UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

FRANCISCO LUAN RODRIGUES DE SOUSA

ANÁLISE DE COMPONENTES PRINCIPAIS E AGRUPAMENTO HIERÁRQUICO PARA A IDENTIFICAÇÃO DE TIPOS DE VIDRO

FORTALEZA
2025

FRANCISCO LUAN RODRIGUES DE SOUSA

RELATÓRIO DE ANÁLISE MULTIVARIADA

Relatório apresentado ao curso de Bacharelado em Estatística do Centro de Ciências da Universidade Federal do Ceará, como parte dos requisitos para a aprovação na disciplina de Análise Multivariada no semestre de 2024.2.
Prof.: Gualberto Segundo Agamez Montalvo.

FORTALEZA
2025

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A análise multivariada desempenha um papel fundamental na compreensão e interpretação de conjuntos de dados complexos, permitindo a identificação de padrões, redução de dimensionalidade e agrupamento de informações. Este relatório concentra-se na aplicação de duas técnicas principais da análise multivariada — Análise de Componentes Principais (ACP) e Análise de Agrupamento (*Clustering*) — no conjunto de dados *Glass Identification*, disponível no repositório UCI Machine Learning.

O *Glass Identification* é um conjunto de dados amplamente utilizado em estudos estatísticos e de aprendizado de máquina devido à sua relevância para a ciência forense. Ele contém informações sobre a composição química de diferentes tipos de vidro, com o objetivo de classificar as amostras em categorias como vidro de janelas, utensílios ou recipientes. Este tipo de análise é especialmente relevante para investigações criminais, onde fragmentos de vidro encontrados em cenas de crime podem fornecer pistas sobre a origem do material ou sua associação com suspeitos e locais. Segundo estudos forenses, a composição química do vidro é uma característica única que pode ser usada como uma "impressão digital" para determinar sua procedência, auxiliando no processo de identificação e correlação de evidências (Koons et al., 2002).

A Análise de Componentes Principais (ACP) é uma técnica estatística de redução de dimensionalidade que transforma variáveis correlacionadas em componentes principais ortogonais, preservando o máximo de variância possível. Segundo Jolliffe (2002), a ACP é amplamente utilizada em contextos exploratórios, permitindo identificar combinações de variáveis que melhor explicam as diferenças entre grupos. No contexto do *Glass Identification*, a ACP pode ser aplicada para simplificar a análise da composição química do vidro, ajudando a entender quais elementos são mais relevantes para a diferenciação entre os tipos de amostras.

A Análise de Agrupamento (*Clustering*), por sua vez, busca identificar padrões e subgrupos naturais nos dados com base na similaridade de seus atributos. Técnicas como o *K-means*, descrito por Hartigan e Wong (1979), são amplamente empregadas para dividir os dados em k clusters, cada um representando um subconjunto de amostras semelhantes. No âmbito da ciência forense, métodos de agrupamento podem ser usados para inferir categorias de vidro com base em características químicas, mesmo na ausência de rótulos previamente definidos, auxiliando em investigações exploratórias e validação de hipóteses.

O objetivo deste trabalho é explorar o dataset *Glass Identification* por meio dessas duas técnicas multivariadas. A ACP será empregada para reduzir a dimensionalidade do conjunto de dados e facilitar a visualização dos padrões, enquanto a Análise de Agrupamento será utilizada para identificar potenciais grupos de vidro relacionados a suas categorias. A aplicação dessas técnicas não apenas demonstra a sua utilidade em contextos analíticos, mas também evidencia como ferramentas estatísticas podem ser aplicadas em áreas práticas, como a ciência forense, para ajudar na interpretação de dados complexos e na tomada de decisões informadas.

O relatório está organizado da seguinte forma: na seção de Metodologia, são

apresentadas as etapas e técnicas utilizadas; na seção de Aplicação, detalhamos como cada método foi implementado; em Resultados e Discussão, analisamos os achados principais.

2 Metodologia

Nesta seção, são apresentados os procedimentos e etapas necessários para a análise multivariada do conjunto de dados *Glass Identification*. Inicialmente, descreve-se o conjunto de dados utilizado, destacando suas características principais e os atributos considerados na análise. Em seguida, detalham-se as etapas de pré-processamento, que incluem a preparação dos dados para garantir sua adequação às técnicas multivariadas empregadas, como normalização e verificação de valores ausentes.

Além disso, as técnicas de análise multivariada utilizadas neste trabalho — Análise de Componentes Principais (ACP) e Análise de Agrupamento (*Clustering*) — são apresentadas em detalhe, com ênfase em seus fundamentos teóricos e desenvolvimento matemático. A ACP, baseada na decomposição em valores singulares (*Singular Value Decomposition* - SVD), será discutida como uma ferramenta de redução de dimensionalidade e análise de variância. A Análise de Agrupamento, por sua vez, será abordada sob a ótica de cluster hierárquico, com explicações sobre o cálculo de centroides e medidas de similaridade. O objetivo desta seção é fornecer uma visão clara e detalhada sobre as etapas metodológicas empregadas, assegurando a reprodutibilidade do trabalho e destacando os aspectos teóricos que fundamentam as análises realizadas. Essa abordagem permite compreender não apenas os resultados obtidos, mas também os processos que os embasaram.

2.1 Descrição do Conjunto de Dados

O conjunto de dados *Glass Identification*, disponibilizado pelo repositório UCI Machine Learning, é amplamente utilizado para problemas de classificação em estudos de ciência de dados e estatística. Ele foi originalmente projetado para auxiliar na identificação de tipos de vidro com base em sua composição química, sendo especialmente relevante na área de ciência forense, onde fragmentos de vidro podem ser usados como evidências em investigações criminais.

O dataset contém 214 observações, cada uma representando uma amostra de vidro. Cada amostra é descrita por 9 atributos contínuos que medem a concentração de diferentes óxidos em percentual de peso, além de uma variável categórica que indica o tipo de vidro. Os tipos de vidro são classificados em seis categorias distintas, como vidro de janelas, recipientes e utensílios, e estão associados a diferentes usos e características químicas.

Atributos do Conjunto de Dados

1. RI (Índice de Refração): Medida do índice de refração do vidro.
2. Na (Sódio): Concentração de sódio (% em peso).

3. Mg (Magnésio): Concentração de magnésio (% em peso).
4. Al (Alumínio): Concentração de alumínio (% em peso).
5. Si (Silício): Concentração de silício (% em peso).
6. K (Potássio): Concentração de potássio (% em peso).
7. Ca (Cálcio): Concentração de cálcio (% em peso).
8. Ba (Bário): Concentração de bário (% em peso).
9. Fe (Ferro): Concentração de ferro (% em peso).
10. Tipo de Vidro: Classe da amostra, com valores de 1 a 7, representando categorias como:
    (i) Janelas de edifícios processadas (vidro float)
    (ii) Janelas de edifícios processadas (vidro não float)
    (iii) Janelas de veículos processadas (vidro float)
    (iv) Janelas de veículos processadas (vidro não float)
    (v) Recipientes
    (vi) Utensílios de mesa
    (vii) Faróis

Exemplo de Tabela do Conjunto de Dados

A Tabela 1 apresenta um exemplo representativo das primeiras linhas do conjunto de dados *Glass Identification*:

Tabela 1: Exemplo das primeiras linhas do conjunto de dados *Glass Identification*.

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Tipo de Vidro
1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.00	0.00	1
1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.00	0.00	1
1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.00	0.00	1
1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.00	0.00	1
1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.00	0.00	1

2.2 Pré-processamento dos Dados

O pré-processamento é uma etapa fundamental para garantir que os dados estejam adequados para as análises subsequentes e para as técnicas multivariadas que serão aplicados. A escolha dos métodos de tratamento deve ser feita com base nas características específicas do conjunto de dados e na análise exploratória dos mesmos. No caso do conjunto de dados `glass`, que contém informações sobre a composição química de diferentes tipos de vidro, realizamos uma série de transformações e tratamentos com o objetivo de preparar os dados para análises mais profundas. Abaixo,

detalho os passos de pré-processamento realizados neste estudo, cujos detalhes completos estão apresentados na seção de *Análise Exploratória dos Dados*.

O dataset `glass` contém 214 instâncias, com 9 variáveis que representam as características químicas de diferentes tipos de vidro. As variáveis incluem a concentração percentual de elementos como silício (Si), sódio (Na), magnésio (Mg), alumínio (Al), silício (Si), potássio (K), cálcio (Ca), bário (Ba), ferro (Fe), e a variável alvo Tipo de vidro (uma variável categórica que indica o tipo de vidro, com valores que variam de 1 a 7). O objetivo deste pré-processamento é garantir que o conjunto de dados esteja pronto para a análise de correlação entre as variáveis e a identificação de padrões de composição entre os tipos de vidro.

Durante a análise exploratória, observamos que os elementos bário (Ba) e ferro (Fe) são raros no conjunto de dados, com muitos valores iguais a zero. Em vez de tratar essas variáveis como contínuas, foi adotada uma abordagem de binarização. Nesse processo:

(i) Para Ba e Fe, atribuímos o valor 1 quando a variável é maior que 0, indicando que o elemento está presente na amostra.

(ii) Caso contrário, o valor foi atribuído como 0, indicando a ausência desses elementos.

Essa transformação resultou em variáveis categóricas binárias para Ba e Fe, o que facilita a análise dos dados.

Um passo essencial no pré-processamento foi a verificação de valores ausentes. A presença de dados faltantes pode comprometer a integridade da análise, e, portanto, todos os registros com valores ausentes foram tratados. Dependendo da quantidade e da importância das variáveis afetadas, as instâncias com dados ausentes foram:

1. Removidas, caso a falta de dados fosse significativa e não fosse possível realizar uma imputação precisa.
2. Imputadas, no caso de valores ausentes em variáveis menos importantes, utilizando técnicas como a imputação pela média ou pela mediana, conforme adequado.

Além disso, também verificamos se existiam valores duplicados no conjunto de dados. A presença de registros duplicados pode introduzir viés nas aplicações, levando a uma super-representação de certas instâncias. Para garantir a integridade dos dados, todas as instâncias duplicadas foram removidas.

A detecção de outliers foi realizada utilizando o método IQR (Interquartile Range), uma técnica estatística robusta para identificar valores que estão fora dos limites de uma distribuição normal. O método baseia-se nos quartis, que dividem os dados em quatro partes iguais. Para calcular os limites para os outliers:

(i) Limite inferior foi definido como $Q1 - 1.5 \times$ IQR

(ii) Limite superior foi definido como $Q3 + 1.5 \times$ IQR

Onde $Q1$ é o primeiro quartil (25%) e $Q3$ é o terceiro quartil (75%) dos dados, e IQR é a diferença entre $Q3$ e $Q1$. Os valores fora desses limites foram considerados outliers.

Embora o tratamento de outliers seja importante para evitar que valores extremos influenciem os resultados dos modelos, é importante destacar que a remoção de outliers pode resultar em perda de informações valiosas. Em nosso caso, as instâncias de outliers podem estar associadas a tipos de vidro com composição química diferenciada, o que poderia conter informações relevantes para distinguir os diferentes tipos de vidro. Portanto, a remoção dos outliers foi feita com cautela, levando em consideração que esses dados poderiam ser indicativos de padrões importantes no comportamento dos tipos de vidro.

Outro ponto relevante a ser destacado é que, embora tenhamos realizado uma análise exploratória detalhada do conjunto de dados e investigado a influência da composição dos elementos químicos nos tipos de vidro, não sou especialista no domínio relacionado aos materiais. Isso implica que o processo de remoção de outliers pode ser impactado pela minha falta de conhecimento aprofundado sobre a química dos vidros. A remoção de outliers, que pode parecer adequada em um primeiro momento, pode, na realidade, afetar a representatividade das técnicas, especialmente se os outliers contiverem informações valiosas sobre os tipos específicos de vidro.

O pré-processamento dos dados é uma etapa essencial para preparar o conjunto de dados para análises mais profundas e a construção das técnicas multivariadas. No caso do conjunto de dados `glass`, as principais etapas de pré-processamento incluíram a binarização das variáveis Ba e Fe, o tratamento de valores ausentes e duplicados, e a detecção e tratamento de outliers. Essas etapas foram realizadas com base nas observações feitas durante a análise exploratória, cujos detalhes completos serão apresentados na seção de Análise Exploratória dos Dados.

O tratamento de outliers, em particular, foi feito com cautela, considerando a possibilidade de perda de informações valiosas sobre os tipos de vidro. Além disso, a falta de conhecimento especializado no domínio dos materiais de vidro significa que algumas escolhas no pré-processamento podem ser revisadas conforme novos conhecimentos sobre o campo sejam adquiridos. Em última análise, o pré-processamento visa garantir que os dados sejam adequados para as análises subsequentes e para a criação de modelos que possam gerar insights significativos sobre os diferentes tipos de vidro.

2.3 Técnicas Utilizadas

2.3.1 Análise de Componentes Principais (ACP)

A Análise de Componentes Principais (ACP), ou *Principal Component Analysis* (PCA), é uma técnica estatística que transforma um conjunto de p variáveis correlacionadas em um conjunto de k variáveis não correlacionadas, com $k < p$, de forma a preservar a maior parte da informação presente nos dados originais. Esta abordagem foi proposta por Karl Pearson em 1901 e, mais tarde, foi formalizada e nomeada por Harold Hotelling em 1933. A ACP é amplamente utilizada para redução de dimensionalidade, permitindo

representar os dados em um espaço de características mais simples e interpretável, mantendo a maior parte da variabilidade dos dados originais.

No caso em que as variáveis originais seguem uma distribuição normal multivariada, as componentes principais resultantes também terão distribuição normal multivariada e serão independentes. Além disso, a primeira componente principal é aquela que captura a maior variação nos dados, enquanto as componentes subsequentes capturam as variações restantes de forma ortogonal às anteriores.

Esse processo é de grande utilidade em diversas áreas, como reconhecimento de padrões, compressão de dados e visualização de dados multidimensionais.

Objetivos principais da Análise de Componentes Principais (ACP):

(i) Redução da dimensionalidade dos dados: A ACP permite reduzir o número de variáveis no conjunto de dados, mantendo a maior parte da variabilidade presente nas variáveis originais.

(ii) Geração de combinações interpretáveis das variáveis originais: As componentes principais são combinações lineares das variáveis originais, facilitando a interpretação e compreensão das relações entre elas.

(iii) Análise da estrutura de correlação entre as variáveis: A ACP ajuda a identificar a correlação entre as variáveis originais, possibilitando uma melhor compreensão da estrutura subjacente aos dados.

O processo de ACP começa com a obtenção das componentes principais exatas, que são derivadas diretamente da matriz de variâncias e covariâncias populacionais, Σ. A partir dessa matriz, as direções que maximizam a variabilidade nos dados podem ser identificadas. No entanto, quando a matriz Σ não está disponível, é possível utilizar as componentes principais estimadas, que são calculadas a partir da matriz de variâncias e covariâncias amostrais, S. Esse procedimento é particularmente útil em situações práticas onde os parâmetros populacionais não são conhecidos.

Além disso, a ACP também é valiosa para:

(i) Eliminação de redundâncias nos dados: Ao identificar as variáveis que carregam as maiores variabilidades, a ACP ajuda a eliminar redundâncias, o que é importante para simplificar modelos de machine learning e melhorar o desempenho computacional.

(ii) Visualização de dados complexos: A redução de dimensionalidade facilita a visualização de dados complexos, permitindo uma interpretação mais intuitiva das relações entre as variáveis.

Seja $\mathbf{X}$ um vetor aleatório de dimensão $p \times 1$, com vetor de médias (populacionais) $\boldsymbol{\mu}$ de dimensão $p \times 1$ e matriz de variâncias e covariâncias populacionais $\boldsymbol{\Sigma}$ de dimensão $p \times p$.

Estamos particularmente interessados no caso em que as variáveis $X_1, X_2, \ldots, X_p$ estão correlacionadas. Isso significa que algumas (ou muitas) das covariâncias Cov(X_i, X_j), para $i, j = 1, 2, \ldots, p$ e $i \neq j$, são não-nulas. Quando isso ocorre, existe

redundância entre as dimensões, o que pode ser um indicativo de que a informação das variáveis originais pode ser comprimida de forma mais eficiente.

Nesse contexto, é comum procurar reduzir a dimensionalidade do problema, criando novas variáveis que sejam combinações lineares das variáveis originais, mas que não estejam mais correlacionadas entre si. Essas novas variáveis são conhecidas como *componentes principais*.

