



Instituto Superior de Engenharia de Lisboa

Departamento de Matemática

Técnicas de Estatística Multivariada

Grupo 7

Afonso Santos, N.º 50484

Carolina Cruz, N.º 50475

Constança Castro, N.º 50523

Rodrigo Meneses, N.º 50542

Docente:

Professor Paulo Ramos

2025-05-31

Índice

INTRODUÇÃO	3
ANÁLISE DESCRITIVA DE DADOS MULTIVARIADOS	4
ANÁLISE DE COMPONENTES PRINCIPAIS (ACP).....	9
TESTE KAISER-MAYER-OLKIN (KMO)	10
TESTE DE ESFERICIDADE DE BARTLETT.....	10
ESCOLHA DA MATRIZ A USAR	11
TESTES PARA A HOMOGENEIDADE DAS VARIÂNCIAS	12
TESTE DE LEVENE	12
TESTE DE BARTLETT	13
DETERMINAR AS COMPONENTES PRINCIPAIS	13
PESO E CORRELAÇÃO ENTRE VARIÁVEIS E COMPONENTES PRINCIPAIS	15
<i>Análise dos ‘Loadings’:</i>	15
SCORES COMPONENTES PRINCIPAIS	17
ROTAÇÃO VARIMAX.....	19
ANÁLISE DE CLUSTERS	20
DIAGRAMA DE PERFIL	20
SCREE PLOT.....	21
MEDIDAS DE PROXIMIDADE	22
<i>Distância Euclidiana</i>	22
‘SINGLE LINKAGE’	23
ANÁLISE COEFICIENTE <i>SILHOUETTE</i>	25
ANÁLISE DISCRIMINATE LINEAR.....	27
Regressão Linear.....	32

Introdução

O presente trabalho tem como objetivo a aplicação de técnicas que foram lecionadas durante o semestre de verão, da unidade curricular, Técnicas de Estatística Multivariada (TEM).

O conjunto estudado contém 40 amostras (grãos de café) pertencentes a duas variedades distintas – Gesha e Caturra – com base em diversas características relevantes para a qualidade do café.

As variáveis em estudo incluem atributos como o aroma (X1), sabor (X2), sabor residual (X3), acidez (X4), corpo (X5), equilíbrio (X6), e a única variável qualitativa, variedade do grão de café (X7).

O trabalho tem cinco pontos para seguir, começando com a Análise Descritiva de Dados, seguida da Análise de Componentes Principais (PCA); a Análise de Clusters; uma Análise Discriminante Linear, e por fim uma regressão linear múltipla com o objetivo de explicar a variável “Sabor” (X2), a partir das restantes variáveis.

Deste modo, este trabalho visa aplicar técnicas estatísticas multivariadas na análise de grãos de café, contribuindo para uma melhor compreensão dos fatores que influenciam a qualidade do café.

Análise Descritiva de Dados Multivariados

De modo a estudar os dados, inicialmente fez-se uma análise descritiva destes, através do uso do Software R.

Começou-se por efetuar os comandos do R, seguindo os exemplos fornecidos nos ficheiros disponibilizados no Moodle. Foram selecionadas as variáveis sensoriais quantitativas (X1 a X6), correspondentes ao *aroma*, *sabor*, *sabor residual*, *acidez*, *corpo* e *equilíbrio*.

Utilizando a função `'basicStats()'` do pacote `'fBasics'`, obtiveram-se medidas estatísticas descritivas como a média, mediana, desvio padrão, entre outras. Observou-se que as variáveis apresentavam distribuições aproximadamente simétricas, com médias e medianas próximas, níveis de dispersão relativamente aproximados, entre outras medidas, como se observa na **Figura 1**.

	aroma	sabor	sabor_res	acidez	corpo	equilibrio
nobs	40.000	40.000	40.000	40.000	40.000	40.000
NAs	0.000	0.000	0.000	0.000	0.000	0.000
Minimum	7.250	7.250	7.170	7.170	7.080	7.080
Maximum	8.500	8.500	8.250	8.330	8.250	8.250
1. Quartile	7.648	7.670	7.420	7.580	7.560	7.500
3. Quartile	8.080	8.080	7.940	7.920	7.830	7.920
Mean	7.833	7.851	7.675	7.748	7.675	7.727
Median	7.835	7.920	7.670	7.750	7.670	7.750
Sum	313.330	314.040	307.000	309.930	306.990	309.090
SE Mean	0.049	0.047	0.048	0.041	0.039	0.046
LCL Mean	7.734	7.757	7.578	7.666	7.596	7.634
UCL Mean	7.932	7.945	7.772	7.830	7.754	7.821
Variance	0.096	0.087	0.092	0.066	0.061	0.085
Stdev	0.310	0.294	0.303	0.257	0.248	0.292
Skewness	-0.006	-0.194	0.029	-0.053	-0.105	-0.106
Kurtosis	-0.889	-0.622	-1.195	-0.480	-0.008	-0.864

Figura 1 – 'Round Data'

Analisando detalhadamente estas estatísticas:

- **Médias e Medianas:** Os valores médios variam entre 7,675 e 7,851, com medianos muito próximas das médias, o que pode levantar suspeitas de uma distribuição simétrica dos dados.

- **Dispersão ***: Os desvios-padrão (*'Stdev'*) variam entre 0,248 (corpo) e 0,310 (aroma), sendo comparativamente “baixos” e próximos entre si. Com estes dados, é possível **supor** uma uniformidade nas avaliações e uma baixa variabilidade.
- **Assimetria**: ou *'Skewness'* é uma medida que avalia a falta de simetria numa distribuição em relação à sua média; neste caso, os valores estão próximos de zero (entre -0,194 e 0,029), insinuando que as distribuições das variáveis são aproximadamente simétricas. Graficamente, os valores negativos, mostram uma assimetria à esquerda, o que significa que a cauda de distribuição se estende mais para valores baixos; por outro lado o único valor positivo, 0,029, indica uma leve assimetria à direita, com uma cauda mais longa para valores maiores. Porém tais valores são próximos de zero, e, portanto, não há evidências estatísticas significativas de uma assimetria acentuada e reforça a possibilidade de uma distribuição normal entre as variáveis.
- **Curtose**: ou *'kurtosis'* avalia se os dados se dispersão entre o centro e as caudas de uma distribuição, com valores maiores indicando que uma distribuição possui muitas observações. Por exemplo podemos ter conjunto de dados que são simétricos, mas que possuem diferentes graus de espalhamento; mantidas as médias constantes o conjunto de dados com maior desvio padrão, tende a afastar-se mais da média. Neste caso, os valores são todos negativos (entre -1,195 e -0,008), indicando distribuições platicúrticas, sugerindo uma maior dispersão dos dados em relação à distribuição normal, principalmente no caso do sabor residual.

Aplicou-se a função de covariância, no software R, para se estudar a relação entre as variáveis, como demonstra a **Figura 2**.

```
> cov(dados)
      X1      X2      X3      X4      X5      X6
X1 0.09603276 0.07854795 0.07075000 0.05155199 0.04698160 0.07097327
X2 0.07854795 0.08656821 0.07634615 0.06297359 0.05384128 0.07706436
X3 0.07075000 0.07634615 0.09210256 0.06260641 0.05827308 0.07923974
X4 0.05155199 0.06297359 0.06260641 0.06599429 0.04629827 0.06280788
X5 0.04698160 0.05384128 0.05827308 0.04629827 0.06139481 0.05840827
X6 0.07097327 0.07706436 0.07923974 0.06280788 0.05840827 0.08508199
```

Figura 2 – Covariância

Com os dados presentes, é difícil concluir-se a associação linear entre as variáveis, assim, de tal modo, utiliza-se o coeficiente de correlação amostral ($|r|$); varia entre -1 e 1. Na **Figura 3**, tem-se a tabela de correlações, de onde é possível tirar-se este coeficiente, e inferir uma conclusão; (ou **Figura 4**, Correlograma, que facilita extremamente a visualização)

```
> cor(dados)
```

	X1	X2	X3	X4	X5	X6
X1	1.0000000	0.8614815	0.7522824	0.6475637	0.6118603	0.7851747
X2	0.8614815	1.0000000	0.8550118	0.8331555	0.7385338	0.8979570
X3	0.7522824	0.8550118	1.0000000	0.8030263	0.7749368	0.8951347
X4	0.6475637	0.8331555	0.8030263	1.0000000	0.7273538	0.8381894
X5	0.6118603	0.7385338	0.7749368	0.7273538	1.0000000	0.8081458
X6	0.7851747	0.8979570	0.8951347	0.8381894	0.8081458	1.0000000

Figura 3 – Correlação

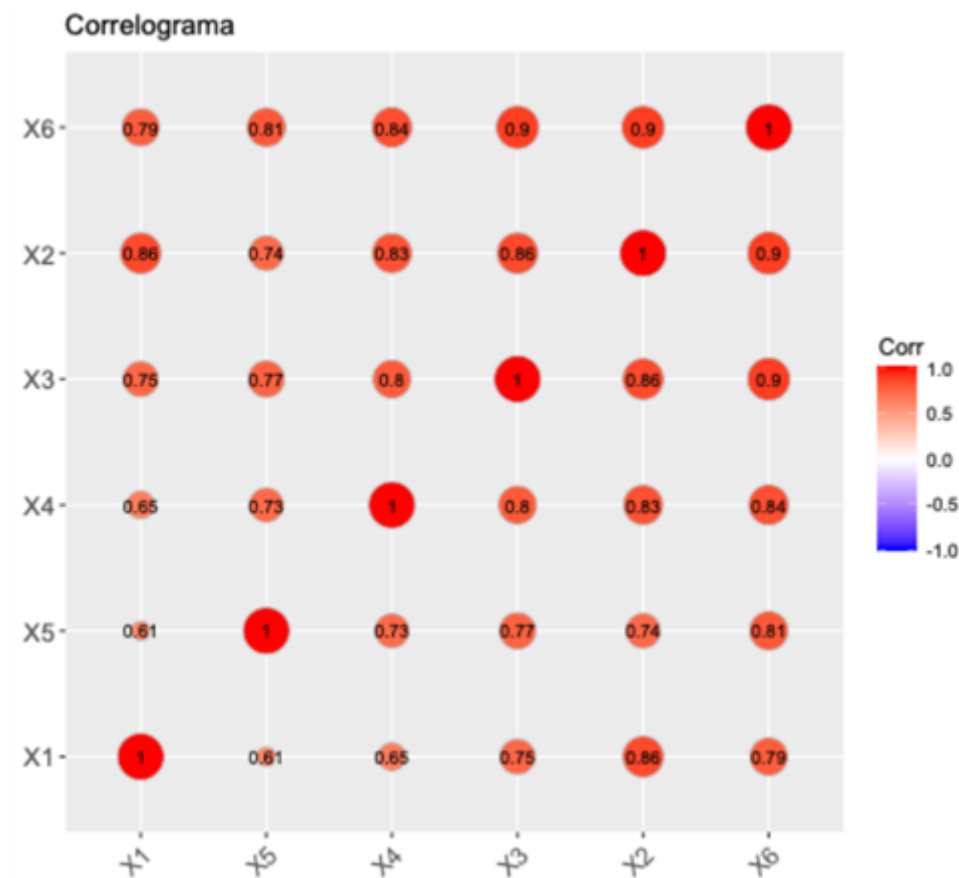


Figura 4 - Correlograma

A partir das figuras, pode-se inferir que todas as correlações são positivas e, portanto, as variáveis tendem a variar para o mesmo sentido.

