Aprendizagem 2024
Homework IV – Group 44
(ist1106827, ist1107245)

Part I: Pen and paper

Consider the bivariate observations:

$$\mathbf{x}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix}0\\2\end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix}3\\-1\end{bmatrix}$$

and the multivariate Gaussian mixture given by:

$$\mathbf{u}_1 = \begin{bmatrix}2\\-1\end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix}1\\1\end{bmatrix}, \quad \mathbf{\Sigma}_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}, \quad \mathbf{\Sigma}_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}, \quad \pi_1 = 0.5, \quad \pi_2 = 0.5.$$

Answer the following questions by presenting all intermediary steps, and use 3 decimal places in each.

1. Perform two epochs of the EM clustering algorithm and determine the new parameters.

médias: $\mu_1 = \begin{bmatrix}2\\-1\end{bmatrix}$, $\mu_2 = \begin{bmatrix}1\\1\end{bmatrix}$

covariâncias: $\Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}$, $\Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$

pesos: $p(K=1) = \pi_1 = 0.5$, $p(K=2) = \pi_2 = 0.5$

observações: $x_1 = \begin{bmatrix}1\\0\end{bmatrix}$, $x_2 = \begin{bmatrix}0\\2\end{bmatrix}$, $x_3 = \begin{bmatrix}3\\-1\end{bmatrix}$

1ª época

E-step:

$\text{Posterior}(c_k \mid x_i) = P(x_i \mid c_k) \cdot P(c_k) = \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \cdot \pi_k$

$P(c_k \mid x_i) = \frac{\text{Posterior}(c_k, x_i)}{\sum_{j=1}^{2} \text{Posterior}(c_j, x_i)}$ ← normalizar

x_1: $\text{Posterior}(K=1 \mid x_1) = P(x_1 \mid K=1) \times P(K=1) = 0,029 \times 0,5 = 0,015$
$\text{Posterior}(K=2 \mid x_1) = P(x_1 \mid K=2) \times P(K=2) = 0,062 \times 0,5 = 0,031$

$P(K=1 \mid x_1) = \frac{0,015}{0,015+0,031} = 0,326$ $\quad P(K=2 \mid x_1) = \frac{0,031}{0,015+0,031} = 0,674$

x_2: $\text{Posterior}(K=1 \mid x_2) = P(x_2 \mid K=1) \times P(K=1) = 0,005 \times 0,5 = 0,003$
$\text{Posterior}(K=2 \mid x_2) = P(x_2 \mid K=2) \times P(K=2) = 0,048 \times 0,5 = 0,024$

$P(K=1 \mid x_2) = \frac{0,003}{0,003+0,024} = 0,111$ $\quad P(K=2 \mid x_2) = \frac{0,024}{0,003+0,024} = 0,889$

x_3: $\text{Posterior}(K=1 \mid x_3) = P(x_3 \mid K=1) \times P(K=1) = 0,036 \times 0,5 = 0,018$
$\text{Posterior}(K=2 \mid x_3) = P(x_3 \mid K=2) \times P(K=2) = 0,011 \times 0,5 = 0,006$

$P(K=1 \mid x_3) = \frac{0,018}{0,018+0,006} = 0,750$ $\quad P(K=2 \mid x_3) = \frac{0,006}{0,018+0,006} = 0,250$

M-step:

updates

$$\mu_c = \frac{\sum_{i=1}^{n} P(K=c|x_i) \cdot x_i}{\sum_{i=1}^{n} P(K=c|x_i)}$$

↳ médias

$$\Sigma_c^{(i,j)} = \frac{\sum_{m=1}^{n} p(K=c|x_m) \times (x_{mi} - \mu_{ci}) \times (x_{mj} - \mu_{cj})}{\sum_{m=1}^{n} P(K=c|x_m)}$$

x_{mi}: valor da feature i na observação x_m; μ_{ci}: valor da feature i no μ_c atualizado

