Segmentação de clientes de uma rede de academias

Table of contents

1 Contextualização

Uma rede de academias de médio porte, presente em várias cidades brasileiras, deseja entender melhor o perfil de seus clientes ativos para:

- Criar planos personalizados
- Melhorar a retenção
- Identificar grupos com risco de evasão
- Direcionar campanhas de marketing

Atualmente, a empresa possui apenas dados operacionais e nunca realizou uma segmentação formal baseada em dados. A equipe de dados foi acionada para realizar uma análise de agrupamentos (clusterização).

---

2 Base de dados

Para cada cliente, foram coletadas as seguintes variáveis quantitativas:

Variável	Descrição
`idade`	Idade do cliente (anos)
`frequencia`	Média de visitas à academia por semana
`tempo_treino`	Tempo médio de treino (minutos)
`gasto_mensal`	Gasto médio mensal (R$)
`tempo_plano`	Tempo de permanência no plano (meses)

Objetivo: identificar grupos homogêneos de clientes com comportamentos semelhantes.

---

3 Preparação e inspeção dos dados

```
load <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}

## Pacotes utilizados nessa análise

packages = c("tidyverse", "dplyr", "ggplot2", "cluster", "seriation", "factoextra", "tibble", "dendexter
load(packages)
```

```
 tidyverse      dplyr    ggplot2    cluster  seriation factoextra     tibble
      TRUE       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE
dendextend    hopkins
      TRUE       TRUE
```

3.1 Leitura e exploração dos dados

os dados para esta análise estão disponíveis aqui.

```
dados %>%
  glimpse()
```

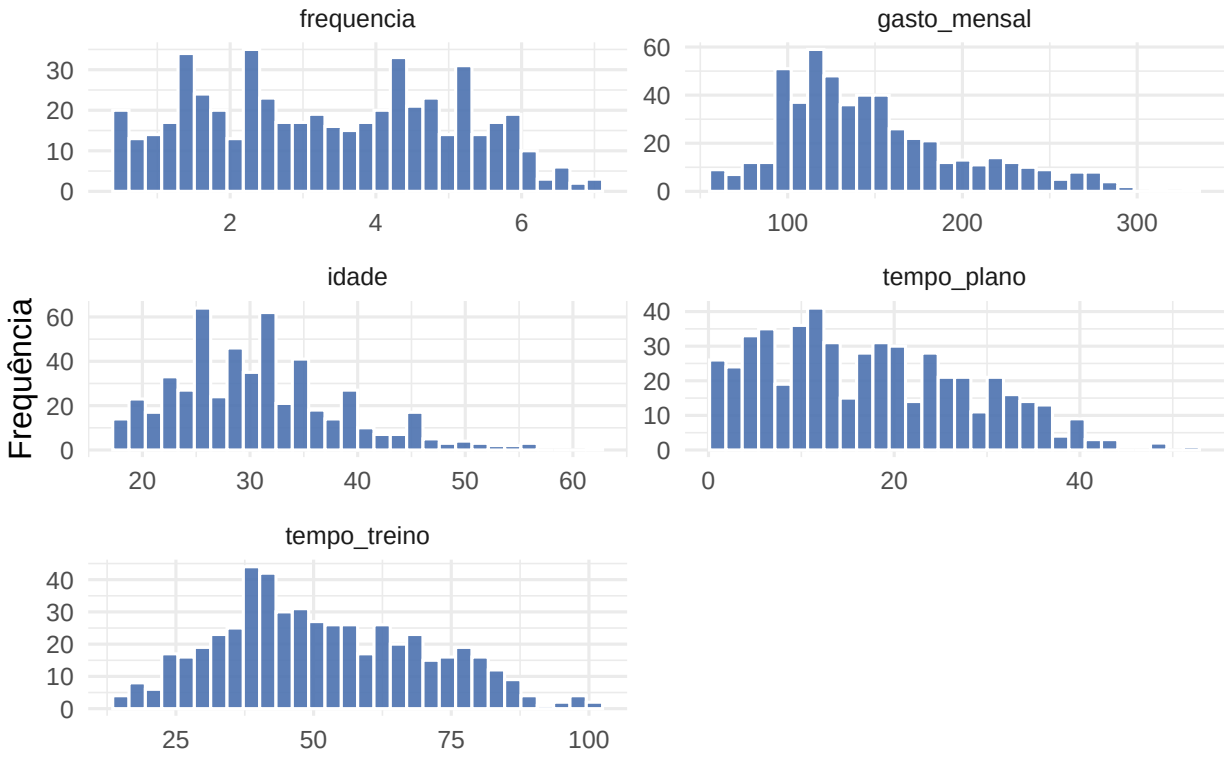
```
Rows: 530
Columns: 5
$ idade        <dbl> 27, 27, 35, 23, 28, 19, 26, 24, 30, 28, 32, 39, 32, 24, 3~
$ frequencia   <dbl> 4.8, 5.7, 4.8, 5.1, 5.8, 5.4, 4.8, 4.9, 6.7, 4.7, 5.7, 5.~
$ tempo_treino <dbl> 80, 78, 82, 49, 63, 79, 66, 91, 90, 82, 73, 85, 84, 61, 6~
$ gasto_mensal <dbl> 233, 242, 233, 268, 253, 242, 241, 204, 172, 254, 249, 23~
$ tempo_plano  <dbl> 32, 43, 31, 34, 15, 34, 37, 37, 48, 15, 19, 35, 20, 14, 2~
```

```r
dados %>%
  summary()
```

```
     idade         frequencia     tempo_treino     gasto_mensal
 Min.   :18.00   Min.   :0.500   Min.   : 15.00   Min.   : 60.0
 1st Qu.:25.00   1st Qu.:1.900   1st Qu.: 38.00   1st Qu.:112.0
 Median :30.00   Median :3.300   Median : 49.50   Median :139.0
 Mean   :30.88   Mean   :3.341   Mean   : 51.78   Mean   :148.9
 3rd Qu.:35.00   3rd Qu.:4.700   3rd Qu.: 65.00   3rd Qu.:175.8
 Max.   :62.00   Max.   :7.000   Max.   :101.00   Max.   :331.0
  tempo_plano
 Min.   : 1.00
 1st Qu.: 8.00
 Median :16.00
 Mean   :17.32
 3rd Qu.:25.00
 Max.   :52.00
```

