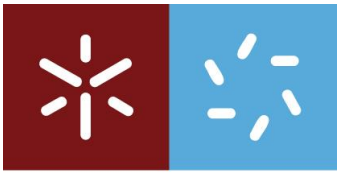

Universidade do Minho
Escola de Ciências

Análise Discriminante, Análise de Componentes Principais e Análise de Clusters da Base de Dados Diamonds

Mestrado em Estatística para a Ciência de Dados
Análise Estatística Multivariada

Ano Letivo 2024/2025
Anita Margarida Antunes Ferreira - pg56093
Inês Margarida Gonçalves Gomes - pg55575
Rui Miguel Pereira Alves - pg55577

Guimarães, abril de 2025

Índice

Índice de figuras

Índice de tabelas

1 Introdução

Num contexto onde a informação é vasta, multidimensional e profundamente interligada, a estatística multivariada surge como uma ferramenta indispensável para revelar estrutura, reduzir complexidade e transformar dados em conhecimento. O presente relatório, realizado no âmbito da unidade curricular de Análise Estatística Multivariada do Mestrado em Estatística para a Ciência de Dados, centra-se na análise da base de dados *diamonds*, composta por variáveis quantitativas e qualitativas que descrevem atributos físicos e comerciais de diamantes, com o objetivo de compreender os fatores que mais contribuem para a sua valorização.

A partir de uma amostra aleatória de 200 observações, recorre-se a técnicas clássicas da Análise Multivariada, como a Análise de Componentes Principais (ACP), a Análise Discriminante e a Análise de Clusters. Estas metodologias permitem não só condensar a informação em eixos de variabilidade máxima, como também identificar padrões ocultos e segmentar grupos de diamantes com características similares. Para além da interpretação estatística, privilegia-se uma análise visual clara e fundamentada, capaz de traduzir relações complexas em representações intuitivas. Este trabalho pretende, assim, ilustrar o poder da análise multivariada na extração de sentido e estrutura em contextos reais, conferindo aos dados um novo nível de legibilidade e relevância.

Por conseguinte, os capítulos que se seguem estão estruturados de forma a refletir as diferentes etapas da análise desenvolvida: após uma exploração inicial dos dados, procede-se à Análise Discriminante, que avalia a capacidade preditiva das variáveis qualitativas. Segue-se a Análise de Componentes Principais (ACP), aplicada às variáveis quantitativas, com o objetivo de reduzir a dimensionalidade dos dados. Por fim, realiza-se a Análise de Clusters, com o intuito de detetar grupos naturais de diamantes com características semelhantes. Cada metodologia é acompanhada de interpretação estatística rigorosa e visualização gráfica, promovendo uma leitura clara, coerente e integrada dos resultados.

2 Descrição dos Dados

A base de dados utilizada neste trabalho é designada por *diamonds* e está disponível na biblioteca *ggplot2* da linguagem R. Esta base contém informações detalhadas sobre 53.940 diamantes de corte redondo, sendo amplamente utilizada em exemplos de visualização e análise de dados devido à sua riqueza e diversidade de variáveis.

Cada observação representa um diamante individual, e as variáveis associadas descrevem tanto as suas características físicas como qualitativas, incluindo o preço, o peso, a qualidade do corte, a cor, a clareza e as suas dimensões.

De forma a tornar a análise mais eficiente e computacionalmente viável, foi extraída uma amostra aleatória simples composta por 200 observações da base de dados original. Para garantir a reprodutibilidade da amostragem, foi definida uma semente aleatória com o valor 3030.

A base de dados apresenta um total de 10 variáveis, que podem ser agrupadas da seguinte forma:

- **Variáveis quantitativas**:
  - *price*: preço do diamante em dólares americanos (US$), variando entre 326 e 18.823.
  - *carat*: peso do diamante, em quilates, com valores entre 0.2 e 5.01.
  - x: comprimento do diamante (em milímetros), variando entre 0 e 10.74.
  - y: largura do diamante (em milímetros), variando entre 0 e 58.9.
  - z: profundidade do diamante (em milímetros), com valores entre 0 e 31.8.
  - *depth*: profundidade total em percentagem, calculada como $\frac{2 \cdot z}{x+y}$, com valores entre 43% e 79%.
  - *table*: largura da face superior do diamante em percentagem relativa ao ponto mais largo, variando entre 43% e 95%.
- **Variáveis qualitativas (categóricas)**:
  - *cut*: qualidade do corte, com cinco níveis ordenados: *Fair*, *Good*, *Very Good*, *Premium* e *Ideal*.
  - *color*: grau de cor do diamante, com níveis que vão de *D* (melhor qualidade) até *J* (pior qualidade).
  - *clarity*: grau de clareza, com oito níveis ordenados: *I1*, *SI2*, *SI1*, *VS2*, *VS1*, *VVS2*, *VVS1* e *IF*, do menos puro ao mais puro.

Esta estrutura rica em variáveis torna o conjunto de dados apropriado para a aplicação de técnicas de análise estatística multivariada como MANOVA, análise discriminante, análise em componentes principais (PCA) e análise de clusters, tal como será desenvolvido nas secções seguintes deste trabalho.

3 Análise Exploratória dos dados

> ”Oh my God, this data is not what I thought it would be!”So already, you’ve discovered something.” ”
>
> Martin Wattenberg

A finalidade deste capítulo é examinar os dados previamente à aplicação de qualquer técnica estatística. Deste modo, é possível um entendimento preliminar dos dados e das relações existentes entre as variáveis analisadas.

3.1 Análise exploratória univariada

A análise exploratória univariada irá consistir em métodos de Estatística Descritiva que permitem a análise de cada variável separadamente. Dado isto, na tabela 3.1 estão apresentadas as estatísticas descritivas relativamente às variáveis quantitativas em estudo.

Tabela 3.1: Estatísticas Descritivas para as variáveis quantitativas

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Peso do diamante (*carat*)	0.2300	0.4100	0.7250	0.8118	1.0400	2.2000
Percentagem total da profundidade (*depth*)	57.50	61.10	61.90	61.83	62.60	66.50
Largura do topo do diamante relativo ao ponto mais largo (*table*)	53.00	55.92	57.00	57.20	59.00	66.00
Preço de cada diamante (*price*)	402	1076	2740	3912	5451	17609
Comprimento (*x*)	3.960	4.750	5.780	5.780	6.503	8.510
Largura (*y*)	3.990	4.768	5.790	5.777	6.510	8.550
Profundidade (*z*)	2.410	2.950	3.550	3.573	4.050	5.230

Analisando a tabela, sumariza-se o seguinte:

- O peso de cada diamante varia entre [0.2300 , 2.2000] carat;
- O valor mais elevado pago por um diamante foi de $17609 e o mais baixo foi de $402. Além disto, em média, o preço de cada diamante foi de $3912;
- O comprimento do diamante varia entre [3.960 , 8.510] mm;
- A largura de cada diamante varia entre [3.990 , 8.550] mm;
- A profundidade de cada diamante varia entre [2.410 , 5.230] mm.

Em relação às variáveis qualidade de corte (*cut*), cor do diamante (*color*) e claridade (*clarity*) observe-se como se distribuem quanto às suas frequências nas figuras 3.1,3.2 e 3.3.

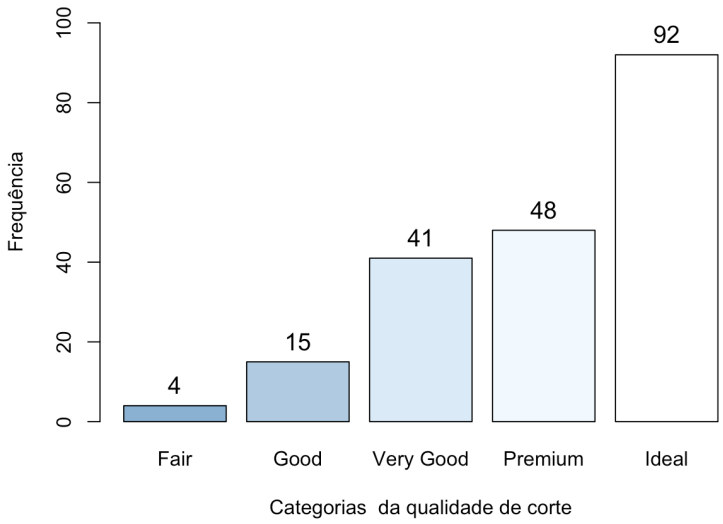

Figura 3.1: Gráfico de barras: Qualidade do corte do diamante

Na figura 3.1, é possível observar que a qualidade de corte da maioria dos diamantes é classificada como ideal. Além disto apenas 4 diamantes tem uma lapidação pobre.

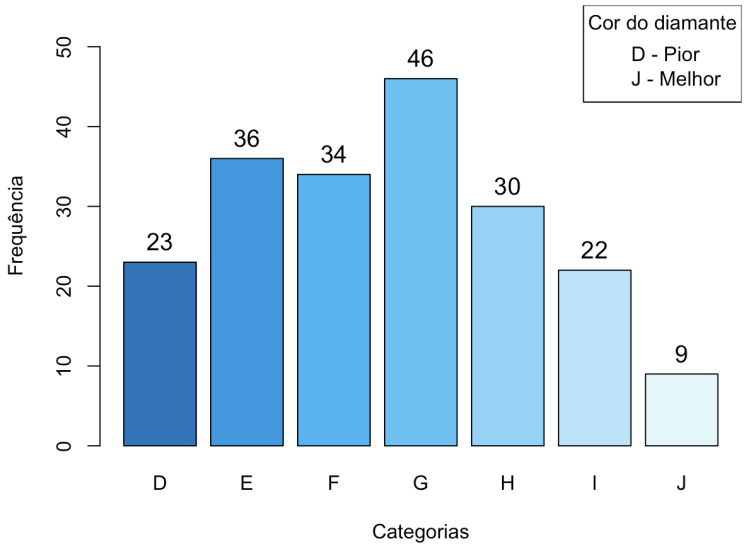

Figura 3.2: Gráfico de barras: Cores do diamante

Nesta amostra, são escassos os diamantes com maior qualidade de cor. Maioritariamente, a cor dos diamantes encontra-se entre *E* a *G*, o que sugere que diamantes com qualidade de cor superior, entre *H-J*, são uma minoria no mercado.

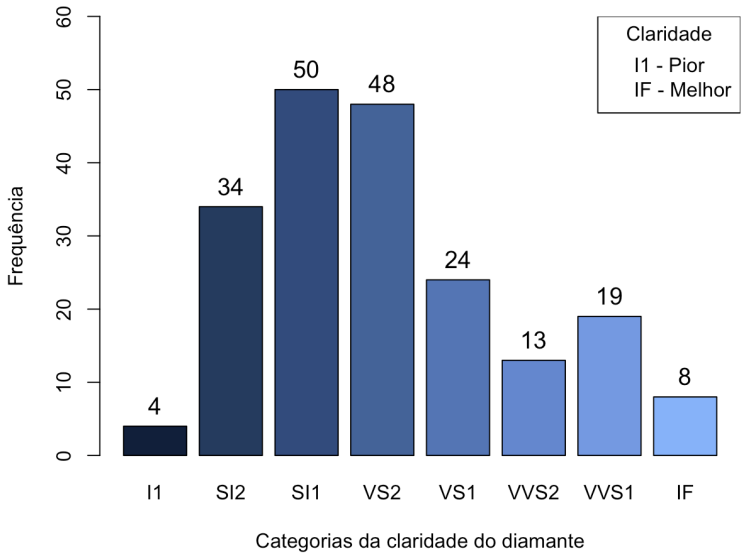

Figura 3.3: Gráfico de barras: Claridade do diamante

Apenas existem 4 diamantes com a pior classificação de claridade (*I1*) e 8 diamantes com a melhor classificação de claridade (*IF*). A maioria dos diamantes está classificada entre uma claridade baixa a média.

3.2 Análise exploratória bivariada

A análise exploratória bivariada inclui métodos de análise de duas variáveis de forma a compreender melhor as relações existentes entre as variáveis em estudo.

A análise bivariada das variáveis visa explorar como as diferentes características dos diamantes, como peso, corte, cor e a claridade, se relacionam com o preço. Ao investigar essas relações, pretende-se entender como cada fator contribui para a valorização do diamante. Assim, foram realizadas diversas inspeções gráficas com o intuito de sugerir respostas a este propósito.

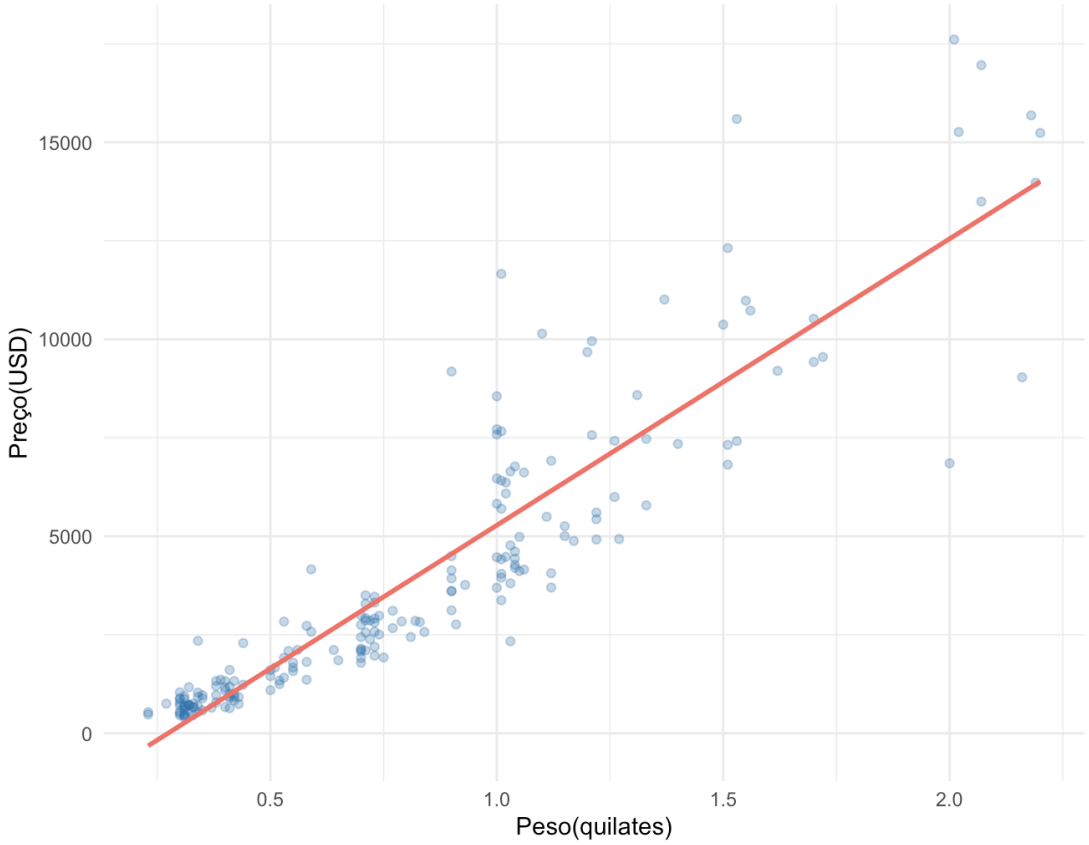

Figura 3.4: Preço do diamante por peso

Primeiramente, na figura 3.4 o gráfico revela uma relação positiva entre o peso do diamante e o seu preço, com uma tendência aproximadamente linear. Nota-se também uma maior dispersão dos preços à medida que o peso aumenta, o que sugere maior variabilidade a partir de 1 quilate.

Relacionar o preço de cada diamante com a qualidade do corte, o preço e a cor do mesmo permite uma análise mais enriquecedora neste estudo. Dado isto, atente-se à figura 3.5.

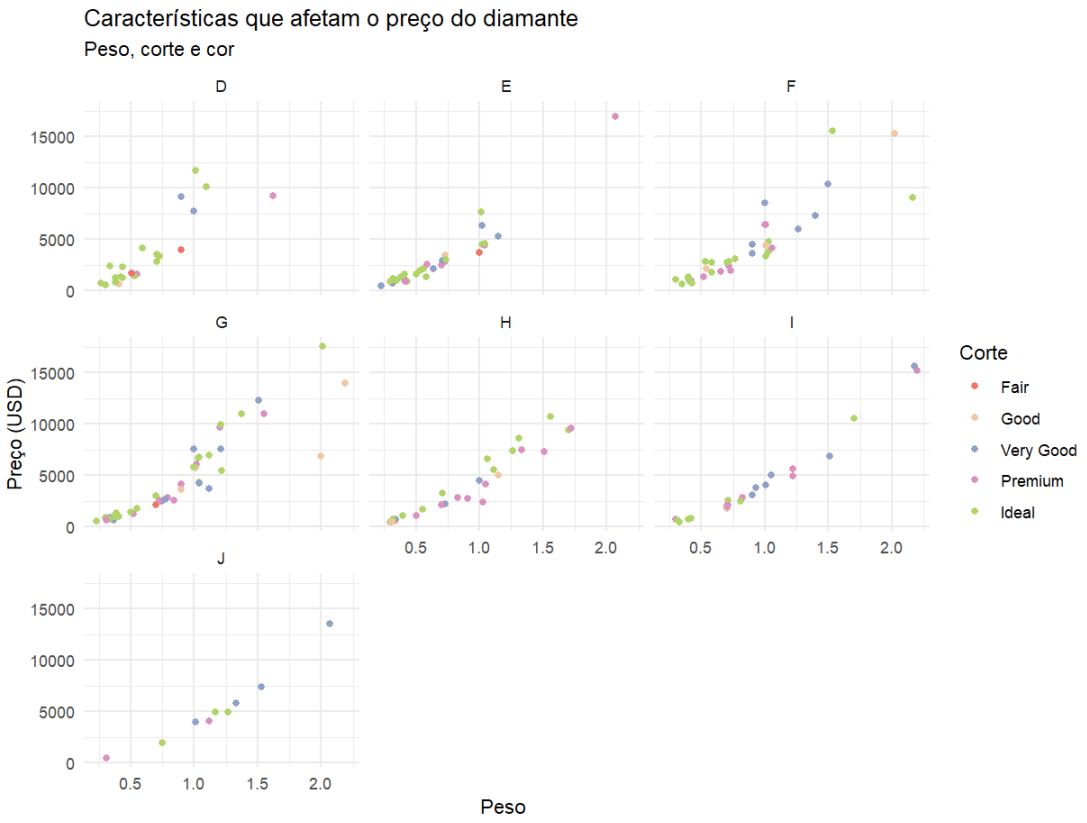

Figura 3.5: Características que afetam o preço do diamante: Peso, corte e cor

Relembre-se que, a cor de cada diamante é classificada de *D* a *J*, sendo que *D* indica a melhor cor e *J* a pior. Assim, indica que:

- Diamantes com a melhor cor (*D*) possuem qualidade de corte ideal maioritariamente em pesos baixo. Além disto, esta qualidade de corte é a mais observada;
- O diamante mais pesado, com a melhor classificação de cor tem qualidade de corte *premium*. Contudo, também é possível observar diamantes com a melhor qualidade de cor e a pior classificação de corte;
- O número de diamantes com a pior cor (*J*) é mais reduzido e observa-se as qualidades de corte ideal, muito boa e premium. Assim, os diamantes com pior classificação de cor não apresentam nesta amostra qualidade de corte pobre;
- De forma geral, está presente em todas as categorias de cor todos os tipos de corte do diamante;
- O gráfico sugere que existe uma predominância do corte ideal pelas diversas cores dos diamantes.

A figura 3.6 relaciona o preço de cada diamante com o seu peso, corte e claridade.