↳ covariâncias

$$p(K=c) = \frac{\sum_{i=1}^{n} P(K=c|x_i)}{n}$$

↳ pesos (π_c)

cluster 1:

$$\mu_1 = \frac{0,326 \times x_1 + 0,111 \times x_2 + 0,750 \times x_3}{0,326+0,111+0,750} = \begin{bmatrix} 2,170 \\ -0,445 \end{bmatrix}$$

$$\Sigma_1^{(1,1)} = \frac{0,326 \times (1-2,170)^2 + 0,111 \times (0-2,170)^2 + 0,750 \times (3-2,170)^2}{0,326+0,111+0,750} = 1,252$$

$$\Sigma_1^{(1,2)} = \Sigma_1^{(2,1)} = \frac{0,326 \times (1-2,170) \times (0+0,445) + 0,111 \times (0-2,170) \times (2+0,445) + 0,750 \times (3-2,170) \times (0-(-0,445))}{0,326+0,111+0,750} = -0,931$$

$$\Sigma_1^{(2,2)} = \frac{0,326 \times (0+0,445)^2 + 0,111 \times (2+0,445)^2 + 0,750 \times (-1+0,445)^2}{0,326+0,111+0,750} = 0,808$$

$$\Sigma_1 = \begin{bmatrix} 1,252 & -0,931 \\ -0,931 & 0,808 \end{bmatrix} \qquad P(K=1) = \frac{0,326+0,111+0,750}{3} = 0,396$$

cluster 2:

$$\mu_2 = \frac{0,674 \times x_1 + 0,889 \times x_2 + 0,250 \times x_3}{0,674+0,889+0,250} = \begin{bmatrix} 0,785 \\ 0,843 \end{bmatrix}$$

$$\Sigma_2^{(1,1)} = \frac{0,674 \times (1-0,785)^2 + 0,889 \times (0-0,785)^2 + 0,250 \times (3-0,785)^2}{0,674+0,889+0,250} = 0,996$$

$$\Sigma_2^{(1,2)} = \Sigma_2^{(2,1)} = \frac{0,674 \times (1-0,785) \times (0-0,843) + 0,889 \times (0-0,785) \times (2-0,843) + 0,250 \times (3-0,785) \times (0-0,843)}{0,674+0,889+0,250} = -1,076$$

$$\Sigma_2^{(2,2)} = \frac{0,674 \times (0-0,843)^2 + 0,889 \times (2-0,843)^2 + 0,250 \times (-1-0,843)^2}{0,674+0,889+0,250} = 1,389$$

$$\Sigma_2 = \begin{bmatrix} 0,996 & -1,076 \\ -1,076 & 1,389 \end{bmatrix} \qquad P(K=2) = \frac{0,674+0,889+0,250}{3} = 0,604$$

2ª época

E-step:

x_1: $\text{Posterior}(K=1|x_1) = P(x_1|K=1) \times P(K=1) = 0,111 \times 0,396 = 0,044$

$\text{Posterior}(K=2|x_1) = P(x_1|K=2) \times P(K=2) = 0,144 \times 0,604 = 0,087$

$P(K=1|x_1) = \frac{0,044}{0,044+0,087} = 0,336$ $\quad P(K=2|x_1) = \frac{0,087}{0,044+0,087} = 0,664$

x_2: $\text{Posterior}(K=1|x_2) = P(x_2|K=1) \times P(K=1) = 0,003 \times 0,396 = 0,001$

$\text{Posterior}(K=2|x_2) = P(x_2|K=2) \times P(K=2) = 0,199 \times 0,604 = 0,120$

$P(K=1|x_2) = \frac{0,001}{0,001+0,12} = 0,008$ $\quad P(K=2|x_2) = \frac{0,120}{0,001+0,120} = 0,992$

x_3: $\text{Posterior}(K=1|x_3) = P(x_3|K=1) \times P(K=1) = 0,312 \times 0,396 = 0,124$

$\text{Posterior}(K=2|x_3) = P(x_3|K=2) \times P(K=2) = 0,015 \times 0,604 = 0,009$

$P(K=1|x_3) = \frac{0,124}{0,124+0,009} = 0,932$ $\quad P(K=2|x_3) = \frac{0,009}{0,124+0,009} = 0,068$

