500m (min);

x_6: 3000m (min);

x_7: Maratona (min).

A figura 1 apresenta o gráfico com correlações, onde podemos observar que as variáveis possuem correlações altas e positivas. Nos gráficos de dispersão, vemos alguns pontos fora da nuvem de pontos e numa análise para detecção de valores atípico, foram encontrados 4 observações

(COK, KOR.N, PNG e SAM) atípicas que foram removidas do conjunto de dados. Após a remoção dos outliers, foi feito o teste shapiro-wilk para dados multivariados que forneceu p-valor igual a 0,061 com nível de significância de 95%, não rejeitamos a hipótese de que os dados tem distribuição normal. Na figura 2, temos as correlações, gráficos de dispersão e densidade após remoção de discrepâncias.

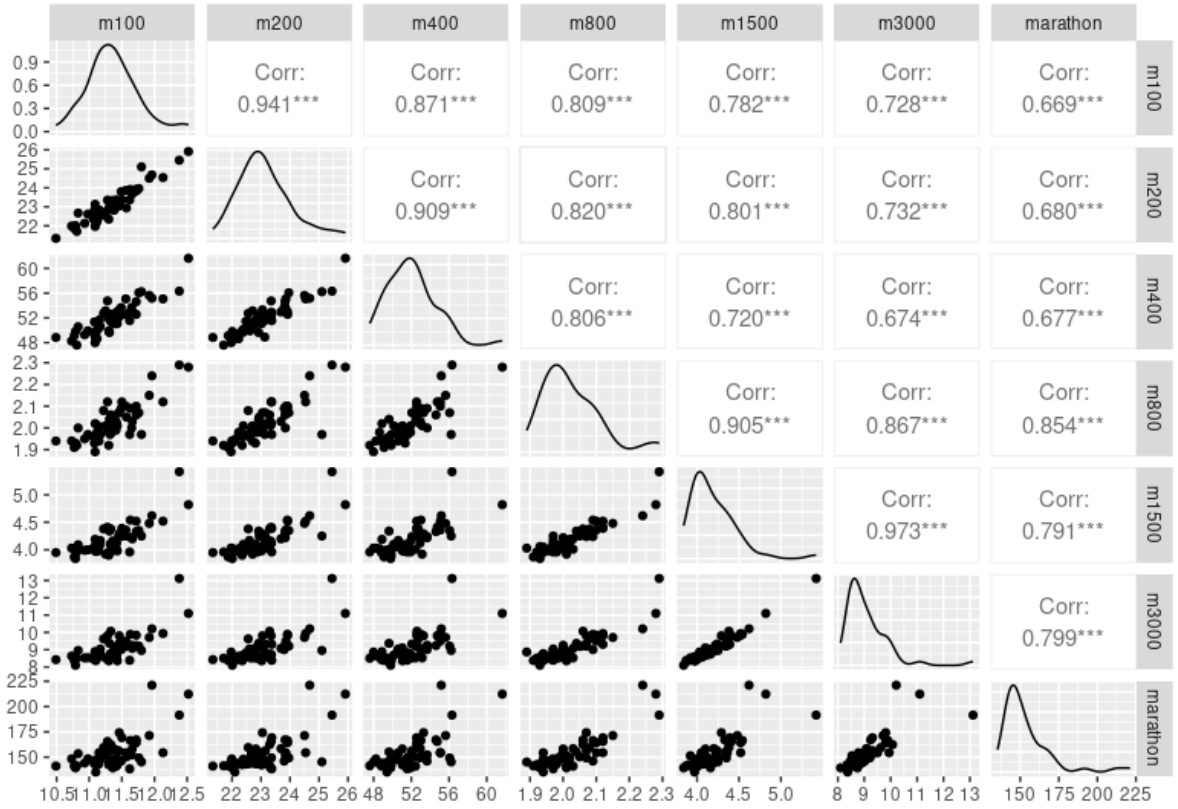

Figura 1: Correlações, gráficos de densidade e dispersão.

