IEE062: Estatística Multivariada II
Exercício Escolar 6

Maria Nilza de Sousa Ramos

Entrega: 25/07/2024

Análise de Agrupamento

Dados Utilizados

O conjunto de dados para esta atividade consiste nos recordes masculinos na modalidade de atletismo referente a 54 países em 8 categorias.

Teste Multivariado

A distância de Mahalanobis é uma métrica que mede a distância entre um ponto e a média de um conjunto de pontos, considerando a variabilidade das diferentes dimensões e é definida como:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^\top \mathbf{S}^{-1} (\mathbf{x} - \mu)}$$

onde:

- $\mathbf{x}$ é um vetor de observações.
- μ é o vetor de médias das variáveis.
- $\mathbf{S}$ é a matriz de covariância das variáveis.
- $\mathbf{S}^{-1}$ é a inversa da matriz de covariância.
- $(\mathbf{x} - \mu)^\top$ é o vetor transposto da diferença entre $\mathbf{x}$ e μ.

Ao calcular a distância de Mahalanobis para cada ponto em um conjunto de dados e criar um boxplot dessas distâncias, podemos identificar outliers multivariados da mesma forma que identificaríamos outliers univariados usando um boxplot.

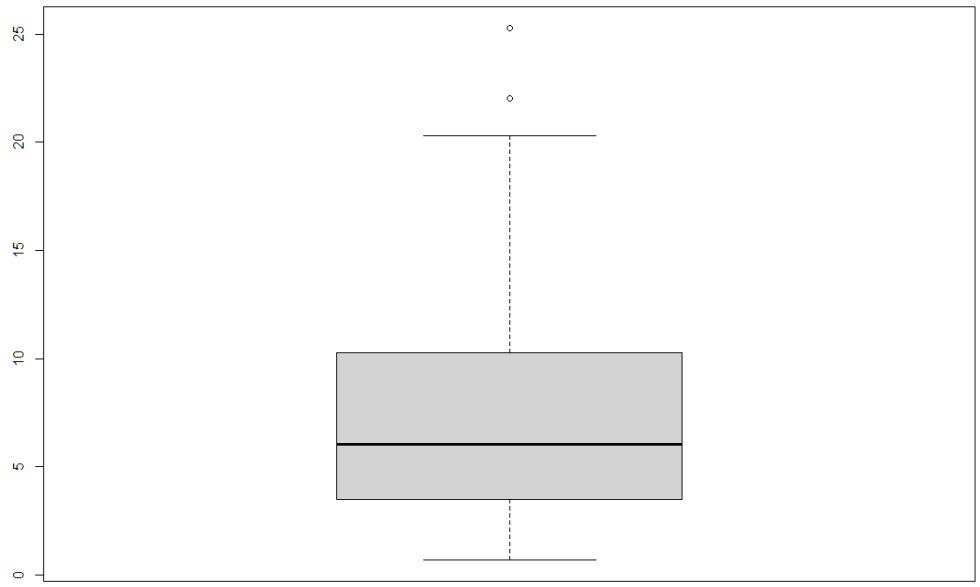


Figura 1: Boxplot das distâncias de Mahalanobis para identificar outliers multivariados

Para este conjunto dados, conforme visto na Figura 1 há 2 possíveis outliers. Todavia, como os dados são registros de recordes de corrida e são poucos dificilmente é um erro ou outlier. Desta forma, todos os dados serão mantidos, pois muito provavelmente trata-se de um destaque superior as nações adversárias nas mesmas categorias.

Distâncias Euclidianas entre os pares dos países.

Após essa primeira análise visual, foi realizada a padronização dos dados e calculas as distâncias euclidianas entre os pares dos países. Para facilitar o entendimento optou-se por colocar em um gráfico de dispersão que pode ser visto na Figura 2 abaixo.