Há a existência de correlações positivas fortes ($|r_{jk}| \in [0,85; 0,95[$), por exemplo, entre X6 (equilíbrio) e X3 (sabor residual), em que r é, aproximadamente, **0,90**; e com sensivelmente o mesmo valor de coeficiente entre X6 (equilíbrio) e X2 (sabor). Quanto maior o círculo e mais vermelho, maior a associação linear entre as variáveis.

A correlação entre as variáveis **X5** (corpo) e **X1** (aroma), $r_{x1,x5} = 0,61$; **X4** (acidez) e **X1** (aroma), $r_{x1,x5} = 0,65$, são “não aceitáveis”, já que os coeficientes se encontram abaixo de 0,7.

* Apenas com o desvio padrão (s) não é possível chegar a uma conclusão definitiva da dispersão dos dados. Assim, através do coeficiente de variação, ou desvio padrão relativo é que se consegue chegar a um desfecho.

Obtém-se este coeficiente com a seguinte fórmula: $cv_j = \frac{s_j}{|\bar{x}|} * 100\%$

Elaborou-se uma função no software, **Figura 5**, que atua sobre esta fórmula, pegando nos valores do desvio padrão (efetuar raiz quadrada de s_{jj}) e média, já obtidos previamente.

```
# Coeficiente de variação
calcular_cv <- function(x) {
  ((sqrt(sd(x)))/mean(x)) * 100
  #(sd(x) / mean(x)) * 100
}

cv_aroma <- calcular_cv(aroma)
cv_sabor <- calcular_cv(sabor)
cv_sabor_res <- calcular_cv(sabor_res)
cv_acidez <- calcular_cv(acidez)
cv_corpo <- calcular_cv(corpo)
cv_equilibrio <- calcular_cv(equilibrio)

cv_results <- data.frame(
  Variavel = c("Aroma", "Sabor", "Sabor_Residual", "Acidez", "Corpo", "Equilibrio"),
  CV_Percentagem = round(c(cv_aroma, cv_sabor, cv_sabor_res, cv_acidez, cv_corpo, cv_equilibrio), 2)
)

print(cv_results)
```

Figura 5 – Função geradora do coeficiente de variabilidade

Aplicando a função adquirem-se os seguintes resultados:

	Variavel	CV_Percentagem
1	Aroma	7.11
2	Sabor	6.91
3	Sabor_Residual	7.18
4	Acidez	6.54
5	Corpo	6.49
6	Equilibrio	6.99

> |

Figura 6 – Coeficiente de Variabilidade (%)

Os coeficientes encontram-se todos abaixo de 15%, aferindo-se que a variabilidade dos dados é baixa; ou seja, estes estão relativamente mais concentrados em torno

da média e há menos flutuação entre eles. A dispersão é baixa e assim conclui-se que os dados são mais consistentes e menos heterogêneos.

Por último, um *outlier* é ‘uma observação que apresenta um grande afastamento das restantes observações e que é inconsistente com estas (...) têm uma grande influência no valor da média e do desvio padrão (...)’[1] .

```
analise_outliers <- function(vetor, nome_var) {
  Q1 <- quantile(vetor, 0.25, na.rm = TRUE)
  Q3 <- quantile(vetor, 0.75, na.rm = TRUE)
  IQ <- IQR(vetor, na.rm = TRUE)


  lim_inf <- Q1 - 1.5 * IQ
  lim_sup <- Q3 + 1.5 * IQ


  outliers <- vetor[vetor < lim_inf | vetor > lim_sup]
  indices <- which(vetor %in% outliers)


  # Print dos dados
  cat("\n Variável:", nome_var, "\n")
  cat(" Q1:", Q1, "\n")
  cat(" Q3:", Q3, "\n")
  cat(" IQ:", IQ, "\n")
  cat(" Limite Inferior:", lim_inf, "\n")
  cat(" Limite Superior:", lim_sup, "\n")
  cat(" Outliers:", outliers, "\n")
  cat(" Índices:", indices, "\n")


  # Boxplot com título informativo
  boxplot(vetor, main = paste("Boxplot -", nome_var),
    ylab = nome_var, col = "lightblue")
  points(indices, outliers, col = "red", pch = 19)
}
```


Figura 7 – Função outliers

 Variável: aroma
 Q1: 7.6475
 Q3: 8.08
 IQ: 0.4325
 Limite Inferior: 6.99875
 Limite Superior: 8.72875
 Outliers:
 Índices:

 Variável: sabor
 Q1: 7.67
 Q3: 8.08
 IQ: 0.41
 Limite Inferior: 7.055
 Limite Superior: 8.695
 Outliers:
 Índices:

 Variável: sabor_res
 Q1: 7.42
 Q3: 7.94
 IQ: 0.52
 Limite Inferior: 6.64
 Limite Superior: 8.72
 Outliers:
 Índices:

 Variável: acidez
 Q1: 7.58
 Q3: 7.92
 IQ: 0.34
 Limite Inferior: 7.07
 Limite Superior: 8.43
 Outliers:
 Índices: |

 Variável: corpo
 Q1: 7.56
 Q3: 7.83
 IQ: 0.27
 Limite Inferior: 7.155
 Limite Superior: 8.235
 Outliers: 8.25 7.08
 Índices: 5 39


 Variável: equilibrio
 Q1: 7.5
 Q3: 7.92
 IQ: 0.42
 Limite Inferior: 6.87
 Limite Superior: 8.55
 Outliers:
 Índices:

Figura 8 – Output com possíveis outliers

A **Figura 8**, mostra cada variável, com o valor correspondente ao primeiro, terceiro quartil, ao intervalo interquartil (**'IQ'**) e também como, aos limites inferiores e superiores.

A variável **corpo**, é a única que contém pelo menos dois outliers: 7.08 e 8.25.

Com base nestes valores, conclui-se que esta variável tem dois outliers moderados, devido a ultrapassarem o montante de $1,5 * IQ$.

Análise de Componentes Principais (ACP)

Após a análise descritiva dos dados, prossegue-se para a dos componentes principais, que tem como objetivo reduzir a dimensão dos dados.

As componentes principais são novas variáveis, independentes, criadas a partir das variáveis originais do conjunto de dados - são combinações lineares das variáveis iniciais. A redução da dimensão facilita a análise e a interpretação dos dados, mantendo o máximo de informação relevante possível.

Verifica-se então a adequação dos dados através de dois testes estatísticos:

1. Teste Kaiser-Mayer-Olkin (KMO)
2. Teste de Esfericidade de Bartlett

Teste Kaiser-Mayer-Olkin (KMO)

1. O Teste Kaiser-Mayer-Olkin(KMO), usa a matriz de correlações para comparar as correlações entre as variáveis. Através do valor do KMO para cada variável consegue-se detetar quais são as variáveis que não estão correlacionadas com outras.

A figura abaixo, apresenta os resultados adquiridos no software R:

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(dados))
Overall MSA = 0.89
MSA for each item =
```

Aro	Sab	SabRes	Aci	Cor	Equi	Var_numeric
0.83	0.85	0.91	0.90	0.93	0.93	0.88

Figura 9 – Resultado Teste KMO

Decisão: Através dos resultados obtidos consegue-se aferir que a adequação da técnica da análise de componentes principais, para analisar os dados, é boa visto que KMO (0.88) está no intervalo [0,80; 0,90

Teste de Esfericidade de Bartlett

2. O Teste de Esfericidade de Bartlett verifica se existe correlação suficiente entre as variáveis para que a análise de componentes principais seja apropriada. São formuladas duas hipóteses:

$$\left\{ \begin{array}{l} H_0 : \text{A matriz de correlação é igual à matriz de identidade} \\ H_1 : \text{A matriz de correlação é diferente à matriz identidade} \end{array} \right.$$

A **Figura 10**, apresenta os resultados obtidos no Software R. Neste teste foi utilizado um nível de significância de 10 por cento (10%). Normalmente a significância escolhida é 1 por cento (1%), 5 por cento (5%) ou o valor escolhido pelo nosso grupo para o correspondente teste. Este valor está associado ao erro de tipo 1 que corresponde à probabilidade de rejeitar H_0 quando esta

hipótese é verdadeira. Se for escolhido um valor baixo para o nível de significância, a possibilidade de se cometer um erro de tipo 1 diminui.