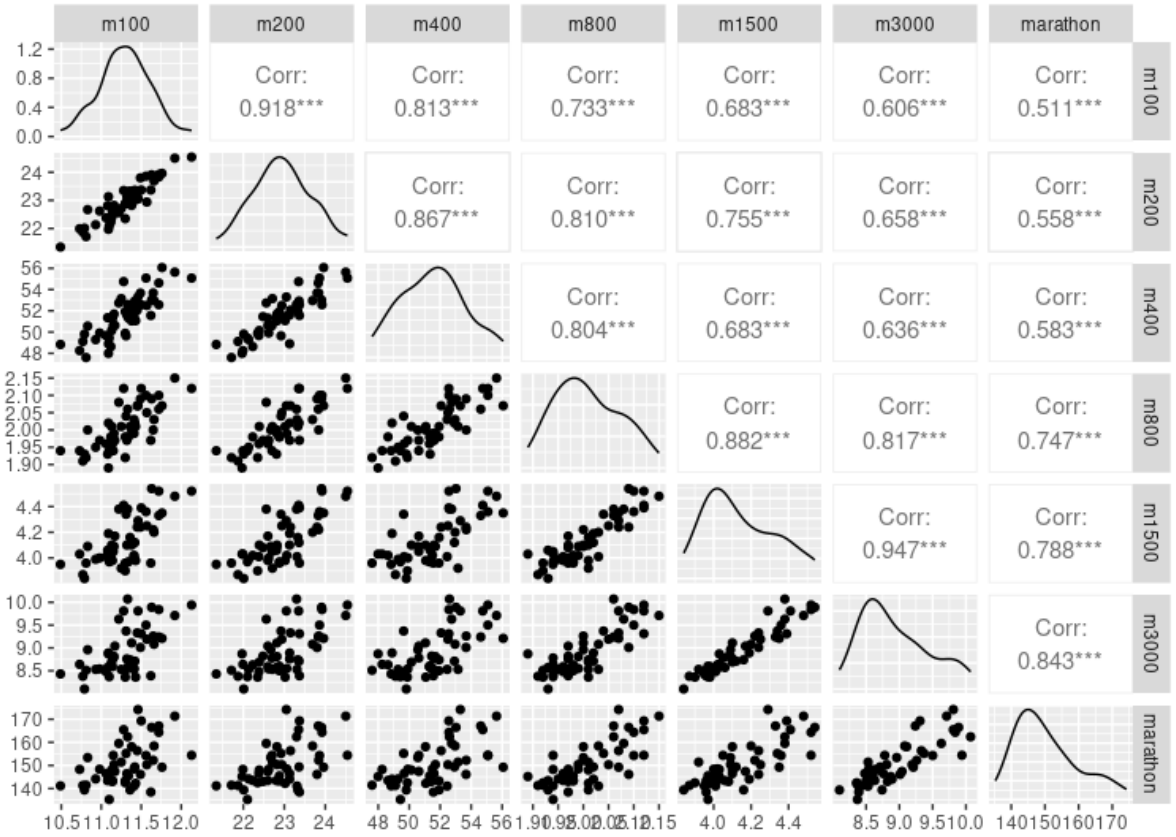

Figura 2: Correlações, gráficos de densidade e dispersão após remoção de outliers.

Nesta análise, prosseguimos com os dados contendo as observações atípicas, considerando que pode ser interessante observar como eles irão se agrupar. Vale ressaltar que na maioria dos métodos de agrupamento, as fontes de erro e variação não são formalmente consideradas em **procedimentos hierárquicos**. Isso significa que um método de agrupamento será sensível a valores discrepantes ou pontos de ruído (Johnson, 2007).

Quando se pensa em construir uma estrutura de agrupamento simples a partir de um conjunto de dados complexo, na maior parte das vezes, leva-se em consideração uma medida de proximidade ou similaridade. Muitas vezes há muita subjetividade envolvida na escolha da medida de similaridade. Considerações importantes incluem a natureza das variáveis (discretas, contínuas, binárias), escalas de medição (nominal, ordinal, intervalar) e conhecimento do assunto.

A distância euclidiana é frequentemente preferida para agrupamento. As distâncias euclidianas foram calculadas para os pares de países, parte delas podem ser visualizadas na figura 3.

	ARG	AUS	AUT	BEL	BER	BRA	CAN	CHI	CHN	COL	COK	CRC	CZE	DE
ARG	0.000000	7.9013290	4.497722	7.3840436	23.882006	3.509045	3.3899263	2.421817	11.381779	5.657305	62.79102	14.062923	6.928542	1.
AUS	7.901329	0.0000000	11.032339	2.8751870	31.059929	4.415903	5.0340044	10.239687	4.340495	11.773203	70.20004	21.305882	1.834748	7.
AUT	4.497722	11.0323388	0.000000	11.3327314	20.044084	6.946719	6.0399089	3.958598	15.018988	1.477227	59.17443	10.324064	9.558766	5.
BEL	7.384044	2.8751870	11.332731	0.0000000	31.208347	4.446381	5.5405505	9.569174	4.123857	12.298504	70.16602	21.395179	4.101853	6.
BER	23.882006	31.0599292	20.044084	31.2083466	0.000000	26.920609	26.0823082	21.975427	35.031427	19.345596	39.19741	9.921421	29.511067	24.
BRA	3.509045	4.4159031	6.946719	4.4463806	26.920609	0.000000	1.3122500	5.866498	8.152490	7.858340	65.98339	17.114707	3.507635	3.
CAN	3.389926	5.0340044	6.039909	5.5405505	26.082308	1.312250	0.0000000	5.629813	9.005548	6.908697	65.19591	16.319663	3.779696	3.
CHI	2.421817	10.2396875	3.958598	9.5691745	21.975427	5.866498	5.6298135	0.000000	13.617118	5.106202	60.69923	12.169215	9.270841	3.
CHN	11.381779	4.3404954	15.018988	4.1238574	35.031427	8.152490	9.0055483	13.617118	0.000000	15.894729	74.08586	25.251794	6.137964	10.
COL	5.657305	11.7732026	1.477227	12.2985040	19.345596	7.858340	6.9086974	5.106202	15.894729	0.000000	58.50560	9.672099	10.200098	6.
COK	62.791018	70.2000434	59.174430	70.1660160	39.197414	65.983393	65.1959071	60.699231	74.085862	58.505596	0.00000	48.915810	68.685463	63.
CRC	14.062923	21.3058818	10.324064	21.3951794	9.921421	17.114707	16.3196630	12.169215	25.251794	9.672099	48.91581	0.000000	19.817681	15.
CZE	6.928542	1.8347479	9.558766	4.1018532	29.511067	3.507635	3.7796958	9.270841	6.137964	10.200098	68.68546	19.817681	0.000000	6.
DEN	1.258531	7.3333417	5.560108	6.5264079	24.869843	3.125524	3.2874762	3.086406	10.552374	6.758032	63.70192	15.055547	6.604468	0.
DOM	16.195376	23.4718384	12.474025	23.5403675	7.779113	19.273401	18.4796429	14.263282	27.406647	11.825088	46.74296	2.179587	21.979752	17.
FIN	3.423770	4.7409704	6.376904	5.1230557	26.404859	0.867871	0.5403702	5.756457	8.655813	7.265005	65.50993	16.636400	3.570476	3.
FRA	4.923017	4.7972388	6.579042	6.1589285	26.456389	2.665089	1.7985828	7.061126	9.034368	7.187406	65.63997	16.822506	3.120801	5.
GER	10.263567	2.3859170	13.293137	4.2658293	33.283365	6.774304	7.3448077	12.611653	3.066056	13.984338	72.45356	23.580127	3.797512	9.
GBR	15.434351	8.3040111	19.151060	8.0797215	39.163080	12.248526	13.1308035	17.623030	4.180682	19.987774	78.20152	29.354587	10.057351	14.
GRE	3.735171	10.0959844	1.023230	10.3955471	20.991062	6.001491	5.1024014	3.640261	14.074331	2.139065	60.11886	11.261834	8.636573	4.
GUA	21.311244	28.8147931	17.845100	28.6877639	3.995585	24.538066	23.8013340	19.221152	32.629428	17.306513	41.49136	7.669570	27.384076	22.
HUN	2.195108	5.8237874	5.937205	5.4946519	25.775413	1.579462	1.7108770	4.458195	9.358077	7.001614	64.76012	15.952921	4.967816	1.
INA	4.834718	12.7101652	4.695636	11.9227933	19.988709	8.320475	8.0520743	2.509841	16.001800	5.600946	58.47850	10.357225	11.722670	5.

