Análisis Multivariado

Producto académico 03

Kevin Heberth Haquehua Apaza

30 de julio del 2025

Table of Contents

# Ejercicios AFE, Cluster y análisis de correspondencia múltiple

## CASO 1: Postulantes (6 puntos)

Se tienen datos de las notas de alumnos postulantes a un colegio de alto rendimiento, se desea agrupar las notas de los cursos y ver que grupos podrían haber de cursos.

Las notas de los siguientes cursos son Razonamiento verbal, Razonamiento matemático, Matemáticas, Psicología y filosofía, Física, Lógica, Biología, Historia y Química.

Archivo a utilizar **postulantes.sav**

library(foreign)  
postulantes <- read.spss(here("9 Analisis Multivariado/Trabajo 3/Postulantes.sav"),  
 to.data.frame = TRUE,  
 use.value.labels = TRUE)

1. Realizar análisis factorial exploratorio
2. Decida cuantos factores retener explique el por qué.
3. Decida el método de rotación y explique el por qué.
4. Explicar los resultados y de sus conclusiones del ejercicio.

### Solución

1. **Realizar análisis factorial exploratorio**

Veamos un resumen de los datos

summary(postulantes)

## ID RV RM MAT   
## Length:541 Min. : 0.50 Min. :-1.08 Min. :-1.250   
## Class :character 1st Qu.: 8.75 1st Qu.: 7.85 1st Qu.: 6.660   
## Mode :character Median :10.50 Median :10.71 Median : 9.790   
## Mean :10.27 Mean :10.14 Mean : 9.301   
## 3rd Qu.:12.50 3rd Qu.:12.85 3rd Qu.:12.290   
## Max. :16.75 Max. :18.57 Max. :18.950   
## PSI FIS LOG BIO   
## Min. : 4.18 Min. :-2.150 Min. :-0.90 Min. :-3.930   
## 1st Qu.: 7.53 1st Qu.: 5.350 1st Qu.: 7.69 1st Qu.: 6.420   
## Median :10.31 Median : 8.920 Median :10.17 Median :10.350   
## Mean :10.35 Mean : 8.351 Mean :10.07 Mean : 9.703   
## 3rd Qu.:13.08 3rd Qu.:11.780 3rd Qu.:12.76 3rd Qu.:13.570   
## Max. :16.53 Max. :20.000 Max. :19.37 Max. :20.000   
## HIS QUI   
## Min. :-1.36 Min. :-2.50   
## 1st Qu.: 8.72 1st Qu.: 8.92   
## Median :10.97 Median :12.85   
## Mean :10.65 Mean :11.57   
## 3rd Qu.:13.10 3rd Qu.:15.00   
## Max. :19.42 Max. :20.00

No se tienen datos vacíos por lo cual no es necesario realizar una técnica de imputación, por lo que empezemos omitiendo la primera columna que representa a los códigos de los postulantes

postulantes <- postulantes[,-1]

#### Prueba de esfericidad de Bartlet

library(psych)  
cortest.bartlett(cor(postulantes), n=nrow(postulantes))

## $chisq  
## [1] 1968.098  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 36

Se nos muestra un p valor menor a 0.05, justifica el uso de reducción de datos.

#### Indicador Kaiser-Meyer-Olkinn KMO y MSA

KMO(postulantes)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = postulantes)  
## Overall MSA = 0.78  
## MSA for each item =   
## RV RM MAT PSI FIS LOG BIO HIS QUI   
## 0.61 0.86 0.81 0.47 0.85 0.61 0.86 0.69 0.86

Los valores son menores a 0.5 a excepción de PSI (Psicología) por lo que se puede extraer para que se considere aceptable la aplicación del análisis factorial al conjunto de datos

data\_AFE <- postulantes[,-4]

Veamos nuevamente el test de bartlet y KMO

cortest.bartlett(cor(data\_AFE), n=nrow(data\_AFE))

## $chisq  
## [1] 1866.656  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 28

Significativo, ahora veamos el KMO

KMO(data\_AFE)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = data\_AFE)  
## Overall MSA = 0.79  
## MSA for each item =   
## RV RM MAT FIS LOG BIO HIS QUI   
## 0.61 0.86 0.81 0.85 0.84 0.86 0.69 0.86