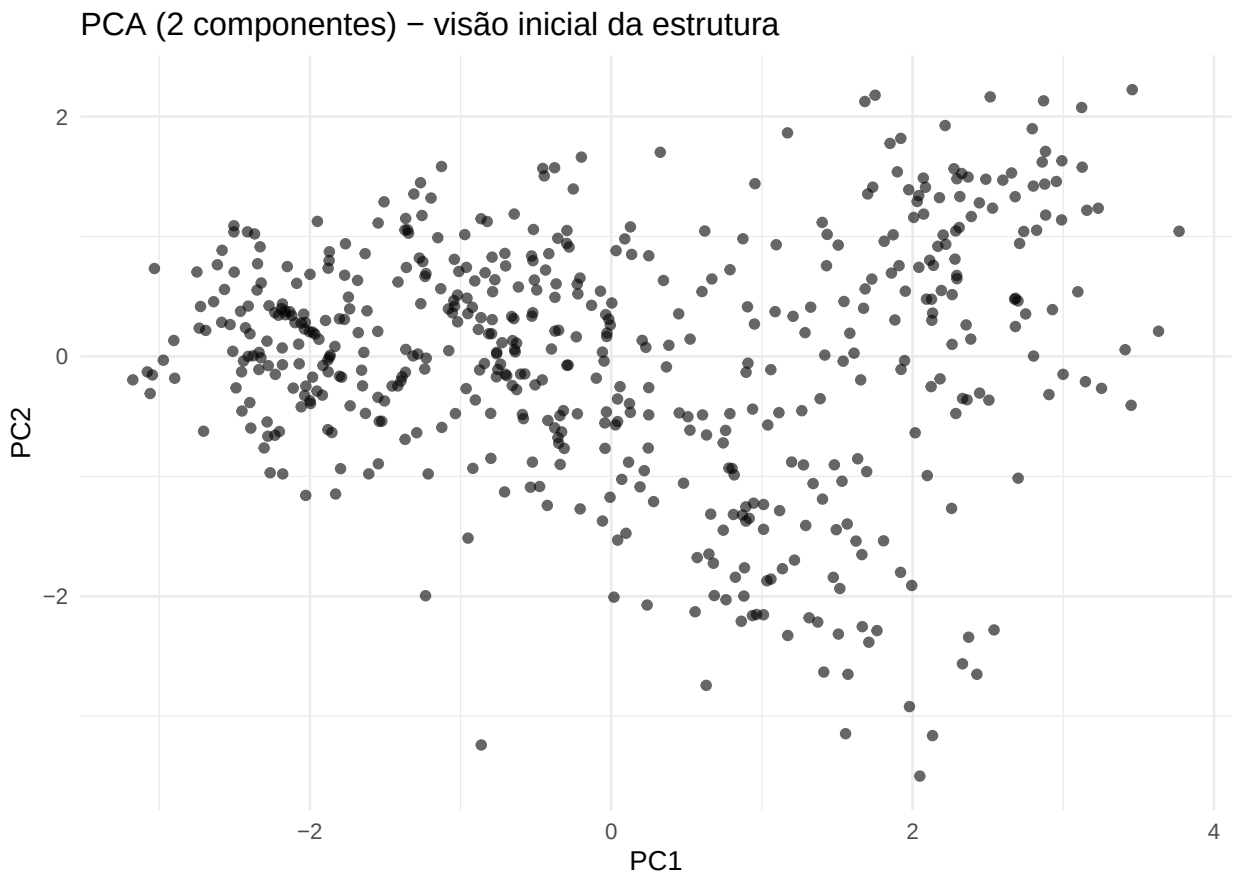
A análise descritiva inicial revela elevada heterogeneidade entre os clientes da academia, tanto em termos de comportamento (frequência e tempo de treino) quanto de vínculo econômico e temporal (gasto mensal e tempo de plano). As distribuições apresentam amplitudes consideráveis e assimetrias moderadas, especialmente para gasto mensal e tempo de permanência, sugerindo a coexistência de múltiplos perfis de clientes. Esse padrão indica forte plausibilidade da existência de estrutura de agrupamento intrínseca, justificando o uso de métodos de clusterização baseados em distância após adequada padronização das variáveis.

```r
# Reorganiza os dados para facilitar a plotagem
dados_long <- dados %>%
  pivot_longer(
    cols = everything(),
    names_to = "variavel",
    values_to = "valor"
  )

# Histogramas individuais (facilita leitura em aula)
ggplot(dados_long, aes(x = valor)) +
  geom_histogram(bins = 30, fill = "#4C72B0", color = "white", alpha = 0.9) +
  facet_wrap(~ variavel, scales = "free", ncol = 2) +
  labs(
    title = "Histogramas das variáveis observadas",
    subtitle = "Distribuições marginais evidenciam heterogeneidade entre clientes",
    x = NULL,
    y = "Frequência"
  ) +
  theme_minimal(base_size = 13)
```