M-step:

cluster 1:

$$\mu_1 = \frac{0,336 \times x_1 + 0,008 \times x_2 + 0,932 \times x_3}{0,336+0,008+0,932} = \begin{bmatrix} 2,455 \\ -0,718 \end{bmatrix}$$

$$\Sigma_1^{(1,1)} = \frac{0,336 \times (1-2,455)^2 + 0,008 \times (0-2,455)^2 + 0,932 \times (3-2,455)^2}{0,336+0,008+0,932} = 0,812$$

$$\Sigma_1^{(1,2)} = \Sigma_1^{(2,1)} = \frac{0,336 \times (1-2,455) \times (0+0,718) + 0,008 \times (0-2,455) \times (2+0,718) + 0,932 \times (3-2,455) \times (0+0,718)}{0,336+0,008+0,932} = -0,430$$

$$\Sigma_1^{(2,2)} = \frac{0,336 \times (0+0,718)^2 + 0,008 \times (2+0,718)^2 + 0,932 \times (-1+0,718)^2}{0,336+0,008+0,932} = 0,240$$

$$\Sigma_1 = \begin{bmatrix} 0,812 & -0,430 \\ -0,430 & 0,240 \end{bmatrix}$$

$$P(K=1) = \frac{0,336+0,008+0,932}{3} = 0,425$$

cluster 2:

$$\mu_2 = \frac{0,664 \times x_1 + 0,992 \times x_2 + 0,068 \times x_3}{0,664 + 0,992 + 0,068} = \begin{bmatrix} 0,503 \\ 1,111 \end{bmatrix}$$

$$\Sigma_2^{(1,1)} = \frac{0,664 \times (1-0,503)^2 + 0,992 \times (0-0,503)^2 + 0,068 \times (3-0,503)^2}{0,664 + 0,992 + 0,068} = 0,487$$

$$\Sigma_2^{(1,2)} = \Sigma_2^{(2,1)} = \frac{0,664 \times (1-0,503) \times (0-1,111) + 0,992 \times (0-0,503) \times (2-1,111) + 0,068 \times (3-0,503) \times (0-1,111)}{0,664 + 0,992 + 0,068} = -0,678$$

$$\Sigma_2^{(2,2)} = \frac{0,664 \times (0-1,111)^2 + 0,992 \times (2-1,111)^2 + 0,068 \times (-1-1,111)^2}{0,664 + 0,992 + 0,068} = 1,106$$

$$\Sigma_2 = \begin{bmatrix} 0,487 & -0,678 \\ -0,678 & 1,106 \end{bmatrix} \qquad P(K=2) = \frac{0,664 + 0,992 + 0,068}{3} = 0,575$$

2. Using the final parameters computed in previous question:
   a) perform a hard assignment of observations to clusters under a MAP assumption.

$$\mu_1 = \begin{bmatrix} 2,455 \\ -0,718 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 0,503 \\ 1,111 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 0,812 & -0,430 \\ -0,430 & 0,240 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 0,487 & -0,678 \\ -0,678 & 1,106 \end{bmatrix}$$

$P(K=1) = 0,425 \qquad P(K=2) = 0,575$

novos parâmetros

hard assignment under a MAP assuption

$$h_{MAP} = \arg\max_k \{P(K|x_i)\} = \arg\max_k \left\{\frac{P(K) \times P(x_i|K)}{P(K)}\right\} = \arg\max_k \{P(K) \times P(x_i|K)\}$$

$$P(x_i|K) = N(x_i | \mu = \mu_i, \Sigma = \Sigma_i)$$

$$P(K|x_i) = P(x_i|K) \times P(K) = N(x_i, \mu = \mu_i, \Sigma = \Sigma_i) \times P(K)$$

x_1: $P(K=1|x_1) = 0,387 \times 0,425 = 0,164$
$P(K=2|x_1) = 0,256 \times 0,575 = 0,147$

$P(K=1|x_1) > P(K=2|x_1)$, logo x_1 pertence ao Cluster 1

x_2: $P(K=1|x_2) = 0 \times 0,425 = 0$
$P(K=2|x_2) = 0,391 \times 0,575 = 0,225$

$P(K=1|x_2) < P(K=2|x_2)$, logo x_2 pertence ao Cluster 2

x_3: $P(K=1|x_3) = 1,325 \times 0,425 = 0,563$
$P(K=2|x_3) = 0 \times 0,575 = 0$

$P(K=1|x_3) > P(K=2|x_3)$, logo x_3 pertence ao Cluster 1

Cluster 1 tem as observações x_1 e x_3
Cluster 2 tem as observações x_2

b) compute the silhouette of the larger cluster (the one that has more observations assigned to it) using the Euclidean distance.

b)

