

**I. Pen-and-paper**

1)

$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, u_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, u_3 = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$   
 $u_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, u_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \pi_1 = 0.5, \pi_2 = 0.5$

E-Step

$K_1 = \text{Cluster 1 com } u_1, \Sigma_1 \in \Pi_1$   
 $K_2 = \text{Cluster 2 com } u_2, \Sigma_2 \in \Pi_2$

$\underbrace{\text{Posterior}(K|u)}_{\text{Posterior}} = \underbrace{P(u|K) P(K)}_{\sim N(u, \Sigma)} \rightarrow \text{fórmula usada nos próximos passos}$

$P(u|K) = \frac{1}{(2\pi)^{|\Sigma|/2}} e^{-\frac{1}{2}(u-u)^T \Sigma^{-1}(u-u)}$

$u_n$ :  $T = 7$  etapas

$\text{Posterior}(K_1|u_n) = P(u_n|K_1) P(K_1) = 0,023 \times 0,5 = 0,0115$   
 $\text{Posterior}(K_2|u_n) = P(u_n|K_2) P(K_2) = 0,062 \times 0,5 = 0,031$

$P(K_1|u_n) = \frac{0,015}{0,015+0,031} = 0,322$   
 $P(K_2|u_n) = \frac{0,031}{0,015+0,031} = 0,678$

$h_0$ :

$\text{Posterior}(K_1|u_2) = P(u_2|K_1) P(K_1) = 0,005 \times 0,5 = 0,002$   
 $\text{Posterior}(K_2|u_2) = P(u_2|K_2) P(K_2) = 0,048 \times 0,5 = 0,024$

$P(K_1|u_2) = \frac{0,002}{0,002+0,048} = 0,042$   
 $P(K_2|u_2) = \frac{0,048}{0,002+0,048} = 0,958$

$u_3$ :

$$\text{Posterior}(K_1|u_3) = P(u_3|K_1)P(K_1) = 0,036 \times 0,5 = 0,018$$

$$\text{Posterior}(K_2|u_3) = P(u_3|K_2)P(K_2) = 0,011 \times 0,5 = 0,005$$

$$P(K_1|u_3) = \frac{0,018}{0,018 + 0,005} = 0,770$$

$$P(K_2|u_3) = \frac{0,005}{0,005 + 0,018} = 0,230$$

M-Steponde  $K = c \rightarrow \text{cluster } c$ 

$$u_c = \sum_{i=1}^N P(K=c|u_i) u_i$$

$$\sum_{k=1}^N P(K=c|u_i)$$

valor da feature  
número da obs. num  $i$  em  $u_c$  atualizada

$$\sum_c^{(n_j)} \sum_{m=1}^N P(K=c|u_m) \times (u_{mj} - u_{ci}) \times (u_{mj} - u_{cj})$$

$$\sum_{m=1}^N P(K=c|u_m)$$

$$P(K=c) = \sum_{i=1}^N P(K=c|u_i)$$

$$u_1 = \frac{\sum_{i=1}^3 P(K_1|u_i) u_i}{\sum_{i=1}^3 P(K_1|u_i)} = \frac{0,332 \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} + 0,092 \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} + 0,770 \begin{bmatrix} 3 \\ -1 \end{bmatrix}}{0,332 + 0,092 + 0,770}$$

$$= \begin{bmatrix} 2,223 \\ -0,496 \end{bmatrix}$$

$$u_2 = \frac{\sum_{i=1}^3 P(K_2|u_i) u_i}{\sum_{i=1}^3 P(K_2|u_i)} = \frac{0,678 \times \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0,908 \begin{bmatrix} 0 \\ 2 \end{bmatrix} + 0,230 \begin{bmatrix} 3 \\ -1 \end{bmatrix}}{0,678 + 0,908 + 0,230}$$