Ahora si es justificable el uso de un análisis factorial exploratorio

1. **Decida cuantos factores retener explique el por qué.**

Empezemos realizando la primera seleccionando tomando en cuenta todos los factores

facto=principal(r=data\_AFE,nfactors=8,rotate="none")  
facto

## Principal Components Analysis  
## Call: principal(r = data\_AFE, nfactors = 8, rotate = "none")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 h2 u2 com  
## RV 0.52 0.76 -0.16 -0.18 -0.03 0.03 0.01 0.32 1 0.0e+00 2.4  
## RM 0.66 -0.28 0.26 -0.52 0.36 0.06 -0.11 -0.01 1 1.0e-15 3.4  
## MAT 0.78 -0.35 0.15 -0.17 -0.27 0.00 0.38 0.01 1 1.8e-15 2.4  
## FIS 0.80 -0.29 0.02 0.12 -0.32 0.24 -0.31 0.03 1 1.8e-15 2.2  
## LOG 0.25 0.42 0.83 0.28 0.02 -0.03 0.00 -0.02 1 2.2e-16 2.0  
## BIO 0.74 -0.04 -0.24 0.46 0.33 0.23 0.14 -0.01 1 1.0e-15 2.8  
## HIS 0.66 0.61 -0.22 -0.14 -0.08 -0.04 -0.02 -0.34 1 5.6e-16 2.9  
## QUI 0.81 -0.22 -0.13 0.19 0.05 -0.48 -0.10 0.06 1 1.4e-15 2.1  
##   
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8  
## SS loadings 3.66 1.46 0.92 0.69 0.42 0.35 0.28 0.22  
## Proportion Var 0.46 0.18 0.11 0.09 0.05 0.04 0.03 0.03  
## Cumulative Var 0.46 0.64 0.75 0.84 0.89 0.94 0.97 1.00  
## Proportion Explained 0.46 0.18 0.11 0.09 0.05 0.04 0.03 0.03  
## Cumulative Proportion 0.46 0.64 0.75 0.84 0.89 0.94 0.97 1.00  
##   
## Mean item complexity = 2.5  
## Test of the hypothesis that 8 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0   
## with the empirical chi square 0 with prob < NA   
##   
## Fit based upon off diagonal values = 1

facto$loadings

##   
## Loadings:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8   
## RV 0.515 0.757 -0.160 -0.183 0.316  
## RM 0.664 -0.278 0.258 -0.517 0.363 -0.109   
## MAT 0.782 -0.350 0.146 -0.171 -0.265 0.380   
## FIS 0.804 -0.290 0.123 -0.321 0.235 -0.307   
## LOG 0.250 0.418 0.826 0.280   
## BIO 0.741 -0.242 0.458 0.327 0.234 0.141   
## HIS 0.659 0.612 -0.217 -0.142 -0.341  
## QUI 0.807 -0.221 -0.127 0.194 -0.480   
##   
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8  
## SS loadings 3.662 1.458 0.918 0.691 0.423 0.349 0.279 0.221  
## Proportion Var 0.458 0.182 0.115 0.086 0.053 0.044 0.035 0.028  
## Cumulative Var 0.458 0.640 0.755 0.841 0.894 0.937 0.972 1.000

Por la mayor explicación de la varianza, se recomienda usar 3 o 4 factores. Decidamos el uso de 4 factores

1. **Decida el método de rotación y explique el por qué.**

Teniendo en cuenta que se tendrán 4 factores veamos las cargas factoriales

facto=principal(r=data\_AFE,nfactors=4,rotate="none")  
facto

## Principal Components Analysis  
## Call: principal(r = data\_AFE, nfactors = 4, rotate = "none")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PC1 PC2 PC3 PC4 h2 u2 com  
## RV 0.52 0.76 -0.16 -0.18 0.90 0.1021 2.0  
## RM 0.66 -0.28 0.26 -0.52 0.85 0.1483 2.6  
## MAT 0.78 -0.35 0.15 -0.17 0.79 0.2146 1.6  
## FIS 0.80 -0.29 0.02 0.12 0.75 0.2534 1.3  
## LOG 0.25 0.42 0.83 0.28 1.00 0.0019 2.0  
## BIO 0.74 -0.04 -0.24 0.46 0.82 0.1816 1.9  
## HIS 0.66 0.61 -0.22 -0.14 0.88 0.1239 2.3  
## QUI 0.81 -0.22 -0.13 0.19 0.75 0.2461 1.3  
##   
## PC1 PC2 PC3 PC4  
## SS loadings 3.66 1.46 0.92 0.69  
## Proportion Var 0.46 0.18 0.11 0.09  
## Cumulative Var 0.46 0.64 0.75 0.84  
## Proportion Explained 0.54 0.22 0.14 0.10  
## Cumulative Proportion 0.54 0.76 0.90 1.00  
##   
## Mean item complexity = 1.9  
## Test of the hypothesis that 4 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.06   
## with the empirical chi square 95.01 with prob < 2.3e-21   
##   
## Fit based upon off diagonal values = 0.98