Silhueta: Avaliar a qualidade dos clusters

queremos otimizar

- coesão: minimizar distâncias intra-cluster
- separação: maximizar distâncias inter-cluster

$$s(x_i) = \begin{cases} 1 - \frac{a}{b}, & \text{se } a < b \\ \frac{b}{a} - 1, & \text{se } a \geq b \end{cases}$$

$a \rightarrow$ media das distâncias de x_i aos pontos do seu cluster

$b \rightarrow \min_K$ (media das distâncias x_i aos pontos do cluster K)

$$S(\text{cluster}) = \frac{\sum_{i=1}^{N} s(x_i)}{N}$$

$$d(a,b) = \sum_{i=1}^{p} \sqrt{(a_i - b_i)^2}$$ distância Euclediana

O maior cluster é o Cluster 1 (x_1, x_3)

x_1:

$a = d(x_1, x_3) = \sqrt{(1-3)^2 + (0-(-1))^2} = \sqrt{5}$

$b = d(x_1, x_2) = \sqrt{(1-0)^2 + (0-2)^2} = \sqrt{5}$

] $a = b$

$$s(x_1) = \frac{\sqrt{5}}{\sqrt{5}} - 1 = 1 - 1 = 0$$

x_3:

$a = d(x_3, x_1) = \sqrt{(3-1)^2 + (-1-0)^2} = \sqrt{5}$

$b = d(x_3, x_2) = \sqrt{(3-0)^2 + (-1-2)^2} = \sqrt{18}$

] $a < b$

$$s(x_3) = 1 - \frac{\sqrt{5}}{\sqrt{18}} = 0,473$$

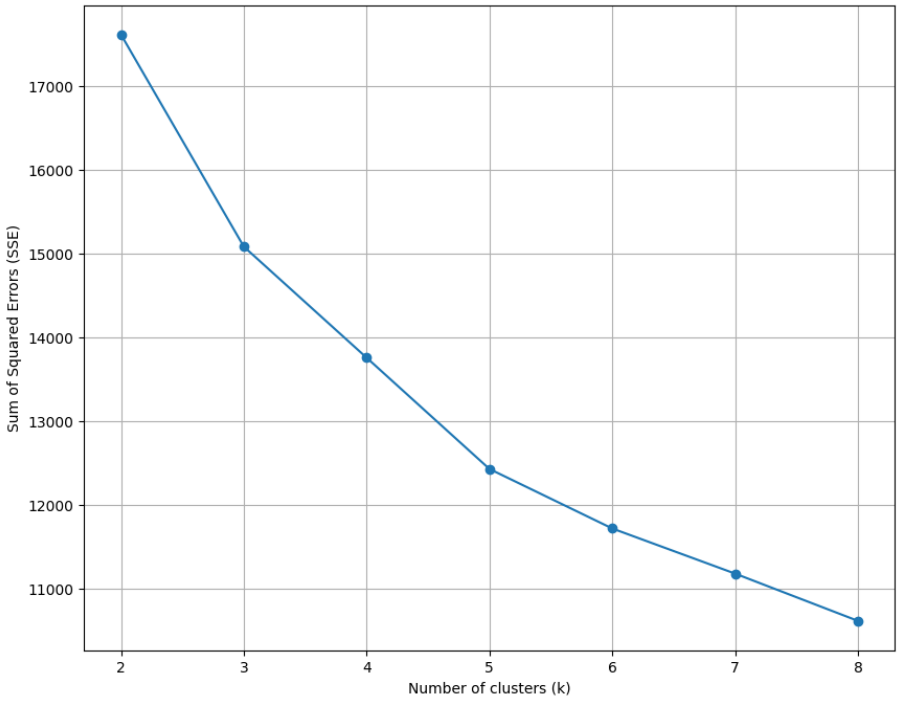
$$S(\text{cluster } 1) = \frac{0 + 0,473}{2} = \boxed{0,237}$$

Part II: Programming

In the next exercise you will use the accounts.csv dataset. This dataset contains account details of bank clients, and the target variable y is binary ('has the client subscribed a term deposit?').

Hint: You can use get_dummies() to change the feature type from categorical to numerical (e.g. pd.get_dummies(data, drop_first=True)).