```
$chisq
[1] 319.6658

$p.value
[1] 3.107208e-55

$df
[1] 21
```

Figura 10 – Output teste de esfericidade de Bartlett

Poder-se-ia calcular o valor crítico (X^2) e compará-lo ao valor que obtemos na figura acima, e assim tomar uma decisão; porém com o p-value obtido, que é aproximadamente 0, e tendo o nível de significância que definimos (10% ou 0,1) tem-se o seguinte:

Regra de decisão: Rejeitar H_0 se $\alpha \geq p - \text{value}$

Decisão: Visto que $\alpha \geq p - \text{value}$, ou seja: $0,1 \geq 3,107208e^{-55}$, deve-se rejeitar H_0 ao nível de significância de 10 por cento (10%), isto é, existe evidência estatística para rejeitar que a matriz de correlações seja idêntica à matriz de identidade. Deste modo, conclui-se que a análise de componentes principais é adequada para analisar estes dados.

Escolha da matriz a usar

Para realizar a ACP, pode-se utilizar tanto a matriz de correlações como a matriz de covariância. A escolha entre uma e outra é determinada através das unidades de medida das variáveis, bem como os valores de variância das mesmas.

No caso das variáveis em estudo, apresentarem unidades de medida diferentes e variâncias diferentes entre si, será mais adequado utilizar a matriz correlações caso contrário, será mais acertado utilizar a matriz de covariância.

Testes para a homogeneidade das variâncias

Os testes a utilizar são o de Levene e Bartlett (homocedasticidade); que servem, superficialmente, para verificar se as variâncias de várias amostras são estatisticamente iguais, ou seja, se as amostras são de populações com homogeneidade de variâncias.

Foi escolhido um nível de significância de 5 por cento (5%) para a resolução dos mesmos; foram formuladas as seguintes hipóteses:

$$\left\{ \begin{array}{l} H_0 : \sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_k \\ H_1 : \exists i, j : \sigma^2_i \neq \sigma^2_j , \quad i = 1 \dots 7 \text{ e } j = 1 \dots 7 \text{ sendo } i \neq j \end{array} \right.$$

Teste de Levene

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value    Pr(>F)
group   6  20.834 < 2.2e-16 ***
      273
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 10 – Output Teste de Levene

Como apresenta a figura acima, o valor de estatística de teste é $F_0 = 20.834$, e com o p-value com o valor de $2.2e^{-16}$. Para o valor crítico em que $k = 7$ e $n = 280$ e, portanto, $F_{6; 280; 0,95}$, este iguala a 2,1025. É também possível fazer uma decisão com base no p-value e no nível de significância, já que é bastante perceptível que α é superior ao esse valor.

Decisão: Sendo que $F_0 > F_{6; 280; 0,95}$ e $\alpha \geq p - value$, deve-se rejeitar H_0 ao nível de significância de 5%, e deste modo, existe evidência estatística significativa para rejeitar que as variâncias da amostra são idênticas.

Teste de Bartlett

```
Bartlett test of homogeneity of variances  
  
data: list(Aro, Sab, SabRes, Aci, Cor, Equi, Var_numeric)  
Bartlett's K-squared = 31.408, df = 6, p-value = 2.118e-05
```

Figura 11 – Output Teste de Bartlett

Através da figura sabe-se que o valor de estatística de teste é $Q_0 = 31.408$ e que o p-value tem como valor, $2.118e - 05$. Para o valor crítico em que $k = 7$ e, portanto, $X_{6; 0,95}$ este iguala a 12,5916. É também possível fazer uma decisão com base no p-value e no nível de significância, já que é bastante perceptível que α é superior ao esse valor.

Decisão: Sendo que $Q_0 > X_{6; 0,95}$ e $\alpha \geq p - value$, deve-se rejeitar H_0 ao nível de significância de 5%, e deste modo, existe evidência estatística significativa para rejeitar que as variâncias da amostra são idênticas.

Posteriormente à realização destes testes e ao resultado da decisão obtido, é seguro aferir que as variâncias da amostra são diferentes. Assim, devemos realizar o estudo utilizando a matriz de correlações, o que equivale a usar a matriz de variâncias covariâncias com os dados estandardizados.

Determinar as componentes principais

A partir da matriz de correlações, foram calculados os valores próprios (*'eigenvalues'*) e vetores próprios (*'eigenvectors'*) que definem as componentes principais. Os primeiros representam a variância explicada por CP, enquanto os vetores determinam os coeficientes das combinações lineares.

```
[1] 5.73433092 0.45785156 0.29131753 0.23248554 0.13011296 0.08912949 0.06477200
```

```
$vectors
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	0.3591014	0.60894481	0.46194888	-0.184541699	-0.2390758	-0.15257972	-0.41514118
[2,]	0.3955841	0.14862235	0.10731501	-0.433537053	0.2795057	-0.06352698	0.73470717
[3,]	0.3924937	-0.05524405	-0.13403616	0.411931831	0.5655246	-0.54365796	-0.19965269
[4,]	0.3704299	-0.28654030	-0.56742422	-0.522456637	-0.2928457	-0.16498341	-0.26975359
[5,]	0.3489524	-0.64838207	0.54072701	0.181386260	-0.3439126	-0.07107111	0.09601674
[6,]	0.4000747	-0.09457207	0.01924717	-0.001286654	0.3751184	0.78384034	-0.27478122
[7,]	0.3761146	0.30429784	-0.37735214	0.549944164	-0.4473845	0.17448058	0.30085054

Figura 12 – Valores e Vetores Próprios

Através da figura acima e segundo o critério de Kaiser, as componentes principais que devem ser retidas são aquelas com valor acima de 1. Neste caso existe apenas uma que cumpre esse critério, e é a primeira componente principal ($\lambda_1 = 5.7343$), e explica 81,9 por cento (%) dos dados, um valor de grande dimensão.

O 'scree plot' é o gráfico onde se representa a percentagem de variância explicada por cada componente. Quando esta percentagem reduz e a curva passa a ser praticamente paralela ao eixo das abcissas, deve-se excluir os componentes correspondentes. Deste modo, devem-se seleccionar todos os componentes até que o efeito referido anteriormente comece a "surgir".

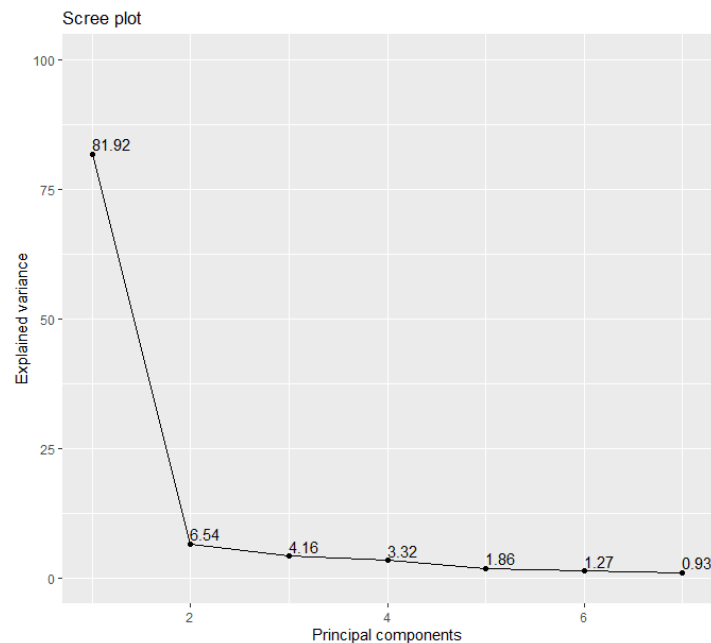


Figura 13 – Scree Plot

Observando o gráfico, a primeira componente principal (PC1), explica aproximadamente 81,9 por cento (%) da variância total dos dados, representando um valor extremamente elevado, e que indica uma forte correlação entre as variáveis originais.

A segunda componente principal (PC2), explica cerca de 6,54 por cento (%) da variância total, constituindo um decréscimo acentuada em relação à anterior.

As restantes componentes, (PC3-PC6), apresentam contribuições progressivamente menores, variando entre 4,16 por cento e 1,27 por cento.

O ponto de inflexão ocorre após a segunda componente, insinuando que PC1 e PC2 detêm a maior parte da variação significativa.

Deste modo, retém-se as duas primeiras componentes; conjuntamente explicam cerca de 88,5 por cento (%) da variância total, proporcionando uma interpretação satisfatória dos dados originais.

A passagem de seis componentes para duas, irá facilitar análises posteriores, tais como a de clusters e discriminante.

Peso e correlação entre variáveis e componentes principais

Análise dos '*Loadings*':

Representam os coeficientes das combinações lineares que definem cada componente principal, designando o peso de cada variável original na formação das novas componentes. Permite compreender quais variáveis contribuem de forma significativa para cada componente e qual a importância interpretativa das mesmas.

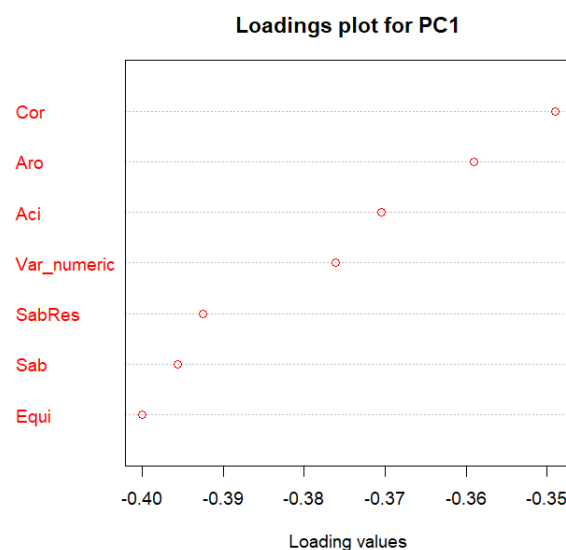


Figura 14 – '*Loading*' PC1

Observando o gráfico acima (PC1), verifica-se que:

- **“Cor” (Corpo)** – apresenta o menor peso em valor absoluto.
- **“Equi” (Equilíbrio)** – apresenta o maior peso em valor absoluto.

Todos os *‘loadings’* apresentam valores negativos e relativamente próximos entre si, indicando que as variáveis variam no mesmo sentido em relação a PC1. Quando PC1 aumenta, todas tendem a diminuir e vice-versa. Isto sugere que a primeira componente principal representa uma medida geral de qualidade do café, onde todas as características sensoriais estão fortemente relacionadas. Esta componente, representa 81,9 por cento (%) da variância total dos dados.