facto$loadings

##   
## Loadings:  
## PC1 PC2 PC3 PC4   
## RV 0.515 0.757 -0.160 -0.183  
## RM 0.664 -0.278 0.258 -0.517  
## MAT 0.782 -0.350 0.146 -0.171  
## FIS 0.804 -0.290 0.123  
## LOG 0.250 0.418 0.826 0.280  
## BIO 0.741 -0.242 0.458  
## HIS 0.659 0.612 -0.217 -0.142  
## QUI 0.807 -0.221 -0.127 0.194  
##   
## PC1 PC2 PC3 PC4  
## SS loadings 3.662 1.458 0.918 0.691  
## Proportion Var 0.458 0.182 0.115 0.086  
## Cumulative Var 0.458 0.640 0.755 0.841

Vemos que las cargas aportan a cada factor siendo posible su distinción entre factores teniendo los siguientes resultados

* PC1: **RM (confuso)**, MAT, FIS, BIO **HIS (confuso)** y QUI
* PC2: RV, **HIS (confuso)**
* PC3: LOG
* PC4: **RM (confuso)** y BIO

Siendo casos en donde no se observan las diferencias de una variable hacia el factor, por lo que es necesario explicar la máxima varianza, veamos con varimax

facto=principal(r=data\_AFE,nfactors=4,rotate="varimax")  
facto

## Principal Components Analysis  
## Call: principal(r = data\_AFE, nfactors = 4, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC1 RC2 RC4 RC3 h2 u2 com  
## RV 0.09 0.93 0.05 0.13 0.90 0.1021 1.1  
## RM 0.15 0.13 0.90 0.04 0.85 0.1483 1.1  
## MAT 0.49 0.06 0.73 0.05 0.79 0.2146 1.8  
## FIS 0.71 0.08 0.49 0.07 0.75 0.2534 1.8  
## LOG 0.03 0.14 0.06 0.99 1.00 0.0019 1.1  
## BIO 0.87 0.24 0.05 0.03 0.82 0.1816 1.2  
## HIS 0.27 0.88 0.15 0.06 0.88 0.1239 1.2  
## QUI 0.77 0.16 0.37 -0.01 0.75 0.2461 1.5  
##   
## RC1 RC2 RC4 RC3  
## SS loadings 2.19 1.78 1.76 1.00  
## Proportion Var 0.27 0.22 0.22 0.13  
## Cumulative Var 0.27 0.50 0.72 0.84  
## Proportion Explained 0.33 0.26 0.26 0.15  
## Cumulative Proportion 0.33 0.59 0.85 1.00  
##   
## Mean item complexity = 1.3  
## Test of the hypothesis that 4 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.06   
## with the empirical chi square 95.01 with prob < 2.3e-21   
##   
## Fit based upon off diagonal values = 0.98

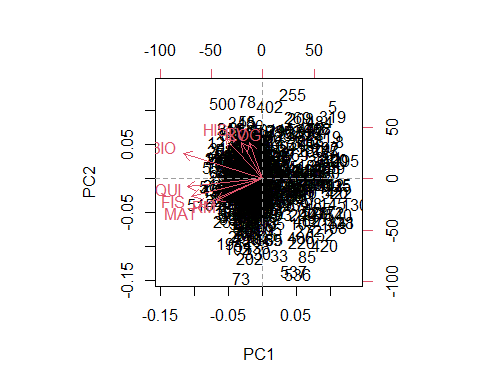
facto$loadings

##   
## Loadings:  
## RC1 RC2 RC4 RC3   
## RV 0.933 0.125  
## RM 0.146 0.127 0.901   
## MAT 0.491 0.734   
## FIS 0.705 0.489   
## LOG 0.142 0.987  
## BIO 0.870 0.239   
## HIS 0.267 0.883 0.145   
## QUI 0.770 0.156 0.369   
##   
## RC1 RC2 RC4 RC3  
## SS loadings 2.191 1.779 1.756 1.003  
## Proportion Var 0.274 0.222 0.220 0.125  
## Cumulative Var 0.274 0.496 0.716 0.841