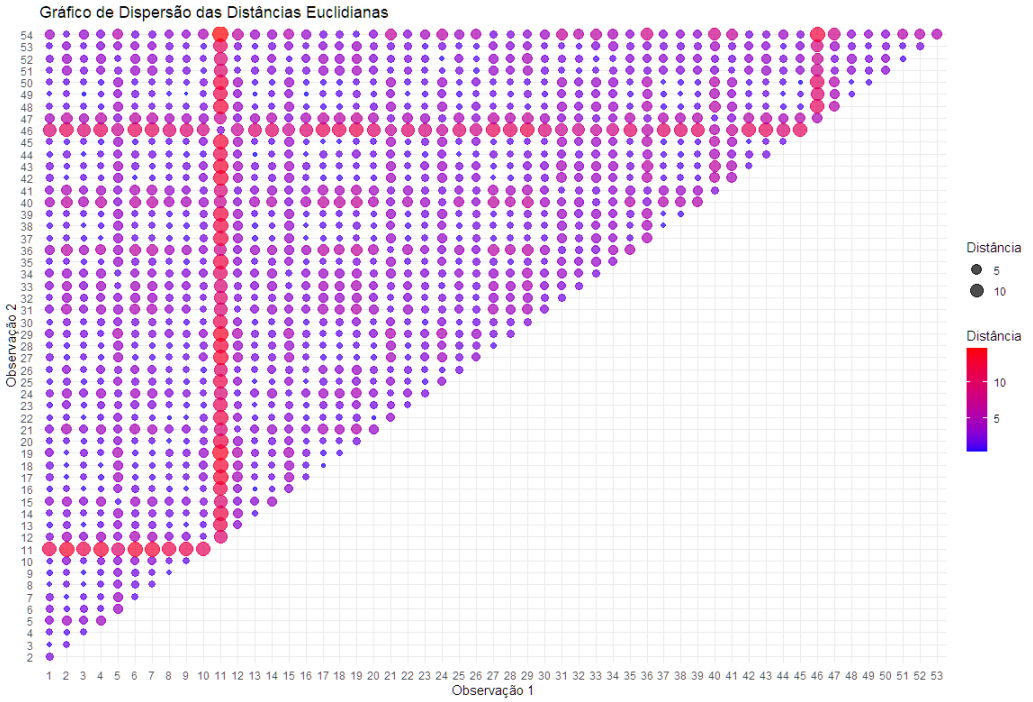
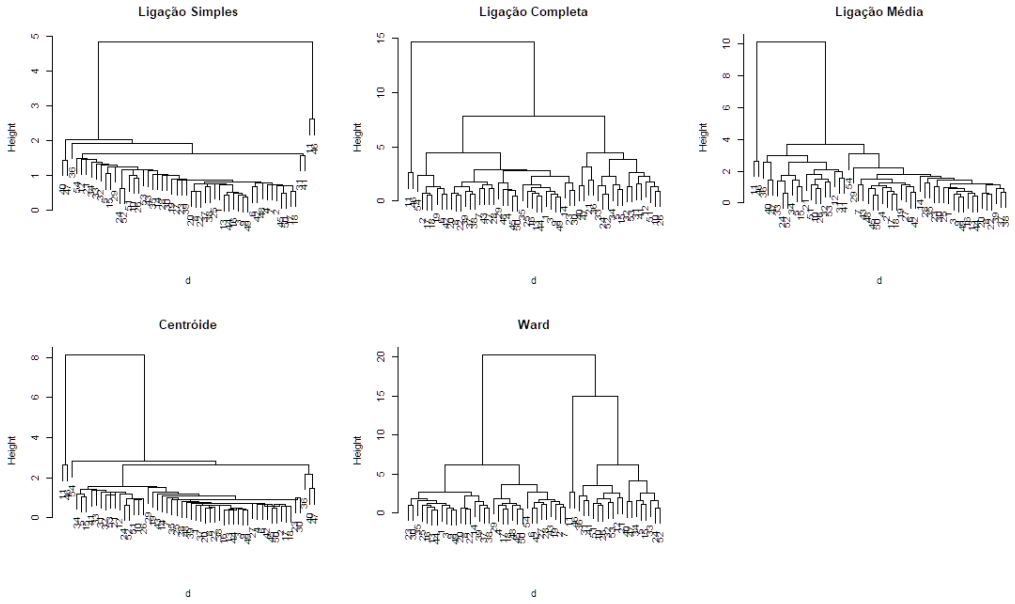


Figura 2: Dispersão das Distâncias Eucliadianas para todos o Conjunto de Dados

As maiores distâncias foram as 11 e 46, por serem duas levantou-se o questionamento de se seriam as mesmas possíveis outliers da Figura 1.