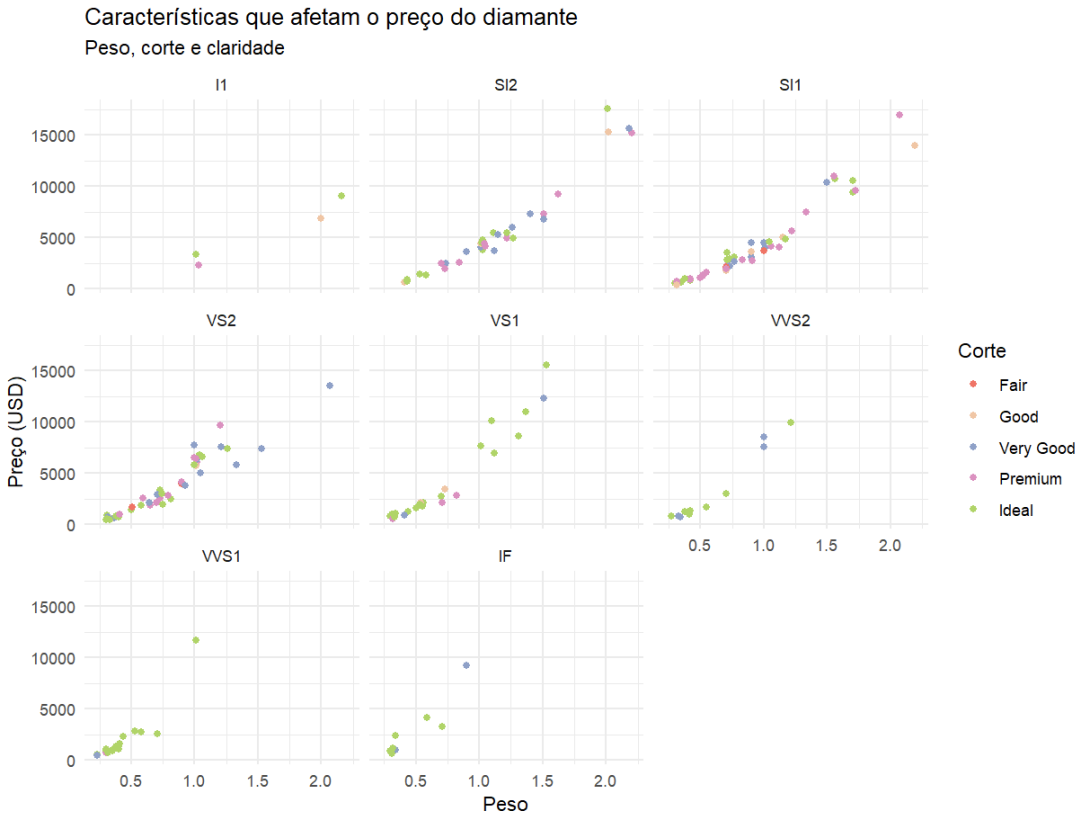

Figura 3.6: Características que afetam o preço do diamante: Peso, corte e claridade

Em relação à claridade do diamante relembre que, *I1* indica a pior claridade e *IF* a melhor claridade. Assim, a figura 3.6 indica o seguinte:

- Só observamos cortes ideais e muito bons em diamantes com a pior claridade (*IF*, *WS1* e *WS2*). Além disto, estes diamantes tem pesos baixos;
- Diamantes com a melhor claridade (*I1*) apresentam cortes ideais, bons e premium;
- As claridades *SI2*, *SI1*,*VS2* E *VS1* apresentam dispersos pesos e de modo geral todos os tipos de corte.

A distribuição do peso em relação à qualidade de corte de cada diamante indica a quantidade de diamantes em cada uma das classes. Dado isto, atente-se à figura 3.7.

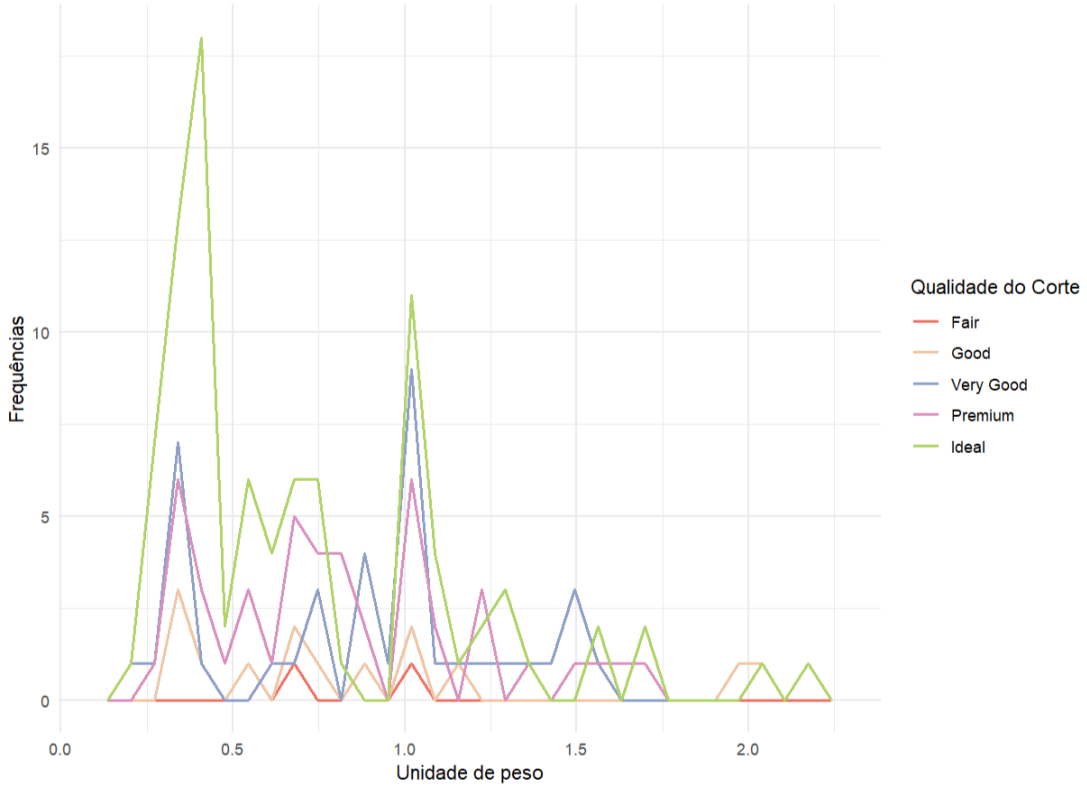

Figura 3.7: Frequências da qualidade de corte por peso (carat)

A figura 3.7 sugere o seguinte:

- Diamantes mais leves, tem maioritariamente uma qualidade de corte classificada como ideal ;
- Os diamantes mais pesados apresentam cortes pobres, ideias ou bons;
- A classificação de corte ideal observa-se em qualquer intervalo de peso e a classificação de um corte pobre é o menos observado;
- De modo geral, as diversas qualidades de corte encontram-se nos diversos pesos.

Nas figuras 3.8,3.9 e 3.10 é possível observar como, respetivamente, o corte, cor e claridade dos diamantes se distribuem quanto ao preço.

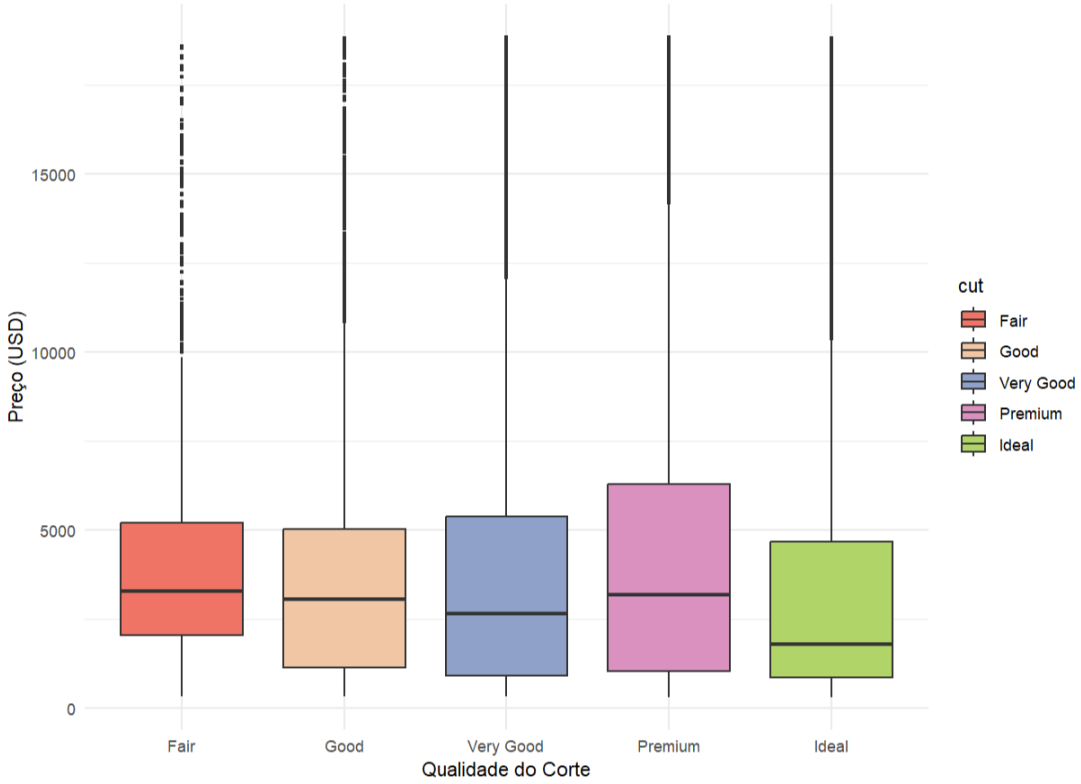

Figura 3.8: Boxplot do preço por qualidade de corte

A figura 3.8 sugere o seguinte:

- Diamantes com qualidade de corte (*cut*) premium são aqueles que apresentam uma maior dispersão de preços e os que podem atingir preços mais elevados;
- A qualidade de corte ideal é aquela que apresenta diamantes com preços mais baixos.
- A mediana do preço dos diamantes com qualidade de corte pobre e boa é aproximadamente a mesma. Além disto, diamantes com qualidade de corte pobre tem uma menor dispersão de preços;
- Nas diversas qualidades de corte são visíveis *outliers*.

Atente-se à próxima figura.

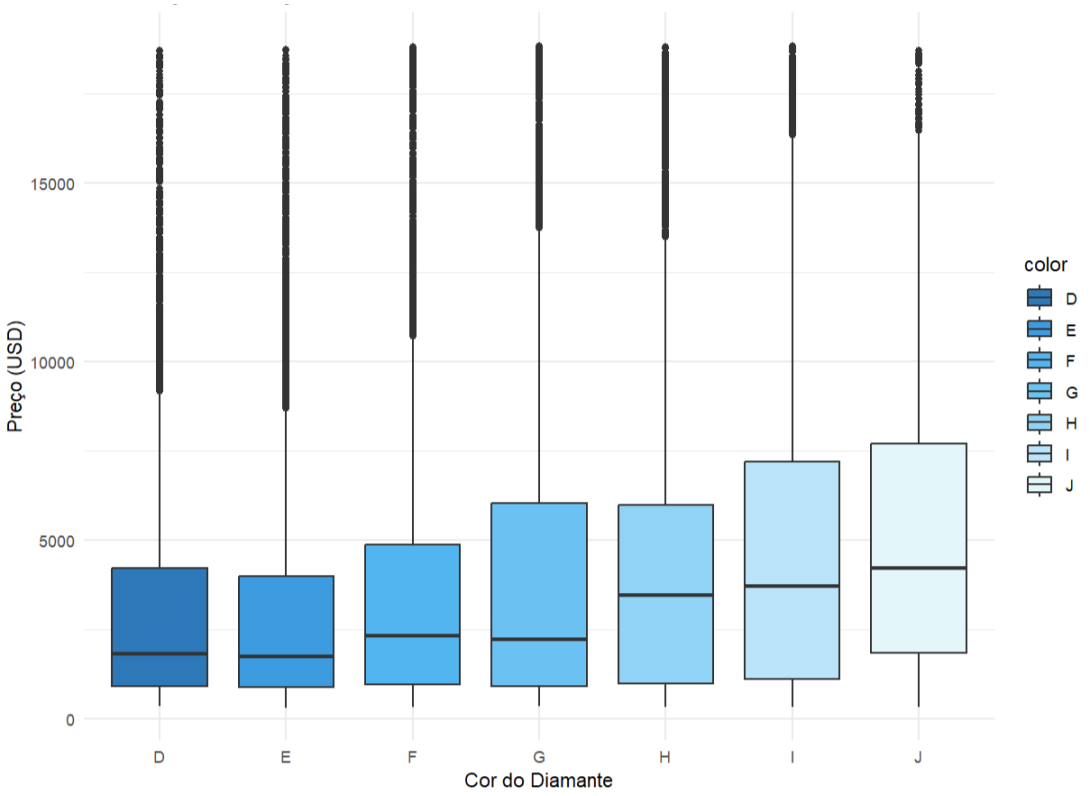

Figura 3.9: Boxplot cor do diamante e o preço

A figura 3.9 sugere o seguinte:

- Existe uma maior dispersão de preços nas cores *G* a *J*. Note-se que *J* indica a pior cor;
- Diamantes com melhor coloração (*D* e *E*) apresentam preços mais baixos e menor dispersão dos mesmos;
- A mediana do preço é mais elevada em diamantes com pior cor. Isto sugere que, diamantes com preços mais elevados podem não apresentar a melhor coloração.
- São observados outliers nas diversas colorações, sendo que, as piores colorações revelam uma quantidade mais reduzida dos mesmos.

Por último, nesta secção, observe-se a figura 3.10 que descreve como se distribui o preço dos diamantes em relação à sua claridade.

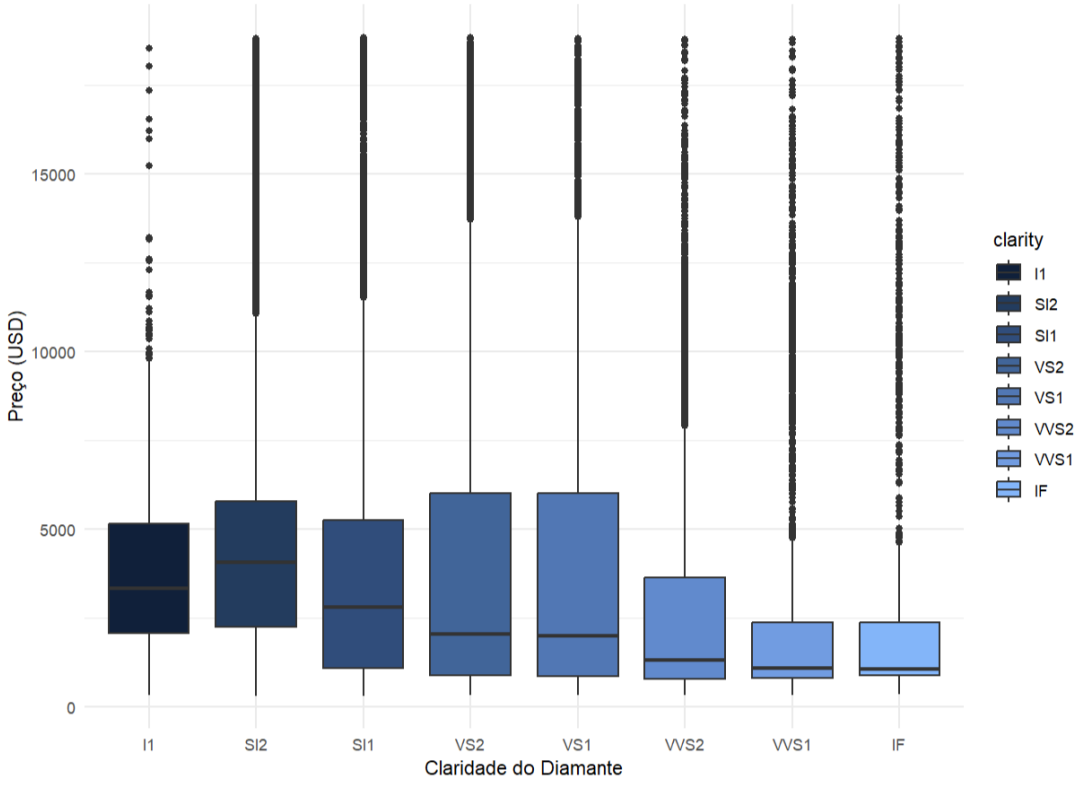

Figura 3.10: Boxplot da claridade e preço do diamante

A figura 3.10 sugere o seguinte:

- A melhor claridade (*IF*) revela uma baixa dispersão de preços e os mesmos pouco elevados;
- A pior claridade (*I1*) apresenta preços superiores aos diamantes com melhor claridade (*IF*);
- Diamantes com claridades intermédias (*VS2* e *VS1*) são os que apresentam uma maior dispersão de preços;
- Em mediana, os preços são mais elevados em claridades baixas. A claridade *SI2* é a que revela os preços mais elevados em relação à mediana dos mesmos;
- Existem *outliers* nas diversas claridades, sendo menos presentes na pior claridade (*I1*).

3.3 Medidas descritivas para amostras multivariadas

Muita informação contida num conjunto de dados pode ser aferida através do cálculo de medidas de estatística descritiva. Assim, é fundamental a descrição da informação através do cálculo de parâmetros de localização, dispersão e associação linear de uma população multivariada.

As 200 observações constituem um subconjunto de todas as observações para a população em estudo, as estatísticas apresentadas para estimar os parâmetros da população são designadas de estimadores. Para uma população multivariada, os estimadores dos parâmetros são os seguintes:

Vetor de médias amostrais

Dado o conjunto das 7 variáveis aleatórias quantitativas da base de dados, o vetor valor médio

$$\mu = (\mu_1, \mu_2, \ldots, \mu_7)^T. \tag{1}$$

Uma estimativa para μ baseada no estimador usual $E(X)$, é dada por

$$\bar{\mathbf{X}}_{7\times 1} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \\ \bar{X}_6 \\ \bar{X}_7 \end{pmatrix} = \begin{pmatrix} 0.812 \\ 61.825 \\ 57.197 \\ 3912.245 \\ 5.780 \\ 5.777 \\ 3.573 \end{pmatrix} \tag{2}$$

A $\bar{\mathbf{X}}_{7\times 1}$, apresentado em (2), designamos de vetor das médias ou centróide. Analisando os resultados obtidos, para o vetor das médias, é possível constatar o seguinte:

- Em média, cada diamante pesa 0.812 carat;
- A média dos preços de cada diamante é de $3912.245;
- O comprimento, largura e profundidade de cada diamante é respetivamente, em média, 5.780 mm, 5.777 mm e 3.573 mm. Note-se que a percentagem total da profundidade, em média, é de 61.825% e ainda que, a largura do topo do diamante relativo ao ponto mais largo em percentagem é de 51.197%.

Matriz de variância/covariância amostral

A matriz de covariâncias populacional, simétrica definida positiva, tem como estimador a matriz **S** cujo elemento s_{kj} representa a covariância amostral das variáveis k e j. Assim, temos que,

$$\mathbf{S}_{p\times p} = \begin{pmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_{22} & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_{33} & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_{pp} \end{pmatrix} =$$

$$\mathbf{S}_{7\times 7} = \begin{pmatrix} 0.215 & & & & & & \\ 0.021 & 1.593 & & & & & \\ 0.161 & -1.226 & 4.702 & & & & \\ 1562.733 & -192.613 & 937.549 & 13634098.005 & & & \\ 0.497 & -0.009 & 0.445 & 3562.454 & 1.199 & & \\ 0.493 & -0.015 & 0.436 & 3549.579 & 1.189 & 1.181 & \\ 0.307 & 0.087 & 0.200 & 2184.397 & 0.738 & 0.732 & 0.460 \end{pmatrix} \tag{3}$$

Ora, a covariância, uma medida de associação linear entre duas variáveis, torna-se difícil de interpretar uma vez que depende das unidades em que as duas variáveis foram medidas. Neste caso, como as variáveis comprimento, largura e profundidade são todas medidas em mm, observamos associações lineares positivas entre elas.