A redução de dimensionalidade pode ser significativa, pois poucas (k) novas variáveis podem explicar uma grande parte da variabilidade presente nas p variáveis originais. Essa redução tem implicações práticas, como a diminuição de custos em termos de tempo computacional e espaço de armazenamento, além de permitir uma análise mais simples e eficiente dos dados.

Componentes Principais

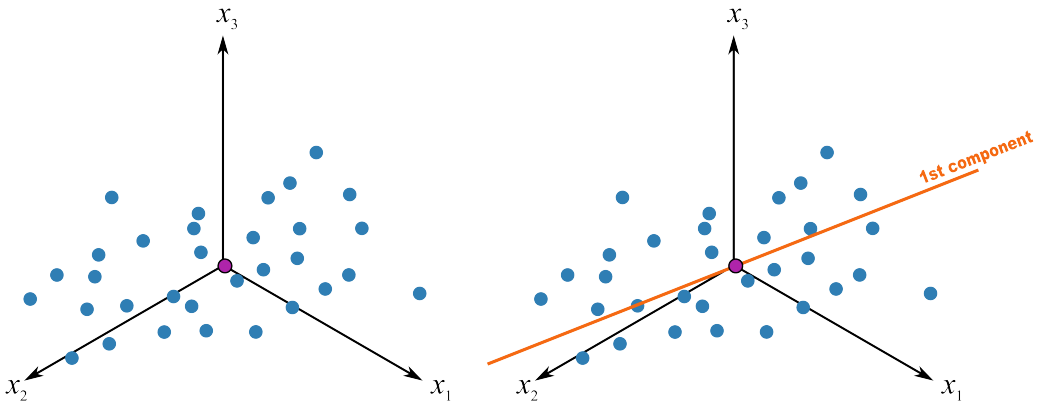
Seja $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sejam $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ os autovalores de $\boldsymbol{\Sigma}$, com os autovetores correspondentes $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$, tais que:

1. $\mathbf{e}_i^\top \mathbf{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\mathbf{e}_i^\top \mathbf{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p\times p} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p)$. Então, o vetor de componentes principais de $\boldsymbol{\Sigma}$ é dado por:

$$\mathbf{Y}_{p\times 1} = O^\top \mathbf{X}.$$

Figura 1: *Interpretação geométrica* com $p = 3$.



Propriedades:

(i) A j-ésima componente principal de $\boldsymbol{\Sigma}$ é dada por:

$$Y_j = \mathbf{e}_j^\top \mathbf{X}.$$

(ii) A esperança de Y_j é:

$$\mathbb{E}(Y_j) = \mathbf{e}_j^\top \boldsymbol{\mu}.$$

(iii) A variância de Y_j é:

$$\text{Var}(Y_j) = \mathbf{e}_j^\top \boldsymbol{\Sigma} \mathbf{e}_j = \lambda_j.$$

(iv) A covariância entre Y_i e Y_j é:

$$\text{Cov}(Y_i, Y_j) = \text{Cor}(Y_i, Y_j) = 0, \quad \text{para } i, j = 1, \ldots, p \text{ e } i \neq j.$$

(v) A proporção da variância total de $\mathbf{X}$ explicada pela j-ésima componente principal é:

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}.$$

Estimativa das Componentes Principais

Como, em geral, a matriz $\boldsymbol{\Sigma}$ é desconhecida, utiliza-se a matriz $\mathbf{S}$, de variâncias e covariâncias amostrais, para estimar as componentes principais. Considere $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ os autovalores de $\mathbf{S}$, com os autovetores correspondentes padronizados $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p$.

A j-ésima componente principal amostral é dada por:

$$\hat{Y}_j = \hat{\mathbf{e}}_j^\top \mathbf{X}.$$

Propriedades:

(i) (1) A variância de $\hat{Y}_j$ é:

$$\text{Var}(\hat{Y}_j) = \hat{\lambda}_j.$$

(ii) A covariância entre $\hat{Y}_i$ e $\hat{Y}_j$ é:

$$\text{Cov}(\hat{Y}_i, \hat{Y}_j) = \text{Cor}(\hat{Y}_i, \hat{Y}_j) = 0, \quad \text{para } i, j = 1, \ldots, p \text{ e } i \neq j.$$

(iii) A proporção da variância total explicada pela j-ésima componente principal amostral é:

$$\frac{\hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i}.$$

(iv) A correlação entre $\hat{Y}_j$ e X_i é:

$$\text{Cor}(\hat{Y}_j, X_i) = \hat{\mathbf{e}}_{ji} \sqrt{\frac{\hat{\lambda}_j}{s_{jj}}}.$$

Pelo Teorema da Decomposição Espectral, temos:

$$\mathbf{S}_{p \times p} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j^\top,$$

que pode ser aproximada por:

$$\mathbf{S}_{p \times p} \approx \sum_{j=1}^{k} \hat{\lambda}_j \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j^\top.$$

Análise de Componentes Principais via Matriz de Correlação

Seja $\mathbf{X} \in \mathbb{R}^{p\times 1} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\mathbf{X} = (X_1, \ldots, X_p)^\top$. Seja $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$ tal que

$$Z_i = \frac{X_i - \mu_i}{\sigma_i},$$

em que $\mu_i = \mathbb{E}(X_i)$ e $\sigma_i^2 = \text{Var}(X_i)$.

Temos que $\text{Cov}(\mathbf{Z}) = \mathbf{P} = \text{Cor}(\mathbf{X})$. Sejam $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ os autovalores de $\mathbf{P}$, com autovetores correspondentes $\mathbf{e}_1, \ldots, \mathbf{e}_p$, tais que:

(i) ($\mathbf{e}_i^\top \mathbf{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,

(ii) $\mathbf{e}_i^\top \mathbf{e}_i = 1$, para $i = 1, \ldots, p$,

(iii) $\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $\mathbf{O}_{p\times p} = (\mathbf{e}_1, \ldots, \mathbf{e}_p)$. Então, o vetor de componentes principais de $\mathbf{P}$ é dado por

$$\mathbf{Y} = \mathbf{O}^\top \mathbf{Z}.$$

Propriedades:

(i) A j-ésima componente principal de $\mathbf{P}$ é dada por:

$$Y_j = \mathbf{e}_j^\top \mathbf{Z}.$$

(ii) A esperança de Y_j é:

$$\mathbb{E}(Y_j) = 0.$$

(iii) A variância de Y_j é:

$$\text{Var}(Y_j) = \mathbf{e}_j^\top \mathbf{P} \mathbf{e}_j = \lambda_j.$$

(iv) A covariância entre Y_i e Y_j é:

$$\text{Cov}(Y_i, Y_j) = \text{Cor}(Y_i, Y_j) = 0, \quad \text{para } i, j = 1, \ldots, p \text{ e } i \neq j.$$

(v) A proporção da variância total de $\mathbf{Z}$ explicada pela j-ésima componente principal é:

$$\frac{\lambda_j}{p}.$$

Estimação das Componentes Principais

Como, em geral, a matriz $\mathbf{P}$ é desconhecida, utiliza-se a matriz $\mathbf{R}$, de correlações amostrais, para estimar as componentes principais.

Considere $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ os autovalores de $\mathbf{R}$, com autovetores correspondentes padronizados $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_p$.

A j-ésima componente principal amostral é dada por:

$$\hat{Y}_j = \hat{\mathbf{e}}_j^\top \mathbf{Z}.$$

Propriedades:

(i) A variância de $\hat{Y}_j$ é:

$$\text{Var}(\hat{Y}_j) = \hat{\lambda}_j.$$

(ii) A covariância entre $\hat{Y}_i$ e $\hat{Y}_j$ é:

$$\text{Cov}(\hat{Y}_i, \hat{Y}_j) = \text{Cor}(\hat{Y}_i, \hat{Y}_j) = 0, \quad \text{para } i, j = 1, \ldots, p \text{ e } i \neq j.$$

(iii) A proporção da variância total explicada pela j-ésima componente principal amostral é:

$$\frac{\hat{\lambda}_j}{p}.$$

(iv) A correlação entre $\hat{Y}_j$ e Z_i é:

$$\text{Cor}(\hat{Y}_j, Z_i) = \hat{e}_{ji}\sqrt{\frac{\hat{\lambda}_j}{p}}.$$

Pelo teorema da decomposição espectral, a matriz $\mathbf{R}_{p\times p}$ pode ser escrita como:

$$\mathbf{R}_{p\times p} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j^\top,$$

e pode ser aproximada por:

$$\mathbf{R}_{p\times p} \approx \sum_{j=1}^{k} \hat{\lambda}_j \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j^\top.$$

Determinação do Número de Componentes Principais

Alguns métodos comumente utilizados para determinar o número de componentes principais são:

1. Proporção da variância total explicada: Seleciona-se o número de componentes principais de modo que a soma acumulada das variâncias explicadas seja maior ou igual a um limiar pré-definido (por exemplo, 70% ou 90%).

2. Análise gráfica da variância explicada (scree-plot): Consiste em analisar o gráfico dos autovalores em ordem decrescente, identificando o "cotovelo" no gráfico, onde a contribuição marginal de cada componente principal se estabiliza.

3. Aproximação da matriz $\mathbf{S}$ (ou $\mathbf{R}$): Avalia-se a qualidade da aproximação de $\mathbf{S}$ (ou $\mathbf{R}$) utilizando um número reduzido de componentes principais.

4. Análise prática das componentes principais: Verifica-se a utilidade prática das componentes principais em relação ao problema específico, como a interpretabilidade ou o impacto em modelos subsequentes.

Inferência Assintótica sobre as Componentes Principais

Seja $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\lambda_1 \geq \cdots \geq \lambda_p > 0$ os autovalores de $\boldsymbol{\Sigma}$, com autovetores correspondentes $\mathbf{e}_1, \ldots, \mathbf{e}_p$, e $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p > 0$ os autovalores de $\mathbf{S}$, com autovetores correspondentes $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_p$. Sejam $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ e $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)^\top$. Os seguintes resultados se aplicam:

1. Se $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$, então, para n suficientemente grande, temos:

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \approx N_p(\mathbf{0}, 2\boldsymbol{\Lambda}^2).$$

2. Se $\mathbf{E}_i = \lambda_i \sum_{k=1,k\neq i}^{p} \frac{\lambda_k}{(\lambda_k-\lambda_i)^2}\mathbf{e}_k\mathbf{e}_k^\top$, então:

$$\sqrt{n}(\hat{\mathbf{e}}_i - \mathbf{e}_i) \approx N_p(\mathbf{0}, \mathbf{E}_i).$$

3. A distribuição de cada $\hat{\lambda}_i$ não depende dos elementos de $\hat{\mathbf{e}}_i$ correspondentes. Segue-se que:

$$\hat{\lambda}_i \overset{\text{ind}}{\sim} N\left(\lambda_i, \frac{2\lambda_i^2}{n}\right).$$

Assim, a probabilidade de $|\hat{\lambda}_i - \lambda_i| \leq z_{\alpha/2}\lambda_i\sqrt{\frac{2}{n}}$ é dada por:

$$P\left(|\hat{\lambda}_i - \lambda_i| \leq z_{\alpha/2}\lambda_i\sqrt{\frac{2}{n}}\right) = 1 - \alpha.$$

Podemos construir um intervalo de confiança com $100(1-\alpha)\%$ para λ_i usando as propriedades da distribuição normal:

$$IC_{100(1-\alpha)\%}(\lambda_i) = \left(\frac{\hat{\lambda}_i}{1 + z_{\alpha/2}\sqrt{\frac{2}{n}}}, \frac{\hat{\lambda}_i}{1 - z_{\alpha/2}\sqrt{\frac{2}{n}}}\right),$$

que também pode ser usado para fazer inferência sobre λ_i, por exemplo, avaliar:

$$H_0 : \lambda_i = \lambda_0 \quad \text{contra} \quad H_1 : \lambda_i \neq \lambda_0.$$

2.3.2 Análise de Agrupamento (*Clustering*)

A análise de agrupamento trata de buscar padrões em um conjunto de dados, agrupando observações multivariadas em clusters para encontrar uma forma de agrupamento ótima. Nesse agrupamento, as observações ou objetos dentro de cada cluster são similares entre si, mas os clusters são distintos uns dos outros. O objetivo da análise de agrupamento é identificar as "agrupamentos naturais" no conjunto de dados que façam sentido para o pesquisador.

A análise de agrupamento também é referida como classificação, reconhecimento de padrões (especificamente, aprendizado não supervisionado) e taxonomia numérica. Classificação e agrupamento são frequentemente usados de forma intercambiável na literatura, mas aqui desejamos diferenciá-los. Na classificação, atribuímos novas observações a um de vários grupos, cujo número já é predefinido. Na análise de agrupamento, nem o número de grupos nem os próprios grupos são conhecidos previamente.

Os dados básicos para a maioria das aplicações de análise de agrupamento consistem na matriz de dados multivariados usual de dimensão $n \times p$, contendo os valores das variáveis que descrevem cada objeto a ser agrupado. A matriz de dados pode ser definida como:

$$Y = \begin{pmatrix} y_1^\top \\ y_2^\top \\ \vdots \\ y_n^\top \end{pmatrix} = \big(y(1), \quad y(2), \quad \ldots, \quad y(p)\big) \,. \tag{1}$$

Onde $y_i^\top$ é um vetor linha de observação e $y(j)$ é uma coluna correspondente a uma variável. Geralmente, desejamos agrupar as n linhas $y_i^\top$ em g clusters, mas também é possível agrupar as colunas $y(j)$, com $j = 1, 2, \ldots, p$.

Muitas técnicas empregadas na análise de agrupamento começam com a avaliação das similaridades entre todos os pares de observações. Isso significa que os pontos de dados precisam ser medidos de forma semelhante ou, pelo menos, a partir de métricas que permitam calcular similaridades. Na prática, existem restrições de tempo que tornam quase impossível determinar o melhor agrupamento de objetos similares a partir de todas as estruturas possíveis. Assim, devemos nos contentar com algoritmos que busquem agrupamentos bons, mas não necessariamente os melhores.

Medidas de Similaridade

Qualquer tentativa de identificar clusters de observações a partir de um conjunto de dados complexo requer o conhecimento de como os itens estão "próximos" uns dos outros, ou seja, uma medida de "proximidade" ou "similaridade". A natureza das variáveis, como discreta, contínua, binária, escalas de medição, como nominal, ordinal, intervalo, razão, e o conhecimento sobre o assunto têm sido sugeridos como considerações importantes. O agrupamento de itens (unidades ou casos) utiliza a medida de proximidade por algum tipo de distância; o agrupamento de variáveis utiliza a medida de proximidade por coeficientes de correlação ou outras medidas de associação.

A distância Euclidiana (em linha reta) entre duas observações p-dimensionais (itens) $x^\top = (x_1, x_2, \ldots, x_p)$ e $y^\top = (y_1, y_2, \ldots, y_p)$ é dada por:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} = \sqrt{(x - y)^\top (x - y)}. \tag{2}$$

A distância estatística entre as mesmas observações é da forma:

$$d(x, y) = \sqrt{(x - y)^\top A (x - y)}, \tag{3}$$

onde $A = S^{-1}$ e S é a matriz de variâncias e covariâncias da amostra. No entanto, a análise de agrupamento tem como objetivo encontrar os grupos distintos no conjunto de dados, então essas quantidades amostrais não podem ser computadas. Portanto, preferimos a distância Euclidiana para o agrupamento.

A terceira medida de distância é a métrica de Minkowski:

$$d(x, y) = \left(\sum_{i=1}^{p} |x_i - y_i|^m \right)^{1/m} . \tag{4}$$

Outras duas medidas populares de “distância” são a métrica de Canberra e o coeficiente de Czekanowski, ambas definidas para variáveis não negativas. Temos:

Métrica de Canberra:

$$d(x,y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}. \tag{5}$$

Coeficiente de Czekanowski:

$$d(x,y) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}. \tag{6}$$

Distância de Manhattan:

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i|. \tag{7}$$

É sempre recomendável usar “verdadeiras” distâncias para agrupar objetos. No entanto, a maioria dos algoritmos de agrupamento aceita números de distância atribuídos subjetivamente, que podem não satisfazer a desigualdade triangular. Em situações onde os itens não podem ser acessados por medições significativas p-dimensionais, os comparamos com base na presença ou ausência de certas características.