Showing 1 to 23 of 54 entries, 54 total columns

Console

Figura 3: Distâncias euclidianas entre os pares de países.

Retornando aos procedimentos hierárquicos, as técnicas de agrupamento hierárquico procedem como uma série de fusões sucessivas ou como uma série de divisões sucessivas. **Métodos hierárquicos aglomerativos** começam com os objetos individuais, portanto, inicialmente existem muitos clusters como objetos. Os objetos que tem maiores similaridades são primeiro agrupados e esses grupos iniciais são mesclados de acordo com suas semelhanças, no caso, as distâncias euclidianas vistas anteriormente. Eventualmente, à medida que a similaridade diminui, todos os subgrupos são fundidos em um único cluster. **Métodos hierárquicos divisivos** funcionam na direção oposta: um único grupo inicial de objetos é dividido em dois subgrupos, de modo que os objetos de um subgrupo estejam "longe" dos objetos do outro. Esses subgrupos são então divididos em subgrupos diferentes; o processo continua até que haja tantos subgrupos quanto objetos, isto é, até que cada objeto forme um grupo.

Os resultados dos métodos aglomerativo e divisivo podem ser exibidos na forma de um diagrama bidimensional conhecido como **dendograma**, que é um gráfico que ilustra as fusões ou divisões que foram feitas em níveis sucessivos.

MÉTODO HIERÁRQUICO

O método hierárquico utilizado nesta análise foi o hierárquico aglomerativo. Para tanto, os países foram agrupados usando dois tipos de ligação (*linkage methods*): simples e completa. A **ligação simples** acontece quando os grupos são fundidos de acordo com a distância entre seus membros mais próximos. A **ligação completa** ocorre quando os grupos são fundidos de acordo com a distância entre os membros mais distantes.

As análises a seguir foram feitas usando os pacotes factoextra e NbClust.

Ligação Simples

A correlação entre a distância cofenética e a distância original quando usamos o método da ligação simples foi igual a 0,89 e os agrupamentos deram-se da seguinte forma:

Table 1: Agrupamentos com ligação simples.