Matriz de correlações amostrais

A estandardização da covariância dá origem ao coeficiente de correlação de Pearson. Ora, o coeficiente de correlação é uma medida de associação linear entre variáveis quantitativas e toma valores entre -1 e 1 inclusive. Note-se que, a matriz de correlação **R** pode ser obtida à custa da matriz de covariância populacional e ainda que, com p variáveis existem $p(p-1)/2$ valores distintos de correlações que se arranjam na matriz de correlações. Assim, a matriz de correlação, simétrica definida positiva, é dada por:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix} =$$

$$\mathbf{R} = \begin{pmatrix} 1 & & & & & & \\ 0.036 & 1 & & & & & \\ 0.160 & -0.448 & 1 & & & & \\ 0.913 & -0.041 & 0.117 & 1 & & & \\ 0.979 & -0.006 & 0.188 & 0.881 & 1 & & \\ 0.978 & -0.011 & 0.185 & 0.885 & 0.999 & 1 & \\ 0.978 & 0.102 & 0.136 & 0.873 & 0.994 & 0.993 & 1 \end{pmatrix} \tag{4}$$

A matriz de correlações amostrais, **R**, indica o seguinte :

- Existe uma forte correlação positiva entre o preço do diamante e o seu tamanho (0.913). Além disto, o comprimento, a largura e a profundidade também estabelecem uma forte correlação positiva com o preço de cada diamante;

- Expectavelmente, existem fortes correlações positivas com o peso do diamante e a seu comprimento, largura e profundidade (0.979, 0.978 e 0.978 respetivamente). Na prática, quanto mais pesado o diamante maior será o seu comprimento, largura e profundidade;

• O aumento da percentagem total de profundidade implica a diminuição da largura do topo do diamante relativo ao ponto mais largo e vice-versa;

• Existe uma correlação positiva, aproximadamente, perfeita entre o comprimento e a largura, o comprimento e a profundidade e a largura e a profundidade.

Graficamente, o gráfico de dispersão seguinte, sugere as conclusões anteriores.

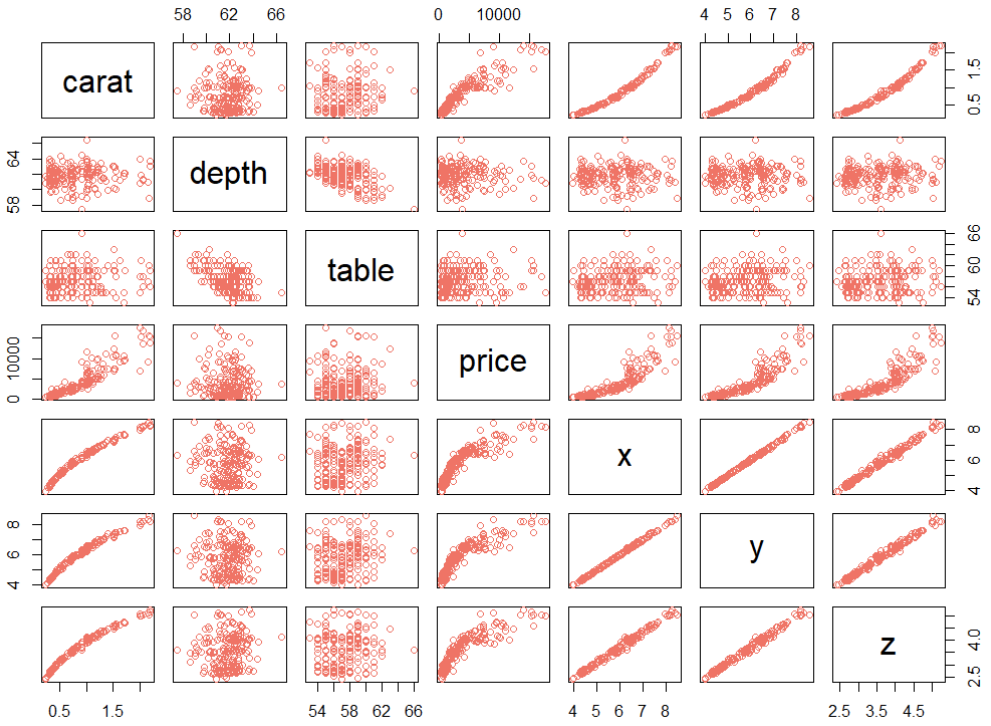

Figura 3.11: Gráfico de dispersão.

4 Inferência Estatística Multivariada

> "Statistics is the science of uncertainty, the art of making inferences in the face of randomness."
>
> David S. Salsburg

A inferência estatística multivariada é uma abordagem essencial na análise de dados que envolvem diversas variáveis dependentes. Ao contrário dos métodos univariados, que analisam cada variável de forma isolada, os testes multivariados permitem avaliar simultaneamente as relações entre variáveis, identificando padrões complexos nos dados. Neste capítulo, iremos explorar a teoria e a aplicação dos testes de hipóteses multivariados, com especial foco na Análise Multivariada de Variância (MANOVA). Os testes de hipóteses desempenham um papel fundamental na estatística, pois permitem validar suposições sobre populações com base em amostras. No contexto multivariado, esses testes são utilizados para determinar se as diferenças entre grupos são estatisticamente significativas quando analisamos um conjunto de variáveis dependentes em simultâneo. A MANOVA é uma extensão da ANOVA que possibilita investigar se uma ou mais variáveis independentes categóricas exercem influência sobre múltiplas variáveis dependentes de forma conjunta. Além disso, discutiremos os pressupostos fundamentais da MANOVA, tais como a normalidade multivariada e a homogeneidade das matrizes de covariância.

4.1 Testes de Hipóteses

Um dos primeiros passos na análise de dados multivariados é testar a normalidade multivariada, uma condição fundamental para muitos testes estatísticos paramétricos. Para isso, utilizamos o teste que avalia se as variáveis seguem uma distribuição normal conjunta. Esse teste é essencial antes da aplicação de testes de hipóteses multivariados, pois muitos deles assumem normalidade para garantir a validade dos resultados. No contexto dos testes de hipóteses, o teste de normalidade multivariada pode ser formulado da seguinte forma:

H_0*: Os dados populacionais seguem uma distribuição normal multivariada.*
versus
H_1*: Os dados populacionais não seguem uma distribuição normal multivariada.*

Os resultados do teste de normalidade multivariada indicam um valor de prova muito baixo, rejeitando a hipótese de que os dados seguem uma distribuição normal conjunta. Isso sugere a presença de desvios significativos da normalidade, o que pode impactar métodos estatísticos que assumem essa propriedade. No entanto, para dar continuidade à análise, será assumida a normalidade dos dados, pois, sem essa suposição, a aplicação de técnicas multivariadas não se podem aplicar.

Após testar a normalidade multivariada, o próximo passo é verificar a homogeneidade das matrizes de covariância entre os grupos, um pressuposto essencial para a maioria dos testes estatísticos multivariados. Para isso vamos recorrer ao teste *M de Box*, que avalia se as matrizes de covariância das variáveis quantitativas são iguais entre os diferentes grupos definidos por uma variável qualitativa.

H_0*: As matrizes de covariância são homogéneas entre os grupos.*
versus
H_1*: Pelo menos um grupo apresenta uma matriz de covariância diferente.*

Para uma análise mais detalhada, realizou-se o teste separadamente para cada variável qualitativa da base de dados. Assim, testámos a homogeneidade das matrizes de covariância considerando, individualmente, os grupos formados pela variável *cut* (tipo de corte), *color* (cor do diamante) e *clarity* (grau de pureza). Essa abordagem permitiu identificar se alguma dessas variáveis categóricas influencia a estrutura de covariância das variáveis numéricas.

Tabela 4.1: Resultados do Teste *M de Box* para diferentes variáveis qualitativas.

Teste *M de Box*	
Variável	**Valor de prova**
cut	2.2e-16
color	6.221e-16
clarity	2.2e-16

A rejeição da hipótese nula para todas as variáveis qualitativas sugere que pelo menos um dos grupos possui uma matriz de covariância significativamente diferente das demais. Essa diferença pode ser resultado de variações estruturais nos dados, como a distribuição desigual das variáveis quantitativas dentro de cada categoria qualitativa.

4.2 MANOVA

A Análise de Variância Multivariada (MANOVA) é uma extensão da *ANOVA* que permite testar, simultaneamente, a influência de uma ou mais variáveis independentes categóricas sobre múltiplas variáveis dependentes quantitativas. Ao invés de avaliar cada variável de forma isolada, como na ANOVA, a MANOVA considera a estrutura conjunta das variáveis, permitindo identificar relações que poderiam ser perdidas em análises univariadas separadas. Para que uma MANOVA possa ser aplicada corretamente, é essencial que os dados satisfaçam dois pressupostos fundamentais: normalidade multivariada das variáveis dependentes dentro de cada grupo e homogeneidade das matrizes de covariância entre os grupos. Além disso, a MANOVA exige que o número de grupos da variável independente seja igual ou superior a três ($g \geq 3$). No entanto, de forma a garantir a realização da MANOVA e dar continuidade à análise, assumimos que os pressupostos referentes à normalidade e à homogeneidade das matrizes de covariância se verificam.

De seguida, elege-se as hipóteses consideradas para a aplicação da MANOVA, que permitem verificar se existem diferenças significativas entre os grupos definidos pelas variáveis qualitativas:

$$H_0 : \mu_{\text{Grupo 1}} = \mu_{\text{Grupo 2}} = \cdots = \mu_{\text{Grupo k}}$$

versus

$$H_1 : \exists \quad i \neq j \quad \text{tal que} \quad \mu_i \neq \mu_j$$

Apresenta-se agora a tabela da MANOVA, onde esta permite compreender como a variação total dos dados é repartida entre os grupos e o erro residual, sendo essencial para a avaliação da significância estatística das diferenças entre grupos. A tabela está organizada em três colunas principais:

- **Fonte de Variação**: Indica os componentes da variabilidade dos dados. Temos três fontes principais:
  - **Grupo**: Representa a variabilidade entre os diferentes grupos definidos pela variável qualitativa.
  - **Resíduo**: Representa a variabilidade dentro dos grupos, ou seja, o erro não explicado pela variável independente.
  - **Total**: Representa a variabilidade total dos dados, sendo a soma das componentes anteriores.

- **Graus de Liberdade (GL)**: Refere-se ao número de valores independentes utilizados no cálculo das variâncias.
  - Para os **grupos**, os graus de liberdade correspondem a $g - 1$, onde g é o número de grupos.
  - Para o **resíduo**, os graus de liberdade são a soma do número de observações em cada grupo menos o número total de grupos.
  - Para o **total**, os graus de liberdade correspondem ao número total de observações menos um.
- **Matriz da Soma de Quadrados e Produtos (SSP)**: Contém as expressões matemáticas das matrizes utilizadas na MANOVA:
  - **Matriz H (Entre Grupos)**: Mede a variabilidade entre os grupos, calculada como

$$\mathbf{H} = \sum_{l=1}^{g} n_l(\bar{x}_l - \bar{x}_{..})(\bar{x}_l - \bar{x}_{..})'$$

  - **Matriz E (Erro)**: Mede a variabilidade dentro dos grupos, dada por

$$\mathbf{E} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \bar{x}_l)'(x_{lj} - \bar{x}_l)$$

  - **Matriz T (Total)**: Representa a variabilidade total dos dados e é calculada como

$$\mathbf{T} = \mathbf{H} + \mathbf{E}$$

Essas matrizes são utilizadas para calcular as estatísticas de teste da MANOVA, tais como *Wilks' Lambda, Hotelling-Lawley e Roy's Largest Root*, que determinam se as diferenças entre grupos são estatisticamente significativas.

4.2.1 MANOVA para a variável *cut*

Para a variável *cut*, consideram-se as seguintes hipóteses, a partir das quais se procederá ao estudo da MANOVA.

$$H_0 : \mu_{\text{Fair}} = \mu_{\text{Good}} = \mu_{\text{Very Good}} = \mu_{\text{Premium}} = \mu_{\text{Ideal}}$$

versus

$$H_1 : \exists \quad i \neq j \quad \text{tal que} \quad \mu_i \neq \mu_j, \quad i, j \in \{\text{Fair, Good, Very Good, Premium, Ideal}\}$$

Utilizámos o *Wilks' Lambda* para avaliar a hipótese de igualdade dos vetores de médias entre os grupos. O valor obtido como valor de prova foi de 2.615×10^{-13}. Dado que este valor de prova é extremamente pequeno e inferior ao nível de significância de 5%, rejeitamos a hipótese nula (H_0), concluindo que pelo menos um dos grupos de *cut* apresenta diferenças significativas nos vetores de médias das variáveis dependentes. Assim, avançamos agora para a construção da tabela da MANOVA:

Tabela 4.2: Tabela da MANOVA para a variável *cut*.

Fonte de Variação	Graus de Liberdade(GL)	Matriz da Soma de Quadrados e Produtos (SSP)
Grupo (**H**)	4	Matriz H
Resíduo (**E**)	195	Matriz E
Total (**T**)	199	Matriz T

As matrizes **H**, **E** e **T** correspondem a:

$$\mathbf{H} = \begin{pmatrix} 2.07 & & & & & \\ 2.57 & 31.33 & & & & \\ 10016.73 & 9315.52 & 59373178.57 & & & \\ 4.63 & 3.45 & 21832.60 & 10.77 & & \\ 4.68 & 4.01 & 22959.37 & 10.75 & 10.84 & \\ 3.02 & 4.03 & 14319.21 & 6.84 & 6.89 & 4.46 \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} 40.70 & & & & & \\ 1.57 & 285.73 & & & & \\ 300967.19 & -47645.57 & 2653812324.43 & & & \\ 94.29 & -5.22 & 687095.83 & 227.74 & & \\ 93.40 & -6.97 & 683406.95 & 225.79 & 224.14 & \\ 58.14 & 13.38 & 420375.83 & 139.95 & 138.72 & 87.01 \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} 42.77 & & & & & \\ 4.14 & 317.06 & & & & \\ 310983.92 & -38330.05 & 2713185502.99 & & & \\ 98.92 & -1.77 & 708928.44 & 238.51 & & \\ 98.08 & -2.96 & 706366.32 & 236.54 & 234.98 & \\ 61.16 & 17.41 & 434695.04 & 146.79 & 145.61 & 91.47 \end{pmatrix}$$

4.2.2 MANOVA para a variável *color*

Para a variável *color*, estabelecem-se as seguintes hipóteses:

$$H_0 : \mu_{\text{D}} = \mu_{\text{E}} = \mu_{\text{F}} = \mu_{\text{G}} = \mu_{\text{H}} = \mu_{\text{I}} = \mu_{\text{J}}$$

versus

$$H_1 : \exists \quad i \neq j \quad \text{tal que} \quad \mu_i \neq \mu_j, \quad i, j \in \{\text{D, E, F, G, H, I, J}\}$$

Recorreu-se ao teste de *Hotelling-Lawley* para avaliar a hipótese de igualdade dos vetores de médias entre os grupos da variável *color*. O valor obtido para a estatística de *Hotelling-Lawley* foi 0.44457, com um valor de prova de 2.491×10^{-5}. Dado que este valor de prova é significativamente pequeno e inferior ao nível de significância de 5%, rejeitamos a hipótese nula (H_0), concluindo que pelo menos um dos grupos de *color* apresenta diferenças significativas nos vetores de médias das variáveis dependentes (*carat, depth, price, x, y* e *z*). Assim, realiza-se agora a construção da tabela da MANOVA:

Tabela 4.3: Tabela da MANOVA para a variável *color*.

Fonte de Variação	**Graus de Liberdade(GL)**	**Matriz da Soma de Quadrados e Produtos (SSP)**
Grupo (**H**)	6	Matriz H
Resíduo (**E**)	193	Matriz E
Total (**T**)	199	Matriz T

As matrizes **H**, **E** e **T** correspondem a:

$$\mathbf{H} = \begin{pmatrix} 3.72 & & & & & \\ 2.36 & 12.98 & & & & \\ 16167.30 & 3478.20 & 91838067.49 & & & \\ 8.55 & 4.38 & 37881.47 & 19.82 & & \\ 8.61 & 4.24 & 38385.08 & 19.99 & 20.17 & \\ 5.47 & 3.38 & 23967.12 & 12.61 & 12.71 & 8.06 \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} 39.05 & & & & & \\ 1.78 & 304.10 & & & & \\ 294816.63 & -41808.25 & 2621347435.51 & & & \\ 90.37 & -6.15 & 671046.97 & 218.69 & & \\ 89.47 & -7.20 & 667981.24 & 216.55 & 214.81 & \\ 55.70 & 14.03 & 410727.92 & 134.17 & 132.89 & 83.41 \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} 42.77 & & & & & \\ 4.14 & 317.06 & & & & \\ 310983.92 & -38330.05 & 2713185502.99 & & & \\ 98.92 & -1.77 & 708928.44 & 238.51 & & \\ 98.08 & -2.96 & 706366.32 & 236.54 & 234.98 & \\ 61.16 & 17.41 & 434695.04 & 146.79 & 145.61 & 91.47 \end{pmatrix}$$

4.2.3 MANOVA para a variável *clarity*

Para a variável *clarity*, formularam-se as seguintes hipóteses:

$$H_0 : \mu_{\text{I1}} = \mu_{\text{SI2}} = \mu_{\text{SI1}} = \mu_{\text{VS2}} = \mu_{\text{VS1}} = \mu_{\text{VVS2}} = \mu_{\text{VVS1}} = \mu_{\text{IF}}$$

versus

$$H_1 : \exists \quad i \neq j \quad \text{tal que} \quad \mu_i \neq \mu_j, \quad i, j \in \{\text{I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF}\}$$

Optou-se pelo teste de *Roy's Largest Root* para examinar a igualdade dos vetores de média entre os diferentes grupos da variável *clarity*. A estatística de *Roy* resultou numa estat´sitica de teste com um valor de 1.1737, o maior entre os testes realizados, com um valor F aproximado de 32.191 e um valor de prova inferior a 2.2×10^{-16}. Este resultado indica que a variável *clarity* tem o maior impacto sobre as variáveis dependentes em comparação com *cut* e *color*. Dado que este valor de prova é extremamente reduzido e significativamente abaixo do nível de significância de 5%, há evidências estatísticas suficientes para rejeitar a hipótese nula. Isso sugere que pelo menos um dos grupos da variável *clarity* apresenta diferenças notáveis nos vetores de médias das variáveis dependentes *carat, depth, price, x, y* e *z*. Com base nesses resultados, prosseguimos agora para a construção da tabela MANOVA:

Tabela 4.4: Tabela da MANOVA para a variável *clarity*.

Fonte de Variação	**Graus de Liberdade(GL)**	**Matriz da Soma de Quadrados e Produtos (SSP)**
Grupo (**H**)	7	Matriz H
Resíduo (**E**)	192	Matriz E
Total (**T**)	199	Matriz T

As matrizes **H**, **E** e **T** correspondem a:

$$\mathbf{H} = \begin{pmatrix} 10.54 & & & & & \\ 7.45 & 10.71 & & & & \\ 42167.41 & 23020.60 & 195947148.89 & & & \\ 25.42 & 17.88 & 104161.09 & 61.95 & & \\ 24.86 & 17.34 & 102252.10 & 60.63 & 59.35 & \\ 15.98 & 11.50 & 65086.28 & 38.93 & 38.10 & 24.48 \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} 32.23 & & & & & \\ -3.31 & 306.35 & & & & \\ 268816.51 & -61350.65 & 251723854.11 & & & \\ 73.51 & -19.65 & 604767.35 & 176.56 & & \\ 73.22 & -20.30 & 604114.22 & 175.91 & 175.63 & \\ 45.18 & 5.90 & 369608.77 & 107.85 & 107.51 & 66.99 \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} 42.77 & & & & & \\ 4.14 & 317.06 & & & & \\ 310983.92 & -38330.05 & 2713185502.99 & & & \\ 98.92 & -1.77 & 708928.44 & 238.51 & & \\ 98.08 & -2.96 & 706366.32 & 236.54 & 234.98 & \\ 61.16 & 17.41 & 434695.04 & 146.79 & 145.61 & 91.47 \end{pmatrix}$$

5 Análise Discriminante

> "The art of being wise is the art of knowing what to ignore."
>
> William James

A Análise Discriminante é um método estatístico multivariado utilizado para diferenciar grupos previamente definidos com base em variáveis quantitativas e para classificar novas observações nesses grupos. O seu principal objetivo é encontrar funções discriminantes, que sejam combinações lineares das variáveis independentes, permitindo maximizar a separação entre os grupos. A construção dessas funções envolve maximizar a variação entre os grupos e minimizar a variação dentro dos grupos, tornando mais fácil distinguir as diferentes categorias. Para que a Análise Discriminante seja aplicada corretamente, é necessário garantir que certos pressupostos estatísticos sejam satisfeitos. Esses pressupostos incluem:

- Deve haver pelo menos dois grupos definidos, ou seja, $g \geq 2$;
- Em cada grupo da amostra recolhida, deve haver pelo menos dois indivíduos, i.e., $n_i \geq 2$, onde n_i representa a dimensão da amostra no grupo i, com $i \in \{1, \ldots, g\}$;
- As p variáveis quantitativas iniciais devem formar um vetor aleatório com distribuição normal p-variada;
- A variabilidade dentro dos g grupos deve ser idêntica, ou seja, as matrizes de covariância dos g grupos devem ser homogéneas;
- Nenhuma das p variáveis quantitativas deve ser uma combinação linear das outras variáveis do conjunto inicial;
- O número total de variáveis quantitativas consideradas deve ser inferior ao número total de observações na amostra menos dois, ou seja, $p < n - 2$, onde $n = n_1 + n_2 + \cdots + n_g$;
- As funções discriminantes a definir devem ser estatisticamente independentes;
- O número de funções discriminantes que podem ser determinadas é dado por $\min(p, g - 1)$.