Métodos de Agrupamento Hierárquico

O agrupamento hierárquico é conduzido por uma série de fusões sucessivas ou uma série de divisões sucessivas. Os métodos hierárquicos aglomerativos começam com os objetos individuais; assim, inicialmente há tantos clusters quanto objetos. Eles primeiro agrupam os objetos mais semelhantes, e então esses grupos são mesclados de acordo com suas similaridades. Eventualmente, todos os subgrupos são combinados em um único cluster. Os métodos hierárquicos divisivos inicialmente dividem todos os objetos em dois subgrupos, nos quais os objetos de um subgrupo estão “distantes” dos objetos do outro grupo. Os subgrupos são ainda mais divididos em subgrupos dissimilares até que cada objeto forme um grupo por si só. Os resultados de ambos os métodos podem ser representados em um dendrograma, que ilustra as fusões ou divisões feitas nos níveis sucessivos.

O foco são os procedimentos hierárquicos aglomerativos, e particularmente, os métodos de ligação, que são adequados para agrupar tanto itens quanto variáveis. Por sua vez, a ligação simples (distância mínima ou vizinho mais próximo), a ligação completa (distância máxima ou vizinho mais distante) e a ligação média (distância média).

Métodos de Ligação

(i) Ligação Simples
As distâncias ou similaridades entre pares de objetos podem ser entradas para um algoritmo de ligação simples. Os clusters são unidos a partir das entidades individuais, combinando os vizinhos mais próximos, o que significa a menor distância ou a maior similaridade. Primeiro, devemos identificar a menor distância

em $D = \{d_{ik}\}$ e unir os objetos correspondentes, digamos, U e V, para formar o cluster (UV). Para o algoritmo acima, a distância entre (UV) e qualquer outro cluster W é calculada por:

$$d(UV, W) = \min\{d_{UW}, d_{VW}\}. \tag{8}$$

As quantidades d_{UW} e d_{VW} são as distâncias entre os vizinhos mais próximos dos clusters U e W e entre os clusters V e W, respectivamente.

Em uma aplicação típica de agrupamento hierárquico, os resultados intermediários, onde os objetos são classificados em um número moderado de clusters, são de grande interesse. Como a ligação simples une os clusters pela menor conexão entre eles, o método não consegue discernir clusters mal separados. Além disso, o método é um dos poucos que pode delinear clusters não elipsoidais. A tendência da ligação simples de selecionar clusters longos e semelhantes a cadeias é conhecida como encadeamento, o que pode ser enganoso se os itens nas extremidades opostas da cadeia forem bastante dissimilares.

(ii) Ligação Completa
O agrupamento por ligação completa é conduzido de maneira quase idêntica ao agrupamento por ligação simples, com uma exceção importante: em cada estágio, a distância entre os clusters é determinada pela distância entre os elementos, um de cada cluster, que estão mais distantes. Assim, a ligação completa garante que todos os itens em um cluster estejam dentro de alguma distância máxima (ou similaridade mínima) uns dos outros.

O algoritmo aglomerativo geral começa encontrando a entrada mínima em $D = \{d_{ik}\}$ e unindo os objetos correspondentes, como U e V, para formar o cluster (UV). Para o Passo 3 do algoritmo acima, a distância entre (UV) e qualquer outro cluster W é calculada por:

$$d(UV, W) = \max\{d_{UW}, d_{VW}\}. \tag{09}$$

Aqui, d_{UW} e d_{VW} são as distâncias entre os membros mais distantes dos clusters U e W e entre os clusters V e W, respectivamente.

Semelhante ao método de ligação simples, uma nova atribuição de distâncias (ou similaridades) que tenha as mesmas ordens relativas das distâncias iniciais não alterará a configuração dos clusters de ligação completa.

O agrupamento por ligação completa é conduzido de maneira quase idêntica ao agrupamento por ligação simples, com uma exceção importante: em cada estágio, a distância entre os clusters é determinada pela distância entre os elementos, um de cada cluster, que estão mais distantes. Assim, a ligação completa garante que todos os itens em um cluster estejam dentro de alguma distância máxima (ou similaridade mínima) uns dos outros.

O algoritmo aglomerativo geral começa encontrando a entrada mínima em $D = \{d_{ik}\}$ e unindo os objetos correspondentes, como U e V, para formar o cluster (UV). Para o Passo 3 do algoritmo acima, a distância entre (UV) e qualquer outro cluster W é calculada por:

$$d(UV, W) = \max\{d_{UW}, d_{VW}\}. \tag{10}$$

Aqui, d_{UW} e d_{VW} são as distâncias entre os membros mais distantes dos clusters U e W e entre os clusters V e W, respectivamente.

Semelhante ao método de ligação simples, uma nova atribuição de distâncias (ou similaridades) que tenha as mesmas ordens relativas das distâncias iniciais não alterará a configuração dos clusters de ligação completa.

(iii) Ligação Média
A ligação média trata a distância entre dois clusters como a distância média entre todos os pares de itens, onde um membro de um par pertence a cada cluster. O método começa procurando na matriz de distâncias $D = \{d_{ik}\}$ para encontrar os objetos mais próximos (ou mais semelhantes), por exemplo, U e V. Esses objetos são unidos para formar o cluster (UV). Para o Passo 3 do algoritmo acima, a distância entre (UV) e o outro cluster W é determinada por:

$$d(UV, W) = \frac{\sum_i \sum_k d_{ik}}{N(UV)N(W)}, \tag{12}$$

onde d_{ik} é a distância entre o objeto i no cluster (UV) e o objeto k no cluster W, e $N(UV)$ e $N(W)$ são os números de itens nos clusters (UV) e W, respectivamente.

Para o agrupamento por ligação média, mudanças na atribuição das distâncias (ou similaridades) podem afetar o arranjo da configuração final dos clusters, mas preservam as ordens relativas.

Método de Agrupamento Hierárquico de Ward

J. H. Ward Jr. [**Ward**] propôs um método de agrupamento hierárquico baseado na minimização da *perda de informação* ao unir dois grupos. Este método considera o aumento na soma dos quadrados dos erros (ESS, do inglês *Error Sum of Squares*) como uma perda de informação. Primeiramente, para um dado cluster k, seja ESS_k a soma dos desvios quadrados de cada item no cluster em relação à média (centroide) do cluster. Se atualmente existem K clusters, define-se ESS como a soma dos ESS_k, ou seja:

$$ESS = ESS_1 + ESS_2 + \cdots + ESS_k.$$

Em cada etapa, considera-se a união de todos os pares possíveis de clusters, e os dois clusters cuja combinação resulte no menor aumento de ESS (perda mínima de informação) são unidos. Inicialmente, cada cluster consiste de um único item e, se houver N itens, $ESS_k = 0$, para $k = 1, 2, \ldots, N$, então $ESS = 0$. No outro extremo, quando todos os clusters são combinados em um único grupo de N itens, o valor de ESS é dado por:

$$ESS = \sum_{j=1}^{N} (x_j - \bar{x})^\top (x_j - \bar{x}), \tag{13}$$

onde x_j é a medição multivariada associada ao j-ésimo item e $\bar{x}$ é a média de todos os itens. Os resultados do método de Ward podem ser exibidos como um dendrograma, no qual o eixo vertical representa os valores de ESS onde as uniões ocorrem.

O método de Ward baseia-se na suposição de que os clusters de observações multivariadas têm formas aproximadamente elípticas. Ele é um precursor hierárquico dos métodos não hierárquicos de agrupamento que otimizam algum critério para dividir os dados em um número determinado de grupos elípticos.

Os procedimentos hierárquicos não levam em consideração fontes de erro e variação, o que significa que um método de agrupamento será sensível a outliers ou ”pontos de ruído”. Além disso, não há provisão para realocação de objetos que possam ter sido ”incorretamente”agrupados em uma etapa inicial; portanto, a configuração final dos clusters deve ser cuidadosamente examinada para verificar se faz sentido ou não. Para um problema específico, é uma boa ideia experimentar vários métodos de agrupamento e, dentro de um método dado, diferentes maneiras de atribuir distâncias. Se os resultados de vários métodos forem (aproximadamente) consistentes entre si, talvez se possa argumentar a favor da existência de agrupamentos ”naturais”.

3 Aplicação

3.1 Análise Exploratória de Dados

Para entender melhor as características do conjunto de dados *Glass Identification*, foi realizada uma análise exploratória inicial. O conjunto contém informações sobre diferentes tipos de vidro, categorizados pela variável alvo `Type`. Cada amostra é descrita por propriedades físico-químicas, incluindo:

- Índice de refração (`RI`);
- Concentrações de elementos químicos, como sódio (`Na`), magnésio (`Mg`), alumínio (`Al`), silício (`Si`), potássio (`K`), cálcio (`Ca`), bário (`Ba`) e ferro (`Fe`).

Os dados foram lidos, tratados para garantir o formato adequado, e suas variáveis foram resumidas estatisticamente. A tabela abaixo apresenta as principais medidas descritivas, como mínimo, máximo, mediana, média e quartis, para cada atributo numérico:

Tabela 2: Resumo estatístico das variáveis do conjunto de dados.

Variável	Média	DP	Mediana	Mín	Máx	25th	75th	Skew	Kurtosis
RI	1.518	0.0030	1.518	1.511	1.534	1.517	1.519	1.625	4.819
Na	13.404	0.8167	13.300	10.730	17.380	12.900	13.810	0.459	2.928
Mg	2.679	1.4437	3.480	0.000	4.490	2.090	3.600	-1.130	-0.468
Al	1.449	0.4959	1.360	0.290	3.500	1.190	1.630	0.925	1.992
Si	72.655	0.7740	72.790	69.810	75.410	72.280	73.090	-0.734	2.868
K	0.499	0.6532	0.560	0.000	6.210	0.130	0.610	6.454	52.716
Ca	8.954	1.4259	8.600	5.430	16.190	8.240	9.150	2.022	6.399
Ba	0.176	0.4982	0.000	0.000	3.150	0.000	0.000	3.359	12.004
Fe	0.057	0.0976	0.000	0.000	0.510	0.000	0.100	1.723	2.493

Além disso, o conjunto de dados apresenta uma distribuição desigual entre os tipos de vidro (`Type`), com o tipo 2 sendo o mais frequente e o tipo 6, o menos comum. Foi criado um gráfico de barras para visualizar a distribuição de cada tipo de vidro no conjunto de dados. O gráfico abaixo mostra a quantidade de amostras por tipo:

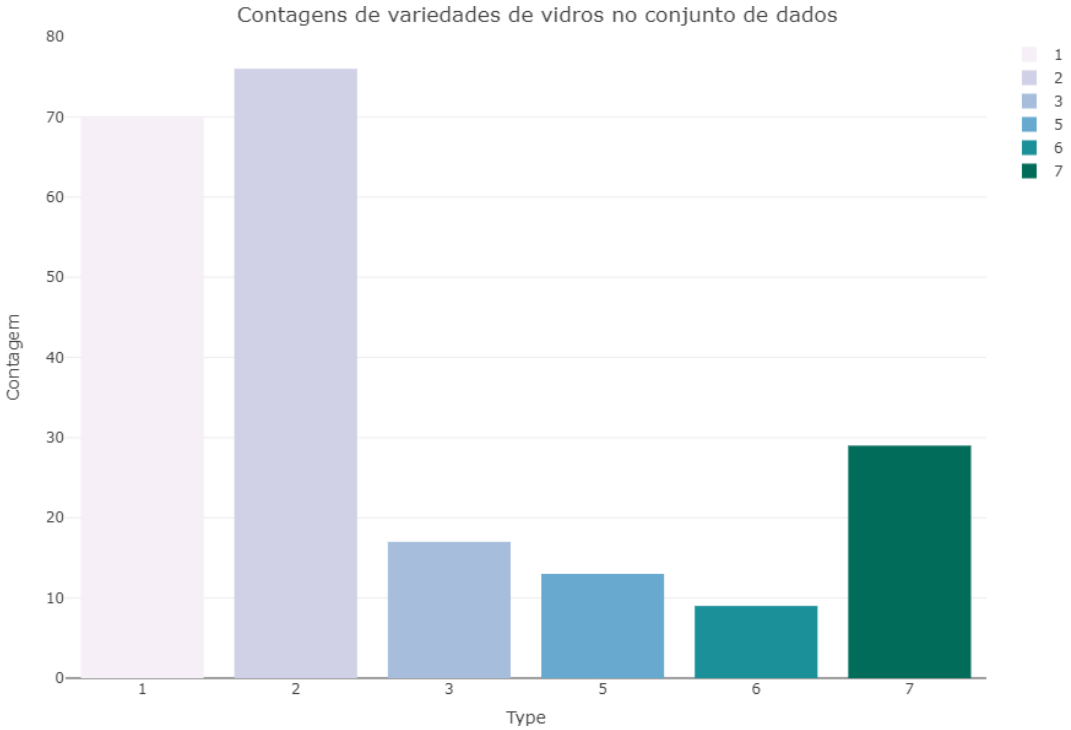


Figura 2: Contagens de variedades de vidros no conjunto de dados.

Tabela 3: Frequência e percentual da variável `Type` no conjunto de dados.

Tipo de Vidro (`Type`)	**Frequência**	**Percentual (%)**
1	70	32.71
2	76	35.51
3	17	7.94
5	13	6.07
6	9	4.21
7	29	13.55
Total	214	100.00

Esse padrão evidencia que os tipos 1 e 2 são os mais prevalentes, representando juntos aproximadamente 68,2% do total de amostras. Por outro lado, o tipo 6 é o menos comum no conjunto de dados.

Além disso, foi calculada a matriz de correlação para os primeiros 9 atributos, representando as variáveis físico-químicas que caracterizam os diferentes tipos de vidro. A matriz de correlação fornece informações sobre a relação linear entre pares de variáveis, com valores que variam entre -1 e 1, onde:

(i) **Valores próximos a 1**: indicam uma forte correlação positiva (à medida que uma variável aumenta, a outra tende a aumentar).

(ii) **Valores próximos a -1**: indicam uma forte correlação negativa (à medida que uma variável aumenta, a outra tende a diminuir).

(iii) **Valores próximos a 0**: sugerem pouca ou nenhuma relação linear entre as variáveis.

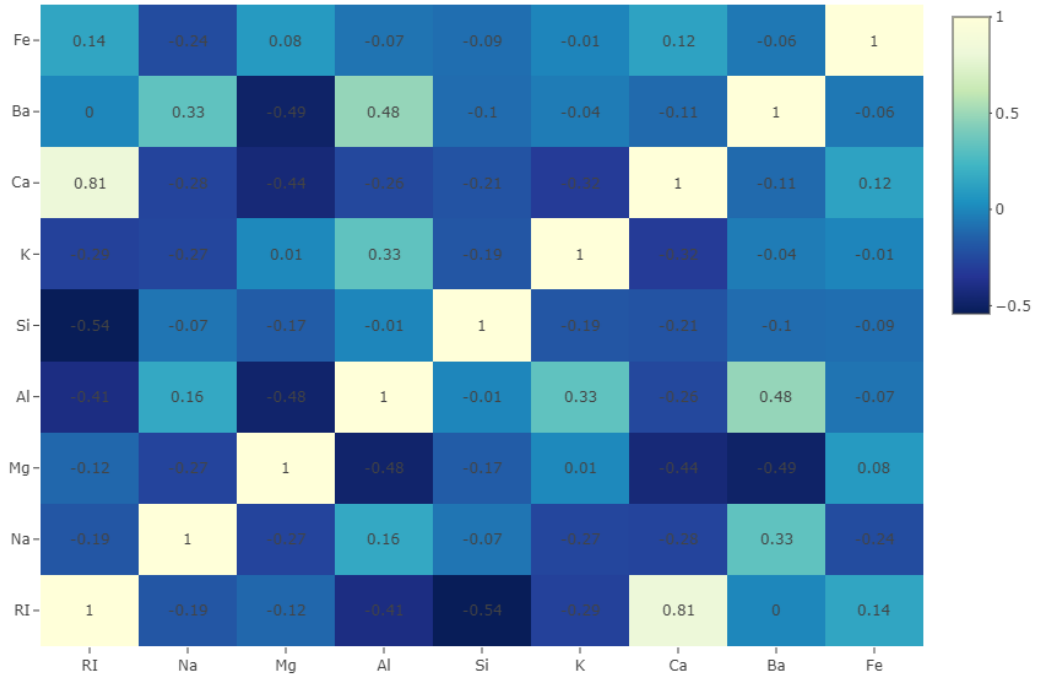


Figura 3: Gráfico de correlação.

1. A maior correlação positiva foi identificada entre RI (índice de refração) e Ca (cálcio), com um coeficiente de $r = 0.81$, indicando que estas duas variáveis possuem uma relação linear forte e direta.