$$= \begin{bmatrix} 0,754 \\ 0,873 \end{bmatrix}$$

$$\begin{aligned}
 \sum_1^{(1,1)} &= \frac{0,332 \times (1-0,223)^2 + 0,092 \times (0-0,223)^2 + 0,770 \times (3-0,223)^2}{0,332 + 0,092 + 0,770} \\
 &= 1,1182 \\
 \sum_1^{(1,2)} &= \sum_1^{(2,1)} = \\
 &= \frac{0,332 \times (1-0,223) \times (0+0,496) + 0,092 \times (0-0,223) \times (2+0,496) + 0,770 \times (3-0,223) \times (-1+0,496)}{0,332 + 0,092 + 0,770} \\
 &= -0,1849 \\
 \sum_1^{(2,2)} &= \frac{0,332 \times (0+0,496)^2 + 0,092 \times (2+0,496)^2 + 0,770 \times (-1+0,496)^2}{0,332 + 0,092 + 0,770} \\
 &= 0,714 \\
 \sum_1 &= \begin{bmatrix} 1,1182 & -0,1849 \\ -0,1849 & 0,714 \end{bmatrix} \\
 \sum_2^{(1,1)} &= \frac{0,678 \times (1-0,754)^2 + 0,908 \times (0-0,754)^2 + 0,230 \times (3-0,754)^2}{0,678 + 0,908 + 0,230} \\
 &= 0,1947 \\
 \sum_2^{(1,2)} &= \sum_2^{(2,1)} = \\
 &= \frac{0,678 \times (1-0,754) \times (0-0,754) + 0,908 \times (0-0,754) \times (2-0,754) + 0,230 \times (3-0,754) \times (-1-0,754)}{0,678 + 0,908 + 0,230} \\
 &= -1,039 \\
 \sum_2^{(2,2)} &= \frac{0,678 \times (0-0,754)^2 + 0,908 \times (2-0,754)^2 + 0,230 \times (-1-0,754)^2}{0,678 + 0,908 + 0,230} \\
 &= 1,364 \\
 \sum_2 &= \begin{bmatrix} 0,1947 & -1,039 \\ -1,039 & 1,364 \end{bmatrix} \\
 P(K_1) &= \frac{0,332 + 0,092 + 0,770}{3} \\
 &= 0,394 \\
 P(K_2) &= \frac{0,678 + 0,908 + 0,230}{3} \\
 &= 0,606
 \end{aligned}$$

2ª iteração

$n_1$ :

$$\text{Posterior}(K_1 | n_1) = P(n_1 | K_1) P(K_1) = 0,110 \times 0,394 = 0,047$$

$$\text{Posterior}(K_2 | n_1) = P(n_1 | K_2) P(K_2) = 0,140 \times 0,606 = 0,080$$

$$P(K_1 | n_1) = \frac{0,047}{0,047 + 0,080} = 0,342$$

$$P(K_2 | n_1) = \frac{0,080}{0,047 + 0,080} = 0,658$$

$n_2$ :

$$\text{Posterior}(K_1 | n_2) = P(n_2 | K_1) P(K_1) = 0,001 \times 0,394 = 0,0005$$

$$\text{Posterior}(K_2 | n_2) = P(n_2 | K_2) P(K_2) = 0,009 \times 0,606 = 0,0057$$

$$P(K_1 | n_2) = \frac{0,0005}{0,0005 + 0,0057} = 0,004$$

$$P(K_2 | n_2) = \frac{0,0057}{0,0005 + 0,0057} = 0,996$$

$n_3$ :

$$\text{Posterior}(K_1 | n_3) = P(n_3 | K_1) P(K_1) = 0,346 \times 0,394 = 0,137$$

$$\text{Posterior}(K_2 | n_3) = P(n_3 | K_2) P(K_2) = 0,011 \times 0,606 = 0,0067$$