A Análise Discriminante pode ser dividida em dois tipos principais:

- **Análise Discriminante Descritiva** – Onde são definidas funções discriminantes com base numa amostra conhecida para analisar as diferenças entre os grupos.
- **Análise Discriminante Preditiva** – Onde as funções discriminantes são utilizadas para prever a qual grupo uma nova observação pertence.

As funções discriminantes permitem identificar as direções nos dados que melhor separam os grupos. Cada função discriminante assume a forma:

$$D_k = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p$$

onde:

- D_k representa o valor da função discriminante para o grupo k;
- X_i são as variáveis independentes;
- b_i são os coeficientes determinados de forma a maximizar a separação entre os grupos.

O número máximo de funções discriminantes extraídas corresponde ao menor valor entre p, o número de variáveis independentes, e $g - 1$, onde g é o número de grupos.

5.1 Análise Discriminante para a variável *cut*

No Capítulo 4.1 foram realizados testes para avaliar os pressupostos necessários à aplicação de métodos estatísticos multivariados. Os resultados indicaram que os dados não seguem uma distribuição normal multivariada, conforme verificado pelo teste de *Shapiro-Wilk* multivariado. Além disso, a homogeneidade das matrizes de covariância foi analisada através do teste *M de Box*, cujos resultados revelaram diferenças significativas entre os grupos, indicando que esse pressuposto também não é satisfeito. No entanto, para dar continuidade à análise, assumiremos que estes pressupostos se verificam, permitindo assim a aplicação da Análise Discriminante para as três variáveis qualitativas da base de dados *diamonds*.

Após a aplicação da Análise Discriminante Linear (LDA) à variável *cut*, obteve-se um conjunto de funções discriminantes que permitem distinguir os grupos com base nas variáveis quantitativas da base de dados.

5.1.1 Probabilidades a Priori

Os valores a priori indicam a proporção de observações em cada grupo antes da aplicação do modelo:

- *Fair*: 2% das observações;
- *Good*: 7,5%;
- *Very Good*: 20,5%;
- *Premium*: 24%;
- *Ideal*: 46%.

Estes valores permitem compreender a distribuição da variável *cut* na amostra e fornecem uma referência para a classificação.

5.1.2 Médias dos Grupos

A Tabela 5.1 apresenta as médias das variáveis quantitativas para cada nível da variável *cut*, demonstrando as diferenças entre os grupos:

Tabela 5.1: Médias das variáveis quantitativas para cada grupo da variável *cut*.

Grupo	Carat	Depth	Table	Price	x	y	z
Fair	0.7775	62.88	58.00	2861.25	5.80	5.76	3.63
Good	0.9547	63.01	57.47	4421.07	5.98	5.99	3.77
Very Good	0.9341	61.86	57.93	4809.85	6.06	6.09	3.75
Premium	0.8656	61.50	58.77	3922.67	5.95	5.91	3.65
Ideal	0.7074	61.74	55.97	3469.52	5.53	5.54	3.42

As diferenças observadas nas médias justificam a separação dos grupos através da Análise Discriminante.

5.1.3 Coeficientes das Funções Discriminantes

A Tabela 5.2 apresenta os coeficientes das funções discriminantes, que indicam como cada variável contribui para a separação dos grupos:

Tabela 5.2: Coeficientes das funções discriminantes para a variável *cut*.

Variável	LD1	LD2	LD3	LD4
Carat	1.88	-4.28	-1.34	9.90
Depth	1.10	-1.09	-2.21	-2.55
Table	0.61	-0.03	0.09	-0.02
Price	-0.0001	1.27	5.82	-0.0001
x	4.26	1.82	-1.76	-14.85
y	1.26	-2.83	4.50	-13.92
z	-9.29	12.18	2.80	4.03

Os coeficientes indicam a contribuição de cada variável na separação dos grupos. Em particular:

- A primeira função discriminante (LD1) tem maior peso nas variáveis z (-9.29) e x (4.26), indicando que estas são as principais responsáveis pela separação entre os grupos.
- A segunda função discriminante (LD2) destaca-se pela influência de *carat* (-4.28) e z (12.18), sugerindo uma nova direção de separação entre os grupos.
- As funções LD3 e LD4 explicam variações menores e têm menor impacto na distinção dos grupos.

5.1.4 Proporção da Variância Explicada

A proporção da variância explicada por cada função discriminante está representada na Tabela 5.3:

Tabela 5.3: Proporção da variância explicada por cada função discriminante.

LD1	LD2	LD3	LD4
55.6%	36.1%	6.7%	1.5%

Verifica-se que:

- A LD1 é a mais relevante, explicando 55,6% da variabilidade entre os grupos;
- A LD2 explica 36,1%;
- As funções LD3 e LD4 possuem uma importância reduzida na separação dos grupos.

5.1.5 Representação Gráfica

Com base nos resultados, verifica-se que as funções discriminantes LD1 e LD2 são as mais relevantes para distinguir os grupos da variável *cut*. O próximo passo será visualizar os resultados graficamente, permitindo uma melhor compreensão da separação entre os grupos da variável *cut* com base nas funções discriminantes.

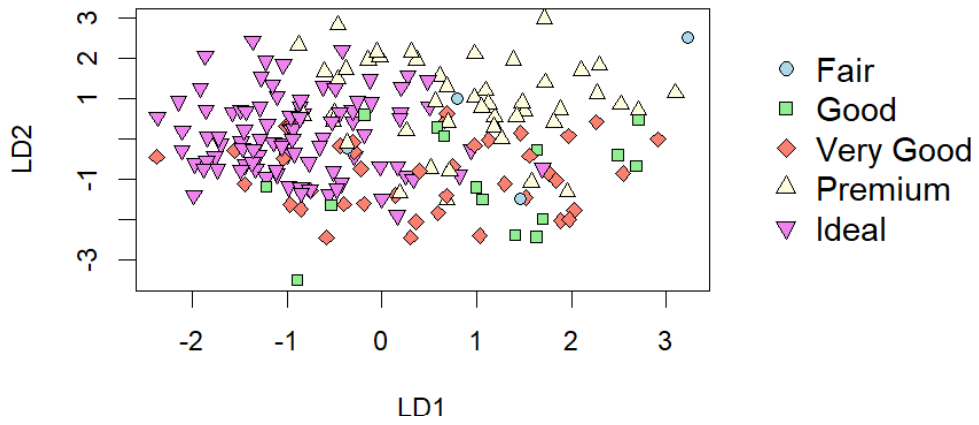

Figura 5.1: Representação das funções discriminantes para a variável *cut*.

O gráfico apresentado na Figura 5.1 ilustra a distribuição dos grupos da variável *cut* em função das duas primeiras funções discriminantes (LD1 e LD2). Observa-se que os grupos não estão totalmente separados, verificando-se alguma sobreposição entre eles. No entanto, a função discriminante LD1 assume um papel predominante na distinção entre os grupos, uma vez que a maior parte da variabilidade ocorre ao longo deste eixo. O grupo *Ideal* apresenta uma maior concentração de observações, enquanto os grupos *Premium*, *Good* e *Very Good* demonstram uma maior dispersão.

5.1.6 Predição das classes para a variável *cut*

Depois de obter as funções discriminantes, vamos agora utilizar o modelo para prever a categoria de *cut* de cada observação. Com esta previsão, podemos avaliar como o modelo classifica os dados e verificar a separação entre os grupos.

A tabela de frequências das previsões mostra que o modelo atribuiu as observações às seguintes classes:

- *Fair*: 0 previsões
- *Good*: 7 previsões
- *Very Good*: 38 previsões
- *Premium*: 50 previsões
- *Ideal*: 105 previsões

Este resultado sugere que o modelo tem uma tendência para classificar a maioria das observações como *Ideal*, enquanto a classe *Fair* não foi prevista em nenhuma observação.

Para avaliar a precisão do modelo, analisamos a matriz de confusão apresentada na Tabela 5.4. Esta matriz permite comparar as classes reais com as classes previstas pelo modelo, identificando padrões de acertos.

Tabela 5.4: Matriz de confusão resultante da classificação da variável *cut*.

Classe Real \ Classe Prevista	**Fair**	**Good**	**Very Good**	**Premium**	**Ideal**
Fair	0	1	0	2	1
Good	0	4	4	4	3
Very Good	0	0	21	7	13
Premium	0	1	5	33	9
Ideal	0	1	8	4	79

- A classe *Ideal* apresentou o maior número de acertos, com 79 observações corretamente classificadas.
- A classe *Premium* também teve um bom desempenho, com 33 acertos, mas houve alguma confusão com as classes *Very Good* e *Ideal*.
- As classes *Fair* e *Good* tiveram um desempenho mais fraco. O modelo não previu nenhuma observação como *Fair*, e apenas 4 observações de *Good* foram corretamente identificadas.

A análise discriminante aplicada à variável *cut* permitiu construir um modelo capaz de diferenciar as categorias desta variável com base em atributos quantitativos. A partir da matriz de confusão, verificamos que o modelo apresentou um desempenho satisfatório, especialmente na classificação da categoria **Ideal**, que obteve o maior número de acertos - 79 observações corretamente classificadas. A categoria *Premium* também apresentou um desempenho aceitável, embora tenha havido alguma confusão com as classes *Very Good* e *Ideal*. A *accuracy* total do modelo foi de 68.5%. Quando analisamos a *accuracy* por classe, os resultados foram os seguintes:

- *Fair*: 0.00%;
- *Good*: 26.67%;
- *Very Good*: 51.22%;
- *Premium*: 68.75%;
- *Ideal*: 85.87%.

5.2 Análise Discriminante para a variável *color*

Após a aplicação da Análise Discriminante à variável *cut*, prosseguimos agora com a mesma abordagem para a variável *color*. O objetivo desta análise é verificar se as variáveis quantitativas permitem distinguir corretamente as categorias de *color* e avaliar a eficácia do modelo na classificação das observações.

5.2.1 Probabilidades a Priori

Os valores a priori para esta variável são os seguintes:

- *D*: 11,5%;
- *E*: 18%;
- *F*: 17%;
- *G*: 23%;
- *H*: 15%;
- *I*: 11%;
- *J*: 4,5%.

A classe dominante trata-se da classe *G* com 23% das observações. Isso significa que, antes da aplicação do modelo, a classe *G* tem maior representatividade na amostra, enquanto *J* é a menos frequente.

5.2.2 Médias dos Grupos

A Tabela 5.5 expõe as médias das variáveis quantitativas para cada nível da variável *color*:

Tabela 5.5: Médias das variáveis quantitativas para cada grupo da variável *color*.

Grupo	Carat	Depth	Table	Price	x	y	z
D	0.6322	61.73	56.43	3564.04	5.38	5.39	3.32
E	0.6292	61.65	57.14	2723.67	5.35	5.34	3.30
F	0.8697	61.64	57.65	4298.03	5.97	5.98	3.68
G	0.8526	61.71	57.39	4456.76	5.86	5.86	3.62
H	0.8330	61.73	58.10	3800.00	5.78	5.78	3.60
I	0.9468	62.37	57.63	4263.95	6.05	6.05	3.77
J	1.1733	61.69	57.30	5212.89	6.61	6.64	4.09

A média das variáveis indica que:

- O peso do diamante (*carat*) aumenta progressivamente de *D* para *J*, sendo *J* a categoria com maior média;
- O preço (*price*) segue a mesma tendência, sugerindo que cores mais intensas podem estar associadas a diamantes de maior valor;
- As dimensões físicas (*x, y, z*) também aumentam ao longo das categorias, indicando uma possível relação entre a cor e o tamanho do diamante.

5.2.3 Coeficientes das Funções Discriminantes

A Tabela 5.6 apresenta os coeficientes das funções discriminantes relativamente À variável *color*:

Tabela 5.6: Coeficientes das funções discriminantes para a variável *color*.

Variável	LD1	LD2	LD3	LD4	LD5	LD6
Carat	8.00	-3.47	-1.91	1.57	7.20	-6.31
Depth	0.54	-3.37	-1.06	-0.71	-2.41	0.73
Table	0.07	-0.02	0.30	0.02	-0.31	-0.21
Price	-0.0006	1.89	2.65	-0.0003	-1.24	0.0001
x	-8.98	-2.79	-3.20	8.29	-6.72	17.68
y	12.72	-3.19	-1.56	-8.79	-1.86	0.00
z	-7.53	52.39	2.59	0.91	3.62	0.00

Os coeficientes mostram que:

- A variável *y* tem o maior impacto na LD1 (12.72), sugerindo que a largura do diamante é o principal fator de separação entre os grupos.
- A variável *z* tem um coeficiente muito alto na LD2 (52.39), indicando que a dimensão vertical do diamante contribui para a distinção entre cores.

5.2.4 Proporção da Variância Explicada

A proporção da variância explicada por cada função discriminante está representada na Tabela 5.7:

Tabela 5.7: Proporção da variância explicada por cada função discriminante.

LD1	LD2	LD3	LD4	LD5	LD6
61.03%	19.73%	8.78%	5.07%	3.97%	1.42%

Verifica-se que:

- A LD1 explica 61,03% da variabilidade entre os grupos, sendo a função discriminante mais relevante;
- A LD2 também tem um peso significativo, explicando 19,73% da variabilidade;
- As funções LD3 a LD6 contribuem menos para a separação dos grupos.

5.2.5 Representação Gráfica

Voltou-se a verificar que as funções discriminantes LD1 e LD2 são as mais relevantes para distinguir os grupos. O próximo passo será visualizar os resultados graficamente, permitindo uma melhor compreensão da separação entre os grupos com base nas funções discriminantes.

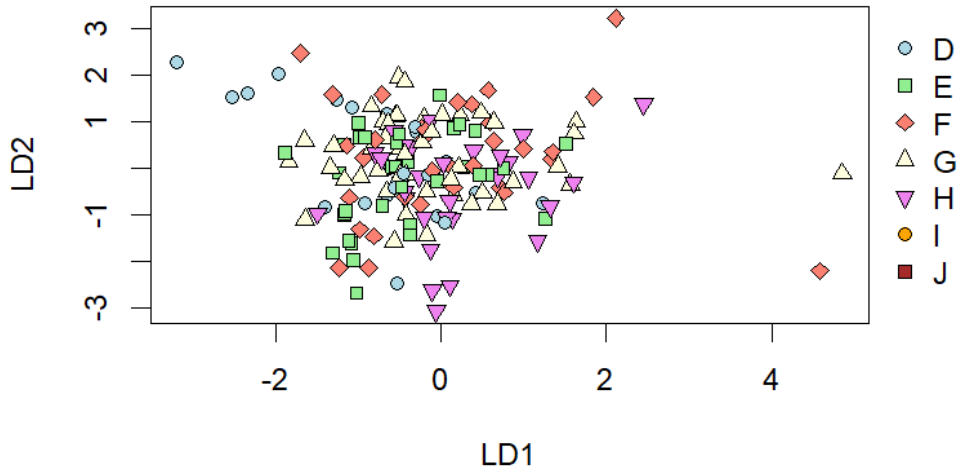

Figura 5.2: Representação das funções discriminantes para a variável *color*.

O gráfico representa a separação dos grupos da variável *color* com base nas funções discriminantes LD1 e LD2. A LD1 tem um papel mais relevante na distinção entre os grupos, enquanto a LD2 contribui de forma menos expressiva, verificando-se uma sobreposição significativa entre algumas categorias. Os grupos *D, E, F, G* e *H* apresentam interseção considerável, sugerindo que as variáveis utilizadas podem não ser totalmente eficazes na separação das categorias. As categorias *I* e *J* não aparecem no gráfico, pois as categorias podem estar sobrepostas a outros grupos, dificultando a sua identificação visual. O grupo *G* mostra maior concentração de observações, enquanto *D, E* e *F* apresentam maior dispersão.

5.2.6 Predição das classes para a variável *color*

Depois de obter as funções discriminantes, utilizamos o modelo para prever a categoria de *color* de cada observação. Com esta previsão, podemos avaliar como o modelo classifica os dados e verificar a separação entre os grupos.

A tabela de frequências das previsões mostra que o modelo atribuiu as observações às seguintes classes:

- *D*: 7 previsões;
- *E*: 44 previsões;
- *F*: 35 previsões;
- *G*: 63 previsões;
- *H*: 20 previsões;
- *I*: 18 previsões;
- *J*: 13 previsões.

Este resultado sugere que o modelo tem uma distribuição mais equilibrada entre as classes, mas algumas categorias, como *G* e *E*, receberam um número maior de previsões, enquanto a classe *D* teve um número significativamente menor de observações classificadas.

Para avaliar a precisão do modelo, analisamos a matriz de confusão apresentada na Tabela 5.8. Esta matriz permite comparar as classes reais com as classes previstas pelo modelo, identificando padrões de acertos e erros.

Tabela 5.8: Matriz de confusão resultante da classificação da variável *color*.

Classe Real \ Classe Prevista	**D**	**E**	**F**	**G**	**H**	**I**	**J**
D	3	7	3	6	2	2	0
E	1	13	8	10	3	1	0
F	1	7	7	12	2	1	4
G	1	8	6	22	3	4	2
H	1	5	6	8	5	4	1
I	0	4	4	4	4	6	0
J	0	0	1	1	1	0	6

- A classe *G* apresentou o maior número de acertos, com 22 observações corretamente classificadas.
- A classe *E* também teve um desempenho razoável, com 13 observações corretamente identificadas, embora haja alguma confusão com as classes *F* e *G*.
- As classes *D* e *J* tiveram um desempenho mais fraco, com apenas 3 e 6 observações corretamente classificadas, respetivamente.
- Observa-se que há uma dispersão significativa das classificações, indicando que algumas categorias apresentam sobreposição entre si.

A análise discriminante aplicada à variável *color* permitiu construir um modelo capaz de diferenciar as categorias desta variável com base em atributos quantitativos. A partir da matriz de confusão, verificamos que o modelo apresentou um desempenho moderado, especialmente na classificação da categoria **G**, que obteve o maior número de acertos com 22 observações corretamente classificadas. No entanto, algumas classes, como *D* e *J*, apresentaram dificuldades na classificação, sugerindo que pode haver interseções nos atributos utilizados para discriminar essas categorias.A *accuracy* total do modelo foi de 31%. A análise da *accuracy* por classe revelou os seguintes resultados:

- *D*: 13.04%;
- *E*: 36.11%;
- *F*: 20.59%;
- *G*: 47.83%;
- *H*: 16.67%;
- *I*: 27.27%;
- *J*: 66.67%.

5.3 Análise Discriminante para a variável *clarity*

Após a aplicação da Análise Discriminante às variáveis *cut* e *color*, procedemos agora com a mesma abordagem para a variável *clarity*

5.3.1 Probabilidades a Priori

As probabilidades a priori para cada categoria de *clarity* são as seguintes:

- *I1*: 2%;
- *SI2*: 17%;
- *SI1*: 25%;
- *VS2*: 24%;
- *VS1*: 12%;
- *VVS2*: 6,5%;
- *VVS1*: 9,5%;
- *IF*: 4%.

A classe predominante é *SI1*, representando 25% das observações, enquanto *I1* é a menos frequente, com apenas 2%. Isso indica que, antes da aplicação do modelo, a classe *SI1* possui maior representatividade na amostra.

5.3.2 Médias dos Grupos

A Tabela 5.9 apresenta as médias das variáveis quantitativas para cada categoria da variável *clarity*:

Tabela 5.9: Médias das variáveis quantitativas para cada categoria da variável *clarity*.