Clusters	1	2	3
No. de países	51	2	1

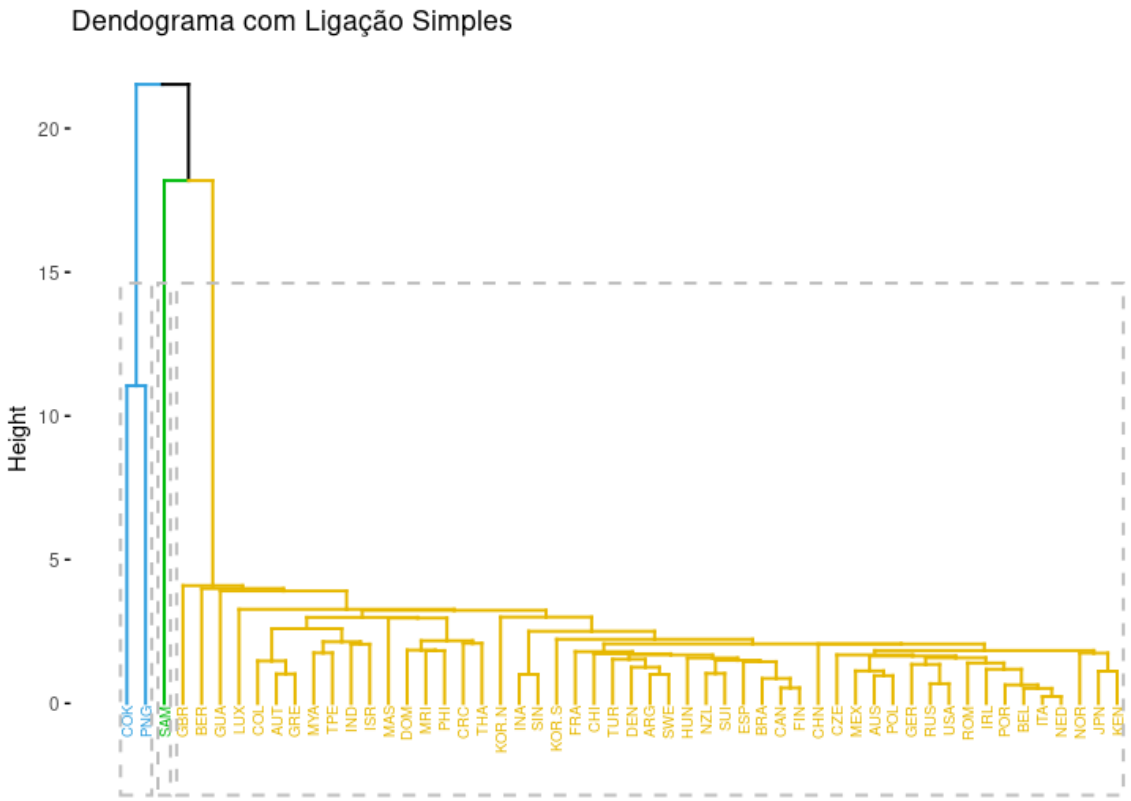

Figura 4: Dendograma usando o método da ligação simples.

Ligação completa

A correlação entre a distância cofenética e a distância original quando usamos o método da ligação simples foi igual a 0,87 e os agrupamentos deram-se da seguinte forma:

Clusters	1	2	3
No. de países	37	17	3

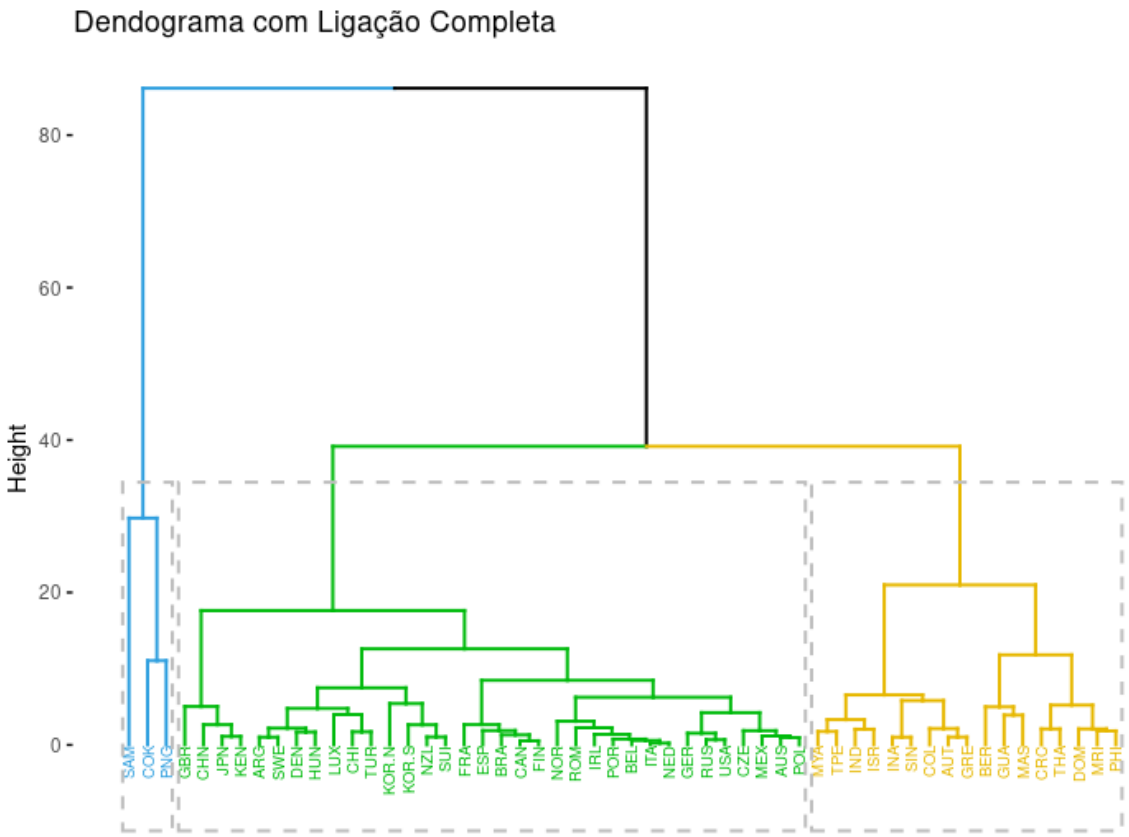

Figura 5: Dendograma usando o método da ligação completa.