Os histogramas das variáveis reforçam a conclusão obtida pela análise descritiva numérica. Observa-se elevada heterogeneidade nas distribuições marginais, com evidências de múltiplos regimes comportamentais, especialmente nas variáveis de frequência semanal, gasto mensal e tempo de permanência no plano. A presença de assimetrias, caudas longas e concentrações distintas sugere que os clientes não formam um grupo homogêneo, mas sim um conjunto de perfis latentes distintos, o que sustenta a aplicação de métodos de agrupamento multivariado.

```
pca <- prcomp(dados, scale. = TRUE)
pca_df <- as.data.frame(pca$x[, 1:2])
names(pca_df) <- c("PC1","PC2")

ggplot(pca_df, aes(PC1, PC2)) +
  geom_point(alpha = 0.6) +
  labs(title = "PCA (2 componentes) - visão inicial da estrutura") +
  theme_minimal()
```



A Figura apresenta a projeção dos dados nas duas primeiras componentes principais, obtidas a partir das variáveis padronizadas. Observa-se que os pontos não se distribuem de forma homogênea no espaço reduzido, mas sim formam regiões de maior densidade separadas por áreas de menor concentração. Esse padrão sugere a presença de estrutura multivariada latente compatível com a existência de grupos distintos de clientes.

Como os métodos baseados em distância são sensíveis à escala, padronizamos (z-score).

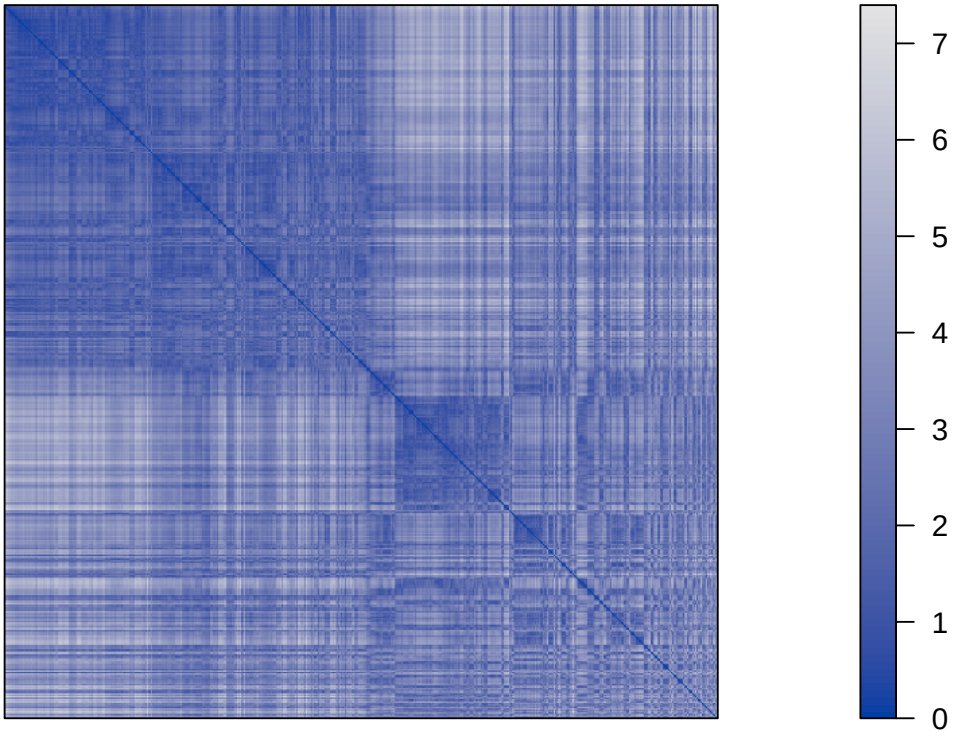
```
X <- scale(dados)

# Distância Euclidiana
D <- dist(X, method = "euclidean")
```