Agrupamento Hierárquico

Os métodos de clusterização utilizados foram os de Ligação Simples, Ligação Completa, da Média, do Centroide e de Ward. Resultando nos seguintes dendogramas:



Para avaliar a qualidade do agrupamento hierárquico usa-se a correlação cophenética. Esta mede a qualidade do dendrograma em representar as distâncias originais entre os pontos de dados. A fórmula utilizada no R é dada por:

$$\text{cor}(d, \text{cophenetic}(hc1))$$

onde:
- d é a matriz de distâncias eclidianas das variáveis padronizadas.
- $hc1$ é a variável que guarda o dendograma do método 1
- cophenetic($hc1$) é a matriz de distâncias cophenéticas derivada do dendrograma.

Resultando nos seguintes valores:

Tabela 1: Valores da Correlação para cada Método de Clusterização

Método	**Correlação**
Simples	0.901
Completa	0.856
Média	0.909
Centroide	0.896
Ward	0.576

O método da média e da ligação simples apresentaram a maior correlação, indicando que esses métodos são mais consistentes em relação às distâncias

entre os pontos dentro dos clusters. Em contrapartida, o método de Ward teve a menor correlação, sugerindo que ele pode formar clusters que são menos consistentes em termos de distância entre pontos.

Tabela 2: Resultados de emaranhamento entre pares de dendrogramas

Combinação de métodos	**Emaranhamento**
Ligação Simples x Ligação Completa	0.79
Ligação Simples x Ligação Média	0.38
Ligação Simples x Centróide	0.36
Ligação Simples x Ward	0.80
Ligação Completa x Ligação Média	0.77
Ligação Completa x Centróide	0.70
Ligação Completa x Ward	0.37
Ligação Média x Centróide	0.34
Ligação Média x Ward	0.96
Centróide x Ward	0.82

A análise de agrupamento hierárquico revelou diferenças notáveis entre os métodos de clusterização. Métodos como Ligação Média e Ward mostraram alta similaridade (Figura 2), enquanto métodos como Ligação Média e Centróide (Figura 3) apresentaram diferenças significativas na estrutura dos clusters.