$$P(K_1 | n_3) = \frac{0,137}{0,137 + 0,0067} = 0,953$$

$$P(K_2 | n_3) = \frac{0,0067}{0,137 + 0,0067} = 0,047$$

$$\underline{U_1} = \frac{\sum_{i=1}^3 P(K_1 | n_i) n_i}{\sum_{i=1}^3 P(K_1 | n_i)} = \frac{0,342 [1] + 0,004 [2] + 0,953 [-1]}{0,342 + 0,004 + 0,953} = \begin{bmatrix} 0,465 \\ -0,728 \end{bmatrix}$$

$$\underline{U_2} = \frac{\sum_{i=1}^3 P(K_2 | n_i) n_i}{\sum_{i=1}^3 P(K_2 | n_i)} = \frac{0,658 [1] + 0,996 [2] + 0,047 [-1]}{0,658 + 0,996 + 0,047} = \begin{bmatrix} 0,469 \\ 1,144 \end{bmatrix}$$

$$\sum_1^{(1,2)} = \frac{0,342 \times (1 - 0,465)^2 + 0,004 \times (0 - 0,465)^2 + 0,953 \times (3 - 0,465)^2}{0,342 + 0,004 + 0,953} = \underline{0,793}$$

$$\sum_1^{(1,2)} = \sum_1^{(2,1)} = \frac{0,342 \times (1 - 0,465)(0 + 0,728) + 0,004 \times (0 - 0,465)(2 + 0,728) + 0,953 \times (3 - 0,465)(-1 + 0,728)}{0,342 + 0,004 + 0,953} \\ = \underline{-0,1407}$$

$$\sum_1^{(2,2)} = \frac{0,342 \times (0 + 0,728)^2 + 0,004 \times (2 + 0,728)^2 + 0,953 \times (-1 + 0,728)^2}{0,342 + 0,004 + 0,953} \\ = \underline{0,1215}$$

$$\underline{\sum_1} = \begin{bmatrix} 0,793 & -0,1407 \\ -0,1407 & 0,1215 \end{bmatrix}$$

$$\sum_2^{(1,1)} = \frac{0,658 \times (1 - 0,469)^2 + 0,996 \times (0 - 0,469)^2 + 0,047 \times (3 - 0,469)^2}{0,658 + 0,996 + 0,047} \\ = \underline{0,474}$$

$$\sum_2^{(1,2)} = \sum_2^{(2,1)} = \frac{0,658 \times (1 - 0,469)(0 - 1,144) + 0,996 \times (0 - 0,469)(2 - 1,144) + 0,047 \times (3 - 0,469)(-1 - 1,144)}{0,658 + 0,996 + 0,047} \\ = \underline{-0,1619}$$

$$\sum_2^{(2,2)} = \frac{0,658 (0 - 1,144)^2 + 0,996 (2 - 1,144)^2 + 0,047 \times (-1 - 1,144)^2}{0,658 + 0,996 + 0,047}$$

$$= 1,061$$

$$\Sigma_2 = \begin{bmatrix} 0,658 & -0,047 \\ -0,047 & 1,061 \end{bmatrix}$$

$$P(K_1) = \frac{0,342 + 0,004 + 0,953}{3} \quad P(K_2) = \frac{0,658 + 0,996 + 0,047}{3}$$

$$= 0,433 \quad = 0,567$$

2. a)

[2]

a)

$$u_1 : \text{Posterior}(k_1 | u_1) = P(u_1 | k_1) P(k_1) = 0,564 \times 0,433 = 0,244$$

$$\text{Posterior}(k_2 | u_1) = P(u_1 | k_2) P(k_2) = 0,304 \times 0,567 = 0,173$$

$\text{Posterior}(k_1 | u_1) > \text{Posterior}(k_2 | u_1)$ , logo  $u_1 \in \text{Cluster 1}$

$$u_2 : \text{Posterior}(k_1 | u_2) = P(u_2 | k_1) P(k_1) = 7,941 \times 10^{-78} \times 0,433 = 3,43 \times 10^{-78}$$

$$\text{Posterior}(k_2 | u_2) = P(u_2 | k_2) P(k_2) = 0,473 \times 0,567 = 0,268$$