3.2 Avaliação da tendência de agrupamento

```
# Tendência a agrupamento: Hopkins
# - H ~ 0.5 sugere aleatoriedade (sem clusters)
# - H alto (ex: > 0.70) sugere forte estrutura de agrupamento
# Observação: Hopkins é sensível ao tamanho amostral e ao "m" (nº pontos amostrados).
# Aqui usamos m = ~10% do N (regra prática razoável).
set.seed(12345)
m <- floor(0.1 * nrow(X))
hop <- hopkins::hopkins(X, m = m)
hop
```

```
[1] 0.9970434
```

```
VAT(D)
```



A estatística de Hopkins apresentou valor extremamente elevado ($H = 0,997$), indicando que os dados estão muito distantes de um padrão aleatório no espaço multivariado e exibem forte tendência a agrupamento. Esse resultado é corroborado pelo método VAT, cuja matriz de distâncias reorganizada revela blocos bem definidos ao longo da diagonal principal, separados por regiões de menor similaridade. A concordância entre a evidência estatística formal e a avaliação visual fornece suporte robusto para a existência de estrutura de agrupamento intrínseca nos dados, legitimando a aplicação subsequente de métodos de clusterização hierárquica.

4 Métos de agrupamento

```
# Vamos testar: single, complete, average, ward.D2
# Critério de comparação:
#   (a) correlação cofenética: quão bem o dendrograma preserva distâncias
#       (maior tende a ser melhor, mas não é o único critério)
#   (b) coerência com interpretação e separação por silhueta (mais adiante)
metodos <- c("single", "complete", "centroid", "average", "ward.D2")

hcl_list <- lapply(metodos, function(met) hclust(D, method = met))
```

```r
names(hcl_list) <- metodos


# Correlação cofenética
coph_corr <- sapply(hcl_list, function(hc) {
  cor(D, cophenetic(hc))
})
coph_corr
```

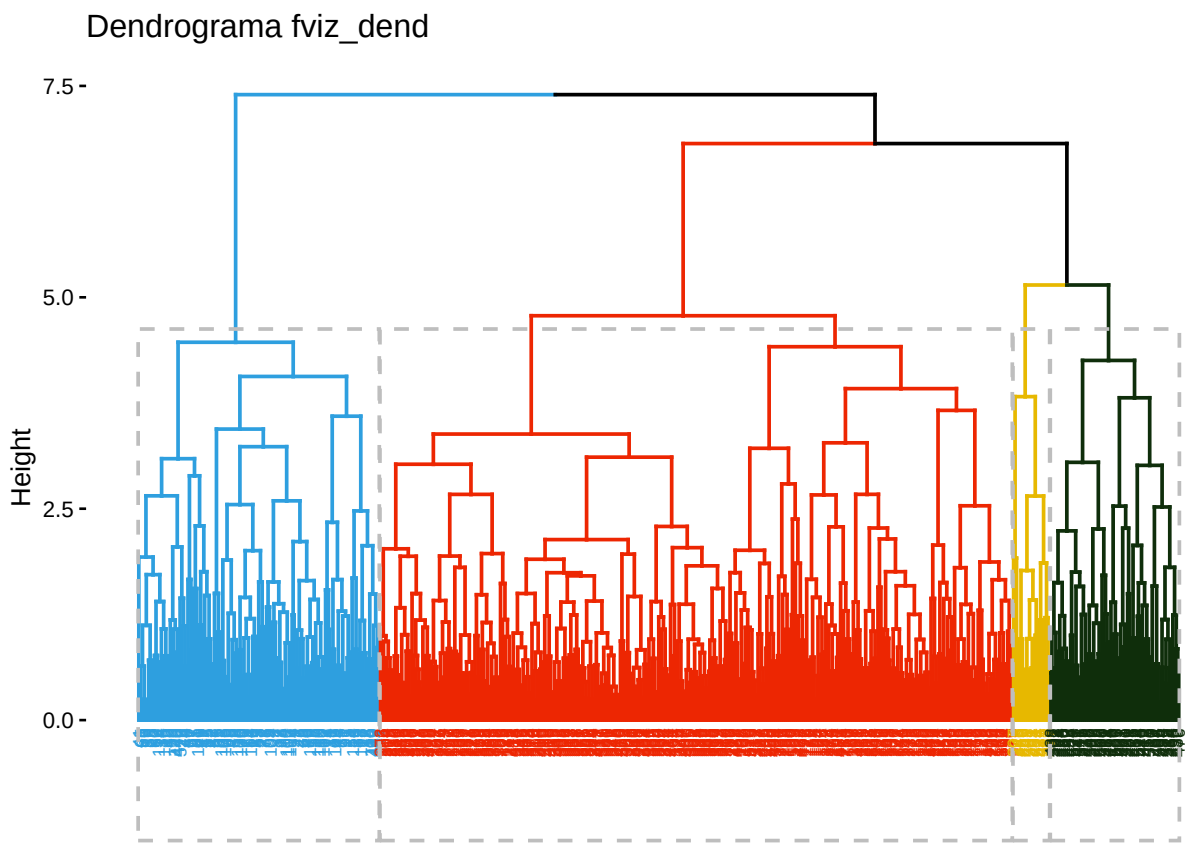
```
   single  complete  centroid   average   ward.D2
0.4504459 0.7146673 0.6463892 0.7016730 0.6990017
```

```r
# Regra prática: maior correlação cofenética => dendrograma mais fiel às distâncias
melhor_met <- names(which.max(coph_corr))
melhor_met
```

```
[1] "complete"
```