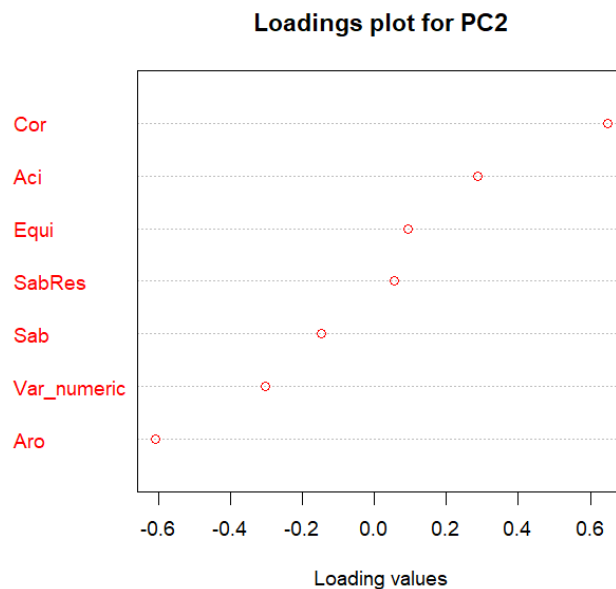


Figura 15 – ‘Loading’ PC2

Para a segunda componente os loading apresenta maior variabilidade, sendo:

- **“Cor” (Corpo)** – apresenta o valor mais elevado.
- **“Equi” (Equilíbrio)** – maior contribuição negativa.

A PC2 representa um contraste entre características físicas e aromáticas do café. Valores positivos estão relacionados a café com maior corpo e acidez, enquanto valores negativos correspondem a cafés com maior intensidade aromática. Esta componente explica uma menor porção da variância: 6,54 por cento (%).

A correlação entre a variável original X_i e a componente principal Y_j é calculada por:

- Para dados não estandardizados:

$$r_{Y_j, X_i} = a_{ij} \frac{\sqrt{\lambda_j}}{\sqrt{s_{ii}}} = a_{ij} \frac{\sqrt{\lambda_j}}{\sqrt{s_i^2}}$$

- Para dados estandardizados:

$$r_{Y_j, Z_i} = a_{ij} \sqrt{\lambda_j}$$

No caso deste específico projeto, que se utiliza a matriz de correlações, aplica-se a fórmula para dados estandardizados:

- Para PC1($\lambda_1 = 5.7343$), por exemplo, escolhendo equilíbrio, tem-se
 $r = -0.399 \cdot 2.395 = -0.956$; e para sabor: $r = -0.392 \cdot 2.395 = -0.939$
- Para PC2($\lambda_2 = 0.4579$), escolhendo corpo, tem-se
 $r = 0.65 \cdot 0.676 = 0.4394$; para equilíbrio temos: $r = 0.15 \cdot 0.676 = 0.1014$

Scores Componentes Principais

Representam as coordenadas de cada observação no novo sistema de eixos definido pelas componentes principais. Esta permite:

1. Observações com scores semelhantes apresentam características semelhantes;
2. Agrupamentos de pontos no espaço das componentes;
3. Pontos distanciados;

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
[1,]	-3.7588444	-0.801830470	0.763450960	-0.8125392145	0.1618194321	-0.43772025	-0.201535752
[2,]	-3.6311888	1.113607299	-0.111255564	-0.0615157367	-0.2123951017	0.42117981	0.229818049
[3,]	-3.2784629	-0.772898606	0.476092513	0.2673185723	0.9622388034	0.11301812	-0.006668062
[4,]	-3.1983434	-0.507466377	0.122600584	-0.0737613500	0.4025486534	-0.33186623	0.534902618
[5,]	-2.6978931	0.849353265	1.186738215	1.0332697248	-0.1159965537	-0.46008031	0.071239298
[6,]	-2.6919545	0.522689603	-0.134285277	0.2884587254	-0.2147347637	0.28029032	0.407119261
[7,]	-2.7024383	1.506766279	-1.200323453	-0.3349020684	-0.1709943055	0.13237101	-0.356751331
[8,]	-2.2712462	0.165712543	0.017343172	0.2442072472	0.0499521073	0.22357308	-0.389559016
[9,]	-2.0236050	-0.246568952	0.055884852	-0.2090058007	-0.2498267431	-0.42568018	-0.336616750
[10,]	-1.9088984	-0.112354486	-0.474884304	0.5963662470	0.3003721651	0.04916416	0.428583685
[11,]	-1.7156698	0.213688513	-0.389501728	0.5943357617	-0.0009047522	0.10439075	0.162419787
[12,]	-1.7568600	-0.967091300	-0.651355358	-0.2442003928	-0.0507988276	0.34023991	0.141262052
[13,]	-1.8032679	-0.193265527	-0.415651438	0.3702525693	0.1244127195	0.13810645	0.100823569
[14,]	-1.4553078	-0.806138703	-0.064436353	-0.2513304546	-0.1771122549	-0.45770426	-0.497113302
[15,]	-1.4650713	-0.003639204	-0.201104209	0.3163441742	-0.4951505978	0.08109673	0.072969412
[16,]	-1.3896996	-1.209956271	0.250973181	0.1606723903	-0.2230396315	0.29864311	-0.183529155
[17,]	-1.3107079	0.591467073	-0.709206001	0.1935360055	-0.4268090211	-0.37578054	-0.145780393
[18,]	-1.2608358	-0.152873483	-0.481725337	-0.1220941620	-0.7734589767	-0.17709406	0.082715858
[19,]	-1.1816391	-1.141986601	0.308422137	0.3710125332	-0.0701250765	0.20787593	-0.374704651
[20,]	-1.0907755	-0.798628138	-0.518706164	0.1984907255	0.0304370309	-0.03113840	-0.124426470
[21,]	0.6803850	0.527276661	0.132679608	-0.5606545251	0.4391629330	0.49330791	-0.232254791
[22,]	0.5492105	0.775154501	-0.229953432	-0.8963354047	0.1079064642	0.17997605	0.196569120
[23,]	0.7967814	0.510893711	0.172428829	-0.6828153624	0.2714532407	0.33208291	-0.291462986
[24,]	0.7121845	0.570137688	1.480763856	-0.2700213543	-0.4185238713	-0.14747650	0.172088731
[25,]	1.2641830	0.408466425	-0.014261449	-0.3997903012	0.4715029415	0.19114504	-0.236420511
[26,]	1.1656344	0.628127642	0.994392215	-0.1253325211	-0.2320030224	0.34315098	-0.128290154
[27,]	1.3460770	0.218318900	-0.207862035	-0.6568155908	0.1903897498	-0.51411226	0.453574578
[28,]	1.4581657	0.265528264	-0.136140815	-0.5583995576	0.4270170051	-0.59253005	0.049235141
[29,]	1.5487642	-0.517410634	0.812364564	-0.4475755228	-0.4058344970	0.44893047	0.375270506
[30,]	2.1862414	-0.380461060	-0.247976112	-0.0009536017	0.7574108735	0.19069314	0.006118619
[31,]	1.8849483	1.101076224	0.099723371	0.5286823602	0.6114959338	-0.31114073	-0.075657769
[32,]	2.2858239	-0.185358605	0.471716277	0.1480379291	0.1595691776	-0.07366501	0.096283989
[33,]	2.2573410	-0.389958809	0.002151812	-0.9381998324	-0.5593433565	-0.11010085	0.032401560
[34,]	2.6282072	-0.453654591	-0.247388102	-0.3057383635	-0.1334562941	0.01681186	0.395366391
[35,]	3.0409228	0.440555592	0.402845141	0.7750231491	-0.0266230644	0.19067723	-0.124741725
[36,]	3.0640019	-0.238012694	-0.330038970	-0.0414058486	-0.0268506412	0.12358474	-0.001159537
[37,]	3.0083920	0.730591876	0.115580703	0.5230234086	-0.1117525430	-0.19891329	-0.125228200
[38,]	3.4606287	0.268166327	-0.204191566	0.2977055810	-0.1544937891	-0.14626458	-0.200639744
[39,]	4.6036910	-0.519347115	-0.898584612	0.4217461206	0.0682796557	-0.30601280	0.025950664
[40,]	4.6511258	-1.008677157	0.012680286	0.6649037408	-0.2857412014	0.19697058	-0.002172589

Figura 16 – Valores ‘MeanCentered’

A figura acima, representa a tabela de scores das componentes principais (valores ‘mean-centered’), e observa-se:

Para PC1:

- Os valores variam entre a observação 1 e a observação 40, sendo a primeira o valor mais baixo;
- Valores negativos: indicam amostras com características sensoriais acima da média geral;
- Valores positivos: representam amostras com características sensoriais abaixo da média geral;

Para PC2:

- Os valores oscilam entre a observação 4 e 7;
- Valores positivos: correspondem a maior corpo e acidez;
- Valores negativos: relacionados com a intensidade aromática e equilíbrio;

Rotação VARIMAX

De modo a melhorar a interpretação das componentes, aplica-se a rotação VARIMAX, uma técnica ortogonal que maximiza a variância dos pesos, simplificando a estrutura fatorial.

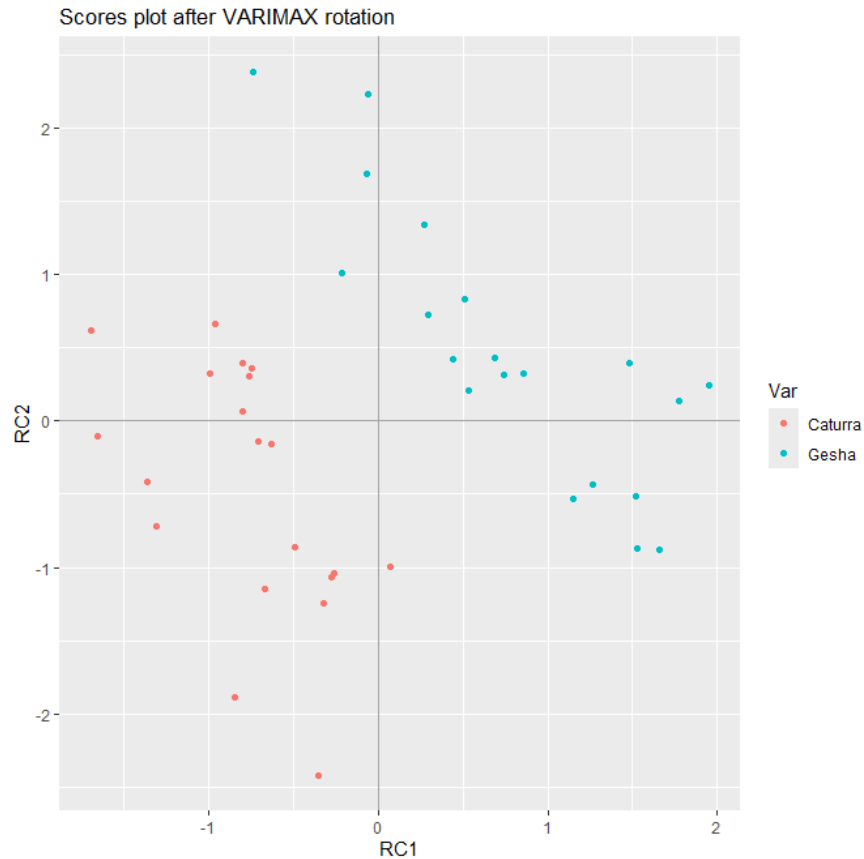


Figura 17 – Gráfico VARIMAX

O gráfico acima, revela uma separação mais clara entre as duas variedades dos grãos de café:

- Variedade Caturra (pontos a vermelho): Concentra-se predominantemente na região negativa de RC1
- Variedade Gesha (pontos a azul): Distribui-se mais amplamente, com maior concentração na zona positiva de RC1.