Grupo	Carat	Depth	Table	Price	x	y	z
I1	1.55	62.50	57.00	5400.75	7.30	7.24	4.55
SI2	1.13	61.87	57.46	497.79	6.55	6.54	4.05
SI1	0.87	62.10	57.16	3933.20	5.90	5.89	3.69
VS2	0.79	61.84	57.52	3758.98	5.79	5.79	3.58
VS1	0.71	61.64	57.08	4153.71	5.58	5.58	3.44
VVS2	0.57	61.66	56.79	2986.54	5.18	5.21	3.19
VVS1	0.42	61.43	56.73	1880.42	4.74	4.76	2.92
IF	0.48	61.46	55.88	2823.50	4.94	4.96	3.04

A análise das médias revela que:

- O peso do diamante (*carat*) diminui progressivamente de *I1* para *VVS1*, com um leve aumento em *IF*;
- O preço (*price*) não segue um padrão linear claro, mas diamantes da categoria *I1* apresentam um valor médio consideravelmente superior aos demais;
- As dimensões físicas (*x, y, z*) diminuem gradualmente ao longo das categorias, sugerindo que diamantes de maior clareza tendem a ser menores.

5.3.3 Coeficientes das Funções Discriminantes

A Tabela 5.10 apresenta os coeficientes das funções discriminantes para a variável *clarity*:

Tabela 5.10: Coeficientes das funções discriminantes para a variável *clarity*.

Variável	LD1	LD2	LD3	LD4	LD5
Carat	-7.57	10.28	-4.33	4.56	-3.23
Depth	-0.99	-2.52	-3.04	1.93	-1.86
Table	0.02	-2.01	-1.89	-1.42	0.25
Price	0.0007	-2.59	2.39	-1.30	-2.72
x	-7.65	-1.45	-1.18	-3.69	-2.06
y	16.35	3.57	4.47	-3.87	2.94

Os coeficientes indicam que:

- A variável *y* exerce forte influência na LD1 (16.35), sugerindo que a largura diamante é um fator chave na diferenciação das categorias de clareza.
- A variável *carat* tem um impacto relevante na LD2 (10.28), indicando que o peso pode contribuir para a distinção entre grupos.

5.3.4 Proporção da Variância Explicada

A proporção da variância explicada por cada função discriminante é apresentada na Tabela 5.11:

Tabela 5.11: Proporção da variância explicada por cada função discriminante para a variável *clarity*.

LD1	LD2	LD3	LD4	LD5	LD6	LD7
82.86%	11.19%	2.61%	2.04%	0.91%	0.35%	0.05%

Os resultados indicam que:

- A LD1 explica 82,86% da variabilidade entre os grupos, sendo a mais relevante para a separação das categorias;
- A LD2 também tem um peso significativo, explicando 11,19% da variabilidade;
- As funções LD3 a LD7 possuem menor influência na separação das classes.

5.3.5 Representação Gráfica

As funções discriminantes LD1 e LD2 destacam-se como as mais relevantes na distinção dos grupos. A seguir, a visualização gráfica facilitará a interpretação da separação entre as categorias.

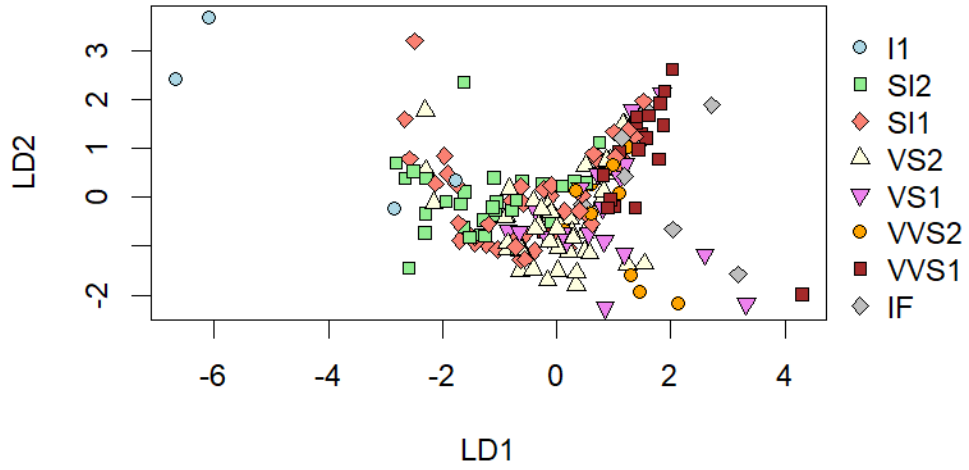

Figura 5.3: Representação das funções discriminantes para a variável *clarity*.

O gráfico representa a separação dos grupos da variável *clarity* com base nas funções discriminantes LD1 e LD2. Observa-se que LD1 tem um papel predominante na distinção entre os grupos, enquanto LD2 contribui de forma menos expressiva. Existe uma sobreposição significativa entre algumas categorias, indicando que as variáveis preditoras não são totalmente eficazes na separação dos grupos. Os grupos *SI1, VS2, VS1 e VVS2* apresentam maior dispersão, enquanto *IF e VVS1* mostram uma distribuição mais concentrada, sugerindo menor variabilidade interna.

5.3.6 Predição das classes para a variável *clarity*

Após a obtenção das funções discriminantes, aplicámos o modelo para prever a categoria de *clarity* de cada observação. Esta previsão permite avaliar a capacidade do modelo em classificar corretamente os dados e analisar a separação entre os grupos.

A tabela de frequências das previsões mostra que o modelo atribuiu as observações às seguintes classes:

- *I1*: 3 previsões;
- *SI2*: 35 previsões;
- *SI1*: 44 previsões;
- *VS2*: 64 previsões;
- *VS1*: 12 previsões;
- *VVS2*: 7 previsões;
- *VVS1*: 34 previsões;
- *IF*: 1 previsão.

Estes resultados indicam que o modelo distribui as previsões de forma relativamente equilibrada entre as classes, com uma maior atribuição às categorias *VS2* e *SI1*. No entanto, algumas categorias, como *IF* e *I1*, receberam um número significativamente menor de previsões, o que pode indicar dificuldades na sua correta identificação.

Para avaliar a precisão do modelo, analisamos a matriz de confusão apresentada na Tabela 5.12. Esta matriz permite comparar as classes reais com as classes previstas, identificando padrões de acerto e erro.

Tabela 5.12: Matriz de confusão resultante da classificação da variável *clarity*.

Classe Real \ Classe Prevista	**I1**	**SI2**	**SI1**	**VS2**	**VS1**	**VVS2**	**VVS1**	**IF**
I1	2	2	0	0	0	0	0	0
SI2	1	20	6	6	1	0	1	0
SI1	1	10	20	13	0	1	5	0
VS2	0	3	12	24	2	2	5	0
VS1	0	0	4	10	4	1	5	0
VVS2	0	0	0	3	0	3	6	0
VVS1	0	0	0	0	0	0	12	1
IF	0	0	0	1	3	0	4	0

A análise da matriz de confusão revela que:

- A categoria *VS2* apresentou o maior número de acertos, com 24 observações corretamente classificadas.
- A classe *SI1* também obteve um desempenho razoável, com 20 classificações corretas, apesar de haver alguma confusão com *VS2* e *SI2*.
- As categorias *I1* e *IF* foram as menos reconhecidas corretamente pelo modelo, sugerindo maior dificuldade na distinção destes grupos.
- Existe uma dispersão significativa entre as classificações, indicando que algumas categorias apresentam sobreposição nos atributos analisados.

A aplicação da análise discriminante à variável *clarity* permitiu construir um modelo capaz de diferenciar as categorias com base em atributos quantitativos. No entanto, a sobreposição entre algumas classes sugere que a separação não é completamente eficaz, podendo ser necessário o uso de variáveis adicionais ou abordagens alternativas para melhorar a precisão do modelo. A *accuracy* total obtida foi de 42.50%, evidenciando um desempenho moderado. A *accuracy* por classe apresenta valores distintos entre os diferentes níveis de *clarity*, como se pode ver de seguida:

- *I1*: 50%;
- *SI2*: 58.82%;
- *SI1*: 40%;
- *VS2*: 50%;
- *VS1*: 16.67%;
- *VVS2*: 23.08%;
- *VVS1*: 63.16%;
- *IF*: 0%.

6 Análise de componentes principais

"Uma solução aproximada para o problema certo é bem mais valioso do que a solução exata para um problema aproximado."

John Tukey

A análise de componentes principais(ACP) é um método estatístico multivariado, que surgiu através de Pearson e Hotelling (Jollife, 2002), que permite transformar um conjunto de variáveis iniciais correlacionadas entre si, num outro conjunto de variáveis não correlacionadas (ortogonais), as designadas componentes principais, que resultam de combinações lineares do conjunto inicial. Atente que, o objetivo na análise de componentes principais não é explicar as correlações entre as variáveis mas apenas encontrar funções matemáticas entre as variáveis iniciais que expliquem o máximo possível da variação existente nos dados e os permitam descrever e reduzir.

Na análise das componentes principais, a representação matemática das combinações lineares não pressupõe a imposição de qualquer modelo causal mas também não permite detetar relações de causa-efeito entre as variáveis iniciais, mesmo se estas existirem.

Dado isto, as componentes principais são expressas como combinações lineares das variáveis originais. Por exemplo, para m componentes e p variáveis ($m \leq p$) :

$$\begin{aligned} CP_1 &= a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ CP_2 &= a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ &\vdots \\ CP_m &= a_{1m}X_1 + a_{2m}X_2 + \cdots + a_{pm}X_p \end{aligned} \tag{5}$$

As novas variáveis não correlacionadas (componentes principais) apresentam-se por ordem decrescente de importância da variabilidade.

Na análise de componentes principais, no conjunto das equações (5) o erro está ausente, uma vez que, as variáveis observáveis são medidas sem erro e as variáveis latentes são combinações lineares perfeitas dessas variáveis.

6.1 Teste de Esfericidade de Bartlett e Estatística de KMO

Antes de se proceder à extração das componentes principais, testa-se a legitimidade da utilização deste método de análise estatística multivariada. Existem várias formas de validar este tipo de análise, iremos validar através do teste de Esfericidade de Bartlett e da estatística de KMO (Kaiser-Meyer-Olkin).

A estatística de KMO compara as correlações entre as componentes. O valor obtido para a estatística de KMO é 0.72, superior a 0.50, indicando a adequação do método de análise das componentes principais.

Tabela 6.1: Índice do valor de KMO

KMO	Análise das componentes principais
1-0.90	Muito boa
0.80-0.90	Boa
0.70-0.80	Média
0.60-0.70	Razoável
0.50-0.60	Má
< 0.50	Inaceitável

O teste de esfericidade de Bartlett testa a hipótese de a matriz de correlações ser uma matriz identidade e o seu determinante ser igual a 1, logo, de as variáveis não estarem correlacionadas entre si.A aplicação da análise de componentes principais pressupõe que se rejeite a hipótese nula, H_0: $\boldsymbol{P=I}$. Como o valor de prova é muito próximo de zero e inferior ao nível de significância(0.05), rejeita-se a hipótese nula, isto é, há evidência estatística de que existe correlação significativa entre as variáveis. Por conseguinte, tendo em conta os indicadores para a realização da ACP, avança-se com o estudo.

6.2 Extração e seleção das componentes principais

O procedimento matemático usado para determinar as componentes principais baseou-se na utilização dos valores próprios da matriz de correlações. Assim, as componentes principais são os vetores próprios associados a esses valores próprios. Estas novas variáveis, componentes principais, têm conjuntamente a mesma variabilidade das variáveis originais. Além disso, cada componente contribui para a variância total com uma quantidade igual ao seu valor próprio. Consequentemente, a proporção de variância total explicada por cada componente principal é dada pelo quociente entre o valor próprio associado ao vetor próprio que a define e a soma dos valores próprios da matriz.

Tabela 6.2: Valores Próprios e proporções de Variância das Componentes Principais

Componentes Principais	Valores Próprios	Proporção de Variância (%)	Proporção Acumulada de Variância (%)
CP1	**4.825**	**68.925**	**68.925**
CP2	**1.440**	**20.573**	**89.498**
CP3	0.549	7.840	97.339
CP4	0.163	2.330	99.669
CP5	0.022	0.317	99.986
CP6	0.001	0.011	99.997
CP7	0.000	0.003	100.000

Na Tabela 6.2 pode observar-se o valor próprio associado a cada componente e a proporção de variância total explicada por cada componente, individual e acumulada.

Tendo em mente que o primeiro objetivo é reduzir a dimensionalidade dos dados, é importante fixar o número de componentes a reter. Para tal é necessário não esquecer que essa redução da dimensão tem de ser ponderada, uma vez que, tem de explicar uma proporção bastante significativa da variância total. Existem alguns critérios para determinar o número de componentes principais a reter. Um deles é o critério de de Kaiser, o qual consiste em excluir as componentes cujos valores próprios são inferiores ao seu valor médio. Uma vez que a Análise de Componentes Principais foi feita partindo da matriz de correlações, este critério traduz-se na exclusão das componentes cujos valores próprios são menores do que 1. Aplicando este critério ao estudo, conclui-se que se devem reter as 2 primeiras componentes principais.

Alternativamente, um critério utilizado neste estudo, para decidir o número de componentes a reter, consiste em incluir apenas as componentes necessárias para explicar mais do que 80% da proporção de variância total (de notar a subjetividade neste critério). De acordo com este critério, foram selecionadas as duas primeiras componentes

principais, que explicam cumulativamente 89.498% da variância total, superando o limiar de 80% para garantir uma representação adequada dos dados com menor dimensionalidade.

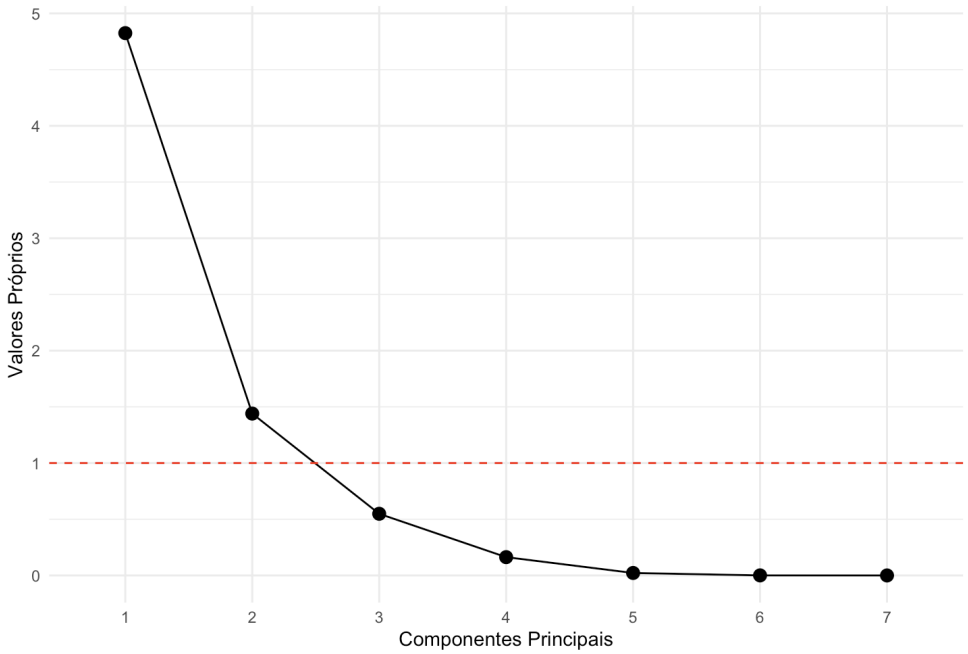

Figura 6.1: Gráfico do Cotovelo (Scree Plot)

Um último critério, consiste na representação gráfica dos pontos que relacionam o número de ordem de cada componente com o seu valor próprio associado (denominado screeplot (Figura 6.1)). Ao unir esses pontos obtém-se uma linha poligonal. O ponto no qual ocorre uma mudança abrupta no declive da linha, indica a ordem da última componente a considerar. Por observação da Figura 6.1, conclui-se que se devem reter as duas primeiras componentes. É notória a existência de um ponto de inflexão na poligonal do screeplot(Figura 6.1), após a segunda componente.

Por conseguinte, a decisão de reter duas primeiras componentes principais foi sustentada por três critérios: o *critério de Kaiser*, o gráfico do cotovelo e a variância explicada acumulada. Em todos os casos, os resultados indicaram que as duas primeiras componentes são suficientes, explicando conjuntamente 89.5% da variância total dos dados.

6.3 Os Pesos (Loadings) e as Comunalidades

Na apresentação dos resultados da ACP é habitual apresentar, em vez dos vetores próprios (CP), os vetores dos pesos (*loadings*). Estes vetores são obtidos multiplicando cada vetor próprio pela raíz quadrada do respetivo valor próprio. Estes vetores podem ser interpretados de duas maneiras: são vetores de pesos das variáveis iniciais nas componentes respetivas, e os seus elementos medem as correlações entre as componentes e as variáveis originais estandardizadas.

Tabela 6.3: Matriz dos Pesos e Comunalidades das Componentes Principais Retidas

Matriz dos Pesos			**Comunalidades**	
Variáveis	**CP1**	**CP2**	**Inicial**	**Extração**
carat	0.451	-0.039	1	0.205
depth	-0.001	-0.721	1	0.519
table	0.092	0.687	1	0.480
price	0.422	-0.012	1	0.178
x	0.452	-0.005	1	0.204
y	0.452	-0.004	1	0.204
z	0.449	-0.084	1	0.209

Na Tabela 6.3 apresentam-se os pesos das duas componentes principais extraídas. Observa-se que a correlação entre as variáveis originais estandardizadas e as componentes principais tende a ser tanto mais elevada quanto maior for o valor próprio (variância explicada) associado à componente. Verifica-se ainda que, ao excluir as cinco componentes restantes, está-se essencialmente a prescindir de componentes pouco correlacionadas com as variáveis originais, e, portanto, com reduzida capacidade explicativa. A soma dos quadrados dos pesos (*loadings*) de cada variável nas componentes retidas,denominada comunalidade, representa a proporção da variância dessa variável explicada pelas componentes selecionadas. Esta comunalidade atinge o valor 1 apenas quando todas as componentes principais são consideradas, ou seja, quando a totalidade da variância original é mantida.

A análise dos pesos revela que a primeira componente principal (CP1) está fortemente associada às variáveis *carat*, *price*, *x*, *y* e *z*, todas com valores de *loading* superiores a 0.42. Estas variáveis estão diretamente relacionadas com o tamanho e o valor do diamante, sugerindo que a CP1 representa uma dimensão de volume e valor. Já a segunda componente (CP2) apresenta elevadas correlações com as variáveis *depth* (*loading* = -0.721) e *table* (*loading* = 0.687), que são medidas ligadas à geometria do corte do diamante. Assim, a CP2 pode ser interpretada como uma dimensão associada às proporções estruturais do diamante.

No que diz respeito às comunalidades, observa-se que as variáveis *depth* e *table* são razoavelmente bem explicadas pelas duas componentes principais retidas (comunalidades de 0.519 e 0.480, respetivamente), enquanto variáveis como *price*, *carat* e *z* apresentam comunalidades mais baixas (entre 0.17 e 0.21), indicando que uma parte significativa da sua variância permanece não explicada pelas duas componentes principais.

Por conseguinte, as duas componentes principais retidas capturam as principais fontes de variação nos dados: uma relacionada com o tamanho/valor e outra com a proporção geométrica, assegurando uma representação simplificada mas interpretável do espaço original de variáveis.