Foram avaliados os métodos hierárquicos aglomerativos single, centroid, average, complete e Ward.D2, utilizando a correlação cofenética como critério de comparação. Os resultados indicaram que o método complete apresentou a maior correlação cofenética (0,715), sugerindo melhor preservação das distâncias originais entre as observações. O método single apresentou desempenho insatisfatório, evidenciando o efeito de encadeamento (chaining), enquanto average e Ward.D2 mostraram desempenho intermediário. Assim, com base no critério adotado, o método complete foi selecionado para a construção da solução hierárquica final.

```r
hc_best <- hcl_list[[melhor_met]]

# Visualização de alta qualidade
fviz_dend(hc_best, k = 4, # Número de grupos
          cex = 0.5, # Tamanho da label
          k_colors = c("#2E9FDF", "#ed2602", "#E7B800", "#0f2e0b"),
          color_labels_by_k = TRUE, # Colorir labels
          rect = TRUE, # Adicionar retângulo
          main = "Dendrograma fviz_dend")
```




```
cluster_hc <- cutree(hc_best, k = 4)
table(cluster_hc)
```

```
cluster_hc
  1   2   3   4
123  66 322  19
```

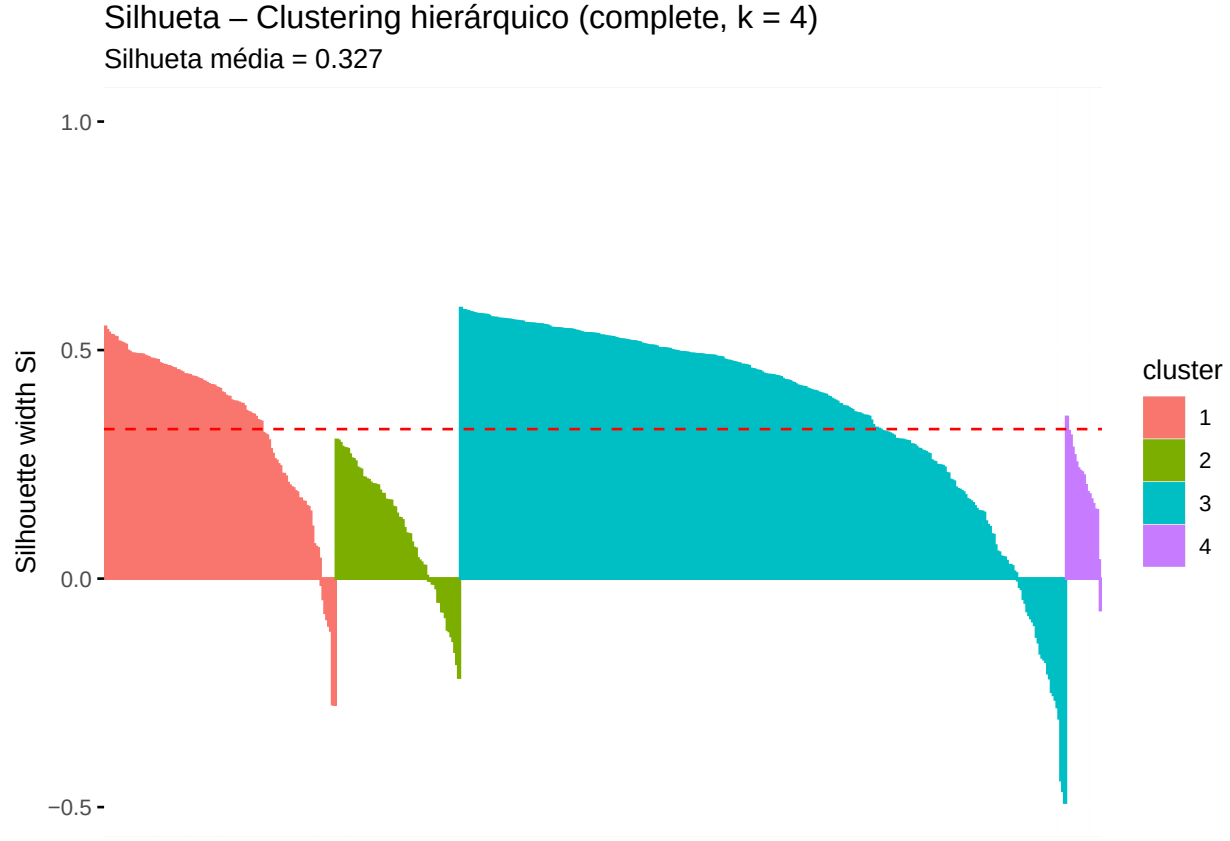
O dendrograma obtido pelo método hierárquico com ligação completa evidencia a presença de quatro ramos principais bem definidos. O corte realizado em um nível de altura relativamente elevado resulta em quatro grupos com boa separação intergrupos e elevada coesão interna. Observa-se um grupo majoritário, representando o perfil predominante dos clientes, dois grupos intermediários com comportamentos distintos e um grupo minoritário altamente isolado, sugerindo um perfil extremo ou especializado. A estrutura observada é consistente com os resultados previamente obtidos pela estatística de Hopkins e pelo método VAT, reforçando a adequação da escolha de $k = 4$.

```
# Silhueta para a solução hierárquica
sil_hc <- silhouette(cluster_hc, D)

# Silhueta média global
mean_sil <- mean(sil_hc[, "sil_width"])
mean_sil
```

```
[1] 0.3270658
```

```r
fviz_silhouette(sil_hc) +
  labs(
    title = "Silhueta - Clustering hierárquico (complete, k = 4)",
    subtitle = paste("Silhueta média =", round(mean_sil, 3))
  )
```

```
  cluster size ave.sil.width
1       1  123          0.35
2       2   66          0.11
3       3  322          0.37
4       4   19          0.21
```




```r
sil_df <- as.data.frame(sil_hc)
sil_df$cluster <- factor(cluster_hc)

sil_df %>%
  group_by(cluster) %>%
  summarise(
    n = n(),
    sil_media = mean(sil_width),
    sil_min = min(sil_width),
    sil_max = max(sil_width)
  )
```