Ligação Simples x Ligação Completa

Usando o pacote NbClust, função NbClust e parâmetros de mínimo igual a 2 e máximo igual a 7, tanto para o método da ligação simples quanto para a completa, ele propôs, para ambos,

3 clusters. Quando observamos os dois dendogramas, notamos que o dendograma da ligação completa consegue agrupar os países de forma mais equilibrada.

MÉTODO NÃO-HIERÁRQUICO

As técnicas de agrupamento não-hierárquico são projetadas para agrupar itens, em vez de variáveis, em uma coleção de K clusters. Os grupos de números K, podem ser especificados antecipadamente ou determinados como parte do procedimento de agrupamento.

O termo K-means descreve um algoritmo que atribui cada item ao cluster que possui o centróide (média) mais próximo. Na sua versão mais simples, o processo é composto por estas três etapas:

1. Particione os itens em K clusters iniciais;

2. Percorra a lista de itens, atribuindo um item ao cluster cujo centróide está mais próximo. Recalcule o centróide para o cluster que recebe o novo item e para o cluster que perde o item;

3. Repita a etapa 2 até que não ocorra mais nenhuma reatribuição.

Em vez de começar com uma partição de todos os itens em K grupos preliminares na etapa 1, poderíamos especificar K centróides iniciais (pontos sementes) e então prosseguir para a etapa 2. A atribuição final de itens aos clusters será, até certo ponto, dependente na partição inicial ou na seleção inicial de pontos iniciais.

É importante ressaltar que existem fortes argumentos para não fixar um número K de clusters:

1. Se dois ou mais pontos sementes estiverem inadvertidamente dentro de um único cluster, seus clusters resultantes serão pouco diferenciados;

2. A existência de um outlier pode produzir pelo menos um grupo com itens muito dispersos;

3. Mesmo que se saiba que a população consiste em K grupos, o método de amostragem pode ser tal que os dados do grupo mais raro não apareçam na amostra. Forçar os dados em grupos K levaria a clusters sem sentido. Nos casos em que uma única execução do algoritmo exige que o usuário especifique K, é sempre uma boa ideia executar novamente o algoritmo para várias escolhas.

Usando o pacote factoextra, podemos analisar quantos cluster são recomendados através do gráfico de otimização da figura 6.