Análise de Clusters

A análise de clusters, tem como objetivo agrupar observações em grupos homogêneos. No contexto deste projeto, pretende-se identificar grupos de grãos de café com características sensoriais semelhantes, independentemente da sua variedade. Gesha ou Caturra.

Diagrama de Perfil

É uma representação gráfica que permite ter ideia do número de clusters a formar, quando se tem um número moderado de variáveis. Este diagrama lista as variáveis ao longo do eixo horizontal e a escala de valores ao longo do eixo vertical.

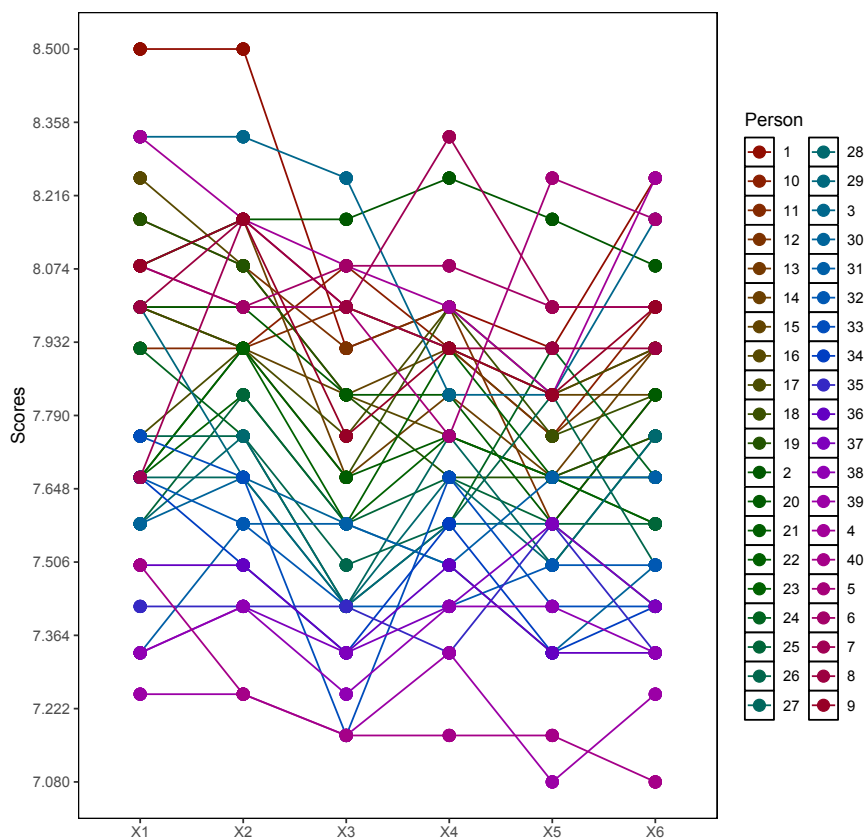


Figura 18 – Diagrama de Perfil

Através da análise deste diagrama conseguimos concluir que X1 e X2 apresentam consistentemente as maiores pontuações, enquanto X3 representa por norma um ponto crítico onde as pontuações tendem a baixar ligeiramente. Conseguimos observar também que X3, X4 e X6 apresentam uma larga variabilidade o que nos indica que provavelmente irão ter um impacto elevado na formação de clusters.

Scree Plot

O Scree Plot representa a variação da soma de quadrados dos desvios dentro dos grupos (WSS) em função do número de clusters. O ponto de inflexão no gráfico sugere o número “ótimo” de clusters, seguindo o critério do “cotovelo”.

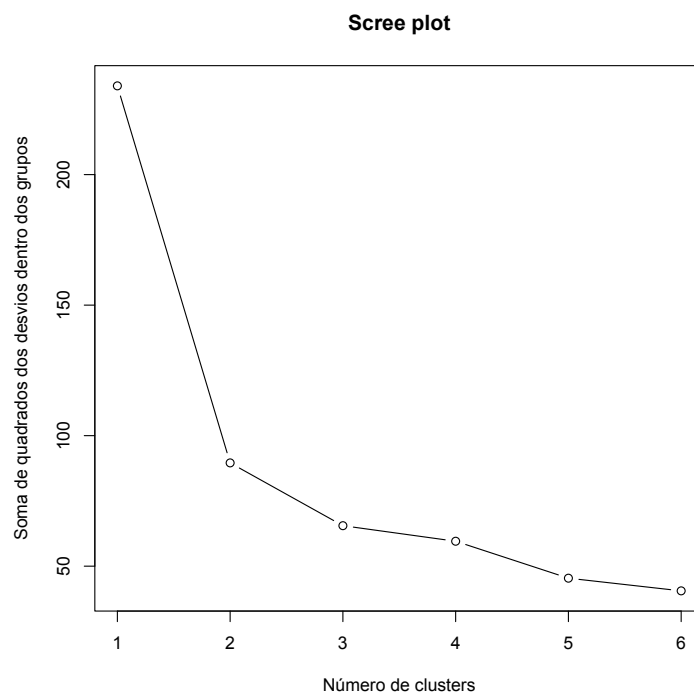


Figura 19 – Scree Plot

Através da figura acima é possível aferir que há uma redução muito significativa da WSS (Within-cluster SUM of Squares), quando se passa de um para dois clusters; sendo de facto a maior diminuição presente no gráfico.

Portanto, o ponto de inflexão mais evidente ocorre quando $k = 2$, após esse valor esta redução torna-se bastante gradual e menos pronunciada. Adicionar mais clusters não proporciona melhorias substanciais na homogeneidade dos grupos.

Medidas de Proximidade

Distância Euclidiana

Para medir a similaridade entre as observações, utilizou-se a medida referida acima; que é a forma de distância mais comum na análise de clusters.

Dado que as variáveis apresentam escalas semelhantes, ou seja, todas avaliadas numa escala de contínua de qualidade sensorial, mas valores de variâncias assimilares, conforme verificado nos testes de homogeneidade, os dados foram estandardizados antes de se prosseguir.

Distância Mahalanobis

Esta distância não foi o método usado para a determinação da proximidade entre as variáveis pois apesar de esta fazer quase o mesmo que a distância euclidiana, não seria a escolha mais acertada visto que Mahalanobis é mais usada em contextos onde as variáveis apresentam diferentes escalas de medida

Métodos hierárquicos

De modo a se ter uma representação visual da formação dos clusters, aplica-se diferentes métodos hierárquicos.

Dendrograma e métodos de ligação

Para representar a formação dos clusters, utiliza-se um dendrograma, um método visual.

Optou-se por métodos hierárquicos aglomerativos para a criação dos dendrogramas devido à sua simplicidade e capacidade de formar clusters de maneira intuitiva. Esses métodos permitem uma análise clara da estrutura dos dados e das relações entre os grupos formados.

Considerou-se o uso do método de ligação média, porque este tende a produzir melhores resultados (**Figuras 20**).

‘Average Linkage’

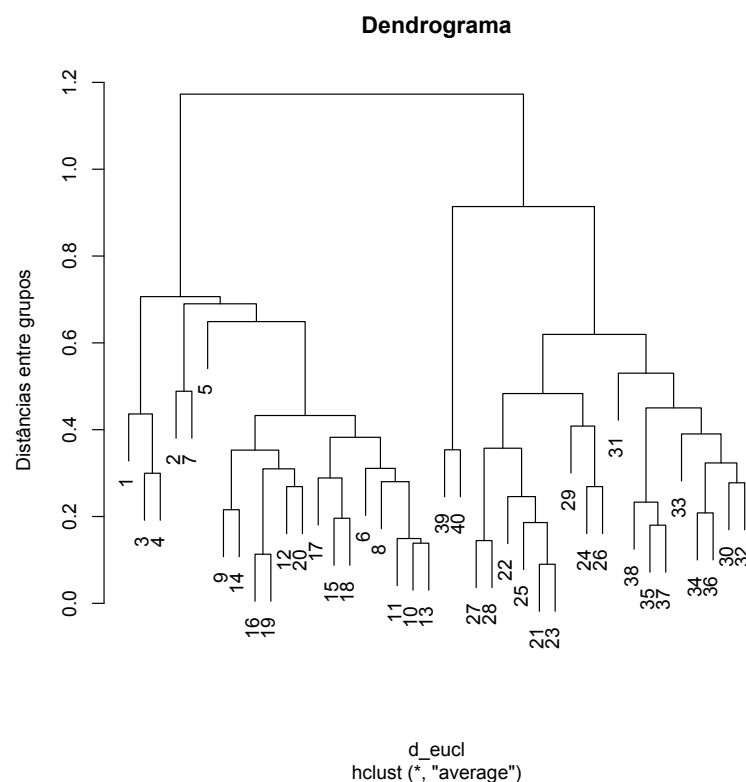


Figura 20 – Dendrograma

O dendrograma acima foi construído com o método ‘*average linkage*’, em que o eixo vertical representa as distâncias entre grupos, variando aproximadamente entre 0.0 e 1.2.

É possível observar que a maioria das fusões internas ocorre entre as distâncias de 0.,1 e 0,6.

Esta diferença salientada entre as alturas das fusões evidencia a existência de dois grupos distintos.

Este método fundamenta-se na média das distâncias entre todos os pares de elementos de grupos diferentes, e assim, tende a gerar clusters mais equilibrados e compactos.