cluster	n	sil_media	sil_min	sil_max
1	123	0.3474965	-0.2757926	0.5509433
2	66	0.1086300	-0.2166009	0.3035377
3	322	0.3711726	-0.4900442	0.5919047
4	19	0.2060862	-0.0689395	0.3537319

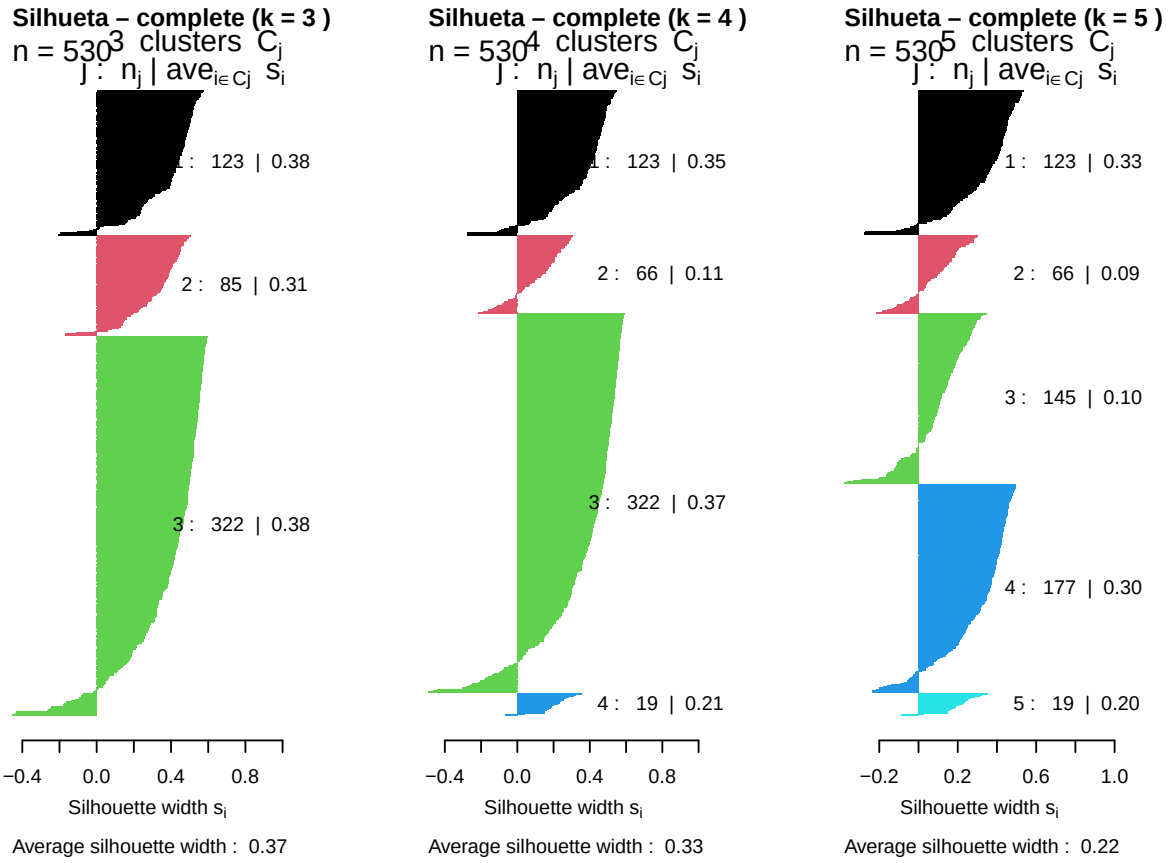
A qualidade da solução hierárquica com quatro grupos foi avaliada por meio do método da silhueta, resultando em valor médio de 0,327, o que indica estrutura de agrupamento moderada. A análise por grupo revelou que dois clusters apresentam coesão razoável, enquanto um grupo intermediário mostrou baixa silhueta média, sugerindo sobreposição com outros perfis. Esses resultados indicam que, embora exista forte estrutura de agrupamento intrínseca nos dados, evidenciada previamente pela estatística de Hopkins e pelo método VAT, a partição em quatro grupos não maximiza completamente a separação entre todos os perfis, motivando a avaliação de soluções alternativas com diferentes valores de k.