2. Uma correlação negativa significativa foi observada entre RI e Si (silício) ($r = -0.54$), sugerindo que, à medida que o índice de refração aumenta, a quantidade de silício diminui.

3. A variável Ba (bário) não apresenta relações lineares muito fortes com outras variáveis, com coeficientes próximos de 0, indicando baixa dependência linear.

4. Outras relações dignas de nota incluem:

    - Al (alumínio) apresenta correlação negativa com Mg (magnésio) ($r = -0.48$) e positiva com Ba ($r = 0.48$).
    - Ca tem correlação negativa com Mg ($r = -0.44$).

Para facilitar a análise e interpretação das relações, será inserido um gráfico de calor (heatmap) que ilustra a matriz de correlação. Nesse gráfico:

- As cores variam de tons claros a escuros, representando correlações mais fracas e mais fortes, respectivamente.

- Cada célula contém o valor arredondado do coeficiente de correlação, permitindo uma visão intuitiva das relações.

O gráfico ajuda a destacar padrões e relações importantes entre as variáveis, servindo como um recurso complementar para entender o comportamento dos dados.

Ademais, com ajuda dos histogramas, ilustramos a distribuição das variáveis físico-químicas no conjunto de dados. Cada variável é representada em um histograma separado, permitindo observar o comportamento dos dados, como concentração, dispersão e assimetrias nas distribuições.

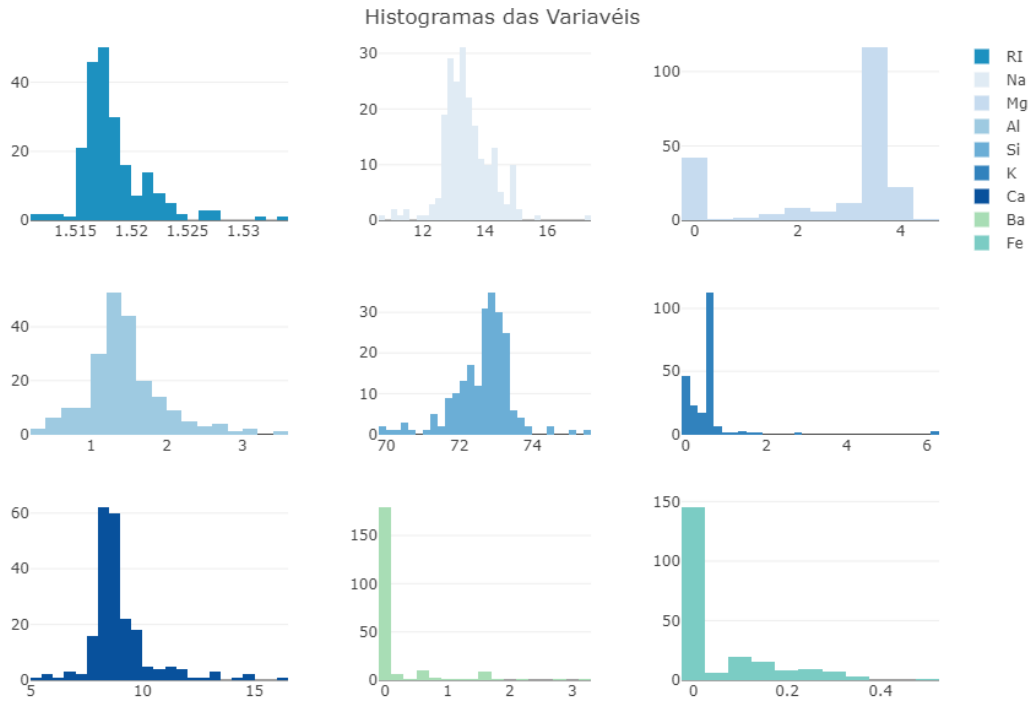


Figura 4: Distribuição das variáveis - Histogramas

As distribuições das variáveis no conjunto de dados apresentam características distintas. O índice de refração (RI) tem uma distribuição aproximadamente simétrica, centrada em torno de 1.52, com leve assimetria positiva. O sódio (Na) tem uma distribuição ampla e levemente assimétrica, com pico em torno de 14, com um número significativo de observações entre 12 e 15. O magnésio (Mg) apresenta uma distribuição assimétrica, com maior frequência em torno de 1.0 e valores decrescendo rapidamente à medida que aumentam. O alumínio (Al) também tem uma distribuição assimétrica, concentrando-se em torno de 1, com valores diminuindo gradualmente até cerca de 3. O silício (Si) tem valores fortemente concentrados em torno de 72, com uma leve cauda esquerda. O potássio (K) mostra uma forte concentração em torno de 0, com poucas observações acima de 1. O cálcio (Ca) tem um pico claro em torno de 9, com a maioria das observações concentradas entre 8 e 11. O bário (Ba) apresenta valores concentrados em torno de 0, sugerindo que essa variável é rara ou possui valores muito baixos na maioria dos casos. O ferro (Fe) tem uma distribuição altamente assimétrica, com a maior parte dos valores próximos de zero. As variáveis como Mg, Al, K, Ba e Fe apresentam distribuições assimétricas, indicando predominância de valores baixos, enquanto Ba, K e Fe são fortemente concentradas em valores próximos de zero, sugerindo baixa presença desses elementos. As variáveis RI, Si e Ca apresentam distribuições mais uniformes, sugerindo que esses componentes são mais consistentemente distribuídos no vidro, enquanto Na e Mg têm distribuições mais amplas, indicando maior variabilidade nas concentrações nos diferentes tipos de vidro.

Quando plotamos o boxplot de cada elemento, pode-se observar que o silício é

o componente principal do vidro, estando presente, em média, em mais de 70% da composição. Também é importante destacar que aproximadamente 90% da composição do vidro é formada por silício, sódio e cálcio. Por outro lado, pode-se observar que o ferro e o bário são os elementos com a menor presença.

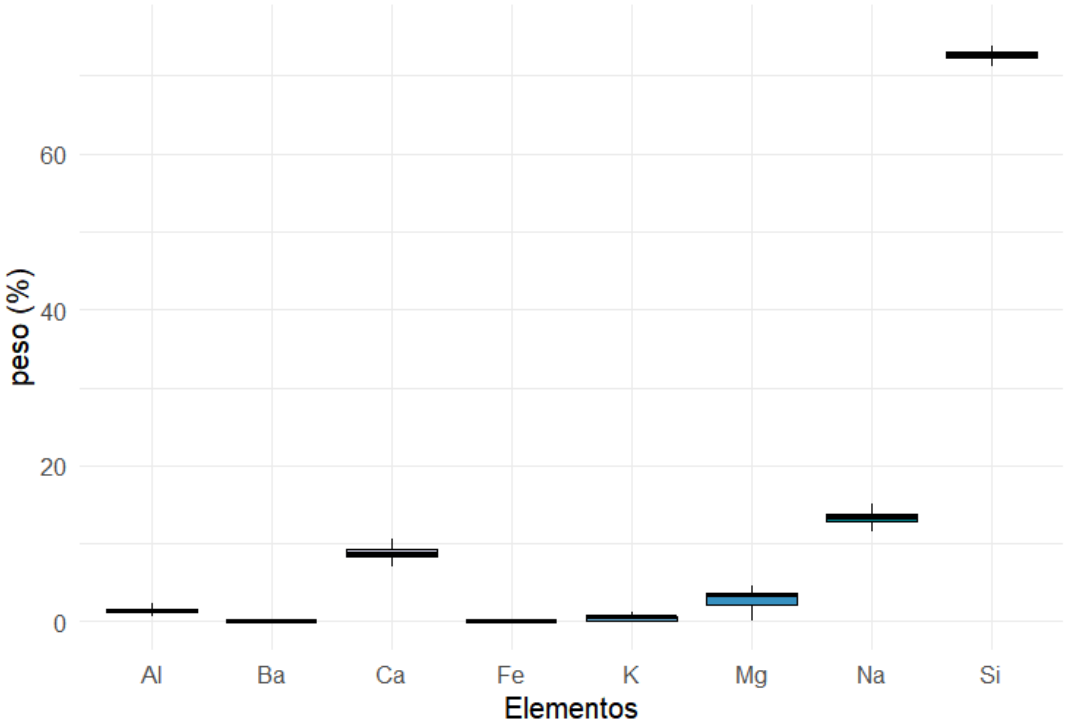


Figura 5: Concetração de peso % nos Elementos - boxplot

A análise do gráfico revela várias observações notáveis sobre a composição das amostras de vidro. Primeiro, o índice de refração está dentro de uma faixa estreita, variando de 1,51 a 1,54, indicando propriedades ópticas consistentes entre as amostras.

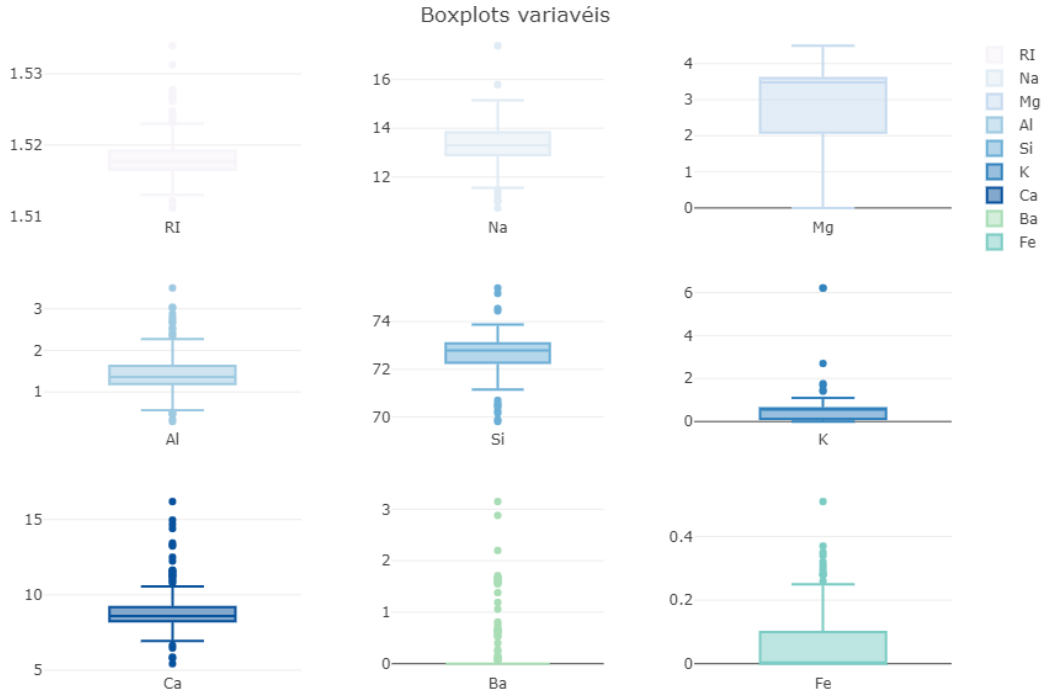


Figura 6: Boxplot - composições

Ao examinar a composição elementar, notamos que os vidros do Tipo 6 e Tipo 7 apresentam percentuais mais elevados de sódio (Na) em comparação com outros tipos. Em contrapartida, os vidros do Tipo 1, Tipo 2 e Tipo 3 exibem percentuais mais elevados de magnésio (Mg). Além disso, os vidros do Tipo 5 e Tipo 7 apresentam percentuais mais altos de alumínio (Al). Curiosamente, o percentual de silício (Si) permanece relativamente consistente entre todos os tipos, sugerindo seu papel integral como um componente fundamental da composição do vidro. Notavelmente, o vidro Tipo 6 não possui composição de potássio (K), bário (Ba) e ferro (Fe) de forma alguma. Além disso, os vidros Tipo 5 e Tipo 6 exibem composições mais altas de cálcio (Ca). O bário (Ba), por outro lado, aparece predominantemente nos vidros do Tipo 7. Por fim, o ferro (Fe) é encontrado principalmente nos vidros Tipo 1, 2 e 3.

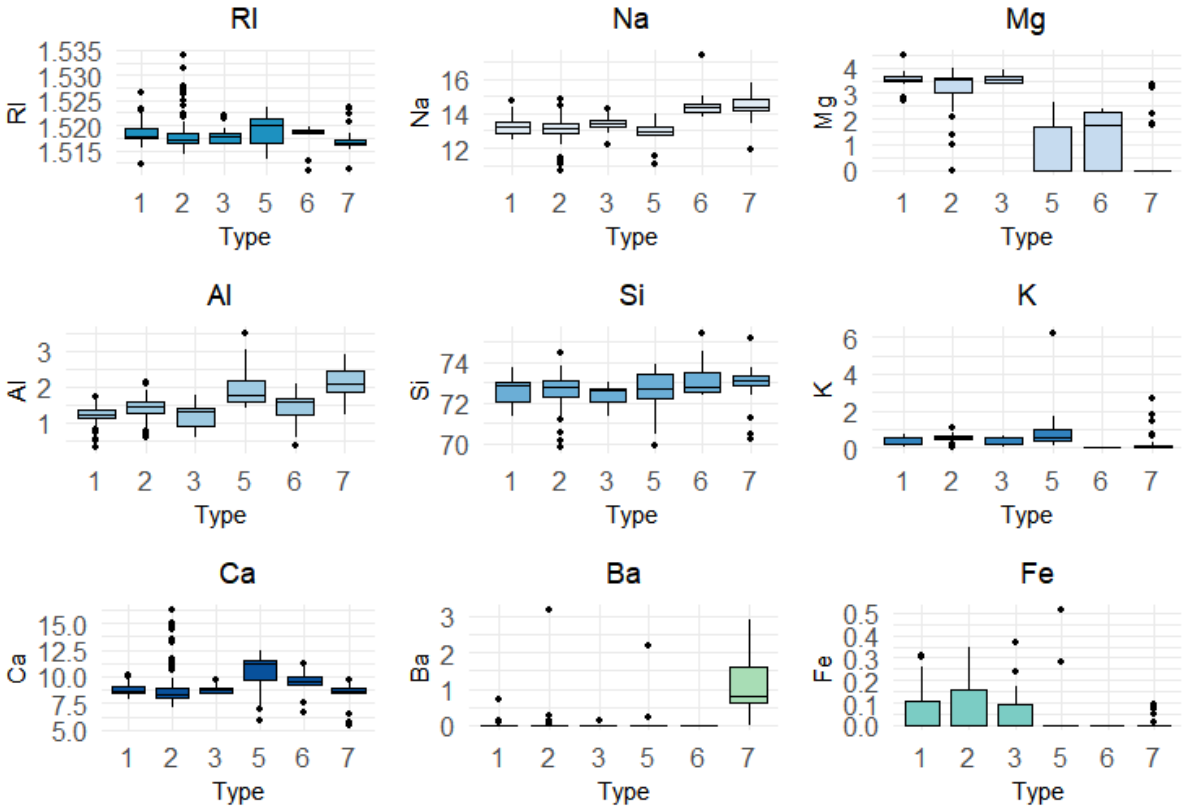


Figura 7: Distribuição das variáveis - Histogramas

3.1.1 Pré-processamento dos Dados

Durante a análise exploratória do conjunto de dados `glass`, foram realizadas as seguintes verificações:

(i) Valores Ausentes: A função `sum(is.na(glass))` indicou a ausência de valores ausentes, retornando o valor **0**. Esse resultado confirma que nenhuma variável possui registros faltantes, eliminando a necessidade de imputação ou exclusão de observações.

(ii) Valores Duplicados: A função `anyDuplicated(glass)` retornou o valor **40**, indicando que o registro localizado na linha 40 é duplicado em relação a um registro anterior no conjunto de dados. Para tratar essa duplicação, o comando `glass <- glass[!duplicated(glass),]` foi aplicado, removendo o registro duplicado e garantindo que o conjunto de dados não contenha observações redundantes.

Essas verificações garantem uma melhor compreensão da integridade dos dados. A ausência de valores ausentes é um aspecto positivo, pois facilita o trabalho analítico. Por outro lado, a identificação de um registro duplicado deve ser analisada para evitar possíveis impactos nos resultados da análise.