1. Normalize the data using MinMaxScaler:
   a) Using sklearn, apply k-means clustering (without targets) on the normalized data with k = {2,3,4,5,6,7,8}, max iter = 500 and random state=42. Plot the different sum of squared errors (SSE) using the _inertia attribute of k-means according to the number of clusters.



   b) According to the previous plot, how many underlying customer segments(clusters) should there be? Explain based on the trade-off between the clusters and inertia.

   To determine the optimal number of clusters, we typically use the "elbow method," which involves plotting the Sum of Squared Errors (SSE) - also known as inertia - across different values of k. Inertia represents the sum of squared distances between data points and their assigned cluster center. The optimal cluster count is suggested by an "elbow" in the plot, marking the point where additional clusters yield diminishing reductions in SSE.

   In this plot, there is not a very defined "elbow," but we can argue that k = 3 is the point that most resembles one. This suggests that 3 clusters provide a suitable balance between minimizing inertia and avoiding excessive complexity. Additionally, the SSE drop between k = 2 and k = 3 is larger than the subsequent drop from k = 3 to k = 4, reinforcing k = 3 as a natural stopping point. A minor shift in inertia reduction is also observed around k = 5, but it is less pronounced, further supporting 3 clusters as the most reasonable choice based on the trade-off between interpretability and SSE reduction.

c) Would k-modes be a better clustering approach? Explain why based on the dataset features.

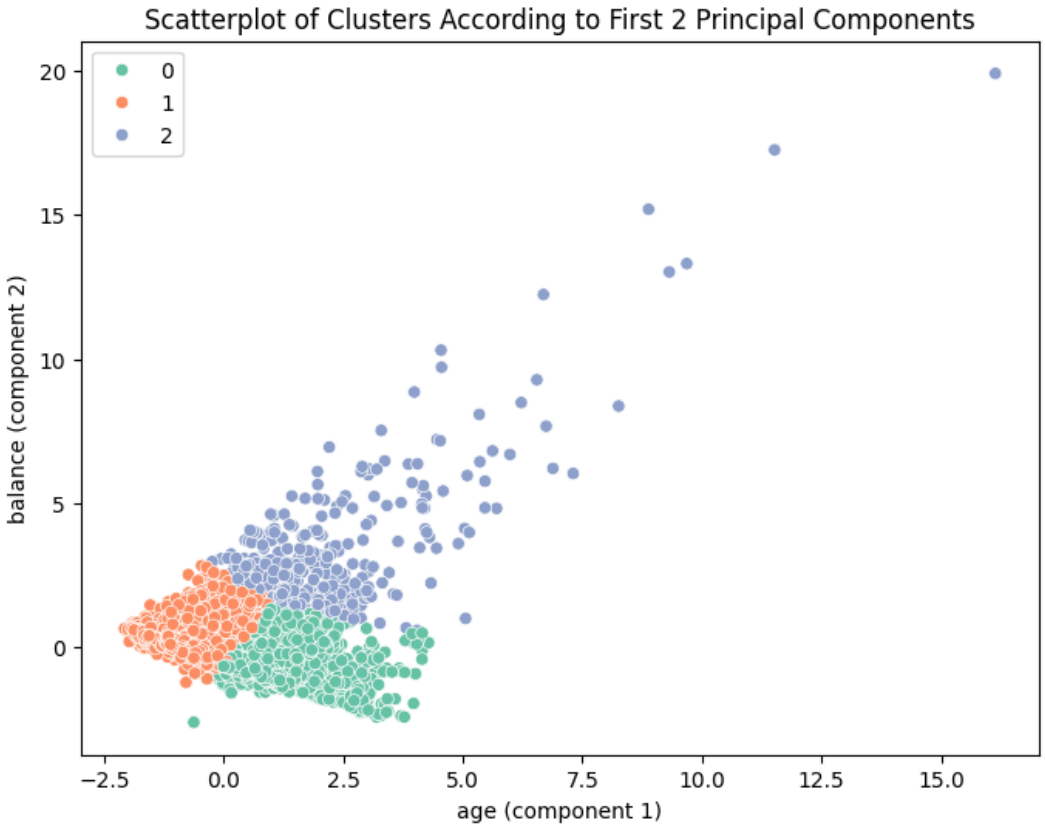
K-modes not be a better clustering approach for this dataset because it is designed specifically for clustering categorical data, forming clusters based on the mode (most frequent category) rather than the mean. Since this dataset includes both numerical and categorical features, k-modes would not be suitable.