4.1 Avaliando outros valores de k

```
ks <- c(3, 4, 5)

sil_results <- lapply(ks, function(k) {
  cl <- cutree(hc_best, k = k)
  sil <- silhouette(cl, D)

  tibble(
    k = k,
    sil_media = mean(sil[, "sil_width"]),
    sil_min = min(sil[, "sil_width"]),
    sil_max = max(sil[, "sil_width"])
  )
})

sil_results <- bind_rows(sil_results)
sil_results
```

k	sil_media	sil_min	sil_max
3	0.3721753	-0.4527945	0.5977440
4	0.3270658	-0.4900442	0.5919047
5	0.2200544	-0.3775301	0.5398297

```
par(mfrow = c(1,3))
for (k in ks) {
  cl <- cutree(hc_best, k = k)
  sil <- silhouette(cl, D)
  plot(
    sil,
    main = paste("Silhueta - complete (k =", k, ")"),
    col = 1:k,
    border = NA
  )
}
```




```
par(mfrow = c(1,1))
```

Comparação objetiva dos valores de silhueta

k	Silhueta média	Silhueta mínima	Silhueta máxima
3	**0,372**	−0,453	0,598
4	0,327	−0,490	0,592
5	0,221	−0,378	0,540

A escolha do número de grupos foi baseada na comparação da silhueta média para diferentes valores de k. Observou-se que a solução com três clusters apresentou o maior valor de silhueta média (0,372), além de uma distribuição mais equilibrada e coerente dos valores individuais, indicando melhor separação e coesão dos grupos. As soluções com quatro e cinco clusters resultaram em redução da silhueta média, evidenciando a presença de clusters fracos e fragmentação artificial. Assim, considerando critérios estatísticos e evidência visual, adotou-se a solução com três clusters como a partição final dos dados.

```
# Rótulos hierárquicos com k = 3
cluster_hc_k3 <- cutree(hc_best, k = 3)

dados_hc3 <- dados %>%
  mutate(cluster = factor(cluster_hc_k3))

descr_hc3 <- dados_hc3 %>%
```

```
  group_by(cluster) %>%
  summarise(
    n = n(),
    idade_media = mean(idade),
    freq_media = mean(frequencia),
    treino_media = mean(tempo_treino),
    gasto_media = mean(gasto_mensal),
    plano_media = mean(tempo_plano),
    .groups = "drop"
  )

descr_hc3
```

cluster	n	idade_media	freq_media	treino_media	gasto_media	plano_media
1	123	30.56098	5.193496	74.91057	221.6423	27.51220
2	85	42.81176	4.614118	58.72941	120.7176	26.21176
3	322	27.85404	2.297515	41.10248	128.4783	11.07143

4.1.0.1 Cluster 1: "Clientes intensivos / premium" (123 clientes, ~23%)

- **Padrão observado**
  - Frequência muito alta (5,2x/semana)
  - Sessões longas (75 min)
  - Maior gasto mensal (R$ 222)
  - Tempo de plano elevado (28 meses)
- **Leitura substantiva**
  - Clientes altamente engajados, com forte uso do serviço e alto valor econômico.
- **Coerência metodológica**
  - Grupo bem separado no dendrograma
  - Silhueta média alta
  - Perfil extremo esperado dado Hopkins 0,997
- Cluster forte, estável e claramente interpretável

4.1.0.2 Cluster 2: "Clientes fiéis e econômicos" (85 clientes, ~16%)

- **Padrão observado**
  - Idade mais elevada (43 anos)
  - Frequência alta, mas menor que o cluster 1
  - Treinos moderados
  - Gasto baixo
  - Tempo de plano longo
- **Leitura substantiva**
  - Clientes consistentes e leais, sensíveis a preço, com padrão de uso regular, porém sem consumo premium.
- **Ponto importante**
  - Esse cluster não apareceu bem separado em k = 4, mas emerge de forma clara em k = 3, o que explica o aumento da silhueta média.
- Cluster legítimo, que o $k = 4$ fragmentava artificialmente

4.1.0.3 Cluster 3: "Clientes ocasionais / risco" (322 clientes, ~61%)

- **Padrão observado**
  - Frequência baixa
  - Treinos curtos
  - Gasto mensal reduzido
  - Tempo de plano curto
  - Idade mais jovem
- **Leitura substantiva**
  - Representa o comportamento mais comum da base: clientes pouco engajados e com maior risco de evasão.
- **Leitura estatística**
  - Maior heterogeneidade interna
  - Silhueta moderada (esperado para clusters grandes)
- Cluster realista e indispensável do ponto de vista gerencial