O método de ligação por média baseia-se na média das distâncias entre todos os pares de elementos de grupos diferentes, o que tende a gerar clusters mais equilibrados e compactos. Assim, a escolha de $k=2$ clusters é justificada pela diferença significativa nas distâncias de fusão e pela ausência de ligações intermédias entre os dois principais grupos.

A **Figura 21**, mostra dois gráficos (complementares) acerca dos valores do D índice. Funcionam de modo a determinar o número ideal de clusters e fortificar a conclusão retirada na **Figura 19**.

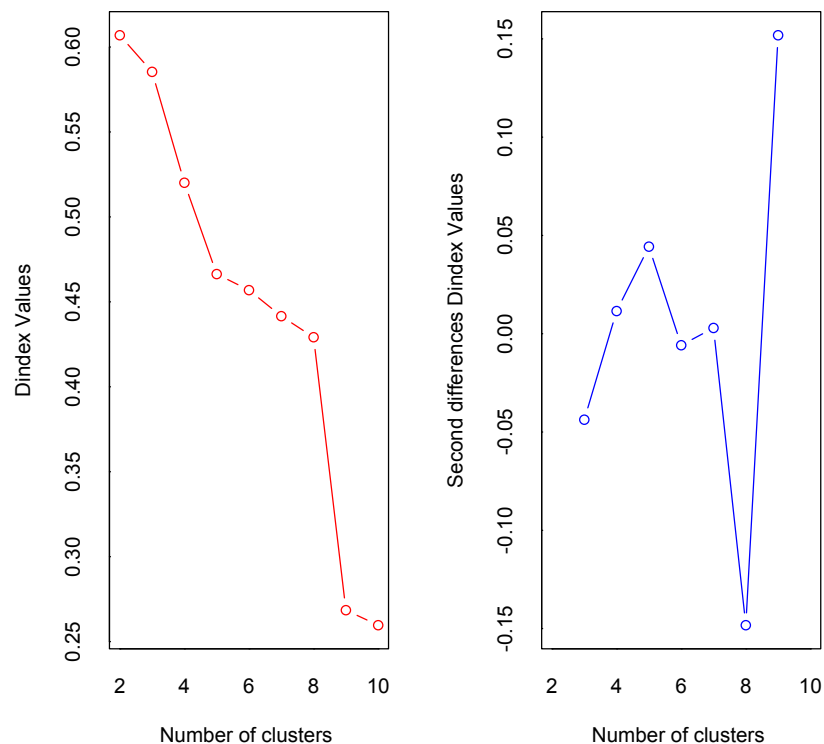


Figura 21

O gráfico à esquerda, “*Dindex values*”, ou valores do índice de D, mostra uma curva decrescente contínua destes valores, conforme o número de clusters aumenta.

Começa aproximadamente em 0.60, quando $k = 2$, e diminui progressivamente até cerca de 0.25 quando $k = 10$.

O gráfico à direita, é mais revelador; quando $k = 2$ há um pico muito pronunciado, indicando uma mudança significativa na curvatura.

Assim tanto o scree plot, quanto os gráficos, convergem para $k = 2$ como número ideal de clusters.

Análise Coeficiente *Silhouette*

Esta análise mede o quão bem uma observação está relacionada a um cluster e estima a distância médias entre eles.

Para cada observação i , calcula-se a dissimilaridade média a_i entre i e todos os outros pontos do cluster A ao qual i pertence.

Para todos os outros clusters C , calcula-se a dissimilaridade média $d(i,C)$ de i para todas as observações de C . O menor destes valores é definido como b_i .

O coeficiente de silhouette para a observação i é então dado por:

$$s_i = \frac{b_i - a_i}{\max \{a_i, b_i\}}$$

Um valor pequeno do coeficiente, aproximadamente 0, significa que a observação está entre dois clusters, como mostra a **Figura 22**. Isto poderá indicar que o método “Single Linkage” com a distância euclidiana está a produzir um resultado abaixo do ótimo:

1. Única observação forma um cluster isolado, enquanto as outras se agrupam num cluster de grande dimensão;
2. A separação entre clusters não é muito clara;

cluster	size	ave.sil.width
1	1	39
2	2	1

Figura 23

Análise Discriminante Linear

A análise discriminante é uma técnica usada em estatística multivariada, que se aplica quando a variável dependente é qualitativa e as variáveis independentes são quantitativas.

Deste modo, são criadas funções discriminantes, provenientes de combinações lineares das variáveis iniciais, que maximizam as diferenças entre as médias dos grupos e minimizam a probabilidade de classificações incorretas dos casos nos grupos.

Foi realizado um teste de Shapiro-Wilks, para a normalidade multivariada, para averiguar se as variedades do café são provenientes de uma população normal multivariada.

Neste caso, usou-se um nível de significância de 5 por cento (%); a **Figura 24** demonstra os resultados de output.

<pre>> mshapiro.test(t(caturra))</pre>	<pre>> mshapiro.test(t(gesha))</pre>
Shapiro-Wilk normality test	Shapiro-Wilk normality test
data: Z	data: Z
W = 0.87932, p-value = 0.0172	W = 0.70364, p-value = 4.319e-05

Figura 24 – Teste de Shapiro-Wilks

É possível aferir que os p-value dos dois testes, estão abaixo de 2 por cento(%), inclusive o do grão gesha, é aproximadamente 0. Assim, deve-se rejeitar H_0 , ao nível de significância de 5% (já que $0,05 > 0,0172$ & $0,05 > 4,319e^{-5}$, e assumir que os dados proveem de uma população normal multivariada.

Para além do teste acima, deve-se realizar o teste M de Box, para averiguar a homogeneidade das matrizes de variâncias covariâncias; mais uma vez utilizou-se um nível de significância de 5 por cento (%); tem-se H_0 como: “As matrizes de variâncias covariâncias são idênticas para todos os grupos” e H_1 : “As matrizes de variâncias covariâncias não são idênticas para todos os grupos”. A Figura 25, tem os resultados deste teste:

```
> boxM(dados, variedade)
```

Box's M-test for Homogeneity of Covariance Matrices

data: dados
Chi-Sq (approx.) = 36.193, df = 21, p-value = 0.0208

Figura 25 – Teste M de Box

Com os resultados, conclui-se que se deve rejeitar H_0 , ao nível de significância de 5% (já que $0,05 > 0,0208$), e assim aferir que as matrizes não são idênticas para todos os grupos.

Função Discriminante

A análise discriminante é realizada através de uma ou mais combinações lineares das variáveis independentes utilizadas (X_i). Cada combinação linear (Y_i) constitui uma função discriminante:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p,$$

em que:

- a_{ij} são os coeficientes de ponderação
- X_j são as variáveis discriminantes não normalizadas

Os coeficientes servem para identificar quais as variáveis que mais contribuem para distinguir os grupos dentro de uma mesma função Y_i , e, portanto, quanto maior o seu valor, maior será a contribuição da variável na discriminação dos grupos.

Os valores abaixo, na **Figura 26**, são as probabilidades a priori dos elementos de cada grupo. No caso do projeto, cada grupo tem 20 elementos e, portanto, o resultado será o seguinte:

```
> modelo_lda$prior
Caturra  Gesha
      0.5    0.5
```

Figura 26

Os resultados específicos das funções discriminantes, sabem-se, no Software R, utilizando “\$means, \$scaling e \$counts”.

Se se tivesse um grupo um grupo com mais observações, o modelo ajustaria essas probabilidades proporcionalmente.

```
> modelo_lda$means|
      X1    X2    X3    X4    X5    X6
Caturra 7.5920 7.6175 7.417 7.5545 7.5165 7.4875
Gesha   8.0745 8.0845 7.933 7.9420 7.8330 7.9670
```

Figura 27 – Modelo “means”

Acima, é possível observar-se a média de cada variável sensorial para as duas variedades:

- **Gesha:** Apresenta valores médios mais elevados em todas as características, sugerindo que essa variedade possui um perfil sensorial mais intenso;

- **Caturra:** tem médias ligeiramente mais baixas, indicando diferenças perceptivas nas propriedades sensoriais.

Como os grupos têm a mesma dimensão, a Função Linear Discriminante ajusta-se sem precisar de correções adicionais.

Os coeficientes discriminantes que variáveis mais contribuem para a distinção entre grupos, e os resultados abaixo, mostram que X1 e a X3 são os fatores mais relevantes.

```
> modelo_lda$scaling
      LD1
X1  3.092076
X2 -2.084192
X3  4.393511
X4  2.291927
X5 -1.557190
X6  1.790304
```

Figura 28 – Modelo “scaling”

Coeficiente de correlação canónica

Este coeficiente dá-nos uma medida da importância da função discriminante e avalia como é que a função discriminante está relacionada com os grupos.

Se a correlação canónica de uma dada função discriminante for 0 então não existe qualquer relação entre a função discriminante correspondente e os grupos.

Se a correlação canónica for próxima de 1 então existe uma relação significativa entre a função discriminante correspondente e os grupos.

O coeficiente de correlação canónica de uma função discriminante é dado por:

$$r_j^* = \sqrt{\frac{\lambda_j}{1 + \lambda_j}}$$

O quadrado desse coeficiente representa a proporção da variância da função discriminante explicada pelos grupos.

O cálculo foi realizado da seguinte forma, e resultado como output:

```
> sqrt((eigen(solve(withinSS(dados,variedade))%*(totalSS(dados)-withinSS(dados,variedade)))$values[1:2])/(1+
(eigen(solve(withinSS(dados,variedade))%*(totalSS(dados)-withinSS(dados,variedade)))$values[1:2])))
[1] 0.9006117      NaN
```

Figura 29 – Coeficiente de correlação canónica

A correlação canónica é próxima de 1 (0.9006117), e então é exequível referir que existe uma relação significativa entre a função discriminante correspondente e os grupos.

Testes de Significância para as funções discriminantes

De modo a avaliar a significância das funções discriminantes usa-se o teste lambda de Wilks, com a distribuição Qui-Quadrado.