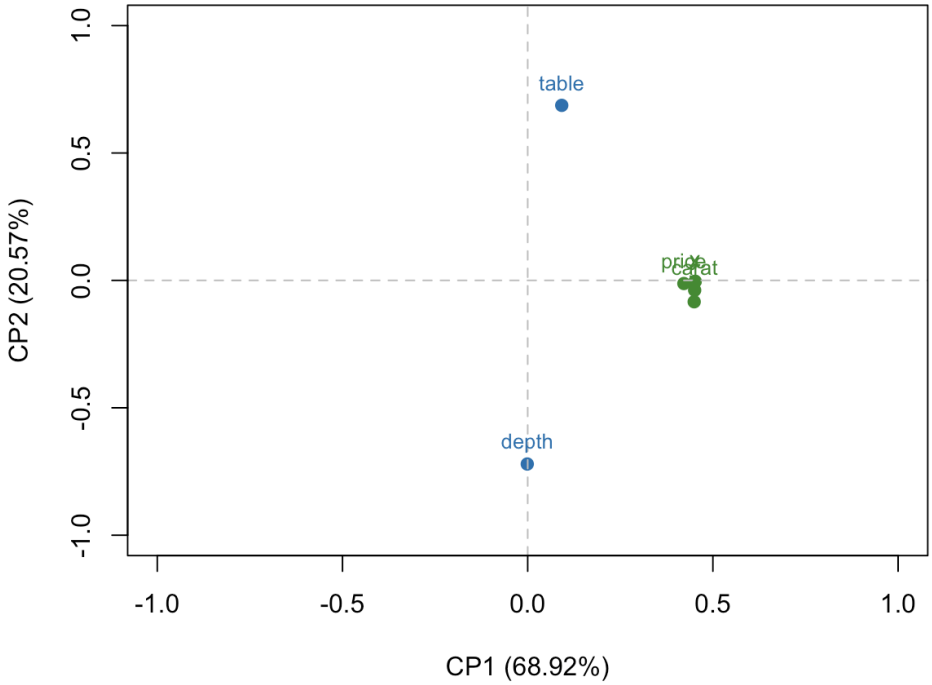

Figura 6.2: Gráfico da CP1 vs CP2(loadings não rodados para todas as variáveis)

O gráfico da figura 6.2 permite auxiliar visualmente à comparação dos pesos de cada variável nas respetivas CP. O gráfico evidencia que a primeira componente principal (CP1), responsável por 68,92% da variância total, está fortemente associada às variáveis *carat*, *price*, *x*, *y* e *z*, refletindo essencialmente a dimensão de tamanho e valor do diamante. A segunda componente (CP2), por sua vez, representa 20,57% da variância e está principalmente relacionada com as variáveis *depth* e *table*, sugerindo uma dimensão associada às proporções geométricas do corte. As cores no gráfico reforçam esta distinção, identificando visualmente a componente mais representativa de cada variável.

Tabela 6.4: Scores das CP retidas para todas as variáveis não rodadas

Observação	**CP1**	**CP2**
1	-2.947	-0.001
2	-2.735	-0.016
3	-0.166	0.801
4	2.246	-0.266
5	-0.219	-0.329
6	-1.386	-0.189

Na tabela 6.4 estão representados os *scores* das CP das 6 primeiras observações. Estes *scores* são nada mais do que os valores de cada indivíduo para cada componente. Cada *score* é calculado a partir da soma dos produtos entre o peso de cada variável e o valor observado estandardizado do indivíduo em cada variável original. No anexo A, é possível visulizar os *scores* das 200 observações.

6.4 Biplot

O *biplot* é uma representação gráfica associada à Análise de Componentes Principais (ACP), que permite visualizar simultaneamente as observações (indivíduos) e as variáveis num mesmo plano. O prefixo "bi" refere-se à sobreposição destas duas projeções: os *scores* dos indivíduos nas componentes principais e os pesos das variáveis (ou *loadings*). Esta sobreposição facilita a interpretação dos dados, embora seja importante reconhecer que as nuvens de indivíduos e variáveis têm significados distintos. As proximidades entre um indivíduo e uma variável não têm um valor matemático rigoroso, mas podem fornecer pistas úteis quando analisadas à luz das correlações entre variáveis e componentes.

A direção de cada seta indica com que componente principal a variável está mais correlacionada, enquanto o seu comprimento representa a contribuição relativa da variável para o plano formado pelas componentes. Assim, variáveis com setas mais longas e bem alinhadas com os eixos principais são as que mais influenciam a estrutura dos dados naquele espaço reduzido.

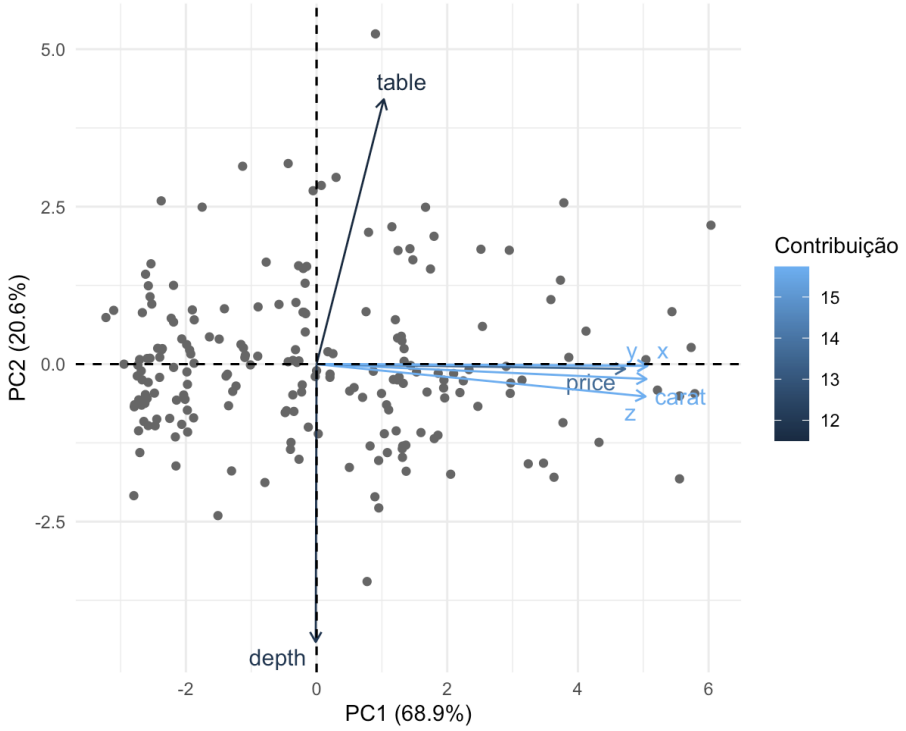

Figura 6.3: Biplot-CP1 vs CP2

O *biplot* obtido representa graficamente as observações e as variáveis projetadas no plano definido pelas duas primeiras componentes principais (CP1 e CP2), que explicam, respetivamente, 68,9% e 20,6% da variância total dos dados. No total, estas duas componentes retêm cerca de 89,5% da variância total.

A primeira componente principal (CP1) apresenta uma forte correlação com as variáveis *carat*, *price*, *x*, *y* e *z*, cujas setas se orientam aproximadamente na mesma direção. Este padrão sugere que CP1 capta essencialmente uma dimensão de volume e valor dos diamantes, refletindo a variabilidade associada ao seu tamanho físico e preço. Por outro lado, a segunda componente principal (CP2) está associada, de forma mais marcada, às variáveis *table* e *depth*, com orientações opostas, indicando uma correlação negativa entre estas duas variáveis. Esta componente parece capturar aspetos relacionados com proporções geométricas, nomeadamente o corte e a profundidade relativa dos diamantes.

A coloração das setas no gráfico representa a contribuição, em percentagem, relativa de cada variável para o plano CP1 vs CP2. Variáveis com coloração mais escura, como *table* e *depth*, apresentam menor influência na estrutura da variância representada.

Note-se que a contribuição de uma variável para uma componente principal quantifica o quanto essa variável influencia a formação dessa componente. Matematicamente, ela é calculada a partir do quadrado do coeficiente de

correlação (*loading*) da variável com a componente, normalizado pela variância explicada por essa componente (isto é, seu valor prórpio). Esta medida permite identificar quais variáveis têm maior peso na explicação da variância capturada por cada componente principal. Visualmente, em um **biplot**, isso se traduz em vetores mais longos e alinhados com os eixos principais, indicando maior influência na direção correspondente.

Relativamente às observações (representadas por pontos), verifica-se uma maior dispersão ao longo da CP1, o que reforça a preponderância desta componente na explicação da variabilidade presente no conjunto de dados.

6.5 Rotação Ortogonal das Componentes Principais (CP)

A rotação ortogonal das componentes principais é um procedimento frequentemente aplicado após a extração das componentes, com o objetivo de facilitar a interpretação da estrutura fatorial. Segundo Thurstone (1947), a rotação visa transformar a matriz de pesos (ou *loadings*) numa estrutura simplificada, em que cada variável esteja fortemente associada a uma única componente, enquanto mantém pesos residuais nas restantes.

Para que essa estrutura seja considerada simplificada, devem ser satisfeitas algumas condições propostas por Harman (1976), nomeadamente:

- Cada linha da matriz de pesos deve conter, pelo menos, um valor próximo de zero, indicando que cada variável não se correlaciona com pelo menos uma das componentes;
- Cada coluna da matriz deve ter, pelo menos, tantos zeros quanto o número de componentes retidas;
- Para qualquer par de colunas, deve existir um conjunto de variáveis com pesos próximos de zero numa coluna e não na outra, sendo desejável, em casos com mais de quatro componentes, que cada coluna apresente mais coeficientes nulos do que não nulos.

O método de rotação mais utilizado é o *Varimax*, proposto por Kaiser (1958). Trata-se de uma rotação ortogonal que preserva a independência entre componentes, e tem como objetivo maximizar a variância dos quadrados dos pesos em cada componente. Em termos práticos, isso significa tornar os pesos elevados ainda mais pronunciados e os restantes próximos de zero, facilitando a associação clara entre variáveis e componentes. Esta técnica resulta numa nova base de componentes que continua não correlacionada entre si, mas cuja interpretação se torna mais evidente: valores de peso próximos de 1 ou -1 indicam uma forte associação entre a variável e a componente; por outro lado, pesos próximos de zero sugerem fraca contribuição dessa variável. Em geral, são considerados como significativos os pesos de módulo igual ou superior a 0,5.

Tabela 6.5: Matriz dos Pesos antes e depois da Rotação Varimax.

Variável	CP1	CP2	CP1_rot	CP2_rot
carat	0.451	-0.039	0.452	-0.012
depth	-0.001	-0.720	0.040	-0.720
table	0.092	0.687	0.052	0.691
price	0.422	-0.012	0.423	0.012
x	0.452	-0.005	0.451	0.021
y	0.452	-0.004	0.451	0.021
z	0.449	-0.083	0.454	-0.058

A Tabela 6.5 apresenta a matriz dos pesos das variáveis nas duas primeiras componentes principais, antes e depois da aplicação da rotação ortogonal *Varimax*. Os valores realçados a azul correspondem a pesos com módulo superior a 0,5, indicando uma forte associação entre a variável e a respetiva componente. Antes da rotação, observa-se que as variáveis *depth* e *table* já estavam fortemente associadas à segunda componente (CP2), enquanto *carat*, *price*, *x*, *y* e *z* apresentavam pesos mais elevados na primeira componente (CP1).

Após a aplicação da rotação *Varimax*, esta estrutura torna-se ainda mais clara: *depth* e *table* mantêm-se destacadas na $CP2_{rot}$, enquanto as demais variáveis mantêm o seu peso na $CP1_{rot}$, com pouca redistribuição entre componentes. Este resultado evidencia que a rotação não altera a variância explicada, mas reorganiza os

loadings para favorecer uma estrutura mais simples, na qual cada variável contribui principalmente para uma única componente.

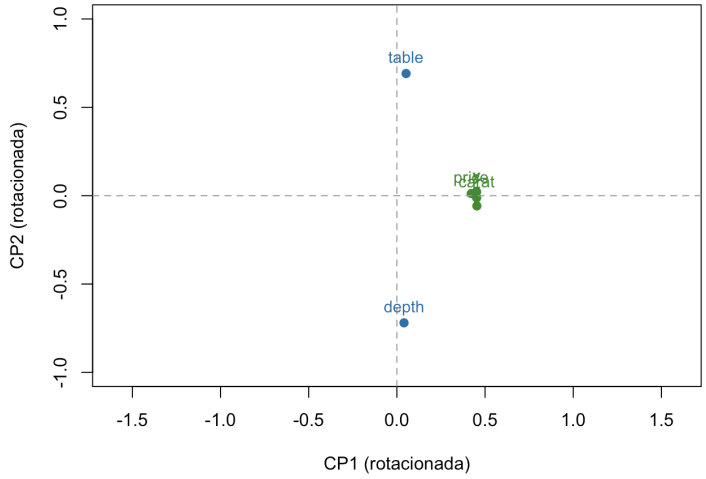

Figura 6.4: Gráfico da CP1 vs CP2(loadings rodados para todas as variáveis)

Neste caso, a aplicação da rotação *Varimax* não alterou significativamente a estrutura observada nos *loadings* originais. As variáveis já se encontravam bem diferenciadas entre as componentes principais, pelo que a rotação serviu essencialmente para reforçar a simplicidade interpretativa da estrutura fatorial.

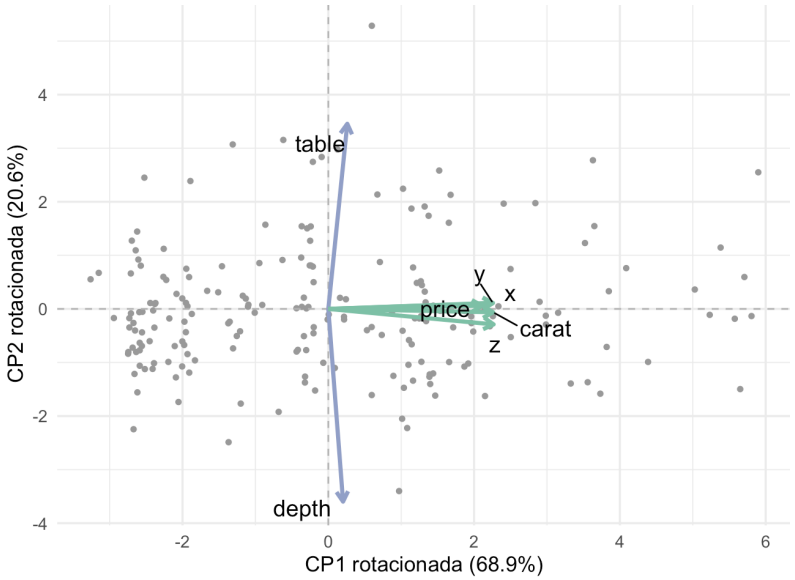

Figura 6.5: Biplot - CP1 vs CP2 após rotação Varimax

Apesar da aplicação da rotação ortogonal *Varimax* às componentes principais retidas, observa-se que a configuração do *biplot* se manteve praticamente inalterada. Tal deve-se ao facto de que, mesmo antes da rotação, as variáveis já se encontravam bem agrupadas por componente: *carat*, *x*, *y*, *z* e *price* associavam-se fortemente à primeira componente, enquanto *depth* e *table* apresentavam maior alinhamento com a segunda. Assim, a rotação apenas refinou ligeiramente os pesos, sem alterar significativamente a estrutura geométrica nem a interpretação das componentes principais.

7 Análise de Clusters

"Em vez de amor, dinheiro, fé, fama, ou beleza... deem-me a verdade."
Henry Thoreau

Nesta secção será realizada uma análise de *clusters* com o objetivo de identificar grupos homogéneos de observações com base nas variáveis quantitativas disponíveis. Serão aplicados tanto métodos hierárquicos como não-hierárquicos, bem como uma análise exploratória dos agrupamentos formados.

7.1 Clustering de Variáveis

Antes de proceder ao agrupamento das observações, é útil estudar as relações entre as variáveis. Para isso, é possível utilizar métodos de agrupamento hierárquico aplicados às variáveis em vez das observações. O objetivo é identificar grupos de variáveis que estejam fortemente associadas entre si.

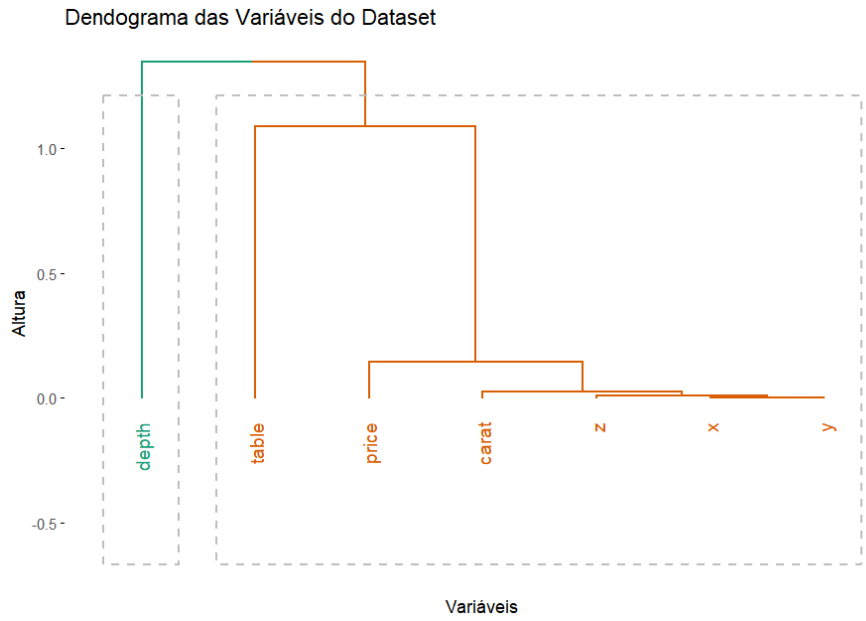

Figura 7.1: Dendograma das variáveis quantitativas

Este dendograma parece indicar que as variáveis estão associadas entre si. Não é chocante dado a relação entre as variáveis relacionadas com a física e as dimensões do diamante (*x, y, z, depth, table, carat*) e o seu valor monetário (*price*).

7.2 Clustering Hierárquico

O *clustering* hierárquico é um método que constrói uma hierarquia de grupos. Existem duas abordagens principais: aglomerativa (*bottom-up*) e divisiva (*top-down*). A abordagem aglomerativa, que será abordada inicialmente, começa por considerar cada observação como um *cluster* separado e, a cada iteração, funde os dois *clusters* mais semelhantes até que reste apenas um. A representação gráfica do processo é feita através de um dendrograma.

7.2.1 Método Aglomerativo

Neste subcapítulo, será aplicada a abordagem aglomerativa de agrupamento hierárquico. Vários métodos de ligação podem ser considerados, como o método da média, do vizinho mais próximo, do centróide, entre outros. Neste

trabalho, será utilizado o método de ligação `ward.D` em conjunto com uma medida de distância euclidiana, uma vez que a junção dos dois originou os melhores resultados.

Existem diferentes métodos para determinar o número ótimo de clusters, entre os quais se destacam:

- **Método do Cotovelo (*Elbow Method*)**: avalia a redução da variância intra-cluster em função do número de clusters k, procurando um ponto de inflexão onde a melhoria começa a ser marginal.
- **Método da Silhueta (*Silhouette Method*)**: avalia a coesão interna de cada cluster e a separação entre clusters, maximizando a largura média da silhueta.

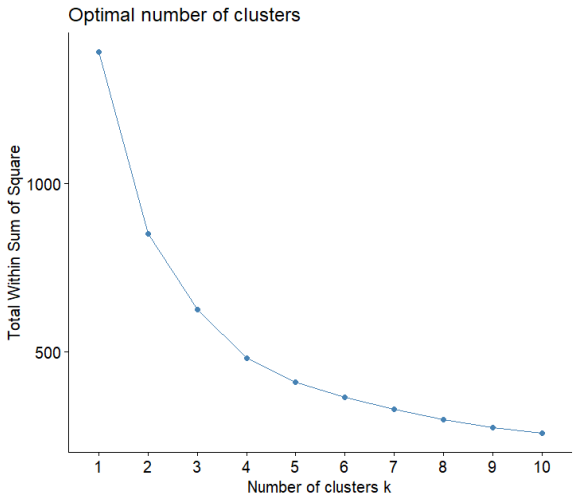

Figura 7.2: Método do Cotovelo para número ótimo de clusters no método Aglomerativo

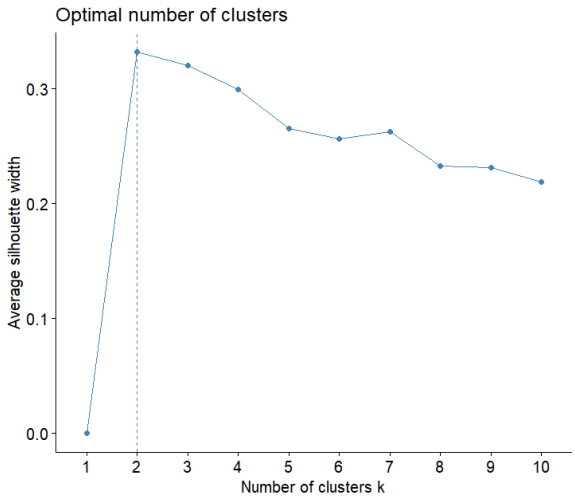

Figura 7.3: Método *Silhouette* para número ótimo de clusters no método Aglomerativo

O método do cotovelo parece sugerir que o número ótimo de clusters é entre 3 e 4, já o método *Silhouette* indica 2 clusters como o número ótimo, não obstante 3 ou 4 clusters parecerem também viáveis.