$\text{Posterior}(k_1 | u_2) < \text{Posterior}(k_2 | u_2)$ , logo  $u_2 \in \text{Cluster 2}$

$$u_3 : \text{Posterior}(k_1 | u_3) = P(u_3 | k_1) P(k_1) = 1,903 \times 0,433 = 0,824$$

$$\text{Posterior}(k_2 | u_3) = P(u_3 | k_2) P(k_2) = 1,1345 \times 10^{-8} \times 0,567 = 7,628 \times 10^{-9}$$

$\text{Posterior}(k_1 | u_3) > \text{Posterior}(k_2 | u_3)$ , logo  $u_3 \in \text{Cluster 3}$

Cluster 1 - }  $u_1, u_3$

Cluster 2 - }  $u_2$

2. b)

b) Silhueta

$$S(u_i) = \begin{cases} 1 - \frac{a}{b}, & \text{se } a < b \\ \frac{b}{a} - 1, & \text{se } a > b \end{cases}$$

Onde:

$a \rightarrow$  média das distâncias  
de  $u_i$  aos pontos do seu cluster  
 $b \rightarrow \min$  (média da d. de  
 $u_i$  aos pontos de  
outros clusters)

 $\underline{u_1}: u_1 \in \text{cluster 1}$ 

$$a \rightarrow d(u_1, u_2) = \sqrt{(1-3)^2 + (0-(-1))^2} = \sqrt{5}$$

$$b \rightarrow d(u_1, u_3) = \sqrt{(1-0)^2 + (0-2)^2} = \sqrt{5}$$

$$S(u_1) = 0$$

 $\underline{u_3}: u_3 \in \text{cluster 1}$ 

$$a \rightarrow d(u_3, u_2) = \sqrt{5}$$

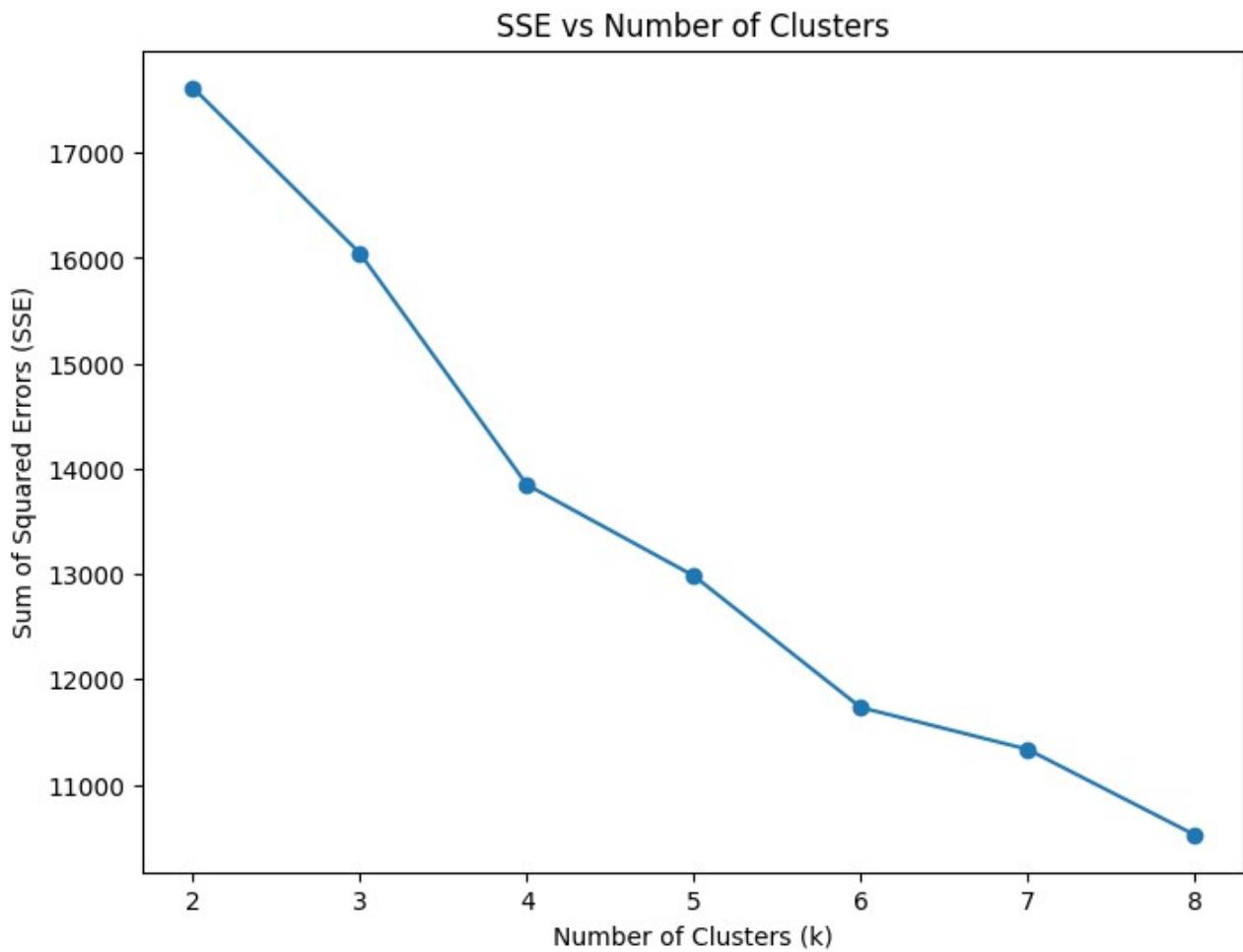
$$b \rightarrow d(u_3, u_1) = \sqrt{(3-0)^2 + (-1-2)^2} = \sqrt{18}$$