Tem como objetivo avaliar as funções que são significativas e assim decidir as que devem ser consideradas. Para este teste existem as seguintes hipóteses:

$$\begin{aligned}H_0: \lambda_1 = \lambda_2 = \dots = \lambda_m = 0 \\ H_1: \exists \lambda_j \neq 0 \\ \text{com } j = 1, \dots, m.\end{aligned}$$

A **Figura 30**, mostra o output deste teste:

```
> ### TESTE DE SIGNIFICÂNCIA: LAMBDA DE WILKS
> Wilks.test(dados, grouping = variedade, method = "c")

One-way MANOVA (Bartlett Chi2)

data: x
Wilks' Lambda = 0.1889, Chi2-Value = 58.329, DF = 6.000,
p-value = 9.828e-11
sample estimates:
      X1      X2      X3      X4      X5      X6
Caturra 7.5920 7.6175 7.417 7.5545 7.5165 7.4875
Gesha   8.0745 8.0845 7.933 7.9420 7.8330 7.9670
```

Figura 30 - Teste Lambda de Wilks

O valor de Wilks obtido é bastante reduzido, o que sugere uma boa separação entre os grupos; o p-value é aproximadamente 0 e, portanto, com qualquer nível de significância 1%, 5% ou 10%, iria-se rejeitar H_0 .

Conclui-se que o modelo discriminante é estatisticamente válido para diferenciar as variedades as médias das variáveis sensoriais mostram diferenças entre grupos.

Classificação dos indivíduos em k grupos

A análise discriminante é uma técnica de classificação importante, que possibilita identificar o grupo mais provável entre os k, a que um dado indivíduo pertence, a partir dos seus valores para as variáveis discriminantes.

Quando se tem k grupos tem-se que definir k funções de classificação ou funções classificatórias.

As funções classificatórias ou de classificação podem ser usadas para:

- Avaliar a eficácia classificativa da análise discriminante
- Classificar novos indivíduos nos grupos definidos

Para avaliar a eficiência do processo de classificação podem calcular-se probabilidades de classificação incorreta.

Existem vários processos para estimar estas probabilidades, nomeadamente:

- Método de resubstituição
- Método de validação cruzada
- Método de Jackknife

Foi criada a seguinte matriz para avaliar a precisão do modelo:

```
> ### CLASSIFICAÇÃO DOS INDIVÍDUOS
> pred <- predict(modelo_lda)
> conf_matrix <- table(variedade, pred$class)
> conf_matrix
```

variedade	Caturra	Gesha
Caturra	20	0
Gesha	0	20

Figura 31

Os resultados mostram uma precisão de 100%, ou seja, o modelo separou perfeitamente os grupos.

Todos os 20 indivíduos de Caturra foram corretamente classificados como Caturra e todos os 20 indivíduos de Gesha foram corretamente classificados como Gesha, ou seja, não houve erros na classificação.

```
> prop.table(conf_matrix, 1)

variedade Caturra Gesha
Caturra    1      0
Gesha      0      1
> accuracy <- sum(diag(prop.table(conf_matrix))) # % de classificação correta
> accuracy
[1] 1
```

Figura 32

Obteve-se 100% de precisão na classificação ($\text{accuracy} = 1$), isso significa que o modelo discriminante separou completamente Gesha e Caturra com base nas variáveis sensoriais.

Gráficos dos Scores Discriminantes

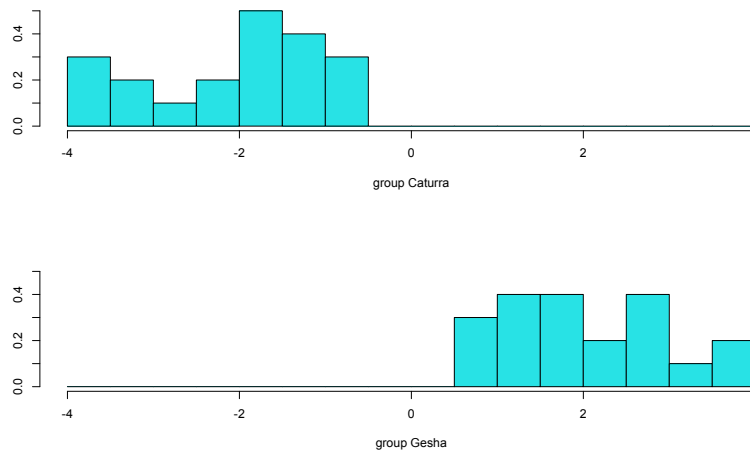
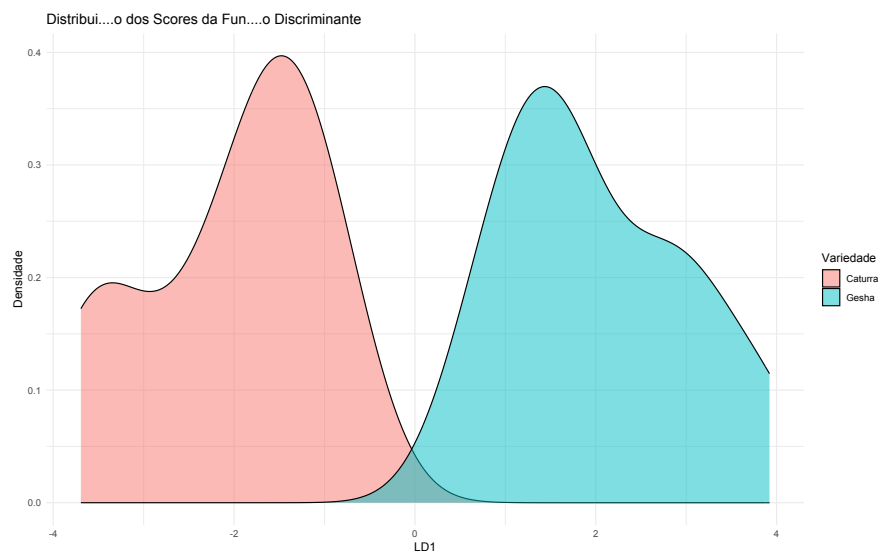


Figura 33 - Gráfico dos Scores Discriminantes

Há dois, um para cada grupo (Caturra e Gesha), o eixo X representa os scores discriminantes, que indicam como os dados foram separados pelas funções discriminantes e o eixo Y representa a frequência desses scores. A separação visível entre os grupos indica que a LDA conseguiu distinguir bem as variedades de café.

A separação visível confirma a eficácia do modelo.

Gráficos de Densidade dos Scores Discriminantes



O eixo X representa os valores da primeira função discriminante (LD1) e o eixo Y representa a densidade, indicando a frequência relativa dos scores. As áreas coloridas indicam como os grupos de café estão distribuídos e é observável a separação entre os grupos

As curvas de densidade mostram que Gesha e Caturra estão bem separados e a separação é quase perfeita, reforçando a eficácia do modelo.

Modelo de regressão linear múltipla

Procura-se realizar uma regressão linear múltipla com o objetivo de explicar a variação do Sabor (X2), função das restantes variáveis fornecidas.

Para este efeito, ajustou-se o modelo de regressão utilizando a função **lm** do software estatístico R, com a variável X2 definida como variável dependente, sendo as restantes independentes. Na figura seguinte está apresentado o resultado obtido.

```
Call:
lm(formula = X2 ~ X1 + X3 + X4 + X5 + X6, data = dados)

Residuals:
    Min       1Q   Median       3Q      Max
-0.159654 -0.077059 -0.009161  0.072809  0.179177

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.38402    0.56517  -0.679   0.5014
X1           0.38492    0.08691   4.429 9.33e-05 ***
X3           0.07987    0.12784   0.625   0.5363
X4           0.30729    0.12021   2.556   0.0152 *
X5           0.01504    0.11517   0.131   0.8969
X6           0.27313    0.16054   1.701   0.0980 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1018 on 34 degrees of freedom
Multiple R-squared:  0.8957,    Adjusted R-squared:  0.8804
F-statistic: 58.41 on 5 and 34 DF,  p-value: 1.023e-15
```

Figura 34 – Chamada da função lm

Conseguiu-se obter que a equação do modelo global de regressão linear múltipla ajustado aos dados fornecidos é a seguinte:

$$\hat{y}_i = -0.38402 + 0.38492x_{i1} - 0.07987x_{i3} - 0.3026x_{i4} + 0.01504x_{i5} + 0.27313 x_{i6}$$

Análise de modelo populacional e ajustado

Neste caso, o coeficiente de determinação r^2 não é o melhor critério para avaliar a qualidade do ajuste do modelo, pois ele tende a aumentar sempre que se adiciona uma nova variável independente, mesmo que essa variável não contribua significativamente para explicar a variável dependente. Consequentemente, é mais adequado utilizar o coeficiente de determinação

ajustado (r^2 ajustado), que corrige essa limitação. Esse indicador só aumenta quando a nova variável realmente melhora o modelo, ou seja, quando traz informação relevante.

Com base no valor do r^2 ajustado, pode-se afirmar que o modelo de regressão linear múltipla apresenta um bom ajuste aos dados. Isso porque aproximadamente 88,04% da variação no sabor do café é explicada pelas outras variáveis. Além de tal, como os coeficientes são relativamente estáveis, é possível concluir que as variáveis utilizadas têm, em geral, relevância estatística no contexto do modelo.

Análise de resíduos

Obtivemos a partir da figura anterior as medidas descritivas dos resíduos e observámos que efetivamente a média dos resíduos é aproximadamente 0, enquanto o desvio padrão dos resíduos é de 0.1018. Em seguida, utilizando o software estatístico R construímos um gráfico de comparação de quantis para tentarmos tirar conclusões quanto à normalidade dos resíduos, este encontra-se na figura abaixo:

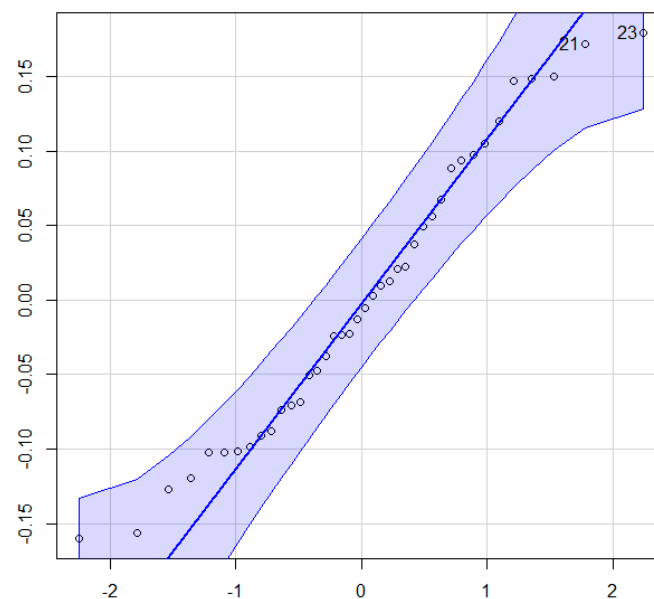


Figura 35 – Comparação de quantis

O gráfico de quantis apresentado compara os quantis dos resíduos padronizados com os quantis de uma distribuição normal teórica. Ao analisar, conseguimos reparar que os pontos seguem de forma bastante próxima a linha reta diagonal, sugerindo que a distribuição dos resíduos está alinhada com a normalidade esperada.