A more appropriate choice would be k-prototypes, which can directly handle mixed data, or alternatively, k-means with one-hot encoding (dummification) of categorical variables. These approaches better accommodate the dataset's numerical and categorical information, making them more effective options.

2. Normalize the data using StandardScaler:
a) Apply PCA to the data. How much variability is explained by the top 2 components?

The top two components explain 52.14% of the data variability, with the first component accounting for 29.14% and the second for 23%.

b) Apply k-means clustering with k=3 and random_state = 42 (all other arguments as default) and use the original 8 features. Next, provide a scatterplot according to the first 2 principal components. Can we clearly separate the clusters? Justify.

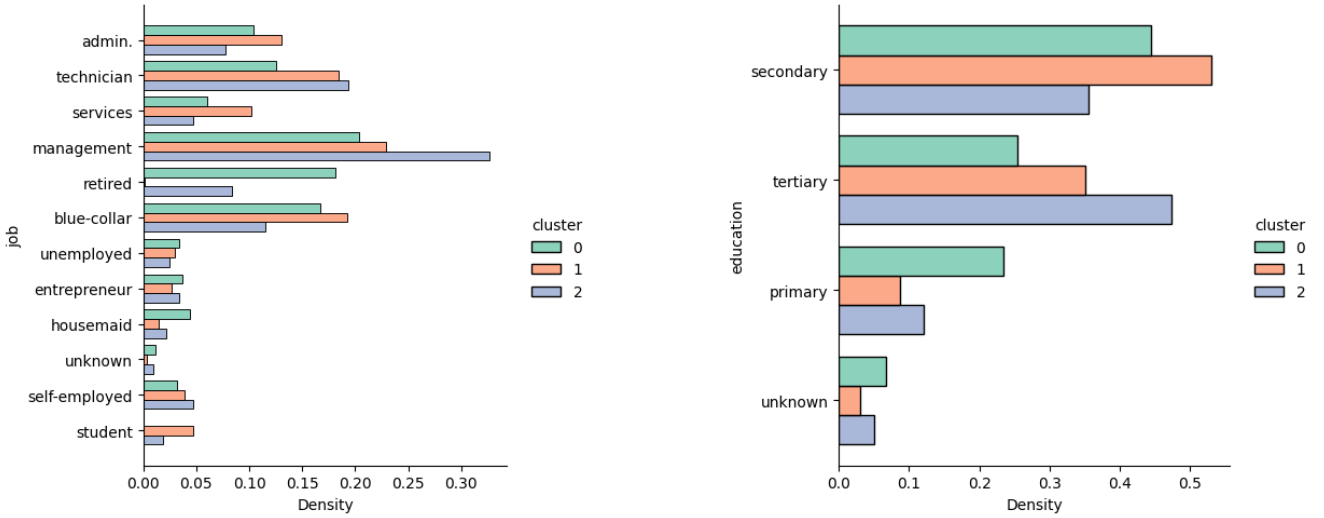


The scatterplot of the clusters projected onto the first two principal components indicates that the clusters can be moderately separated, but the boundaries are not entirely distinct, with some overlap present.

Cluster 1 (in orange) shows strong cohesion, with points tightly clustered together, suggesting low internal variability. Cluster 0 (in green) is also relatively compact, though it appears slightly more dispersed than Cluster 1, suggesting a moderate level of cohesion.Cluster 2 (in blue), however, is more spread out, indicating a weaker cohesion and higher internal variability. This spread might reflect aspects of the original data that are not fully captured by the first two principal components.

Despite the proximity of the clusters, minimal overlap suggests they are reasonably well-separated in this two-dimensional projection. Overall, while the clusters are not perfectly distinct, they display enough separation to indicate that each group possesses unique characteristics, even if Cluster 2 shows greater variability.

c) Plot the cluster conditional features of the frequencies of “job” and ”education” according to the clusters obtained in the previous question (2b.). Use sns.displot (see Data Exploration notebook), with multiple= ”dodge”, stat=’density’, shrink=0.8 and common_norm=False. Describe the main differences between the clusters in no more than half a page.