$$S(u_3) = 1 - \frac{\sqrt{5}}{\sqrt{18}}$$

$$S(\text{cluster 1}) = \frac{S(u_1) + S(u_3)}{2} = \frac{1 - \frac{\sqrt{5}}{\sqrt{18}}}{2} \approx 0,136$$

## II. Programming and critical analysis

1.a)



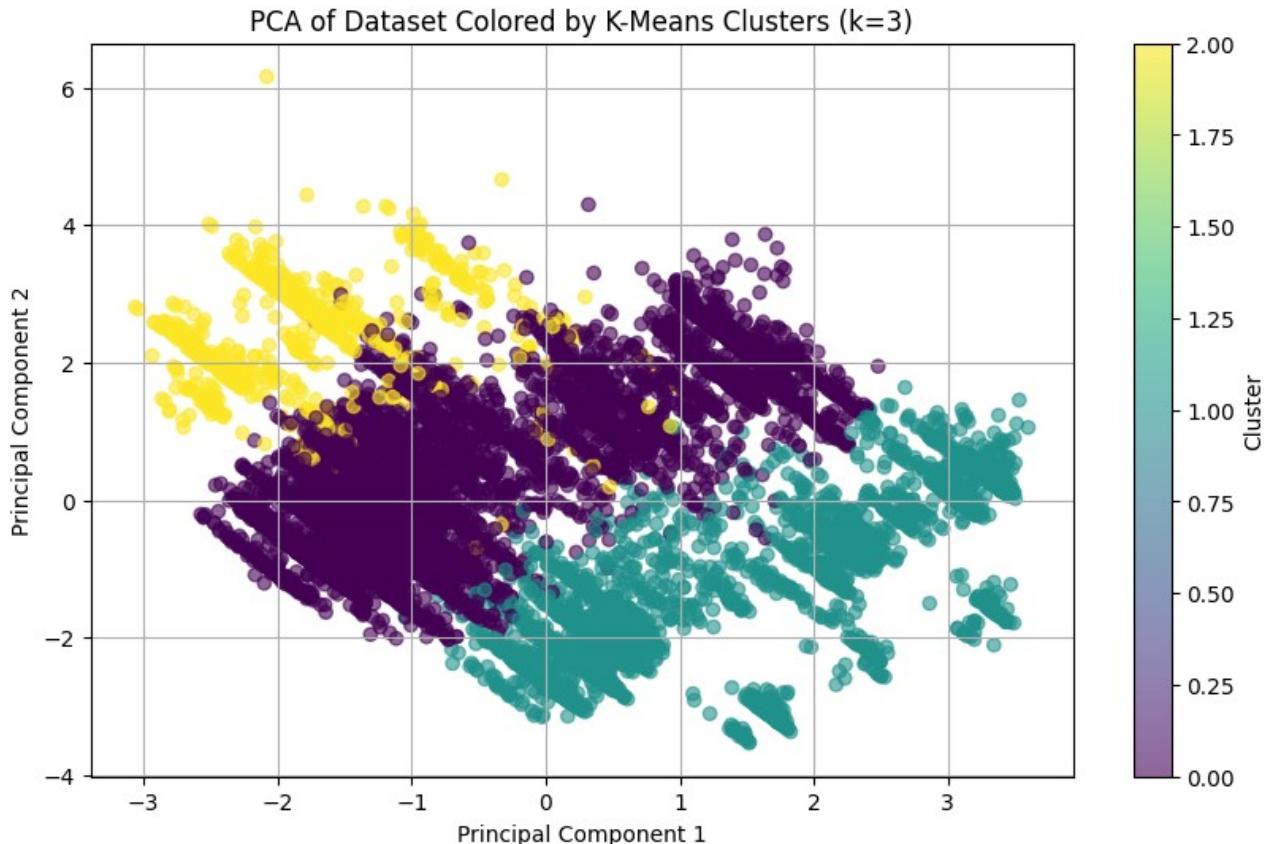
1.b)