Durante o pré-processamento do conjunto de dados `glass`, diversas etapas foram realizadas para preparar os dados para análises posteriores, conforme descrito abaixo:

(i) Binarização das Variáveis `Ba` e `Fe`: Observou-se na análise exploratória que os elementos Ba (Bário) e Fe (Ferro) não são amplamente distribuídos entre as instâncias. Para simplificar a análise, essas variáveis foram binarizadas: valores maiores que 0 foram codificados como 1, indicando a presença do elemento na instância, enquanto valores iguais a 0 foram codificados como 0, indicando a ausência. Essa transformação foi realizada utilizando a função `ifelse()` no R, e as variáveis Ba e Fe foram convertidas para o tipo categórico (`factor`). A variável de saída `Type` também foi transformada para o tipo categórico, a fim de garantir consistência nos métodos de análise.

(ii) Detecção e Tratamento de Outliers: A identificação de outliers foi conduzida utilizando o método do Intervalo Interquartil (IQR). Para cada variável numérica, os limites inferior e superior foram definidos como:

$$\text{Limite Inferior} = Q1 - 3 \cdot \text{IQR} \quad \text{e} \quad \text{Limite Superior} = Q3 + 3 \cdot \text{IQR},$$

onde $Q1$ e $Q3$ correspondem ao primeiro e terceiro quartis, respectivamente, e $\text{IQR} = Q3 - Q1$. Essa abordagem é mais rigorosa em relação ao método convencional, que utiliza o multiplicador de $1,5$, pois utiliza 3 como multiplicador para detectar outliers extremos. Foram identificadas **16 instâncias** como outliers. Essas instâncias foram removidas do conjunto de dados utilizando os índices retornados pela análise.

(iii) Impactos do Tratamento de Outliers: Embora a remoção de outliers seja uma prática comum para melhorar a qualidade dos dados, reconhece-se que essa etapa pode levar à perda de informações potencialmente valiosas. Amostras discrepantes podem conter características importantes relacionadas ao tipo de vidro ou à sua composição química. Além disso, a ausência de conhecimento aprofundado na área de vidros e sua composição pode ter influenciado as decisões tomadas durante o tratamento, introduzindo possíveis vieses na análise subsequente.

(iv) Conjunto de Dados Limpo: Após a binarização e remoção dos outliers, obteve-se um conjunto de dados pré-processado e limpo (`df_cat_clean`), pronto para análise e construção de modelos e das técnicas.

Tabela 4: Resultados dos testes de normalidade (Multivariada e Univariada).

Tipo de Teste	Variável/Teste	Estatística	p-valor	Resultado
3*Multivariada	Mardia Skewness	4077.31	0	NÃO
	Mardia Kurtosis	81.28	0	NÃO
	MVN	NA	NA	NÃO
9*Univariada	Anderson-Darling (RI)	8.54	< 0.001	NÃO
	Anderson-Darling (Na)	3.12	< 0.001	NÃO
	Anderson-Darling (Mg)	29.12	< 0.001	NÃO
	Anderson-Darling (Al)	3.85	< 0.001	NÃO
	Anderson-Darling (Si)	5.14	< 0.001	NÃO
	Anderson-Darling (K)	25.76	< 0.001	NÃO
	Anderson-Darling (Ca)	14.28	< 0.001	NÃO
	Anderson-Darling (Ba)	55.50	< 0.001	NÃO
	Anderson-Darling (Fe)	32.46	< 0.001	NÃO

Os resultados indicam que os dados não seguem uma distribuição normal multivariada, conforme evidenciado pelos testes de Mardia (Assimetria: estatística = 4077,31, $p < 0,001$; Curtose: estatística = 81,28, $p < 0,001$), resultando em **NO** para normalidade multivariada. Além disso, os testes de normalidade univariada (Anderson-Darling) revelaram que nenhuma das variáveis individuais (**RI**, **Na**, **Mg**, **Al**, **Si**, **K**, **Ca**, **Ba**, **Fe**) segue uma distribuição normal ($p < 0,001$ para todas).

3.2 Análise de Componentes Principais

Sabemos que para aplicar corretamente a Análise de Componentes Principais (PCA), nossos dados não necessariamente precisam atender a certos pressupostos. Observamos que, no nosso caso, a nromalidade não é atendido para as variáveis, particularmente para Ba e Fe. Como mencionado anteriormente, essas variáveis possuem um grande número de valores zero. Além disso, é necessário que haja um certo nível de correlação entre as variáveis. No entanto, a variável Fe não exibe correlações significativas com as demais variáveis. Portanto, decidimos excluir as variáveis Ba e Fe desta análise. Adicionalmente, é essencial que nossos dados estejam livres de outliers. Assim, antes de aplicar a PCA, removeremos os outliers utilizando o método do Intervalo Interquartil (IQR). Utilizamos uma margem maior para evitar a remoção de muitas observações.

Aplicaremos a PCA nas características do conjunto de dados, utilizando a variável Tipo como variável suplementar. Como estamos buscando relações entre os elementos e o tipo, também utilizaremos a variável índice de reflexão como variável suplementar. Aplicamos a PCA nos dados limpos e, para a análise dos componentes resultantes, focaremos nos três primeiros componentes. Esses três componentes foram escolhidos porque possuem autovalores acima da média. Ao utilizar os três primeiros componentes, conseguimos explicar 85% da variância dos dados. Na próxima seção, será feita uma explicação mais detalhada sobre a interpretação dos resultados obtidos através da análise de componentes principais (PCA). A abordagem envolverá a análise dos

autovalores, a contribuição das variáveis para os componentes principais, e como as variáveis suplementares foram incorporadas na análise, além de uma visão geral das implicações desses resultados para a compreensão do conjunto de dados.]

O objetivo do PCA é encontrar um novo sistema de coordenadas em que as direções das novas variáveis maximizam a variância dos dados. Cada componente principal Y_j pode ser escrita como uma combinação linear das variáveis originais:

$$Y_j = e_{j1}X_1 + e_{j2}X_2 + \cdots + e_{jp}X_p$$

$$Y_1 = 0.7292 \cdot Na - 0.9075 \cdot Mg + 0.5353 \cdot Al + 0.1475 \cdot Si - 0.7501 \cdot K + 0.4094 \cdot Ca$$

$$Y_2 = -0.2324 \cdot Na - 0.2726 \cdot Mg + 0.6769 \cdot Al + 0.7118 \cdot Si + 0.4363 \cdot K - 0.5323 \cdot Ca$$

$$Y_3 = -0.5814 \cdot Na - 0.1558 \cdot Mg - 0.2615 \cdot Al + 0.5453 \cdot Si - 0.1040 \cdot K + 0.6450 \cdot Ca$$

$$Y_4 = -0.1662 \cdot Na - 0.1560 \cdot Mg + 0.3417 \cdot Al - 0.4079 \cdot Si + 0.3864 \cdot K + 0.3582 \cdot Ca$$

$$Y_5 = 0.2072 \cdot Na - 0.2108 \cdot Mg - 0.2620 \cdot Al + 0.0698 \cdot Si + 0.2929 \cdot K + 0.0178 \cdot Ca$$

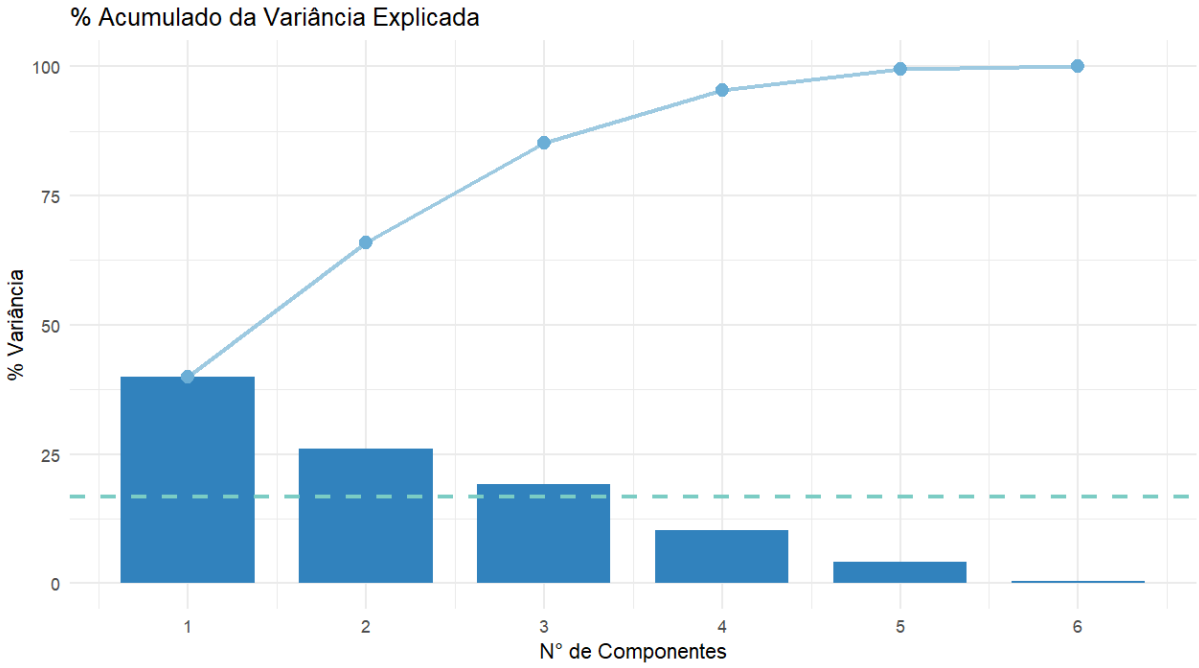


Figura 8: Variância Acumulada

Tabela 5: Autovalor, Percentual de Variância e Percentual Acumulado de Variância das Componentes Principais.

Componente	Autovalor	Variância (%)	% Acumulado de Variância (%)
Comp 1	2.3939123	39.90	39.90
Comp 2	1.5669594	26.12	66.01
Comp 3	1.1547896	19.25	85.26
Comp 4	0.6126891	10.21	95.47
Comp 5	0.2469323	4.12	99.59
Comp 6	0.0247173	0.41	100.00
Total	-	100.00	-

Como os dados estão escalonados, o valor médio do autovalor é 1. Portanto, extrairemos os primeiros 3 componentes principais, pois são os únicos componentes que possuem um autovalor acima de 1. Infelizmente, não existe uma maneira objetiva bem aceita para decidir quantos componentes principais são suficientes. Isso dependerá do campo de aplicação específico e do conjunto de dados específico. Na prática, tendemos a olhar para os primeiros componentes principais a fim de encontrar padrões interessantes nos dados. Na nossa análise, os três primeiros componentes principais explicam 85% da variação. Esta é uma porcentagem aceitavelmente grande. Uma alternativa para determinar o número de componentes principais é olhar para o Scree Plot, que é o gráfico dos autovalores ordenados do maior para o menor. De acordo com Jolliffe (2002) o número de componentes é determinado no ponto além do qual os autovalores restantes são todos relativamente pequenos e de tamanho comparável.

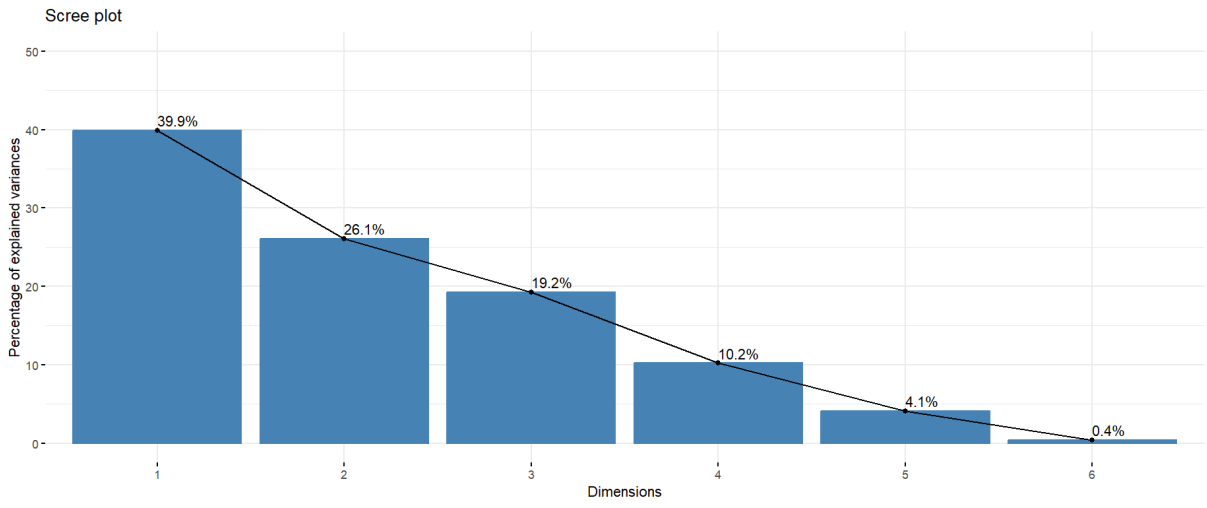


Figura 9: Scree plot dos autovalores dos componentes principais

3.3 Análise de Cluster

A clusterização hierárquica aglomerativa foi utilizada para identificar padrões de similaridade entre os diferentes tipos de vidro do nosso conjunto de dados sobre vidros. A aplicação combinada de Análise de Componentes Principais (PCA) e Clustering Hierárquico buscou torna-se a a ánalise mais interpretável, reduzir a dimensionalidade

dos dados e identificar padrões robustos, mesmo na presença de outliers. Essa abordagem é particularmente útil quando há variáveis correlacionadas, como no caso do conjunto de dados utilizado, pois facilita a separação entre os grupos de interesse.

O PCA foi empregado para condensar as variáveis originais em três componentes principais (`Dim.1`, `Dim.2` e `Dim.3`), que capturam a maior variância dos dados, transformando-os em um espaço ortogonal onde as relações multivariadas são mais claras. A escolha da distância euclidiana para a matriz de dissimilaridades justifica-se por sua adequação a dados contínuos em espaços reduzidos, preservando a geometria das relações entre observações. O método de Ward (Ward.D2) foi selecionado por minimizar a variância intra-clusters, alinhando-se ao PCA, que também prioriza a retenção de variância. Após a aplicação da PCA, foram calculadas as medianas das coordenadas dos dados em cada um dos três primeiros componentes principais, agrupadas por tipo de vidro. A escolha da mediana, em vez da média, foi motivada pelo fato de que a mediana é menos sensível a outliers, garantindo que os valores extremos não distorçam a representação central de cada grupo. Essa abordagem melhora a estabilidade da análise ao fornecer uma melhor estimativa da localização central dos grupos de vidro, especialmente quando há assimetrias ou distribuições não normais nos dados que é o nosso caso.

O dendrograma gerado a partir dessa análise permitiu visualizar a estrutura hierárquica dos agrupamentos, facilitando a interpretação das relações entre os diferentes tipos de vidro. A escolha do número de clusters foi baseada na inspeção da estrutura do dendrograma, onde foi possível identificar três grandes agrupamentos distintos, refletindo a similaridade entre os tipos de vidro de acordo com suas características principais extraídas via PCA. Essa abordagem permitiu uma segmentação mais precisa dos diferentes tipos de vidro, tornando a análise mais robusta e interpretável.

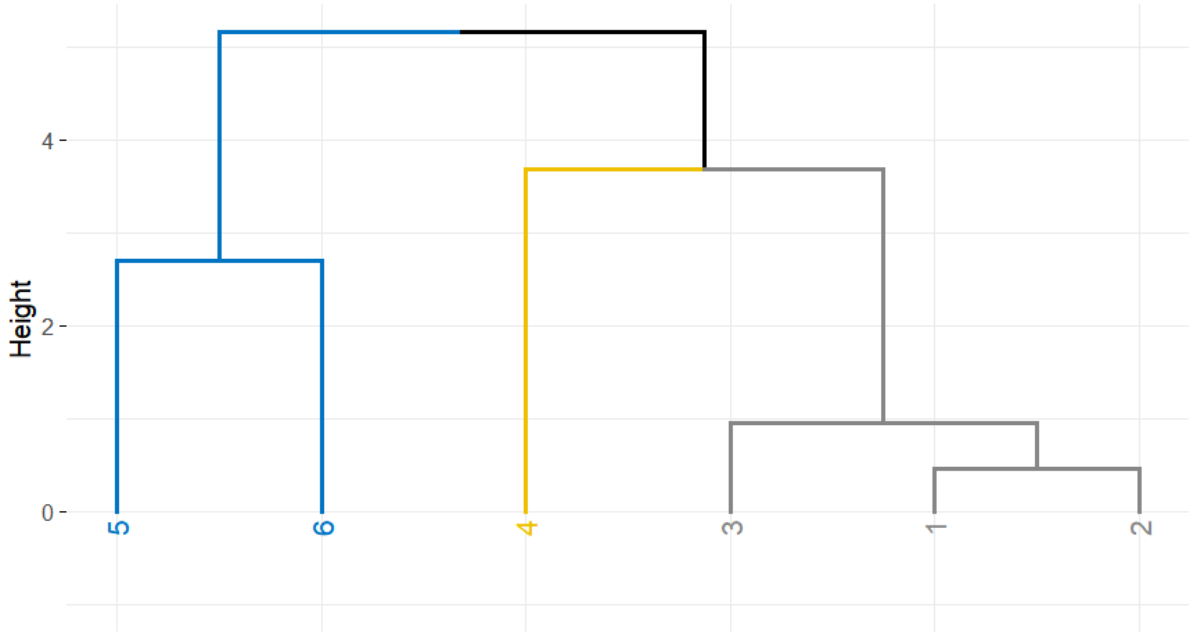


Figura 10: Dendograma com os cluster

Além disso, o processo de definição do número de clusters é uma parte fundamental da análise de agrupamento. Para isso, foram empregados diferentes métodos para avaliar a qualidade e a adequação da divisão dos dados em clusters. Cada abordagem possui vantagens distintas e oferece uma perspectiva única sobre como os dados devem ser agrupados.