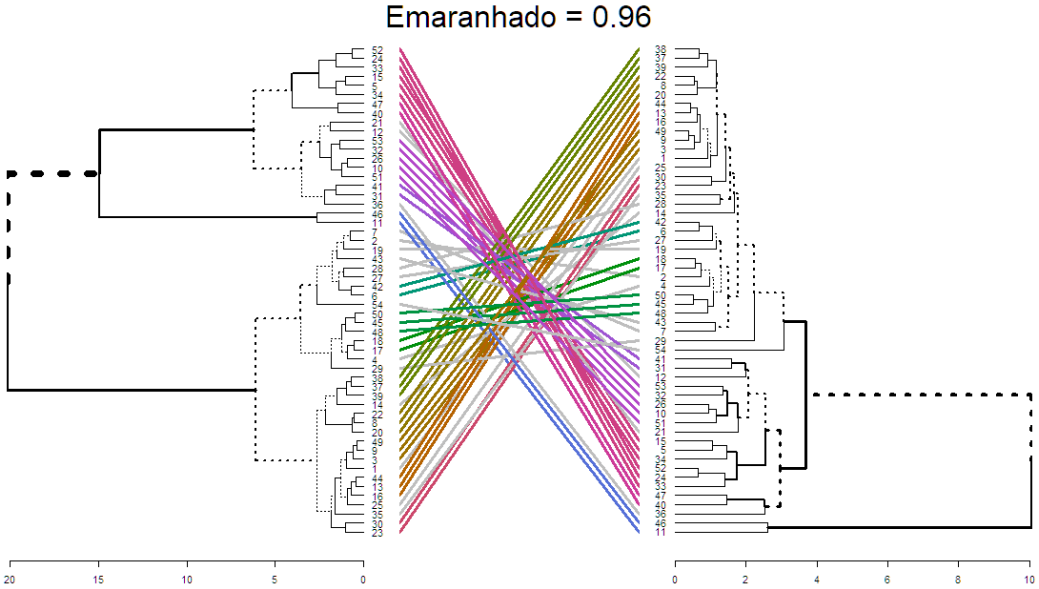


Figura 3: Gráfico compartivo de Dendogramas com Alto Emaranhamento

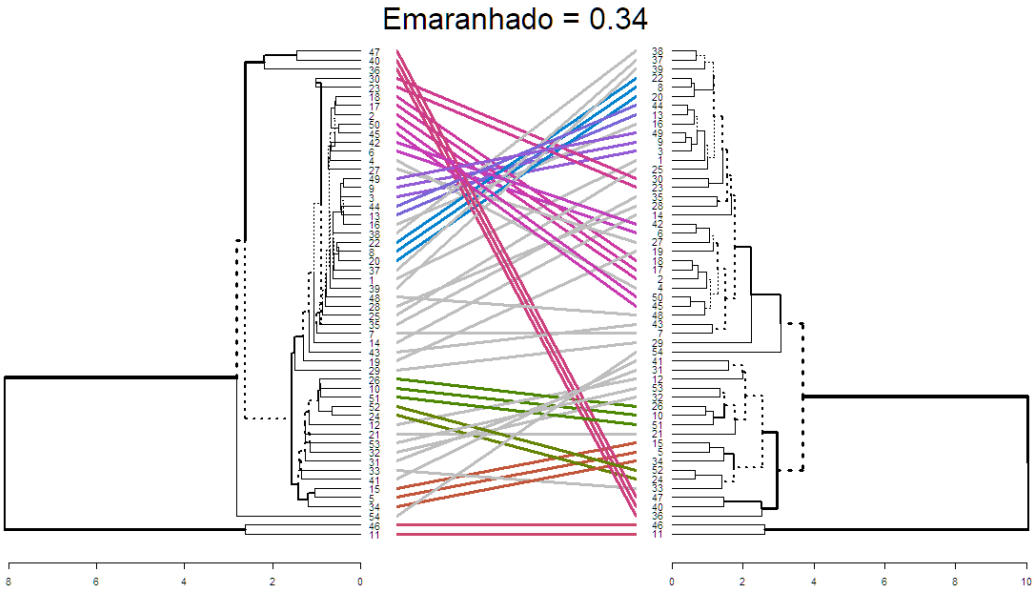


Figura 4: Gráfico compartivo de Dendogramas com Baixo Emaranhamento

Agrupamento Não-Hierárquico (K-means)

O algoritmo K-means é uma técnica de agrupamento que divide um conjunto de dados em k clusters com o objetivo de minimizar a variância dentro dos clusters. O algoritmo foi executado 20 vezes com diferentes centroides iniciais num intervalo de 1 à 10 clusters resultando no screeplot a seguir:

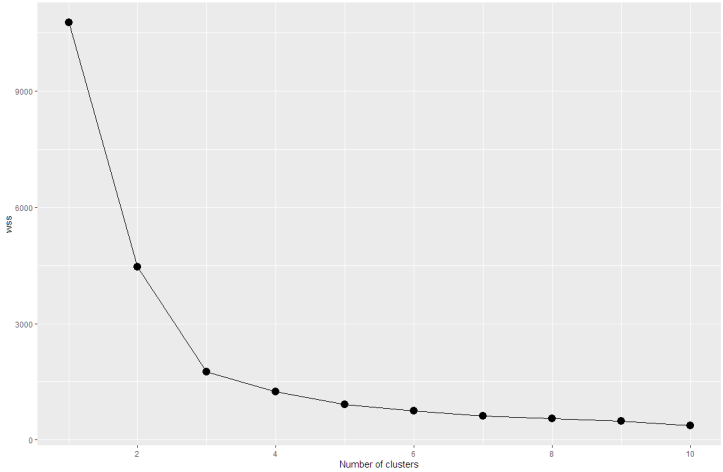


Figura 5: Gráfico compartivo de Dendogramas com Baixo Emaranhamento

Nos agrupamento com mais de 3 clusters há uma angulação bem baixa com relação ao eixo horizontal. Desta forma, adota-se para o método K-means 3 clusters com 20 execuções diferentes de centroides tendo como resultado a seguinte representação:

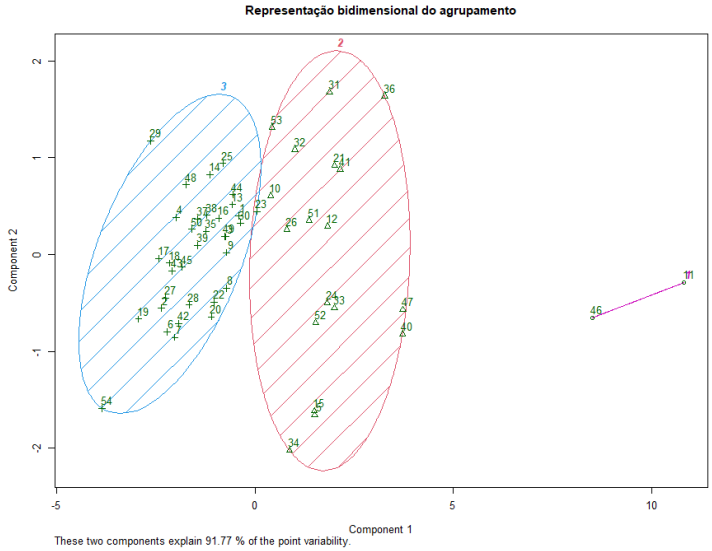


Figura 6: Gráfico compartivo de Dendogramas com Baixo Emaranhamento

No primeiro agrupamento estão os países de Bermudas, Colômbia, Costa Rica, República Dominicana, Guatemala, Indonésia, Israel, Coreia do Norte, Luxemburgo, Malásia, Maurício, Mianmar (Birmânia), Papua-Nova Guiné, Filipinas, Singapura, Taiwan, Tailândia e Turquia. No segundo agrupamento encontram-se os países 11 e 46 do conjunto de dados: Ilhas Cook e Samoa. No terceiro agrupamento estão Argentina, Austrália, Áustria, Bélgica, Brasil, Canadá, Chile, China, República Tcheca, Dinamarca, Finlândia, França, Alemanha, Grã-Bretanha, Grécia, Hungria, Índia, Irlanda, Itália, Japão, Quênia, Coreia do Sul, México, Países Baixos, Nova Zelândia, Noruega, Polônia, Portugal, Romênia, Rússia, Espanha, Suécia, Suíça e EUA.

Além disso, o cálculo das médias de cada variávl para cada cluster pode ser visto na tabela abaixo, de tal forma que é visível sempre os valores do cluster 2 possúirem os maiores valores.

Tabela 3: Médias das variáveis por cluster

	V2	V3	V4	V5	V6	V7	V8	V9
Cluster1	10.34	20.91	46.54	1.81	3.77	14.07	29.47	138.48
Cluster2	10.88	22.16	50.69	1.94	4.13	16.49	35.05	166.38
Cluster3	10.11	20.25	45.17	1.74	3.56	13.20	27.66	128.90

Por curiosidade tentou-se reduzir o agrupamento para 2 clusters e o algoritmo formou dois grupos, o primeiro sem os possíveis outliers e o segundo com as nações Ilhas Cook e Samoa.

Por conseguinte, o número ideal de agrupamentos realmente deve incluir um para os dois valores mais discrepante no mínimo, e por ainda ter 56 dados dispersos para diferenciações mais dois clusters parece o ideal.