Veamos ahora los factores, asimismo como su importancia

* RC1: FIS, BIO y QUI
* RC2: RV e HIS
* RC4: RM, MAT
* RC3: LOG

biplot(prcomp(data\_AFE, scale = FALSE))  
abline(h = 0, v = 0, lty = 2, col = 8)



Como se observa hay mejor distinción de las variables con respecto a sus factores asimismo podemos explicar cada factor de la siguiente manera

* **RC1 (Ciencias naturales):** Conformada por Física, biología y química
* **RC2 (Comprensión información):** Conformada por razonamiento verbal e historia
* **RC4 (Ciencias matemáticas):** Conformada por razonamiento matemático y matemáticas
* **RC3 (Lógica):** Conformada por lógica

Concluyendo que la rotación varimax permite una mejor distinción entre factores a través de la explicación de su varianza máxima asimismo como los componentes creados tienen lógica con el contexto del ejercicio.

1. **Explicar los resultados y de sus conclusiones del ejercicio.**

Saquemos ahora los scores

scores <- as.matrix(data\_AFE) %\*% as.matrix(facto$loadings)  
scores <- data.frame(scores) ; head(scores)

## RC1 RC2 RC4 RC3  
## 1 13.376612 26.61663 17.41696 15.688767  
## 2 34.545960 27.36652 23.76605 15.217669  
## 3 38.127554 23.16733 33.66225 6.645127  
## 4 51.435037 34.52837 42.72932 14.062324  
## 5 5.254995 24.61105 0.42748 12.582034  
## 6 23.483688 29.39205 22.55278 15.244419

Realizemos una transformación manteniendo sus características

Zscores<-scale(scores)  
transScore <- Zscores\*100+500 # Proceso de baremación de PISA  
transScore <- data.frame(transScore)

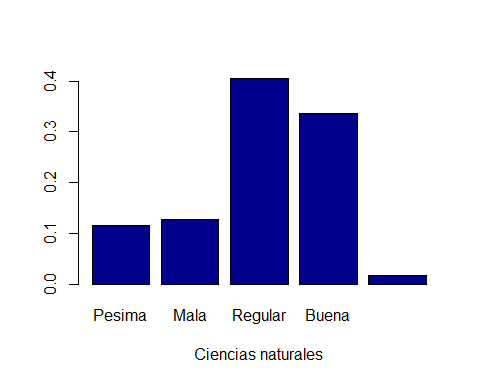
Recodifiquemos para la interpretación

#### RC1 (Ciencias naturales)

transScore$RNC1[transScore$RC1<350] <-1  
transScore$RNC1[transScore$RC1>=350 & transScore$RC1<450] <-2  
transScore$RNC1[transScore$RC1>=450 & transScore$RC1<550] <-3  
transScore$RNC1[transScore$RC1>=550 & transScore$RC1<650] <-4  
transScore$RNC1[transScore$RC1>=650] <-5  
  
  
# Etiquetar  
transScore$RNC1 <- factor(transScore$RNC1,   
 labels = c("Pesima", "Mala", "Regular",  
 "Buena", "Excelente"))  
  
  
fi=table(transScore$RNC1)  
probabilidad=prop.table(table(transScore$RNC1))\*100  
cbind(fi,probabilidad)

## fi probabilidad  
## Pesima 62 11.460259  
## Mala 69 12.754159  
## Regular 219 40.480591  
## Buena 182 33.641405  
## Excelente 9 1.663586

barplot(prop.table(table(transScore$RNC1)), col = "darkBlue", xlab = "Ciencias naturales")



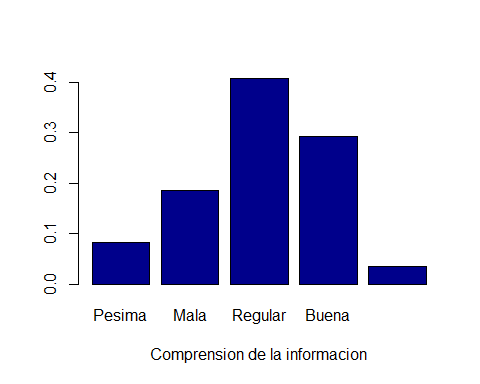
Se observa que la mayor parte se encuentra en un nivel regular y bueno en la parte de ciencias naturales, seguido de malos y pésimos y por último una pequeña parte tiene notas excelentes en las ciencias naturales.