O método do cotovelo é uma das técnicas mais utilizadas para determinar o número de clusters. Ele consiste em plotar a soma dos quadrados das distâncias dentro dos clusters (também conhecida como ”within-cluster sum of squares”) em função do número de clusters. O objetivo é observar onde ocorre uma diminuição abrupta dessa soma, o que indica que adicionar mais clusters não resulta em uma melhoria significativa na qualidade do agrupamento.

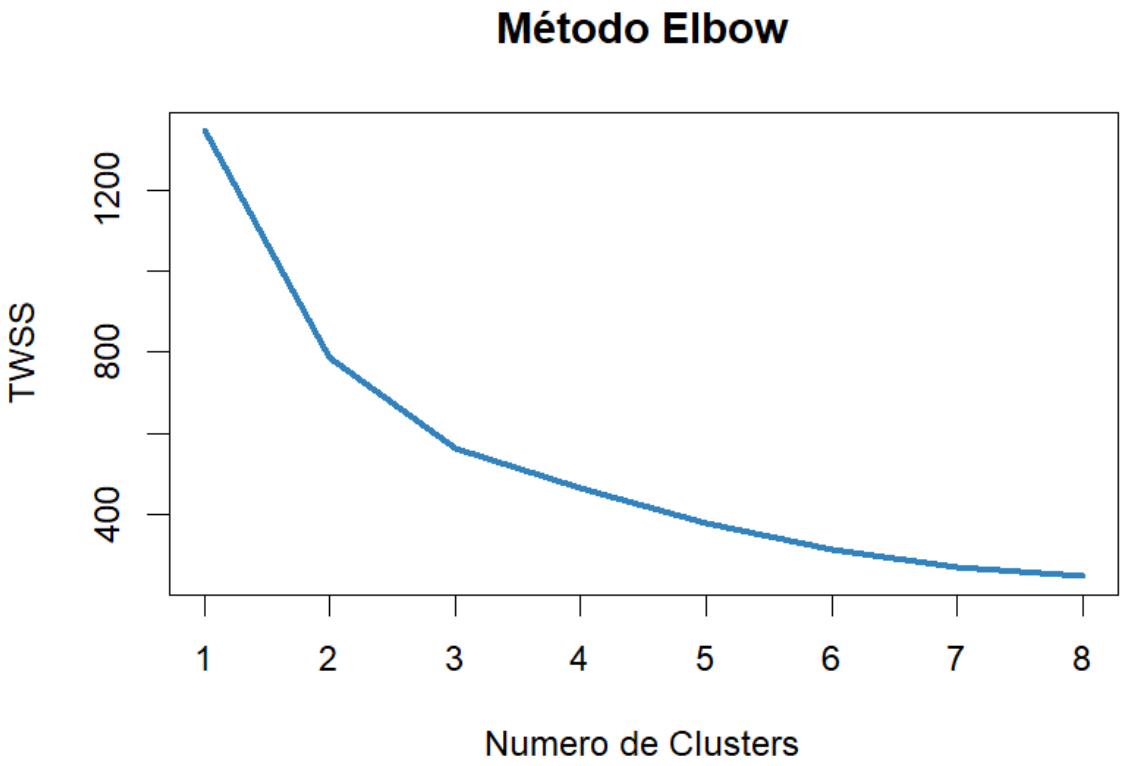


Figura 11: Numero de Clusters vs TWSS.

No gráfico do cotovelo, o ponto onde a curva se ”aplana”e a redução da soma dos quadrados começa a ser menos pronunciada é considerado o número ótimo de clusters. Esse ponto é muitas vezes chamado de ”cotovelo”. No nosso caso, o gráfico indicou que o número ideal de clusters deve ser três, o que corrobora os resultados obtidos com os métodos hierárquicos. Ao observar a diminuição gradual da soma dos quadrados, foi possível identificar que adicionar mais clusters além de três não proporcionava uma melhoria substancial, tornando esse valor o mais apropriado para a análise.

Além disso, foi aplicado também o método da silhueta e o método da estatística gap. O coeficiente de silhueta mede a qualidade da divisão em clusters, balanceando coesão

interna e separação entre grupos. O valor ótimo é identificado pelo pico da curva de silhueta média. Em dados com outliers, este método pode ser menos confiável, mas serve como referência visual.

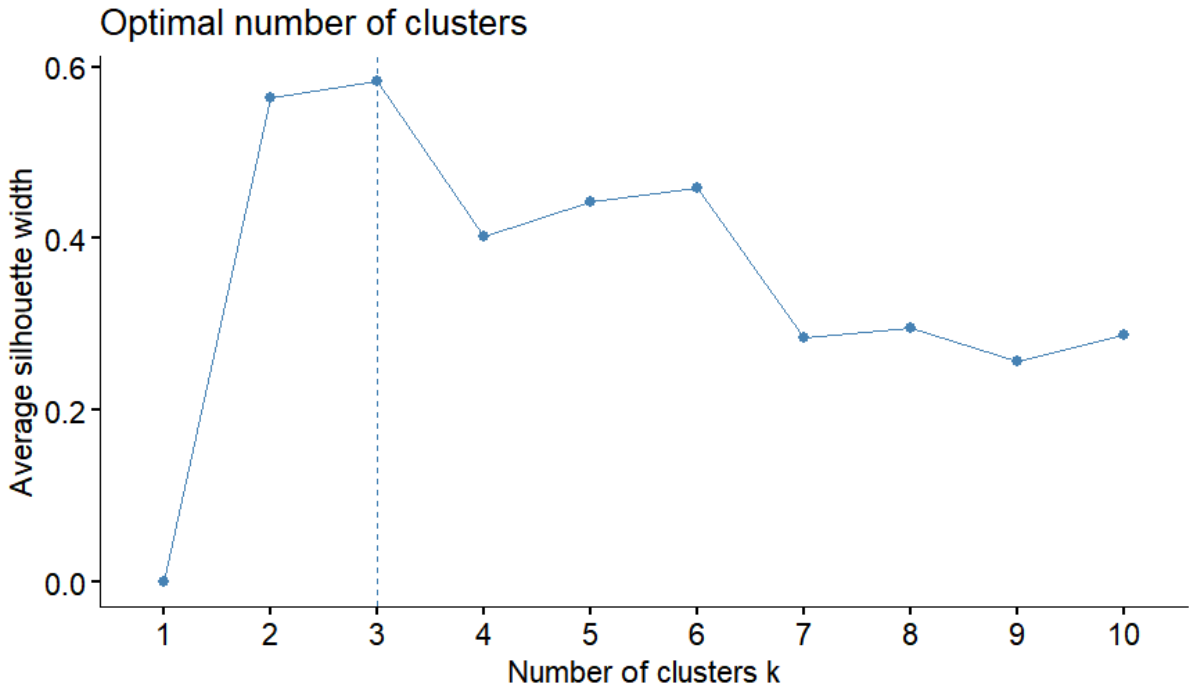


Figura 12: Método da Silhueta

Já estatística gap compara a dispersão intra-cluster dos dados observados com uma distribuição nula aleatória. O número ideal de clusters é o menor valor onde a curva gap atinge um platô. Este método é particularmente robusto em cenários com alta variabilidade ou outliers, alinhando-se à escolha da mediana na etapa anterior.

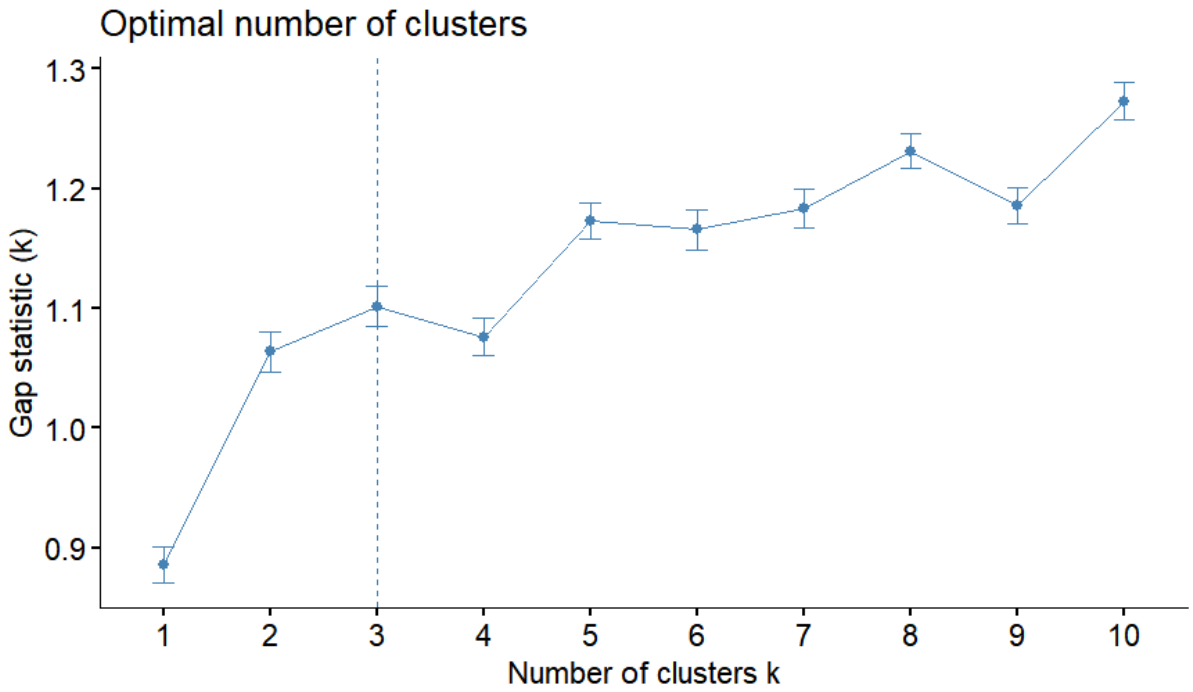


Figura 13: Método Gap

Diante da análise realizada, integrando PCA, clustering hierárquico e métodos de validação (silhueta e gap statistic), poderíamos estender a investigação comparando os resultados com o algoritmo K-Means. Este método particiona os dados em k clusters pré-definidos, minimizando a variância intra-clusters através da iteração de centróides. Embora o método K-Means seja uma alternativa viável para análise de clusters, optou-se por excluí-lo do escopo deste relatório devido à sua sensibilidade a outliers e à natureza não hierárquica, que limitariam a interpretação robusta requerida pelos dados. A abordagem hierárquica com PCA, aliada à agregação por medianas, já forneceu uma partição estável e validada estatisticamente (via silhueta e *gap statistic*), garantindo coerência metodológica e clareza nos resultados. Eventuais comparações com K-Means poderiam ser exploradas em estudos futuros, com ajustes específicos para mitigação de outliers.

4 Resultados

4.1 Resultados da ACP

A análise de componentes principais (PCA) revelou que as variáveis possuem contribuições significativas nos três primeiros componentes principais. As coordenações e contribuições podem ser interpretadas com base nos gráficos de círculo de correlação e nas métricas de Coordenação, Correlação, Coseno Quadrado e Contribuição dos elementos. O Componente 1 (Dim 1 - 39,90%): Este componente é fortemente influenciado por Na (Sódio) e Ca (Cálcio), que possuem alta correlação positiva com o eixo. Por outro lado, Mg (Magnésio) tem uma correlação negativa. Esses elementos químicos destacam as principais diferenças observadas nos dados originais. Já o Componente 2 (Dim 2 - 26,12%): As variáveis Si (Silício) e Al (Alumínio) contribuem positivamente para este componente, enquanto K (Potássio) e Mg possuem contribuições negativas. Este componente reflete uma nova dimensão de variância explicada, associada a elementos estruturais do vidro. E o Componente 3 (Dim 3 - 19,25%): Neste componente, Ca tem maior contribuição positiva, enquanto RI (Índice de Refração), destacado em azul nos gráficos, também se apresenta como uma variável importante. K e Mg estão negativamente correlacionados com este eixo, indicando sua menor influência em algumas dimensões específicas.

Tabela 6: Coordenação, Correlação, Coseno Quadrado e Contribuição dos Elementos nas Dimensões Principais.

Elemento	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Coord.					
Na	0.7292	-0.2324	-0.5814	-0.1662	0.2072
Mg	-0.9075	-0.2726	-0.1558	-0.1560	-0.2108
Al	0.5353	0.6769	-0.2615	0.3417	-0.2620
Si	0.1475	0.7118	0.5453	-0.4079	0.0698
K	-0.7501	0.4363	-0.1040	0.3864	0.2929
Ca	0.4094	-0.5323	0.6450	0.3582	0.0178
Cor.					
Na	0.7292	-0.2324	-0.5814	-0.1662	0.2072
Mg	-0.9075	-0.2726	-0.1558	-0.1560	-0.2108
Al	0.5353	0.6769	-0.2615	0.3417	-0.2620
Si	0.1475	0.7118	0.5453	-0.4079	0.0698
K	-0.7501	0.4363	-0.1040	0.3864	0.2929
Ca	0.4094	-0.5323	0.6450	0.3582	0.0178
Cos2					
Na	0.5317	0.0540	0.3380	0.0276	0.0429
Mg	0.8236	0.0743	0.0243	0.0244	0.0444
Al	0.2865	0.4582	0.0684	0.1168	0.0686
Si	0.0218	0.5067	0.2973	0.1664	0.0049
K	0.5627	0.1904	0.0108	0.1493	0.0858
Ca	0.1676	0.2834	0.4160	0.1283	0.0003
Contrib.					
Na	22.21	3.45	29.27	4.51	17.38
Mg	34.40	4.74	2.10	3.97	17.99
Al	11.97	29.24	5.92	19.06	27.79
Si	0.91	32.34	25.75	27.16	1.98
K	23.51	12.15	0.94	24.36	34.74
Ca	7.00	18.08	36.02	20.94	0.13

4.1.1 Coordenadas

As coordenadas mostram como cada variável original contribui para cada componente principal. Valores absolutos maiores indicam uma contribuição mais significativa. O Comp 1 composto por Mg tem a maior contribuição negativa (-0.9075), enquanto Na e Al têm contribuições positivas signifcativas. Isso sugere que o Comp 1 está fortemente relacionado à oposição entre Mg e Na/Al. Além disso, o Comp 2 composto por Si e Al têm as maiores contribuições positivas, enquanto Ca tem uma contribuição negativa significativa. Isso indica que o Comp 2 contrasta Si/Al com Ca. E por último, o Comp 3: Ca e Si têm as maiores contribuições positivas, enquanto Na tem uma contribuição negativa significativa. Isso sugere que o Comp 3 contrasta Ca/Si com Na.

4.1.2 Importância do Cos^2

O coseno ao quadrado (Cos^2) foi utilizado como métrica de importância para verificar a qualidade da representação das variáveis nos componentes principais. Variáveis com Cos^2 próximo de 1 estão bem representadas nos componentes, enquanto valores baixos indicam menor contribuição ou alinhamento com aquele componente específico. A análise revelou que Na, Ca, Si, e Al apresentam altos valores de Cos^2 para os primeiros dois componentes, indicando que são bem explicados pela variância total capturada pela PCA. Já RI mostrou maior Cos^2 no Componente 3, destacando-se em dimensões mais específicas do problema. Além disso, Mg e K apresentam Cos^2 mais distribuído, refletindo sua menor dominância em um único componente, mas contribuindo em múltiplas dimensões.