In terms of job attributes, Cluster 0 shows the highest density in categories such as retired, housemaid, entrepreneur, and unemployed. The significant presence of retired and unemployed individuals indicates that this cluster comprises many people who are not actively participating in the workforce. Additionally, it has a notable representation of individuals in admin and blue-collar jobs. Cluster 1 displays a higher density of individuals in blue-collar, admin, and services roles, suggesting a trend toward manual labor and service-oriented employment. This cluster has the highest density of students. Cluster 2 is primarily populated by individuals in management and technician roles, indicating a concentration in professional and skilled technical occupations. Interestingly, this cluster has the lowest density of unemployed individuals, implying a more stable employment environment.

Regarding education, Cluster 0 is characterized by the highest density of individuals with primary education and the lowest density of those with tertiary education, reflecting generally lower educational attainment. Cluster 1 shows a higher density of individuals with secondary education, along with some representation of those holding tertiary education. Cluster 2 stands out with the highest density of individuals possessing tertiary education, indicating a group with significantly higher educational qualifications.

Overall, Cluster 2 represents individuals with higher education levels and stable, high-status jobs, suggesting a more affluent socioeconomic profile. Cluster 1 features individuals with moderate educational attainment and a prevalence of blue-collar or service roles, indicating a mixed socioeconomic standing. In contrast, Cluster 0 includes a diverse group, primarily consisting of individuals with lower education levels and a considerable proportion of retired or unemployed individuals, reflecting lower overall workforce participation.

Appendix

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

df = pd.read_csv("./data/accounts.csv")

X = df.iloc[:, :8]

X.dropna(inplace = True)
X.drop_duplicates(inplace = True)
X = pd.get_dummies(X, drop_first = True)

numeric_features = ["age", "balance"]

X_normalized = X.copy()
X_normalized[numeric_features] = MinMaxScaler().fit_transform(X[numeric_features])

k_values = [2, 3, 4, 5, 6, 7, 8]
sse = []

for k in k_values:
    kmeans = KMeans(n_clusters = k, init = "random", random_state = 42, max_iter = 500)
    kmeans_model = kmeans.fit(X_normalized)
    sse.append(kmeans_model.inertia_)

plt.figure(figsize = (10, 8))
plt.plot(k_values, sse, marker = 'o')
plt.xlabel("Number of clusters (k)")
plt.ylabel("Sum of Squared Errors (SSE)")
plt.grid(True)

plt.show()
```

Listing 1: Line Graph of Sum of Squared Errors (SSE) for K-Means Clustering with Varying Cluster Counts

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

numeric_features = ["age", "balance"]

X_normalized = X.copy()
```

```
X_normalized[numeric_features] = StandardScaler().fit_transform((X[numeric_features
    ]))

pca = PCA(n_components = 2)
pca.fit(X_normalized)
explained_variance = pca.explained_variance_ratio_

print(explained_variance)
```

Listing 2: Explained Variance of Top 2 Principal Components Using PCA

```
import seaborn as sns

kmeans = KMeans(n_clusters = 3, random_state = 42)
clusters = kmeans.fit_predict(X_normalized)

pca = PCA(n_components = 2)
pca_components = pca.fit_transform(X_normalized)

selected_features = list(X_normalized.var().sort_values(ascending = False).head(2).
    index)


plt.figure(figsize = (8, 6))
sns.scatterplot(x = pca_components[:, 0], y = pca_components[:, 1], hue = X["
    cluster"], palette = "Set2")
plt.xlabel(selected_features[0] + " (component 1)")
plt.ylabel(selected_features[1] + " (component 2)")
plt.title("Scatterplot of Clusters According to First 2 Principal Components")

plt.show()
```

Listing 3: K-Means Clustering with k=3 and Scatterplot of First Two Principal Components

```
df = pd.read_csv("./data/accounts.csv")

X = df.iloc[:, :8]

X.dropna(inplace = True)
X.drop_duplicates(inplace = True)

X["cluster"] = clusters

sns.displot(
    data = X,
    y = "job",
    hue = "cluster",
    multiple = "dodge",
    stat = "density",
    shrink = 0.8,
    common_norm = False,
    palette = "Set2"
)

sns.displot(
    data = X,
    y = "education",
```

```
    hue = "cluster",
    multiple="dodge",
    stat = "density",
    shrink = 0.8,
    common_norm = False,
    palette = "Set2"
)

plt.show()
```

Listing 4: Cluster Conditional Features of Job and Education Frequencies