A solução hierárquica com três grupos revelou perfis claramente distintos de clientes. O primeiro grupo é composto por clientes intensivos, com elevada frequência, longas sessões de treino, alto gasto mensal e forte fidelização. O segundo grupo reúne clientes fiéis e consistentes, caracterizados por tempo de permanência elevado, frequência regular e menor gasto, indicando sensibilidade a preço. O terceiro grupo, majoritário, representa clientes ocasionais, com baixo engajamento e maior risco de evasão. A escolha de três clusters é sustentada pela maximização da silhueta média e pela concordância com a solução obtida via K-médias, reforçando a robustez da segmentação.

4.1.1 K-means

```
# K-médias depende de inicialização. Use nstart alto para estabilidade.
set.seed(12345)
km <- kmeans(X, centers = 3, nstart = 50)

cluster_km <- km$cluster

table(cluster_km)
```

```
cluster_km
  1   2   3
284 125 121
```

```
cat("\nInércia total (tot.withinss):", round(km$tot.withinss, 2), "\n")
```

```
Inércia total (tot.withinss): 1093.67
```

O algoritmo K-médias foi aplicado com três clusters, utilizando múltiplas inicializações para garantir estabilidade da solução. Os tamanhos dos grupos obtidos foram relativamente equilibrados, refletindo a tendência do método em minimizar a variância intra-cluster. A escolha de $k = 3$ é consistente com a solução hierárquica previamente selecionada com base no critério da silhueta, indicando concordância entre métodos fundamentados em princípios distintos. As diferenças observadas entre as soluções concentram-se principalmente em observações de fronteira, o que é esperado dado o caráter particional do K-médias.

```r
# Dados com rótulos do K-médias
dados_km <- dados %>%
  mutate(cluster = factor(cluster_km))

# Estatísticas descritivas por cluster
descr_km <- dados_km %>%
  group_by(cluster) %>%
  summarise(
    n = n(),
    idade_media = mean(idade),
    freq_media = mean(frequencia),
    treino_media = mean(tempo_treino),
    gasto_media = mean(gasto_mensal),
    plano_media = mean(tempo_plano),
    .groups = "drop"
  )

descr_km
```

cluster	n	idade_media	freq_media	treino_media	gasto_media	plano_media
1	284	27.23239	2.072887	39.69014	127.0669	9.552817
2	125	29.80800	5.128000	74.10400	221.6960	27.672000
3	121	40.55372	4.471901	57.07438	124.7438	24.834711

4.1.1.1 Cluster 1: "Clientes ocasionais / risco" (284 clientes, ~54%)

- **Padrão observado**
  - Frequência baixa
  - Sessões curtas
  - Menor tempo de plano
  - Gasto mensal baixo–moderado
  - Idade mais jovem
- **Leitura substantiva**
  - Representa clientes pouco engajados, com uso esporádico da academia e maior risco de evasão.
- **Coerência estatística**
  - Cluster grande → silhueta moderada (esperado)
  - Corresponde ao cluster 3 da solução hierárquica
- Perfil realista e central do negócio

4.1.1.2 Cluster 2: "Clientes intensivos / premium" (125 clientes, ~24%)

- **Padrão observado**
  - Maior frequência semanal
  - Sessões mais longas
  - Maior gasto mensal
  - Maior tempo de permanência
  - Idade intermediária

- **Leitura substantiva**
  - Clientes altamente engajados, que utilizam intensamente o serviço e apresentam maior valor econômico.
- **Coerência estatística**
  - Cluster bem separado
  - Alto alinhamento com o cluster 1 do hierárquico
- Cluster forte, estável e estrategicamente valioso

4.1.1.3 Cluster 3: "Clientes fiéis e econômicos" (121 clientes, ~23%)

- **Padrão observado**
  - Idade mais elevada
  - Frequência relativamente alta
  - Treinos médios
  - Gasto baixo
  - Tempo de plano longo
- **Leitura substantiva**
  - Clientes consistentes e leais, que mantêm vínculo prolongado, mas sem consumo premium.
- **Coerência estatística**
  - Perfil emerge claramente em $k = 3$
  - Equivale ao cluster 2 da solução hierárquica
- Perfil importante que seria perdido em soluções mais grosseiras

A análise descritiva dos clusters obtidos pelo algoritmo K-médias com três grupos revelou perfis claramente distintos de clientes. O primeiro cluster é composto majoritariamente por clientes ocasionais, com baixa frequência, sessões curtas e menor tempo de permanência no plano. O segundo cluster reúne clientes intensivos, caracterizados por elevada frequência semanal, treinos longos, maior gasto mensal e forte fidelização. O terceiro cluster representa clientes fiéis e econômicos, com idade mais elevada, frequência regular, gasto moderado e longo tempo de vínculo. Esses perfis são consistentes com aqueles identificados pela abordagem hierárquica, indicando concordância entre métodos baseados em princípios distintos.