#### RC2 (Comprensión de la información)

transScore$RNC2[transScore$RC2<350] <-1  
transScore$RNC2[transScore$RC2>=350 & transScore$RC2<450] <-2  
transScore$RNC2[transScore$RC2>=450 & transScore$RC2<550] <-3  
transScore$RNC2[transScore$RC2>=550 & transScore$RC2<650] <-4  
transScore$RNC2[transScore$RC2>=650] <-5  
  
  
# Etiquetar  
transScore$RNC2 <- factor(transScore$RNC2,   
 labels = c("Pesima", "Mala", "Regular",  
 "Buena", "Excelente"))  
  
  
fi=table(transScore$RNC2)  
probabilidad=prop.table(table(transScore$RNC2))\*100  
cbind(fi,probabilidad)

## fi probabilidad  
## Pesima 44 8.133087  
## Mala 100 18.484288  
## Regular 220 40.665434  
## Buena 158 29.205176  
## Excelente 19 3.512015

barplot(prop.table(table(transScore$RNC2)), col = "darkBlue", xlab = "Comprension de la informacion")



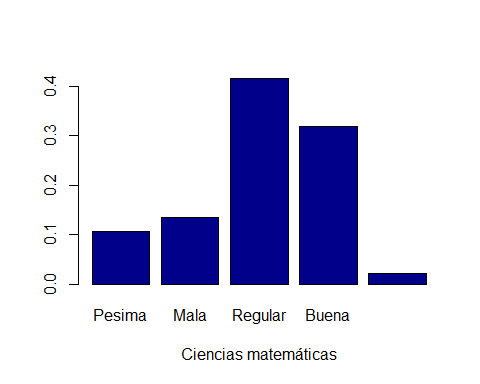
Se observa que respecto a comprension de la información la mayor parte se encuentra en un nivel regular, seguido de bueno, mala, pésima y una pequeña parte en el nivel excelente.

#### RC4 (Ciencias matemáticas)

transScore$RNC4[transScore$RC4<350] <-1  
transScore$RNC4[transScore$RC4>=350 & transScore$RC4<450] <-2  
transScore$RNC4[transScore$RC4>=450 & transScore$RC4<550] <-3  
transScore$RNC4[transScore$RC4>=550 & transScore$RC4<650] <-4  
transScore$RNC4[transScore$RC4>=650] <-5  
  
  
# Etiquetar  
transScore$RNC4 <- factor(transScore$RNC4,   
 labels = c("Pesima", "Mala", "Regular",  
 "Buena", "Excelente"))  
  
  
fi=table(transScore$RNC4)  
probabilidad=prop.table(table(transScore$RNC4))\*100  
cbind(fi,probabilidad)

## fi probabilidad  
## Pesima 58 10.720887  
## Mala 73 13.493530  
## Regular 225 41.589649  
## Buena 173 31.977819  
## Excelente 12 2.218115

barplot(prop.table(table(transScore$RNC4)), col = "darkBlue", xlab = "Ciencias matemáticas")



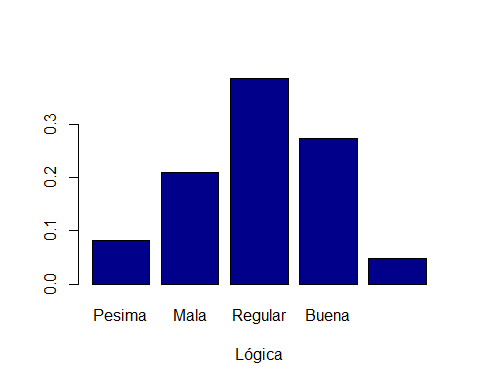
Se observa que la mayor parte se encuentra en un nivel regular y bueno, seguido de malo y pésimo, y una pequeña parte tiene un término excelente con respecto a los resultados enfocados a las ciencias matemáticas.