De modo a retirar conclusões mais acertadas, recorreremos à realização de um teste de normalidade dos resíduos de Shapiro-Wilk, cujo resultado se encontra abaixo.

shapiro-wilk normality test

```
data: residuos  
W = 0.96256, p-value = 0.2046
```

Figura 36 – Teste Shapiro-Wilk

Formulação de hipóteses:


$$\left\{ \begin{array}{l} H_0 : \text{A amostra dos resíduos é proveniente de uma população com} \\ \text{distribuição normal.} \\ \\ H_1 : \text{A amostra dos resíduos é proveniente de uma população com distribuição} \\ \text{diferente da normal.} \end{array} \right.$$

- Nível de significância $\alpha = 0.01$
- $p - value = 0.2046$

Decisão: Como o p-valor (0.2046) é superior ao nível de significância de 1%, não se rejeita a hipótese nula. Logo, existem evidências estatísticas suficientes para afirmar que a amostra de resíduos provém de uma população com distribuição normal.

Teste de significância para os coeficientes de regressão:

Através da **Figura 34** (resultado lm), é possível ainda ver os resultados dos testes da significância de cada parâmetro, através do valor da estatística de teste ou através do seu $p-value$.

Assim, e com base nos $p-values$, pode-se concluir que:

1. O intercepto não é estatisticamente significativo ($p-value = 0.5014$), indicando que, quando todas as variáveis independentes são zero, o valor esperado da variável dependente não é significativamente diferente de zero.
2. As variáveis X1 ($p-value = 9.33e-05$) e X4 ($p-value = 0.0152$) são estatisticamente significativas ao nível de 1% e 5%, respectivamente, sugerindo uma forte influência sobre a variável dependente.
3. A variável X6, com um $p-value = 0.0980$, apresenta significância ao nível de 10%, indicando uma possível relação com a variável dependente, embora com menor evidência estatística.

4. As variáveis X3 ($p\text{-value} = 0.5363$) e X5 ($p\text{-value} = 0.8969$) não são estatisticamente significativas ao nível de 5%, pois seus valores de $p\text{-value}$ são maiores que 0.05.

Esses resultados sugerem que, no modelo ajustado, X1 é a variável mais significativa, seguida por X4 e, em menor grau, por X6. As variáveis X3 e X5 não mostram uma relação estatisticamente significativa com a variável dependente dentro do nível de confiança de 5%.

Intervalos de confiança para os coeficientes de regressão

Os intervalos de confiança para os coeficientes de regressão foram calculados com um nível de significância de 1% ($\alpha = 0.01$). Estes intervalos fornecem uma faixa dentro da qual se espera que os verdadeiros valores dos coeficientes de regressão se situem.

Para este efeito pode-se observar na figura seguinte os resultados obtidos:

	2.5 %	97.5 %
(Intercept)	-1.53258900	0.7645492
X1	0.20829211	0.5615387
X3	-0.17993257	0.3396758
X4	0.06299898	0.5515715
X5	-0.21901602	0.2490869
X6	-0.05312356	0.5993853

Figura 37 – Intervalos de Confiança

Conclui-se, então, que os intervalos de confiança para os coeficientes estimados são:

- **Intercepto (β_0)**
Intervalo de confiança: [-1.53258900; 0.7645492]
- **X1**
Intervalo de confiança: [0.20829211; 0.5615387]
- **X3**
Intervalo de confiança: [-0.17993257; 0.3396758]
- **X4**
Intervalo de confiança: [0.06299898; 0.5515715]
- **X5**
Intervalo de confiança: [-0.21901602; 0.2490869]
- **X6**
Intervalo de confiança: [-0.05312356; 0.5993853]

Os intervalos de confiança mostram que ao nível de 95% de confiança, consegue-se esperar que os verdadeiros valores dos coeficientes estejam dentro dessas faixas.

Modelos de regressão linear

Critério de Informação de Akaike (AIC): O AIC é usado para comparar modelos, contudo é uma medida de ajustamento que penaliza o modelo por ter demasiadas variáveis. Sabe-se que se o AIC aumenta, o ajustamento piora. Se o AIC diminui, o ajustamento melhora. A regressão stepwise pode ser feita e foi realizada para três direções: forward, backward, both.

Método Forward (Figura 38)

```
Start:  AIC=-96.89
X2 ~ 1

      Df Sum of Sq  RSS    AIC
+ X6   1   2.7223 0.6539 -160.549
+ X1   1   2.5056 0.8705 -149.101
+ X3   1   2.4681 0.9080 -147.414
+ X4   1   2.3436 1.0326 -142.272
+ X5   1   1.8415 1.5347 -126.422
<none>      3.3762 -96.886

Step:  AIC=-160.55
X2 ~ X6

      Df Sum of Sq  RSS    AIC
+ X1   1  0.215421 0.43845 -174.53
+ X4   1  0.073551 0.58032 -163.32
+ X3   1  0.044567 0.60930 -161.37
<none>      0.65387 -160.55
+ X5   1  0.001608 0.65226 -158.65

Step:  AIC=-174.54
X2 ~ X6 + X1

      Df Sum of Sq  RSS    AIC
+ X4   1  0.081714 0.35674 -180.78
<none>      0.43845 -174.53
+ X3   1  0.016922 0.42153 -174.11
+ X5   1  0.004773 0.43368 -172.97

Step:  AIC=-180.79
X2 ~ X6 + X1 + X4

      Df Sum of Sq  RSS    AIC
<none>      0.35674 -180.78
+ X3   1  0.0045087 0.35223 -179.29
+ X5   1  0.0006434 0.35609 -178.86
```

Método Backward (Figura 39)

```

Start:  AIC=-177.31
X2 ~ X1 + X3 + X4 + X5 + X6

      Df Sum of Sq    RSS    AIC
- X5   1  0.000176  0.35223 -179.29
- X3   1  0.004042  0.35609 -178.86
<none>          0.35205 -177.31
- X6   1  0.029971  0.38202 -176.05
- X4   1  0.067665  0.41972 -172.28
- X1   1  0.203102  0.55515 -161.10

Step:  AIC=-179.29
X2 ~ X1 + X3 + X4 + X6

      Df Sum of Sq    RSS    AIC
- X3   1  0.004509  0.35674 -180.78
<none>          0.35223 -179.29
- X6   1  0.035582  0.38781 -177.44
- X4   1  0.069301  0.42153 -174.11
- X1   1  0.203756  0.55598 -163.04

Step:  AIC=-180.79
X2 ~ X1 + X4 + X6

      Df Sum of Sq    RSS    AIC
<none>          0.35674 -180.78
- X6   1  0.073164  0.42990 -175.32
- X4   1  0.081714  0.43845 -174.53
- X1   1  0.223584  0.58032 -163.32

```

Método Both (Figura 40)

```

Start:  AIC=-177.31
X2 ~ X1 + X3 + X4 + X5 + X6

      Df Sum of Sq    RSS    AIC
- X5   1  0.000176  0.35223 -179.29
- X3   1  0.004042  0.35609 -178.86
<none>          0.35205 -177.31
- X6   1  0.029971  0.38202 -176.05
- X4   1  0.067665  0.41972 -172.28
- X1   1  0.203102  0.55515 -161.10

Step:  AIC=-179.29
X2 ~ X1 + X3 + X4 + X6

      Df Sum of Sq    RSS    AIC
- X3   1  0.004509  0.35674 -180.78
<none>          0.35223 -179.29
- X6   1  0.035582  0.38781 -177.44
+ X5   1  0.000176  0.35205 -177.31
- X4   1  0.069301  0.42153 -174.11
- X1   1  0.203756  0.55598 -163.04

Step:  AIC=-180.79
X2 ~ X1 + X4 + X6

      Df Sum of Sq    RSS    AIC
<none>          0.35674 -180.78
+ X3   1  0.004509  0.35223 -179.29
+ X5   1  0.000643  0.35609 -178.86
- X6   1  0.073164  0.42990 -175.32
- X4   1  0.081714  0.43845 -174.53
- X1   1  0.223584  0.58032 -163.32

```

Seleção do 'melhor' método

Com base nos métodos de seleção de variáveis Forward, Backward e Both, avaliou-se o desempenho dos modelos ajustados e a inclusão ou exclusão de variáveis, com foco na melhoria do ajuste do modelo.

- **Modelo Forward**

Selecionou as variáveis: X_1 , X_4 e X_6

- AIC: -180.79
- Este modelo apresentou o menor valor de AIC entre os três métodos, indicando melhor ajuste em relação aos demais.
- As variáveis escolhidas são coerentes com os resultados de significância observados nos testes dos coeficientes.

- **Modelo Backward**

Selecionou as variáveis: X_1 , X_4 e X_6

- AIC: -180.79
- Coincide com o modelo Forward, o que sugere estabilidade na seleção de variáveis relevantes.
- Reforça que essas três variáveis contribuem de forma significativa para a explicação da variável dependente.

- **Modelo Both**

Selecionou as variáveis: X_1 , X_4 e X_6

- AIC: -180.79
- O resultado idêntico aos modelos Forward e Backward demonstra robustez no processo de seleção.
- Indica que não há ganho expressivo em complexidade ou performance ao permitir inclusão e exclusão simultânea.

Todos os três métodos identificaram o mesmo subconjunto ótimo de variáveis: X_1 (Aroma), X_4 (Acidez) e X_6 (Equilíbrio). Isso evidencia que essas variáveis possuem o maior poder explicativo sobre a variável dependente X_2 , considerando o critério AIC.

Consegue-se, então, chegar à conclusão de que a equação será a seguinte:

$$\hat{y}_i = -0.38402 + 0.38492 X_{i1} - 0.30729 X_{i4} - 0.27313 X_{i6}$$