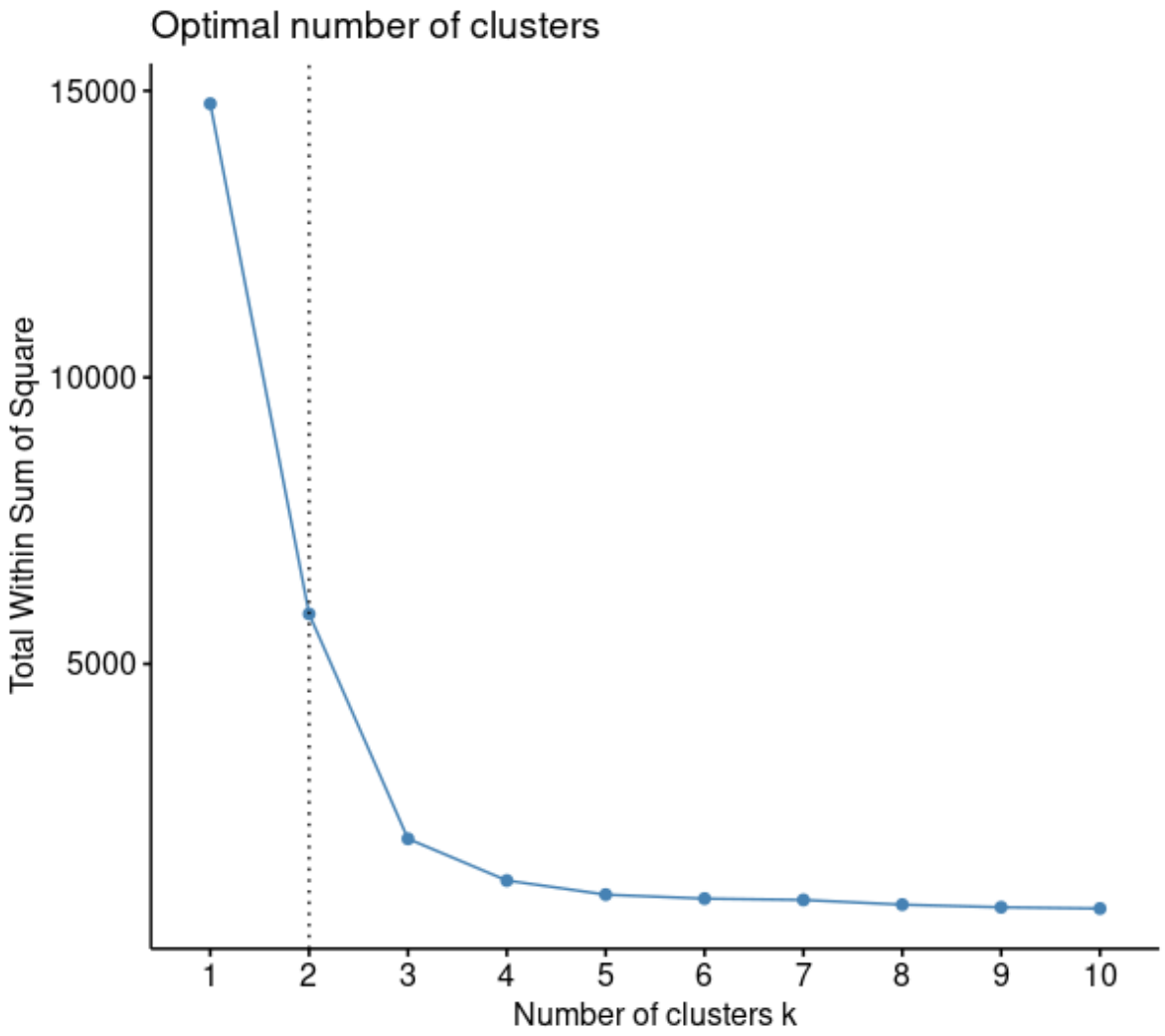

Figura 6: Otimização do número de clusters.

Fixada a semente, o algoritmo foi reproduzido três vezes com números de clusters 2, 3, 4.

Para K = 2

Quando K é igual a 2, ou seja, dois centróides, dois clusters, os países ficam divididos em um grupo de 45 e outro de 9 países.

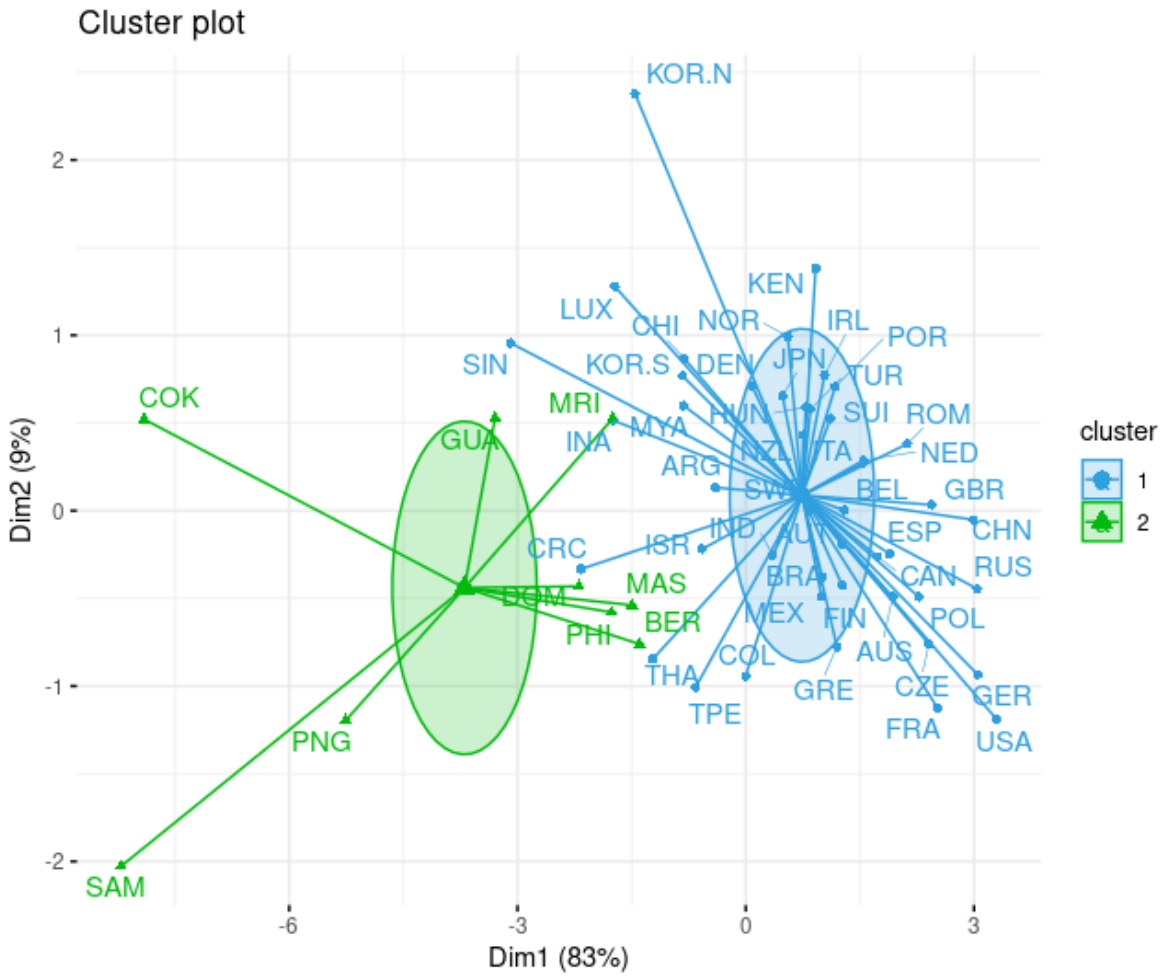

Figura 7: Clusters não-hierárquico para K = 2.

As médias dos clusters são as seguintes:

Table 3: Médias dos clusters para K = 2.

Clusters	m100	m200	m400	m800	m1500	m3000	maratona
1	11,26	22,89	51,34	1,99	4,10	8,84	147,92
2	11,81	24,22	55,22	2,15	4,58	10,24	182,09

Para K = 3

Quando K é igual a 3, ou seja, três centróides, três clusters, os países ficam divididos em grupos de 35, 16 e 3 países.