Tendo em conta os resultados obtidos através destes métodos, estão de seguida representados os dendogramas e os gráficos dos clusters nas duas principais componentes para 2, 3 e 4 clusters.

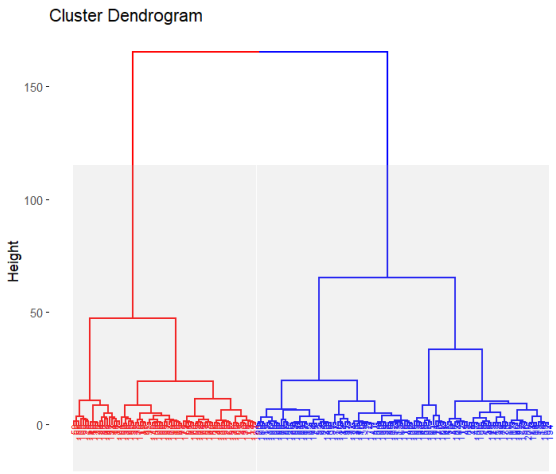

Figura 7.4: Dendograma das observações para 2 clusters

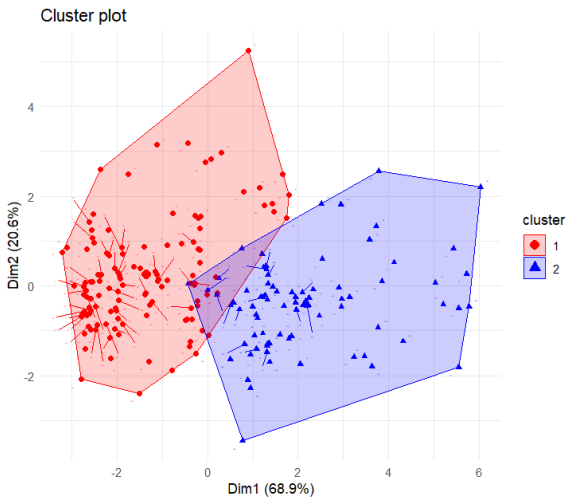

Figura 7.5: 2 Clusters nas duas principais componentes

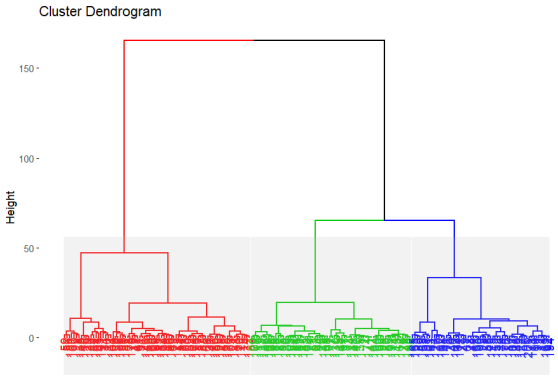

Figura 7.6: Dendograma das observações para 3 clusters

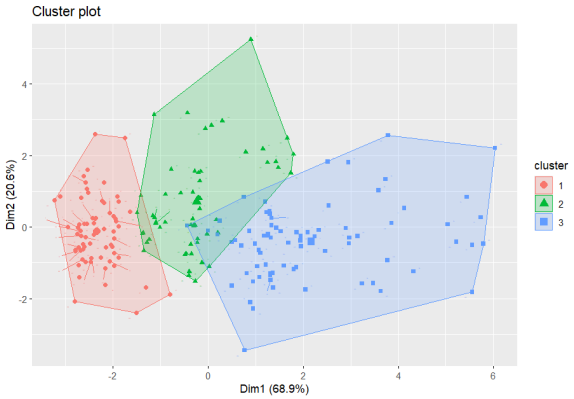

Figura 7.7: 3 Clusters nas duas principais componentes

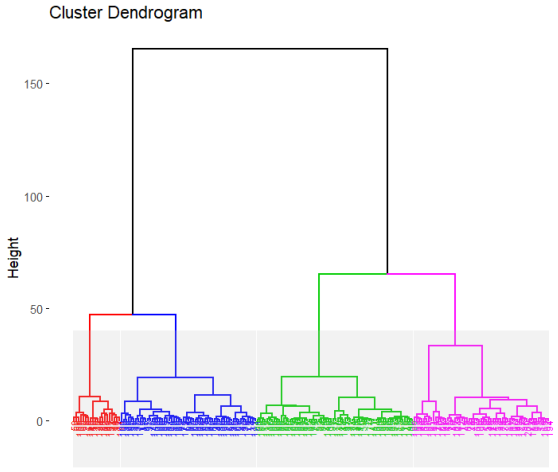

Figura 7.8: Dendograma das observações para 4 clusters

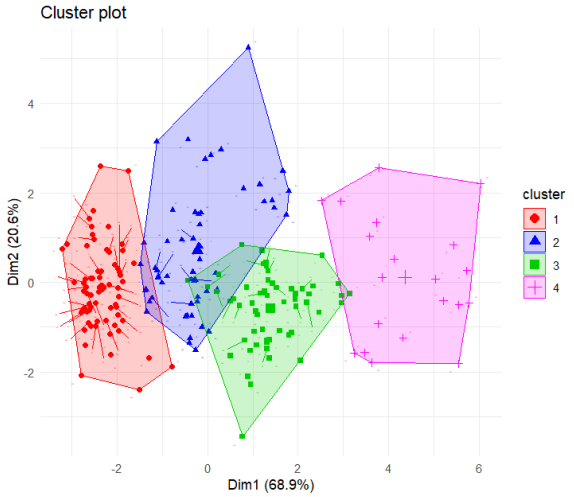

Figura 7.9: 4 Clusters nas duas principais componentes

O método hierárquico revelou que é possível dividir as observações, agrupando as mesmas por clusters, originando grupos distintos, ainda que por vezes sobrepostos. É possível que isto ocorra dado à presença de três variáveis qualitativas onde na totalidade estão presentes 20 categorias. No entanto, vai ser possível verificar no próximo capítulo, que o método de *K-Means* obteve uma melhor separação dos grupos, pelo que irá ser feita uma análise mais profunda dos clusters obtidos nesse tópico. Não obstante é possível visualizar as tabelas sumário para o método aglomerativo no segundo anexo.

7.3 Clustering Não-Hierárquico

7.3.1 Método *K-Means*

O método *K-Means* é um algoritmo de agrupamento não hierárquico baseado na minimização da variância dentro dos clusters. Dado um número fixo de clusters k, o objetivo é dividir os dados em k grupos de forma a que a soma das distâncias quadradas entre as observações e o centro do seu respetivo cluster seja minimizada.

Serão aplicados os mesmos métodos já utilizados para sugerir o valor ótimo de k e posteriormente será realizada a aplicação do algoritmo *K-means* para diferentes valores de k, analisando-se a qualidade dos agrupamentos formados.

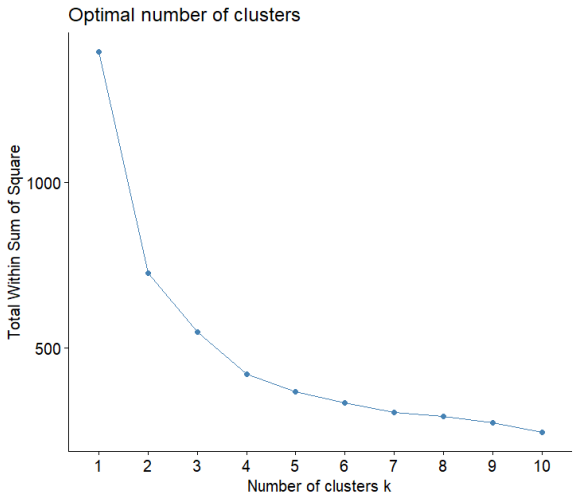

Figura 7.10: Método do Cotovelo para número ótimo de clusters no método K-means

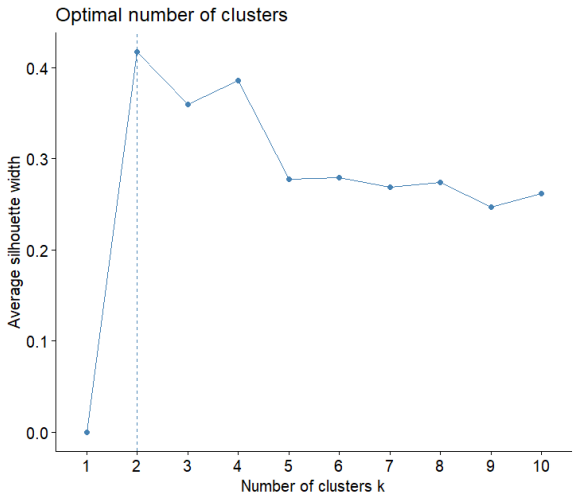

Figura 7.11: Método *Silhouette* para número ótimo de clusters no método K-means

O método do cotovelo parece sugerir que o número ótimo de clusters é entre 2 e 3, já o método *Silhouette* indica 2 clusters como o número ótimo, não obstante 3 ou 4 clusters parecerem também viáveis, sendo que neste método 4 clusters até obtém melhor resultado que 3 clusters.

Tendo em conta os resultados obtidos através destes métodos, estão de seguida representados os gráficos dos clusters nas duas principais componentes para 2, 3 e 4 clusters.

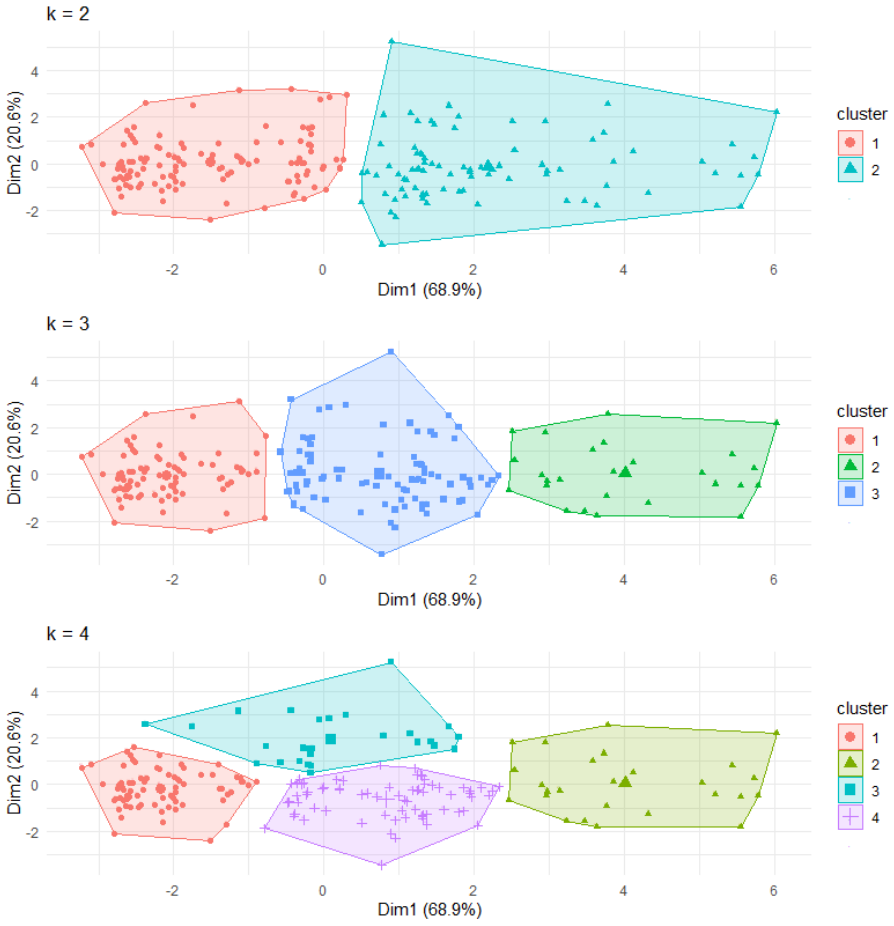

Figura 7.12: Gráficos do K-means para 2, 3 e 4 Clusters nas duas componentes principais

O método de K-Means obtém uma separação mais clara entre os diferentes grupos. Em nenhum dos três casos os clusters se intercetam, o que é um indicador bastante positivo.

Na divisão em 2 clusters para a divisão em 3 clusters foi criado um grupo entre os dois grupos já existentes, sendo que o número de observações incluídas nestes diminuiu.

Do segundo para o terceiro gráfico o último grupo apenas aumentou em duas observações, mantendo-se praticamente constante. Os restantes dois grupos já existentes transformaram-se em três, sendo que o azul tem a segunda componente superior aos outros dois.

Tabela 7.1: Tabela descritiva das variáveis pelos 2 clusters. Para as variáveis quantitativas está representado a média e o desvio-padrão entre parênteses. Para as variáveis qualitativas está representado a frequência absoluta e a relativa (para cada cluster), em percentagem.

	1 N=82	2 N=118
carat	1.26 (0.36)	0.50 (0.18)
cut:		
Fair	2 (2.44%)	2 (1.69%)
Good	7 (8.54%)	8 (6.78%)
Very Good	26 (31.7%)	15 (12.7%)
Premium	19 (23.2%)	29 (24.6%)
Ideal	28 (34.1%)	64 (54.2%)
color:		
D	6 (7.32%)	17 (14.4%)
E	8 (9.76%)	28 (23.7%)
F	16 (19.5%)	18 (15.3%)
G	22 (26.8%)	24 (20.3%)
H	13 (15.9%)	17 (14.4%)
I	10 (12.2%)	12 (10.2%)
J	7 (8.54%)	2 (1.69%)
clarity:		
I1	4 (4.88%)	0 (0.00%)
SI2	25 (30.5%)	9 (7.63%)
SI1	21 (25.6%)	29 (24.6%)
VS2	20 (24.4%)	28 (23.7%)
VS1	7 (8.54%)	17 (14.4%)
VVS2	3 (3.66%)	10 (8.47%)
VVS1	1 (1.22%)	18 (15.3%)
IF	1 (1.22%)	7 (5.93%)
depth	62.0 (1.43)	61.7 (1.12)
table	57.5 (2.43)	57.0 (1.94)
price	7288 (3571)	1567 (910)
x	6.87 (0.62)	5.03 (0.61)
y	6.86 (0.61)	5.03 (0.60)
z	4.25 (0.38)	3.10 (0.37)

Pelos dados da tabela podemos verificar que a variável *cut* influencia a divisão dos grupos, sendo que o segundo cluster contém maior percentagem de diamantes com melhor corte (*Premium* e *Ideal* - 79% contra 57% do primeiro grupo). No segundo grupo a percentagem de diamantes com as melhores cores (D e E) também é bastante superior (aprox. 38% contra 17%). O mesmo acontece com a claridade, observando-se maior número de diamantes com melhor claridade neste grupo (aprox. 30% contra 6%).

Nas variáveis quantitativas é de realçar a diferença do valor médio do preço entre os dois grupos (7288$ do grupo 1 contra 1567$ do grupo 2) e da diferença do peso (1.26 carat contra 0.50 carat).

Tabela 7.2: Tabela descritiva das variáveis pelos 3 clusters.

	1 N=92	2 N=82	3 N=26
carat	0.93 (0.18)	0.39 (0.10)	1.71 (0.31)
cut:			
Fair	3 (3.26%)	1 (1.22%)	0 (0.00%)
Good	7 (7.61%)	5 (6.10%)	3 (11.5%)
Very Good	23 (25.0%)	11 (13.4%)	7 (26.9%)
Premium	26 (28.3%)	15 (18.3%)	7 (26.9%)
Ideal	33 (35.9%)	50 (61.0%)	9 (34.6%)
color:			
D	9 (9.78%)	13 (15.9%)	1 (3.85%)
E	13 (14.1%)	22 (26.8%)	1 (3.85%)
F	17 (18.5%)	12 (14.6%)	5 (19.2%)
G	22 (23.9%)	17 (20.7%)	7 (26.9%)
H	12 (13.0%)	12 (14.6%)	6 (23.1%)
I	13 (14.1%)	5 (6.10%)	4 (15.4%)
J	6 (6.52%)	1 (1.22%)	2 (7.69%)
clarity:			
I1	2 (2.17%)	0 (0.00%)	2 (7.69%)
SI2	21 (22.8%)	5 (6.10%)	8 (30.8%)
SI1	24 (26.1%)	17 (20.7%)	9 (34.6%)
VS2	30 (32.6%)	16 (19.5%)	2 (7.69%)
VS1	7 (7.61%)	13 (15.9%)	4 (15.4%)
VVS2	3 (3.26%)	9 (11.0%)	1 (3.85%)
VVS1	2 (2.17%)	17 (20.7%)	0 (0.00%)
IF	3 (3.26%)	5 (6.10%)	0 (0.00%)
depth	61.9 (1.44)	61.8 (1.02)	61.6 (1.29)
table	57.5 (2.31)	56.6 (1.75)	57.8 (2.50)
price	4395 (2061)	1090 (567)	11104 (3353)
x	6.23 (0.40)	4.68 (0.37)	7.65 (0.47)
y	6.22 (0.40)	4.69 (0.37)	7.62 (0.46)
z	3.85 (0.27)	2.90 (0.23)	4.71 (0.30)

Pelos dados da tabela podemos verificar que o segundo grupo possui maior percentagem de diamantes com corte *Premium* ou *Ideal* (aprox. 79% contra 64% e 62% dos grupos 1 e 3 respetivamente) e cor melhor - D ou E (aprox. 42% contra 25% e 8% dos grupos 1 e 3 respetivamente).

Nas variáveis quantitativas é de realçar a diferença do valor médio do peso entre os três grupos (0.93, 0.39 e 1.71 carat), sendo que o grupo 2 possui os diamantes com peso inferior e o grupo 3 com peso superior. Simultaneamente o valor médio do preço do grupo 2 (1090$) também é bastante inferior ao dos outros (4395$ e 11104$) onde o grupo 3 possui o valor médio do preço mais elevado.

Tabela 7.3: Tabela descritiva das variáveis pelos 4 clusters.