Algoritmo em R

```
# Carregando as bibliotecas necessárias
library(cluster)
library(dendextend)
library(reshape2)
library(ggplot2)

# Lendo os dados
promo <- read.table("T8-6.DAT")
promo_0 <- promo
promo <- promo[-1]

# Calculando a matriz de covariância e a média
p.cov <- var(promo)
p.mean <- apply(promo, 2, mean)

# Calculando a distância de Mahalanobis
p.mah <- mahalanobis(promo, p.mean, p.cov)

# Visualização do boxplot das distâncias de Mahalanobis
boxplot(p.mah)

# Padronizando as variáveis
Dados.pad <- scale(promo)
```

```
# Calculando a matriz de distâncias Euclidianas
d <- dist(Dados.pad, method = "euclidean")

# Convertendo a matriz de distâncias para um data frame
dist_matrix <- as.matrix(d)
dist_df <- as.data.frame(as.table(dist_matrix))

# Filtrando a parte triangular superior da matriz de distâncias
dist_df <- dist_df[upper.tri(dist_matrix), ]

# Criando um gráfico de dispersão das distâncias Euclidianas
scatter_plot <- ggplot(dist_df, aes(Var1, Var2, size = Freq, color = Freq)) +
  geom_point(alpha = 0.7) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(title = "Gráfico de Dispersão das Distâncias Euclidianas",
       x = "Observação 1",
       y = "Observação 2",
       size = "Distância",
       color = "Distância") +
  theme_minimal()
print(scatter_plot)

# Definindo os clusters a partir do método escolhido
# Métodos disponíveis: "average", "single", "complete", "centroid", "ward.D"
hc1 <- hclust(d, method = "single")
hc2 <- hclust(d, method = "complete")
hc3 <- hclust(d, method = "average")
hc4 <- hclust(d, method = "centroid")
hc5 <- hclust(d, method = "ward.D")

# Calculando a correlação cophenética para cada método de clusterização
colunas <- c("Simples", "Completa", "Media", "Centroide", "Ward")
correlacoes <- c(cor(d, cophenetic(hc1)), cor(d, cophenetic(hc2)),
                 cor(d, cophenetic(hc3)), cor(d, cophenetic(hc4)),
                 cor(d, cophenetic(hc5)))
x <- matrix(c(colunas, round(correlacoes, 3)), nrow = 2, byrow = T)
x

# Desenhando os dendrogramas
plot(hc1, cex = 0.6, hang = -1)
plot(hc2, cex = 0.6, hang = -1)
```

```
plot(hc3, cex = 0.6, hang = -1)
plot(hc4, cex = 0.6, hang = -1)
plot(hc5, cex = 0.6, hang = -1)

# K-means clustering
library(cluster)
set.seed(42)
library(dplyr)

# Determinando o número ótimo de clusters pelo método Elbow
wss <- numeric(10)
for (i in 1:10) {
  km.out <- kmeans(d, centers = i, nstart = 20)
  wss[i] <- km.out$tot.withinss
}

wss_df <- tibble(clusters = 1:10, wss = wss)

scree_plot <- ggplot(wss_df, aes(x = clusters, y = wss, group = 1)) +
  geom_point(size = 4) +
  geom_line() +
  scale_x_continuous(breaks = c(2, 4, 6, 8, 10)) +
  xlab('Number of clusters')
scree_plot

# Aplicando K-means com 3 clusters
km.out <- kmeans(d, centers = 3, nstart = 20)
clusplot(Dados.pad, km.out$cluster, main = 'Representação 2D do agrupamento',
         color = TRUE, shade = TRUE, labels = 2, lines = 0)

# Aplicando K-means com 2 clusters
km.out2 <- kmeans(d, centers = 2, nstart = 20)
clusplot(Dados.pad, km.out2$cluster, main = 'Representação 2D do agrupamento',
         color = TRUE, shade = TRUE, labels = 2, lines = 0)

# Identificando os países em cada cluster
promo_O[km.out$cluster == 1, ]$V1
promo_O[km.out$cluster == 2, ]$V1
promo_O[km.out$cluster == 3, ]$V1

# Calculando as médias das variáveis para cada cluster
```

```
var <- c("V2", "V3", "V4", "V5", "V6", "V7", "V8", "V9")
cluster1_medias <- colMeans(promo[km.out$cluster == 1, ])
cluster2_medias <- colMeans(promo[km.out$cluster == 2, ])
cluster3_medias <- colMeans(promo[km.out$cluster == 3, ])

medias_df <- data.frame(
  Cluster1 = cluster1_medias,
  Cluster2 = cluster2_medias,
  Cluster3 = cluster3_medias
)
t(medias_df)
```