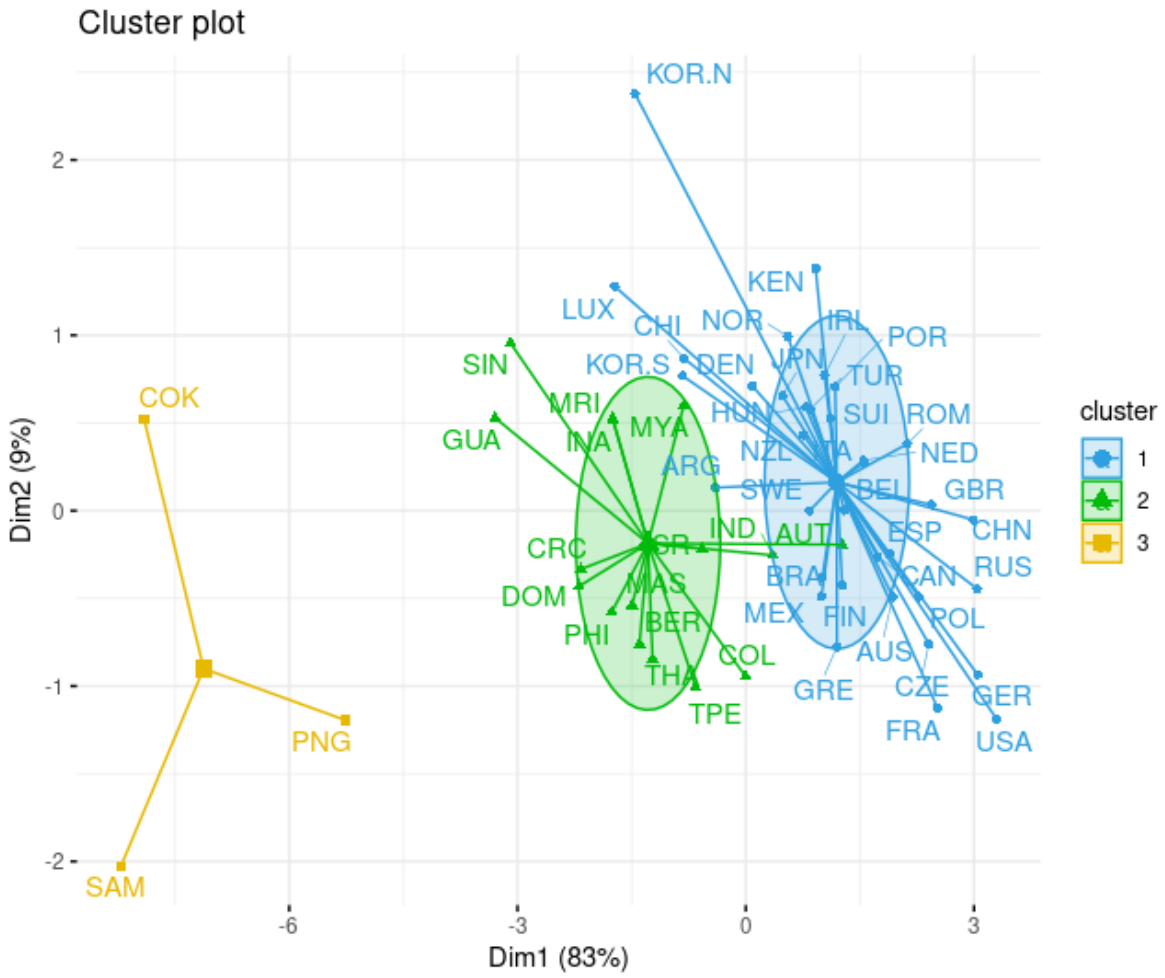

Figura 8: Clusters não-hierárquico para K = 3.

As médias dos clusters são as seguintes:

Table 4: Médias dos clusters para K = 3.

Clusters	m100	m200	m400	m800	m1500	m3000	maratona
1	11,20	22,76	51,02	1,97	4,05	8,67	145,12
2	11,52	23,46	53,01	2,07	4,34	9,52	161,94
3	12,28	25, 34	57,71	2,27	4,95	11,47	208,35

Para K = 4

Quando K é igual a 4, ou seja, quatro centróides, quatro clusters, os países ficam divididos em

grupos de 3, 15, 8 e 28 países.