#### RC3 (Lógica)

transScore$RNC3[transScore$RC3<350] <-1  
transScore$RNC3[transScore$RC3>=350 & transScore$RC3<450] <-2  
transScore$RNC3[transScore$RC3>=450 & transScore$RC3<550] <-3  
transScore$RNC3[transScore$RC3>=550 & transScore$RC3<650] <-4  
transScore$RNC3[transScore$RC3>=650] <-5  
  
  
# Etiquetar  
transScore$RNC3 <- factor(transScore$RNC3,   
 labels = c("Pesima", "Mala", "Regular",  
 "Buena", "Excelente"))  
  
  
fi=table(transScore$RNC3)  
probabilidad=prop.table(table(transScore$RNC3))\*100  
cbind(fi,probabilidad)

## fi probabilidad  
## Pesima 44 8.133087  
## Mala 114 21.072089  
## Regular 209 38.632163  
## Buena 148 27.356747  
## Excelente 26 4.805915

barplot(prop.table(table(transScore$RNC3)), col = "darkBlue", xlab = "Lógica")



Se observa que respecto a lógica la mayor parte se encuentra en un nivel regular, seguido de bueno, mala, pésima y una pequeña parte en el nivel excelente.

## CASO 2: Agrupando clientes mayoristas (7 puntos):

El conjunto de datos se refiere a los clientes de un distribuidor mayorista de Portugal, el cual comercializa distintos tipos de productos.

Cada una de las observaciones hace referencia a un cliente distinto, el cual incluye el gasto anual en unidades monetarias (u.m.) para cada una de las categorías.

Se nos solicita realizar un análisis clúster que nos permita agrupar a nuestros clientes en función de los distintos tipos de productos que adquirieron, para lo cual contamos:

| Variable | Descripción |
| --- | --- |
| Channel | Canal de clientes: 1. Horeca (Hotel/Restaurante/Café) 2. Canal Minorista |
| Región | Región de los clientes: 1. Lisboa, 2. Oporto y 3. Otra |
| Fresh | Gasto anual en productos frescos. |
| Milk | Gasto anual en productos lácteos. |
| Grocery | Gasto anual en productos comestibles. |
| Frozen | Gasto anual en productos congelados. |
| Detergent\_Papers | Gasto anual en detergentes y productos de papel. |
| Delicatessen | Gasto anual en productos preparados (snacks y licor). |

Los datos se encuentran en el archivo “clientes.csv”.

clientes <- read.csv(here("9 Analisis Multivariado/Trabajo 3/clientes.csv"))

Luego de cargar el conjunto de datos en R, realizar las 2 opciones que se presenta:

**Opción 1**:

1. Generar un nuevo dataset solo con las variables numéricas y estandarizarlas.
2. Generar el agrupamiento por particiones utilizando el método kmeans con k=4.
3. Añadir el dataset original la columna cluster, que identificará a los grupos que obtuvimos mediante esta metodología.
4. Graficar y perfilar a nuestros clientes según su agrupación.

**Opción 2**:

1. Generar un nuevo dataset solo con las variables numéricas y estandarizarlas.
2. Encuentre ahora los clusters de forma jerárquica, calculando la matriz de distancias euclidianas y seleccionando en enlace que creas mejor se ajuste a los datos.
3. Comparar los métodos de enlace y determinar cuál es el adecuado.
4. Generar el nuevo agrupamiento jerárquico con el enlace seleccionado.
5. Graficar el dendograma respectivo y determinar el número de clusters.
6. Graficar y perfilar a nuestros clientes según su agrupación jerárquica.

### Solución

**Opción 1**:

1. **Generar un nuevo dataset solo con las variables numéricas y estandarizarlas.**
2. **Generar el agrupamiento por particiones utilizando el método kmeans con k=4.**
3. **Añadir el dataset original la columna cluster, que identificará a los grupos que obtuvimos mediante esta metodología.**
4. **Graficar y perfilar a nuestros clientes según su agrupación.**

**Opción 2**:

1. **Generar un nuevo dataset solo con las variables numéricas y estandarizarlas.**
2. **Encuentre ahora los clusters de forma jerárquica, calculando la matriz de distancias euclidianas y seleccionando en enlace que creas mejor se ajuste a los datos.**
3. **Comparar los métodos de enlace y determinar cuál es el adecuado.**
4. **Generar el nuevo agrupamiento jerárquico con el enlace seleccionado.**
5. **Graficar el dendograma respectivo y determinar el número de clusters.**
6. **Graficar y perfilar a nuestros clientes según su agrupación jerárquica.**

## CASO 3: (7 puntos)

Investigar y realizar un informe monográfico sobre el ***análisis de correspondencia múltiple*** adjuntar un ejercicio aplicando R o Phyton.

### Solución

Aca va la solución