O objetivo do estudo foi compreender as principais diferenças na composição química de diferentes tipos de vidro. Os componentes principais fornecem uma visão clara das variáveis mais influentes. Diante disso, o Componente 1 separa os vidros com base em Na e Ca, elementos importantes na resistência e transparência. O Componente 2 explora as diferenças associadas a Si e Al, elementos-chave na estrutura do vidro e propriedades térmicas. E o Componente 3 adiciona detalhes relacionados ao RI (Índice de Refração), essencial para aplicações ópticas, e Ca, associado à durabilidade. Logo, isso permite identificar as principais características químicas que diferenciam os tipos de vidro analisados, facilitando uma classificação precisa e direcionada às suas propriedades de uso.

Os círculos de correlação apresentados mostram a relação entre as variáveis originais e os componentes principais. Cada círculo corresponde a um par de componentes (Dim 1 x Dim 2, Dim 1 x Dim 3, Dim 2 x Dim 3). Cada variável é representada por um vetor (seta) no círculo, onde o comprimento da seta indica a qualidade da representação (ou o valor do *cos²*) da variável no espaço formado pelas dimensões selecionadas. As setas mais longas, próximas à circunferência, indicam que a variável é bem representada pelo par de componentes e as setas curtas ou próximas à origem indicam que a variável tem pouca influência ou não é bem representada por aquele par de componentes.

(i) Dim 1 x Dim 2 (Gráfico à esquerda): As variáveis Na, Ca, Si, Al e K estão próximas à circunferência, indicando boa representação neste plano.Na e Ca possuem uma forte correlação positiva com o Componente 1, enquanto Si e Al se correlacionam mais fortemente com o Componente 2. Mg apresenta correlação negativa com o Componente 1, estando orientada na direção oposta de Na e Ca. RI (Índice de Refração), em azul, possui menor contribuição relativa no plano Dim 1 x Dim 2, como evidenciado por sua seta mais curta e distante da circunferência.

(ii) Dim 2 x Dim 3 (Gráfico ao centro): Si e Al continuam sendo bem representados, agora com maior influência no Componente 2, enquanto RI ganha destaque no Componente 3. Mg, Na e K possuem vetores inclinados, mostrando que são parcialmente representados por este par de componentes, mas com menor relevância no Componente 3.

(iii) Dim 1 x Dim 3 (Gráfico à direita):

As variáveis Na, Ca e Si são bem representadas neste plano, com Na e Ca correlacionadas positivamente com o Componente 1. RI aparece com maior contribuição no Componente 3, representado por sua posição próxima à circunferência no eixo vertical. Mg apresenta correlação negativa com o Componente 1, alinhando-se opostamente aNa e Ca.

(iv) Correlação entre as variáveis:

Setas próximas indicam alta correlação entre as variáveis. Por exemplo, Na e Ca são fortemente correlacionados em todos os gráficos. Setas em direções opostas indicam correlação negativa, como entre Mg e Na/Ca. Setas formando ângulos próximos a 90° indicam variáveis não correlacionadas, como Si e Mg no gráfico Dim 1 x Dim 2.

(v) Relação com o contexto do problema: A boa representação das variáveis químicas (Na, Ca, Mg, Si, K) nos primeiros componentes reflete sua importância na análise estrutural do vidro. RI (Índice de Refração) é mais bem representado no Componente 3, evidenciando que as propriedades ópticas do vidro têm menor relação com as principais variações químicas, mas são capturadas no terceiro eixo. O alinhamento de Na e Ca indica que essas variáveis frequentemente variam juntas no contexto analisado, enquanto a oposição de Mg destaca uma composição química distinta.

Dessa forma, os círculos de correlação nios permitem uma análise detalhada das relações entre as variáveis e os componentes principais, elucidando os padrões químicos e ópticos mais relevantes no nosso conjunto de dados.

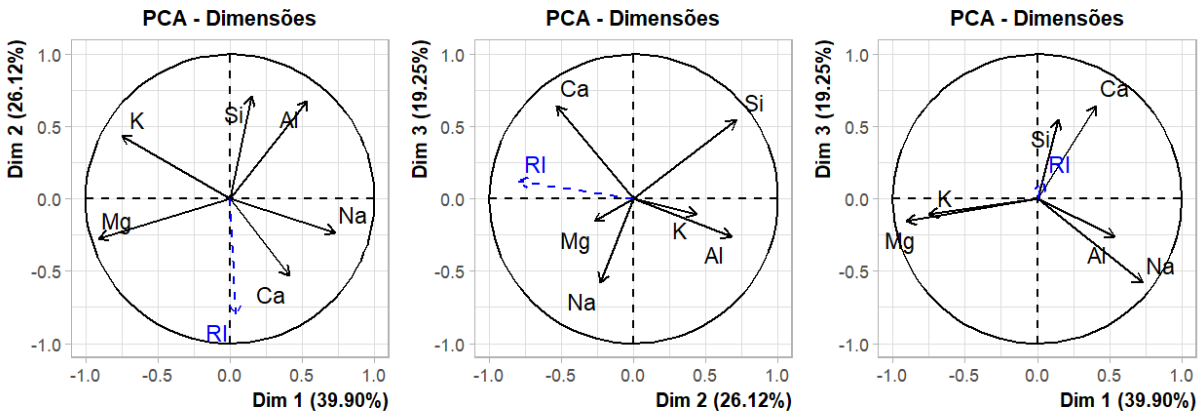


Figura 14: Correlações de variavéis

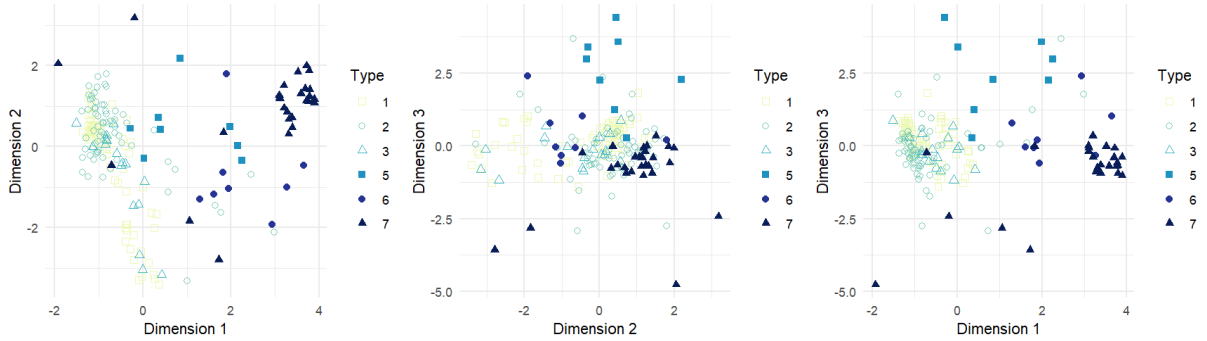


Figura 15: Gráficos Individuais - dimensões

O biplot combina as informações sobre as variáveis originais e as observações em um único gráfico, permitindo interpretar tanto a contribuição de cada variável nos componentes principais quanto a distribuição das observações no espaço reduzido. Os dois primeiros componentes (Dim 1 e Dim 2) explicam juntos 66,02% da variância total dos dados, proporcionando uma boa representação do conjunto original em duas dimensões. A inclusão do terceiro componente (Dim 3) aumenta a variância explicada para 85,27%.

(i) No Componente 1, as variáveis *Na* (Sódio) e *Ca* (Cálcio) têm fortes contribuições positivas, enquanto *Mg* (Magnésio) possui uma contribuição negativa, o que sugere que o Componente 1 está relacionado à variação química entre esses elementos.

(ii) No Componente 2, *Si* (Silício) e *Al* (Alumínio) apresentam as maiores contribuições positivas, enquanto *K* (Potássio) e *Mg* contribuem negativamente. Este componente parece capturar variações relacionadas à estrutura química do vidro.

(iii) O Componente 3 destaca a importância de *RI* (Índice de Refração), em azul nos gráficos, como uma variável fundamental para características ópticas, enquanto *Ca* também tem alta contribuição positiva.

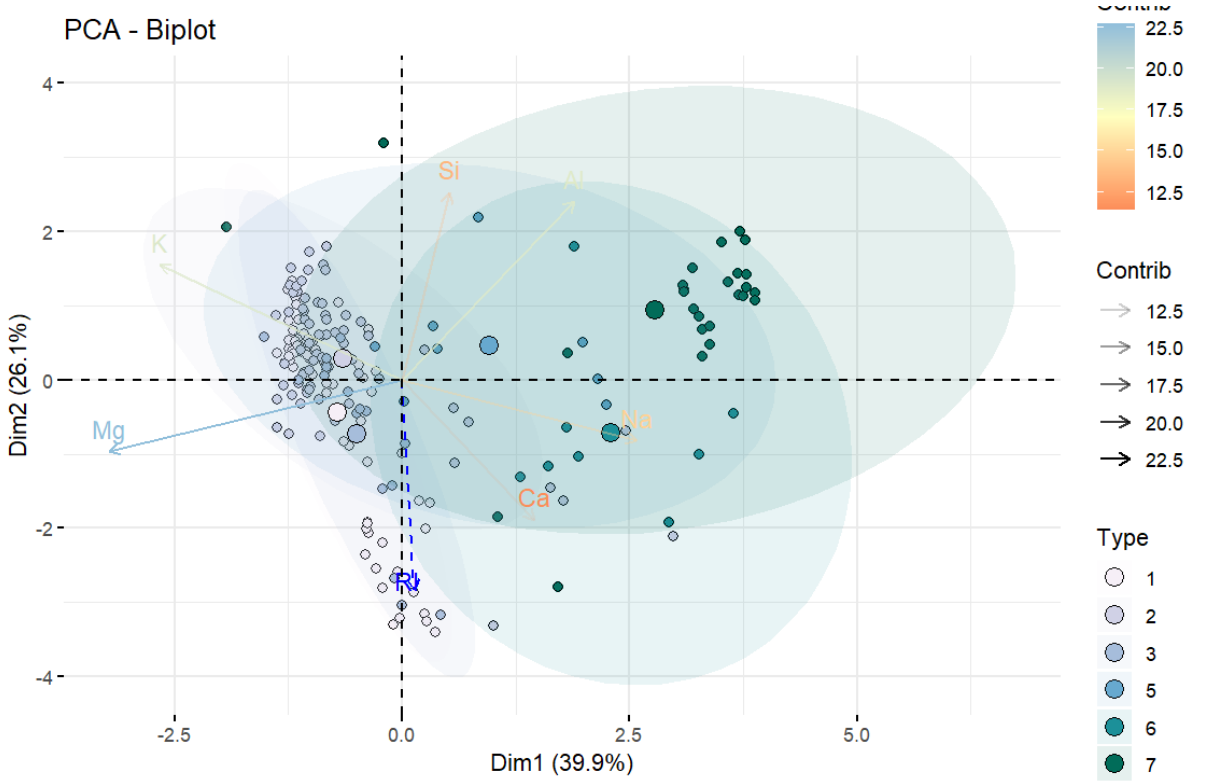


Figura 16: indivíduos por grupos e variáveis por suas contribuições aos componentes principais

A proximidade das setas no biplot reflete a correlação entre as variáveis. Na e Ca apresentam forte correlação positiva no primeiro componente. Mg está negativamente correlacionado com Na e Ca, indicando que são características opostas no contexto analisado. Si e Al mostram uma correlação positiva moderada, refletindo sua contribuição conjunta na estrutura do vidro. No biplot, as observações estão distribuídas ao longo dos componentes principais. Observações próximas a uma variável indicam maior valor relativo para aquela característica, enquanto posições intermediárias refletem combinações de características. Os componentes capturam diferenças significativas na composição química do vidro, importantes para aplicações específicas. Por exemplo, Na e Ca são relevantes para a resistência e transparência, enquanto Si e Al impactam a estabilidade térmica e estrutural.O RI (Índice de Refração) é crítico para aplicações ópticas e, no terceiro componente, destaca variações mais específicas nas propriedades do vidro.

A análise de componentes principais (PCA) revelou que os três primeiros componentes capturam aproximadamente 85,26% da variância total dos dados, sendo suficientes para representar a maior parte da informação contida no conjunto de dados. A PCA demonstrou ser uma ferramenta eficaz para reduzir a dimensionalidade do conjunto de dados e identificar as principais características químicas que diferenciam os tipos de vidro. Os três primeiros componentes principais capturam a maior parte da variância, destacando a importância de elementos como **Na**, **Ca**, **Si**, **Al** e **Mg** na composição e nas propriedades do vidro. A análise dos círculos de correlação e do biplot permitiu visualizar as relações entre as variáveis e as observações, fornecendo insights valiosos para a classificação e aplicações específicas dos vidros. Essa abordagem facilita a compreensão das diferenças químicas e estruturais entre os tipos de vidro, contribuindo

para a tomada de decisões em contextos práticos.

4.2 Resultados da Análise de Cluster

A estrutura hierárquica do dendrograma, combinada à análise das componentes principais (PCA), revelou três clusters distintos, organizados com base nas dissimilaridades calculadas entre os tipos de vidro. A aplicação do método de Ward.D2 na matriz de distâncias euclidianas (calculada sobre as medianas das três componentes principais) permitiu identificar padrões robustos, mesmo em um cenário com presença de outliers. A escolha da mediana, em vez da média, garantiu que os centróides dos grupos representassem tendências centrais estáveis, minimizando distorções causadas por valores extremos.

Tabela 7: Mediana das Coordenadas nos Três Primeiros Componentes Principais por Tipo de Vidro.

Tipo de Vidro	Dimensão 1	Dimensão 2	Dimensão 3
1	-0.832	0.200	0.266
2	-0.918	0.373	-0.153
3	-0.489	-0.400	-0.180
5	0.621	0.435	2.620
6	1.920	-1.020	0.0812
7	3.310	1.180	-0.661

O grupo 1 destaca-se pela elevada dissimilaridade em relação aos demais clusters, especialmente devido ao valor extremo na terceira componente principal (Dim.3 = 2.62 para o tipo 5). A Dim.3, associada a variáveis originais específicas , atua como principal discriminador desse cluster. O tipo 6, embora menos extremo na Dim.3 (Dim.3 = 0.08), compartilha com o tipo 5 valores altos na Dim.1 (1.92), sugerindo uma relação subjacente em outra dimensão. Os vidros desse tipo são chamados containers e provavelmente está associada a propriedades como resistência térmica ou composição química específica (alto teor de óxidos de cálcio ou magnésio), essenciais para garantir durabilidade em aplicações que exigem esterilização (embalagens) ou resistência a impactos (utensílios).