	1 N=26	2 N=76	3 N=70	4 N=28
carat	1.71 (0.31)	0.38 (0.09)	0.96 (0.18)	0.77 (0.21)
cut:				
Fair	0 (0.00%)	1 (1.32%)	2 (2.86%)	1 (3.57%)
Good	3 (11.5%)	5 (6.58%)	5 (7.14%)	2 (7.14%)
Very Good	7 (26.9%)	9 (11.8%)	16 (22.9%)	9 (32.1%)
Premium	7 (26.9%)	12 (15.8%)	17 (24.3%)	12 (42.9%)
Ideal	9 (34.6%)	49 (64.5%)	30 (42.9%)	4 (14.3%)

continued on next page

Tabela 7.3 – *continued from previous page*

	1 N=26	2 N=76	3 N=70	4 N=28
color:				
D	1 (3.85%)	13 (17.1%)	7 (10.0%)	2 (7.14%)
E	1 (3.85%)	19 (25.0%)	6 (8.57%)	10 (35.7%)
F	5 (19.2%)	10 (13.2%)	11 (15.7%)	8 (28.6%)
G	7 (26.9%)	16 (21.1%)	18 (25.7%)	5 (17.9%)
H	6 (23.1%)	12 (15.8%)	11 (15.7%)	1 (3.57%)
I	4 (15.4%)	5 (6.58%)	12 (17.1%)	1 (3.57%)
J	2 (7.69%)	1 (1.32%)	5 (7.14%)	1 (3.57%)
clarity:				
I1	2 (7.69%)	0 (0.00%)	2 (2.86%)	0 (0.00%)
SI2	8 (30.8%)	5 (6.58%)	16 (22.9%)	5 (17.9%)
SI1	9 (34.6%)	15 (19.7%)	18 (25.7%)	8 (28.6%)
VS2	2 (7.69%)	13 (17.1%)	23 (32.9%)	10 (35.7%)
VS1	4 (15.4%)	12 (15.8%)	6 (8.57%)	2 (7.14%)
VVS2	1 (3.85%)	9 (11.8%)	1 (1.43%)	2 (7.14%)
VVS1	0 (0.00%)	17 (22.4%)	2 (2.86%)	0 (0.00%)
IF	0 (0.00%)	5 (6.58%)	2 (2.86%)	1 (3.57%)
depth	61.6 (1.29)	61.9 (0.89)	62.5 (0.88)	60.0 (1.05)
table	57.8 (2.50)	56.4 (1.48)	56.7 (1.81)	60.0 (1.95)
price	11104 (3353)	1052 (540)	4561 (2124)	3376 (1874)
x	7.65 (0.47)	4.65 (0.34)	6.27 (0.40)	5.90 (0.54)
y	7.62 (0.46)	4.65 (0.34)	6.26 (0.40)	5.90 (0.56)
z	4.71 (0.30)	2.88 (0.22)	3.92 (0.25)	3.54 (0.33)

Pelos dados da tabela podemos verificar que o segundo grupo possui maior percentagem de diamantes com corte *Premium* ou *Ideal* (aprox. 80% contra 62%, 67% e 57% dos grupos 1, 3 e 4 respetivamente). A nível de cor os grupos 2 e 4 possuem maior percentagem de diamantes com cor D ou E (aprox. 42% contra 8%, 19% e dos grupos 1 e 3 respetivamente). Na variável *Clarity* o grupo 2 possui 22 observações nos dois níveis de melhor claridade quando o grupo 1 possui 0, o grupo 3 possui 4 e o grupo 4 apenas uma.

Nas variáveis quantitativas é de realçar a diferença do valor médio do peso entre os três grupos (1.71, 0.38, 0.96 e 0.77 carat), sendo que o grupo 2 possui os diamantes com peso inferior e o grupo 1 com peso superior. Em concordância o valor médio do preço do grupo 2 (1052$) também é bastante inferior ao dos outros (11104$, 4561$ e 3376$) onde o grupo 1 possui o valor médio do preço mais elevado.

8 Conclusão

Com base nos resultados obtidos, conclui-se que as variáveis qualitativas *cut*, *color* e *clarity* influenciam significativamente os vetores de médias das variáveis quantitativas em estudo. Através da MANOVA, foi possível confirmar diferenças estatisticamente significativas entre os grupos de cada variável, sendo a variável *clarity* a que evidenciou maior impacto na estrutura multivariada dos dados. Estes resultados reforçam a utilidade da análise multivariada na identificação de padrões relevantes em contextos com múltiplas variáveis dependentes.

Acerca da Análise Discriminante, esta revelou diferentes níveis de eficácia na classificação das variáveis qualitativas. A variável *cut* destacou-se pelo melhor desempenho do modelo, enquanto *color* apresentou maior dificuldade na separação entre grupos, refletindo-se numa maior sobreposição. Já a variável *clarity* mostrou uma capacidade moderada de discriminação, embora com algumas categorias menos bem distinguidas. Estes resultados sugerem que a separação entre classes nem sempre é clara, podendo justificar a utilização de variáveis adicionais ou métodos complementares.

A Análise de Componentes Principais revelou-se uma abordagem eficaz para a redução dimensional dos dados, permitindo condensar a informação das sete variáveis originais em apenas duas componentes principais, que explicam conjuntamente 89.5% da variância total. A CP1 agrega as variáveis relacionadas com as dimensões físicas e o valor do diamante (*carat*, *price*, *x*, *y*, *z*), enquanto a CP2 está fortemente associada às proporções geométricas (*table* e *depth*). O biplot confirmou visualmente esta separação clara.

Embora tenha sido aplicada a rotação ortogonal Varimax com o intuito de facilitar a interpretação, o seu impacto foi praticamente nulo, dada a já evidente estrutura dos dados nas componentes não rotacionadas. As variáveis já se encontravam fortemente agrupadas nas componentes principais originais, pelo que a rotação não acrescentou valor interpretativo substancial.

Conclui-se, portanto, que a Análise de Componentes Principais permitiu identificar duas dimensões fundamentais subjacentes à estrutura dos dados dos diamantes — uma ligada ao tamanho e valor, outra às proporções — proporcionando uma visão sintética dos padrões de variabilidade presentes no conjunto de dados.

A análise de *clusters* foi efetuada recorrendo a um método hierárquico aglomerativo (*bottom-up*) e a um não hierárquico (*K-means*). Ambos os métodos permitiram segmentar os diamantes em grupos com características semelhantes. Os melhores resultados foram obtidos utilizando o método *K-means*. Os métodos *elbow* e da *silhouette* foram aplicados para auxiliar na escolha do número ótimo de grupos.

A análise com o método *K-means* permitiu observar que os *clusters* se distinguem essencialmente pelo *peso (carat)* e pelo *preço* dos diamantes, bem como por variáveis qualitativas como *cut*, *color* e *clarity*. De forma consistente, os diamantes com menor peso e menor qualidade de corte, cor e clareza tendem a concentrar-se num *cluster* com preços significativamente mais baixos. Por outro lado, os diamantes de maior peso e qualidade reúnem-se num ou mais *clusters* associados a preços mais elevados.

Ao aumentar o número de *clusters*, foi possível refinar a segmentação, distinguindo grupos com características específicas, como diamantes com excelente corte e cor, mas com peso e preço moderados. Esta estrutura de agrupamento reflete padrões comerciais reais na indústria dos diamantes, em que múltiplos fatores influenciam o valor do produto final.

Assim, conclui-se que o método *K-means* foi eficaz na identificação de grupos homogéneos de diamantes e na revelação de padrões relevantes para a análise e interpretação dos dados.

Para além disto, como estudantes, este trabalho representou o exercício de lapidar o nosso próprio olhar analítico. A aplicação dos métodos de Análise Multivariada foi essencial não apenas para extrair padrões relevantes, mas sobretudo para desenvolver competências críticas, pensamento estruturado e sensibilidade Estatística. Cada técnica aplicada foi também um passo no fortalecimento da nossa autonomia. No fim, mais do que respostas, ganhámos ferramentas e com elas, a capacidade de transformar complexidade em clareza.

9 Referências

1. Reis, E. *Estatística Multivariada Aplicada*. Edições Sílabo, 2001.
2. Branco, J. A. *Uma Introdução à Análise de Clusters*. Sociedade Portuguesa de Estatística, Évora, 2004.
3. Serra, C., *Slides sobre MANOVA*, unidade curricular de Análise Estatística Multivariada, Universidade do Minho, ano letivo 2019/2020.
4. Serra, C., *Slides de Análise Discriminante*, unidade curricular de Análise Estatística Multivariada, Universidade do Minho, ano letivo 2019/2020.
5. Hélder Armando Gonçalves Pinto. *Analysis of state.x77 Dataset*. University of Aveiro, MAP-PDMA – Multivariate Analysis and Statistical Learning, 2023.
6. Nuno André Costa do Vale. *Análise Estatística Multivariada no Estudo de Dietas Sustentáveis para o Linguado Senegalês*. Universidade do Minho, Licenciatura em Estatística Aplicada, 2013/2014.
7. Jolliffe, I. T. *Principal Component Analysis*. Springer Series in Statistics, Springer-Verlag, 1986.
8. Chavent, M. *Exemple d'interprétation d'une ACP*. Disponível em: `https://marie-chavent.perso.math.cnrs.fr/wp-content/uploads/2013/10/Exemple_interpret_ACP.pdf`. Acesso em: abril de 2025.

A Primeiro anexo

Tabela A.1: Scores das CP retidas para todas as variáveis não rodadas.

Obs	CP1	CP2	Obs	CP1	CP2	Obs	CP1	CP2	Obs	CP1	CP2
1	-2.95	-0.00	51	2.05	-1.75	101	1.47	1.66	151	2.52	1.82
2	-2.73	-0.01	52	1.21	0.70	102	0.57	-0.37	152	-1.01	-0.01
3	-0.17	0.80	53	-2.68	-0.65	103	3.14	-0.25	153	1.60	-1.09
4	2.25	-0.27	54	1.34	0.04	104	5.22	-0.41	154	-2.01	-0.56
5	-0.22	-0.33	55	0.80	2.09	105	-1.97	-1.08	155	-0.32	0.23
6	-1.39	-0.19	56	2.95	1.81	106	-2.19	0.67	156	-2.03	-0.49
7	-2.38	2.59	57	-0.43	0.04	107	-1.16	0.31	157	-1.28	-0.43
8	-0.17	0.51	58	-0.36	-0.49	108	-2.64	-0.91	158	3.86	0.11
9	4.12	0.52	59	3.59	1.02	109	0.77	-3.45	159	-2.59	0.09
10	5.73	0.27	60	-2.55	1.07	110	-2.40	0.11	160	3.78	2.56
11	-1.10	0.10	61	2.47	-0.67	111	-2.43	0.25	161	1.74	1.51
12	-0.13	-1.00	62	-2.19	-0.05	112	-0.20	1.52	162	-0.05	2.75
13	-1.36	-0.16	63	-2.62	-0.48	113	-2.36	0.25	163	-0.40	-1.35
14	0.90	5.24	64	-0.32	0.98	114	-0.02	-0.20	164	-2.58	1.24
15	2.54	0.60	65	-0.43	3.19	115	-2.62	1.43	165	1.67	2.49
16	1.31	-1.48	66	1.07	-0.65	116	3.77	-0.93	166	0.20	-0.21
17	-2.63	-0.62	67	-0.77	1.62	117	-2.16	-1.16	167	-0.15	1.55
18	-0.39	-1.24	68	4.32	-1.24	118	-1.98	-0.12	168	-1.98	-0.32
19	2.96	-0.46	69	1.31	-1.34	119	-2.48	-0.46	169	1.32	-1.30
20	-2.47	-0.98	70	0.07	2.84	120	3.24	-1.58	170	2.90	-0.04
21	5.56	-0.51	71	0.21	-0.16	121	2.97	-0.30	171	-2.40	-0.21
22	-1.09	0.14	72	-2.55	0.09	122	-1.35	-0.66	172	2.10	-0.15
23	-2.79	-0.68	73	1.37	-1.29	123	-3.22	0.74	173	1.94	-0.38
24	5.44	0.83	74	-0.28	1.56	124	-2.58	-0.29	174	-2.59	-0.54
25	-2.45	-0.87	75	-0.17	1.28	125	-0.46	-0.74	175	0.76	0.83
26	-0.19	0.82	76	1.86	-1.13	126	-1.51	-2.40	176	-2.80	-2.09
27	-2.79	-0.65	77	-1.41	0.88	127	-1.13	3.14	177	-1.75	2.49
28	3.73	1.33	78	1.33	-0.30	128	1.30	0.44	178	0.87	-0.11
29	3.47	-1.57	79	-2.19	1.25	129	-1.94	0.22	179	1.95	-0.25
30	1.86	-0.15	80	-1.64	0.43	130	-3.10	0.85	180	-0.90	0.91
31	0.99	-0.47	81	1.31	0.37	131	1.22	-1.06	181	-1.98	-0.73
32	0.51	-0.43	82	-2.23	0.73	132	-2.52	0.95	182	5.79	-0.47
33	2.33	-0.09	83	1.80	2.03	133	1.43	1.83	183	0.00	-0.10
34	5.04	0.07	84	-1.88	-0.85	134	1.37	-1.70	184	1.69	-0.44
35	-2.06	0.40	85	-1.30	-1.70	135	1.24	0.41	185	-2.25	-0.86
36	-2.06	-0.96	86	-0.90	0.13	136	0.95	-2.28	186	1.18	-0.24
37	-2.37	0.23	87	-0.79	-1.88	137	-2.14	-0.57	187	1.96	-0.54
38	1.53	-0.12	88	-2.71	-1.40	138	-2.68	-0.11	188	-1.99	0.31
39	0.89	-2.11	89	-2.15	-1.62	139	-1.49	0.40	189	0.82	-1.30
40	1.24	-0.25	90	-0.27	0.05	140	3.63	-1.79	190	-0.36	0.06
41	-2.67	-0.25	91	-0.23	-0.44	141	0.30	2.96	191	6.04	2.21
42	0.17	0.20	92	0.03	-1.11	142	-1.12	0.26	192	-1.93	0.16
43	-0.48	-0.77	93	-0.35	-0.75	143	5.55	-1.82	193	0.70	-0.53
44	-0.30	0.03	94	-2.72	-0.57	144	-2.53	1.59	194	-0.27	-1.51
45	0.25	0.17	95	1.80	-1.18	145	-2.58	-0.98	195	1.34	0.25
46	-0.21	0.82	96	-1.88	0.01	146	1.04	-1.10	196	1.10	-0.73
47	-2.71	0.07	97	-1.90	0.86	147	-2.74	-0.19	197	-2.72	-1.06
48	2.19	-0.45	98	1.15	2.18	148	-1.87	0.70	198	0.50	-1.64
49	-0.57	0.95	99	-2.67	0.82	149	1.25	1.80	199	1.08	-1.40
50	0.95	-1.53	100	1.26	-0.21	150	1.44	-0.02	200	-1.23	-0.35

B Segundo anexo

Tabela B.1: Tabela descritiva das variáveis pelos 2 clusters para o método aglomerativo. Para as variáveis quantitativas está representado a média e o desvio-padrão entre parênteses. Para as variáveis qualitativas está representado a frequência absoluta e a relativa (para cada cluster), em percentagem.

	1 N=123	2 N=77
carat	0.53 (0.22)	1.27 (0.38)
cut:		
Fair	3 (2.44%)	1 (1.30%)
Good	8 (6.50%)	7 (9.09%)
Very Good	20 (16.3%)	21 (27.3%)
Premium	28 (22.8%)	20 (26.0%)
Ideal	64 (52.0%)	28 (36.4%)
color:		
D	18 (14.6%)	5 (6.49%)
E	31 (25.2%)	5 (6.49%)
F	20 (16.3%)	14 (18.2%)
G	23 (18.7%)	23 (29.9%)
H	17 (13.8%)	13 (16.9%)
I	11 (8.94%)	11 (14.3%)
J	3 (2.44%)	6 (7.79%)
clarity:		
I1	0 (0.00%)	4 (5.19%)
SI2	10 (8.13%)	24 (31.2%)
SI1	29 (23.6%)	21 (27.3%)
VS2	30 (24.4%)	18 (23.4%)
VS1	17 (13.8%)	7 (9.09%)
VVS2	12 (9.76%)	1 (1.30%)
VVS1	18 (14.6%)	1 (1.30%)
IF	7 (5.69%)	1 (1.30%)
depth	61.5 (1.24)	62.3 (1.16)
table	57.2 (2.19)	57.2 (2.14)
price	1826 (1421)	7245 (3763)
x	5.11 (0.70)	6.86 (0.67)
y	5.11 (0.70)	6.84 (0.66)
z	3.14 (0.41)	4.27 (0.39)

Tabela B.2: Tabela descritiva das variáveis pelos 3 clusters para o método aglomerativo.

	1 N=66	2 N=57	3 N=77
carat	0.36 (0.07)	0.72 (0.16)	1.27 (0.38)
cut:			
Fair	1 (1.52%)	2 (3.51%)	1 (1.30%)
Good	5 (7.58%)	3 (5.26%)	7 (9.09%)
Very Good	11 (16.7%)	9 (15.8%)	21 (27.3%)
Premium	10 (15.2%)	18 (31.6%)	20 (26.0%)
Ideal	39 (59.1%)	25 (43.9%)	28 (36.4%)
color:			
D	11 (16.7%)	7 (12.3%)	5 (6.49%)

continued on next page

Tabela B.2 – *continued from previous page*

	1	2	3
	N=66	N=57	N=77
E	17 (25.8%)	14 (24.6%)	5 (6.49%)
F	8 (12.1%)	12 (21.1%)	14 (18.2%)
G	14 (21.2%)	9 (15.8%)	23 (29.9%)
H	10 (15.2%)	7 (12.3%)	13 (16.9%)
I	5 (7.58%)	6 (10.5%)	11 (14.3%)
J	1 (1.52%)	2 (3.51%)	6 (7.79%)
clarity:			
I1	0 (0.00%)	0 (0.00%)	4 (5.19%)
SI2	3 (4.55%)	7 (12.3%)	24 (31.2%)
SI1	13 (19.7%)	16 (28.1%)	21 (27.3%)
VS2	13 (19.7%)	17 (29.8%)	18 (23.4%)
VS1	9 (13.6%)	8 (14.0%)	7 (9.09%)
VVS2	8 (12.1%)	4 (7.02%)	1 (1.30%)
VVS1	15 (22.7%)	3 (5.26%)	1 (1.30%)
IF	5 (7.58%)	2 (3.51%)	1 (1.30%)
depth	61.9 (1.02)	61.1 (1.35)	62.3 (1.16)
table	56.4 (1.69)	58.1 (2.37)	57.2 (2.14)
price	922 (434)	2872 (1450)	7245 (3763)
x	4.55 (0.27)	5.75 (0.43)	6.86 (0.67)
y	4.56 (0.27)	5.75 (0.44)	6.84 (0.66)
z	2.82 (0.18)	3.51 (0.24)	4.27 (0.39)

Tabela B.3: Tabela descritiva das variáveis pelos 4 clusters para o método aglomerativo.

	1	2	3	4
	N=66	N=57	N=57	N=20
carat	0.36 (0.07)	0.72 (0.16)	1.08 (0.17)	1.79 (0.31)
cut:				
Fair	1 (1.52%)	2 (3.51%)	1 (1.75%)	0 (0.00%)
Good	5 (7.58%)	3 (5.26%)	4 (7.02%)	3 (15.0%)
Very Good	11 (16.7%)	9 (15.8%)	16 (28.1%)	5 (25.0%)
Premium	10 (15.2%)	18 (31.6%)	15 (26.3%)	5 (25.0%)
Ideal	39 (59.1%)	25 (43.9%)	21 (36.8%)	7 (35.0%)
color:				
D	11 (16.7%)	7 (12.3%)	4 (7.02%)	1 (5.00%)
E	17 (25.8%)	14 (24.6%)	4 (7.02%)	1 (5.00%)
F	8 (12.1%)	12 (21.1%)	9 (15.8%)	5 (25.0%)
G	14 (21.2%)	9 (15.8%)	17 (29.8%)	6 (30.0%)
H	10 (15.2%)	7 (12.3%)	10 (17.5%)	3 (15.0%)
I	5 (7.58%)	6 (10.5%)	8 (14.0%)	3 (15.0%)
J	1 (1.52%)	2 (3.51%)	5 (8.77%)	1 (5.00%)
clarity:				
I1	0 (0.00%)	0 (0.00%)	2 (3.51%)	2 (10.0%)
SI2	3 (4.55%)	7 (12.3%)	18 (31.6%)	6 (30.0%)
SI1	13 (19.7%)	16 (28.1%)	13 (22.8%)	8 (40.0%)
VS2	13 (19.7%)	17 (29.8%)	17 (29.8%)	1 (5.00%)
VS1	9 (13.6%)	8 (14.0%)	5 (8.77%)	2 (10.0%)
VVS2	8 (12.1%)	4 (7.02%)	0 (0.00%)	1 (5.00%)
VVS1	15 (22.7%)	3 (5.26%)	1 (1.75%)	0 (0.00%)
IF	5 (7.58%)	2 (3.51%)	1 (1.75%)	0 (0.00%)
depth	61.9 (1.02)	61.1 (1.35)	62.6 (0.93)	61.5 (1.36)

continued on next page

Tabela B.3 – *continued from previous page*

	1	2	3	4
	N=66	N=57	N=57	N=20
table	56.4 (1.69)	58.1 (2.37)	57.0 (1.93)	57.8 (2.63)
price	922 (434)	2872 (1450)	5575 (2176)	12005 (3236)
x	4.55 (0.27)	5.75 (0.43)	6.53 (0.36)	7.78 (0.45)
y	4.56 (0.27)	5.75 (0.44)	6.52 (0.35)	7.77 (0.43)
z	2.82 (0.18)	3.51 (0.24)	4.08 (0.21)	4.79 (0.29)