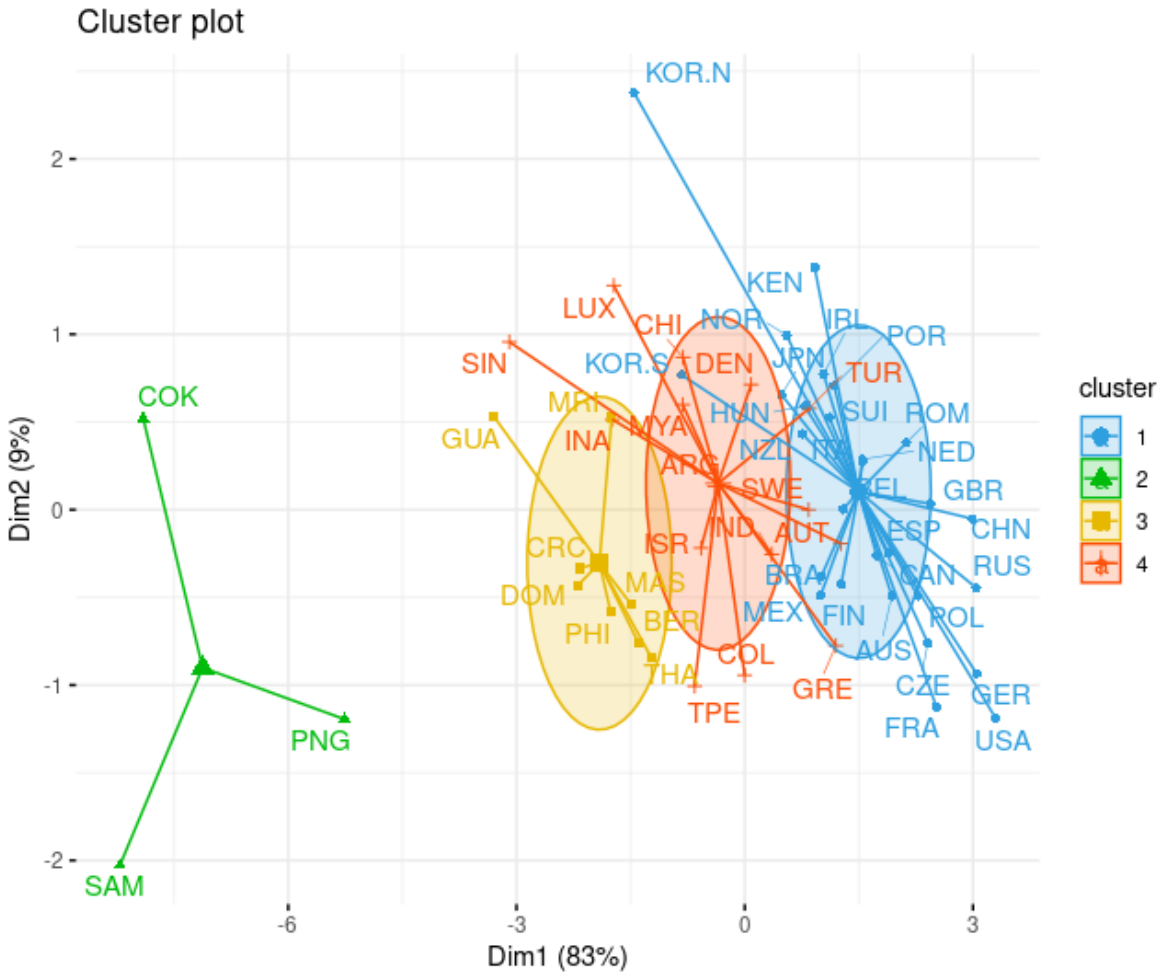

Figura 9: Clusters não-hierárquico para K = 4.

As médias dos clusters são as seguintes:

Table 5: Médias dos clusters para K = 4.

Clusters	m100	m200	m400	m800	m1500	m3000	maratona
1	12,28	25,34	57,71	2,27	4,95	11,47	208,35
2	11,43	23,23	52,65	2,03	4,21	9,16	153,79
3	11,57	23,65	53,63	2,09	4,41	9,71	167,56
4	11,15	22,66	50,55	1,96	4,02	8,60	143,67

Valores das médias para K= 2, 3, 4.

Quando variamos o valor de K, nota-se como as médias tendem a ficar mais próximas ou não e isso, claro, reflete em como os centróides ficam disposto. Quando temos K = 2, temos médias próximas para todas as variáveis (tabela 3), e no gráfico da figura 7, vemos como os clusters são relativamente próximos. Quando temos K = 3, temos dois vetores de médias que são mais próximos e um outro que contém médias superiores aos outros vetores (tabela 4). No gráfico da figura 8, observa-se um cluster mais distante dos outros dois que são mais próximos. Quando temos K = 4, temos três vetores de médias mais próximas (tabela 5) e um outro com médias maiores. No gráfico da figura 9, temos 3 clusters muito próximos e um outro que fica distante dos 3.

MÉTODO HIERÁRQUICO X MÉTODO NÃO HIERÁRQUICO

CONCLUSÃO

Existem vantangens e desvantagens em ambos os métodos, a começar pela visualização dos clusters. No dendograma do método hierárquico, com 54 observações, a visualização dos clusters fica prejudicada, por mais que recursos como cores diferentes para cada cluster tenham sido utilizados. Por outro lado, nos gráficos do método não-hierárquico, o tamanho dos clusters não impediu uma boa visualização dos mesmos. Quanto ao número de clusters adequados, em Kmeans, precisamos reproduzir o algoritmo várias vezes para analisar com qual valor de K teremos o número de clusters mais adequado, enquanto que no modelo hierárquico, pode-se analisar o dendograma e a partir dele tomar a melhor decisão. É interessante observar como esses dados foram sensíveis aos outliers detectados nos gráficos da análise descritiva, um cluster em todas as análises foi sempre formados por eles, o grupo de 3 países (COK, PNG e SAM). Removendo o cluster de outliers, 2 clusters são adequados para estes dados.

Referências

JOHNSON, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2007.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.