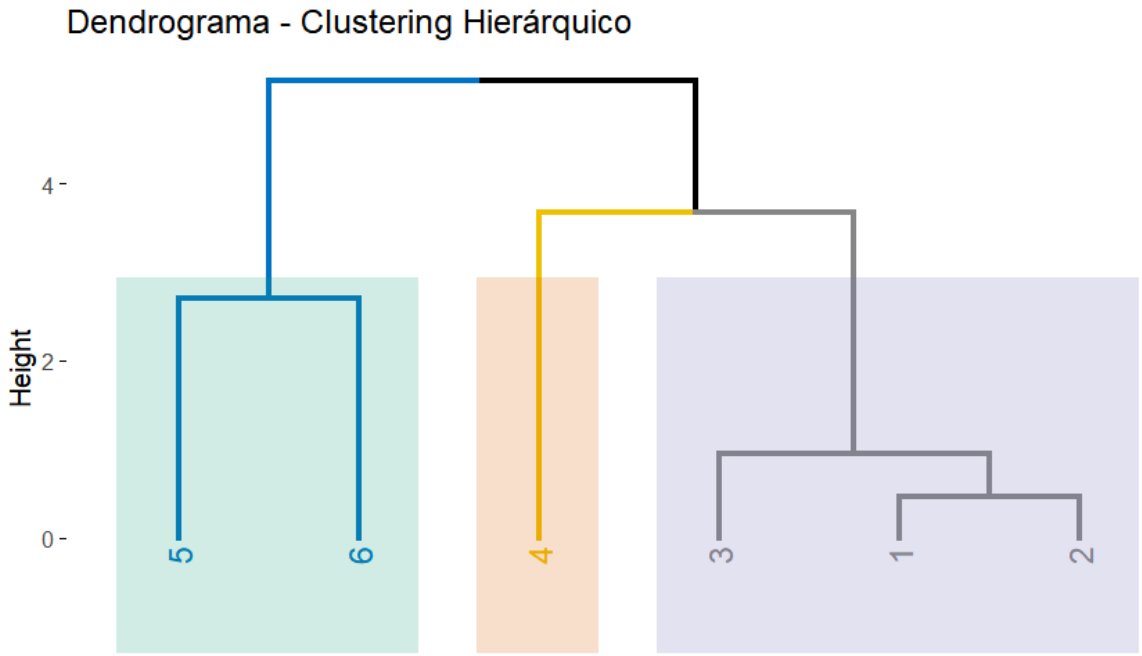


Figura 17: Dendograma

A proximidade entre esses tipos sugere que, apesar de usos distintos, compartilham requisitos técnicos similares. O cluster 2, isolado em um cluster único, que contém o tipo 4 ocupa uma posição intermediária na hierarquia, sugerindo que compartilha parcialmente características com outros grupos, mas mantém particularidades que justificam sua separação. Suas medianas nas componentes principais (Dim.1 = 0.62, Dim.2 = 0.44) indicam um perfil equilibrado, porém distinto — possivelmente associado a uma composição química ou propriedade física que não se alinha completamente com nenhum dos outros clusters. Essa singularidade pode refletir aplicações intermediárias ou uma transição na classificação dos vidros. O grupo 3, agrupa vidros destinados a janelas de edificações (tipos 1 e 2) e janelas de veículos (tipo 3), todos processados por técnica de float (exceto o tipo 2). A homogeneidade do grupo reflete a similaridade na composição química (alto teor de sílica e óxidos de sódio) e nas propriedades funcionais, como transparência, resistência mecânica e isolamento térmico. A pequena variação nas componentes principais (Dim.1 entre -0.83 e -0.92) indica que o processamento float (tipos 1 e 3) e não float (tipo 2) não altera significativamente o perfil multivariado desses vidros, justificando sua fusão em um único cluster. A proximidade entre eles no dendrograma reforça a validade da agregação hierárquica, indicando que pequenas variações nas variáveis originais não são suficientes para justificar sua separação em clusters distintos.

Referências

[1] John A. Hartigan e Manchek A. Wong. “A K-means Clustering Algorithm”. Em: *Applied Statistics* 28.1 (1979), pp. 100–108. DOI: 10.2307/2346830.

[2] Richard A. Johnson e Dean W. Wichern. *Applied Multivariate Statistical Analysis*. 6th. Pearson Prentice Hall, 2007. ISBN: 978-0131877153.

[3] Ian T. Jolliffe. *Principal Component Analysis*. 2nd. New York, NY: Springer Series in Statistics, 2002. DOI: 10.1007/b98835.

[4] Ian T. Jolliffe. *Principal Component Analysis*. 2nd. New York: Springer, 2002. DOI: 10.1007/b98835.

[5] Robert D. Koons, Charles Fiedler e Robert K. Rawley. “Classification and Discrimination of Glass Fragments Using Refractive Index and Elemental Composition”. Em: *Forensic Science International* 128.1-2 (2002), pp. 37–46. DOI: 10.1016/S0379-0738(02)00178-2.

[6] Pedro R. Peres-Neto, Donald A. Jackson e Keith M. Somers. “How many principal components? Stopping rules for determining the number of non-trivial axes revisited”. Em: *Computational Statistics & Data Analysis* 49.4 (2005), pp. 974–997. DOI: 10.1016/j.csda.2004.06.015.

[7] Qiang Xu. *Selected Methods of Cluster Analysis*. https://carleton.ca/math/wp-content/uploads/Xu-Qiang-Honours-Project-May-3rd-2021.pdf: SCHOOL OF MATHEMATICS e STATISTICS, 2021.

5 Apêndice - Códigos em R

```r


   # Dataset utilizado
   ------------------------------------------------------

glass <- read.csv("C:/Users/Luan Sousa/Downloads/glass.csv")
head(glass)

# Cores padronizadas
   ------------------------------------------------------
YlGnBu = c("#ffffd9", "#edf8b1", "#c7e9b4", "#7fcdbb", "#41
   b6c4",
           "#1d91c0", "#225ea8", "#253494", "#081d58")

PuBuGn = c("#0a9ad2", "#e0ebf4", "#c6dbef", "#9ecae1", "#6
   baed6", "#3182bd", "#08519c",
           "#98bba1", "#7bccc4", "#43a2ca", "#0868ac")


# Pacotes necessarios
   ------------------------------------------------------
load_packages <- function(packages) {
  # Verifica quais pacotes n o est o instalados
  not_installed <- packages[!(packages %in% installed.
   packages()[, "Package"])]

  # Instala os pacotes que n o est o instalados
  if (length(not_installed) > 0) {
    install.packages(not_installed)
  }

  # Carrega todos os pacotes
  invisible(lapply(packages, library, character.only = TRUE))
}

# Lista de pacotes pra carregar
my_packages <- c(
  "MASS", "klaR", "kmed", "dplyr", "tidyr", "HSAUR", "httpgd"
   , "plotly",
  "ggpubr", "ggplot2", "qqplotr", "caTools", "cluster", "
   biotools",
  "gridExtra", "factoextra", "FactoMineR"
)

# Chamada da fun   o para carregar os pacotes
```

```r
load_packages(my_packages)

head(glass)
features <- names(glass)[names(glass) != "Type"]
glass[features] <- lapply(glass[features], as.numeric)

glass$Type <- as.factor(glass$Type)

head(glass)


summary(glass)
unique(glass$Type)

bar_plot <- plot_ly(data = glass, x = ~Type,
                    color = ~Type, colors = "YlGnBu") %>%
  add_histogram() %>%
  layout(title = "Count of Types in glass Dataset",
         xaxis = list(title = "Type"), yaxis = list(title = "Count"))

bar_plot

(sum(glass$Type == 2) + sum(glass$Type == 1))/214

corr_matrix <- cor(glass[, 1:9])

heatmap <- plot_ly(
  x = colnames(corr_matrix),
  y = colnames(corr_matrix),
  z = corr_matrix,
  type = "heatmap",
  colorscale = "YlGnBu"
)

heatmap <- heatmap %>%
  add_annotations(
    x = rep(colnames(corr_matrix), each = length(colnames(corr_matrix))),
    y = rep(colnames(corr_matrix), length(colnames(corr_matrix))),
    text = round(corr_matrix, 2),
    showarrow = FALSE)

heatmap


anyDuplicated(glass)
```

```r
glass <- glass[!duplicated(glass),]

df_cat <- glass
# Categorize variable 'Ba'
df_cat$Ba <- ifelse(glass$Ba != 0.0, 1, 0)
df_cat$Ba <- as.factor(df_cat$Ba)

# Categorizando a  variable 'Fe'
df_cat$Fe <- ifelse(glass$Fe != 0.0, 1, 0)
df_cat$Fe <- as.factor(df_cat$Fe)

df_cat$Type <- as.factor(df_cat$Type)

head(df_cat)

num_features <- select_if(df_cat, is.numeric)

outliers <- lapply(num_features, function(feature) {
  q1 <- quantile(feature, 0.25)
  q3 <- quantile(feature, 0.75)
  iqr <- q3 - q1
  lower_bound <- q1 - 3 * iqr
  upper_bound <- q3 + 3 * iqr
  outlier_rows <- which(feature < lower_bound | feature > upper_bound)
  outlier_rows
})

outlier_indexes <- unique(unlist(outliers))

length(outlier_indexes)

df_cat_clean <- df_cat[-outlier_indexes,]


df <- glass
df_pca <- df[-c(8, 9)]

num_features <- select_if(df_pca, is.numeric)


outliers <- lapply(num_features, function(feature) {
  q1 <- quantile(feature, 0.25)
  q3 <- quantile(feature, 0.75)
  iqr <- q3 - q1
  lower_bound <- q1 - 3 * iqr
  upper_bound <- q3 + 3 * iqr
  outlier_rows <- which(feature < lower_bound | feature >
```

```
    upper_bound)
  outlier_rows
})


outlier_indexes <- unique(unlist(outliers))

length(outlier_indexes)


df_pca <- df_pca[-outlier_indexes,]

df_pca == df_cat_clean[,-c(8,9)]
# same df

pca <- PCA(df_pca, scale = T, quali.sup = 8, quanti.sup = 1,
    graph=T)
pca

df_num  <- subset(df_cat, select = -c(Ba, Fe, Type))
summary(df_num)
var(df_num$RI)

d <- dist(df_num, method = "euclidean") # distance matrix
fit <- hclust(d, method="single")
plot(fit, main = "Dendrogram of Single Linkage", labels =
    FALSE)
fit1 <- hclust(d, method="complete")
plot(fit1, main = "Dendrogram of complete Linkage", labels =
    FALSE)
fit2 <- hclust(d, method="average")
plot(fit2, main = "Dendrogram of Average Linkage", labels =
    FALSE)
fit3 <- hclust(d, method="ward.D2")
plot(fit3, main = "Dendrogram of Ward Method", labels = FALSE
    )
fit4 <- hclust(d, method="centroid")
plot(fit4, main = "Dendrogram of Centroid Method", labels =
    FALSE) # Dendogram


plot(fit3,main="Dendrogram Ward Method Linkage", labels=FALSE
    , xlab="", sub="")
groups <- cutree(fit3, k=2 )# c
rect.hclust(fit3, k=2, border="red")

groups <- cutree(fit3, k=3 )# c
rect.hclust(fit3, k=3, border="blue")
```

```
171
172 groups <- cutree(fit3, k=4 )# c
173 rect.hclust(fit3, k=4, border="orange")
174
175 groups <- cutree(fit3, k=5 )# c
176 rect.hclust(fit3, k=5, border="green")
177
178 groups <- cutree(fit3, k=6 )# c
179 rect.hclust(fit3, k=6, border="purple")
180
181
182 aux<-c()
183 for (i in 1:dim(df_num)[2]) {
184   k <- kmeans(df_num, centers = i, nstart = 25)
185   aux[i] <- k$tot.withinss
186 }
187 plot(aux, xlab="Number of Clusters", ylab="TWSS", type="l",
     main="TWSS vs. number of clusters", col=YlGnBu[6], lwd=3)
188
189 fviz_nbclust(df_num, kmeans, method = "silhouette")
190 fviz_nbclust(df_num, kmeans, method = "gap_stat")
191
192
193 res <- fastkmed(d, 6)
194 silhouette <- sil(d, res$medoid, res$cluster)
195
196 silhouette$plot
197
198
199 cluster = cutree(fit3,3)
200
201 p <- ggplot(data = df_num, aes(x = cluster, fill = glass$Type
     )) +
202   geom_bar() +
203   theme_bw() + labs(title="", fill = "Type") + scale_fill_
     brewer(palette = 'YlGnBu')
204
205 plot(p)
206
207 k3 <- kmeans(d, centers = 3, nstart = 25)
208 str(k3)
209 names(k3)
210
211 aggregate(df_num,by=list(k3$cluster),FUN=mean)
212 df_num$cluster<-as.numeric(k3$cluster)
213
214 k3$withinss
215 k3$totss
```

```r
k3$tot.withinss
k3$betweenss + k3$tot.withinss # BSS + Wss


fviz_cluster(k3, data = df_num,
             palette = c(PuBuGn[2], PuBuGn[6], PuBuGn[9]),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw()
)


names(k3)
clusplot(df_num, k3$cluster, color=TRUE, shade=TRUE,
         labels=2, lines=0)

c3<-clara(df_num,3)
names(c3)

clusplot(df_num, c3$cluster, color=TRUE, shade=TRUE,
         labels=2, lines=0)



fviz_cluster(k3, data = df_num,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw(),
)


p <- ggplot(data = df_num, aes(x = k3$cluster, fill = glass$
    Type)) +
  geom_bar() +
  theme_bw()

print(p)


df_sum <- df %>%
  rowwise() %>%
  mutate(Sum = sum(c_across(-Type))) %>%
  ungroup()

distance_matrix <- dist(df_sum$Sum, method = "euclidean") #
    Dist ncia Euclidiana

```

```r
hc <- hclust(distance_matrix, method = "complete")

plot(hc, labels = F, main = "Dendrograma - Cluster Hier rquico",
     xlab = "Tipo de Vidro", ylab = "Dist ncia", cex = 0.7)


clusters <- cutree(hc, k = 3)
df_sum <- df_sum %>%
  mutate(Cluster = clusters)

# Visualizar a distribui  o dos clusters
a = table(df_sum$Cluster)
a


####

# Passo 1: Calcular as medianas por tipo de vidro (excluindo a vari vel Fe)
glass_grouped <- df %>%
  group_by(Type) %>%
  summarise(across(everything(), min, na.rm = TRUE)) %>%
  ungroup()

# Passo 2: Normalizar os dados (excluindo a coluna 'Type')
normalized_data <- glass_grouped %>%
  select(-Fe,-Ba, -RI) %>%
  select(-Type) %>%
  scale()

# Passo 3: Calcular a matriz de dist ncias
distance_matrix <- dist(normalized_data, method = "manhattan") # Dist ncia Euclidiana

# Passo 4: Aplicar o cluster hier rquico
hc <- hclust(distance_matrix, method = "ward.D2")

# Passo 5: Visualizar o dendrograma
fviz_dend(hc,
          k = 3, # N mero de clusters
          cex = 0.7,
          palette = "jco",
          rect = F,
          rect_border = "jco",
          rect_fill = F,
          labels_track_height = 0.8,
          main = "Dendrograma - Similaridade entre Tipos de
```

```
   Vidro (sem Fe)")

# Passo 6: Cortar o dendrograma e visualizar os clusters
clusters <- cutree(hc, k = 3)

# Mapear os clusters aos tipos de vidro
glass_grouped <- glass_grouped %>%
  mutate(Cluster = clusters)

# Visualizar os resultados
glass_grouped %>% arrange(Cluster)

library(MVN)
library(GGally)
ggpairs(glass[,-10])

mardialT = mvn(glass[, -10], mvnTest = 'mardia',
   univariatePlot = "scatter")
mardialT




df2_clean <- df2[, !(colnames(df2) %in% c("Mg", "K", "Ba", "
   Fe", "Type"))]

mvn(df2_clean, mvnTest = 'mardia', bc = T, univariatePlot   =
    'histogram', bcType = 'optimal')


kmeans_result <- kmeans(normalized_data, centers = 3, nstart
   = 25) # N mero de clusters desejado: 3
fviz_cluster(kmeans_result,
             data = normalized_data,
             palette = "jco", # Paleta de cores
             geom = "point", # Mostrar pontos
             ellipse.type = "convex", # Adicionar elipses
   convexas para os clusters
             ggtitle = "Cluster K-means - Tipos de Vidro (sem
    Fe)")


summary(pca)

pca_data <- pca$ind$coord[, 1:3]

head(pca_data)
```

```r

# Calcular a matriz de dist ncias
distance_matrix <- dist(pca_data, method = "euclidean")
hc <- hclust(distance_matrix, method = "ward.D2")


pca_data_with_type <- data.frame(pca_data, Type = df_pca$Type
  )

pca_medians <- pca_data_with_type %>%
  group_by(Type) %>%
  summarise(across(everything(), median, na.rm = TRUE)) %>%
  ungroup()
head(pca_medians)

distance_matrix <- dist(pca_medians[, -1], method = "
  euclidean")  # Excluindo a coluna 'Type'
hc <- hclust(distance_matrix, method = "ward.D2")
hc$height
hc$order
hc$call




  fviz_dend(hc,
            k = 3,  # N mero de clusters desejado
            cex = 0.7,  # Tamanho do texto
            palette = "jco",
            rect = F,
            rect_border = "jco",
            rect_fill = F,
            labels_track_height = 0.8,
            main = "Dendrograma - Clustering Hier rquico")



  fviz_dend(hc,
            k = 3, # N mero de clusters desejado
            cex = 0.7, # Tamanho do texto
            palette = "jco",
            rect = F,
            rect_border = "jco",
            rect_fill = F,
            labels_track_height = 0.8,
            main = "Dendrograma - Similaridade entre Tipos de
  Vidro (sem Fe)")
```

```
393
394
395 fviz_dend(hc,
396           k = 3,
397           cex = 0.9,
398           palette = "jco",
399           rect = TRUE,
400           rect_border = "jco",
401           rect_fill = TRUE,
402           labels_track_height = 0.8,
403           main = "Dendrograma - Clustering Hier rquico",
404           lwd = 1.2,
405           horiz = F)  # Para manter a orienta  o vertical
```