Com base no gráfico, o número ideal de clusters parece ser 4, pelo facto de o declive da curva começar a diminuir (em valor absoluto). Isto indica que a adição de mais clusters não baixa suficientemente o valor do SSE para se justificar aumentar a complexidade do modelo. Assim, com 4 clusters atinge-se um equilíbrio entre divisão dos dados e simplicidade do modelo.

1.c)

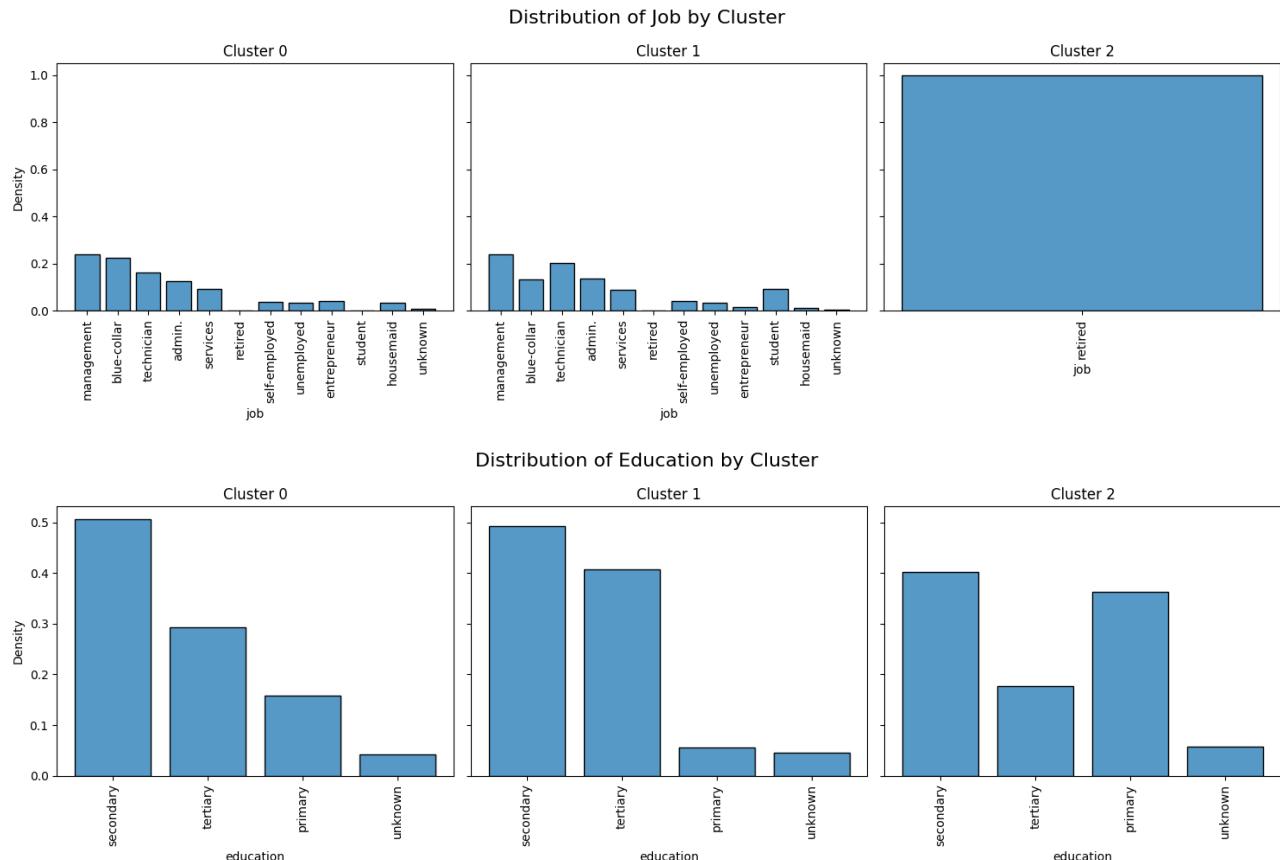
Sim, o k-modes é a abordagem mais adequada para conjuntos de dados predominantemente categóricos. Isso deve-se ao fato de que, neste caso, a utilização da moda para cálculos do algoritmo é mais apropriada do que a média, que é utilizada no k-means. A moda permite uma representação mais fiel das características dos dados categóricos, resultando num clustering mais eficaz.

2. b)



Não conseguimos separar os clusters de forma clara. A mistura de pontos entre os diferentes clusters obtidos pelo k-means (com  $k=3$ ) pode ser explicada pelo facto de apenas 23% da variabilidade ter sido capturada pelas duas principais componentes principais no PCA. Este valor baixo indica que as componentes não refletem adequadamente a informação total do conjunto de dados, sugerindo uma estrutura complexa que não é bem representada na projeção bidimensional. Como resultado, muitos padrões e interações relevantes entre as variáveis permanecem não capturados, comprometendo a separação entre os clusters. Assim, esta mistura deve-se à falta de distinção clara nas características que realmente importam.

2.c)



Os gráficos apresentam a distribuição das categorias de emprego e educação em quatro clusters (0, 1 e 2).

Relativamente à distribuição de empregos, verificamos que os clusters 0 e 1 são amplamente dominado por gestores, técnicos, operários e administradores, sendo que o resto das atividades aparece em menor quantidade e pode ser desprezado. O cluster 2 só contém reformados.

Relativamente à distribuição de educação, verificamos que o cluster 0 possui cerca de metade das pessoas com o ensino secundário, um terço com o ensino superior e uma percentagem baixa de pessoas com o ensino primário. O cluster 1 possui quase só pessoas com o ensino superior e secundário. No cluster 2, predominam pessoas com educação secundária e primária, havendo também algumas (em menor quantidade) com o ensino superior.

Podemos, portanto, concluir que os três clusters representam grupos de indivíduos com características distintas. Os Clusters 0 e 1 são compostos principalmente por pessoas em cargos de gestão, técnicos, operários e administração, com um nível de educação relativamente alto (ensino superior e secundário). O Cluster 2 é composto apenas por reformados, no qual se verifica um grau de educação mais baixo (predominantemente primário e secundário).

**END**